Lecture (chapter 4): Measures of dispersion

Ernesto F. L. Amaral

February 5, 2018
Advanced Methods of Social Research (SOCI 420)

Source: Healey, Joseph F. 2015. ''Statistics: A Tool for Social Research.'' Stamford: Cengage Learning. 10th edition. Chapter 4 (pp. 91–121).

TEXAS A&M UNIVERSITY

Chapter learning objectives

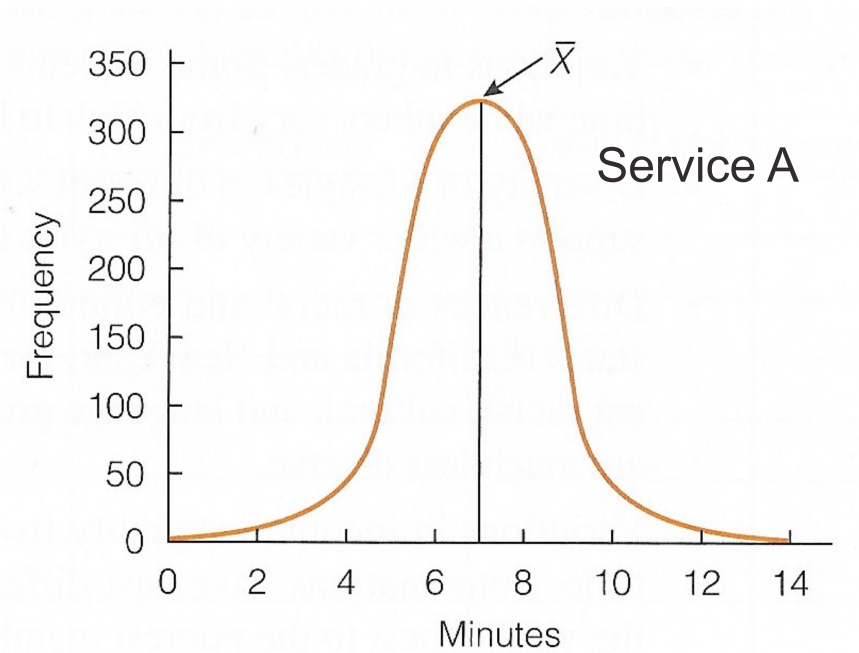
- Explain the purpose of measures of dispersion
- Compute and interpret these measures
  - Range (*R*), interquartile range (*Q* or *IQR*)
  - Standard deviation (s), variance (s^2)
- Select an appropriate measure of dispersion and correctly calculate and interpret the statistic
- Describe and explain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

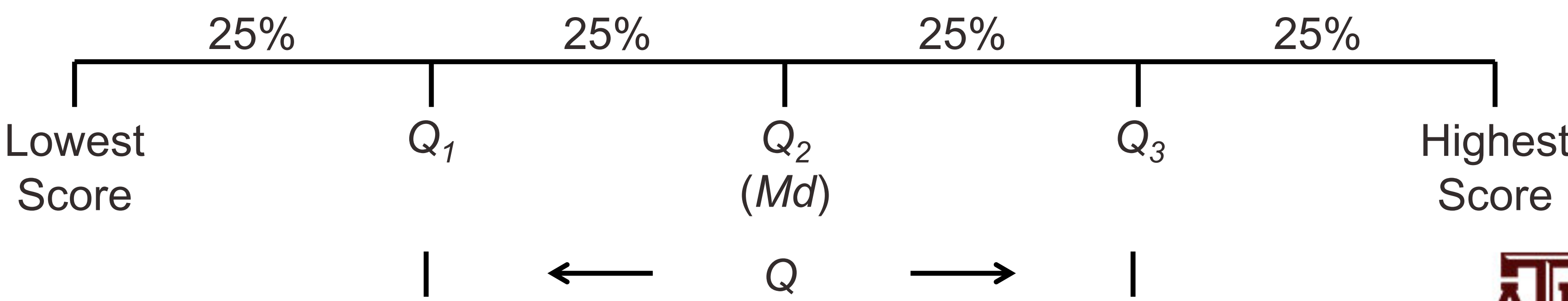
- Range indicates the distance between the highest and lowest scores in a distribution
- Range (*R*) = Highest Score – Lowest Score
- Quick and easy indication of variability
- Can be used with ordinal-level or interval-ratio-level variables
- Why can't the range be used with variables measured at the 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.

Example: 2016 GSS in Stata

- Respondent 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
```

Reporting measures

Descriptive statistics of respondent income in constant dollars, U.S. adult population, 2016

Statistics	**Respondent income**
Minimum	363.00
Mean	34,649.30
Median	23,595.00
Maximum	189,211.50
Range	188,848.50
Interquartile range	28,132.50
Standard deviation	36,722.06
Sample size	**1,632**

Source: 2016 General Social Survey.

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

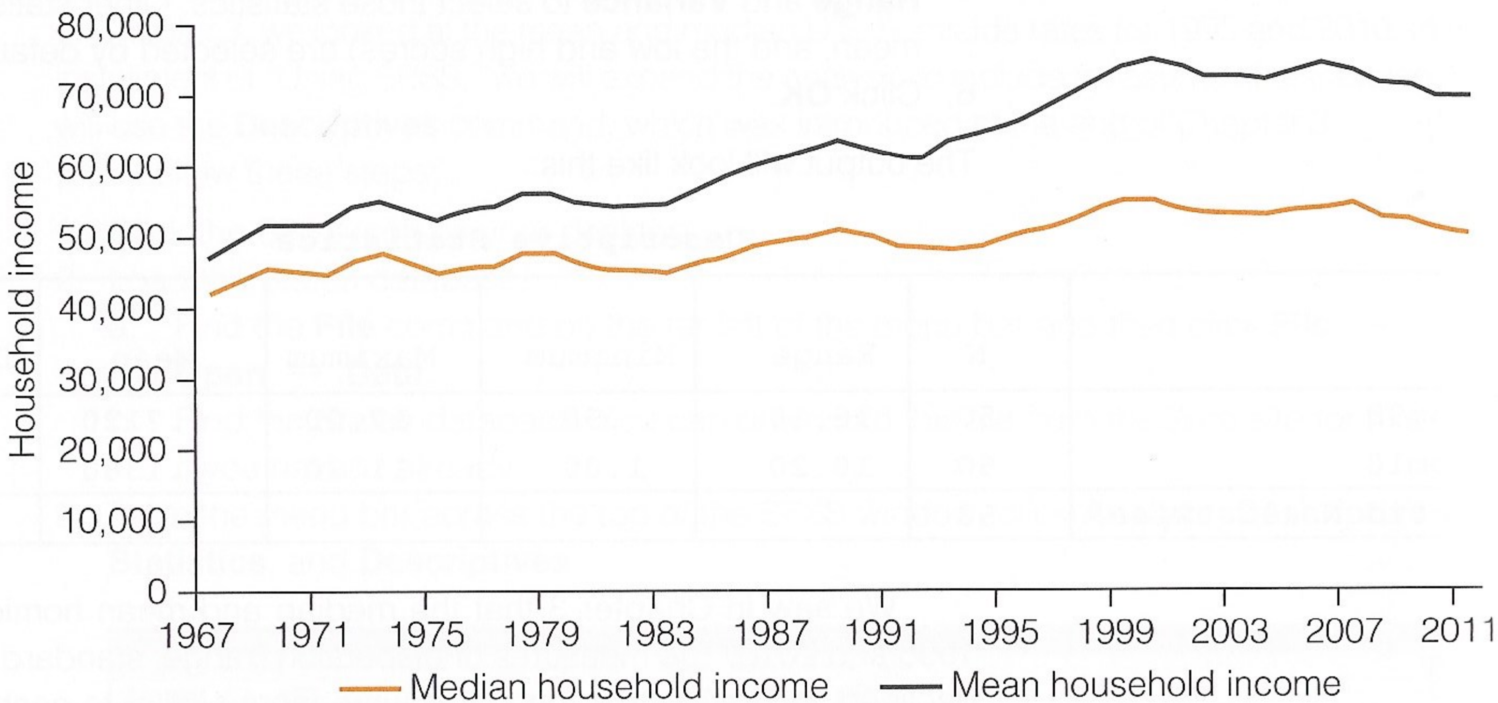
Marital status	**Maternal engagement**				**Paternal engagement**			
	1 year old		3 years old		1 year old		3 years old	
	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s	$\bar{X}$	s
Married	5.30	1.40	4.95	1.33	4.64	1.75	4.01	1.43
Cohabiting	5.23	1.36	4.86	1.38	4.67	1.58	4.04	1.53

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

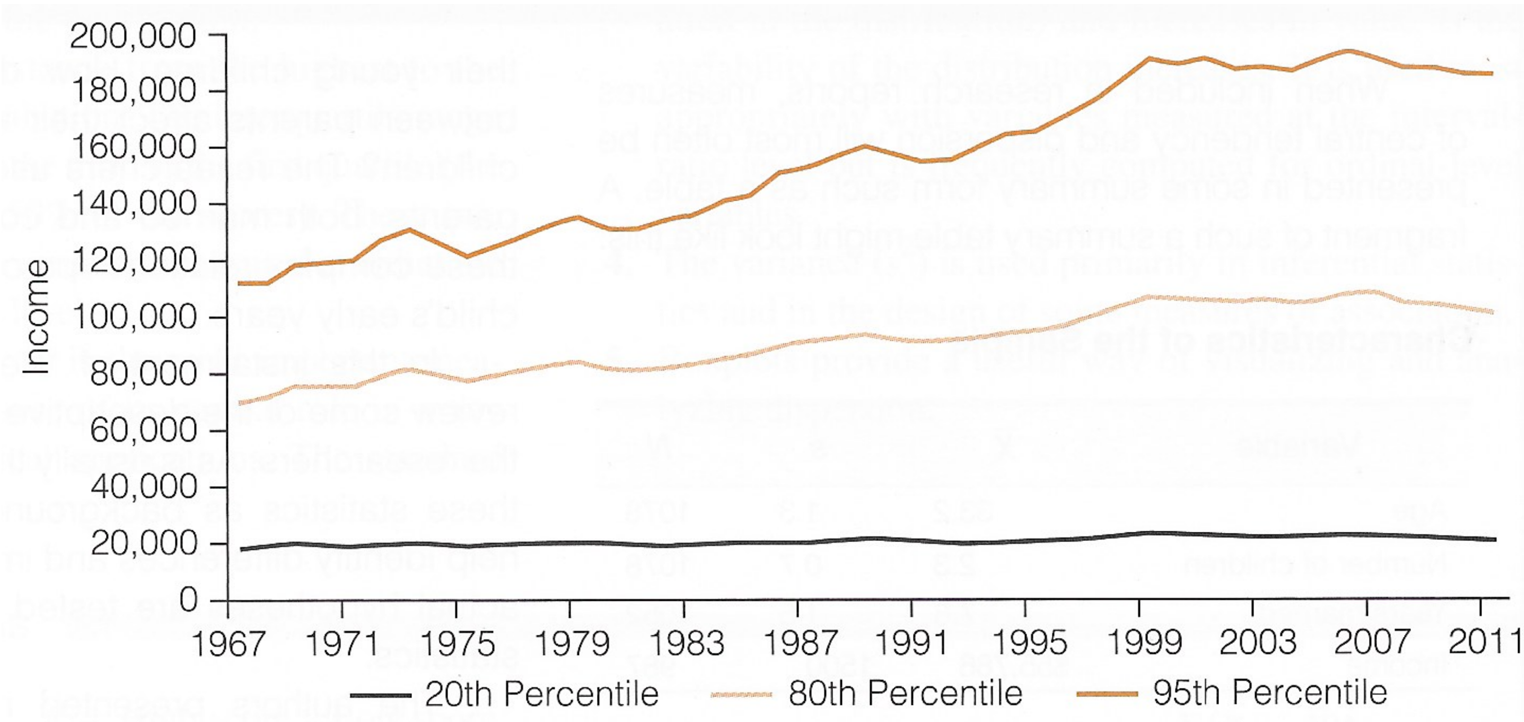


Source: Healey 2015, p.108.

Income: Dispersion increased

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

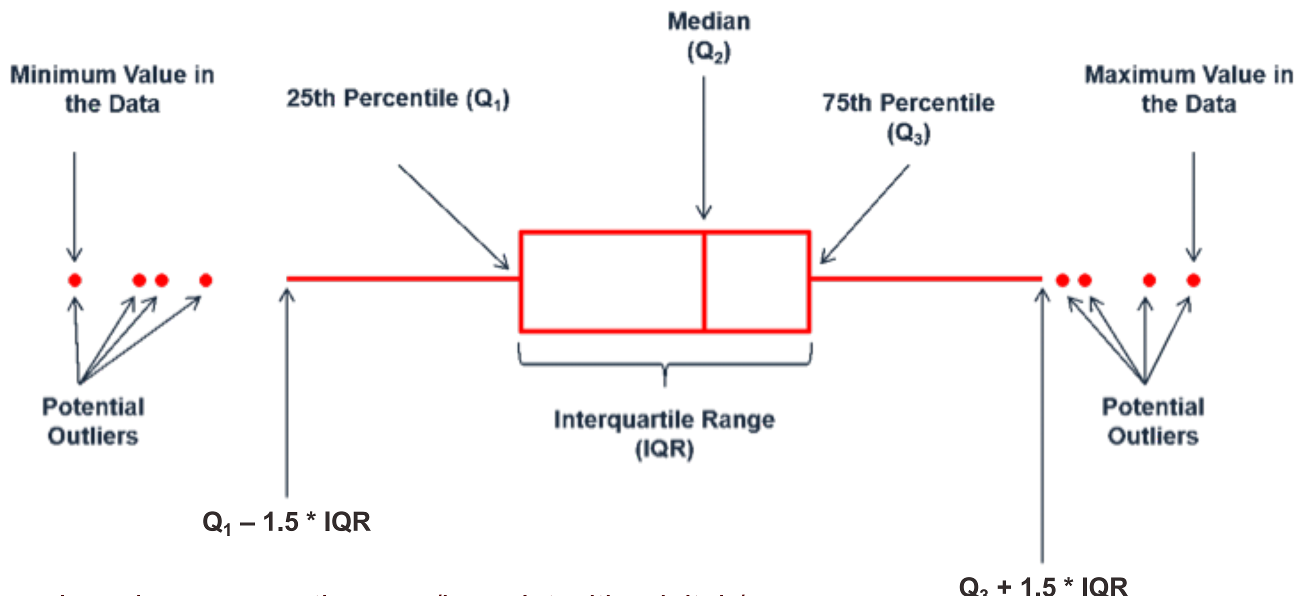
Percentiles of household income, United States, 1967–2011



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



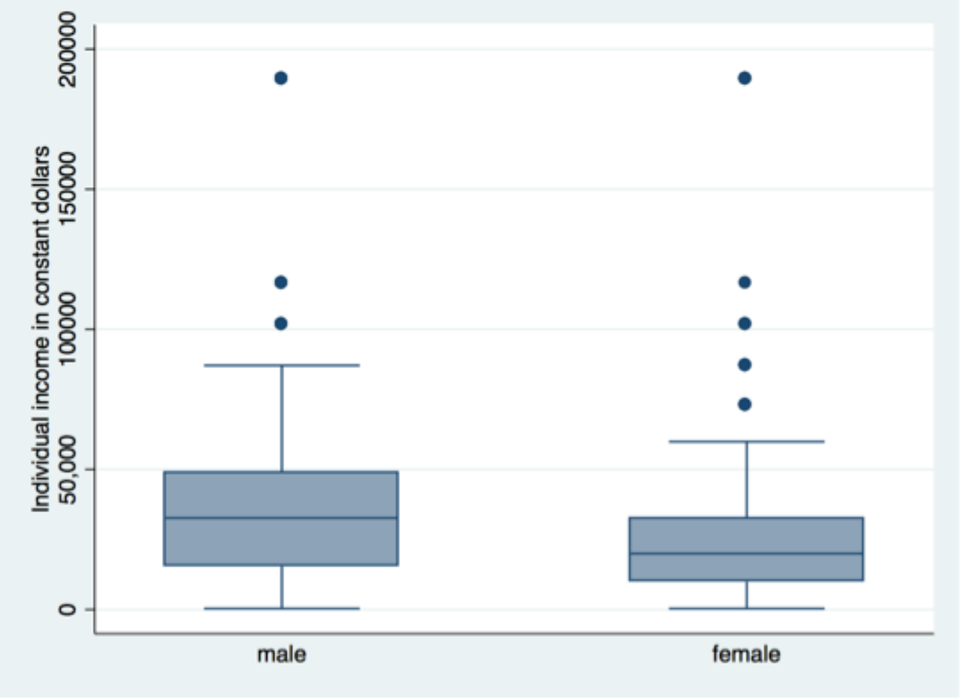
Source: https://www.leansigmacorporation.com/box-plot-with-minitab/

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
Standard deviation	41,295.31	30,201.87
Mean	41,282.78	28,109.34

Commands in Stata

```
table sex [aweight=wtssall], c(min
conrinc p25 conrinc p50 conrinc p75
conrinc max conrinc)

table sex [aweight=wtssall], c(iqr
conrinc sd conrinc mean conrinc)

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



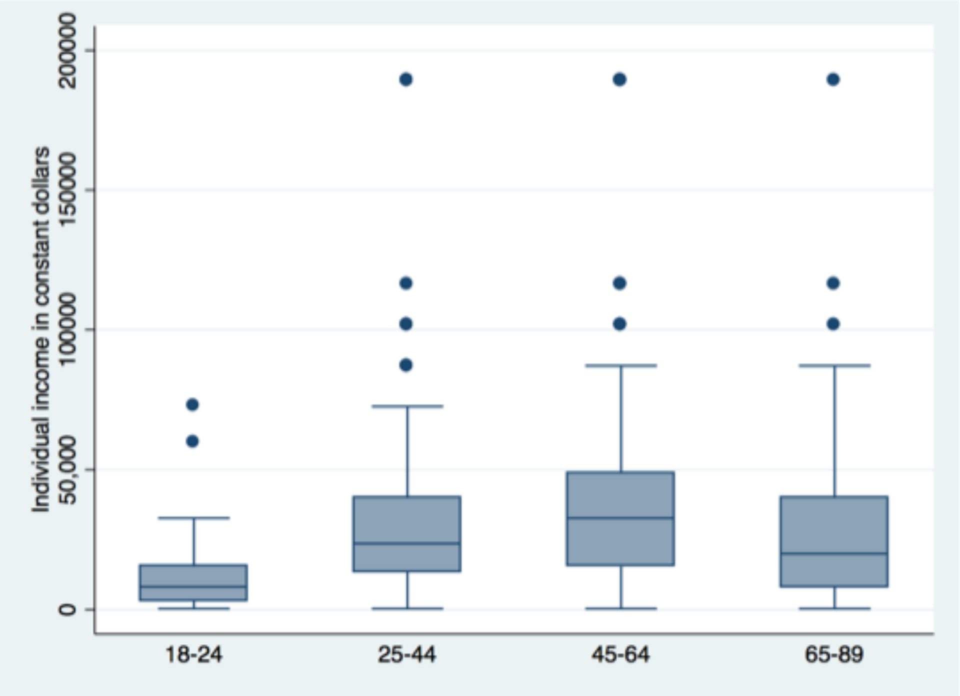
Source: 2016 General Social Survey.

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
Standard deviation	11,787.32	33,269.47	41,486.09	33,303.36
Mean	11,214.16	32,863.93	42,552.21	30,848.29

Commands in Stata

```
table agegr1 [aweight=wtssall], c(min
conrinc p25 conrinc p50 conrinc p75
conrinc max conrinc)

table agegr1 [aweight=wtssall], c(iqr
conrinc sd conrinc mean conrinc)

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



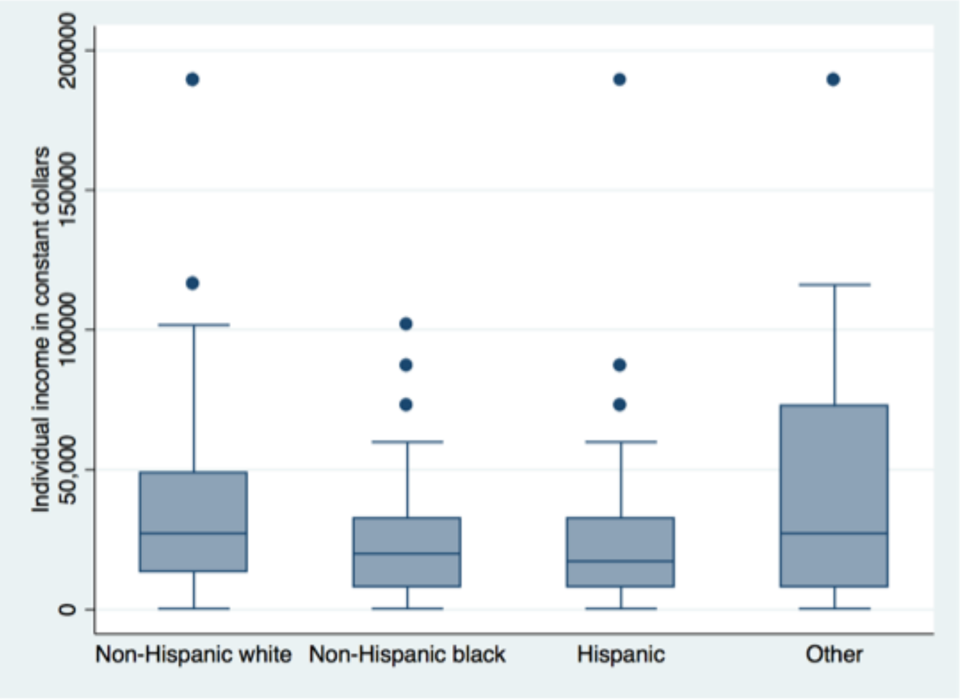
Source: 2016 General Social Survey.

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
Standard deviation	39,157.17	19,671.53	21,406.31	59,219.90
Mean	38,845.62	23,243.04	23,128.92	50,156.35

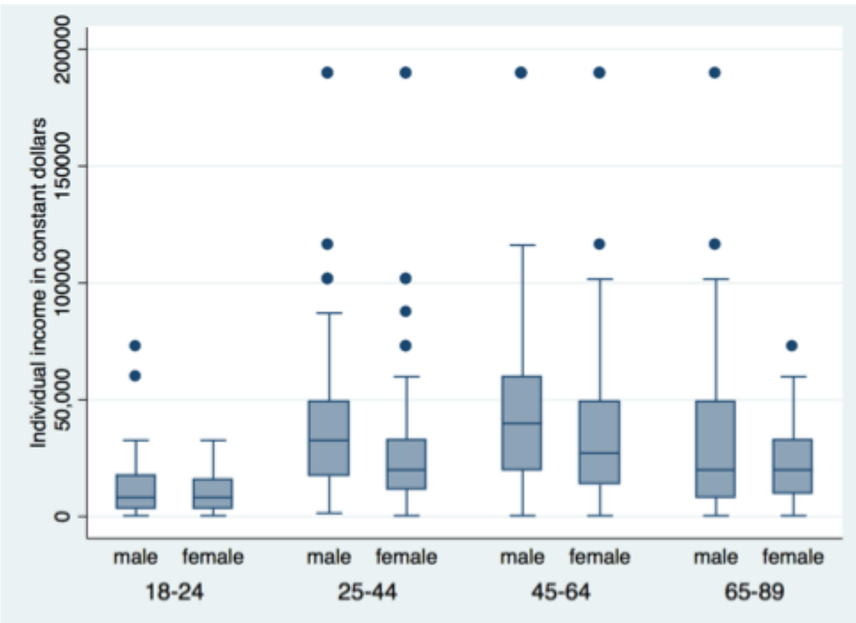
Commands in Stata

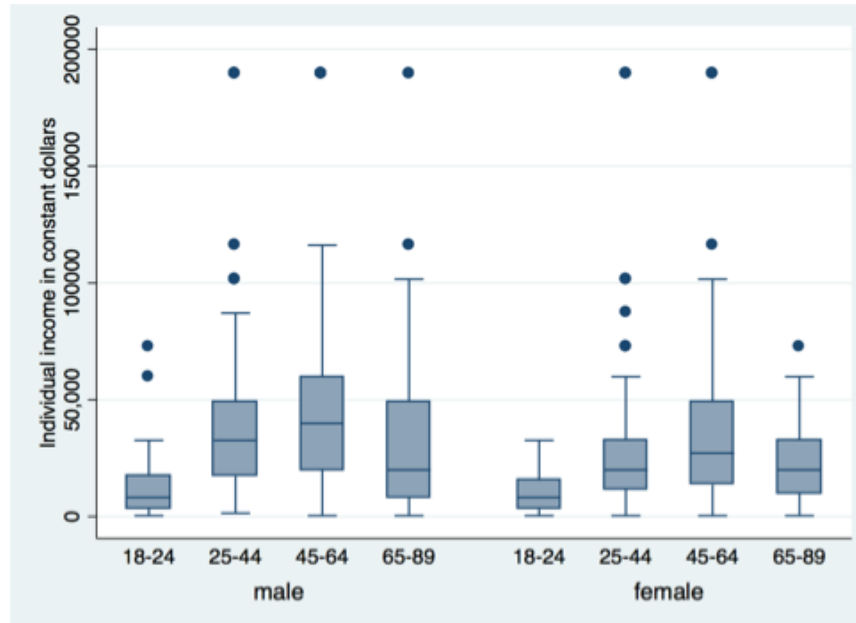
```
table raceeth [aweight=wtssall], c(min
conrinc p25 conrinc p50 conrinc p75
conrinc max conrinc)

table raceeth [aweight=wtssall], c(iqr
conrinc sd conrinc mean conrinc)

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

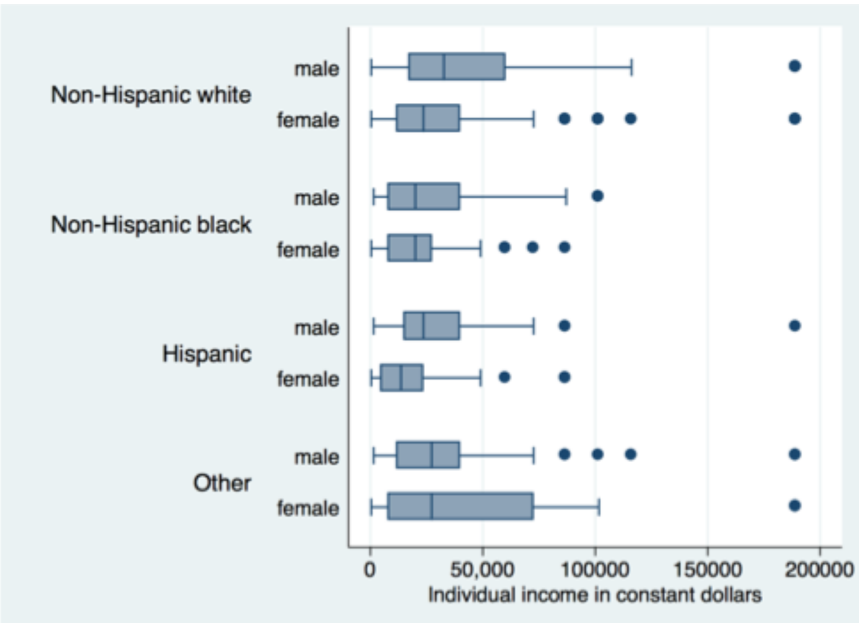
```
graph box conrinc [aweight=wtssall],
over(sex) over(agegr1) ytitle(Individual
income in constant dollars)
```

Command in Stata

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

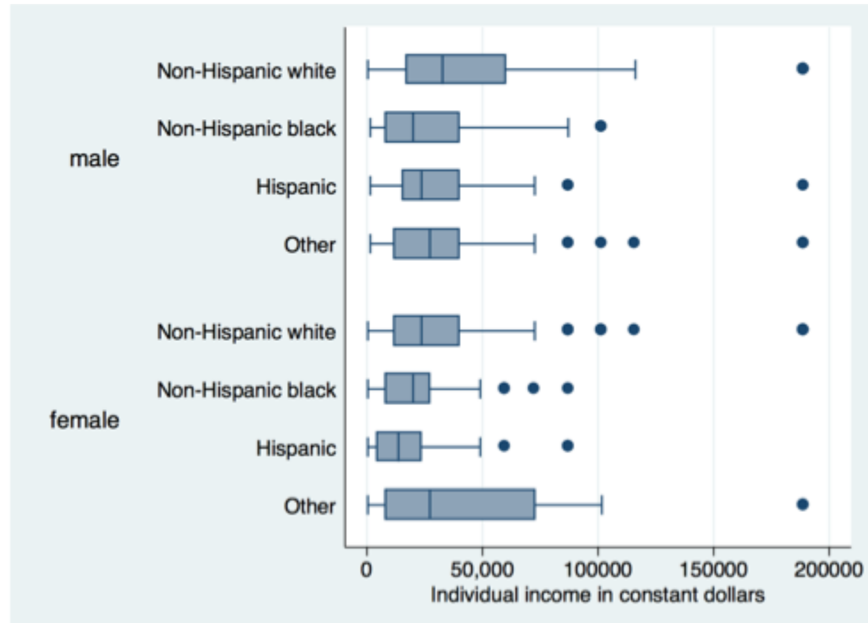
Source: 2016 General Social Survey.

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)
```

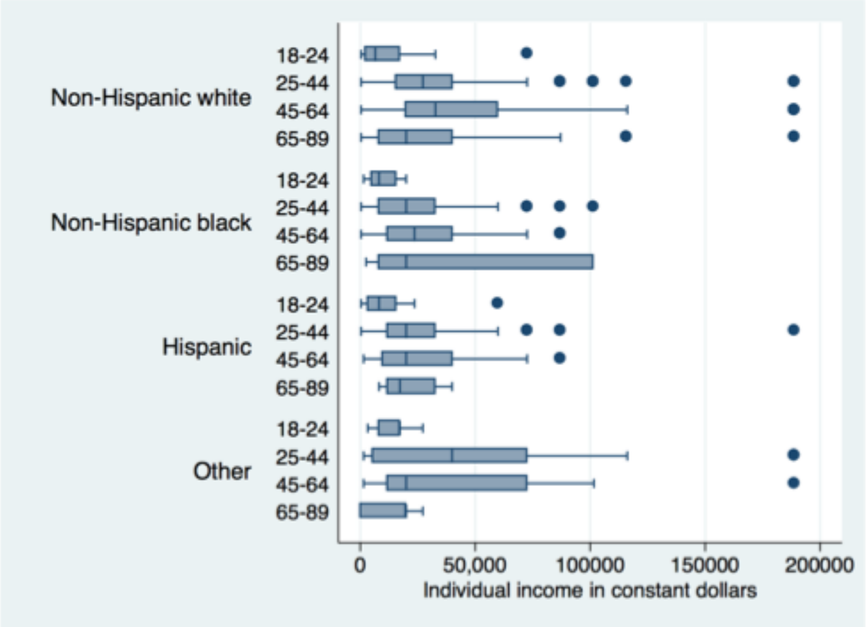


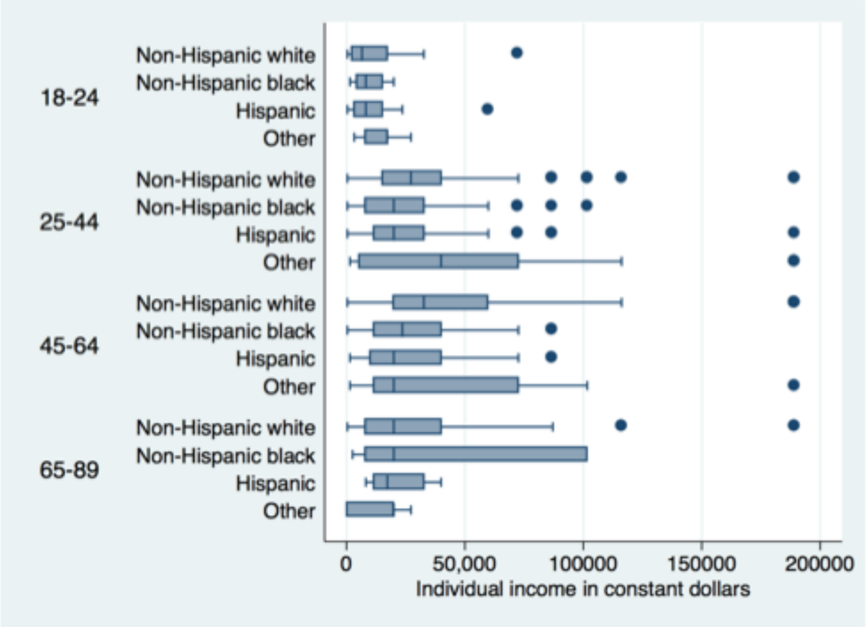
Command in Stata

```
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

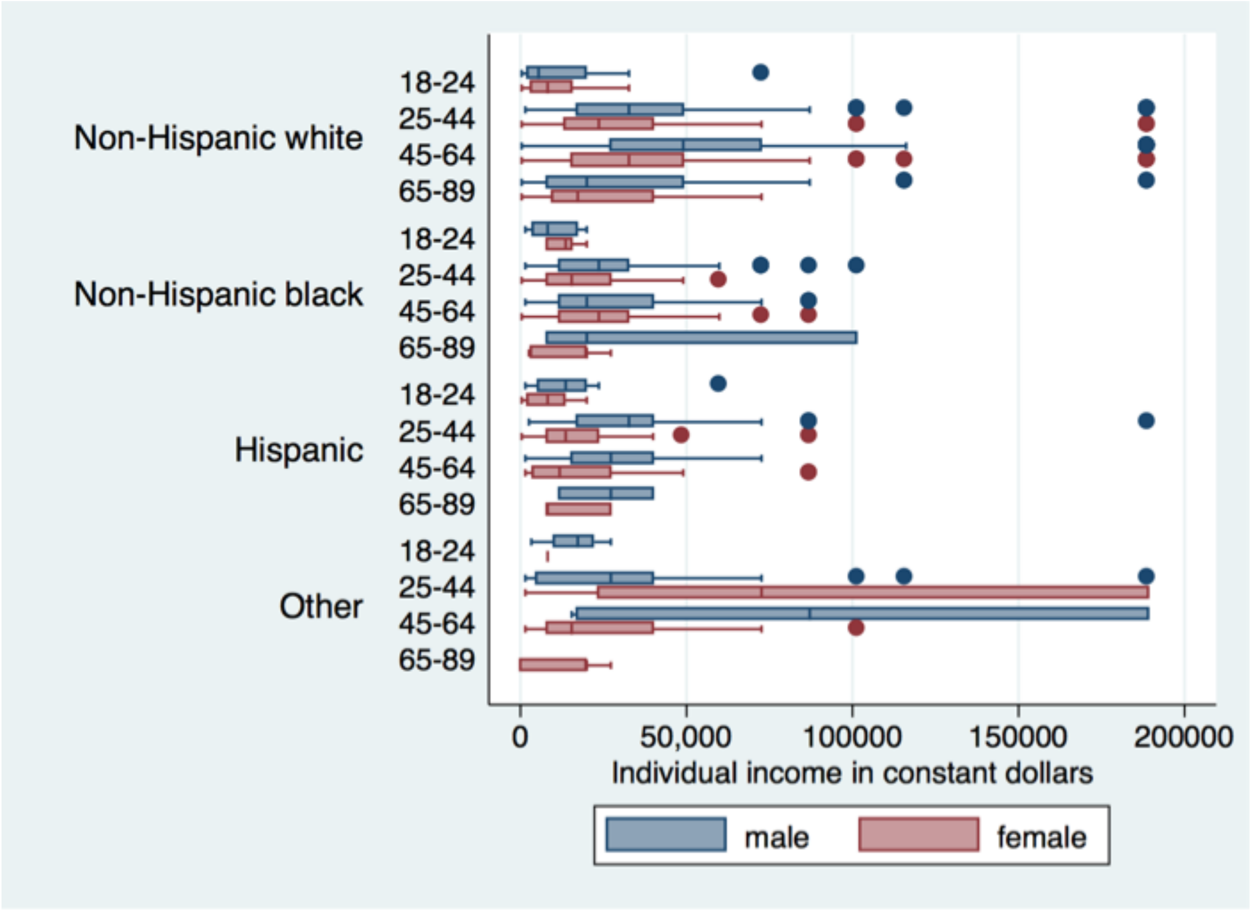
```
graph hbox conrinc [aweight=wtssall],
over(agegr1) over(raceeth)
ytitle(Individual income in constant
dollars)
```

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

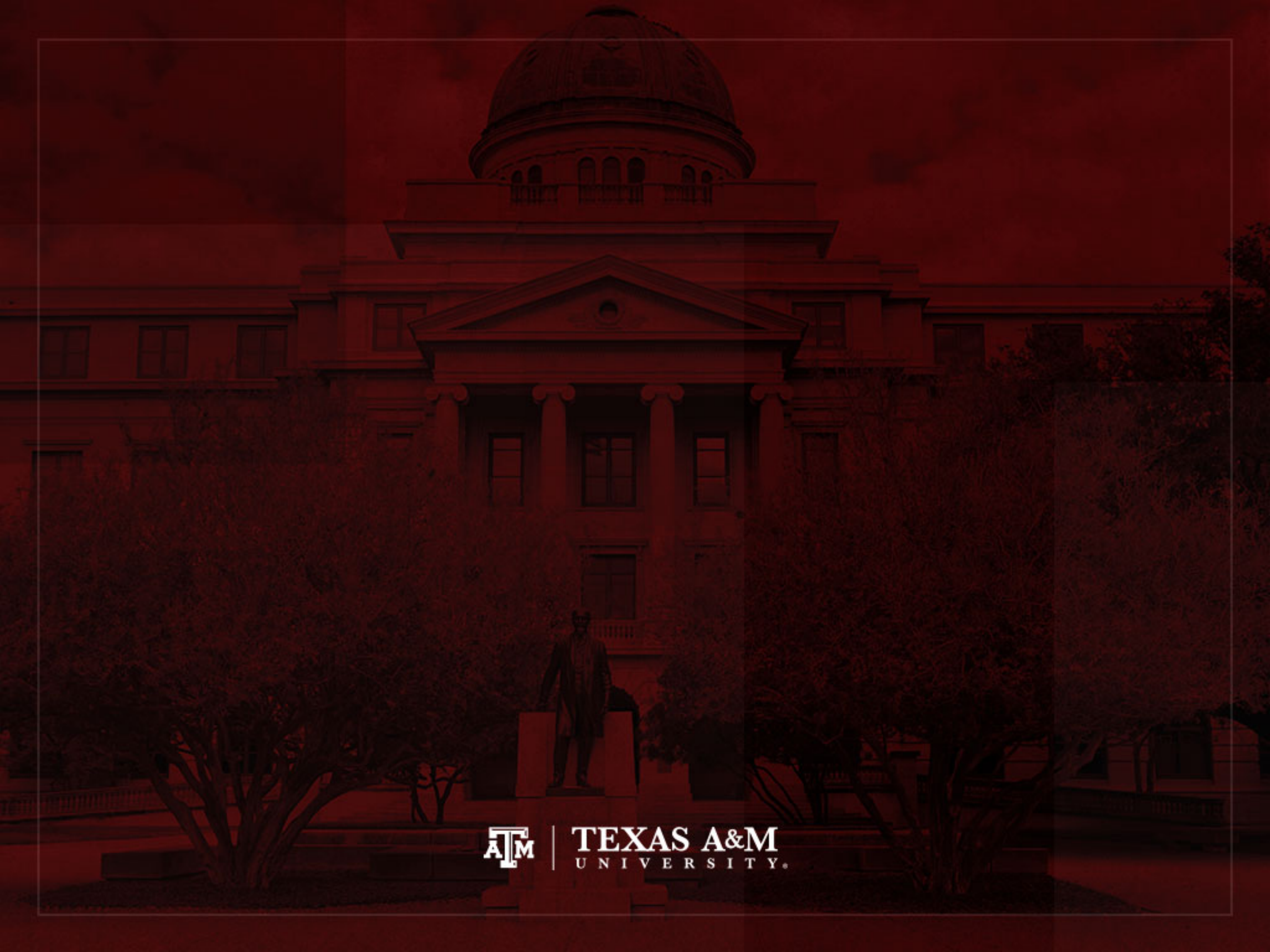



```
graph hbox conrinc [aweight=wtssall], over(sex) over(agegr1) over(raceeth)
ytitle(Individual income in constant dollars)
```

Source: 2016 General Social Survey.

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