[POLS 4150] Measures of Spread, Bivariate Descriptive Statistics

Professor Jason Anastasopoulos
ljanastas@uga.edu

University of Georgia

January 24, 2017

From last time...

- Mean
- Median
- Mode

All ways of describing the center of a distribution...

Mean

$$\text{Sample mean: } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- The sample mean is simply the sum of a number of observations divided by the number of observations.

Median

$$Age = \{22, 18, 19, 60\}$$
$$Age = \{18, \mathbf{19}, \mathbf{22}, 60\}$$

- Median is a simple measure of the center.
- Can be found by organizing the data from lowest to highest value and finding out which value splits the data in two.

Mode

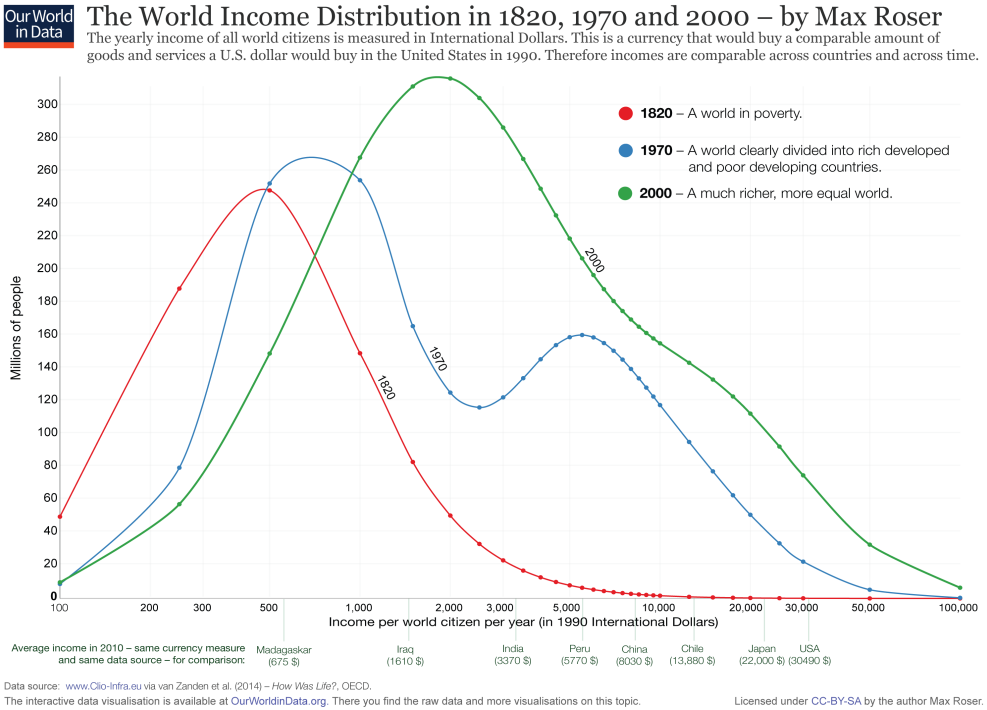
$$\text{Party affiliation} = \{0, 0, 0, 1, 1, 1, 1, 1, 1\}$$

- The mode is simply the value that occurs most frequently.

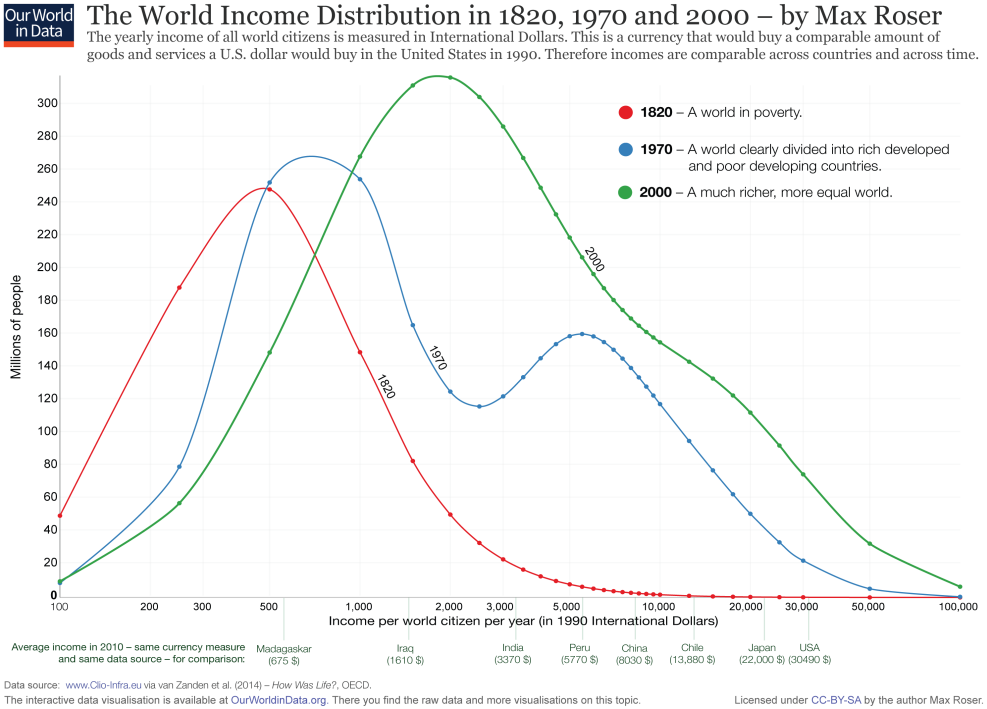
Variability and standard deviation

- Often we're interested in describing the variability or spread of some data.
- This can make a big difference in terms of how we interpret a variable.

Income distribution, then and now

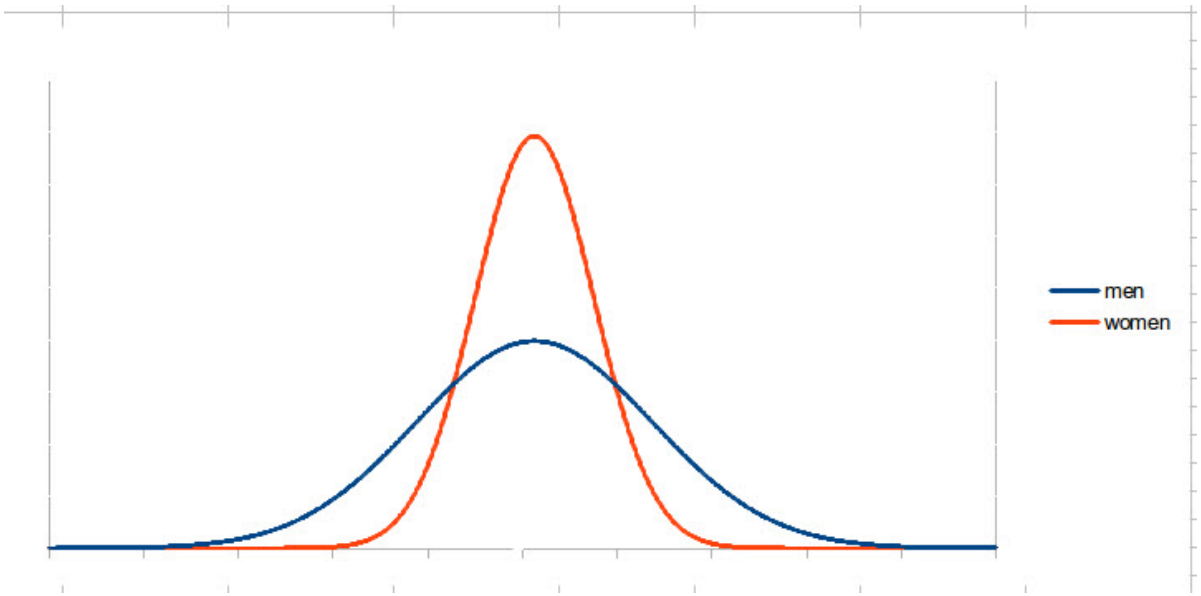


Income distribution, then and now

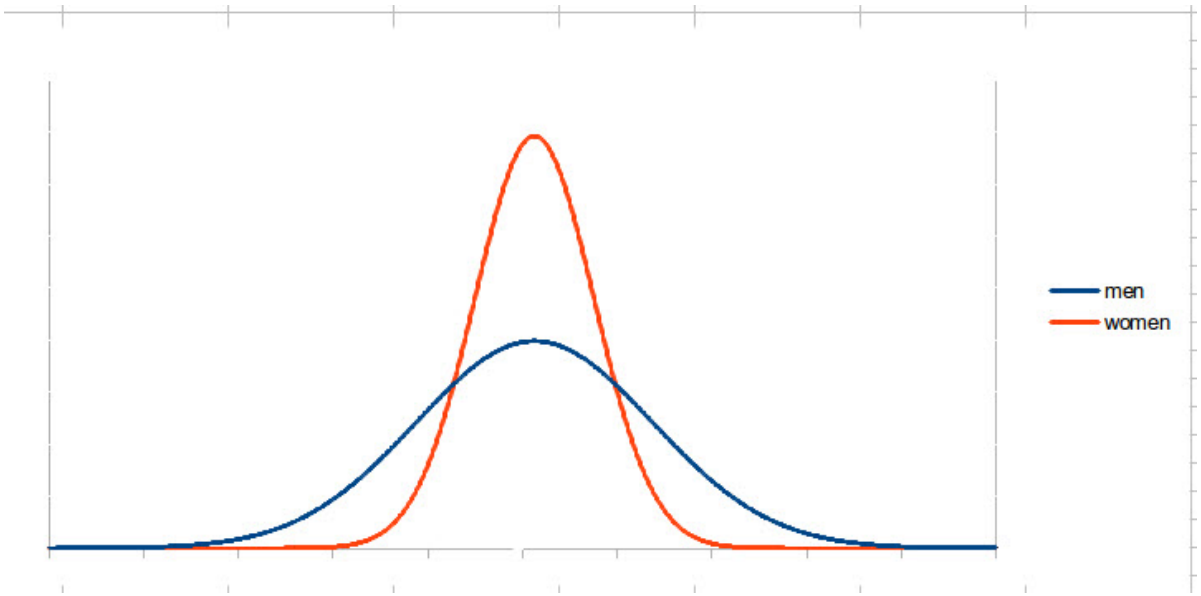


- 1820 – small spread, most people in poverty.
- 1970 – bimodal distribution – very high spread, many rich, many poor.
- 2000 – wealthier more equal income distribution.

Male and female IQ scores



Male and female IQ scores



- Higher spread in male IQ.
- Lower spread in female IQ.

Measuring spread

Two main means of measuring spread:

1. Range.
2. Standard deviation.

Range

$$\text{Range} = max(X) - min(X)$$

- Range is just the difference between the minimum and maximum observation in the data.

Range

$$X = \text{Primary Turnout}_{2016} = \{50\%, \cdots, 74.5\%\}$$
$$\text{Range} = 74.5 - 50 = 24.5\%$$

- Primary turnout in 2016 was close to the record breaking 2008 election.
- 57.6 million Americans turned out to vote.
- Hawaii had the *lowest* turnout and Minnesota, the highest.
- Turnout range is 24.5

Standard deviation

$$\text{Deviation} = x_i - \bar{x}$$

- Another method of measuring spread is how far from the mean the observations are.
- We can think of the distance between one observation and the mean as a *deviation*.

Standard deviation

$$\sum_{i=1}^{N} x_i - \bar{x} \ ?$$

- If we wanted to get total deviation can we just sum the deviations? Why or why not?

Standard deviation

$$X = \{100, -100\}$$
$$\text{Total Deviations} = (100 - 0) + (-100 - 0) = 0$$

- No!
- $Range = 200$
- The large $+$ and $-$ values cancel out and data with a very large spread, has zero deviations.

Standard deviation

$$X = \{100, -100\}$$

$$\text{Deviations} = \sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\text{Deviations} = \sqrt{(100-0)^2 + (-100-0)^2} = 141.42$$

- To correct with this, we *square* the deviations and;
- take the square root to put them back on the same scale as the original variable.

Standard deviation

$$\text{Standard Deviation: } s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

- Rather than get the total number of deviations we a better measure would be the *average* of those deviations.
- Divide by $N - 1$.
- This is the *standard deviation*.

Agreement with statement “we should build a border wall with Mexico”

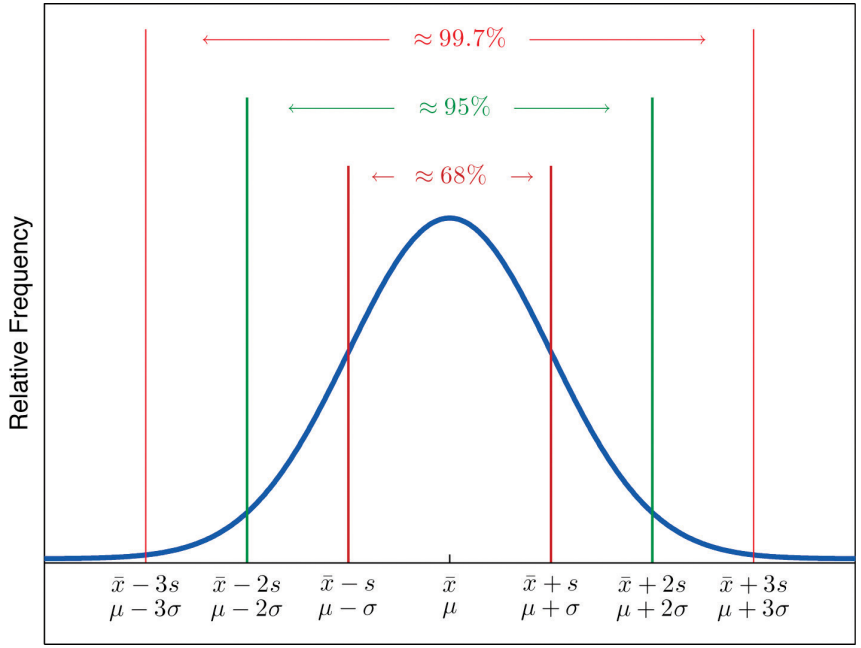
$$1 - 5 \text{ Scale}$$
$$1 = \text{Strongly Disagree}, \cdots, 5 = \text{Strongly Agree}$$

$$\text{New York} = \{1, 1, 1, 1, 2, 3\},\ \bar{x} = 1.5,\ s = 0.08$$
$$\text{Georgia} = \{1, 5, 3, 2, 1, 4\},\ \bar{x} = 2.7,\ s = 1.6$$

Properties of the standard deviation

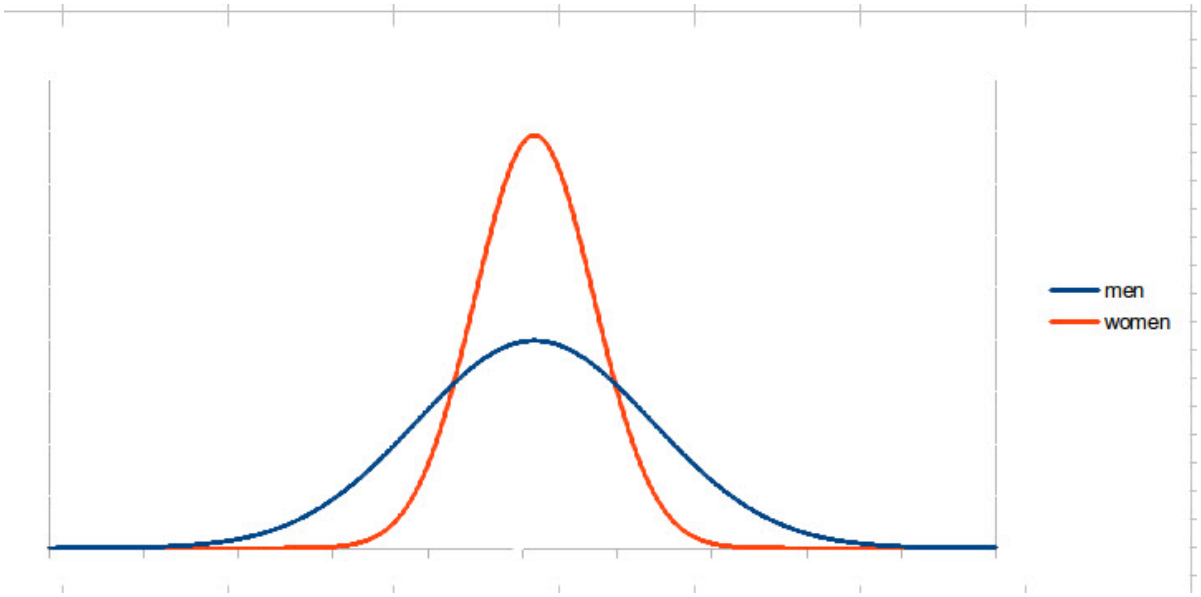
1. $s \geq 0$
2. $s = 0$ iff $x_1 = x_2, \cdots, = x_n$
3. When data are rescaled s is also rescaled.

Empirical rule



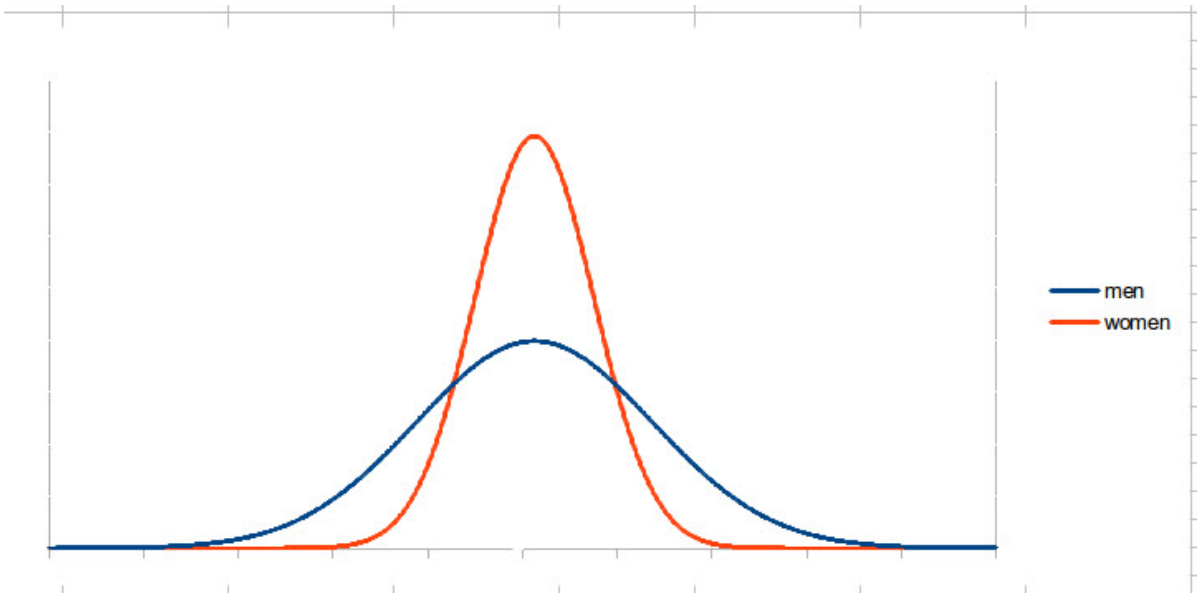
- 68% of observations between, $(\bar{x} - s, \bar{x} + s)$.
- 95% of observations between, $(\bar{x} - 2s, \bar{x} + 2s)$.
- Most/all observations between, $(\bar{x} - 3s, \bar{x} + 3s)$.

Male and female IQ scores



Q: Do men or women have a higher *mean IQ*?

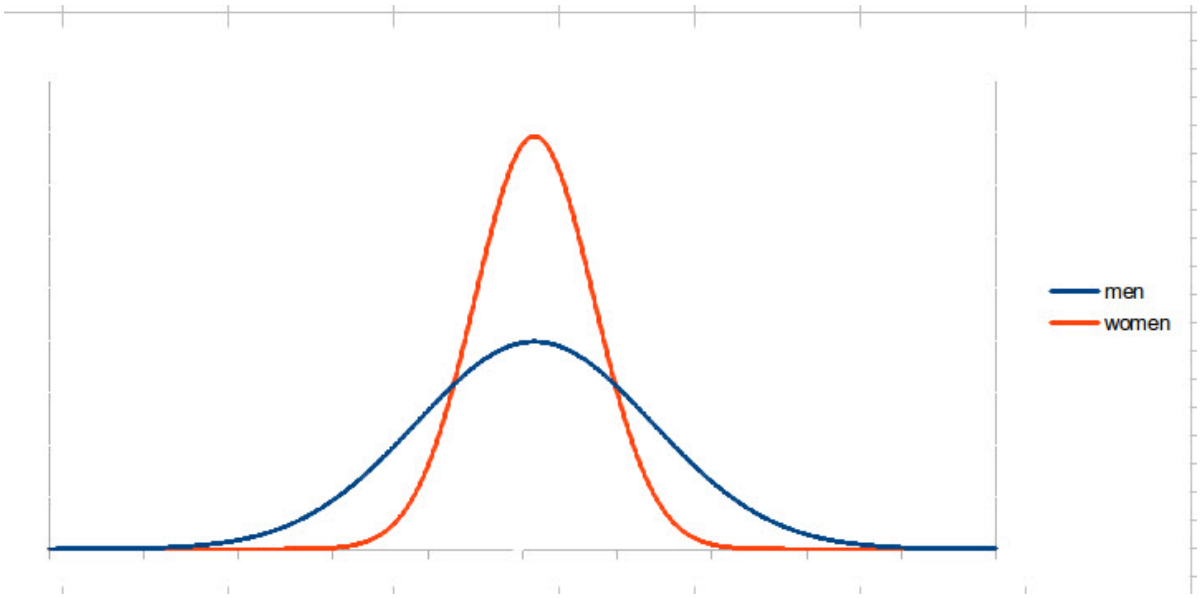
Male and female IQ scores



Q: Do men or women have a higher *mean IQ*?

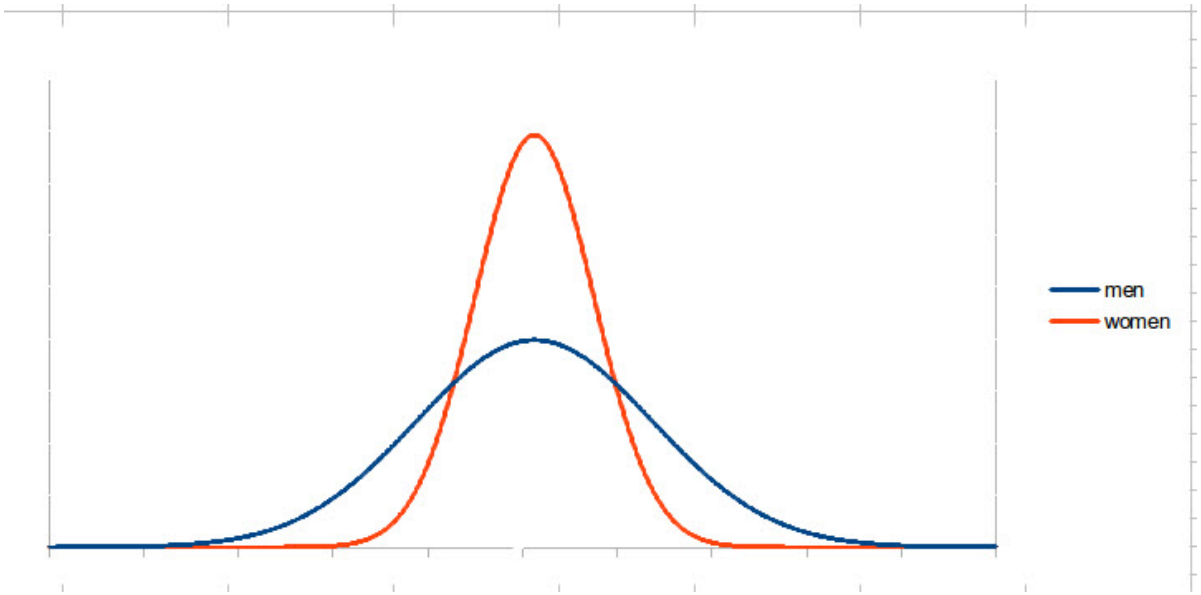
A: They're nearly identical: $\bar{x}_m = \bar{x}_f = 100$

Male and female IQ scores



Q: Which gender has a higher standard deviation?

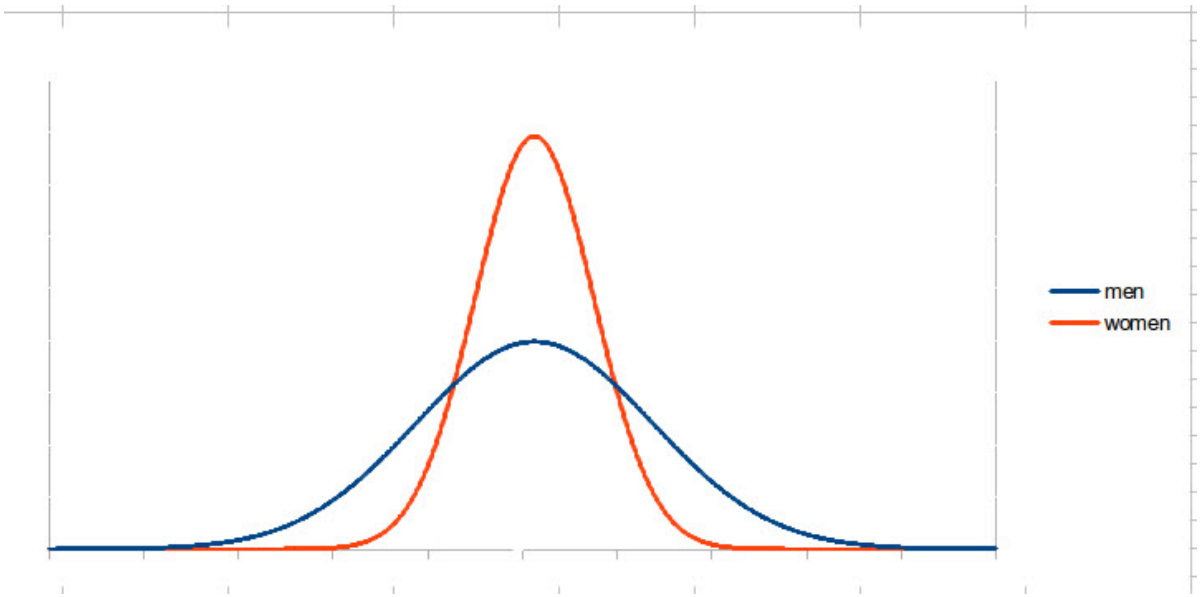
Male and female IQ scores



Q: Which gender has a higher standard deviation?

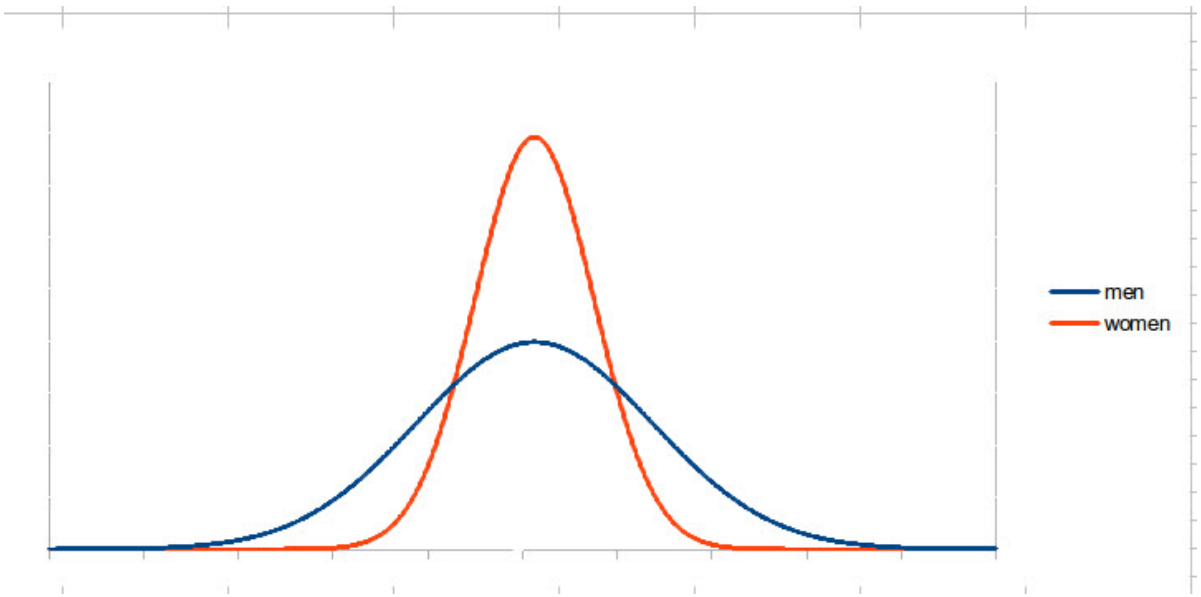
A: **Men**.

Male and female IQ scores



Q: If $\bar{x}_m = 100$ and $s_m = 20$, the IQ scores of 95% of men range from...

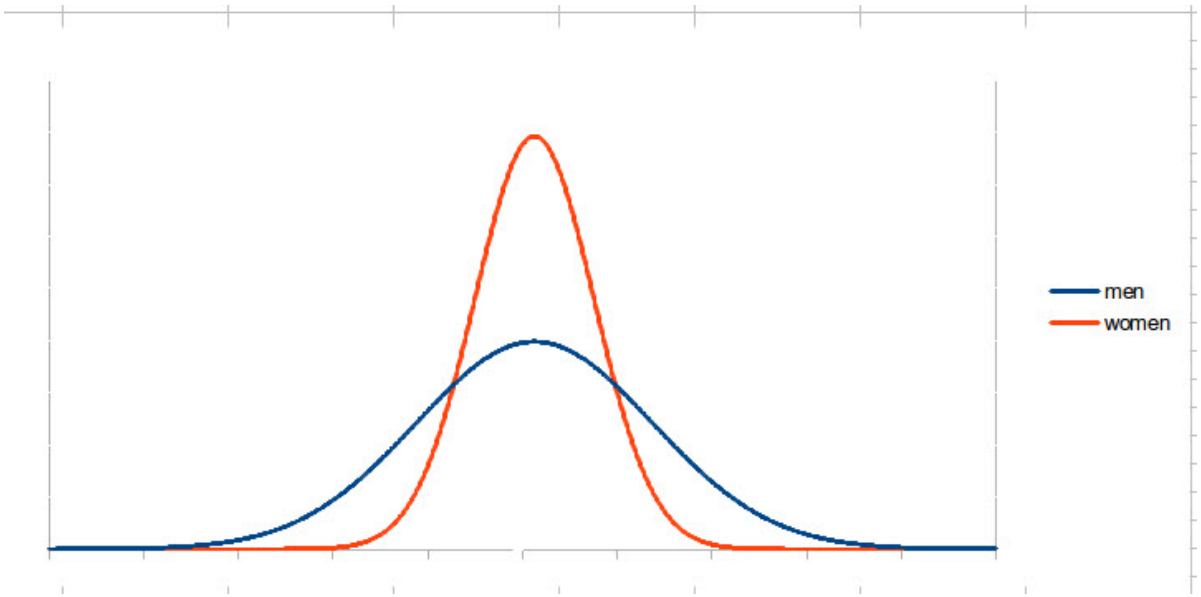
Male and female IQ scores



Q: If $\bar{x}_m = 100$ and $s_m = 20$, the IQ scores of 95% of men range from...

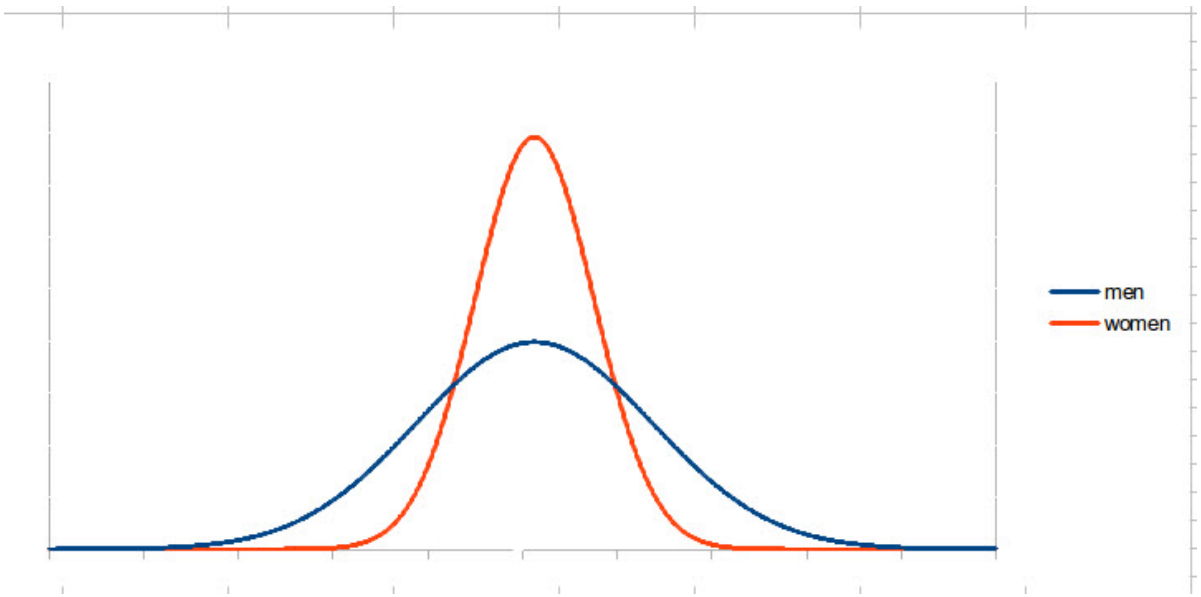
A: 60 to 140

Male and female IQ scores



Q: If $\bar{x}_w = 100$ and $s_w = 15$, the IQ scores of 95% of women range from...

Male and female IQ scores



Q: If $\bar{x}_w = 100$ and $s_w = 15$, the IQ scores of 95% of women range from...

A: 70 to 130

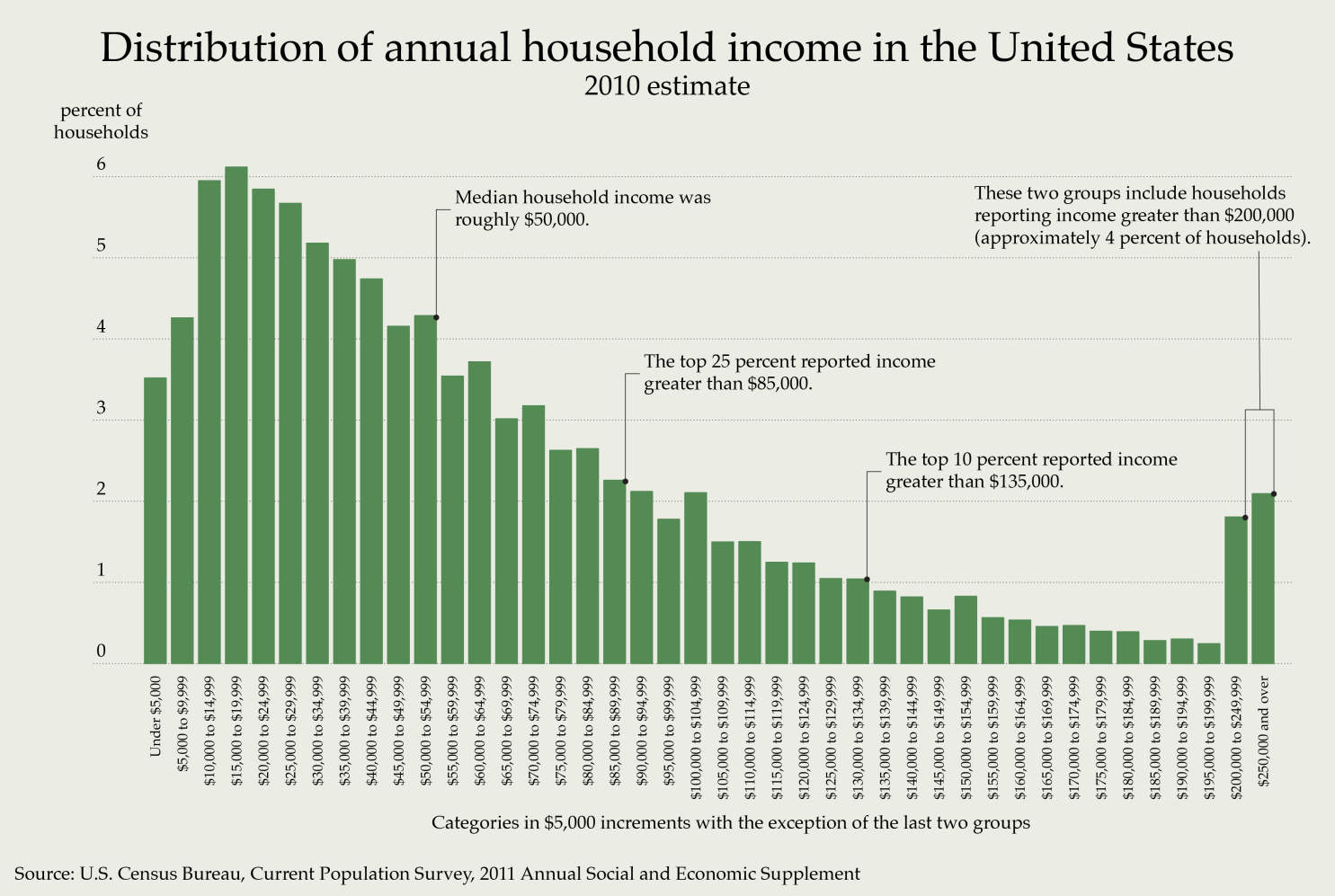
Measures of position

- Another means of describing data are measures of position.
- These tell you what % of data fall above and below a certain point.
- Common measures of position are:
  1. Percentiles
  2. Quartiles

Percentiles

- Point **p** at which **p%** of observations fall below or at that point and $(100 - p)\%$ fall above it

Income distribution



Quartiles

- Chops data up into 3 portions.
- **Lower quartile** – 25th%ile.
- **Inter quartile** – Between 25th and 75th%ile.
- **Upper quartile** – 75th%ile.

Hillary Clinton vote share by county, 2016

```
> summary(vote.share.clinton)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2344  0.3500  0.3620  0.4724  1.0000
```

Outliers and the IQR

$$IQR = X_{75} - X_{25}$$

- Interquartile range is the range of the 25th and 75th%ile of the data.
- For the Hillary Clinton county data this would be: $0.4724 - 0.2344 = 0.238$

Outliers and the IQR

$$\pm 1.5 * IQR = Outlier$$

- Outliers are defined using the IQR.
- Clinton data: $1.5 * 0.238 = 0.357$
- Outlier counties are:
  1. Counties where HRC vote share ≥ 0.8294.
  2. Counties where HRC vote share ≤ 0.