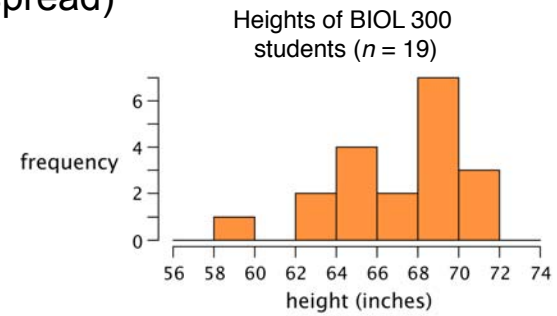
Chapter 3: Describing data

Two common descriptions of data

- Location (or central tendency)
- Width (or spread)



Measures of location

Mean

Median

Mode

Mean

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

n is the size of the sample

Mean

Y_1=56, Y_2=72, Y_3=18, Y_4=42

$\overline{Y}$ = (56+72+18+42) / 4 = 47

Median

- The *median* is the middle measurement in a set of ordered data.

The data:

18 28 24 25 36 14 34

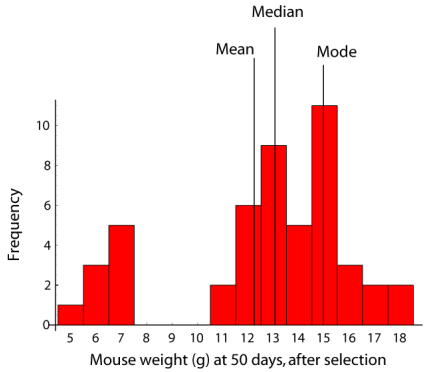
can be put in order:

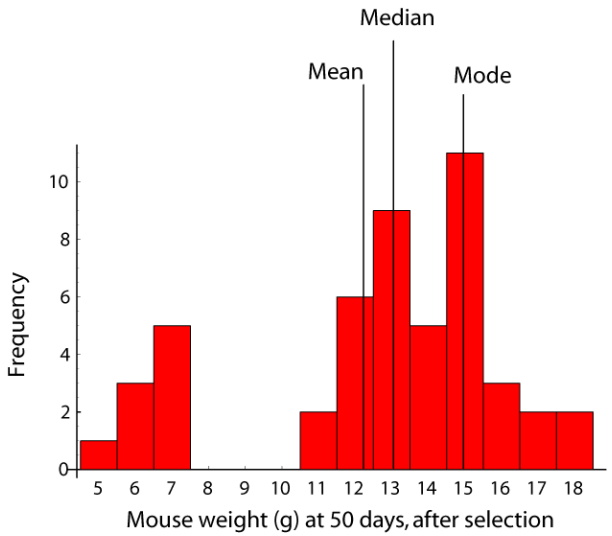
14 18 24 25 28 34 36

Median is 25.

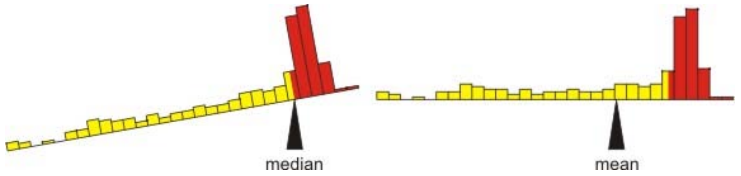
Mode

The mode is the most frequent measurement.





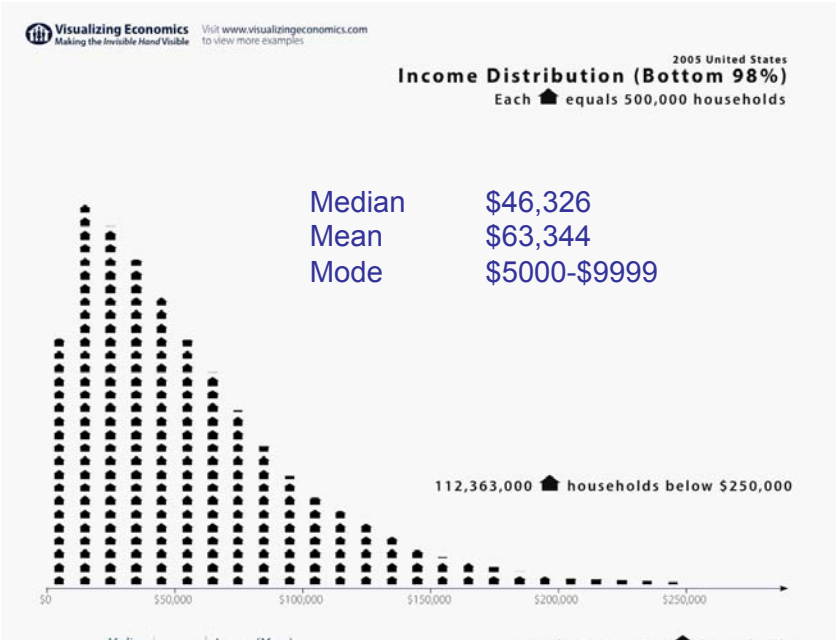
The mean is the center of gravity;
the median is the middle measurement.



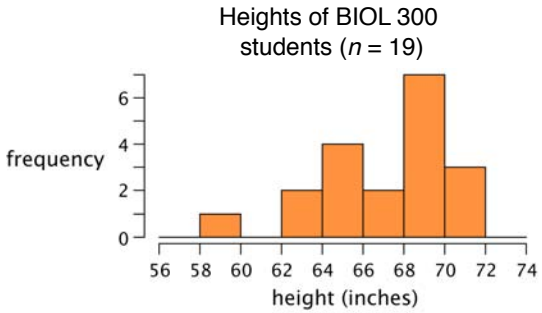
Mean and median for US household income, 2005

Median	$46,326
Mean	$63,344
Mode	$5000-$9999

Why?

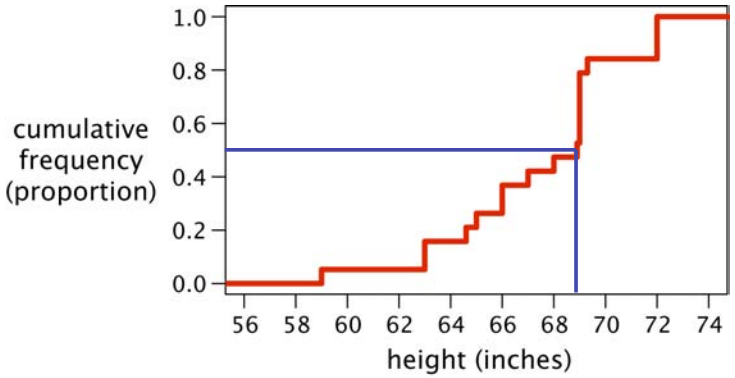


For many distributions, the mean, median, and mode are similar



Mean: 67.4 inches
Median: 68.9 inches
Mode: 68-70 inches

The median (and other quantiles) can be seen on a CFD plot



Measures of width

- Range
- Standard deviation
- Variance
- Coefficient of variation

Range

14 17 18 20 22 22 24
25 26 28 28 28 30 34 36

The range is the maximum minus the minimum:

36 -14 = 22

The range is a poor measure of distribution width

Small samples tend to give lower estimates of the range than large samples

So sample range is a ***biased estimator*** of the true range of the population.

Variance in a population

$$\sigma^2 = \frac{\sum_{i=1}^{N} (Y_i - \mu)^2}{N}$$

N is the number of individuals in the population.
μ is the true mean of the population.

Sample variance

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}$$

n is the sample size

Example: Sample variance

Family sizes of 5 BIOL 300 students: 2 3 3 4 4 (in units of siblings)

Y_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})^2$
2	-1.2	1.44
3	-0.2	0.04
3	-0.2	0.04
4	0.8	0.64
4	0.8	0.64
Sums: 16		2.80

"Sum of squares"

$$\overline{Y} = \frac{(2+3+3+4+4)}{5} = \frac{16}{5} = 3.2$$

$$s^2 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n-1}$$

$$s^2 = \frac{2.80}{4} = 0.70$$ (in units of siblings **squared**)

Shortcut for calculating sample variance

$$s^2 = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\left(Y_i^2\right)}{n} - \bar{Y}^2\right)$$

Example: Sample variance (shortcut)

Family sizes of 5 BIOL 300 students: 2 3 3 4 4

Y_i	Y_i^2	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
2	4	-1.2	1.44
3	9	-0.2	0.04
3	9	-0.2	0.04
4	16	0.8	0.64
4	16	0.8	0.64
Sums: 16	54		2.80

$$\bar{Y} = \frac{(2+3+3+4+4)}{5} = 3.2$$

$$s^2 = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\left(Y_i^2\right)}{n} - \bar{Y}^2\right)$$

$$s^2 = \frac{5}{4}\left(\frac{54}{5} - (3.2)^2\right) = 0.70$$

Standard deviation (SD)

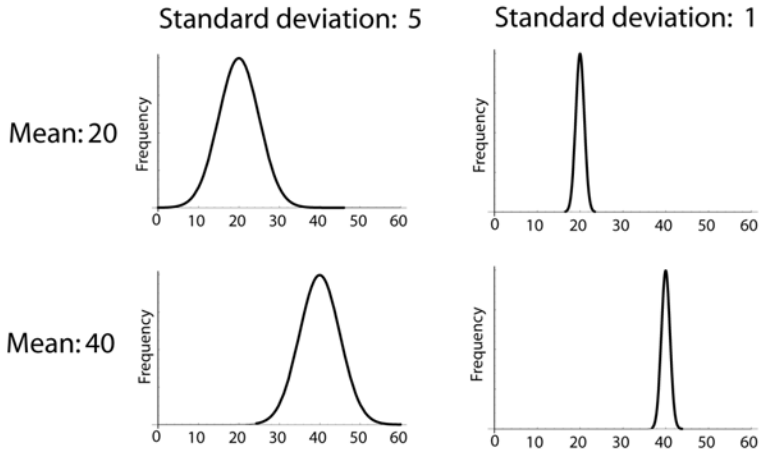
- Positive square root of the variance

σ is the true standard deviation
s is the sample standard deviation:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}}$$

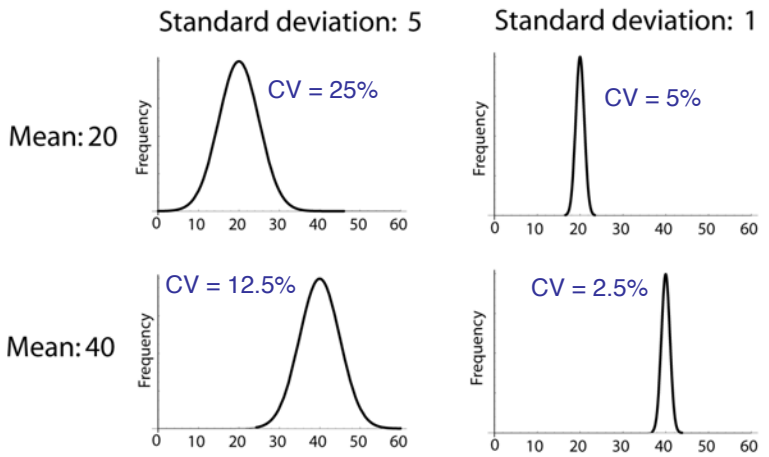
$$s^2 = 0.70$$

$$s = \sqrt{0.70} = 0.84$$



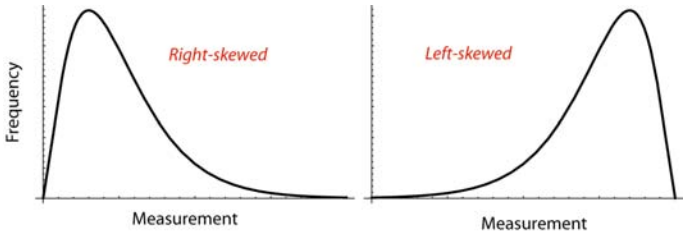
Coefficient of variation (CV)

$$CV = 100\% \frac{s}{\overline{Y}}$$



Skew

- Skew is a measurement of asymmetry
- Skew (as in "skewer") refers to the pointy tail of a distribution



Nomenclature

	Population Parameters	Sample Statistics
Mean	μ	$\overline{Y}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Manipulating means

- E[X] = mean of X (or "expectation of X")
- The mean of the sum of a variable and a constant: E[X + c] = E[X]+ c
- The mean of a product of a variable and a constant: E[c X] = c E[X]
- The mean of the sum of two variables: E[X + Y] = E[X]+ E[Y]
- The mean of a product of two variables: E[X Y] = E[X] E[Y] if and only if X and Y are independent.

Manipulating variance

- The variance of the sum of a variable and a constant: Var[X + c] = Var[X]
- The variance of a product of a variable and a constant: Var[c X] = c^2 Var[X]
- The variance of the sum of two variables: Var [X + Y] = Var[X]+ Var[Y] if and only if X and Y are independent.