Lecture 3: Measures of central tendency and dispersion

Ernesto F. L. Amaral

September 8, 2022
Introduction to Sociological Data Analysis (SOCI 600)

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapters 3 (pp. 66–90), 4 (pp. 91–121).

TEXAS A&M UNIVERSITY

Outline

- Measures of central tendency
- Measures of dispersion

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

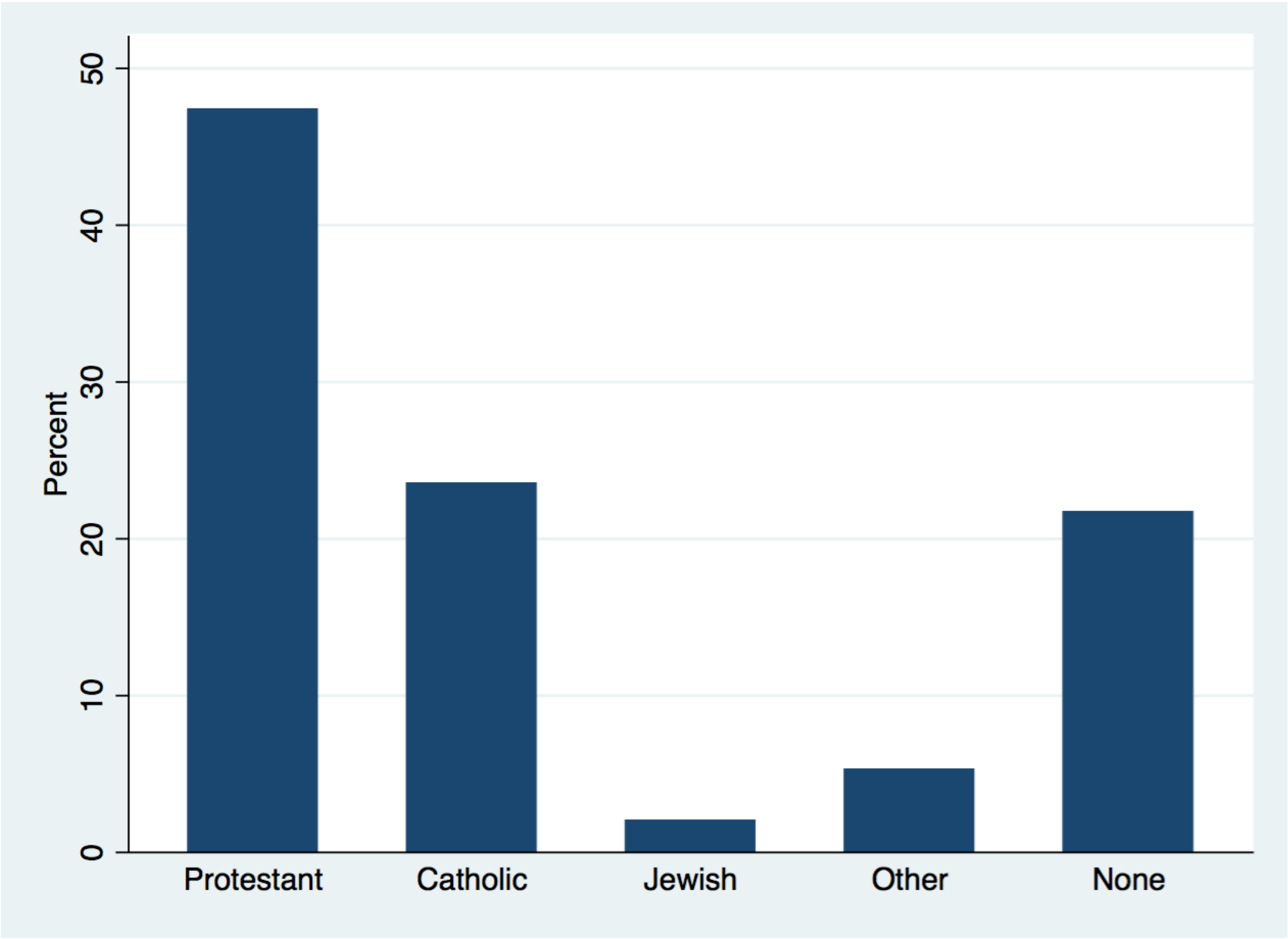
Example of mode

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

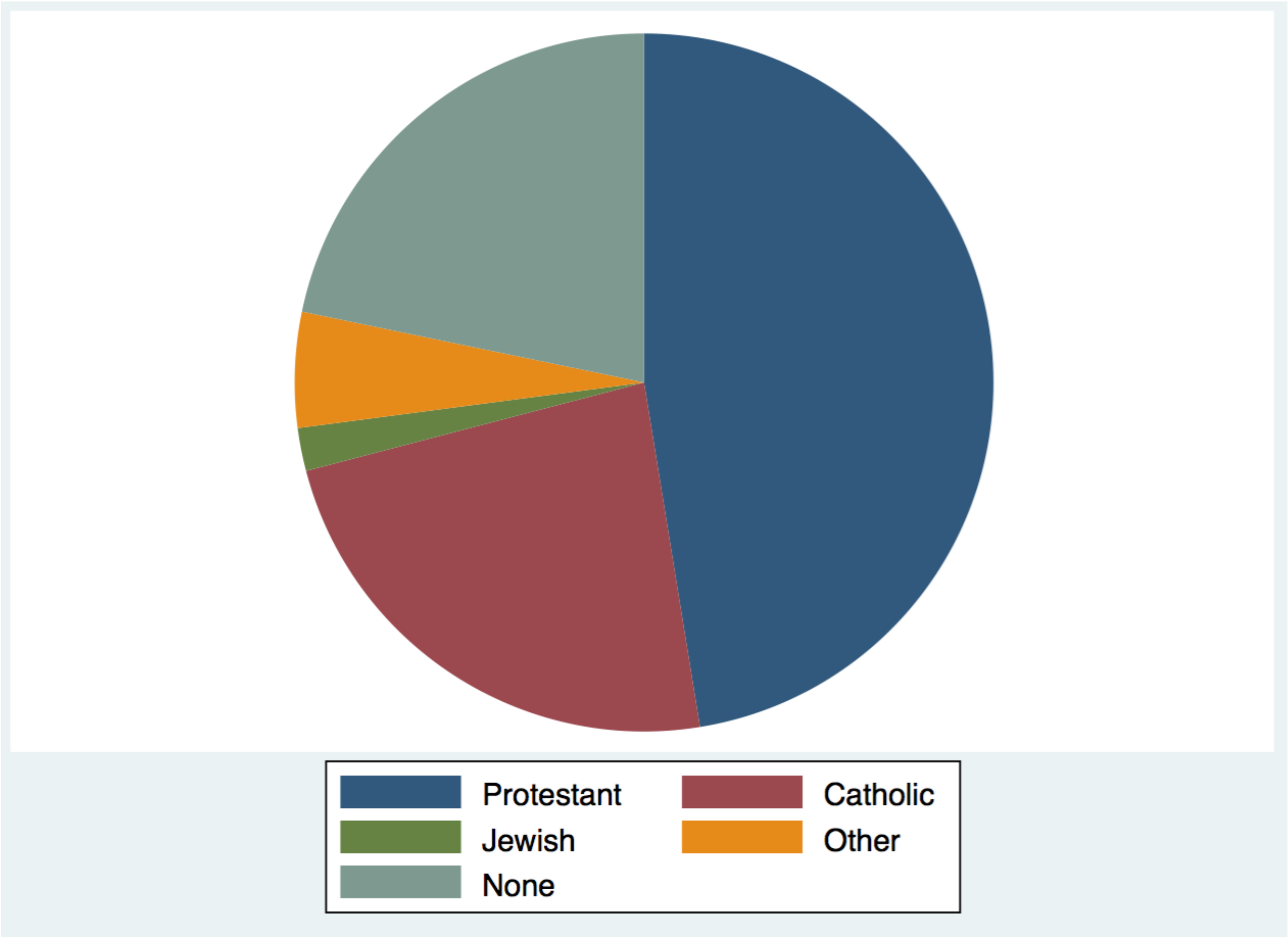
Source: Healey 2015, p.67.

Religious preference, U.S. adult population, 2016



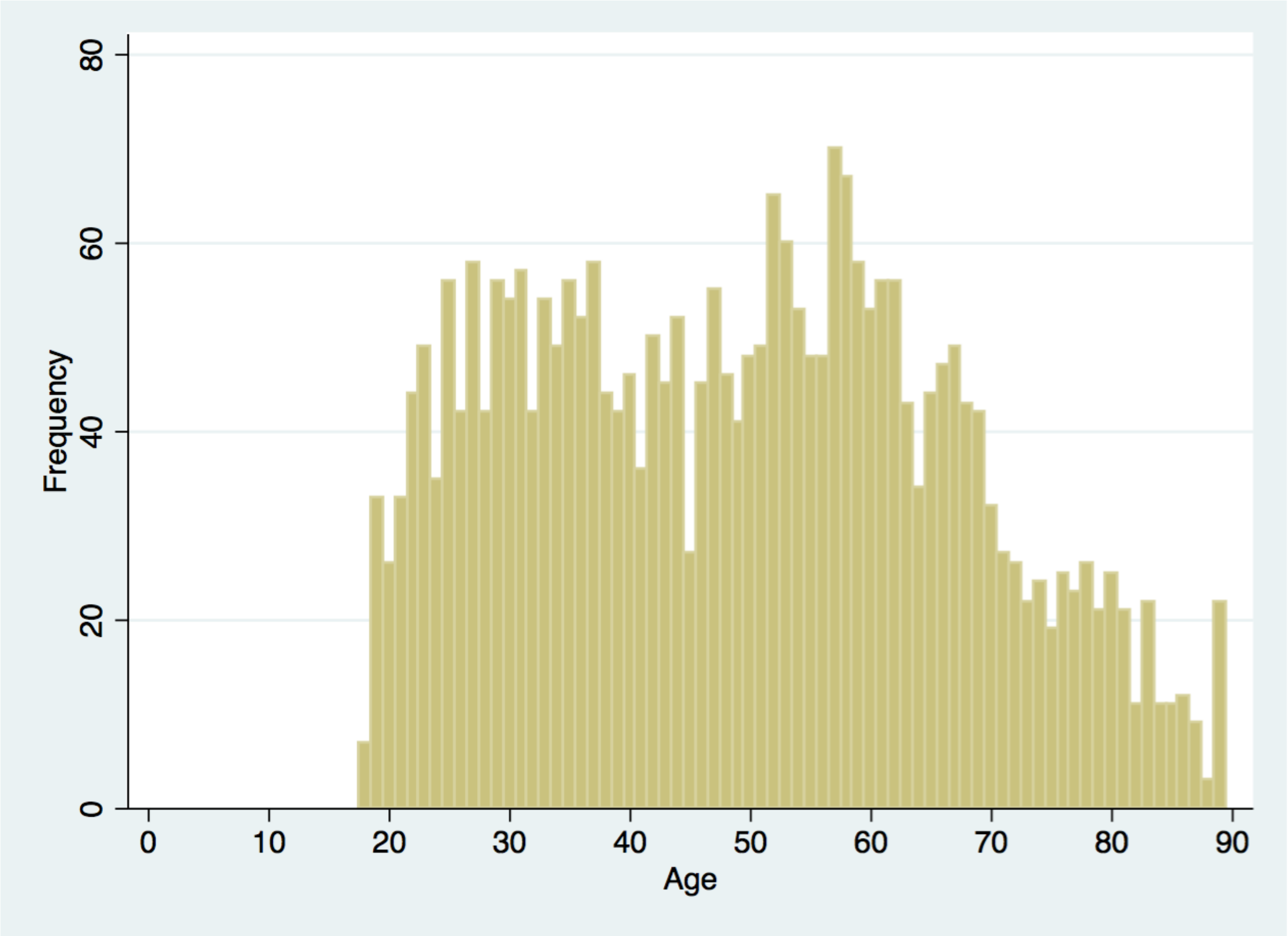
Source: 2016 General Social Survey.

Religious preference, U.S. adult population, 2016



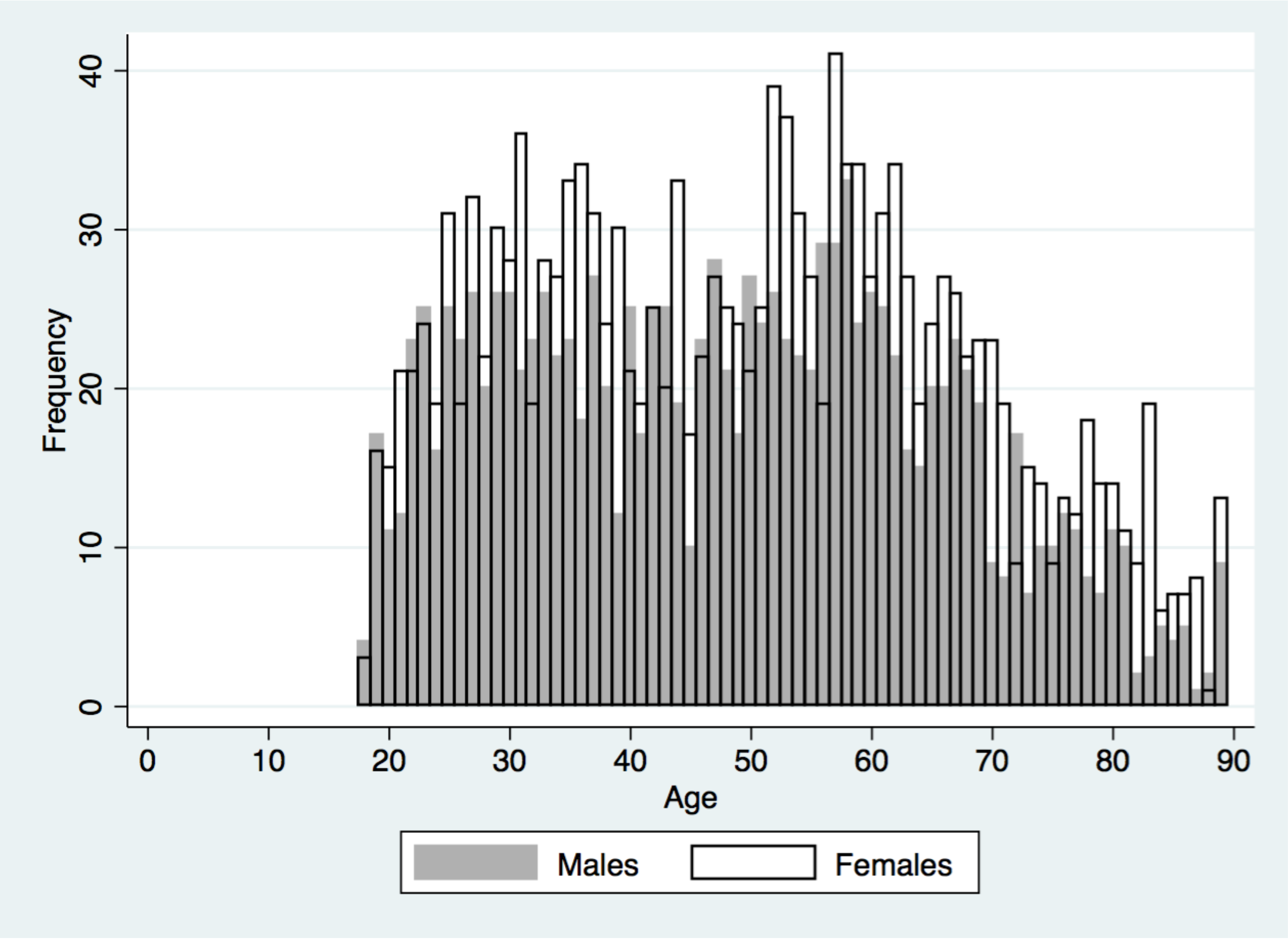
Source: 2016 General Social Survey.

Age distribution, U.S. adult population, 2016

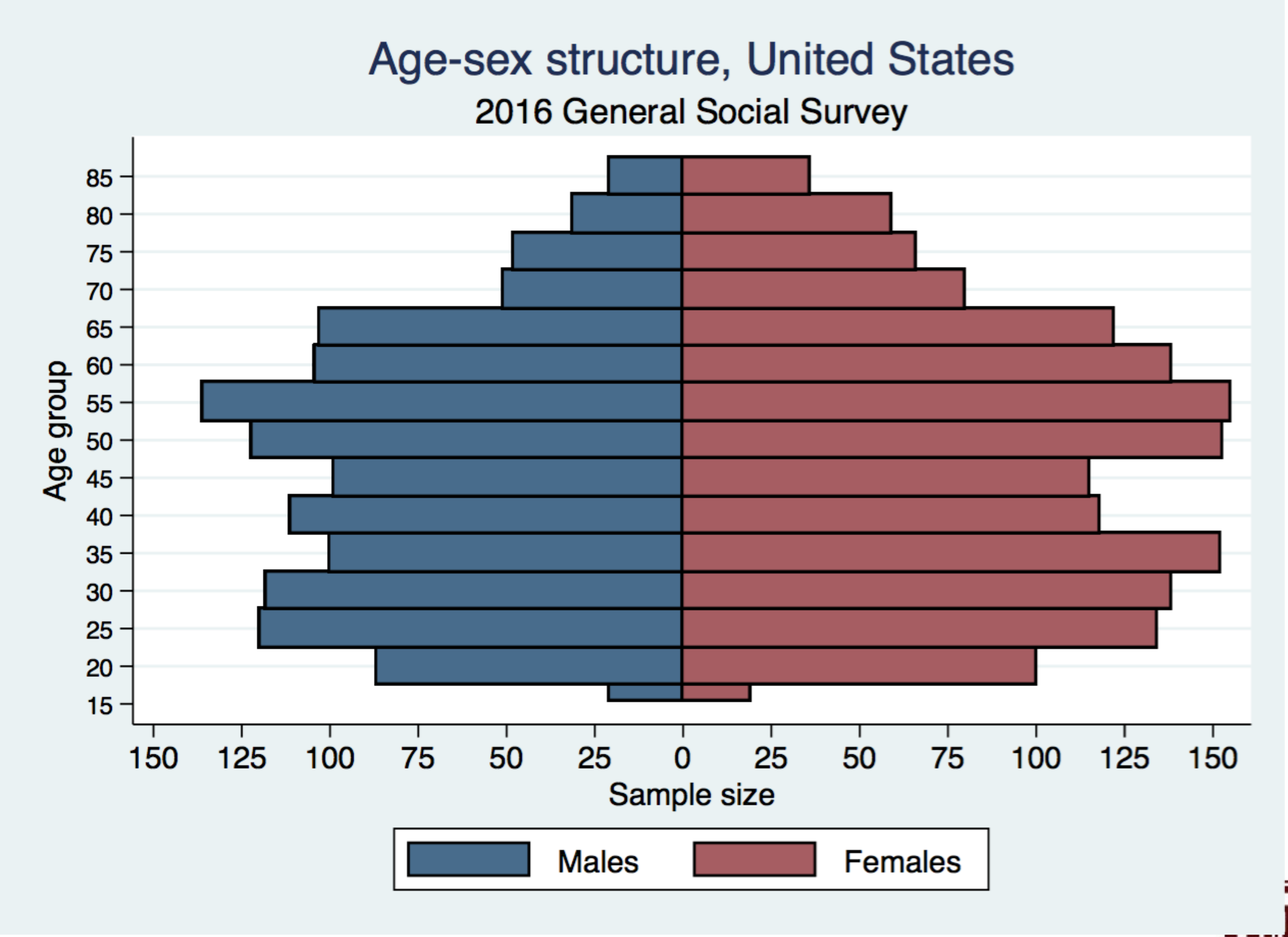


Source: 2016 General Social Survey.

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



Source: 2016 General Social Survey.


Age-sex structure, United States
2016 General Social Survey
Age group
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
150
125
100
75
50
25
0
25
50
75
100
125
150
Sample size
Males
Females

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 test scores

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

Source: Healey 2015, p.68.

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-ratio-level 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**	
A	10	
B	10	
C	8	
D	7	← Median = *Md*
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	
B	10	
C	8	
D	7	
		← Median = *Md* = (7+5) / 2 = 6
E	5	
F	4	
G	2	
H	1	

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 x 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 x 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       Obs                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	Obs	Percentile	Centile	— Binom. Interp. — [95% Conf.	Interval]
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 $\bar{X}$ = the mean

$\sum(X_i)$ = 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**

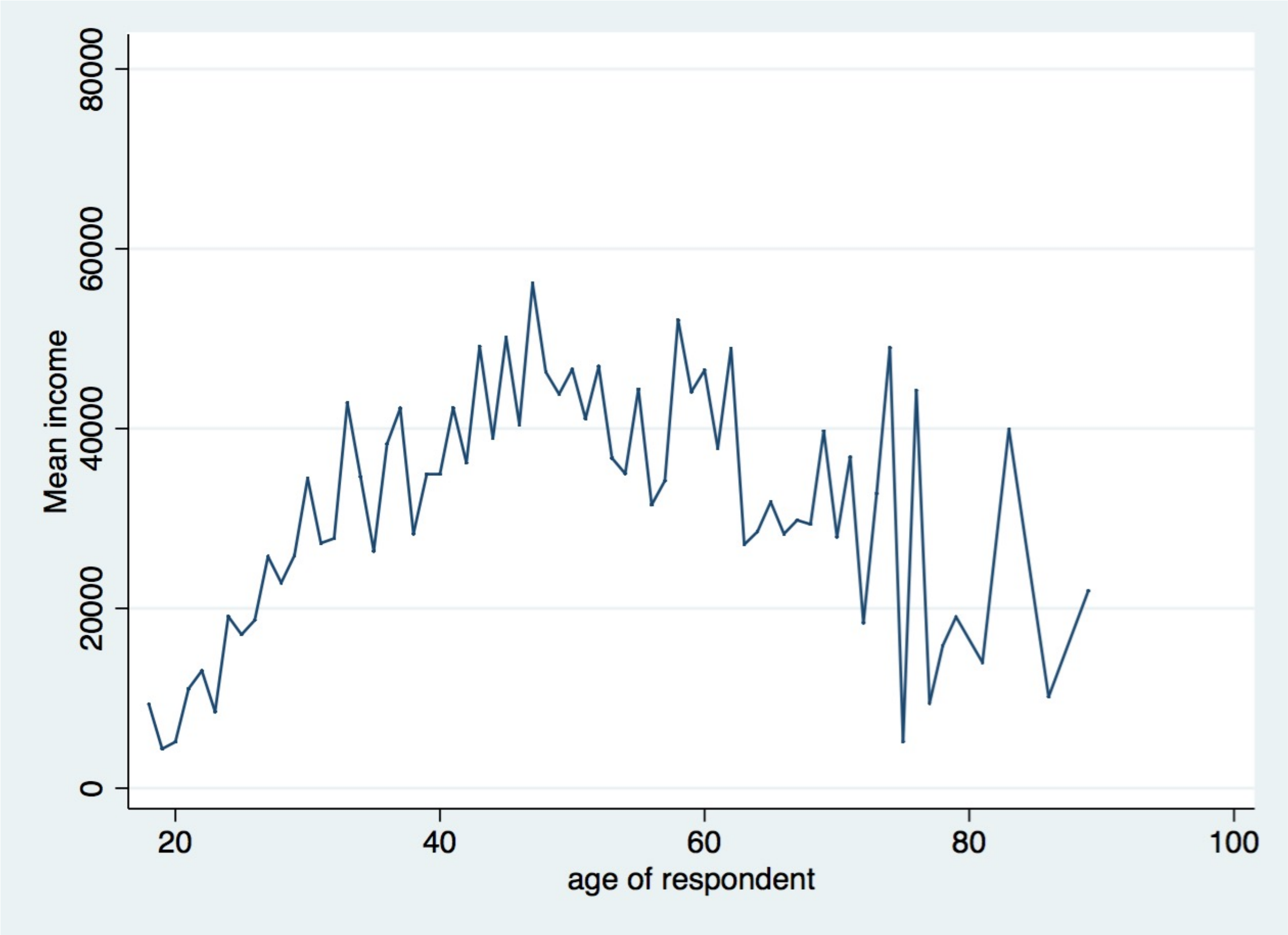
Mean income by age-group

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**

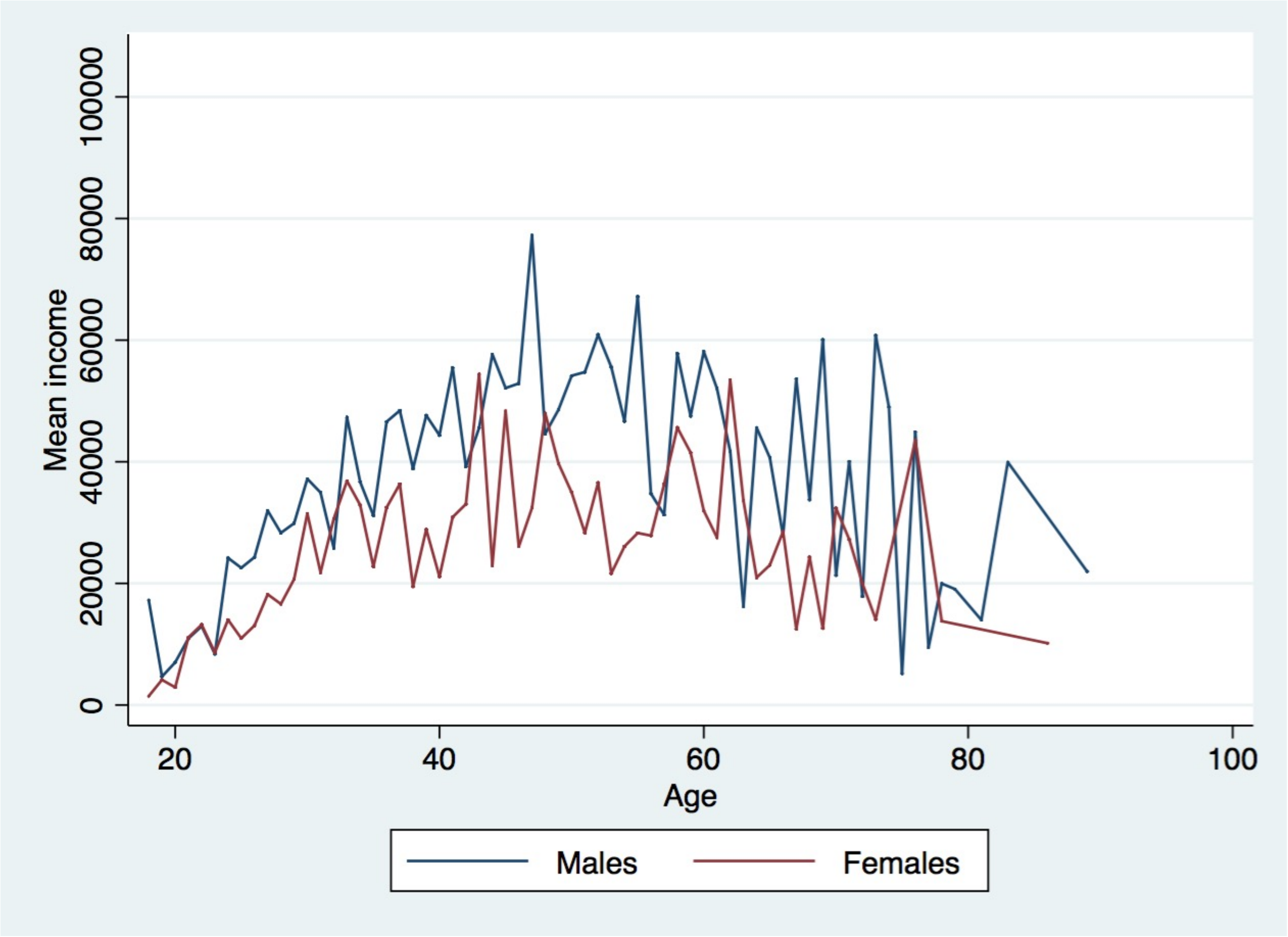
Source: 2016 General Social Survey.

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



Source: 2016 General Social Survey.

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



Source: 2016 General Social 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”

Mean balances all the scores

- A demonstration showing that all scores cancel out around the mean

X_i	$X_i - \bar{X}$
65	65 – 78 = –13
73	73 – 78 = –5
77	77 – 78 = –1
85	85 – 78 = 7
90	90 – 78 = 12
$\sum(X_i)$ **= 390** $\bar{X}$ **= 390 / 5 = 78**	$\sum(X_i - \bar{X})$ **= 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	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$(X_i - 77)^2$
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** $\bar{X}$ **= 78**	$\sum(X_i - \bar{X})$ **= 0**	$\sum(X_i - \bar{X})^2$ **= 388**	$\sum(X_i - 77)^2$ **= 393**

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

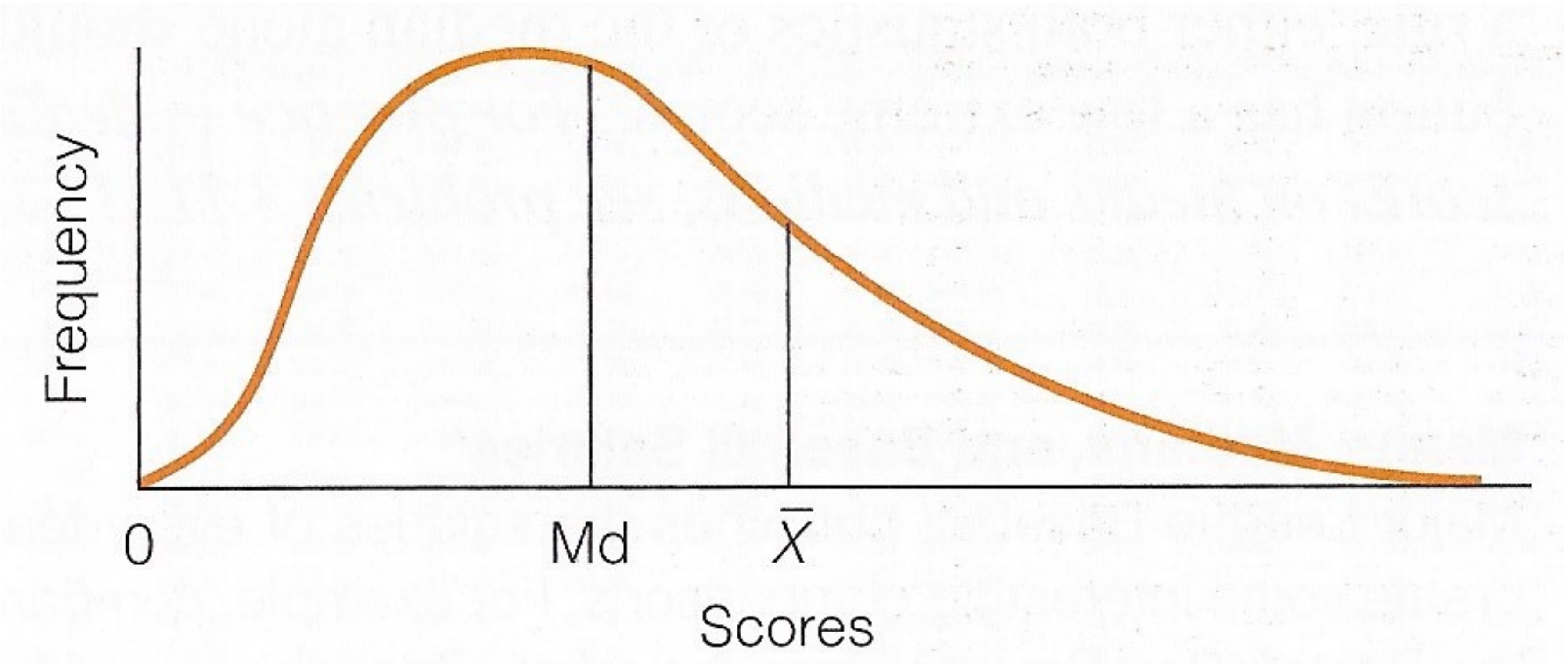
- **Weaknesses**
- The mean is affected by every score
- If there are some very high or low scores
  - Extreme scores: ”outliers”
  - The mean may be misleading
  - This is the case of skewed distributions

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

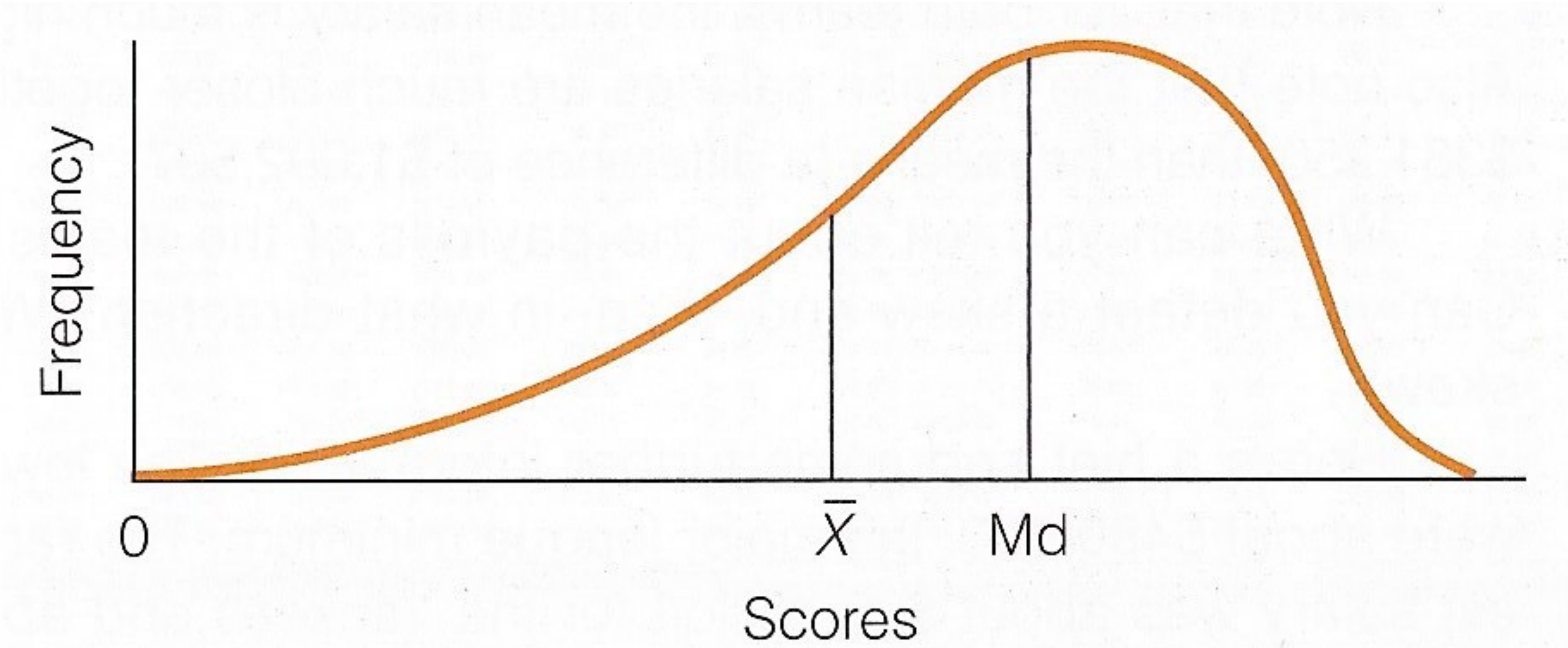
- The mean is greater in value than the median



Source: Healey 2015, p.77.

Negatively skewed distribution

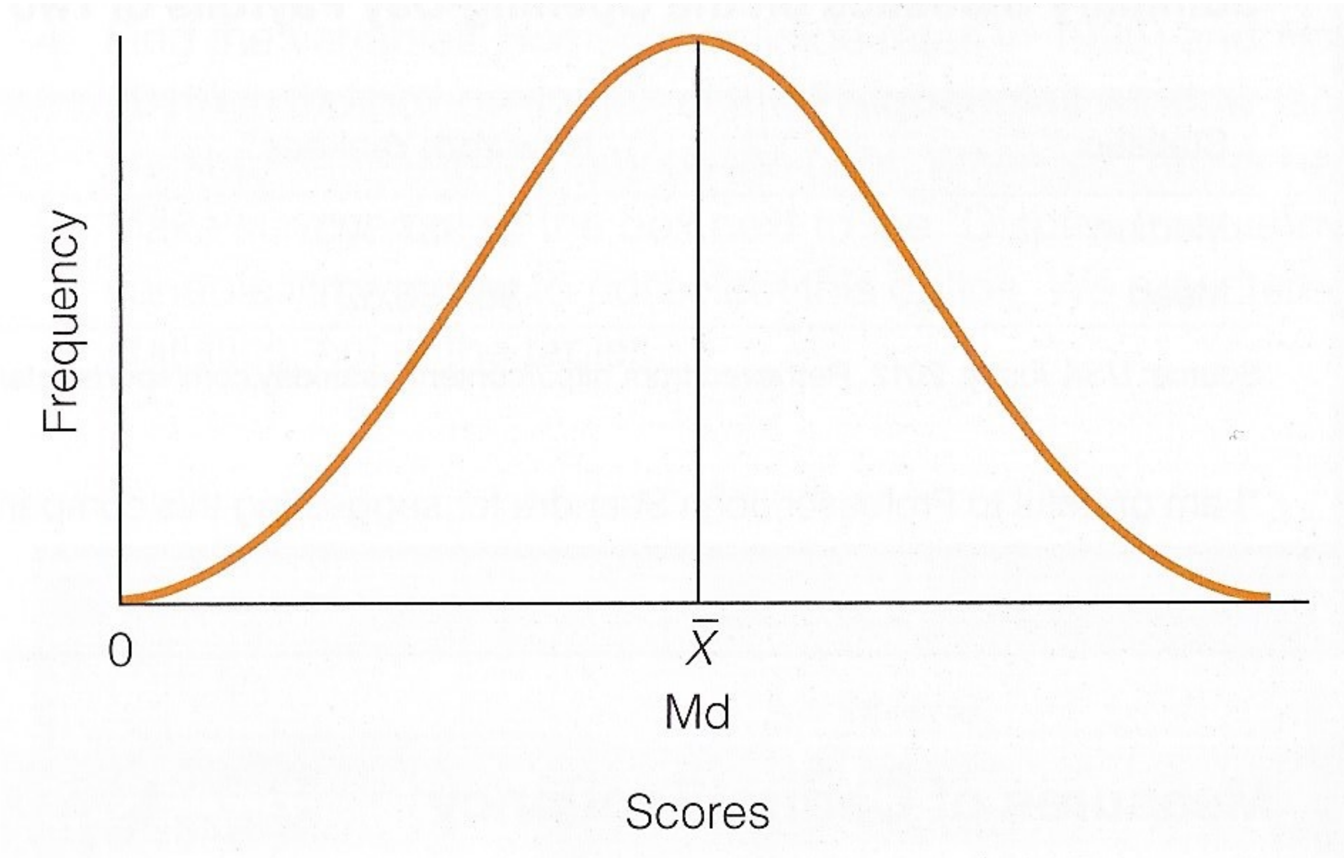
- The mean is less than the median



Source: Healey 2015, p.77.

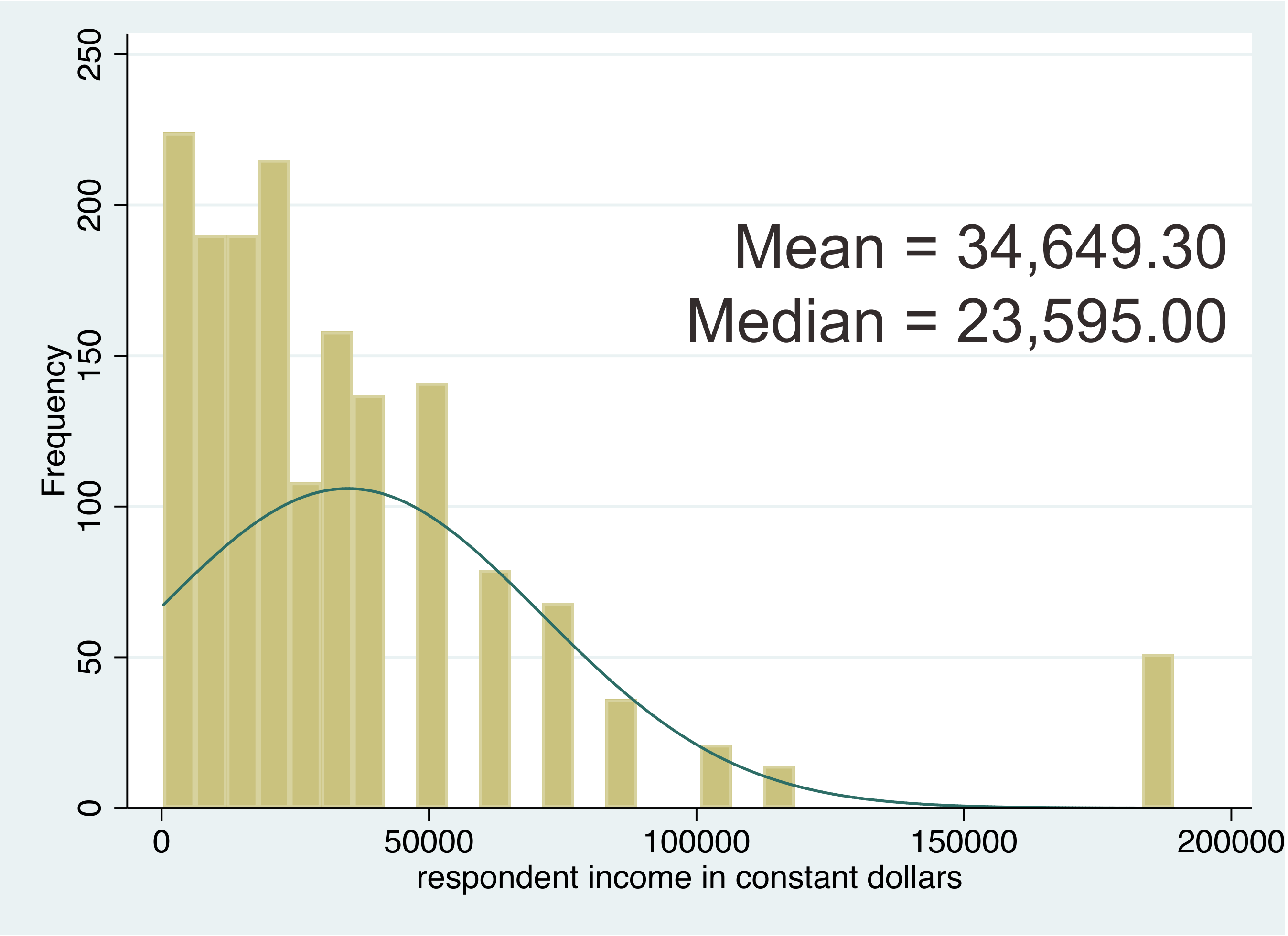
Symmetrical distribution

- The mean and median are equal



Source: Healey 2015, p.77.

Income distribution, U.S. adult population, 2016



Source: 2016 General Social Survey.

Level of measurement

- Relationship between level of measurement and measures of central tendency

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

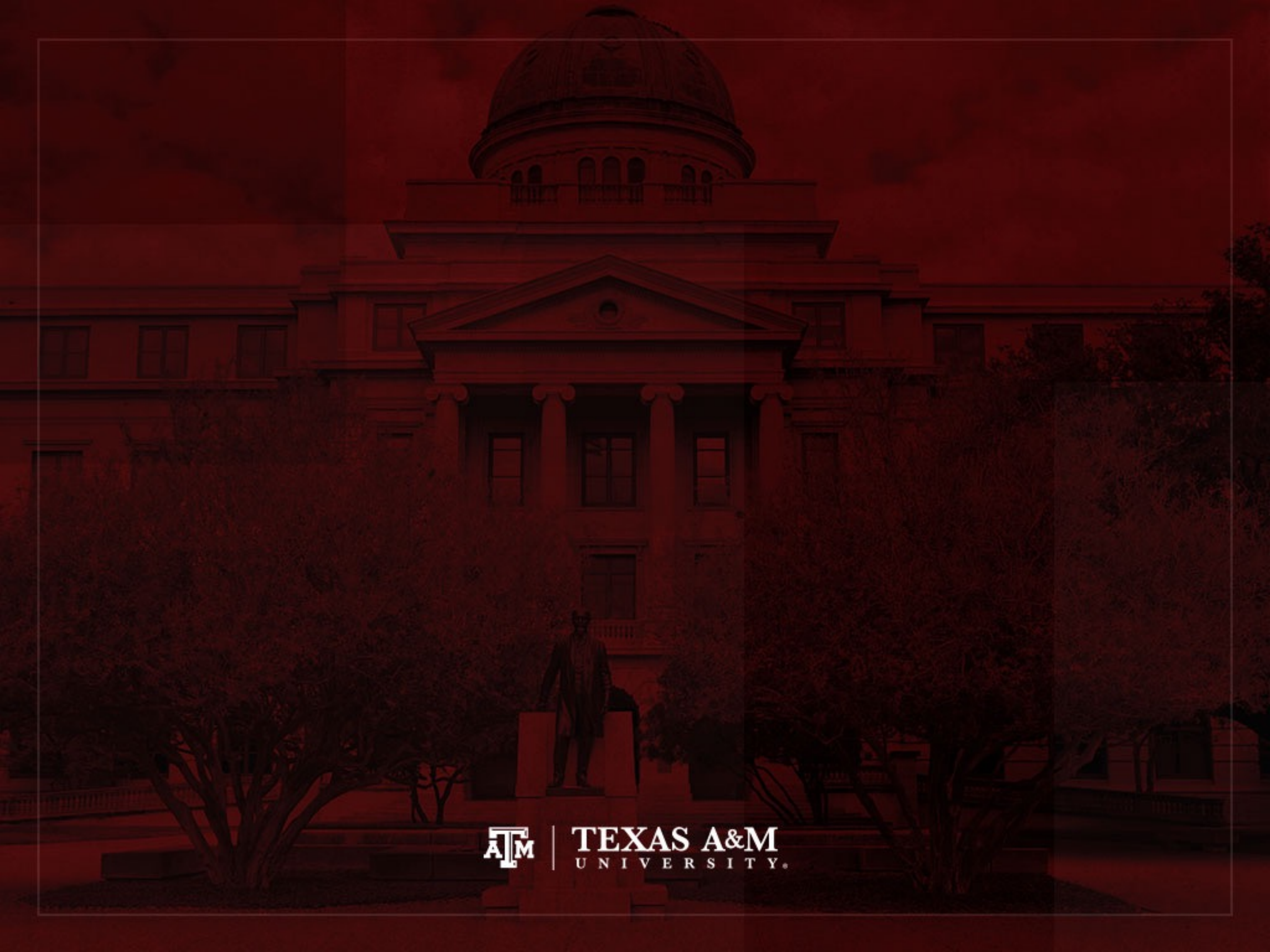
  - **YES**: most appropriate measure for each level
  - Yes: measure is also permitted
  - Yes (?): mean is often used with ordinal-level variables, but this practice violates level-of-measurement guidelines
  - No: cannot be computed for that level

Source: Healey 2015, p.80.

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. 3. You want to report the central score. The median always lies at the exact center of the distribution.
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.


TEXAS A&M
UNIVERSITY

Measures of dispersion

- Explain the purpose of measures of dispersion
- Compute and interpret these measures
  - Range (*R*), interquartile range (*Q* or *IQR*)
  - Standard deviation (s), variance (s^2)
- Select an appropriate measure of dispersion and correctly calculate and interpret the statistic
- Describe and explain the mathematical characteristics of the standard deviation
- Analyze a boxplot

Concept of dispersion

- Dispersion refers to the variety, diversity, or amount of variation among scores
- The greater the dispersion of a variable, the greater the range of scores and the greater the differences between scores
- Examples
  – Typically, a large city will have more diversity than a small town
  – Some states (California, New York) are more racially diverse than others (Maine, Iowa)

Ambulance assistance

- Examples below have similar means
  - 7.4 minutes for service A and 7.6 minutes for service B
- Service A is more consistent in its response
  - Less dispersion than service B

Source: Healey 2015, p.92.

Range (*R*)

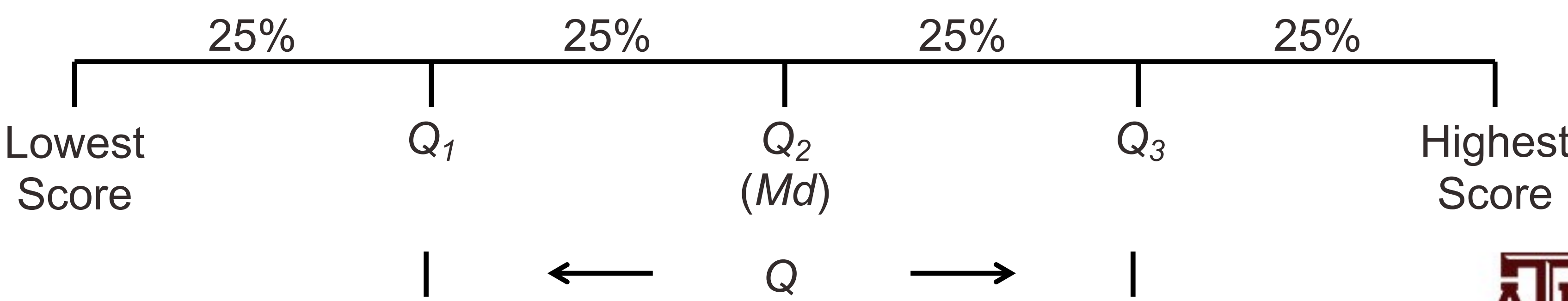
- Range indicates the distance between the highest and lowest scores in a distribution
- Range (*R*) = Highest Score – Lowest Score
- Quick and easy indication of variability
- Can be used with ordinal-level or interval-ratio-level variables
- Why can't the range be used with variables measured at the nominal level?
  - For these variables, use frequency distributions to analyze dispersion

Limitations of range

- Range is based on only two scores
- It is distorted by atypically high or low scores
  - Influenced by outliers
- No information about variation between high and low scores

Interquartile range (*Q* or *IQR*)

- A type of range measure
  - Considers only the middle 50% of the cases in a distribution
- Avoids some of the problems of the range by focusing on just the middle 50% of scores
  - Avoids the influence of outliers



Limitation of interquartile range

- The interquartile range is based on only two scores
- It fails to yield any information from all of the other scores
  - Based only on Q_1 and Q_3

Birth rates for 40 nations, 2012

(number of births per 1000 population)

Rank	Nation	Birth rate	Rank	Nation	Birth rate
40 (highest)	Niger	46	20	Libya	23
39	Uganda	45	19	India	22
38	Malawi	43	18	Venezuela	21
37	Angola	42	17	Mexico	20
36	Mozambique	42	16	Colombia	19
35	Tanzania	41	15	Kuwait	18
34	Nigeria	40	14	Vietnam	17
33	Guinea	39	13	Ireland	16
32	Senegal	38	12	Chile	15
31	Togo	36	11	Australia	14
30	Kenya	35	10	United States	13
29	Ethiopia	34	9	United Kingdom	13
28	Rwanda	33	8	Russia	13
27	Ghana	32	7	France	13
26	Guatemala	29	6	China	12
25	Pakistan	28	5	Canada	11
24	Haiti	27	4	Spain	10
26	Cambodia	26	3	Japan	9
22	Egypt	25	2	Italy	9
21	Syria	24	1 (lowest)	Germany	8

Source: Healey 2015, p.94.

Examples of *R* and *IQR*

- Range = Highest score – Lowest score = 46 – 8 = 38

- Interquartile range (IQR)
  - Locate Q_3 (75th percentile) and Q_1 (25th percentile)
  - Q_3: 0.75 x 40 = 30th case
    - Kenya is the 30th case with a birth rate of 35
  - Q_1: 0.25 x 40 = 10th case
    - United States is the 10th case with a birth rate of 13
  - Difference of these values is interquartile range
    - $IQR = Q3 – Q1 = 35 – 13 = 22$

Standard deviation

- The most important and widely used measure of dispersion
  - It should be used with interval-ratio-level variables, but is often used with ordinal-level variables
- Good measure of dispersion
  - Uses all scores in the distribution
  - Describes the average or typical deviation of the scores
  - Increases in value as the distribution of scores becomes more diverse

Interpreting standard deviation

- It is an index of variability that increases in value as the distribution becomes more variable
- It allows us to compare distributions
- It can be interpreted in terms of normal deviation
  - We will discuss on Chapter 5

Formulas

- Standard deviation and variance are based on the distance between each score and the mean
- Formula for variance

$$s^2 = \frac{\sum(X_i - \bar{X})^2}{N}$$

- Formula for standard deviation

$$s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{N}}$$

Step-by-step calculation of *s*

- Subtract mean from each score: $(X_i - \bar{X})$
- Square the deviations: $(X_i - \bar{X})^2$
- Sum the squared deviations: $\sum(X_i - \bar{X})^2$
- Divide the sum of squared deviations by *N*:

$$\frac{\sum(X_i - \bar{X})^2}{N}$$

- Square root brings value back to original unit:

$$\sqrt{\frac{\sum(X_i - \bar{X})^2}{N}}$$

Residential campus	Age (X_i)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	18	18 – 19 = –1	$(-1)^2 = 1$
	19	19 – 19 = 0	$(0)^2 = 0$
	20	20 – 19 = 1	$(1)^2 = 1$
	18	18 – 19 = –1	$(-1)^2 = 1$
	20	20 – 19 = 1	$(1)^2 = 1$
	$\sum(X_i)$ = 95 **$\bar{X}$ = 95/5 = 19**	**$\sum(X_i - \bar{X})$ = 0**	**$\sum(X_i - \bar{X})^2$ = 4** **$s = \sqrt{4/5}$ = 0.89**

Urban campus	Age (X_i)	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
	20	20 – 23 = –3	$(-3)^2 = 9$
	22	22 – 23 = –1	$(-1)^2 = 1$
	18	18 – 23 = –5	$(-5)^2 = 25$
	25	25 – 23 = 2	$(2)^2 = 4$
	30	30 – 23 = 7	$(7)^2 = 49$
	$\sum(X_i)$ = 115 **$\bar{X}$ = 115/5 = 23**	**$\sum(X_i - \bar{X})$ = 0**	**$\sum(X_i - \bar{X})^2$ = 88** **$s = \sqrt{88/5}$ = 4.20**

This residential campus is less diverse with respect to age (s=0.9) than this urban campus (s=4.2).

Source: Healey 2015, p.100.

Homicides per 100,000 population

New England states

State	Homicide rate	Deviation	Deviation squared
Connecticut	3.6	0.88	0.77
Massachusetts	3.2	0.48	0.23
Rhode Island	2.8	0.08	0.01
Vermont	2.2	–0.52	0.27
Maine	1.8	–0.92	0.85
	$\sum(X_i)$ = 13.6 $\bar{X}$ = 2.72	$\sum(X_i - \bar{X})$ = 0	$\sum(X_i - \bar{X})^2$ = 2.13 $s = \sqrt{2.13/5}$ = 0.66

Western states

State	Homicide rate	Deviation	Deviation squared
Arizona	6.4	2.02	4.08
Nevada	5.9	1.52	2.31
California	4.9	0.52	0.27
Oregon	2.4	–1.98	3.92
Washington	2.3	–2.08	4.33
	$\sum(X_i)$ = 21.9 $\bar{X}$ = 4.38	$\sum(X_i - \bar{X})$ = 0	$\sum(X_i - \bar{X})^2$ = 14.91 $s = \sqrt{14.91/5}$ = 1.73

Source: Healey 2015, p.105.

Reporting several variables

- Measures of central tendency (e.g., mean) and dispersion (e.g., standard deviation)
  - Valuable descriptive statistics
  - Basis for many analytical techniques
  - Most often presented in summary tables

Characteristics of the sample

Variable	Mean	Standard deviation	Number of cases
Age	33.2	1.3	1,078
Number of children	2.3	0.7	1,078
Years married	7.8	1.5	1,052
Income (in dollars)	55,786	1,500	987

Source: Healey 2015, p.110.

Parental engagement

- Means and standard deviations for number of days per week each parent engaged with child
  - How does maternal engagement compare to paternal engagement?
  - How does married engagement compare to cohabiting engagement?
  - How does engagement change over time?

Parental engagement by age of child, gender, and marital status

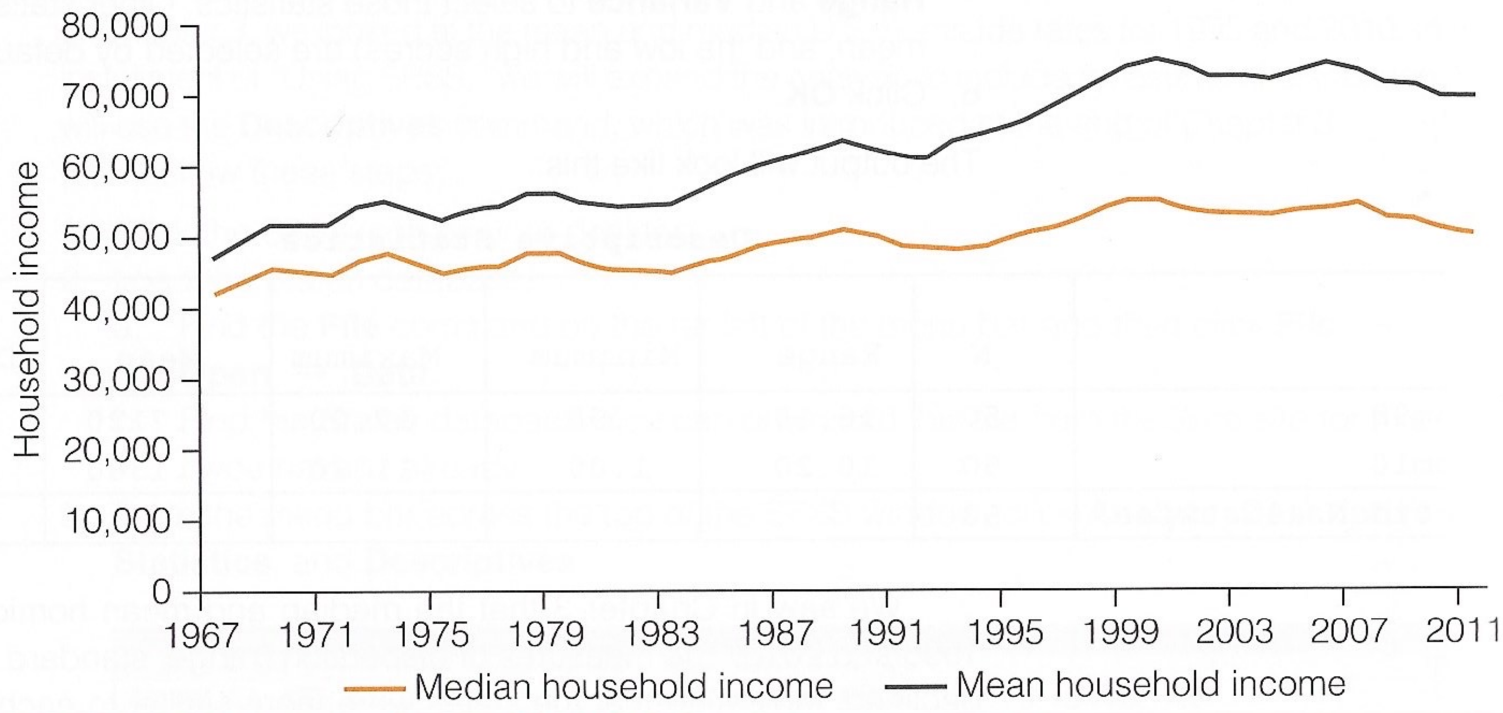
Marital status	**Maternal engagement**				**Paternal engagement**			
	1 year old		3 years old		1 year old		3 years old	
	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
Married	5.30	1.40	4.95	1.33	4.64	1.75	4.01	1.43
Cohabiting	5.23	1.36	4.86	1.38	4.67	1.58	4.04	1.53

Source: Healey 2015, p.110.

Income: Central tendency

- Median
  - Increases in income of the average American household
- Mean
  - Increases in average income for all American households

Median and mean household incomes, United States, 1967–2011

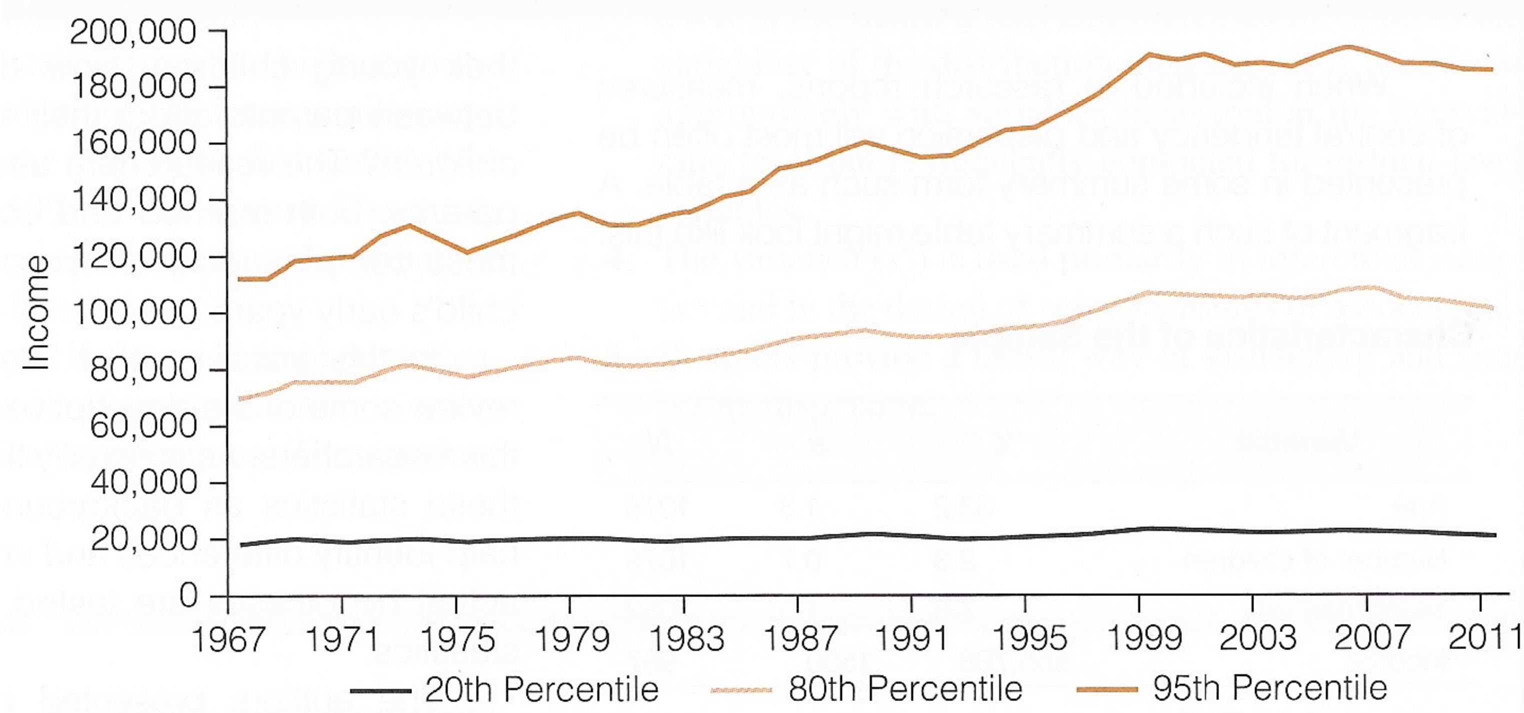


Source: Healey 2015, p.108.

Income: Dispersion increased

- The increase was not shared equally
  - Low-income households: no growth
  - High-income households: robust increases

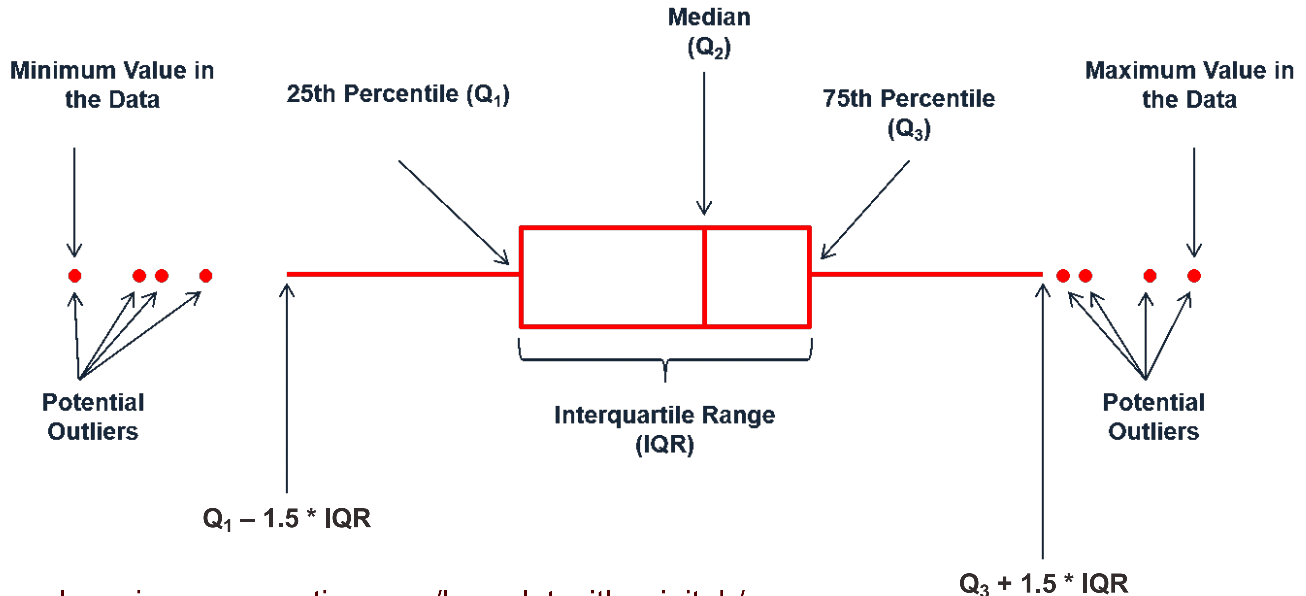
Percentiles of household income, United States, 1967–2011



Source: Healey 2015, p.109.

Boxplots

- Boxplot is also known as ”box and whiskers plot”
  - It provides a way to visualize and analyze dispersion
  - Useful when comparing distributions
  - It uses median, range, interquartile range, outliers
  - Easier to read all this information than in tables



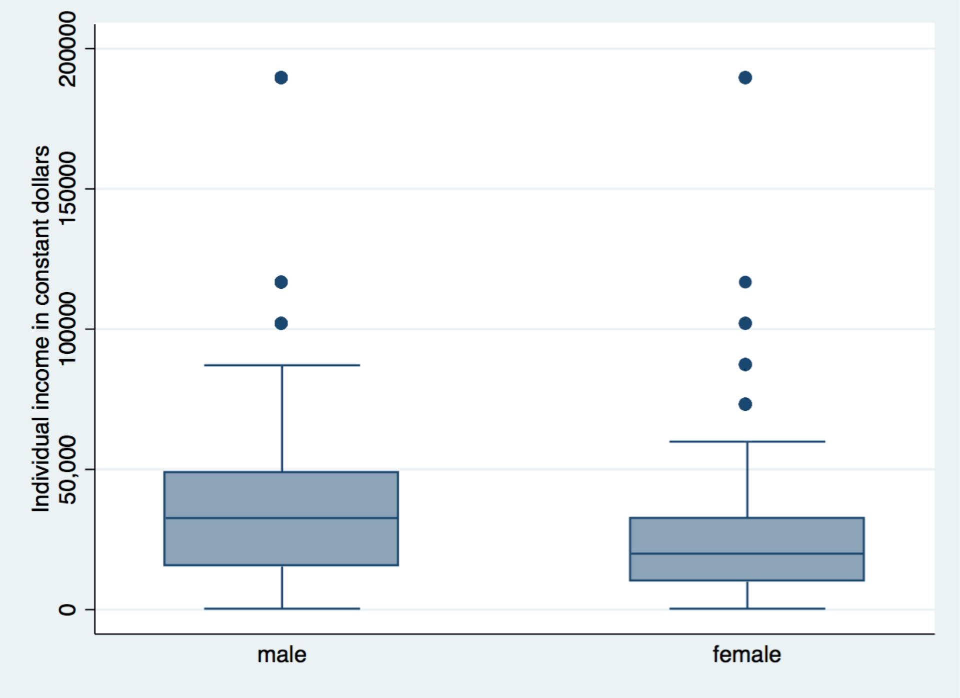
Source: https://www.leansigmacorporation.com/box-plot-with-minitab/

Income by sex, 2016

Statistics for individual income	Male	Female
Lowest score	363.00	363.00
Q1	15,427.50	9,982.50
Median	32,670.00	19,965.00
Q3	49,005.00	32,670.00
Highest score	189,211.46	189,211.46
IQR	33,577.50	22,687.50
Mean	41,282.78	28,109.34
Standard deviation	41,295.31	30,201.87

Commands in Stata

```
tabstat conrinc [aweight=wtssall],
by(sex) stat(min p25 p50 p75 max iqr
mean sd)

graph box conrinc [aweight=wtssall],
over(sex) ytitle(Individual income in
constant dollars)
```



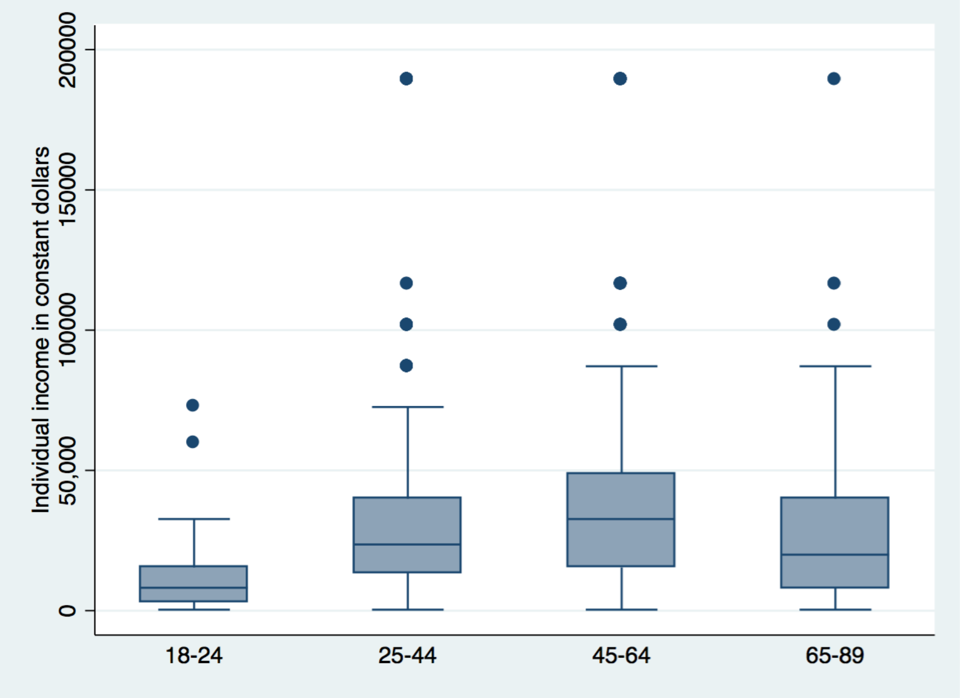
Source: 2016 General Social Survey.

Income by age group, 2016

Statistics for individual income	18–24	25–44	45–64	65–89
Lowest score	363.00	363.00	363.00	363.00
Q1	3,267.00	13,612.50	15,427.50	8,167.50
Median	8,167.50	23,595.00	32,670.00	19,965.00
Q3	15,427.50	39,930.00	49,005.00	39,930.00
Highest score	72,600.00	189,211.46	189,211.46	189,211.46
IQR	12,160.50	26,317.50	33,577.50	31,762.50
Mean	11,214.16	32,863.93	42,552.21	30,848.29
Standard deviation	11,787.32	33,269.47	41,486.09	33,303.36

Commands in Stata

```
tabstat conrinc [aweight=wtssall],
by(agegr1) stat(min p25 p50 p75 max iqr
mean sd)

graph box conrinc [aweight=wtssall],
over(agegr1) ytitle(Individual income in
constant dollars)
```



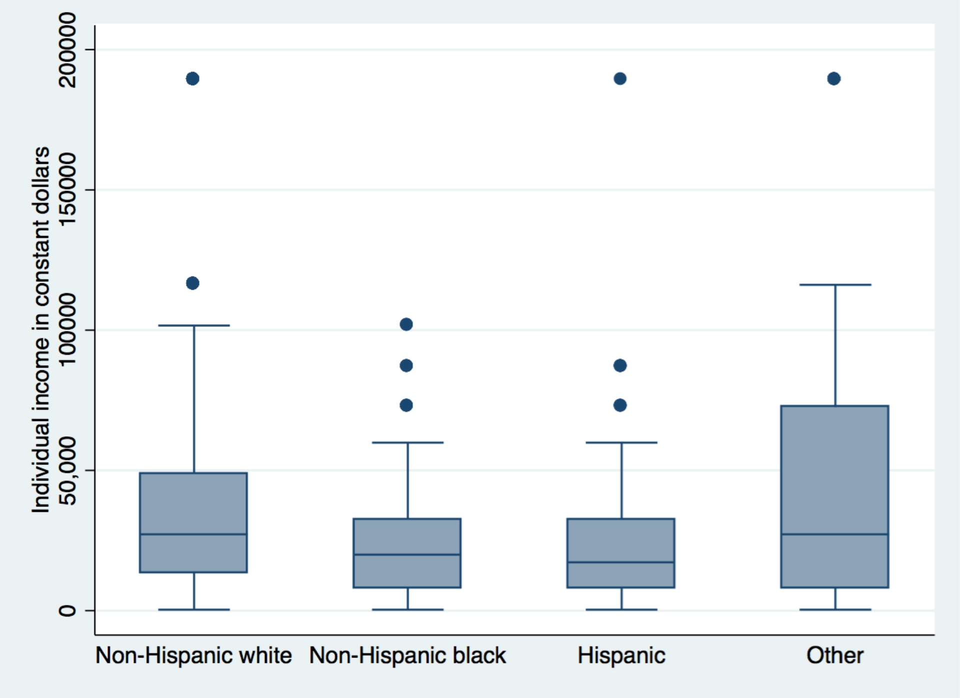
Source: 2016 General Social Survey.

Income by race/ethnicity, 2016

Statistics for individual income	Non-Hispanic white	Non-Hispanic black	Hispanic	Other
Lowest score	363.00	363.00	363.00	363.00
Q1	13,612.50	8,167.50	8,167.50	8,167.50
Median	27,225.00	19,965.00	17,242.50	27,225.00
Q3	49,005.00	32,670.00	32,670.00	72,600.00
Highest score	189,211.46	101,640.00	189,211.46	189,211.46
IQR	35,392.50	24,502.50	24,502.50	64,432.50
Mean	38,845.62	23,243.04	23,128.92	50,156.35
Standard deviation	39,157.17	19,671.53	21,406.31	59,219.90

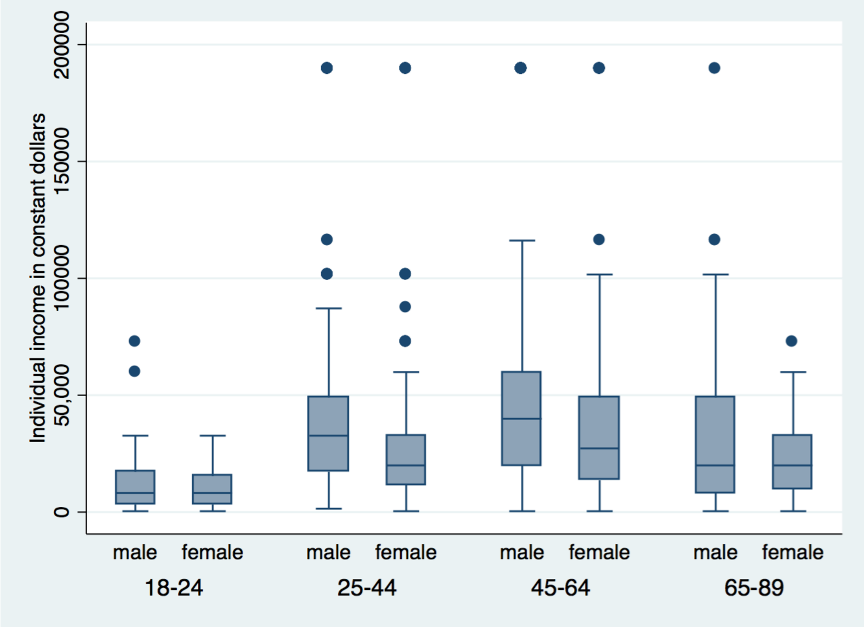
Commands in Stata

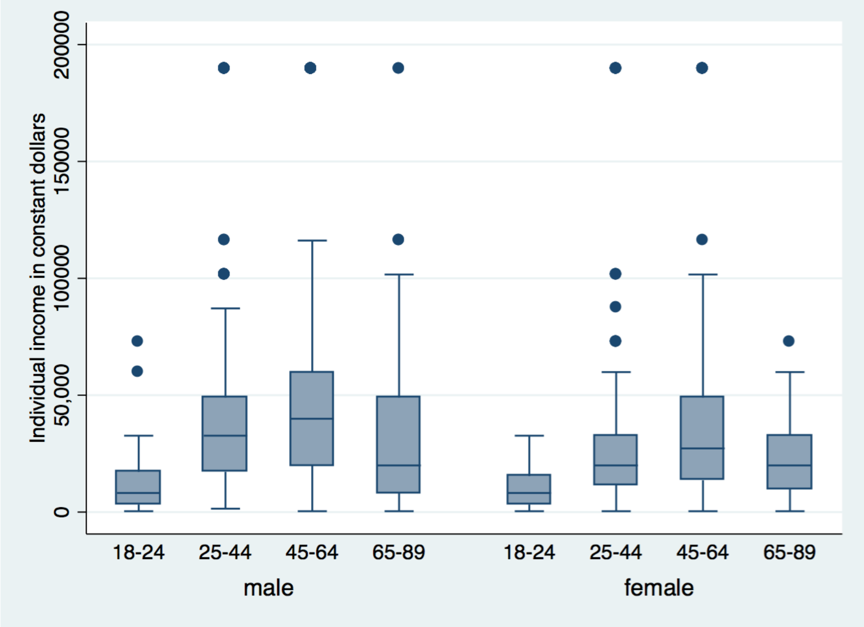
```
tabstat conrinc [aweight=wtssall],
by(raceeth) stat(min p25 p50 p75 max iqr
mean sd)

graph box conrinc [aweight=wtssall],
over(raceeth) ytitle(Individual income
in constant dollars)
```



Source: 2016 General Social Survey.

Income by sex and age group, 2016





Command in Stata

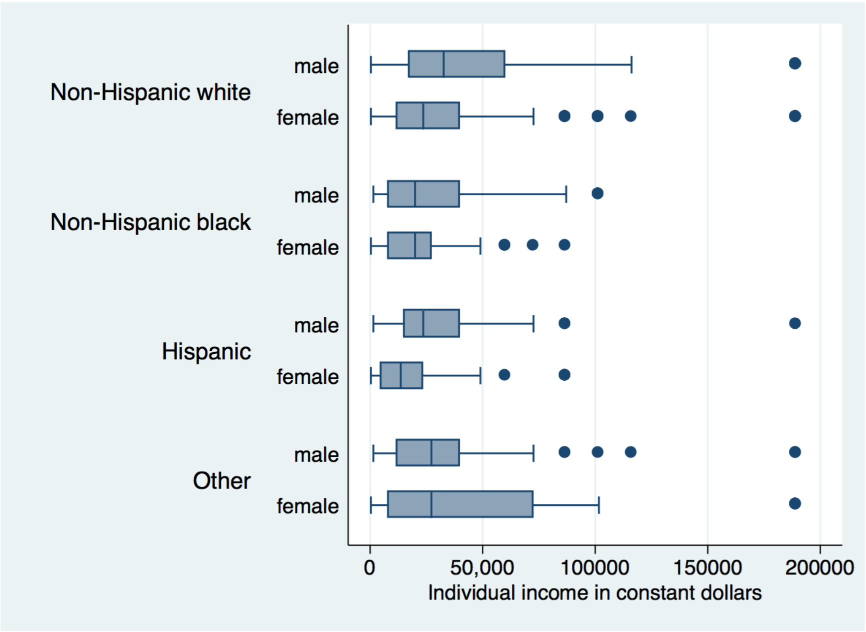
```
graph box conrinc [aweight=wtssall],
over(sex) over(agegr1) ytitle(Individual
income in constant dollars)
```

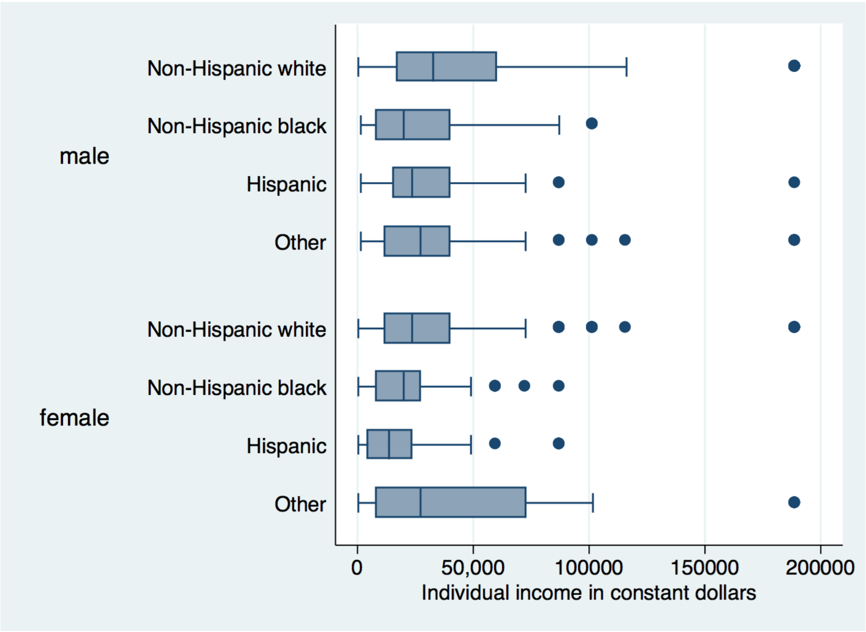
Command in Stata

```
graph box conrinc [aweight=wtssall],
over(agegr1) over(sex) ytitle(Individual
income in constant dollars)
```

Source: 2016 General Social Survey.

Income by sex and race/ethnicity, 2016





Command in Stata

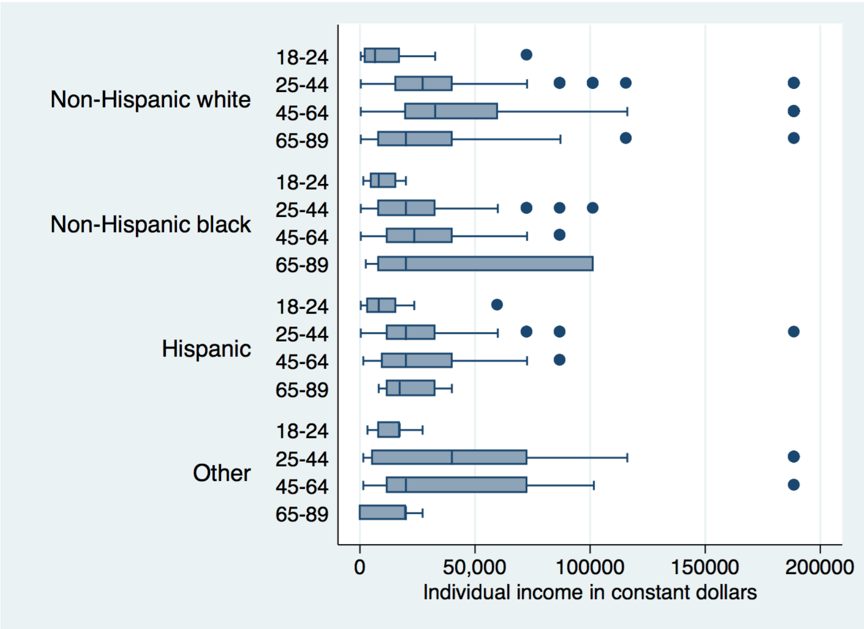
```
graph hbox conrinc [aweight=wtssall],
over(sex) over(raceeth)
ytitle(Individual income in constant
dollars)
```

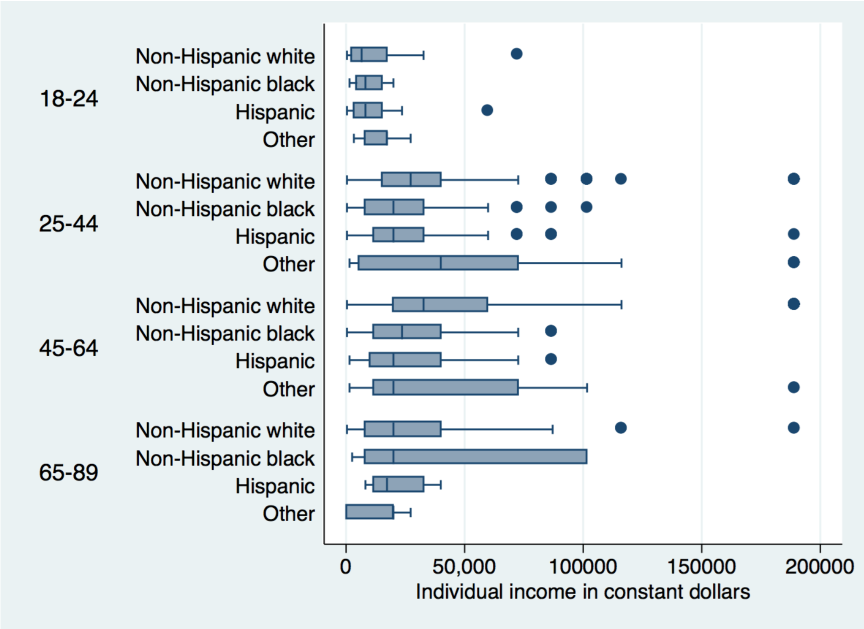
Command in Stata

```
graph hbox conrinc [aweight=wtssall],
over(raceeth) over(sex)
ytitle(Individual income in constant
dollars)
```

Source: 2016 General Social Survey.

Income by age group and race/ethnicity, 2016





Command in Stata

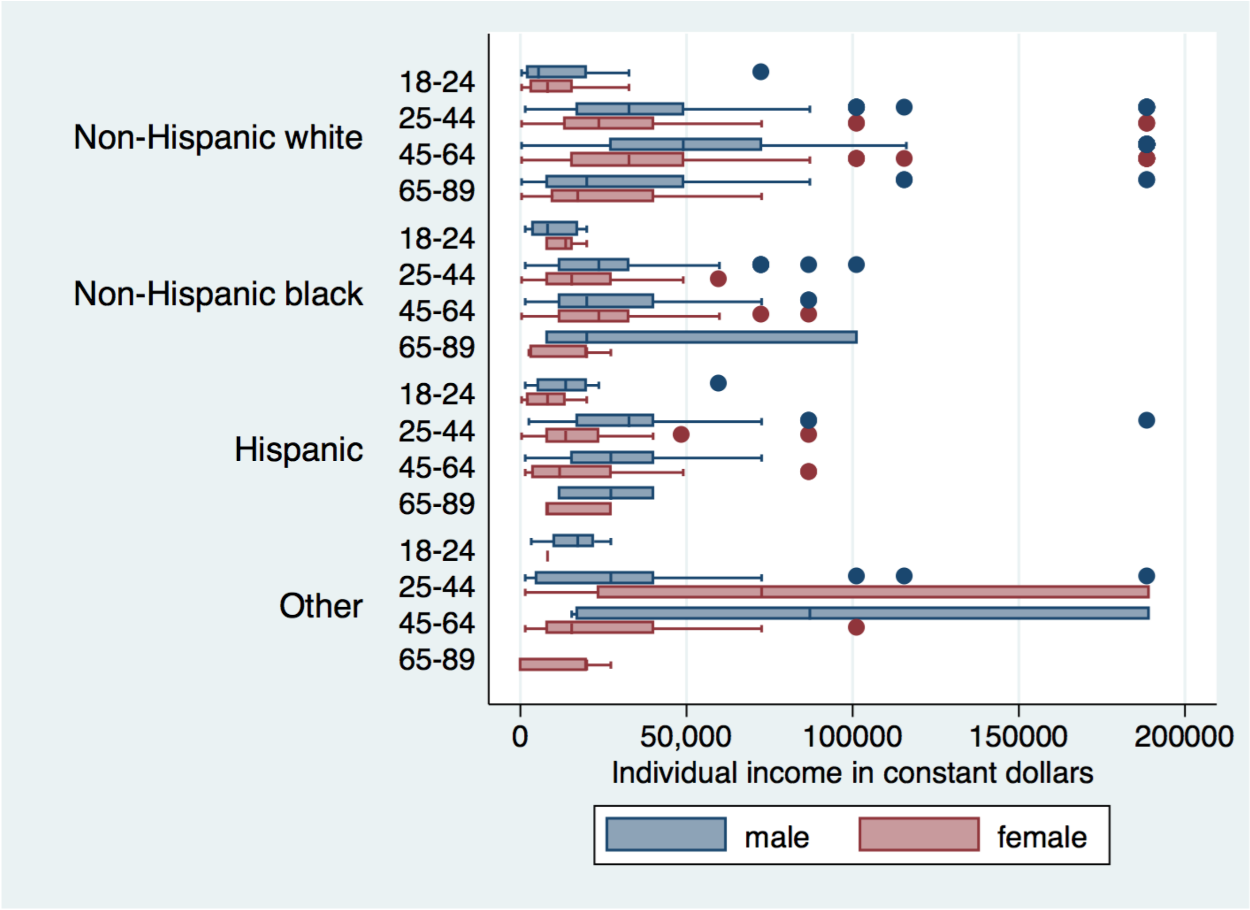
```
graph hbox conrinc [aweight=wtssall],
over(agegr1) over(raceeth)
ytitle(Individual income in constant
dollars)
```

Command in Stata

```
graph hbox conrinc [aweight=wtssall],
over(raceeth) over(agegr1)
ytitle(Individual income in constant
dollars)
```

Source: 2016 General Social Survey.

Income by sex, age group, and race/ethnicity, 2016




```
graph hbox conrinc [aweight=wtssall], over(sex) over(agegr1) over(raceeth)
ytitle(Individual income in constant dollars)
```

Source: 2016 General Social Survey.

Example: 2016 GSS in Stata

- Respondents' income in constant dollars

`sum conrinc [aweight=wtssall], d`

```
                respondent income in constant dollars
-------------------------------------------------------------
      Percentiles      Smallest
 1%          363            363
 5%         1452            363
10%         3993            363       Obs               1,632
25%      11797.5            363       Sum of Wgt.  1,695.2263

50%        23595                      Mean            34649.3
                        Largest       Std. Dev.      36722.06
75%        39930       189211.5
90%        72600       189211.5       Variance       1.35e+09
95%       101640       189211.5       Skewness       2.538394
99%     189211.5       189211.5       Kurtosis       10.63267
```

Example: 2016 GSS in Stata

- Respondents' income in constant dollars

codebook conrinc

```
--------------------------------------------------------------------------------
conrinc                                   respondent income in constant dollars
--------------------------------------------------------------------------------

                  type:  numeric (double)
                 label:  LABW, but 26 nonmissing values are not labeled

                 range:  [363,189211.46]              units:  .01
         unique values:  26                       missing .:  0/2,867
       unique mv codes:  1                       missing .*:  1,235/2,867

              examples:  17242.5
                         39930
                         .i     IAP
                         .i     IAP
```

Edited table

Table 1. Descriptive statistics of respondents' income in constant dollars, U.S. adult population, 2016

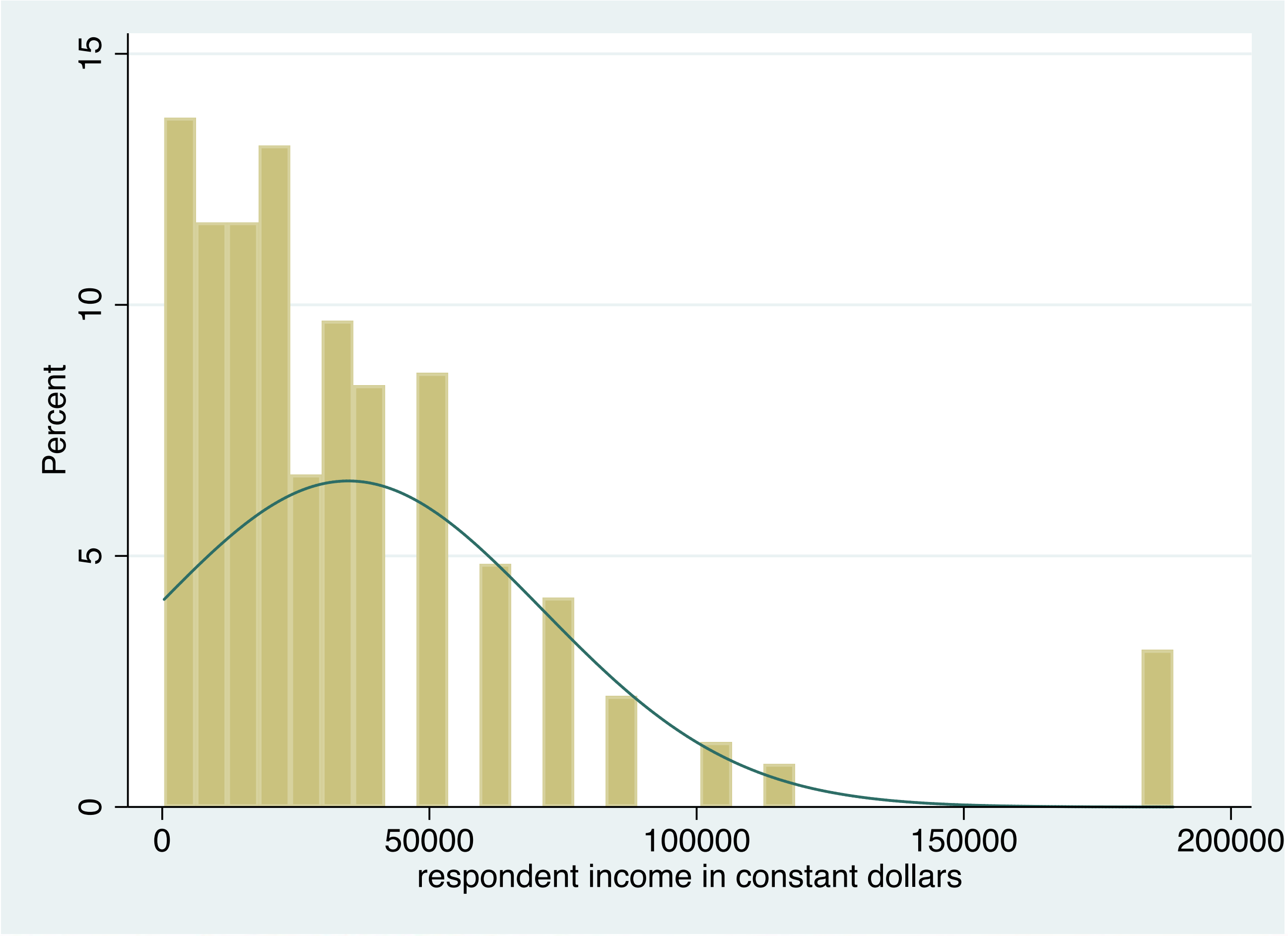
Statistics	Income
Mean	34,649.30
Minimum	363.00
25th percentile	11,797.50
Median	23,595.00
75th percentile	39,930.00
Maximum	189,211.50
Range	188,848.50
Interquartile range	28,132.50
Standard deviation	36,722.06
Sample size	**1,632**
Missing cases	1,235

Source: 2016 General Social Survey.

Example: 2016 GSS in Stata

- Respondents' income in constant dollars

```
hist conrinc, percent normal
```



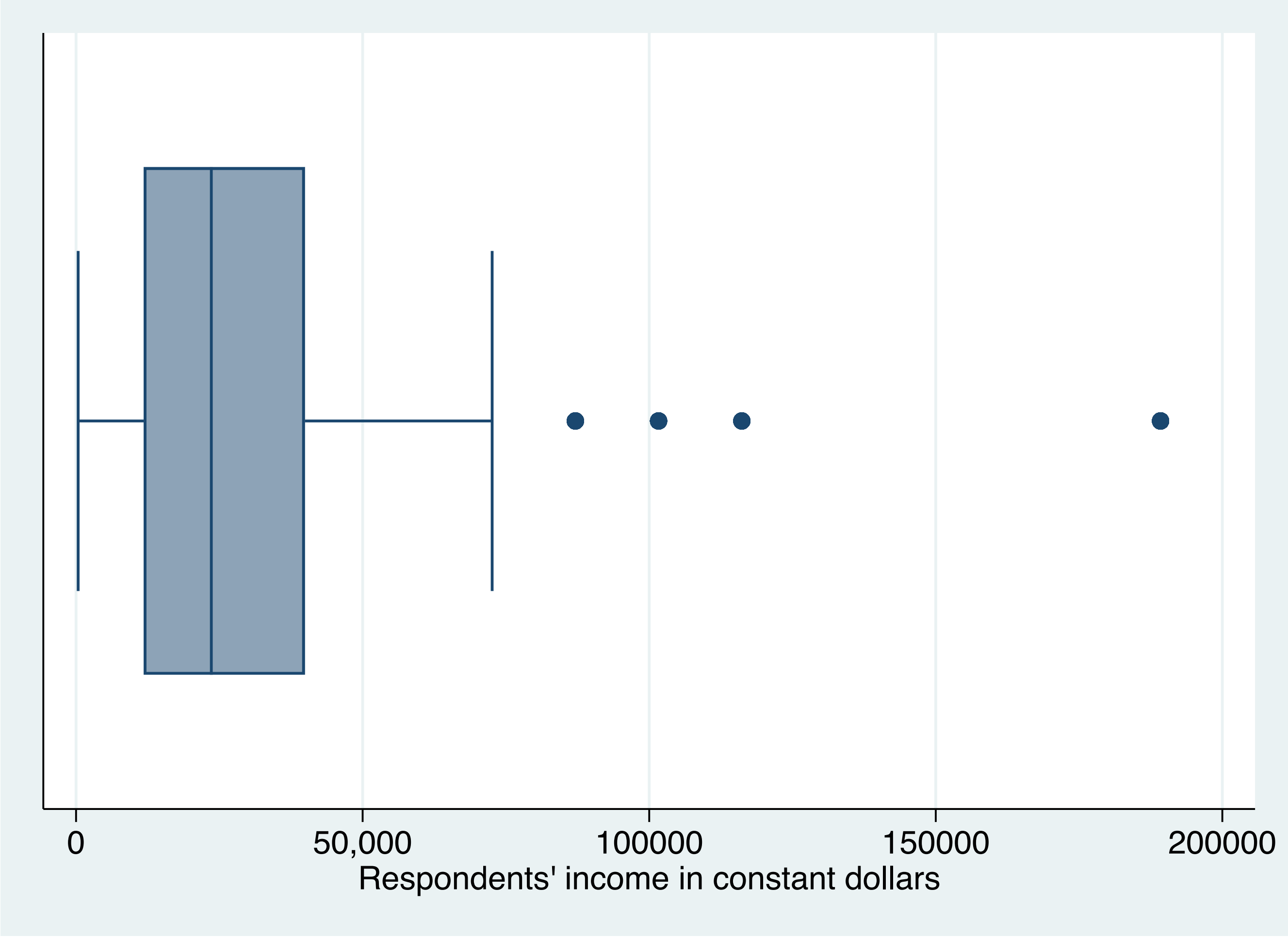
Example: 2016 GSS in Stata

- Generate box plot for respondents' income in constant dollars

```
graph hbox conrinc [aweight=wtssall],
ytitle(Respondents' income in constant dollars)
```

Edited figure

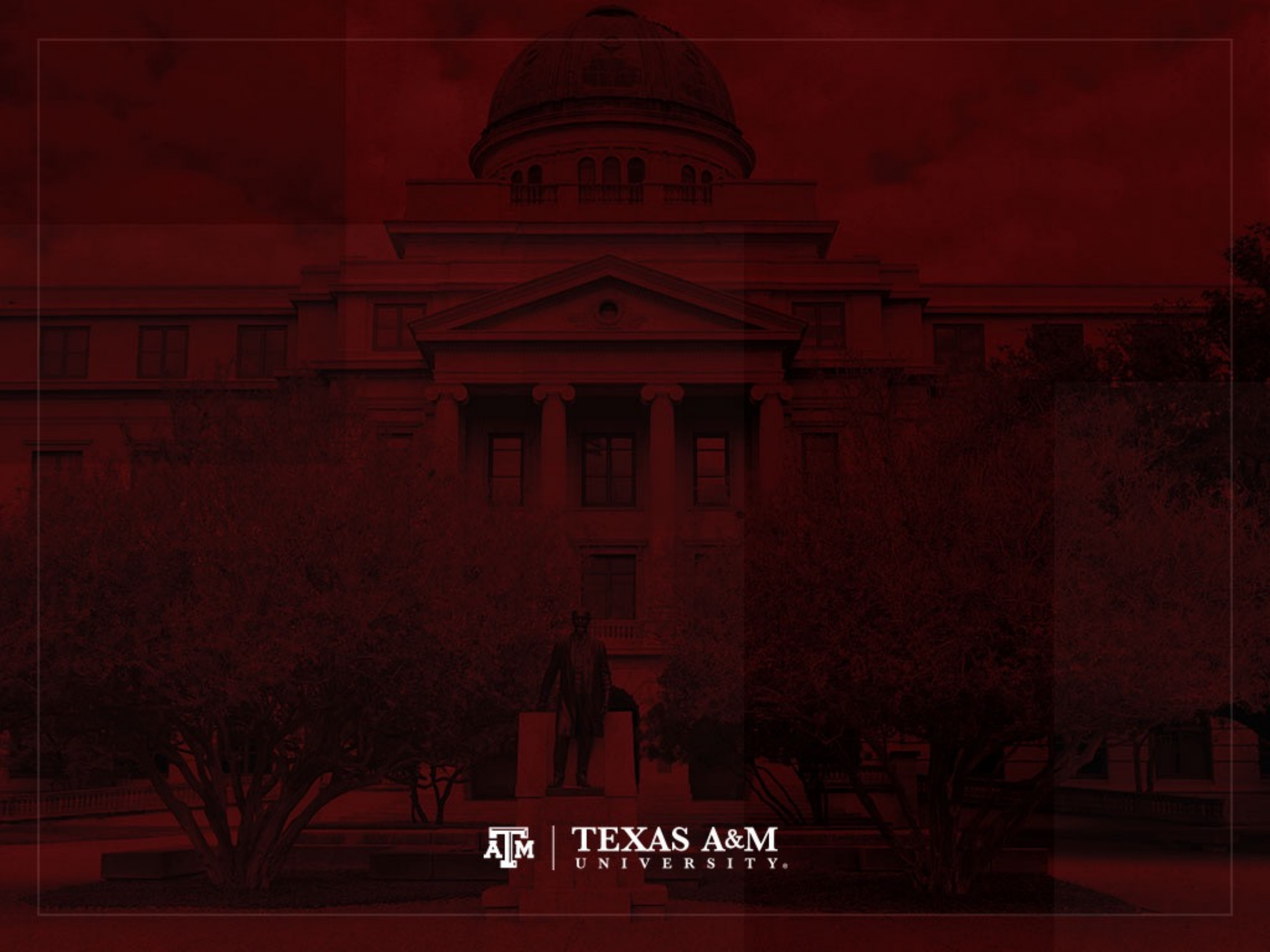
Figure 1. Distribution of respondents' income in constant dollars, U.S. adult population, 2016



Source: 2016 General Social Survey.

Summary

- Measures of dispersions are higher for more diverse groups
  - Larger samples and populations
- Measures of dispersions decrease, as diversity or variety decreases
  - Smaller samples and more homogeneous groups
- The lowest possible value for range and standard deviation is zero
  - In this case, there is no dispersion


TEXAS A&M
UNIVERSITY