

2. Describing Univariate Distributions

LPO.8800: Statistical Methods for Education Research

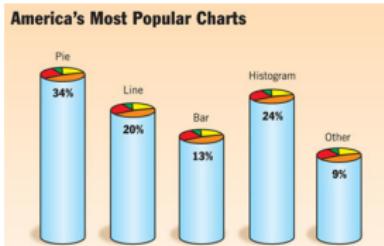
Sean P. Corcoran

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

- Describing univariate distributions: categorical and quantitative data
- Measures of central tendency, variability, skewness, position
- Data transformations



Importance of descriptive analyses

⤤ Prof Dynarski Retweeted

 **Douglas Irwin** @D_A_Irwin · 7h

"The first duty of an economist is to describe correctly what is out there: a valid description without a deeper explanation is worth a thousand times more than a clever explanation of nonexistent facts." - Paul Samuelson

 **Ryan Edwards** @ryanbedw... · 1d

Unpopular opinion: humble descriptive analysis is often much better and more helpful than second rate causal work. Journals that often get more of the latter should appre...
[Show this thread](#)

3 53 247

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

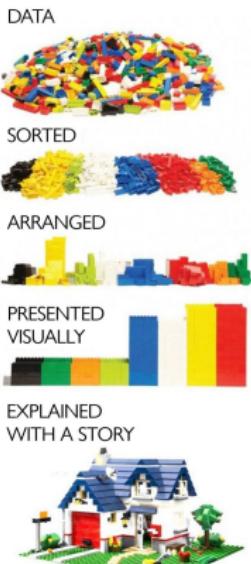
- Descriptive analysis characterizes the world or a phenomenon—answering questions about who, what, where, when, and to what extent. Whether the goal is to identify and describe trends and variation in populations, create new measures of key phenomena, or describe samples in studies aimed at identifying 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



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	

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

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

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).

```
. table parmar18
```

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

The `table` command was substantially overhauled in Stata 17, with lots of options for customizing format and cell contents.

Frequency distributions

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

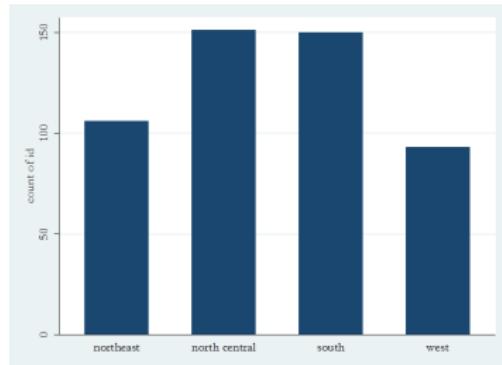
- **mutually exclusive:** being in one category precludes being in another
- **collectively exhaustive:** all possible categories are represented by the defined categories

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.
 - ▶ (count) for frequency
 - ▶ (percent) for relative frequency
- Try: `region, parmarl8, advmath8`

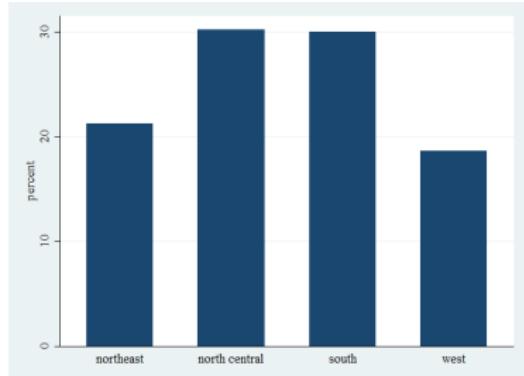
Bar graphs

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



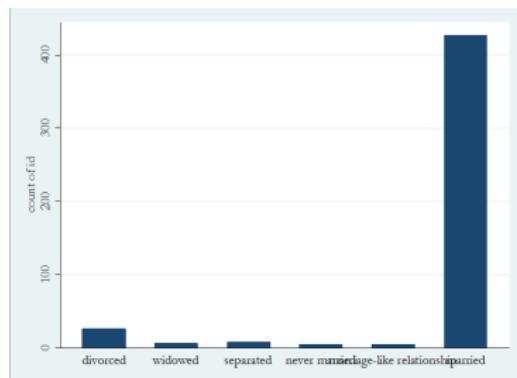
Bar graphs

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



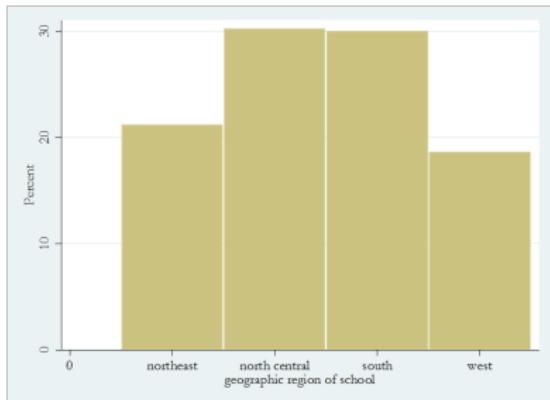
Bar graphs

```
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

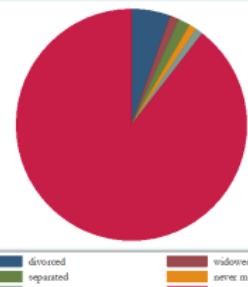
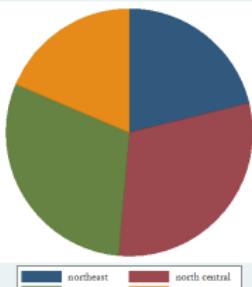
Pie graphs

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

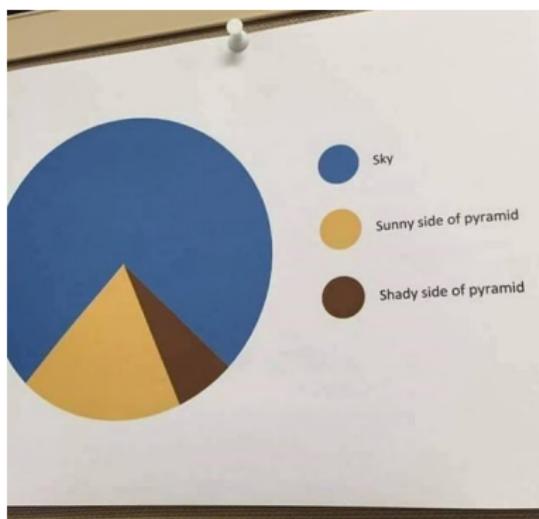
`graph pie, over(region)`



Pie graphs



Pie graphs



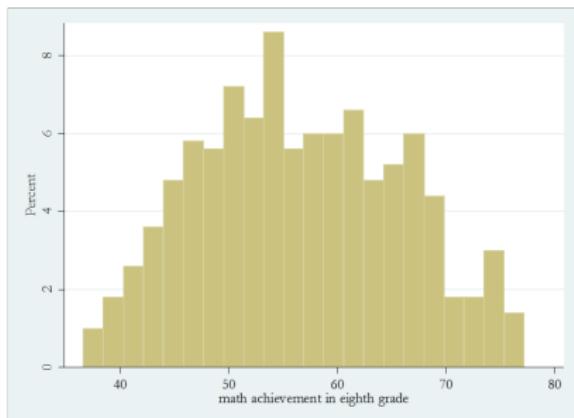
Histograms

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`

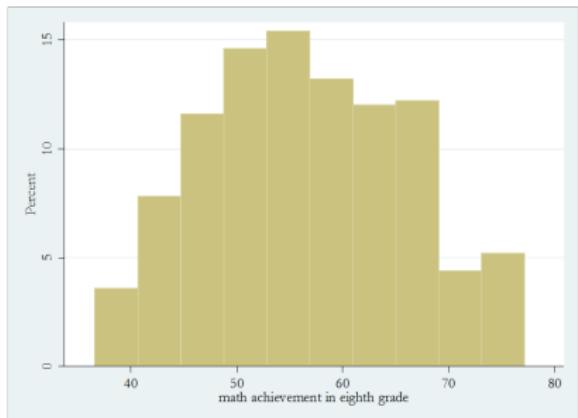
Histograms

`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*	023
40*	0225555799
41*	34689
42*	0335577
43*	0000133445689
44*	00001233344677788
45*	334467778899
46*	3445555667788889
47*	00223444566778888
48*	00000133667889
49*	00000133667889
50*	000112356677888899
51*	000011123333445566889
52*	11111222333445566889
53*	000022456677889
54*	000000123334455577888899
55*	0000000012333445668899
56*	11223346777888889
57*	000011223334444778
58*	00113334677888899999
59*	03444666666666666666
60*	0000123334456677888
61*	000112333345689
62*	000011123334555889
63*	222367777779
64*	01125988888899
65*	22345677889
66*	0000122333344667799
67*	122333456677889
68*	00222346666899
69*	01122446779
70*	0011899
71*	389
72*	146
73*	3335788
74*	02233334
75*	00014
76*	222222

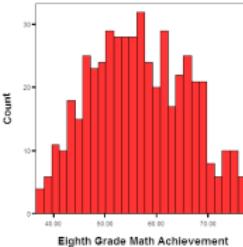
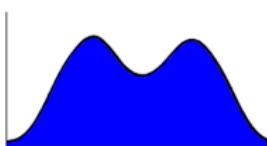
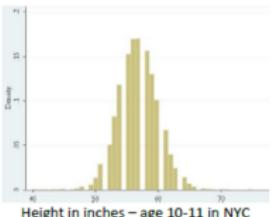
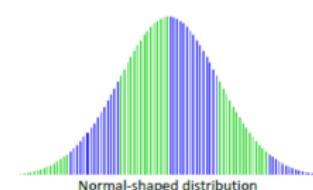
Stem-and-leaf plot

- 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

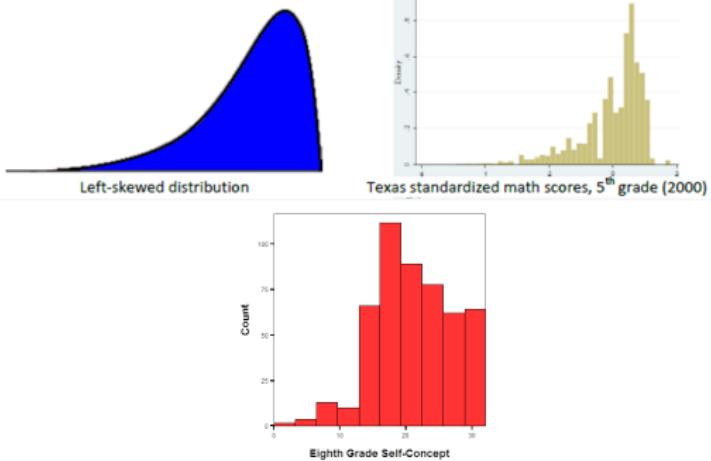
Shape of distributions



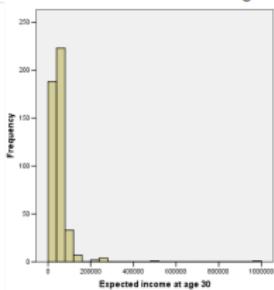
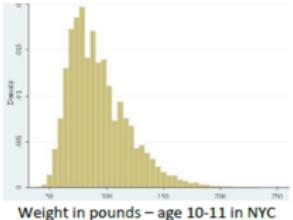
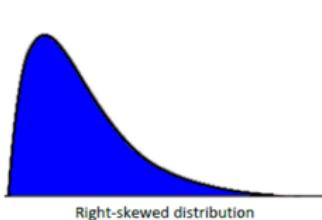
Shape of distributions

- 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



Right-skewed distributions



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
- Later: 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

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

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, 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, 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, 4, 5, 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:

math achievement in eighth grade					
	Percentiles	Smallest	Obs	Sum of wgt.	
1%	38.55	36.61			
5%	41.89	37.14			
10%	44.185	37.2			
25%	49.42	37.24			
50%	56.18				
		Largest	Mean	56.59102	
75%	63.74	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, 4, 5, 5, 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). See also the user-written command modes.

Mode

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

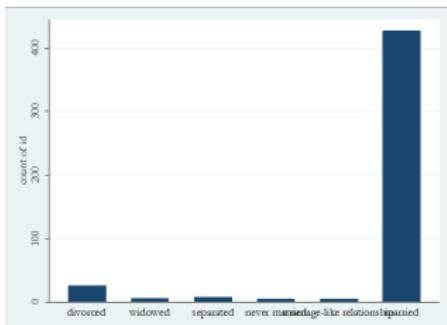
family size	Freq.	Percent	Cum.
4	199	39.80	39.80
5	142	28.40	68.20
6	55	11.00	79.20
3	52	10.40	89.60
7	21	4.20	93.80
9	13	2.60	96.40
2	9	1.80	98.20
8	9	1.80	100.00
Total	500	100.00	

```
. ** 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

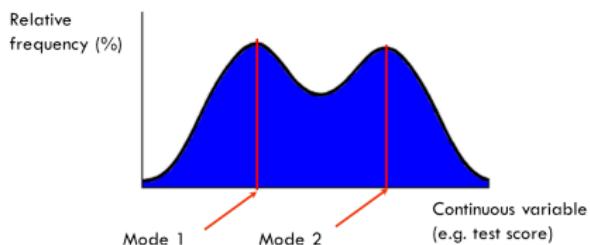
North Central is (technically) the modal region. But region here is more accurately described as *bimodal*—it has a distribution with two values that occur most often (North Central and South).

- . 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:



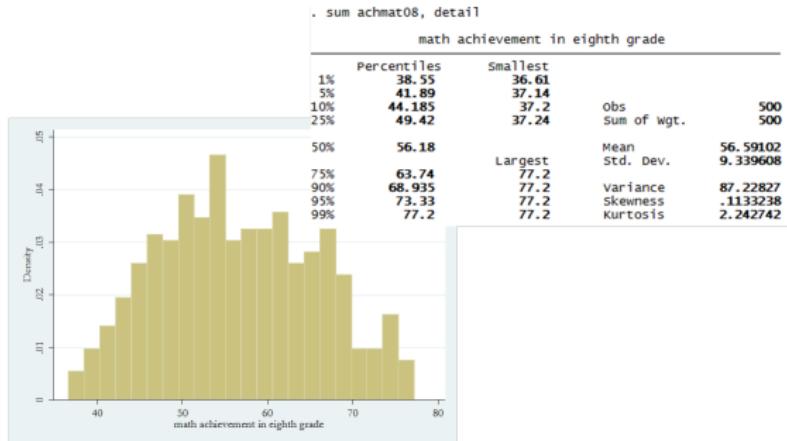
Comparing measures of central tendency

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.



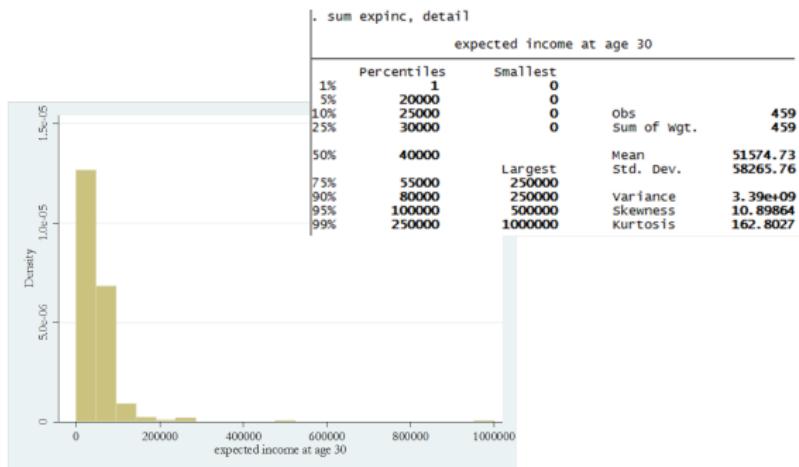
Comparing measures of central tendency

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

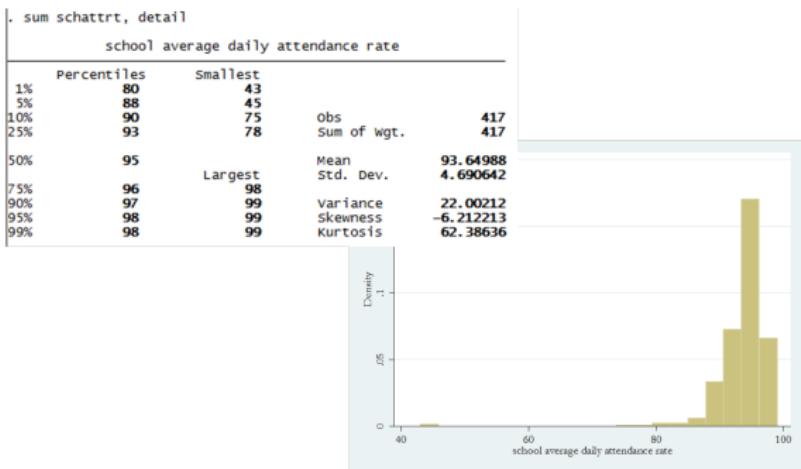
Comparing measures of central tendency

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

		stats		
		achrdg08	achrdg10	achrdg12
mean	mean	56.04906	56.11404	55.60188
	p50	56.445	57.545	57.005
	N	500	500	500

variable	mean	p50	N
achrdg08	56.04906	56.445	500
achrdg10	56.11404	57.545	500
achrdg12	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:

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

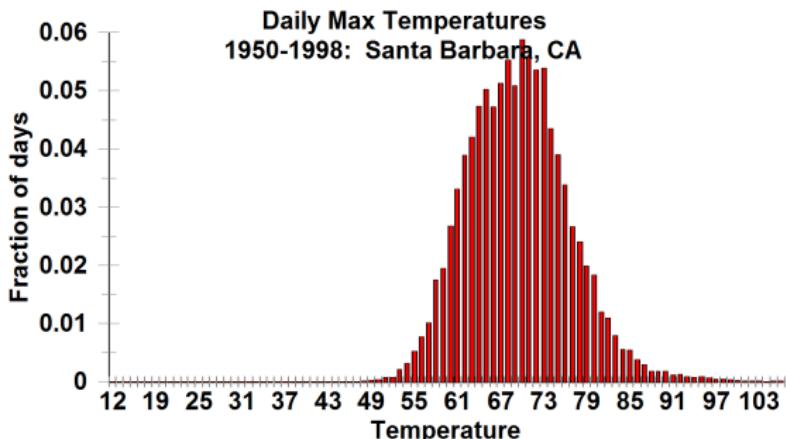
Measures of variability

Measures of **variability** are intended to capture how “spread out” the data points of the distribution are.

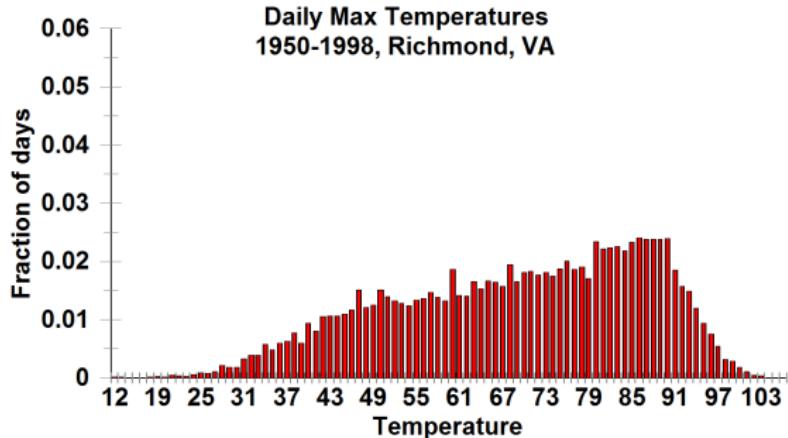
Two distributions can have the same measure of central tendency but very different dispersion. Example: consider mean daily high temperatures for two cities over 49 years:

- Richmond, VA: 69.0
- Santa Barbara: 69.5

Measures of variability



Measures of variability



Measures of variability

Common measures of variability, dispersion, or “spread” include:

- Range
- Variance
- Standard deviation
- Interquartile range (IQR)

Range

The **range** is the difference between the largest and smallest observed values in a distribution

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

For achmat08: $77.2 - 36.61 = 40.59$

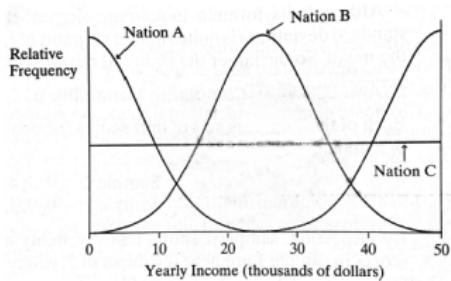
Range

The range is clearly sensitive to extreme values, since it is defined by the largest and smallest values.

expected income at age 30				
Percentiles		Smallest		
1%	1	0		
5%	20000	0		
10%	25000	0	Obs	459
25%	30000	0	Sum of wgt.	459
50%	40000		Mean	51574.73
		Largest	Std. Dev.	58265.76
75%	55000	250000	Variance	3.39e+09
90%	80000	250000	Skewness	10.89864
95%	100000	500000	Kurtosis	162.8027
99%	250000	1000000		

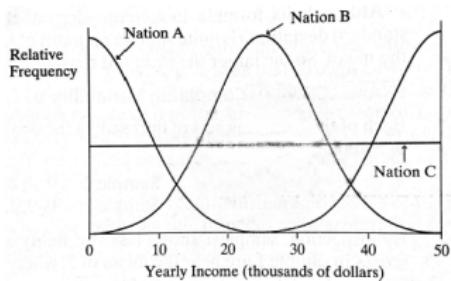
Range

Another disadvantage of the range is that it ignores all of the data beyond the maximum and minimum. In the following figure, the three income distributions have the same range (and mean):



Range

- Nation B: incomes tend to be close to the mean (low variability)
- Nation A: incomes tend to be far from the mean (high variability)



Variance

A more meaningful measure of variability would reflect distance from *every* data point to the mean. Consider the “average deviation from the mean”:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})$$

Variance

Table 2.3: Deviations above and below the mean

	W_i	$(W_i - \bar{W})$	Deviation 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

Variance

- Problem: the sum of all deviations from the mean *will always be zero*—the negative deviations from the mean will cancel out the positive deviations.

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \sum x_i - \frac{1}{n} n\bar{x} = \bar{x} - \bar{x} = 0$$

- Recall characterization of the mean as a “center of gravity,” where deviations above the mean are exactly balanced by deviations below the mean.

Variance

An alternative: the “average squared deviation from the mean,” or variance (s^2):

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- the numerator is the **sum of squares**
- the denominator is $n - 1$ rather than 1 (explained later)

Variance

Table 2.4: Variance and standard deviation of A.L. wins

	W_i	$(W_i - \bar{W})$	$(W_i - \bar{W})^2$
Tampa Bay	62	10	100
Boston	62	10	100
Baltimore	58	6	36
NY Yankees	54	2	4
Toronto	47	-5	25
Detroit	58	6	36
Cleveland	55	3	9
Kansas City	50	-2	4
Minnesota	45	-7	49
Chi White Sox	40	-12	144
Oakland	61	9	81
Texas	56	4	16
Seattle	49	-3	9
LA Angels	48	-4	16
Houston	35	-17	289
SUM	-	0	918
Variance (s^2)	-	-	$(918/14)=65.57$
Std. Dev (s)	-	-	$\sqrt{65.57} = 8.1$

Note: $\bar{W} = 52$

Variance

- A variance of zero means there is no variation in x
- It is hard to interpret the variance in isolation, but it could be used to compare two distributions of a similar measure
- The magnitude of s^2 depends on the units of measurement. E.g.: variance of income in *cents* will be higher than the variance of income in *dollars*
- s^2 is generally hard to interpret (it is the average “squared deviation from the mean”)

Variance: why $n - 1$?

Why do we divide by $n - 1$ when calculating the variance, rather than n ?

- Typically we are *estimating* a population variance from sample data.
- The variance formula included \bar{x} which is *also estimated* and may vary from sample to sample, in ways related to the variance.
- Dividing by n systematically *under-estimates* the variance by a factor of $(n - 1)/n$. It is *biased*.
- $n - 1$ is sometimes called the *degrees of freedom* since the sample mean \bar{x} provides one piece of information about the sample.
- If you are calculating variance for a *population*, dividing by n is appropriate.
- It is common practice just to divide by $n - 1$ in all cases. In large samples, it makes little difference.

For more see: <https://towardsdatascience.com/why-sample-variance-is-divided-by-n-1-89821b83ef6d>

Standard deviation

The **standard deviation** (s) restores the variance measure to units of the original variable:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

- If it helps, think of this as the “average” or “typical” deviation of x from its mean (though this is not strictly correct).
- s^2 and s are always positive, unless there is no variance in x (in which case they are both zero)

Standard deviation

Table 2.4: Variance and standard deviation of A.L. wins

	W_i	$(W_i - \bar{W})$	$(W_i - \bar{W})^2$
Tampa Bay	62	10	100
Boston	62	10	100
Baltimore	58	6	36
NY Yankees	54	2	4
Toronto	47	-5	25
Detroit	58	6	36
Cleveland	55	3	9
Kansas City	50	-2	4
Minnesota	45	-7	49
Chi White Sox	40	-12	144
Oakland	61	9	81
Texas	56	4	16
Seattle	49	-3	9
LA Angels	48	-4	16
Houston	35	-17	289
SUM	-	0	918
Variance (s^2)	-	-	$(918/14)=65.57$
Std. Dev (s)	-	-	$\sqrt{65.57} = 8.1$

Note: $\bar{W} = 52$

Standard deviation

The variance and standard deviation are obtained using `summarize` with the `detail` option:

```
. sum achmat08, detail  
math achievement in eighth grade
```

	Percentiles	Smallest	obs	Sum of wgt.
1%	38.55	36.61		
5%	41.89	37.14		
10%	44.185	37.2	500	500
25%	49.42	37.24		
50%	56.18		Mean	56.59102
		Largest	Std. Dev.	9.339608
75%	63.74	77.2		
90%	68.935	77.2	Variance	87.22827
95%	73.33	77.2	Skewness	.1133238
99%	77.2	77.2	Kurtosis	2.242742

Can use `summarize` alone for standard deviation. Also, `tabstat` with `stat(sd)`

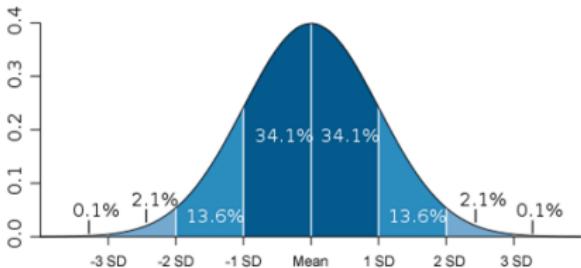
Standard deviation

expected income at age 30				
Percentiles	Smallest			
1%	1	0		
5%	20000	0		
10%	25000	0	Obs	459
25%	30000	0	Sum of wgt.	459
50%	40000		Mean	51574.73
75%	55000	250000	Std. Dev.	58265.76
90%	80000	250000	Variance	3.39e+09
95%	100000	500000	Skewness	10.89864
99%	250000	1000000	Kurtosis	162.8027

Standard deviation

- The variance and standard deviation are especially sensitive to outliers/extreme values, since the variance formula *squares* deviations from the mean
- With **normal** (bell-shaped) distributions, s provides a good rule of thumb (the "Empirical Rule"):
 - ▶ About 68% of observations lie within 1 s.d. of the mean
 - ▶ About 95% of observations lie within 2 s.d. of the mean
 - ▶ Nearly all (99%) lie within 3 s.d. of the mean
 - ▶ Example: IQ mean of 100, standard deviation of 15

Standard deviation



Coefficient of variation

The **coefficient of variation (CV)** expresses the standard deviation as a percentage of the mean:

$$CV = \frac{s}{\bar{x}} * 100$$

- Must have $\bar{x} \neq 0$. If $\bar{x} < 0$, can use the absolute value of \bar{x} .
- The CV adjusts for the scale of units used, allowing for more appropriate comparisons. For example, the CV is often used to measure inequality in school spending across states or districts. Comparing s across these groups would be misleading if, say, the *level* of spending was higher in some districts/states than others.

Choosing a measure of variation

- Choice of dispersion measure should be made carefully—be aware of extreme values / outliers that may influence or distort the picture
 - ▶ The IQR is a “robust” alternative (shown later)
- One possibility when outliers exist: calculate statistics with outliers *dropped* (but be forthcoming about this—e.g. discuss both sets of results—and understand why outliers exist)

Why not use mean absolute deviation?

You may be asking why we square deviations from the mean (variance) and then take the square root (standard deviation) when we could work with absolute values:

$$MAD = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$$

You can do this, or (instead) calculate mean deviations from the *median*. (In Stata, the egen functions `mdev` and `mad` provide the mean deviation from the mean and median, respectively.)

These are rarely used, in part because the standard deviation has nice mathematical and statistical properties. For example, the `sd` is a defining feature of the normal distribution, while the `MAD` is not. But there is a “robust” debate over the value of the `MAD`.

Skewness statistic

The **skewness statistic** is a summary of the positive or negative skewness of a distribution:

$$G = \frac{n}{(n-1)(n-2)} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{s^3}$$

- Notice this is based on *cubed* deviations from the mean (can be positive or negative).
- Positive for positively skewed distributions and negative for negatively skewed distributions (zero for symmetric)

Skewness statistic

The **standard error of the skewness** is:

$$\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}}$$

- The concept of a standard error is introduced later
- For now: if the skewness statistic is more than 2x the standard error of the skewness, we can say the distribution is “significantly skewed”
- Can calculate yourself in Stata using `display`, or install the user-written `summskew` command (see Github for ado file).

Skewness statistic

The skewness statistic is reported using `summarize with detail`:

math achievement in eighth grade					
Percentiles		Smallest		Obs	Sum of wgt.
1%	38.55	36.61			
5%	41.89	37.14			
10%	44.185	37.2			
25%	49.42	37.24			
50%	56.18		Mean	56.59102	
75%	63.74	Largest	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	

- `achmat08` has a slight positive skew
- Std err of skewness not reported here by Stata, but display $\text{sqrt}((6*500*499)/(498*501*503))$ yields 0.109. The skewness statistic is *not* more than twice the std err of the skewness

Quantiles

Quantiles are cutpoints that divide a distribution into continuous intervals of equal size. Common quantiles include:

- **Quartiles:** 4 equally sized groups
- **Deciles:** 10 equally sized groups
- **Terciles:** 3 equally sized groups
- **Percentiles:** 100 equally sized groups

If there are n observations in a dataset, *quartiles* will include $n/4$ observations in each group. Note the term “quantile” is often used to refer to both the cutpoint and the groups themselves.

Quantiles

In real datasets, there may not be the right number of data points to divide into equally sized quantiles. For example, if $n = 10$, you can't divide the 10 observations into 4 equally sized groups. So what rule do you use?

If dividing the data into q groups, and you want to find the k th quantile ($k \leq q$), find the observation ranked $n * k / q$ when ranking the data points from smallest to largest.

- If this is an *fractional* value, take the next-largest ranked observation.
- If this is an *integer* value, take the midpoint of that ranked observation and the next-largest one.

Quantiles

Example: suppose $n = 74$

- 5th percentile is the $74 * (5/100) = 3.7$ th obs (round to 4)
- 10th percentile is the $74 * (10/100) = 7.4$ th obs (round to 8)
- 25th percentile is the $74 * (25/100) = 18.5$ th obs (round to 19)
- Median is the $74 * (50/100) = 37$ th obs (take avg of 37th and 38th)

This operationalizes the p th percentile as the smallest value that is greater than $p\%$ of the observations.

Percentiles

Can sometimes determine percentiles from cumulative frequency distributions:

```
. tabulate famsize
```

family size	Freq.	Percent	Cum.
2	9	1.80	1.80
3	52	10.40	12.20
4	199	39.80	52.00
5	142	28.40	80.40
6	55	11.00	91.40
7	21	4.20	95.60
8	9	1.80	97.40
9	13	2.60	100.00
Total	500	100.00	

e.g., 10th percentile is $500*(10/100) = 50$ th obs (take avg of 50th and 51st) = 3

Percentiles

For quantitative variables, can use `summarize` with detail option in Stata to get common percentiles:

```
. sum achrdg8, detail  
reading achievement in eighth grade
```

	Percentiles	Smallest		
1%	37.525	35.74		
5%	40.705	35.82		
10%	43.71	37.17	Obs	500
25%	49.975	37.29	Sum of wgt.	500
50%	56.445		Mean	56.04906
		Largest	Std. Dev.	8.829726
75%	63.195	70.55	Variance	77.96406
90%	69.15	70.55	Skewness	-1.1485071
95%	70.55	70.55	Kurtosis	2.191284
99%	70.55	70.55		

Alternative 1: `tabstat` with `stat(p25 p50 etc)`.

Alternative 2: create a variable that contains a percentile of `varname` using `egen`. E.g., `egen varnamep10=pctile(varname), p(10)`

Percentiles

A third alternative uses `centile varname, centile(p)`.

Note, however, that `centile` uses a linear interpolation rather than the “next highest rank” method. This will often provide a slightly different answer.

- `centile` will calculate the rank $P = (n + 1)p/100$ where p is the desired percentile.
- If P is fractional, `centile` will interpolate between the two observations ranked $\text{int}(P)$ and $\text{int}(P) + 1$.
- If P is an integer, it will use that ranked observation for the p th percentile.

Percentiles

The **five number summary** is a report of the minimum, Q1, median, Q3, and maximum. In Stata, use `summarize`, `detail` or `tabstat`:

```
. tabstat achrdg08 achmat08 achsci08 achscls08, stat(min p25 p50 p75 max)
```

stats	achrdg08	achmat08	achsci08	achscls08
min	35.74	36.61	34.94	29.2
p25	49.975	49.42	48.69	49.39
p50	56.445	56.18	55.4	54.76
p75	63.195	63.74	62.19	61.2
max	70.55	77.2	80.01	76.7

Percentiles

The p th percentile is sometimes defined as the value that is greater than or equal to $p\%$ of the observations (vs. “greater than”).

This can make a difference in some datasets. In the limit—with continuous distributions—there will be no difference in these definitions.

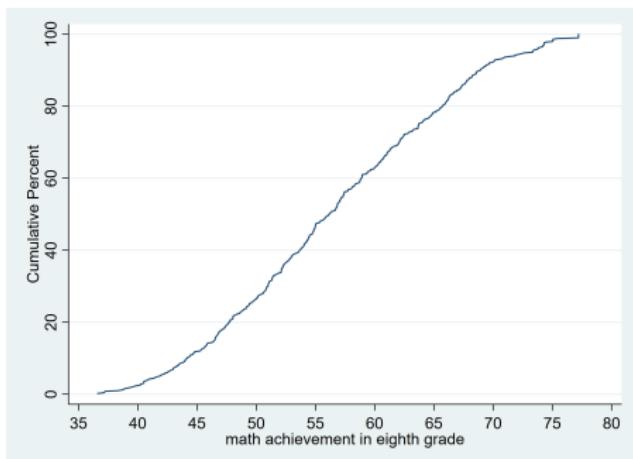
Empirical CDF (“ogive”)

- A graph showing percentiles on the vertical axis and values of x on the horizontal axis is an **ogive**, or **empirical CDF** (cumulative distribution function)
- This graph in Stata requires the extra step of creating a new variable with the cumulative relative frequency, using `cumul`:

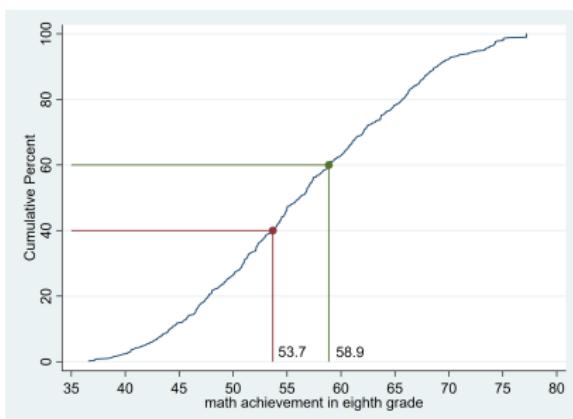
```
▶ cumul achmat08, gen(cumachmat08)
▶ replace cumachmat08=cumachmat08*100
▶ sort cumachmat08
▶ twoway line cumachmat08 achmat08, ytitle("Cumulative Percent")
```

- Alternative: `ssc install distplot`

Empirical CDF (“ogive”)



Empirical CDF (“ogive”)



- The 40th percentile is approximately 53.7
- The 60th percentile is approximately 58.9

Z-scores

Percentiles are a measure of *position* in a distribution. Another is the **z-score** (or standardized score). The z-score for a particular value of x (x_i) is defined as:

$$z_i = \frac{x_i - \bar{x}}{s}$$

(Details later in this lecture).

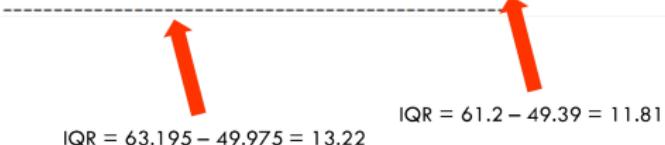
Interquartile range

- The **interquartile range** is the difference between the upper and lower quartiles (75th and 25th percentiles)—a measure of dispersion that is robust to extreme values/outliers
- Pull quartiles from `summarize`, `detail` or `tabstat`, or use the `centile` command.
- You can also use the `iqr` stat option in `tabstat`

Interquartile range

```
. tabstat achrdg08 achmat08 achsci08 achscls08, stat(min p25 p50 p75 max)
```

stats	achrdg08	achmat08	achsci08	achscls08
min	35.74	36.61	34.94	29.2
p25	49.975	49.42	48.69	49.39
p50	56.445	56.18	55.4	54.76
p75	63.195	63.74	62.19	61.2
max	70.55	77.2	80.01	76.7



Interquartile range

Compare the IQR (\$25,000) to s (\$58,265)

```
. sum expinc30, detail
```

expected income at age 30

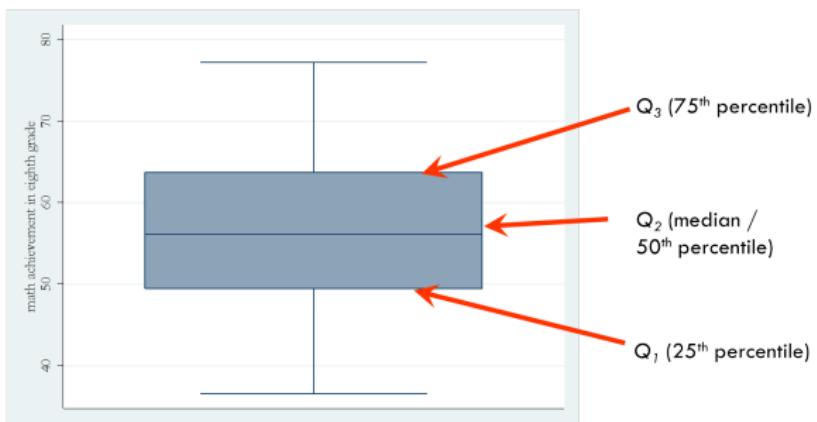
	Percentiles	smallest		
1%	1	0		
5%	20000	0		
10%	25000	0	Obs	459
25%	30000	0	Sum of Wgt.	459
50%	40000		Mean	51574.73
		Largest	Std. Dev.	58265.76
75%	55000	250000	Variance	3.39e+09
90%	80000	250000	Skewness	10.89864
95%	100000	500000	Kurtosis	162.8027
99%	250000	1000000		

Box plots

A **boxplot** (or box-and-whiskers plot) shows features of a variable's distribution: median, Q3, Q1, and extreme values (tails)

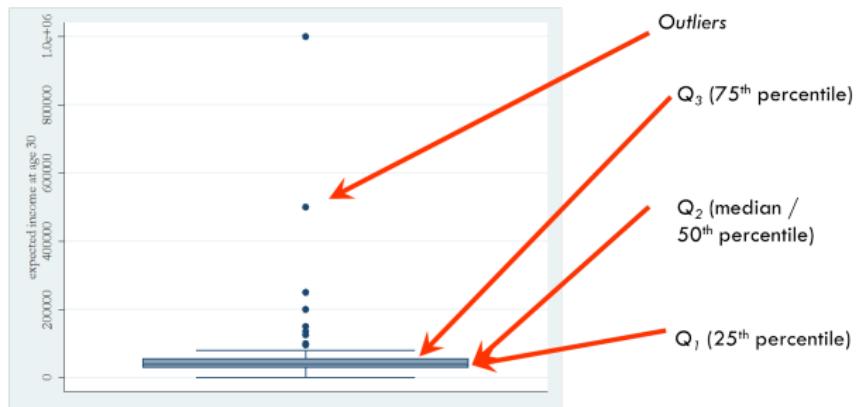
- The tails are "whiskers" that extend to the maximum and minimum in the data *if* there are no outliers (defined here as 1.5 IQR lengths above Q3 or below Q1).
- If there are outliers, the whiskers extend to the observation closest to (but not beyond) these thresholds (the highest and lowest **adjacent value**). Outliers are shown as points or asterisks (*).
- In Stata: `graph box varname`

Box plots



- Median=56.18, $Q_1=49.42$, $Q_3=63.74$, $IQR=14.32$
- Outliers would be above $(63.74 + (1.5 * 14.32)) = 85.22$ or below $(49.42 - (1.5 * 14.32)) = 27.94$

Box plots



- Outliers would be above $(55,000 + (1.5 * 25,000)) = \$92,500$ or below $(55,000 - (1.5 * 25,000)) = -\$7,500$
- Stata can suppress the outliers with the `nooutsides` option

Transforming variables

Transforming a variable simply means expressing it another way. Useful in many circumstances:

- Changing units of measure (e.g. feet to inches)
- Putting an observed value into the context of its distribution (e.g., "is 160 a high LSAT score?")
- Reduce the skewness of a variable—to work with a more symmetric variable
- Compare distributions on different scales
- Combining multiple variables into one index

Transforming variables

The z-score was one example of a transformation: from the units of the original scale to “standard deviation units”

- Example: Boston Red Sox have 62 wins, where the mean in the A.L. is 52. Their z-score is $(62 - 52)/8.1 = 1.235$. Their number of wins is $z = 1.235$ standard deviations above the mean.

Transforming variables

- Transformations can be linear or nonlinear
- Monotonic transformations preserve the *order* of the data points
- Non-monotonic transformations fail to preserve the order of the data points. Example: x^2 , winsorizing
- We will only consider the monotonic transformations here.

Transforming variables

A **linear transformation** involves only addition, subtraction, multiplication, or division, applied to the original variable x_0 :

$$x_1 = a + (b * x_0)$$

- a can be positive (addition) or negative (subtraction)
- $b \neq 0$, but can be negative and $|b|$ can be < 1 (i.e., division)

Transforming variables

In Stata, `generate` (or `gen`) is the all-purpose command for creating new variables. Example: expected income, in thousands of dollars

```
. gen expinc30b = (expinc30 / 1000)  
(41 missing values generated)
```

```
. sum expinc*
```

Variable	Obs	Mean	Std. Dev.	Min	Max
expinc30	459	51574.73	58265.76	0	1000000
expinc30b	459	51.57473	58.26576	0	1000

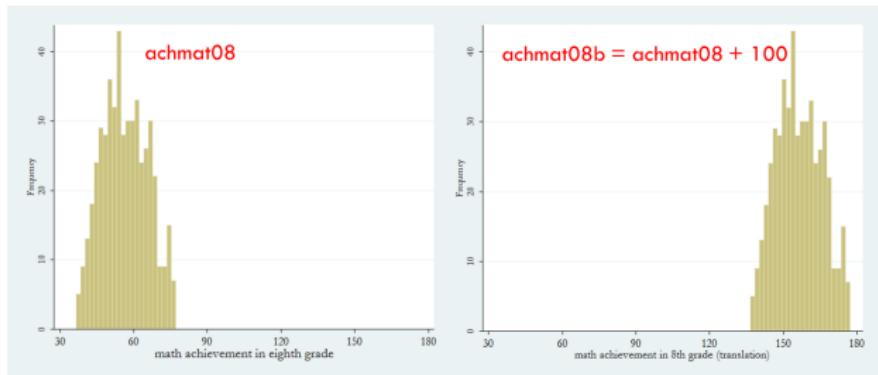
Translation

The simplest linear transformation is a **translation** where $b = 1$ and a is positive or negative:

$$x_1 = a + x_0$$

Shifts the distribution to the right or left. E.g. $achmat08r = achmat08 + 100$

Translation

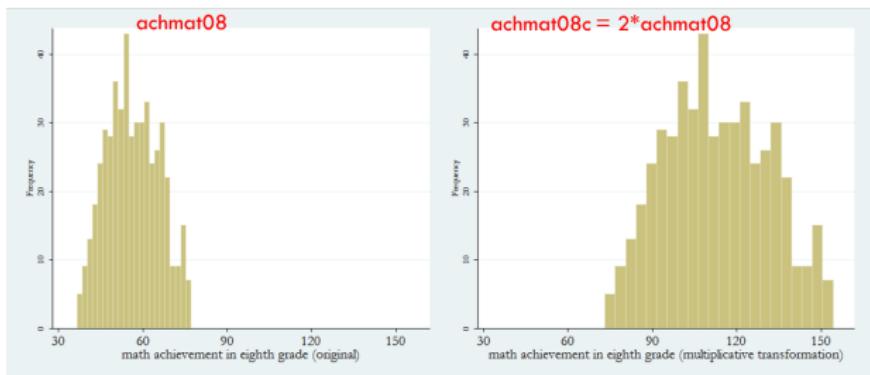


Multiplicative transformation

- The translation did not affect the *shape* or *spread* (variability) of the distribution, but did affect its position, shifting it to the left or right.
- Now consider a **multiplicative** transformation, where $a = 0$ but $b \neq 0$ and $b \neq 1$:

$$x_1 = (b * x_0)$$

Multiplicative transformation

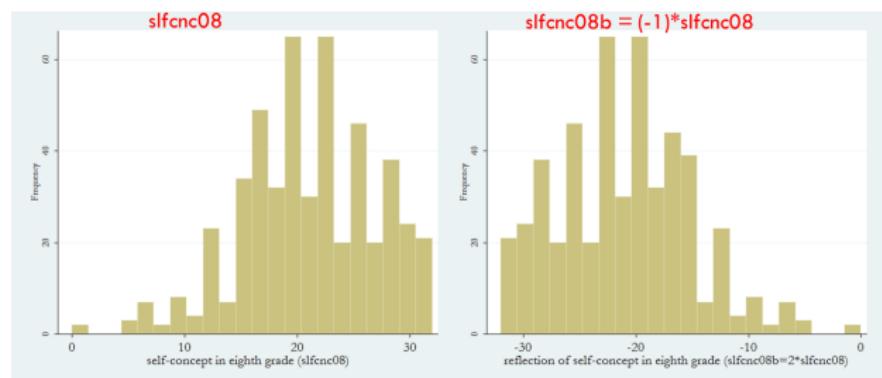


Multiplicative transformation

- The multiplicative transformation has affected *both* the location of the distribution *and* its spread.
 - ▶ The distribution is “stretched” whenever $|b| > 1$
 - ▶ The distribution is “compressed” whenever $|b| < 1$
- While the spread is affected, *relative distance between points is preserved*. *Shape* (symmetry, skewness) is preserved unless $b < 0$, in which case the distribution is flipped to its mirror image. A **reflection** multiplies the original variable by -1 .

Multiplicative transformation

Reflection:



Transforming variables

How will a linear transformation of a variable affect its:

- Mean?
- Variance?
- Standard deviation?
- Range?
- IQR?
- Skewness?

Transforming variables

How a linear transformation affects the mean (\bar{x}_{new} vs. \bar{x}_{old}):

$$\bar{x}_{old} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\bar{x}_{new} = \frac{1}{n} \sum_{i=1}^n (a + bx_i)$$

$$\bar{x}_{new} = \frac{1}{n} \sum_{i=1}^n a + \frac{1}{n} \sum_{i=1}^n bx_i$$

$$\bar{x}_{new} = a + b\bar{x}_{old}$$

Transforming variables

In other words, the mean of the new variable (\bar{x}_{new}) is equal to the old mean (\bar{x}_{old}) multiplied by b , and increased (or decreased) by a . Also true for median, mode, and percentiles.

Transforming variables

Example: use a linear transformation to express Fahrenheit temperature in Celsius.

- The conversion from F to C is: $x_C = -17.78 + 0.556 * x_F$ (a linear transformation)
- The mean daily high in Santa Barbara (in Fahrenheit) is $\bar{x}_F = 69.5$
- So then: $\bar{x}_C = -17.78 + 0.556 * 69.5 = 20.86$

Transforming variables

How a linear transformation affects the variance (s_{new}^2 vs. s_{old}^2):

$$s_{old}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s_{new}^2 = \frac{1}{n-1} \sum_{i=1}^n (a + bx_i - (a + b\bar{x}_{old}))^2$$

$$s_{new}^2 = \frac{1}{n-1} \sum_{i=1}^n (bx_i - b\bar{x}_{old})^2$$

Transforming variables

Continued:

$$s_{new}^2 = \frac{1}{n-1} \sum_{i=1}^n b^2 (x_i - \bar{x}_{old})^2$$

$$s_{new}^2 = b^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_{old})^2$$

$$s_{new}^2 = b^2 s_{old}^2$$

Transforming variables

- In other words, the variance of the transformed variable (s_{new}^2) is equal to the variance of the original variable (s_{old}^2), multiplied by b^2 .

- The standard deviation of the transformed variable is:

$$s_{new} = \sqrt{s_{new}^2} = \sqrt{b^2 s_{old}^2} = |b| s_{old}$$

- Notice a does not affect the variance or standard deviation.
- The range and IQR of the transformed variable are equal to the range and IQR of the original variable, multiplied by $|b|$

Transforming variables

The skewness statistic is unaffected by a linear transformation *unless b* is negative. Then the skewness of the transformed variable is $(-1) * G_{old}$

Standardized variables

Standardized variables are an example of a linear transformation

In general a standardized variable applies a linear transformation to a variable to give it a predefined mean (c) and standard deviation ($v > 0$)

	Variable	Mean
Original	x	\bar{x}
Transformed	$x - \bar{x} + c$	c

	Variable	Standard deviation
Original	x	s
Transformed	$\frac{v}{s}x$	v

Standardized variables

Examples of standardized variables:

- SAT sections: mean 500, sd 100
- GRE sections: 130-170, mean 150, sd 8
- IQ score: mean 100, sd 15
- MCAT section: mean 125, sd 3
- z-score: mean 0, sd 1

Z-score transformation

It is easy to go from standard score to original scale score, given the mean and sd:

$$z = \frac{x - \bar{x}}{s}$$

$$x = \bar{x} + sz$$

Example:

- *achmat08*: mean 56.59 and sd 9.34
- $z = 0.89$ would be $x = 56.59 + 9.34(0.89) = 64.9$

Z-score transformation

There are several ways to obtain z-scores in Stata. Manually, can obtain mean and sd, and then create the z-score using *gen*

```
. sum expinc30
      Variable   Obs    Mean    Std. Dev.     Min     Max
      expinc30   459  51574.73  58265.76      0  1000000
. sum achmat08
      Variable   Obs    Mean    Std. Dev.     Min     Max
      achmat08   500  56.59102  9.339608   36.61  77.2
. gen zachmat08 = (achmat08 - 56.59102) / 9.339608
. sum zachmat08
      Variable   Obs    Mean    Std. Dev.     Min     Max
      zachmat08   500  1.04e-08    1. -2.139385  2.206621
```

Can also use *eegen* with *std* function (*preferable*)

Z-score transformation

Alternatively, `sum` saves its results in “`r()`” macros:

```
. sum achmat08
    Variable      Obs       Mean     Std. Dev.      Min      Max
achmat08 |  500  56.59102   9.339608   36.61    77.2
. return list
scalars:
r(N) = 500
r(sum_w) = 500
r(mean) = 56.59102010345459
r(Var) = 87.22826930507507
r(sd) = 9.339607556266756
r(min) = 36.61000061035156
r(max) = 77.19999694824219
r(sum) = 28295.5100517273
. gen zachmat08b = (achmat08 - r(mean)) / r(sd)
. sum zachmat08b
    Variable      Obs       Mean     Std. Dev.      Min      Max
zachmat08b |  500  -9.77e-10   1  -2.139385  2.206621
```

Nonlinear transformations

- A *nonlinear* transformation involves some mathematical operation other than addition, subtraction, multiplication, and division
- There are lots of examples, but common ones include logarithmic (log), square root, inverse hyperbolic sine transformations

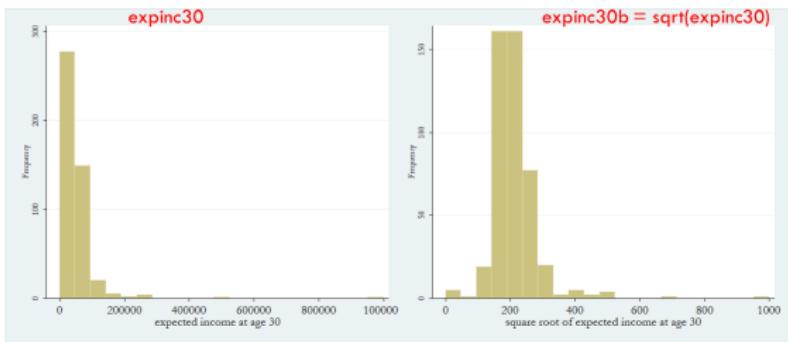
Nonlinear transformations

- Nonlinear transformations alter the shape (symmetry, skewness) of a distribution. In fact, this is often one of the purposes of a nonlinear transformation: to reduce the skewness in a distribution
- Many statistical procedures assume the variable has a normal or approximately normal distribution. In practice this assumption is violated. Transformations can help.

Square root transformation

- Consider the square root transformation of expected income at age 30: `gen expinc30b=sqrt(expinc30)`
- Skewness statistics: $G_{old} = 10.898$, $G_{new} = 3.7$

Square root transformation



Square root transformation

- The square root transformation reduces the magnitude of large values of the original variable by more than it does small values. Examples:
- Original value of 4 becomes 2 (half the size)
- Original value of 100 becomes 10 (1/10 the size)

Square root transformation

- The square root is not defined for negative numbers. If the original variable has negative values, one can first apply a translation ($+a$) that ensures the minimum value is above zero (preferably equal to 1)
- Values between 0 and 1 become larger when applying the square root, unlike values above 1
- Log transformations have a similar effect on the distribution of a variable as the square root

Log transformation

- Recall that logarithms are exponents and depend on the base
- e.g. $\log_2 8 = 3$
- In this example, 2 is the base. $\log_2 8$ asks: “to what power does one take 2 to get 8?”
- The base can be any positive number other than 1

Log transformation

x value	Base 2	"log points"	x value	Base 10	"log points"
2	$\log_2 2$	1	10	$\log_{10} 10$	1
4	$\log_2 4$	2	100	$\log_{10} 100$	2
8	$\log_2 8$	3	1000	$\log_{10} 1000$	3
16	$\log_2 16$	4	10000	$\log_{10} 10000$	4
32	$\log_2 32$	5			
64	$\log_2 64$	6			

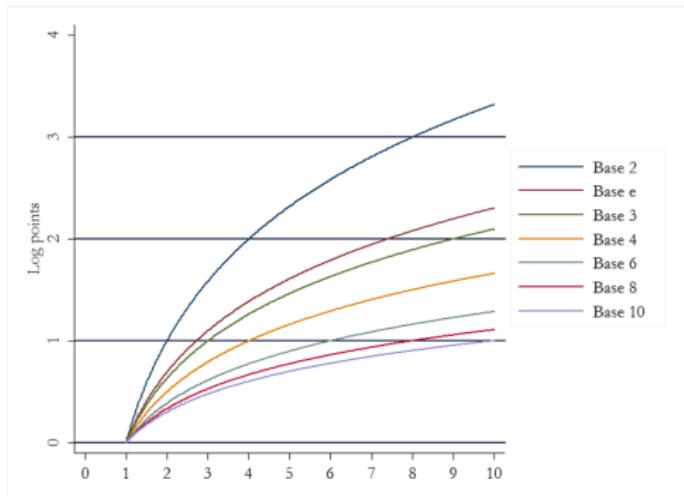
Log transformation

- On the \log_2 scale: each point is associated with 2 to that power
 - Values of 2, 4, 8, 16, ... \rightarrow 1, 2, 3, 4
 - Larger values reduced more than smaller ones
- On the \log_{10} scale: each point is associated with 10 to that power
 - Values of 10, 100, 1000, ... \rightarrow 1, 2, 3
 - Transformation has a larger effect

Log transformation

- Notice it takes larger and larger increases in x to go up by one unit (one “log point”). The square root has a similar property.
- The “natural logarithm” is a commonly used log function, where the base is 2.718 (e). The natural log is denoted $\ln(x)$.
- \ln is so commonly used that whenever you hear that a “log” has been applied, it is almost always \ln .
- The natural log also has a nice interpretation: a 0.01 change is approximately 1% change in the original variable. (This approximation works well for small changes, but not larger ones).
 - ▶ More precisely: an increase if p in the log is equal to a $(e^p - 1) \times 100\%$ increase in the original variable. For small p , $(e^p - 1)$ and p are very similar.

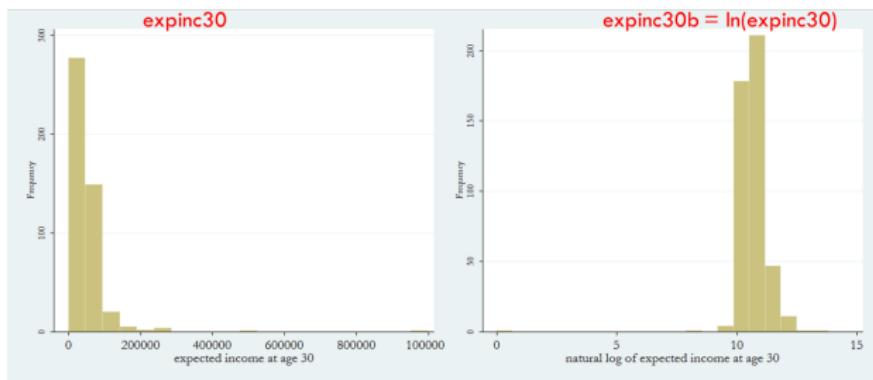
Log transformation



Log transformation

- Note the logarithm is only valid for $x \geq 1$. (The square root was only valid for $x \geq 0$)
- If x doesn't meet this condition, you can do a simple linear transformation first (e.g. add a constant a to all values such that the minimum $x \geq 1$).
- Another option: inverse hyperbolic sine transformation

Log transformation



Skewness statistics: $G_{old} = 10.898$, $G_{new} = -6.76$

Inverse hyperbolic sine transformation

The inverse hyperbolic sine function (or *arcsinh*) is defined both for $x > 0$ and $x \leq 0$:

$$IHS = \ln(x + \sqrt{x^2 + 1})$$

x	ln x	asinh x
-10	undef	-3.00
-5	undef	-2.31
0	undef	0.00
5	1.61	2.31
10	2.30	3.00
100	4.61	5.30
1000	6.91	7.60
50000	10.82	11.52
100000	11.51	12.21

In Stata, the IHS function is `asinh()`. See Bellemare & Wichman (2020) for more on IHS.

Other types of transformations

Other types of transformations:

- “collapsing” or re-coding variables—creating a categorical variable from a continuous variable with many possible values
- ranking
- combining variables—e.g. a composite score (math + reading + social studies + science) or an index of multiple outcomes (Kling, Liebman, & Katz, 2007)

Note: be aware of how missing values affect your calculation!

Recoding into ordinal groups

Example: *unitmath*, units of math classes taken, ranges from 1-6 but has fractional values. Can set up groups/intervals

```
. table unitmath
```

units in mathemati cs (naep)	Freq.
1	1
1.5	2
1.75	1
2	36
2.38	1
2.5	10
2.75	1
2.81	1
2.97	1
3	121
3.5	28
3.6	1
3.63	1
3.96	1
4	229
4.08	1
4.29	1
4.5	22
4.99	2
5	31
5.5	2
5.99	1
6	4

Recoding into ordinal groups

```
, recode unitmath (1/1.99 = 1) (2/2.99 = 2) (3/3.99 = 3) (4/4.99 =  
4) (5/100 - 5), gen(unitmathc)  
(82 differences between unitmath and unitmathc)
```

```
. table unitmathc
```

RECODE of	
unitmath	
(units in	
mathemati	
cs	
(naep)) Freq.	
1 4	
2 50	
3 153	
4 255	
5 38	

Multiple outcome index

Kling, Liebman, & Katz (2007) looked at the effects of the Moving to Opportunity (MTO) experiment in which families in public housing were given Section 8 vouchers to relocate to lower-poverty neighborhoods.

There are lots of possible outcomes which they combine them into indices:

- Index of economic self-sufficiency: equally weighted average of z-score for five measures of earnings, employment, and public assistance receipt
- Index of physical health (separate for adults and children): e.g., distress, depression, worrying, calmness, sleep
- Index of mental health (separate for adults and children): e.g., self-reported health, asthma, obesity, hypertension
- Index of youth risky behavior: e.g., marijuana, smoking, alcohol
- Index of youth education: e.g., graduated, reading score, math score

Winsorizing

Winsorizing is a type of non-monotonic transformation—it is a way of addressing outliers.

- **Trimming** means discarding (or ignoring) outlier observations, for example the bottom and top 1% of values.
- **Winsorizing** means re-coding outlier observations to less extreme values (e.g., setting values above the 95th percentile to the 95th percentile)

See Stata packages `trimmean`, `trimplot` and `winsor`, but use with good judgment! Removing/re-coding outliers has consequences.

Next lecture

- Basic rules of probability; probability as relative frequency
- Conditional probability and Bayes' Rule
- Probability distributions: discrete and continuous
- Binomial distribution
- Normal distribution
- And much, much, more!