# Lecture 3: Measures of central tendency

#### **Ernesto F. L. Amaral**

February 10, 2025
Advanced Methods of Social Research (SOCI 420)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 3 (pp. 66–90).



## Outline

- Mode
- Median
- Mean
- Skewed distributions



## Measures of central tendency

- Univariate descriptive statistics
  - Summarize information about the most typical, central, or common score of a variable
- Mode, median, and mean are different statistics and have same value only in certain situations
  - Mode: most common score
  - Median: score of the middle case
  - Mean: average score
- They vary in terms of
  - Level-of-measurement considerations
  - How they define central tendency



### Mode

The most common score

 Can be used with variables at all three levels of measurement

Most often used with nominal-level variables



# Finding the mode

Count the number of times each score occurred

The score that occurs most often is the mode

- If the variable is presented in a frequency distribution, the mode is the largest category
- If the variable is presented in a line chart, the mode is the highest peak



## Example of mode

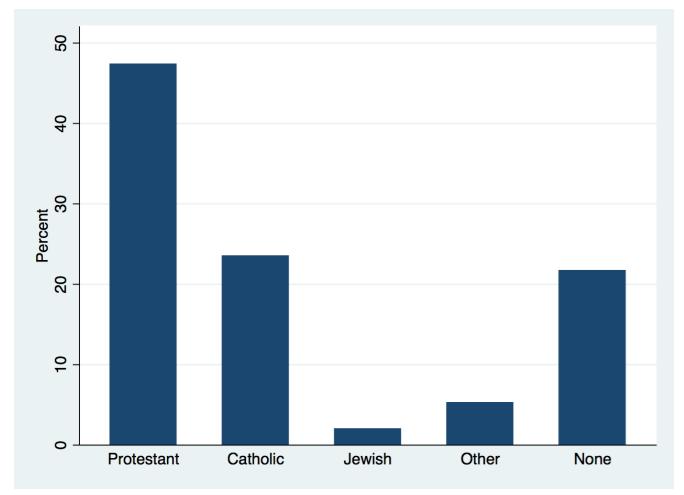
Top ten U.S. cities visited by overseas travelers, 2010

City	Number of visitors
Boston	1,186,000
Chicago	1,134,000
Las Vegas	2,425,000
Los Angeles	3,348,000
Miami	3,111,000
New York City	8,462,000
Oahu / Honolulu	1,634,000
Orlando	2,750,000
San Francisco	2,636,000
Washington, D.C.	1,740,000

Source: Healey 2015, p.67.

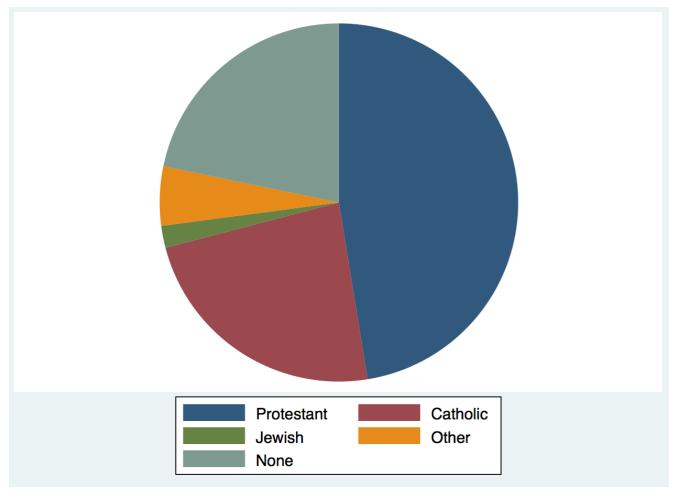


# Religious preference, U.S. adult population, 2016





# Religious preference, U.S. adult population, 2016





## Example: 2018 ACS in Stata

. tab raceth [fweight=perwt]

raceth	Freq.	Percent	Cum.
White	197,034,851	60.22	60.22
African American	40,373,281	12.34	72.56
Hispanic	59,740,273	18.26	90.82
Asian	18,662,293	5.70	96.53
Native American	2,170,486	0.66	97.19
Ohter races	9,186,255	2.81	100.00
Total	327,167,439	100.00	

. count if raceth!=.
3,214,539



## Edited table

Table 1. Distribution of U.S. population by race/ethnicity, 2018

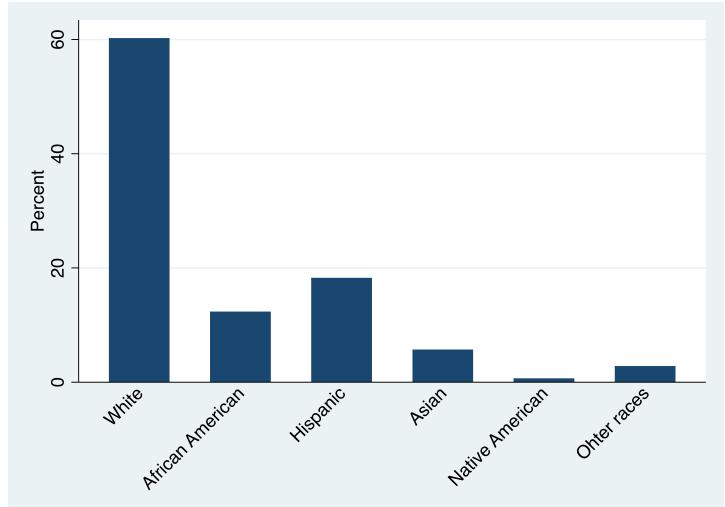
Race/ethnicity	Percentage
Non-Hispanic White	60.22
Non-Hispanic African American	12.34
Hispanic	18.26
Non-Hispanic Asian	5.70
Non-Hispanic Native American	0.66
Other races	2.81
Total	99.99
Population size (N)	327,167,439
Sample size (n)	3,214,539

Source: 2018 American Community Survey.



## Column graph for race/ethnicity, 2018

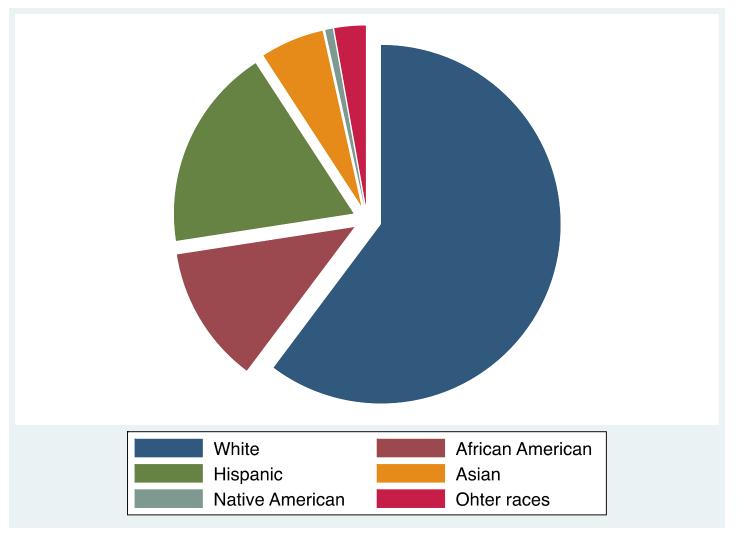
graph bar [fweight=perwt], over(raceth,
 label(angle(45))) ytitle("Percent")





## Pie graph for race/ethnicity, 2018

graph pie [fweight=perwt], over(raceth) pie(\_all, explode)





Source: 2018 American Community Survey.

## Example: 2018 ACS in Stata

. tab educgr [fweight=perwt]

educgr	Freq.	Percent	Cum.
Less than high school	97,758,814	29.88	29.88
High school	92,183,547	28.18	58.06
Some college	60,822,461	18.59	76.65
College	47,865,798	14.63	91.28
Graduate school	28,536,819	8.72	100.00
Total	327,167,439	100.00	

. count if educgr!=.
3,214,539



## Edited table

Table 1. Distribution of U.S. population by educational attainment, 2018

Educational attainment	Percentage
Less than high school	29.88
High school	28.18
Some college	18.59
College	14.63
Graduate school	8.72
Total	100.00
Population size (N)	327,167,439
Sample size (n)	3,214,539

Source: 2018 American Community Survey.



## Limitations of mode

- Some distributions have no mode
- Some distributions have multiple modes

Distributions of scores on two tests

Score (% correct)	Test A Frequency of scores	Test B Frequency of scores
97	14	22
91	14	3
90	14	4
86	14	22
77	14	3
60	14	22
55	14	22
Total	98	98

Source: Healey 2015, p.68.



### Limitations of mode

 The mode of an ordinal or interval-ratio level variable may not be central to the whole distribution

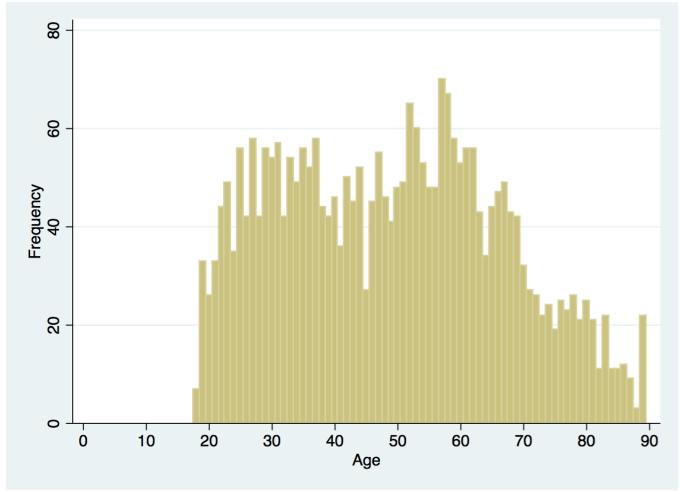
A distribution of tes	st score:	5
-----------------------	-----------	---

Score (% correct)	Frequency
93	8
68	3
67	4
66	2
62	7
Total	24

Source: Healey 2015, p.68.

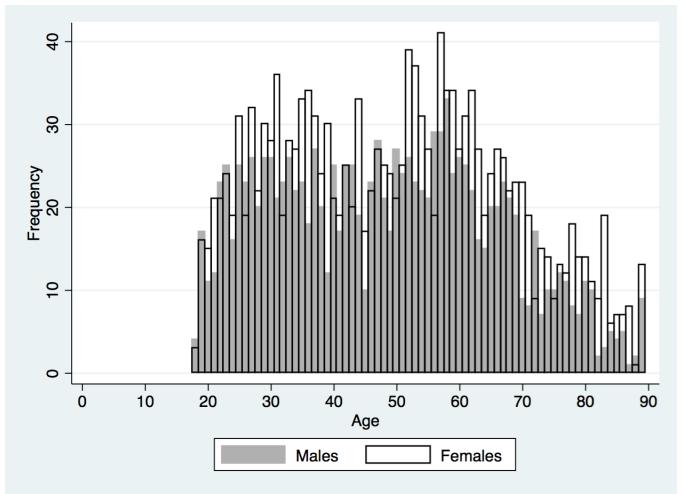


# Age distribution, U.S. adult population, 2016





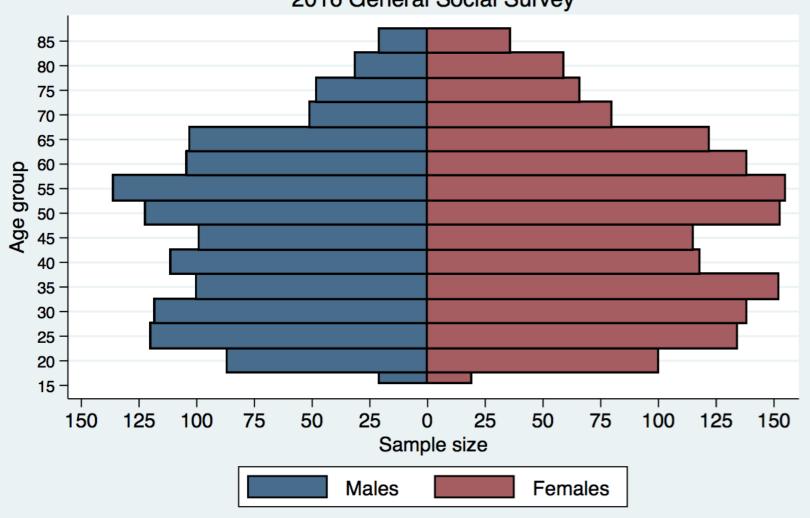
# Age distribution by sex, U.S. adult population, 2016

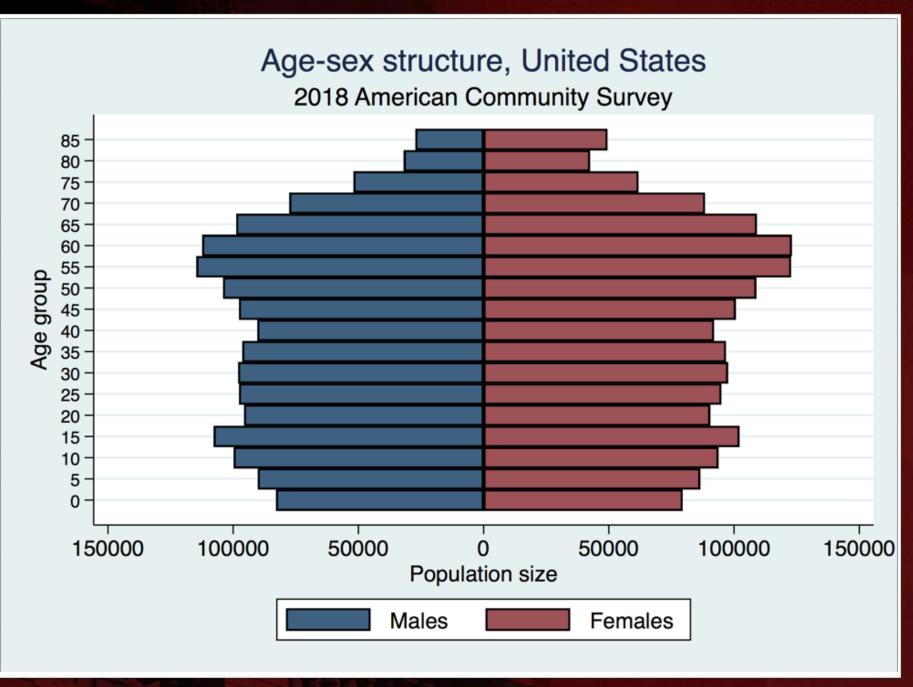




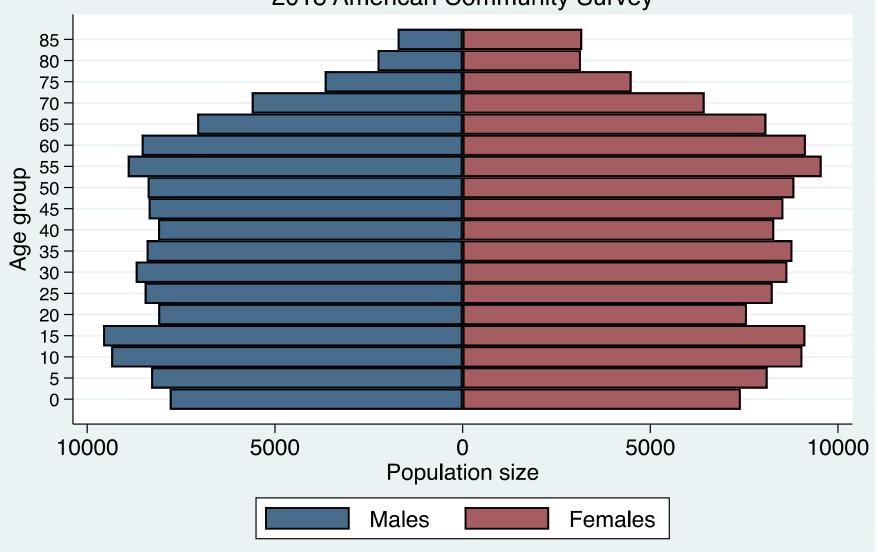
#### Age-sex structure, United States

2016 General Social Survey



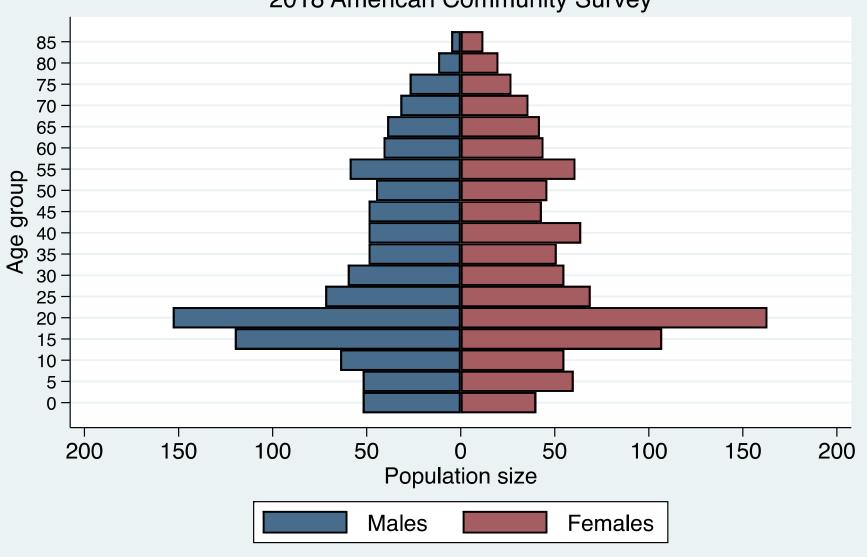


#### Age-sex structure, Texas 2018 American Community Survey



#### Age-sex structure, Brazos county

2018 American Community Survey



### Median

 The median (Md) is the exact center of distribution of scores

The score of the middle case

 It can be used with ordinal-level or interval-ratiolevel variables

It cannot be used for nominal-level variables



# Finding the median

- Arrange the cases from low to high
  - Or from high to low
- Locate the middle case
- If the number of cases (N) is odd
  - The median is the score of the middle case

- If the number of cases (N) is even
  - The median is the average of the scores of the two middle cases



## Example of median

Finding the median with seven cases (N is odd)

Case	Score	
Α	10	
В	10	
С	8	
D	7	← Median = <i>Md</i>
E	5	
F	4	
G	2	

Source: Healey 2015, p.69.



## Example of median

Finding the median with eight cases (N is even)

Case	Score	
A	10	
В	10	
С	8	
D	7	
		← Median = <i>Md</i> = (7+5) / 2 = 6
E	5	
F	4	
G	2	
Н	1	

AM

Source: Healey 2015, p.69.

## Other measures of position

- Percentiles
  - Point below which a specific percentage of cases fall
- Deciles
  - Divides distribution into tenths (10, 20, 30, ..., 90)
- Quartiles
  - Divides distribution into quarters (25, 50, 75)
- The median falls at the 50th percentile or the 5th decile or the 2nd quartile

### Manual calculation

- Arrange scores in order from low to high
- Multiply the number of cases (N) by the proportional value of the percentile
  - For example: the 75th percentile would be 0.75
- The resultant value marks the order number of the case that falls at the percentile



# Examples of manual calculation

- In a sample of 70 test grades we want to find the 4th decile (or 40th percentile)
  - $-70 \times 0.40 = 28$
  - The 28th case is the 40th percentile
- In a sample of 70 test grades we want to find the 3rd quartile (or 75th percentile)
  - $-70 \times 0.75 = 52.5$ , rounding to 53
  - The 53rd case is the 75th percentile



## Example: 2016 GSS in Stata

• 75% of the population is younger than 60 years sum age [aweight=wtssall], d

age of respondent

	Percentiles	Smallest		
1%	19	18		
5%	21	18		
10%	24	18	0bs	2,857
25%	33	18	Sum of Wgt.	2,855.4791
50%	47		Mean	47.56141
		Largest	Std. Dev.	17.58891
75%	60	89		
90%	72	89	Variance	309.3698
95%	78	89	Skewness	.2328772
99%	86	89	Kurtosis	2.161393



## Example: 2016 GSS in Stata

 The "centile" command allows us to estimate any percentile, but weights are not allowed centile age, centile (37)

• 37% of the sample is younger than 41 years

Variable		Percentile	Centile	[95% Conf. Int	
age	2,857	37	41	40	42



### Mean

- The average score
- Requires variables measured at the interval-ratio level, but is often used with ordinal-level variables
- Cannot be used for nominal-level variables
- The mean (arithmetic average) is by far the most commonly used measure of central tendency

# Finding the mean

- Add all of the scores and then divide by the number of scores (N)
- The mathematical formula for the mean is

$$\bar{X} = \frac{\sum (X_i)}{N}$$

where  $\overline{X}$  = the mean  $\Sigma(X_i)$  = the summation of the scores N = the number of cases



# Examples of mean, 2016 GSS

#### Mean income by sex

tabstat conrinc [aweight=wtssall], by(sex) stat(mean)

Sex	Mean income
Male	41,282.78
Female	28,109.34
Overall	34,649.30

#### Mean income by race/ethnicity

tabstat conrinc [aweight=wtssall], by(raceeth) stat(mean)

Race/ethnicity	Mean income
Non-Hispanic white	38,845.62
Non-Hispanic black	23,243.04
Hispanic	23,128.92
Other	50,156.35
Overall	34,649.30

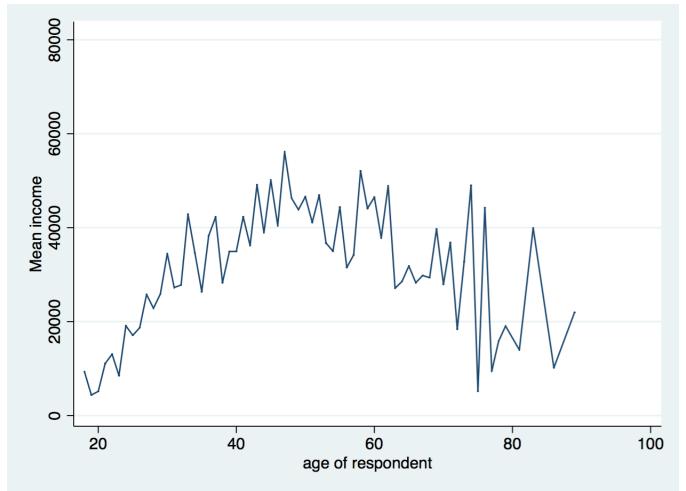
#### Mean income by age-group

tabstat conrinc [aweight=wtssall], by(agegr1) stat(mean)

Age group	Mean income
18–24	11,214.16
25–44	32,863.93
45–64	42,552.21
65–89	30,848.29
Overall	34,649.30

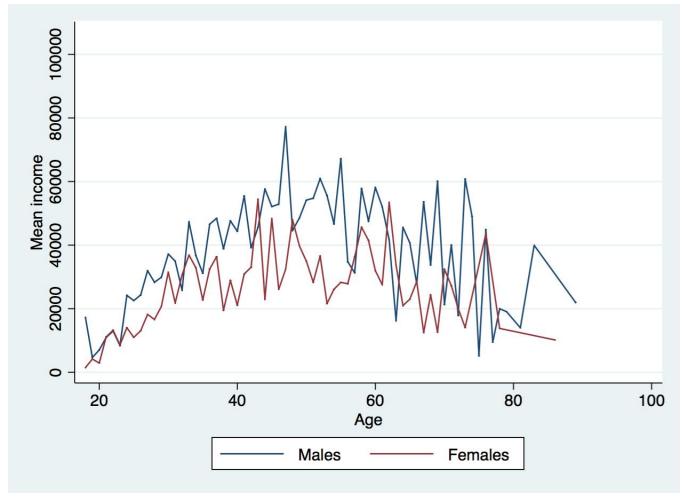


# Mean income by age, U.S. adult population, 2016





# Mean income by age and sex, U.S. adult population, 2016





## Wage and salary income by sex, 2018 ACS

- . \*\*\*Income
- . table year [fweight=perwt] if income!=0, c(mean income p50 income)

Census year	mean(income)	med(income)
2018	50043.98	35000

- . \*\*\*Income by sex
- . table female [fweight=perwt] if income!=0, c(mean income p50 income)

female	mean(income)	med(income)
Male	59014.14	40000
Female	40294.34	30000

## Wage and salary income by race/ethnicity, 2018 ACS

- . \*\*\*Income by race/ethnicity
- . table raceth [fweight=perwt] if income!=0, c(mean income p50 income)

raceth	mean(income)	med(income)
White	55289.18	40000
African American	37183.63	29000
Hispanic	36236.16	27500
Asian	64154.23	43000
Native American	34851.55	27000
Ohter races	44162.79	30000



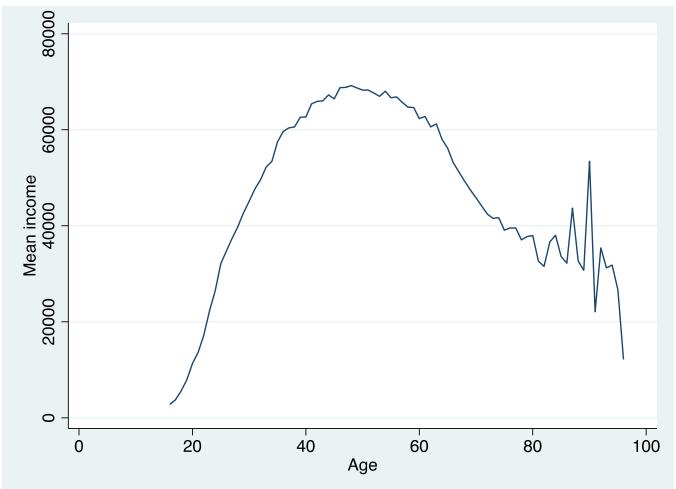
## Wage and salary income by education, 2018 ACS

- . \*\*\*Income by educational attainment
- . table educgr [fweight=perwt] if income!=0, c(mean income p50 income)

educgr	mean(income)	med(income)
Less than high school High school Some college	22750.89 34055.76 39607.05	18000 27000 30300
College Graduate school	67654.84 98541.49	50000 72000



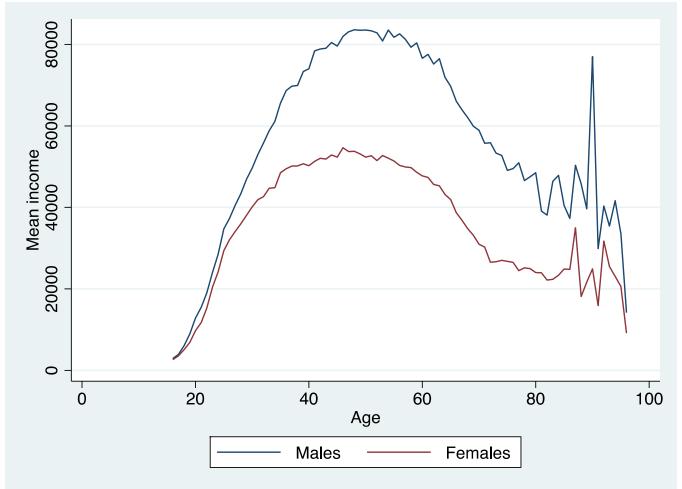
# Mean income by age, U.S. population, 2018





Source: 2018 American Community Survey.

# Mean income by age and sex, U.S. population, 2018





Source: 2018 American Community Survey.

#### Three characteristics of the mean

- Mean balances all the scores in a distribution
  - All scores cancel out around the mean

$$\sum (X_i - \bar{X}) = 0$$

 Mean minimizes the variation of the scores, "least squares principle"

$$\sum (X_i - \bar{X})^2 = minimum$$

- Mean is affected by all scores
  - All scores are used in the calculation of the mean
  - It can be misleading if the distribution has "outliers" April

#### Mean balances all the scores

A demonstration showing that all scores cancel out around the mean

$X_i$	$X_i - \overline{X}$
65	65 - 78 = -13
73	73 - 78 = -5
77	77 - 78 = -1
85	85 - 78 = 7
90	90 - 78 = 12
$\sum_{i} (X_i) = 390$ $\overline{X} = 390 / 5 = 78$	$\sum (X_i - \overline{X}) = 0$

Source: Healey 2015, p.74.



#### Mean minimizes variation

- A demonstration showing that the mean is the point of minimized variation
  - If we performed these operations with any number other than the mean (e.g., 77), the result would be a sum greater than 388

$X_i$	$X_i - \overline{X}$	$(X_i - \overline{X})^2$	$(X_i - 77)^2$
65	65 - 78 = -13	$(-13)^2 = 169$	$(65 - 77)^2 = (-12)^2 = 144$
73	73 - 78 = -5	$(-5)^2 = 25$	$(73 - 77)^2 = (-4)^2 = 16$
77	77 - 78 = -1	$(-1)^2 = 1$	$(77 - 77)^2 = (0)^2 = 0$
85	85 - 78 = 7	$(7)^2 = 49$	$(85 - 77)^2 = (8)^2 = 64$
90	90 - 78 = 12	$(12)^2 = 144$	$(90 - 77)^2 = (13)^2 = 169$
$\sum (X_i) = 390$	$\sum (X_i - \overline{X}) = 0$	$\sum (X_i - \overline{X})^2 = 388$	$\sum (X_i - 77)^2 = 393$
$\overline{X}$ = 78			

Source: Healey 2015, p.75.

### Mean is affected by all scores

 A demonstration showing that the mean is affected by every score

Scores	Measures of central tendency	Scores	Measures of central tendency	Scores	Measures of central tendency
15	Mean = 25	15	Mean = 718	0	Mean = 22
20		20		20	
25	Median = 25	25	Median = 25	25	Median = 25
30		30		30	
35		3500		35	

Source: Healey 2015, p.76.



### Mean is affected by all scores

- Strength
- The mean uses all the available information from the variable

- Weaknesses
- The mean is affected by every score
- If there are some very high or low scores
  - Extreme scores: "outliers"
  - The mean may be misleading
  - This is the case of skewed distributions

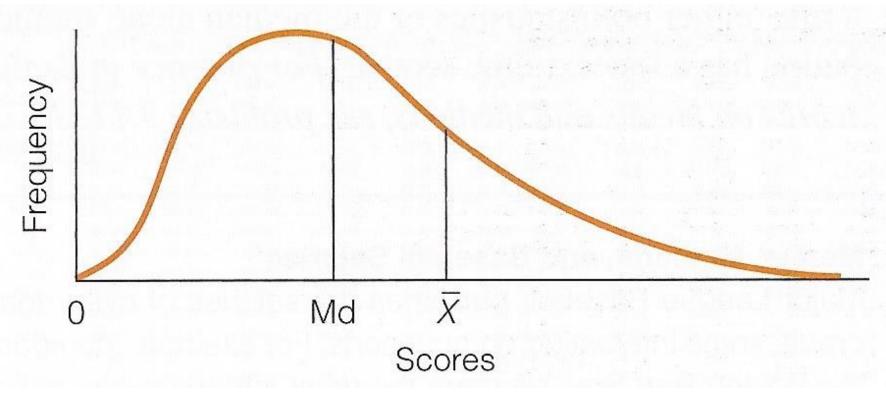


#### Skewed distributions

- When a distribution has a few very high or low scores, the mean will be pulled in the direction of the extreme scores
- For a positive skew
  - The mean will be greater than the median
- For a negative skew
  - The mean will be less than the median
- When an interval-ratio-level variable has a pronounced skew, the median may be the more trustworthy measure of central tendency

### Positively skewed distribution

The mean is greater in value than the median

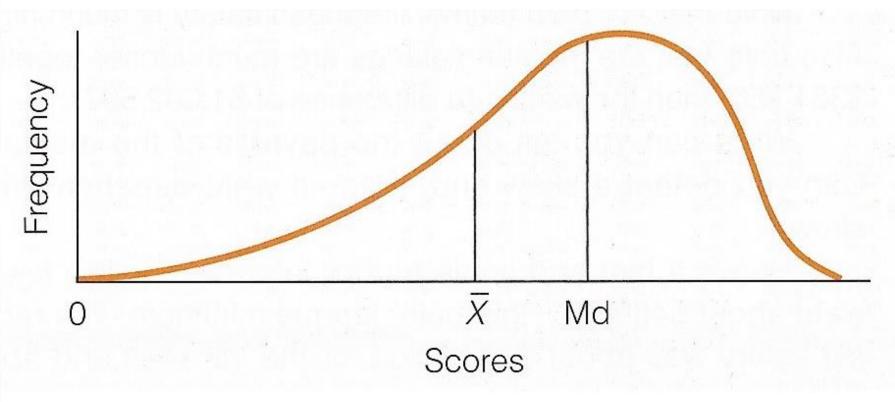


Source: Healey 2015, p.77.



### Negatively skewed distribution

The mean is less than the median

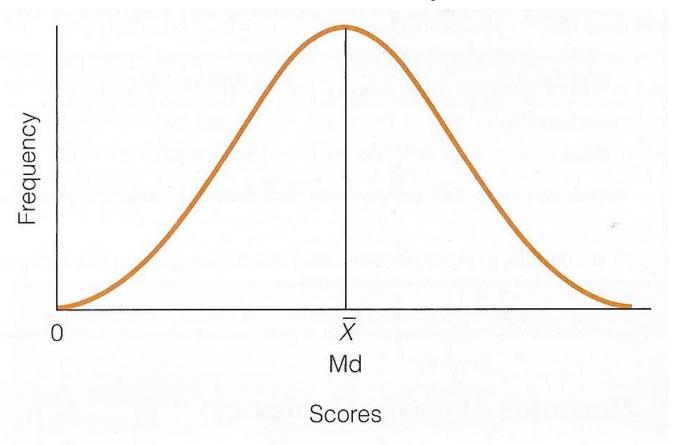


Source: Healey 2015, p.77.



### Symmetrical distribution

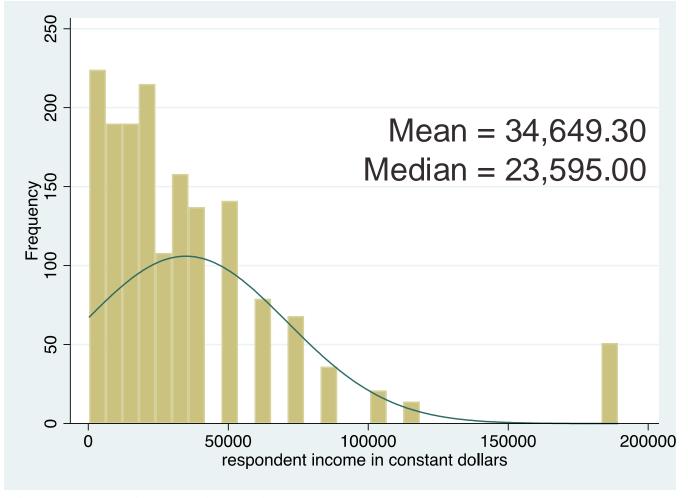
The mean and median are equal







## Income distribution, U.S. adult population, 2016

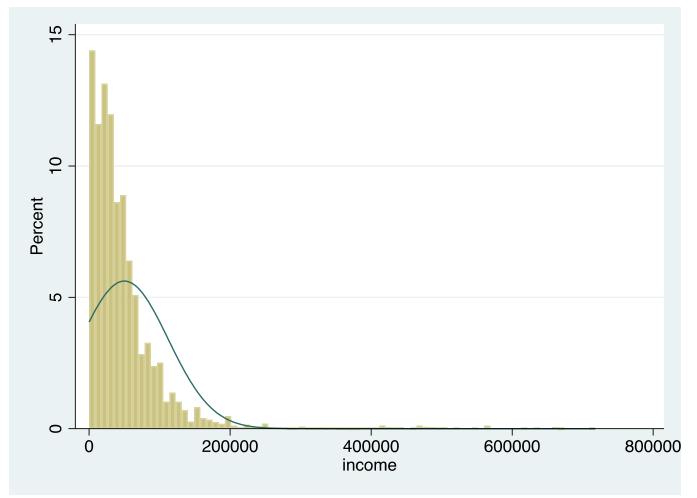




Source: 2016 General Social Survey.

### Histogram of wage and salary income, U.S. population, 2018

hist income [fweight=perwt] if income!=0, percent normal





Source: 2018 American Community Survey.

#### Level of measurement

 Relationship between level of measurement and measures of central tendency

Measure	Level of measurement			
of central tendency	Nominal	Ordinal	Interval-ratio	
Mode	YES	Yes	Yes	
Median	No	YES	Yes	
Mean	No	Yes (?)	YES	

- YES: most appropriate measure for each level
- Yes: measure is also permitted
- Yes (?): mean is often used with ordinal-level variables, but this practice violates level-ofmeasurement guidelines
- No: cannot be computed for that level



### Summary to choose measure

Use the mode when:	1. The variable is measured at the nominal level.	
	2. You want a quick and easy measure for ordinal- and interval-ratio-level variables.	
	3. You want to report the most common score.	
Use the median when:	1. The variable is measured at the ordinal level.	
	2. An interval-ratio variable is badly skewed.	
	<ol><li>You want to report the central score. The median always lies at the exact center of the distribution.</li></ol>	
Use the mean when:	1. The variable is measured at the interval-ratio level (except when the variable is badly skewed).	
	2. You want to report the typical score. The mean is the statistics that exactly balances all of the scores.	
	3. You anticipate additional statistical analysis.	
Source: Healey 2015, p.81.	ĀM	

