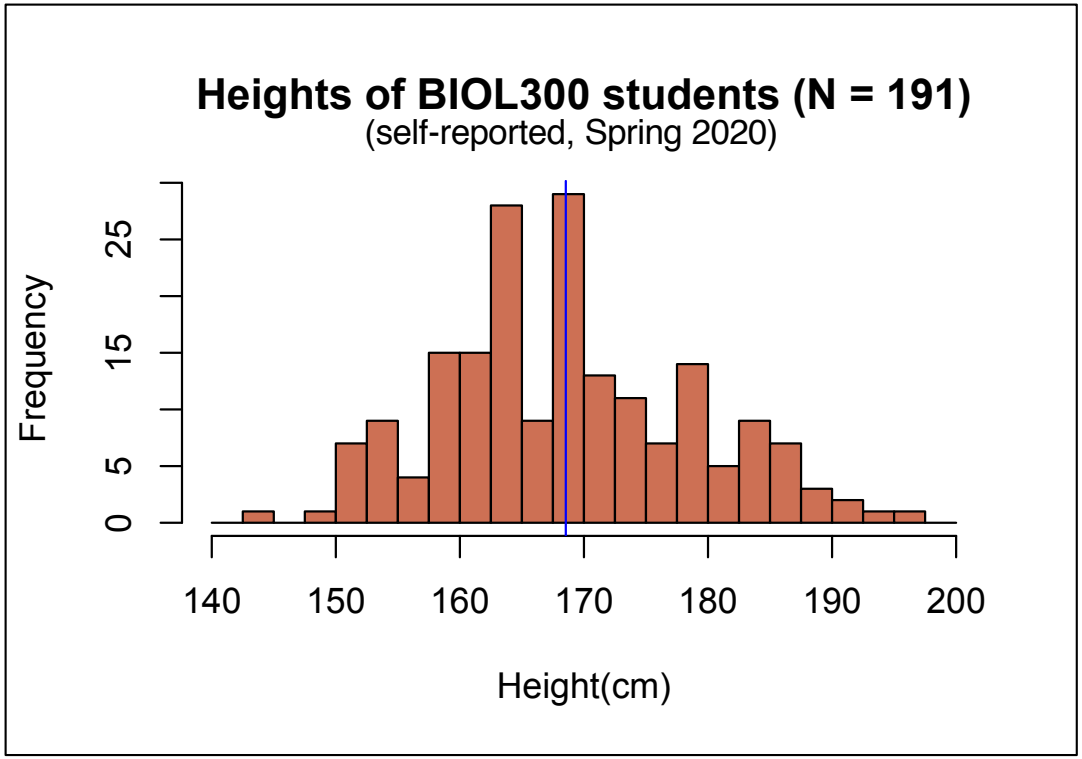
Chapter 3:
Describing data

Two common descriptions of data

1) Location (or central tendency)
2) Width (or spread)



Measures of location

Mean

Median

Mode

Mean of a sample

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

Y is a variable

n is the size of the sample

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

Mean

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

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

Mean of a sample

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

n is the size of the sample

Mean of a population

$$\mu = \frac{\sum_{i=1}^{N} Y_i}{N}$$

N is the size of the population

Median

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

The data:

18 32 24 31 35 7 34

can be put in order:

7 18 24 31 32 34 35

Median is 31.

Median,
when even number of observations

The data:

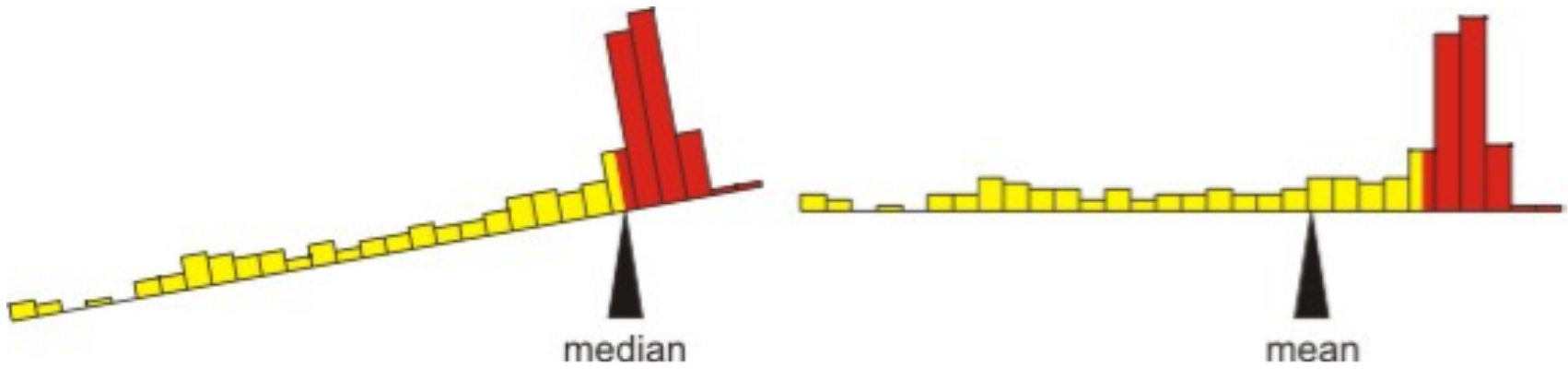
4 8 10 12 39 412

We take halfway between the middle two:

(10 + 12) / 2 = 11

Median is 11.

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?

Visualizing Economics
Making the *Invisible Hand* Visible

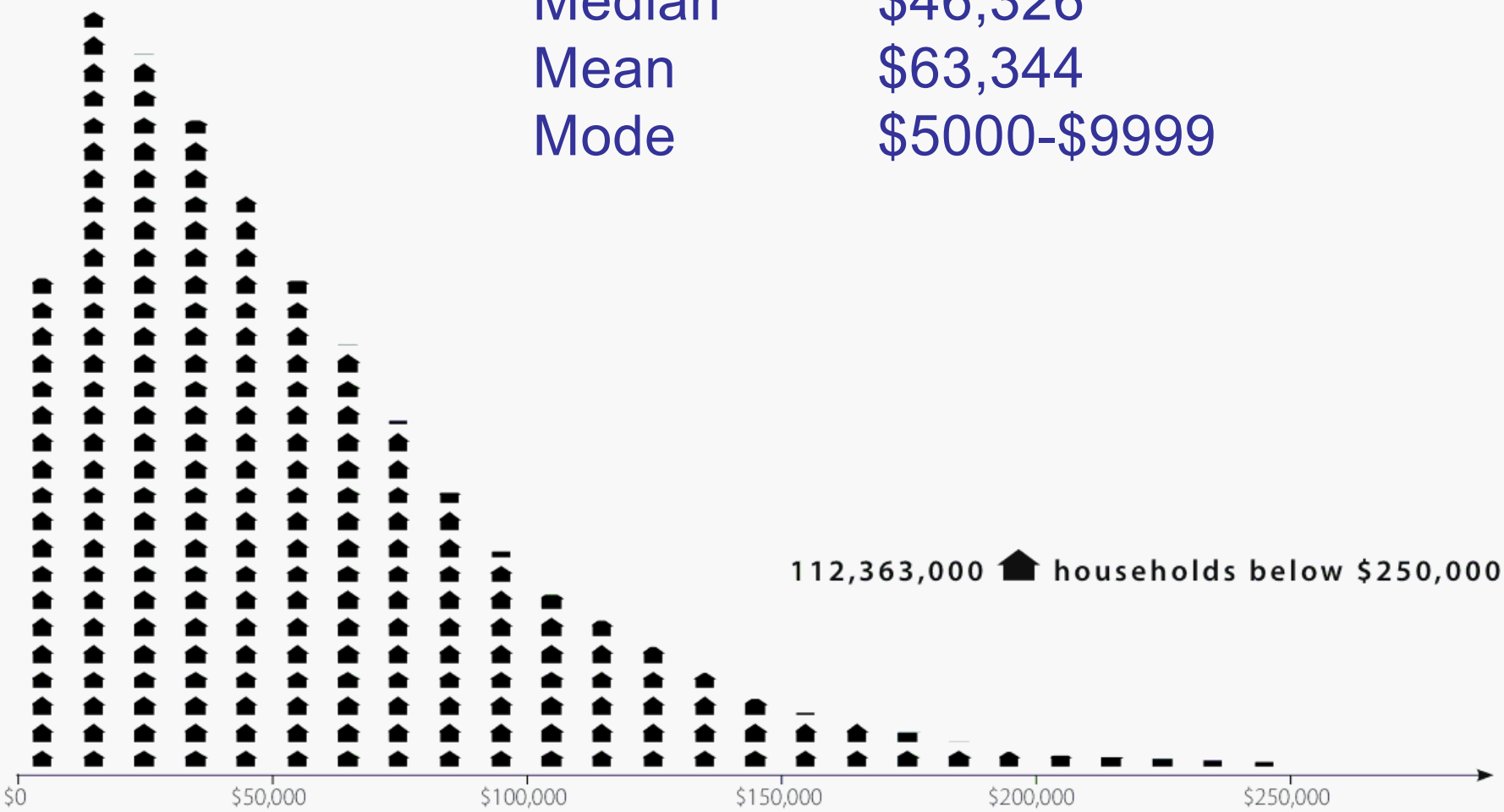
Visit www.visualizingeconomics.com
to view more examples

2005 United States

Income Distribution (Bottom 98%)

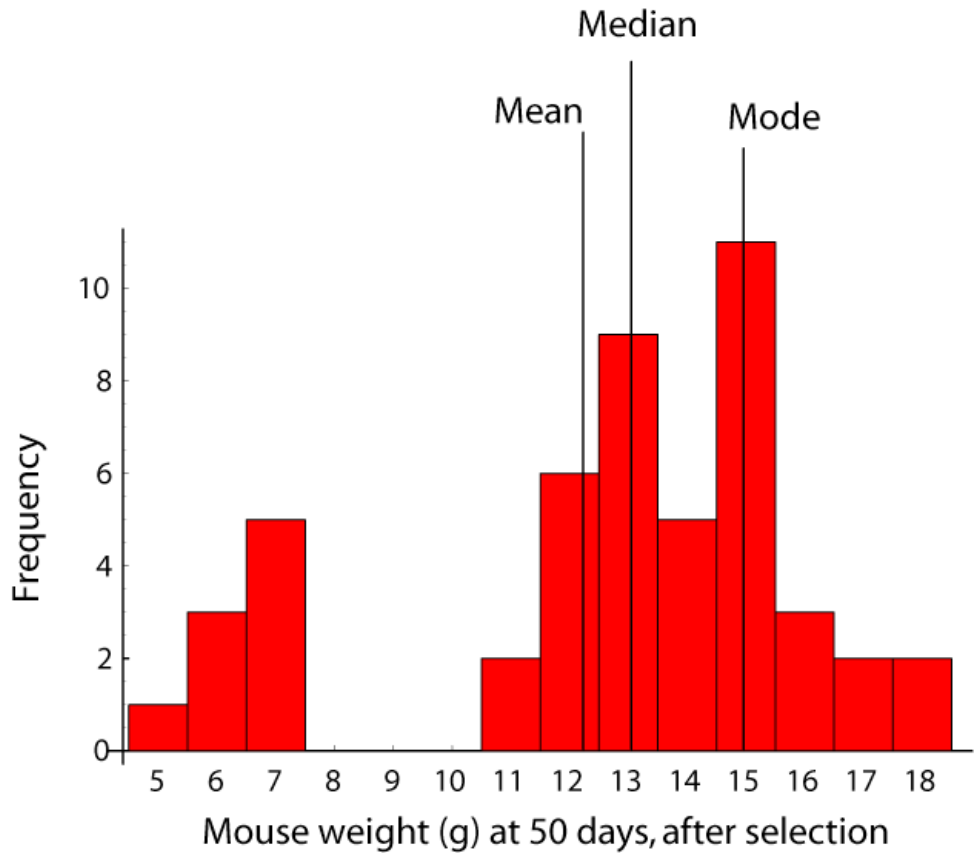
Each ⌂ equals 500,000 households

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



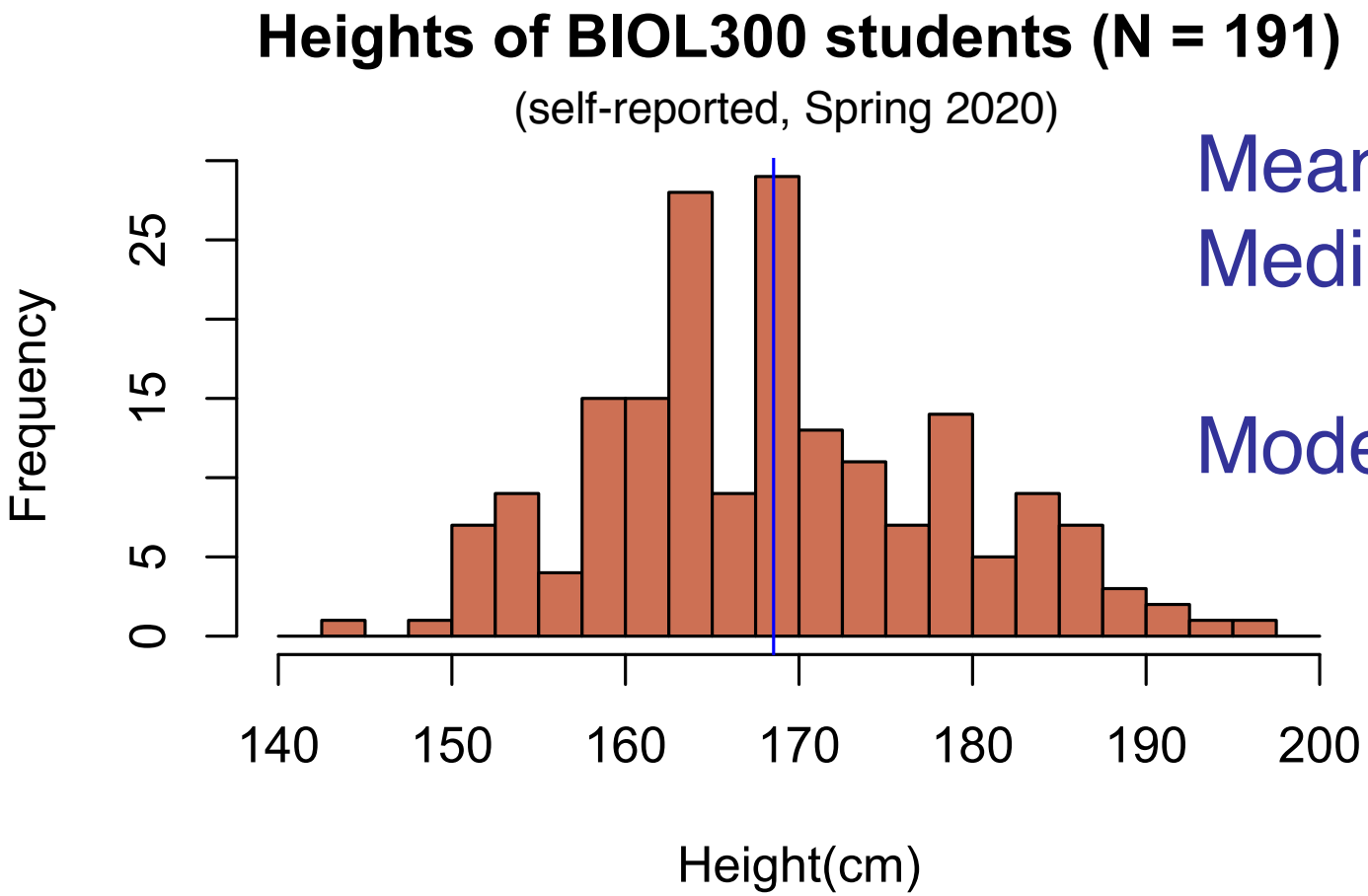
Mode

The *mode* is the most frequent measurement.



Related terms:
unimodal, *bimodal*, *trimodal*

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



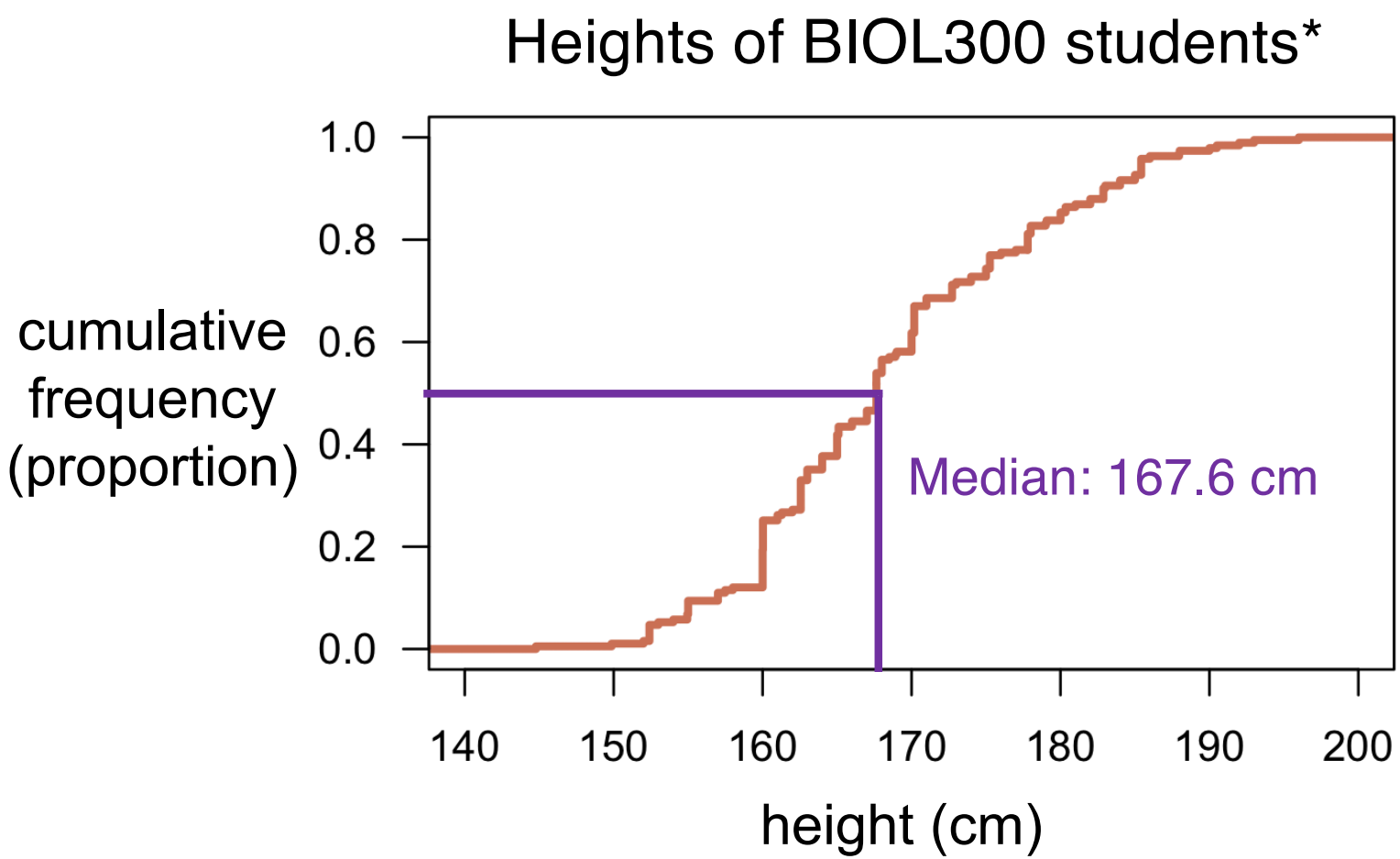
Mean: 168.53 cm

Median: 167.64 cm

(66 inches)

Mode: 167.5-170 cm

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



* 2020 Spring, N=191, self-reported

Measures of width

- Range
- Variance
- Standard deviation
- 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:

$$36 - 14 = 22$$

Often expressed as two numbers: 14 to 36

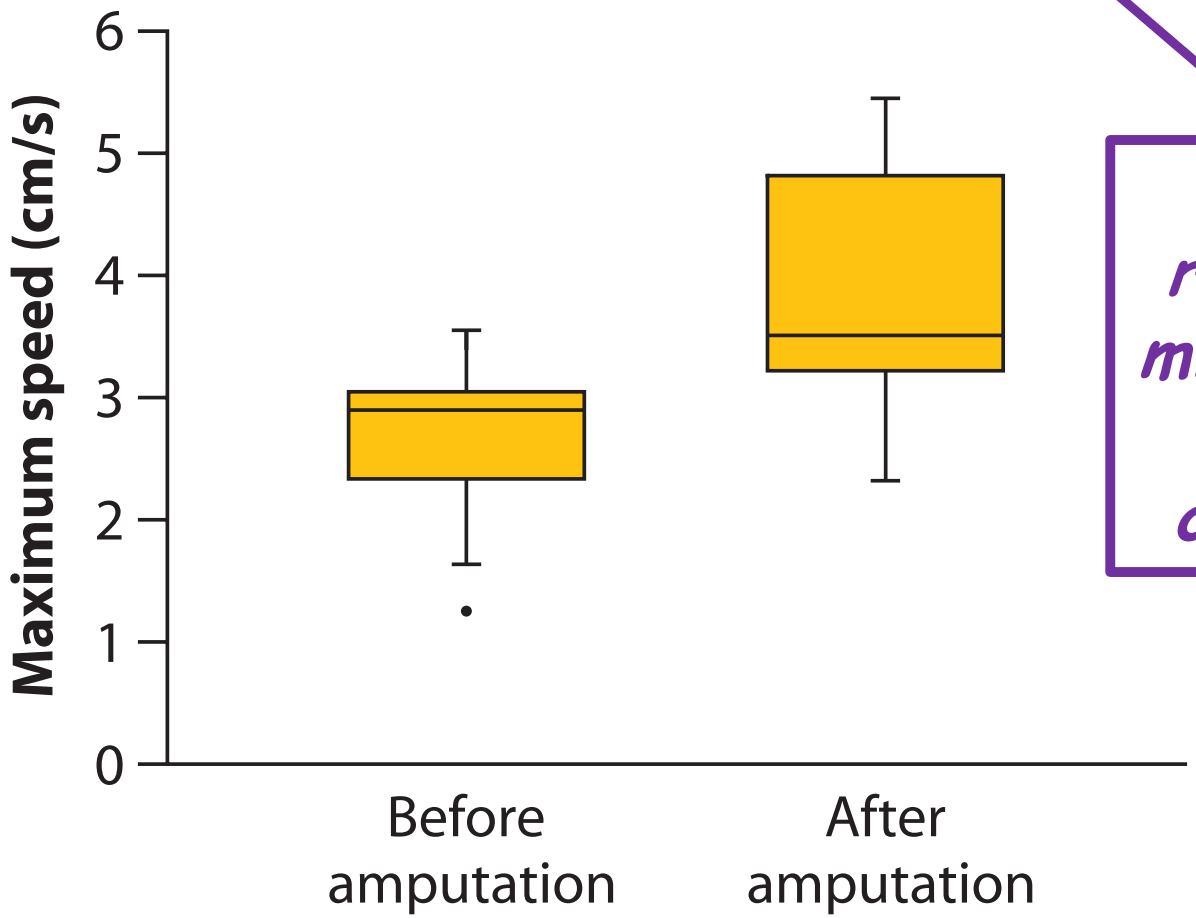
The range is a poor measure of distribution width

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 (sample range tends to be lower).

Box Plot

(uses median and interquartile range)



Shows the range of the middle 50% of ordered observations

Figure 3.2-2 Box plot of the running speed of 16 male spiders before and after self-amputation of a pedipalp. (Ramos et al. 2004)

Variance

- *hugely important in statistics*
- *average squared deviation from the mean*
- *has useful mathematical properties*

Variance in a population

"average squared deviation from the mean"

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

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

Sample variance

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

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	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
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"

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

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

$$s^2 = \frac{2.80}{4} = 0.70$$

(in units of siblings **squared**)

Shortcut for calculating sample variance

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

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)

- A little more intuitive than variance
- Positive square root of the variance

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

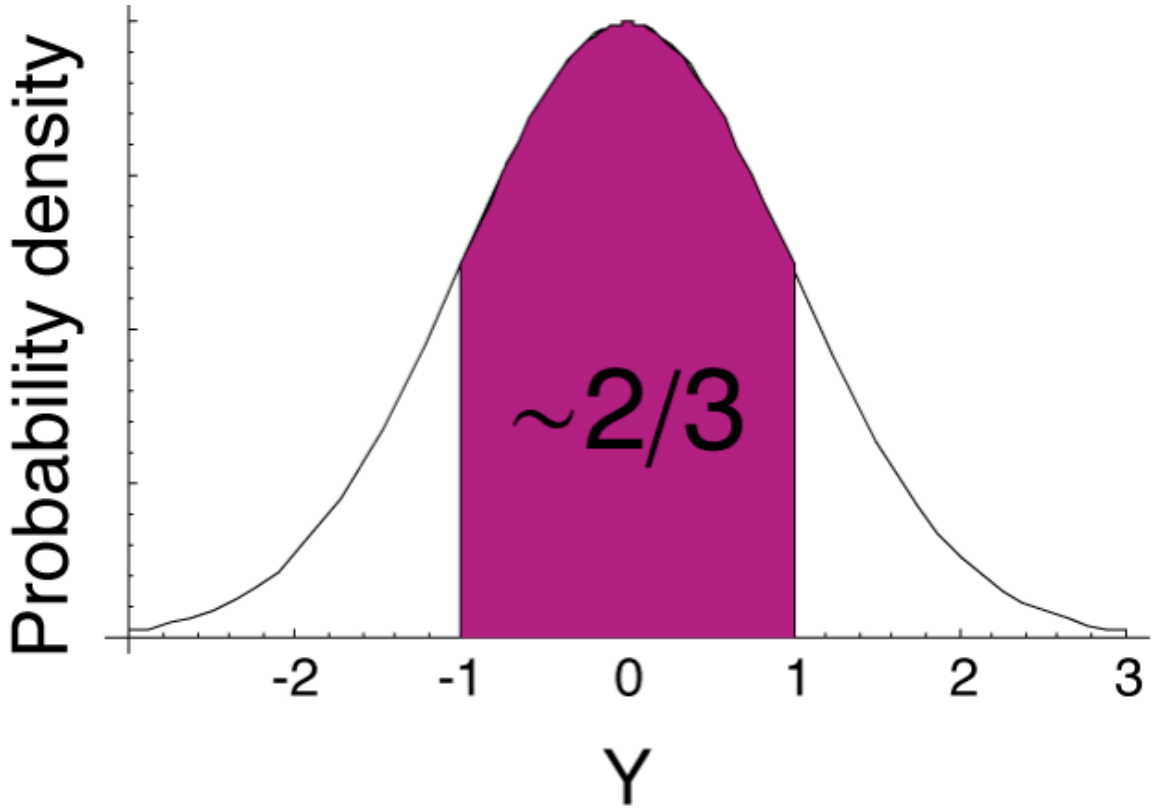
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}{n-1}}$$

$$s^2 = 0.70$$

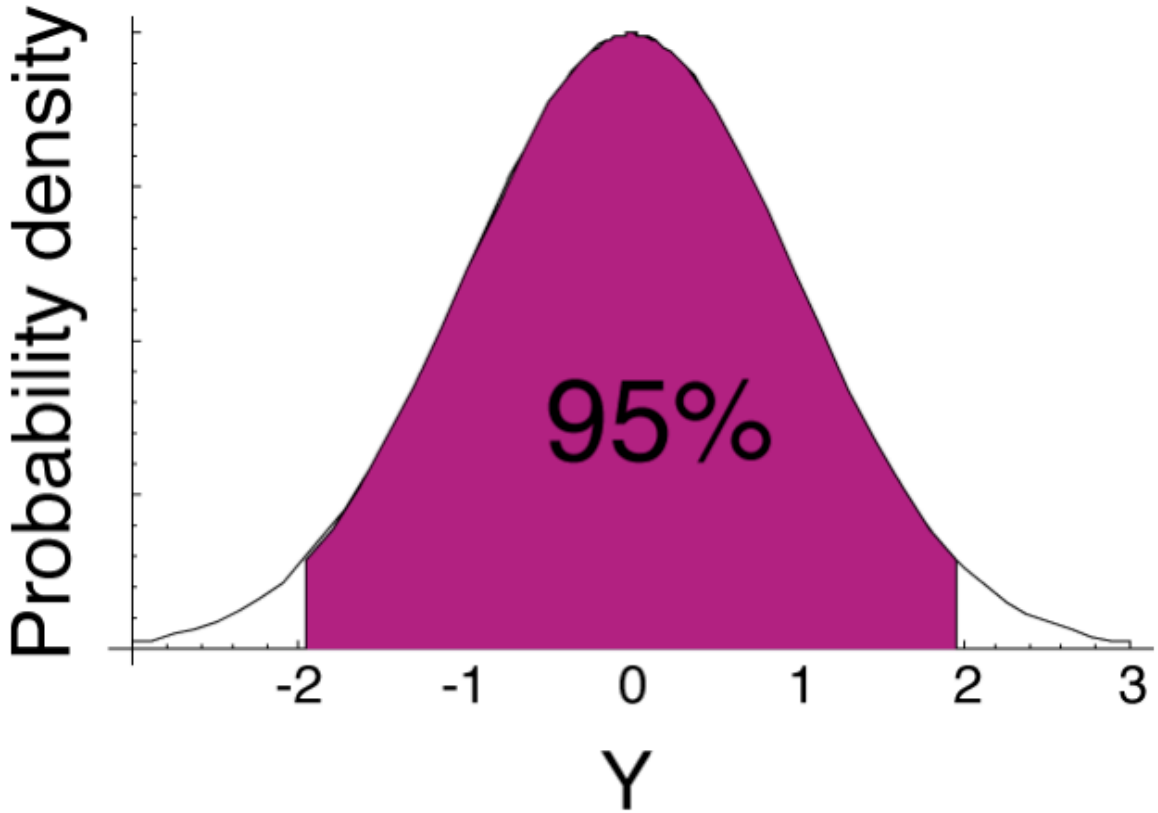
$$s = \sqrt{0.70} = 0.84$$

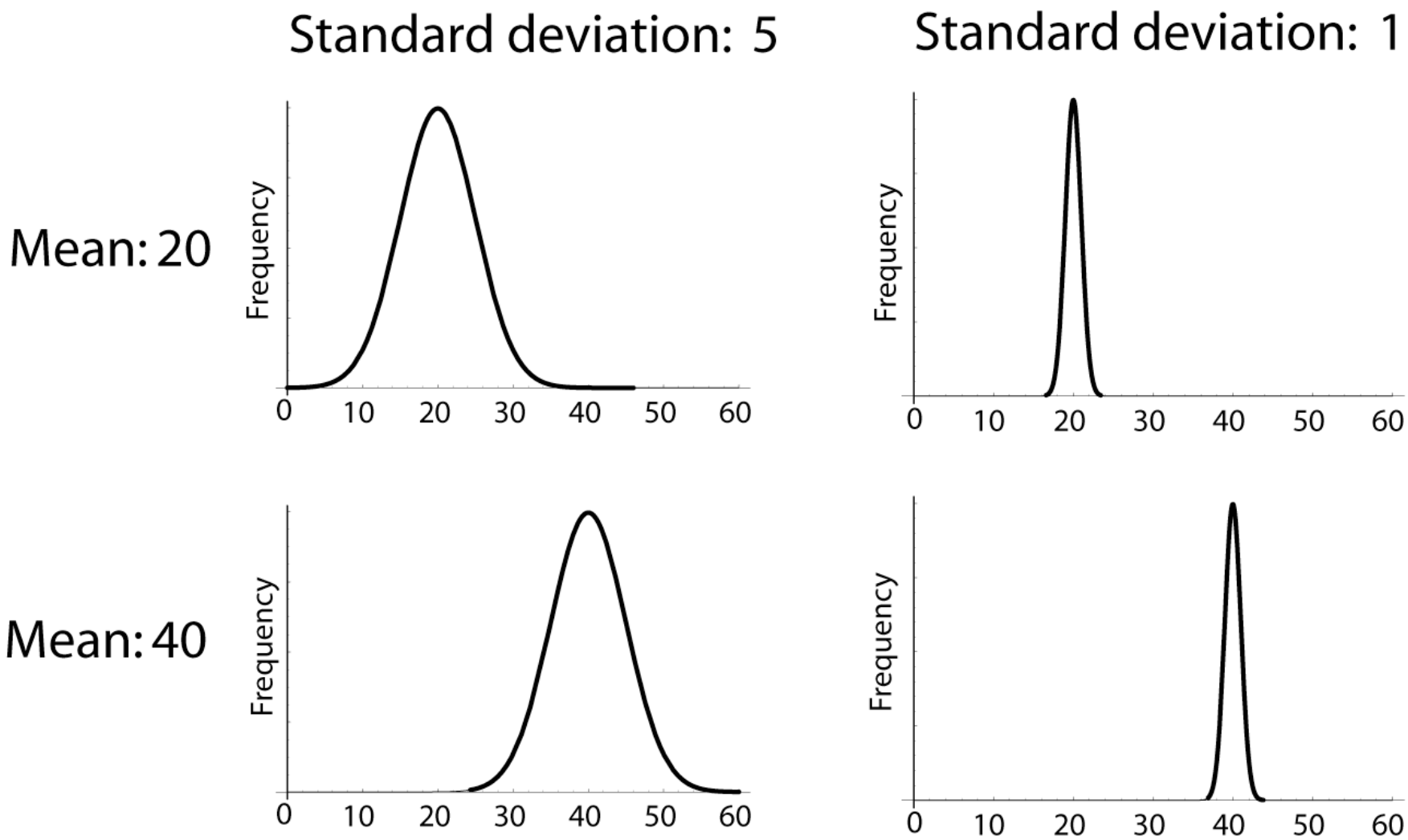
(in units of siblings)

For a bell-shaped population, about 2/3 of individuals are within one standard deviation of the mean



For a bell-shaped population, about 95% of individuals are within two standard deviations of the mean



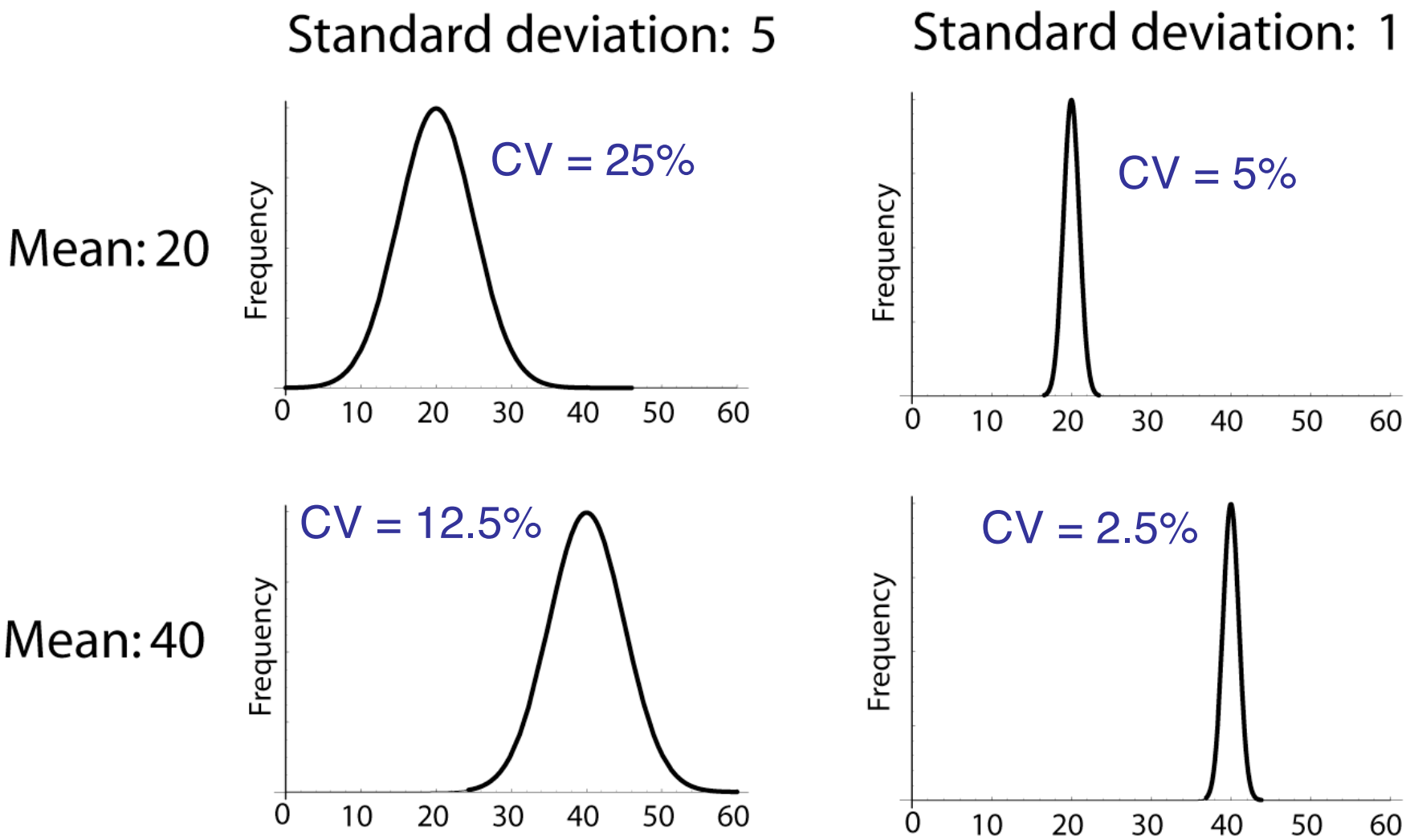

Standard deviation: 5
Standard deviation: 1
Mean: 20
Mean: 40
Frequency
0
10
20
30
40
50
60

Coefficient of variation (CV)

Expresses how big the standard deviation is, in relation to the mean:

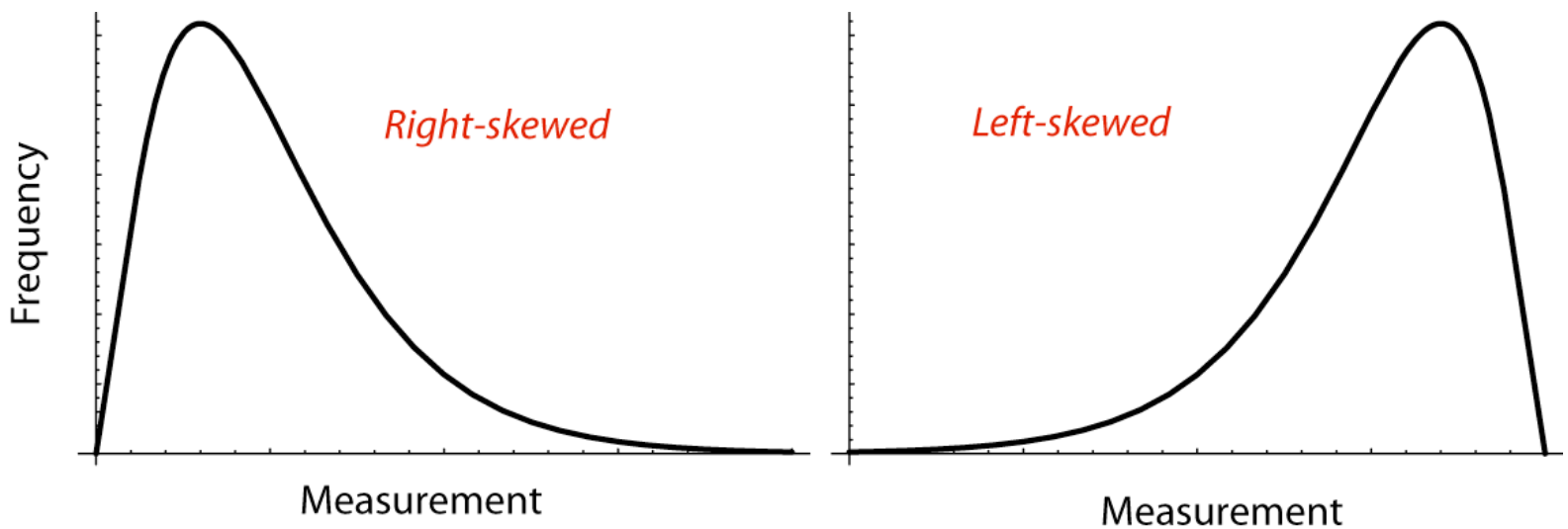
$$CV = 100\% \frac{s}{\bar{Y}}$$

Example: 1 cm SD when studying mice or giraffes


Standard deviation: 5
Standard deviation: 1
Mean: 20
Mean: 40
CV = 25%
CV = 5%
CV = 12.5%
CV = 2.5%
Frequency
0
10
20
30
40
50
60

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

- If X represents a set of measurements, and mean(X) is the mean of all X values:
- Adding a constant (c) to each X:

$$\text{mean}(X + c) = \text{mean}(X) + c$$

- Multiplying each X by a constant:

$$\text{mean}(c\,X) = c\,\text{mean}(X)$$

- Summing two variables:

$$\text{mean}(X + Y) = \text{mean}(X) + \text{mean}(Y)$$

- *For those interested:* Multiplying two variables: $\text{mean}(X\,Y) = \text{mean}(X)\,\text{mean}(Y)$ if and only if X and Y are independent.

Manipulating variance

- Adding a constant (c) to each X:

$$Var(X + c) = Var(X)$$

- Multiplying each X by a constant:

$$Var(c\,X) = c^2\,Var(X)$$

- *For those interested:* Summing two variables: $Var(X + Y) = Var(X) + Var(Y)$ if and only if X and Y are independent.