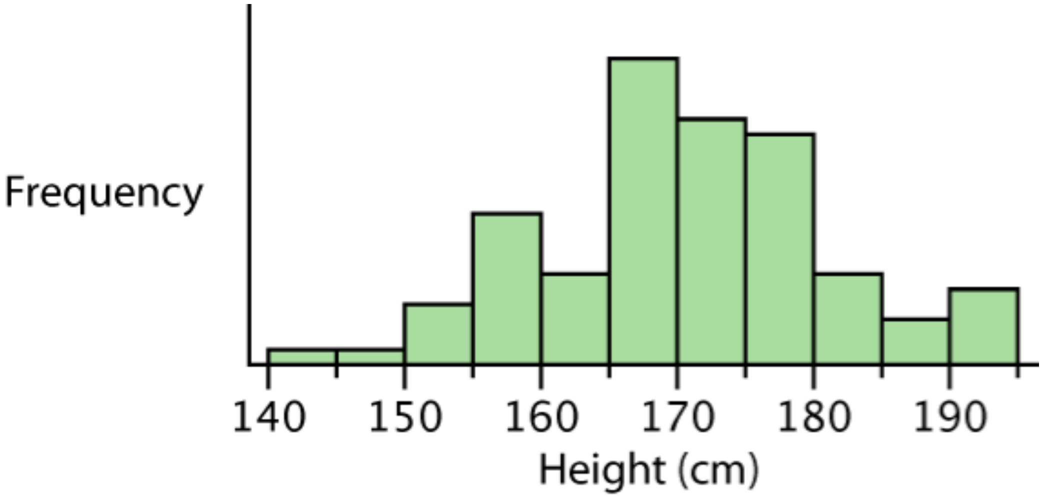
Describing data

Two common descriptions of numerical 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 (n = odd number):

18 28 24 25 36 14 34

can be put in order:

14 18 24 $\underline{25}$ 28 34 36

Median is 25

The data (n = even number):

18 28 24 25 36 14 34 17

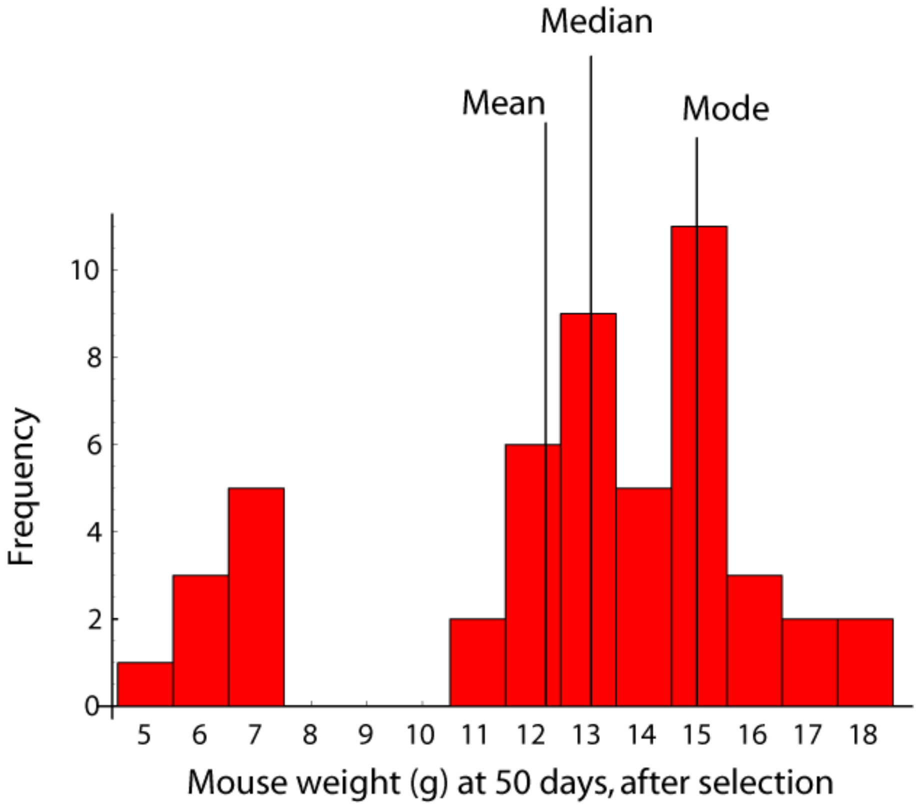
can be put in order:

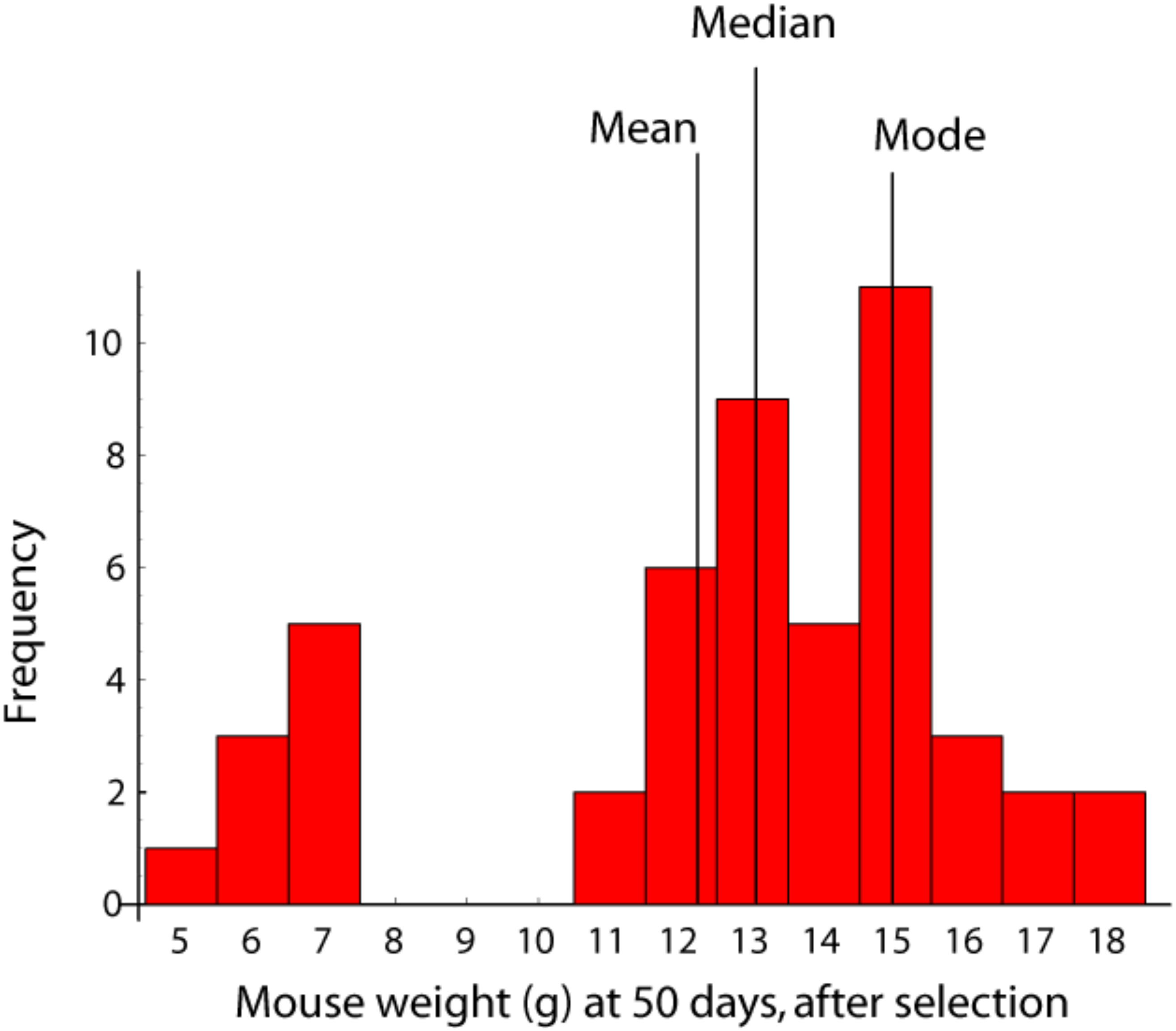
14 17 18 $\underline{24 \quad 25}$ 28 34 36

Median is 24.5

Mode

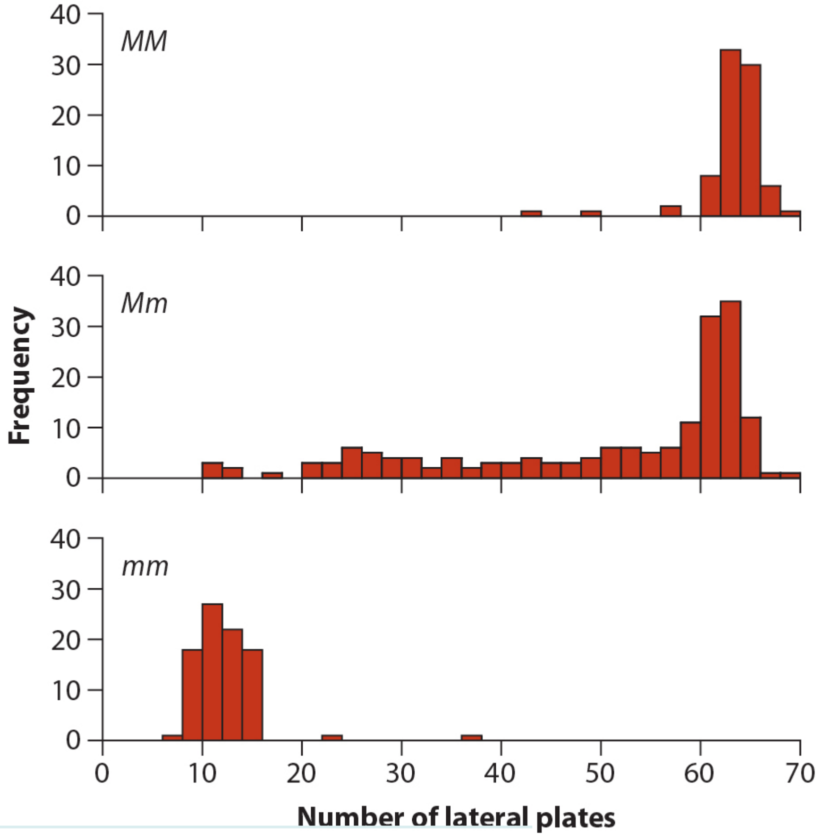
The mode is the most frequent measurement.





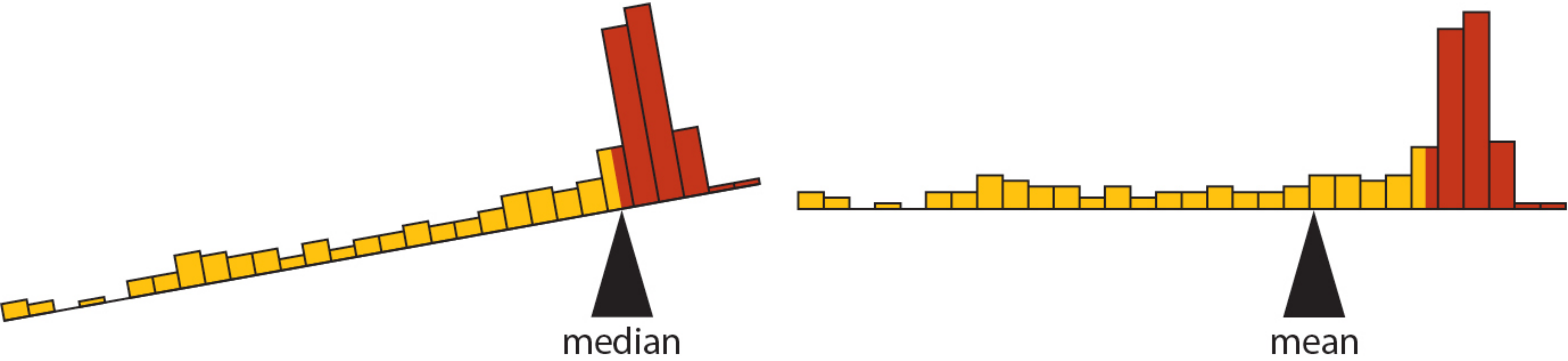
genotype *pygmy mutation*

How do mean and median compare?

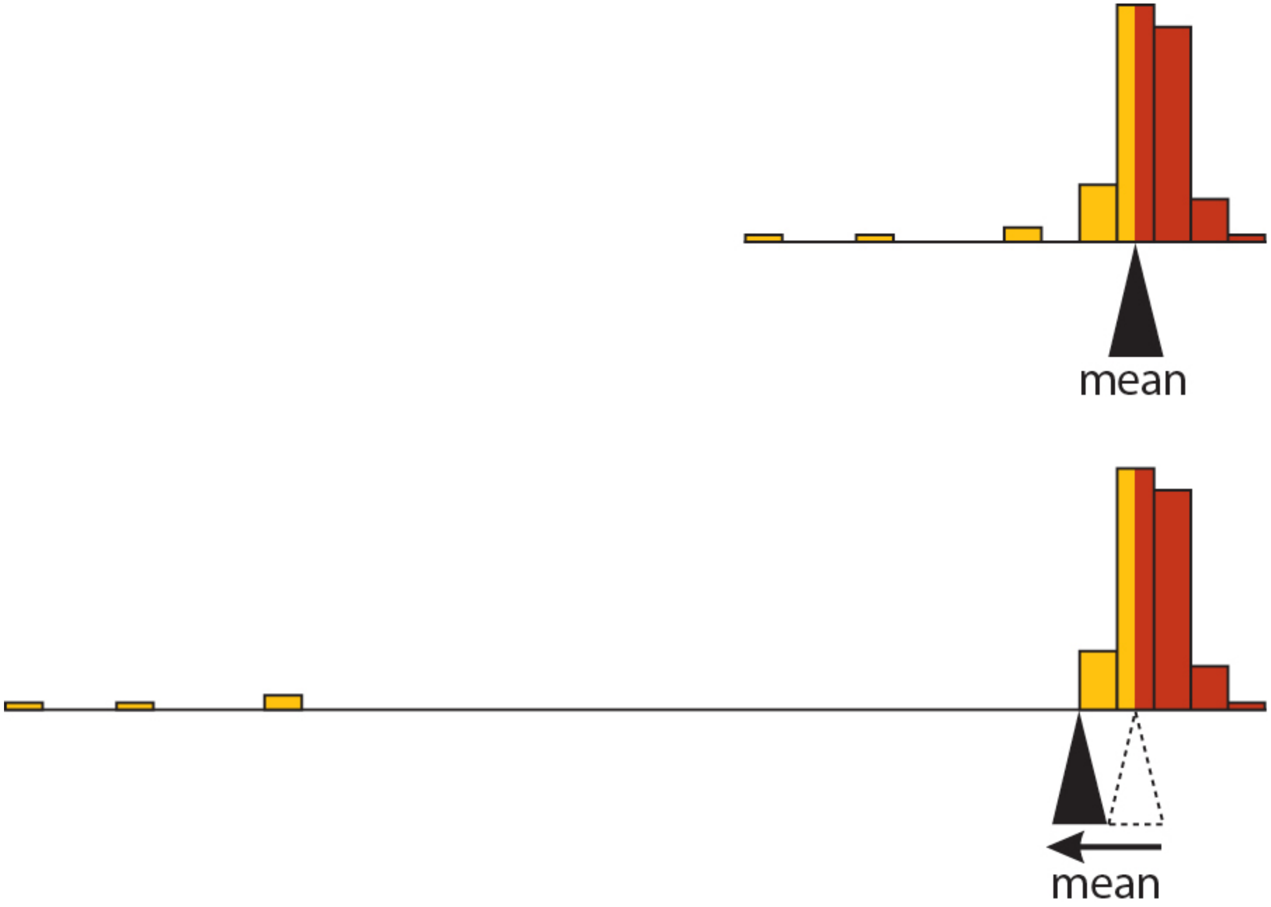


Genotype	Mean	Median
MM	62.8	63
Mm	50.4	59
mm	11.7	11

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



The mean is more sensitive to extreme observations than the median



Mean and median for US household income, 2005

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

Why?

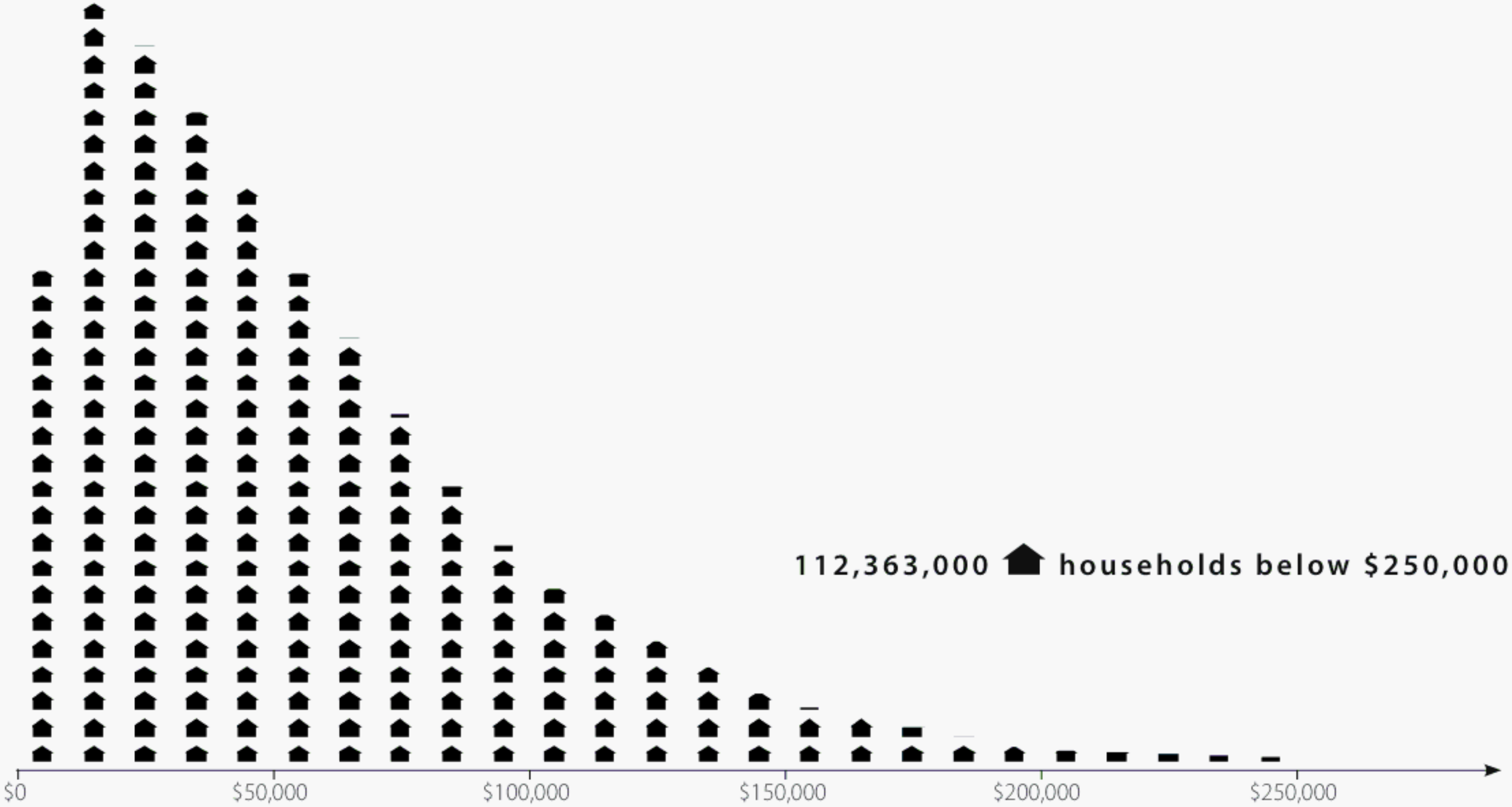
Visualizing Economics
Making the *Invisible Hand* Visible

Visit www.visualizingeconomics.com
to view more examples

2005 United States

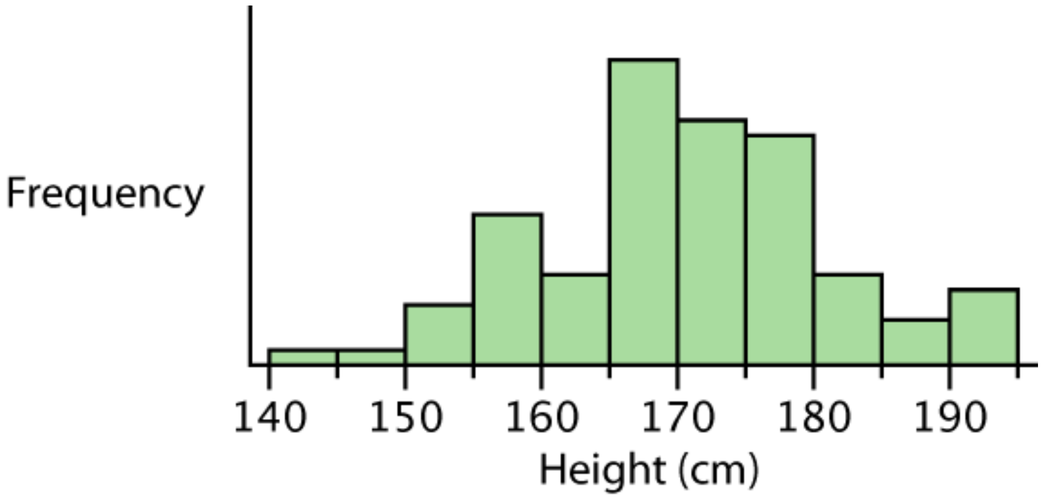
Income Distribution (Bottom 98%)

Each ☗ equals 500,000 households



112,363,000 ☗ households below $250,000

University student heights



Mean 169.3 cm

Median 170 cm

Mode 165-170 cm

Measures of width

- Range
- Standard deviation
- Variance
- Coefficient of variation
- Interquartile range

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

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

n is the sample size

Sample variance

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

n is the sample size

Why n -1?

- Samples value closer to sample mean than true (population) mean
- Thus, standard deviation smaller (=biased)
- The n – 1 makes the estimate unbiased

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”

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

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

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

(in units of siblings **squared**)

Standard deviation (SD)

- Positive square root of the variance

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

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

$$s^2 = 0.70$$

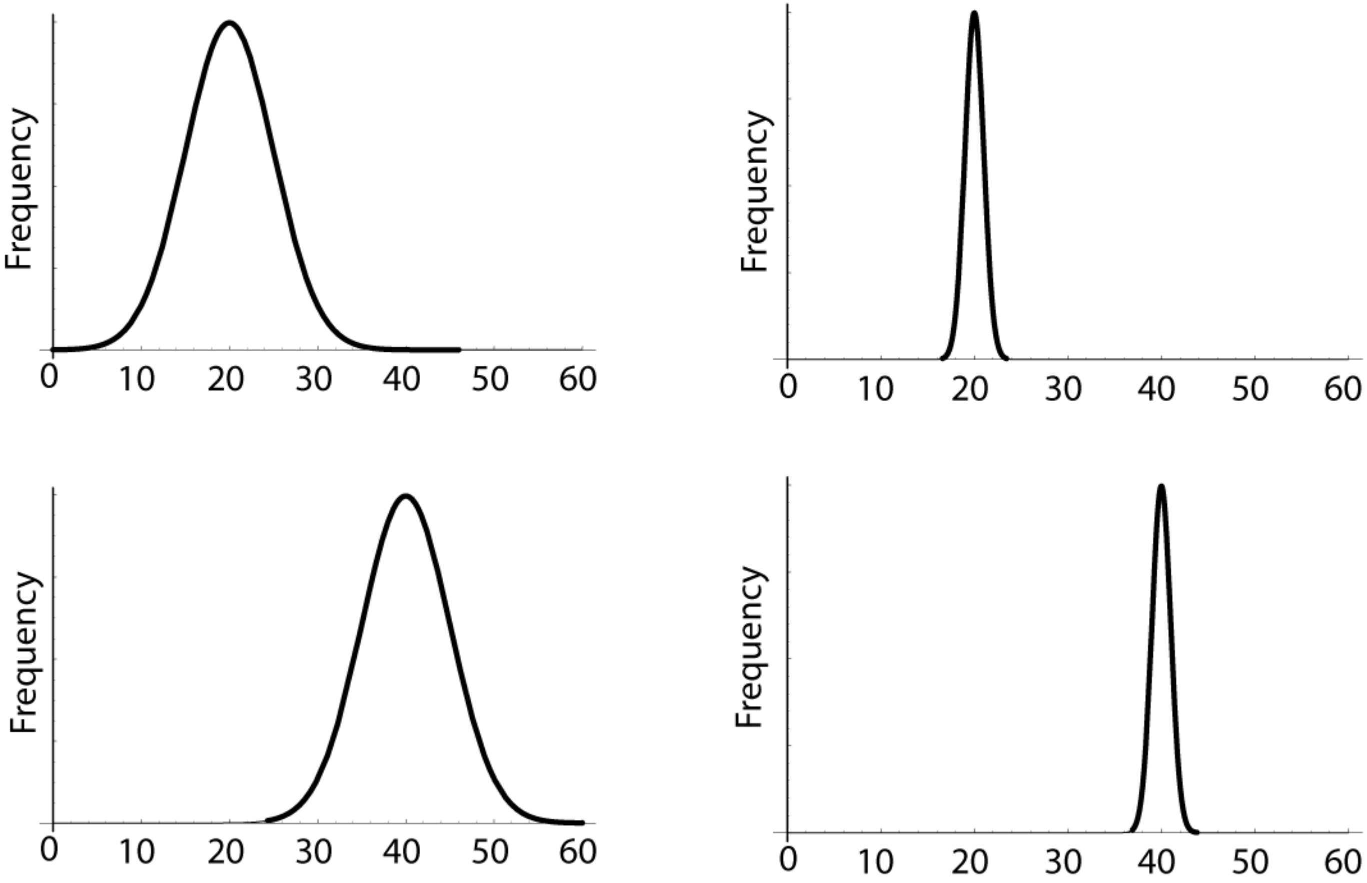
$$s = \sqrt{0.70} = 0.84$$

Standard deviation: 5

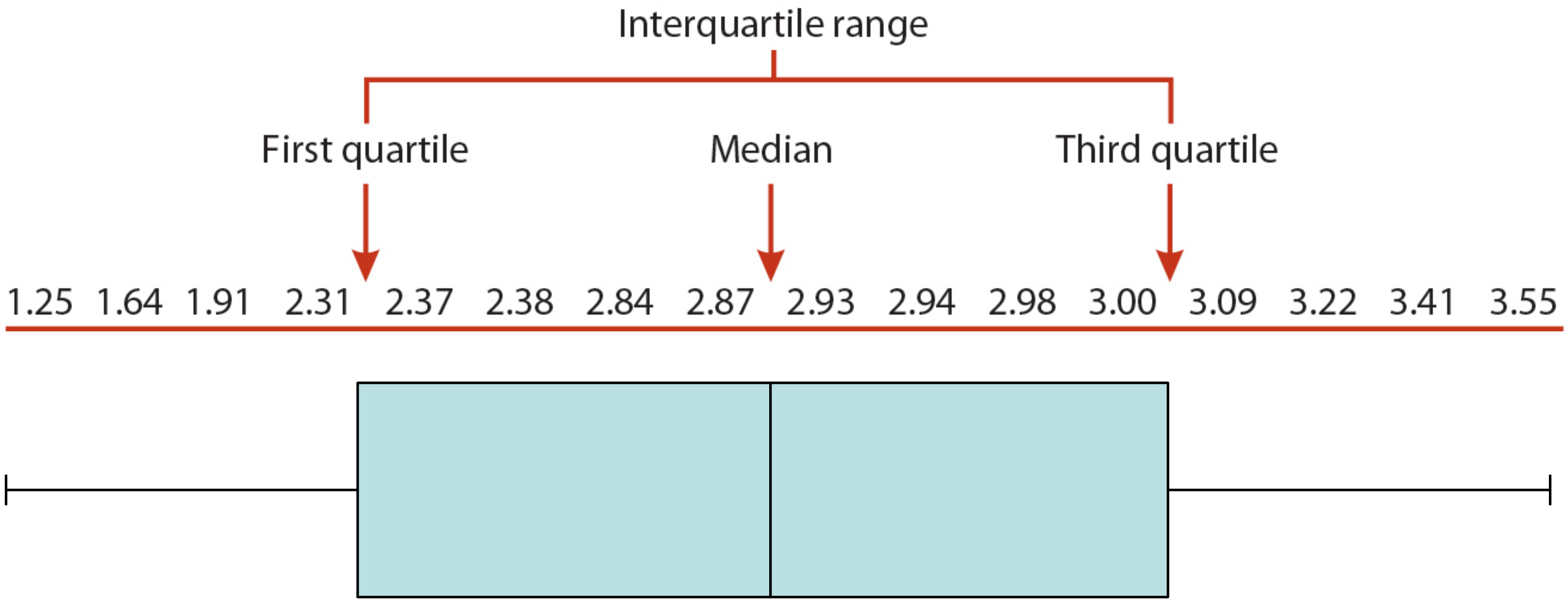
Standard deviation: 1

Mean: 20

Mean: 40



Interquartile Range

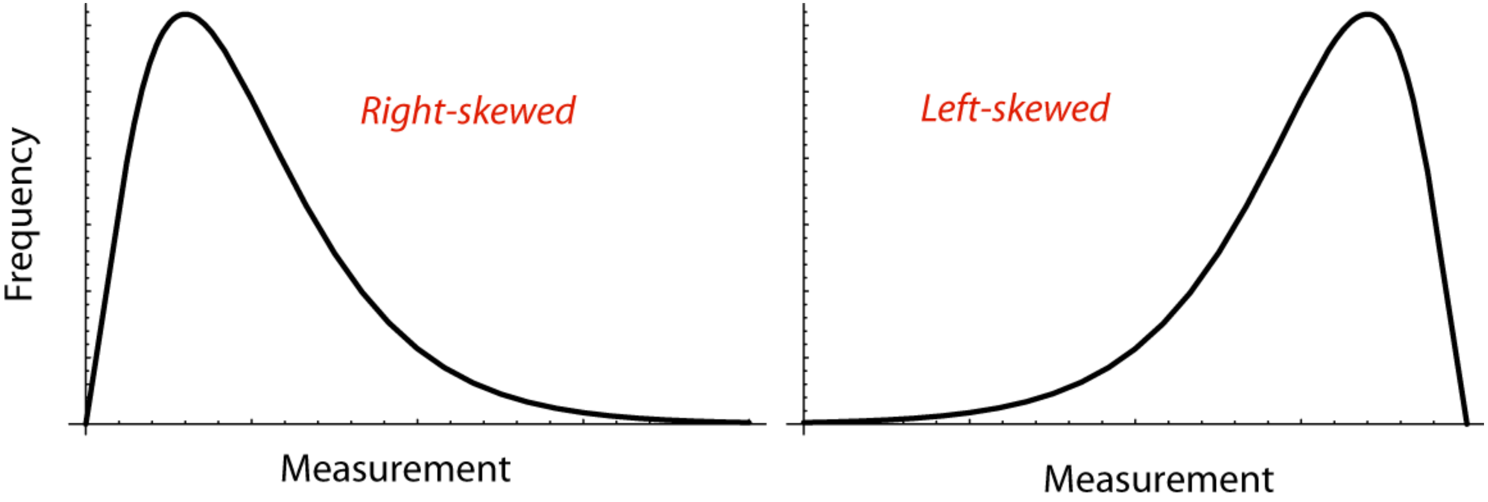


Extreme values on box plots

Values lying farther from the box edge than 1.5 times the interquartile range

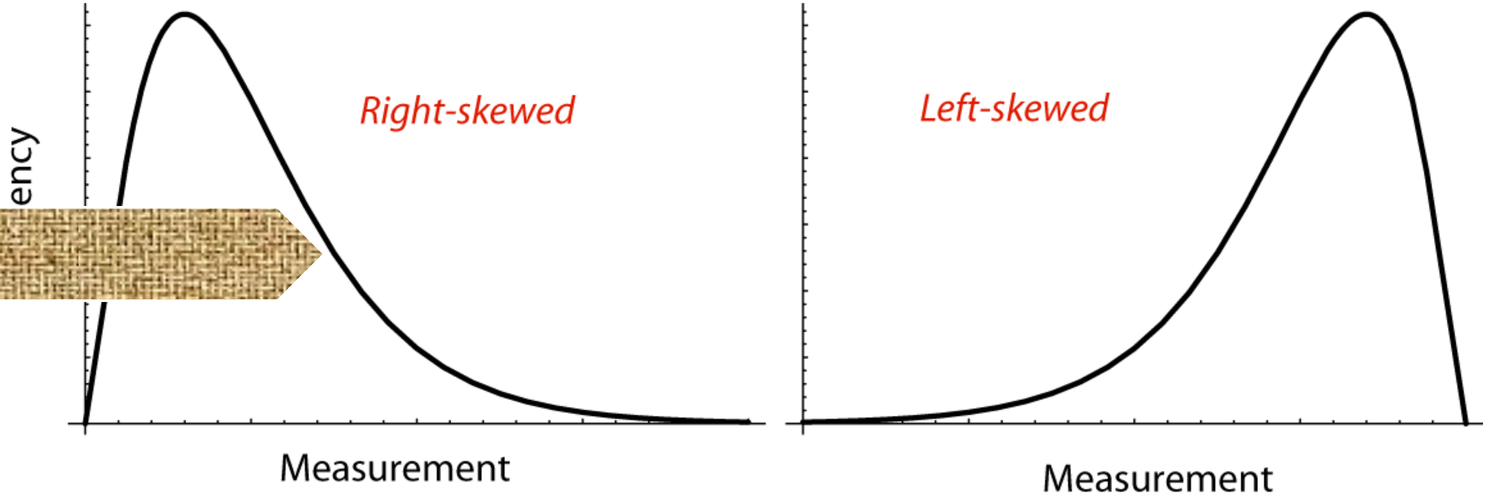
Skew

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



Skew

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



Nomenclature

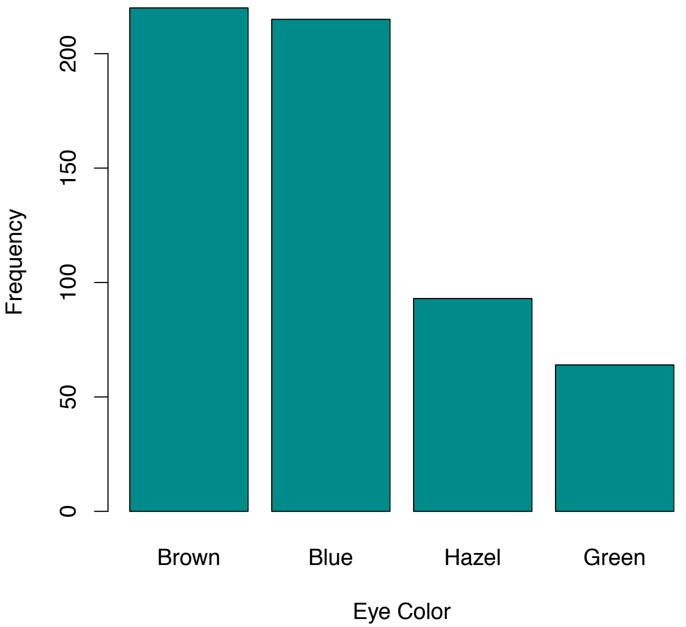
	Population Parameters	Sample Statistics
Mean	μ	$\overline{Y}$
Variance	σ^2	s^2
Standard Deviation	σ	s

Greek Roman

One common description of categorical data

- Proportion

$$\hat{p} = \frac{\text{number in category}}{n}$$



Eye Color	Proportion
Brown	220 / 592 = 0.37
Blue	215 / 592 = 0.36
Hazel	93 / 592 = 0.16
Green	64 / 592 = 0.11

Calculate Mean, Median and Mode

9 14 4 7 2 18 2

Calculate Mean, Median and Mode

9 14 4 7 2 18 2

Mean: 9 + 14 + 4 + 7 + 2 + 18 + 2 = 56

56 / 7 = 8

Calculate Mean, Median and Mode

9 14 4 7 2 18 2

Mean: 9 + 14 + 4 + 7 + 2 + 18 + 2 = 56

56 / 7 = 8

Median: 2 2 4 7 9 14 18

Calculate Mean, Median and Mode

9 14 4 7 2 18 2

Mean: 9 + 14 + 4 + 7 + 2 + 18 + 2 = 56

56 / 7 = 8

Median: 2 2 4 7 9 14 18

Mode: 2 2 4 7 9 14 18

Calculate Sample Variance and Standard Deviation

9 14 4 7 2 18 2

$\bar{Y}$ = 8

Variance:

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

Standard deviation:

$$s = \sqrt{s^2}$$

Calculate Sample Variance and Standard Deviation

	Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
	2	-6	36
	2	-6	36
	4	-4	16
	7	-1	1
	9	1	1
	14	6	36
	18	10	100
Sums:	56		226

$\bar{Y} = 8$

$s^2 = 226 / 6 = 37.7$

$s = \sqrt{37.7} = 6.1$

What are the units?

9 14 4 7 2 18 2 (cm)

Mean: 8 cm

Median: 7 cm

Mode: 2 cm

Variance: 37.7 cm^2

Standard Deviation: 6.1 cm

Experimental vs. Observational Study

- Experimental Study
  - Treatments are assigned randomly
  - Can assign cause – effect relationships

- Observational Study
  - No control over which individual falls into which group
  - Can only show associations but not cause - effect

Does smoking cause cancer?

In a sample of 25 year old Canadian man, living in the city, randomly drawn smokers had a higher chance of getting lung cancer.

Observational or experimental?

Show cause – effect?

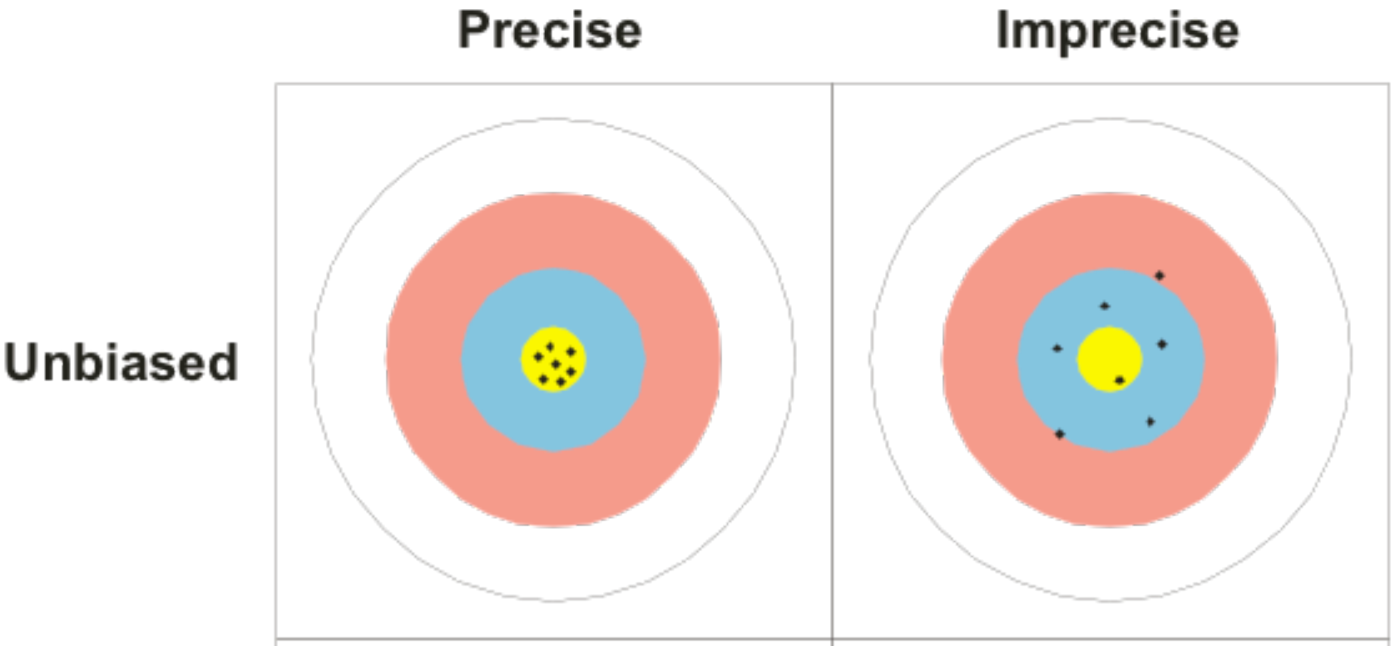
Estimating with uncertainty

Estimation and Sampling Error

- Estimation is inferring a population parameter from sample data

- In the face of chance, how much do we trust an estimate?

  → We need to be able to quantify uncertainty based on sampling error

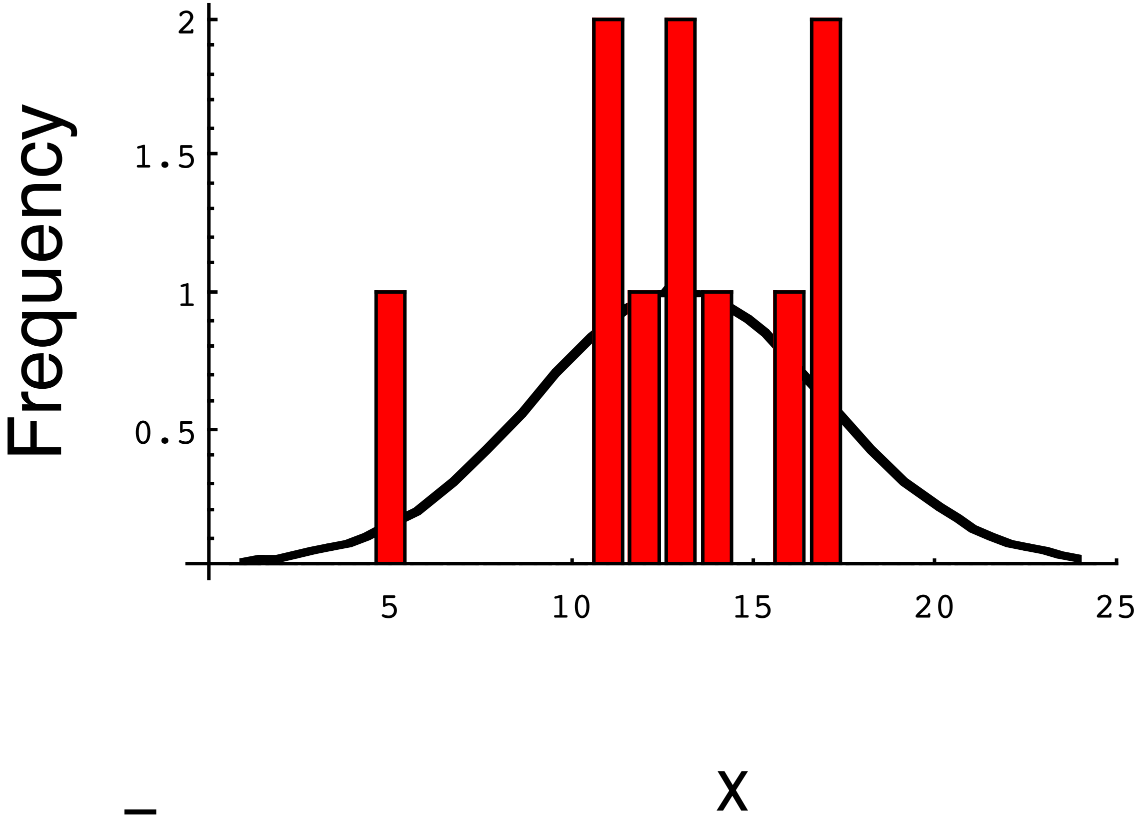


Precise

Imprecise

Unbiased

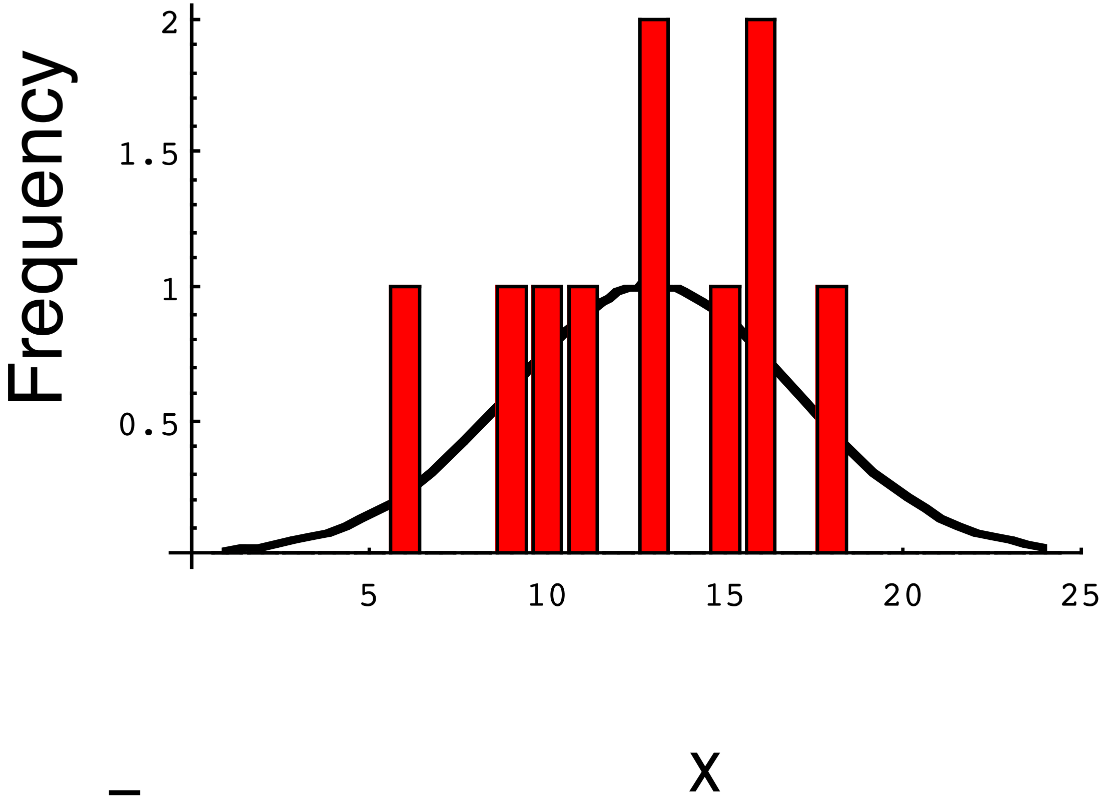
Sample size 10 from Normal distribution with μ=13 and σ^2=16



$\bar{X} = 13.5$

$s^2 = 12.1$

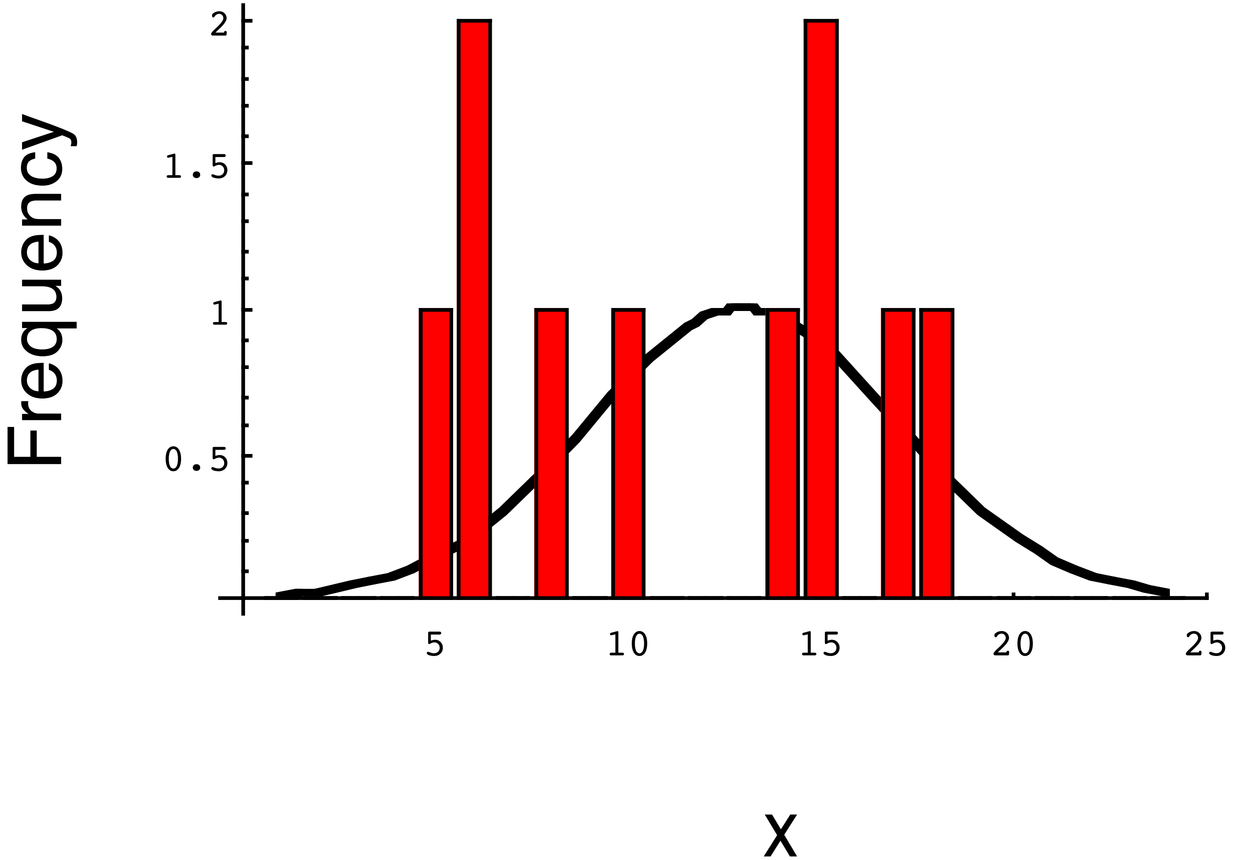
Another sample of 10 from same distribution



$\bar{X} = 13.3$

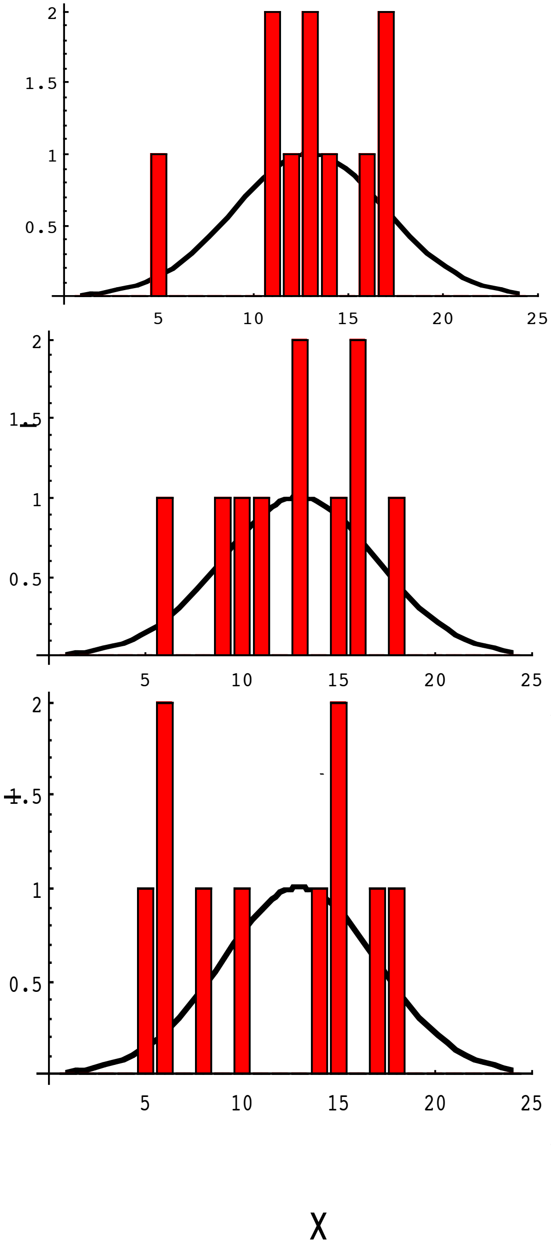
$s^2 = 13.0$

A third sample of 10 from the same distribution



$\bar{X} = 11.9$

$s^2 = 28.3$

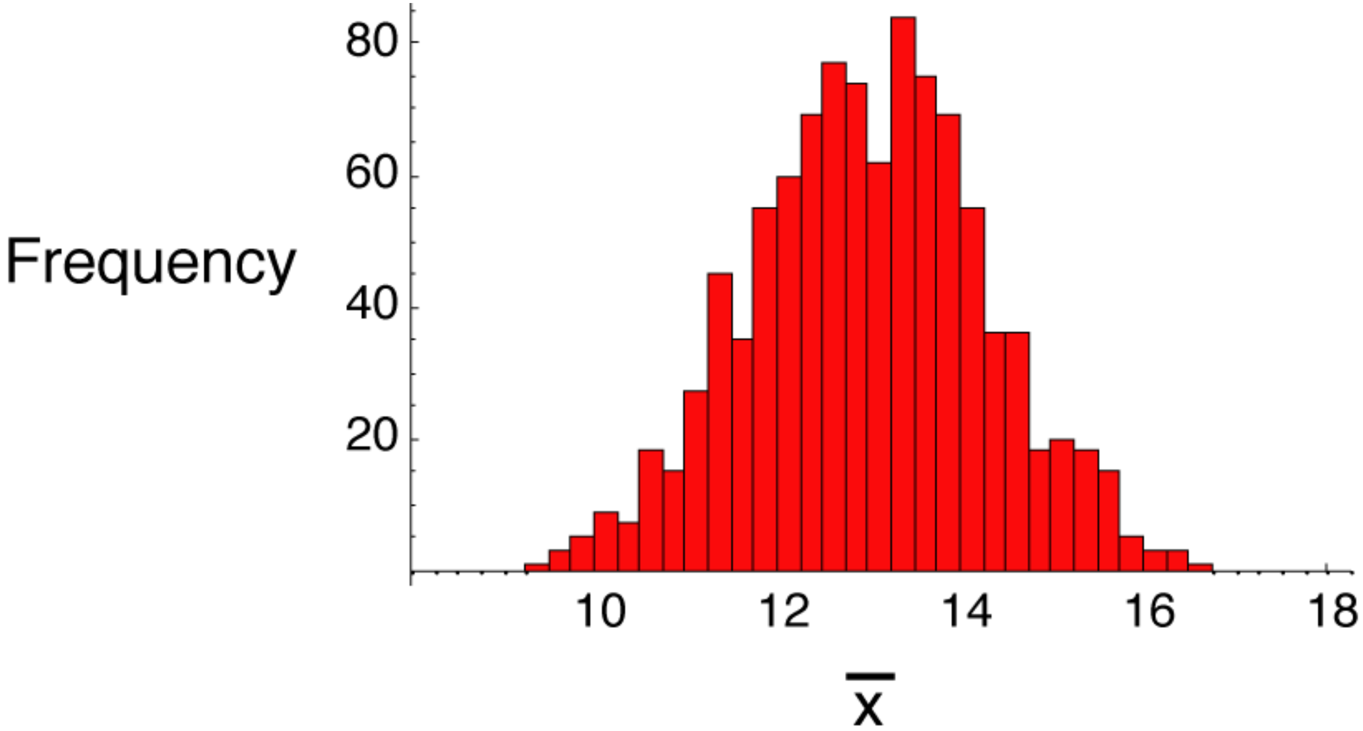


$\overline{X} = 13.5$
$s^2 = 12.1$

$\overline{X} = 13.3$
$s^2 = 13.0$

$\overline{X} = 11.9$
$s^2 = 28.3$

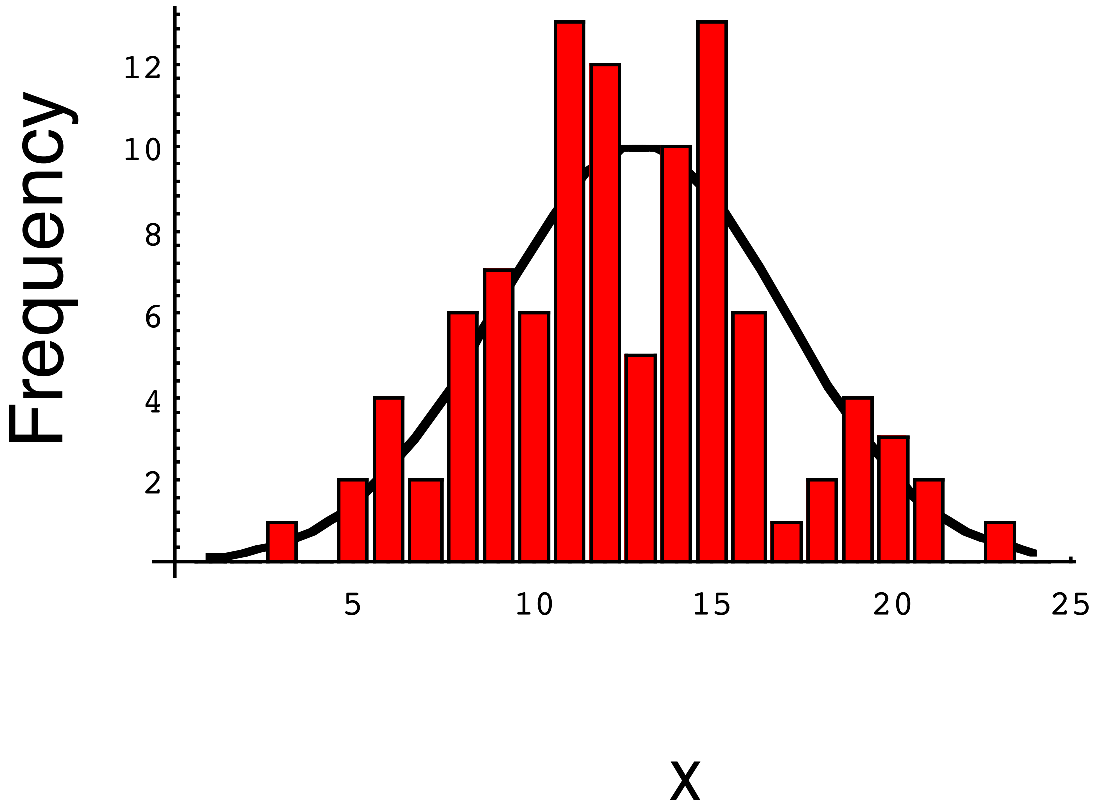
Distribution of the means of many samples, each of sample size 10



Sampling Distribution

The probability distribution of all values for an estimate that we might obtain when we sample a population

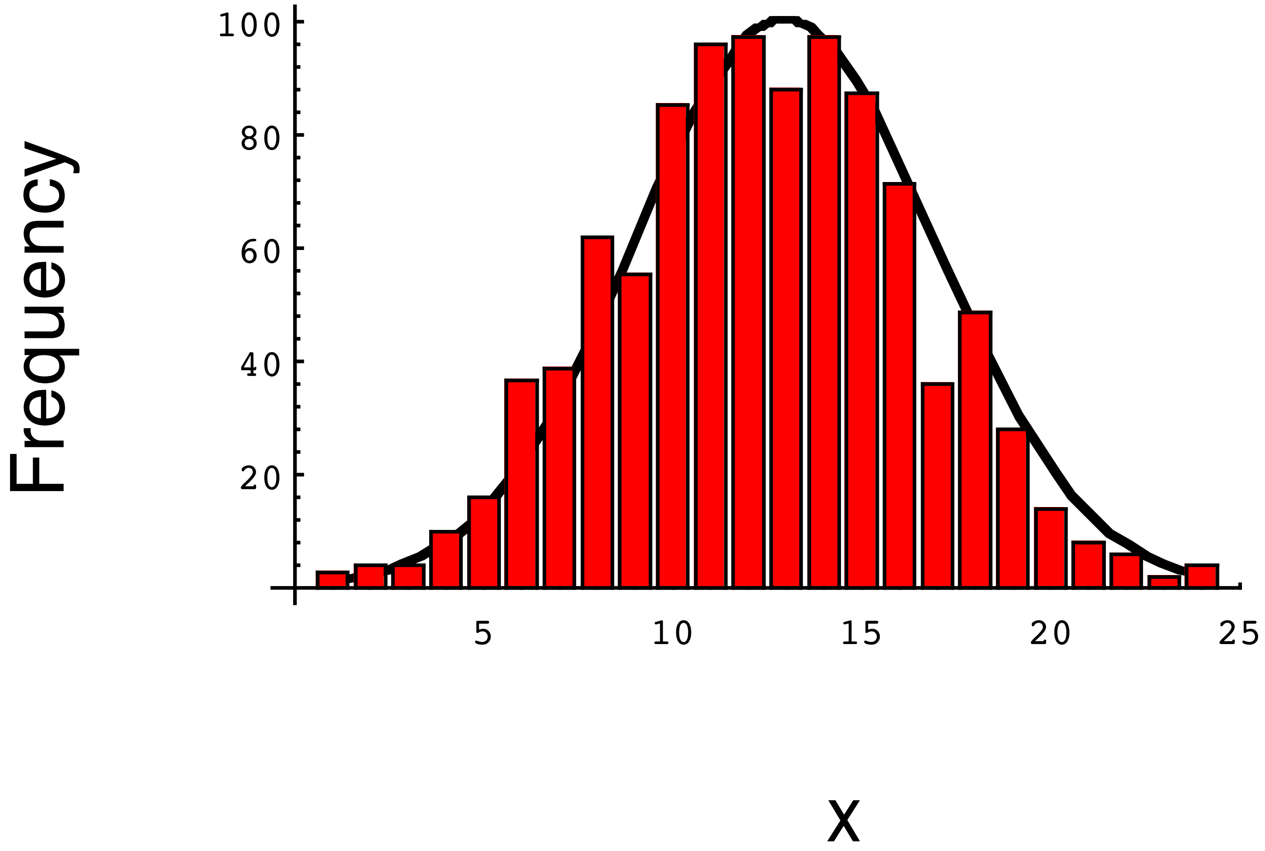
A sample of <u>100</u> from the same population distribution



$\bar{X} = 13.0$

$s^2 = 15.6$

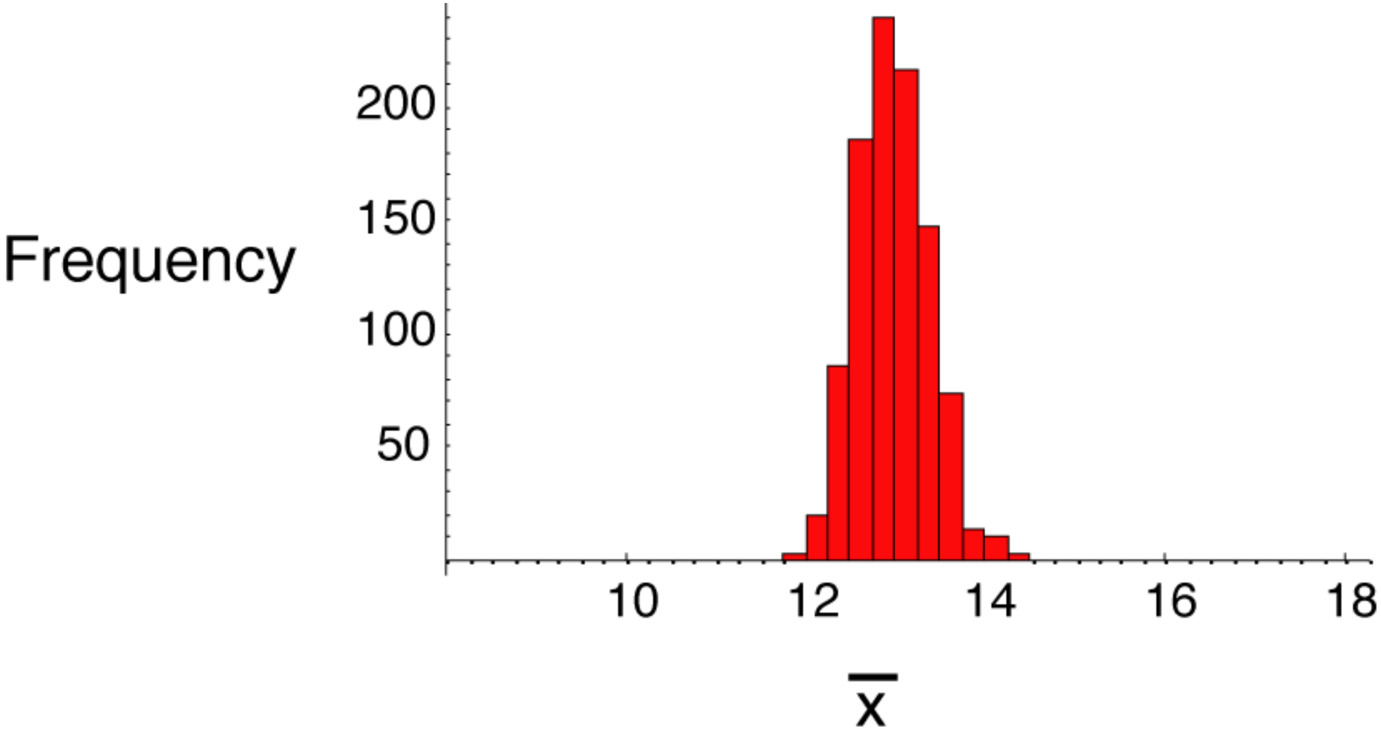
A sample of 1000 from the same population distribution



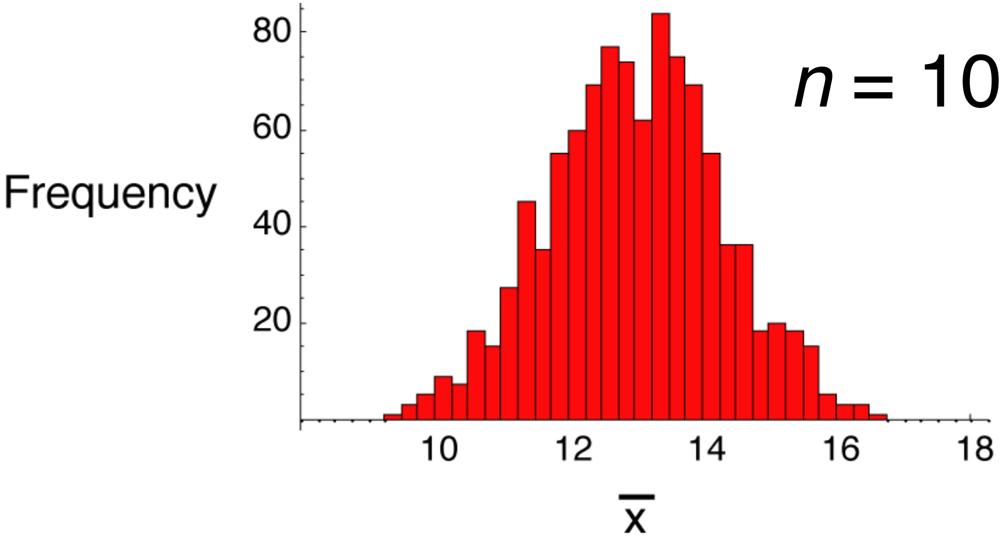
$\bar{X} = 12.9$

$s^2 = 16.3$

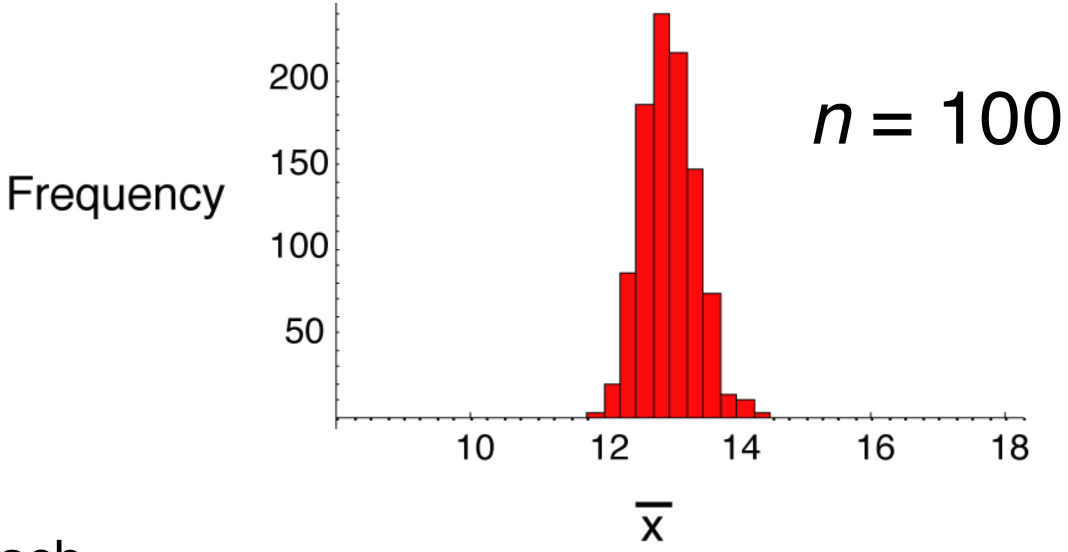
Distribution of the means of many samples, each of sample size 100



Variation in sample means decreases with sample size



$n = 10$



$n = 100$

1000 samples each

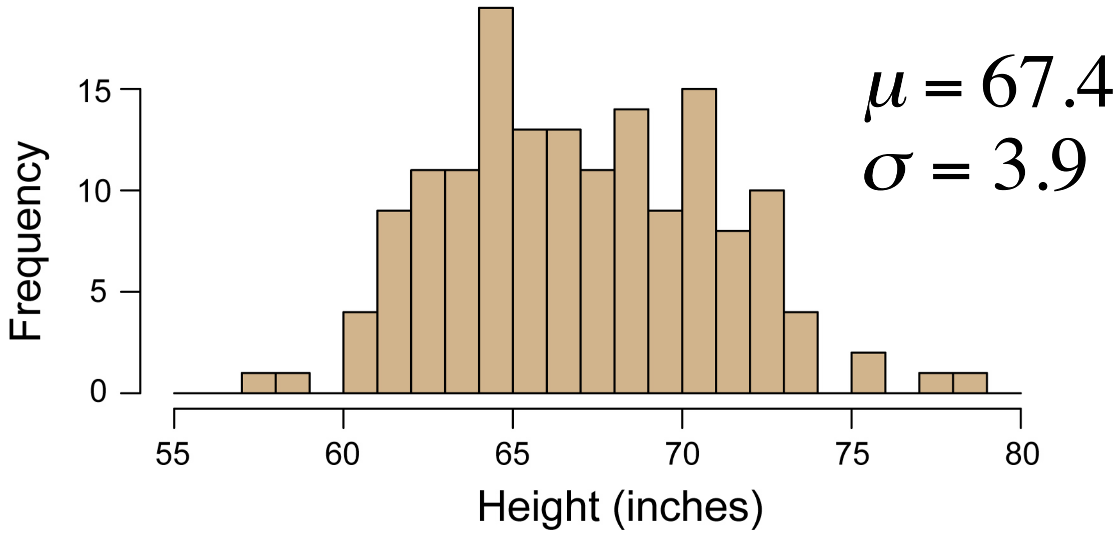
The *standard error* of an estimate is the standard deviation of its sampling distribution.

The standard error predicts the *sampling error* of the estimate.

Standard Deviation vs. Standard Error

- Standard deviation gives you a sense of the spread in your sample (how far typical individuals are from your estimate)
- Standard error gives you a sense of how far your estimate is likely to be from the true parameter

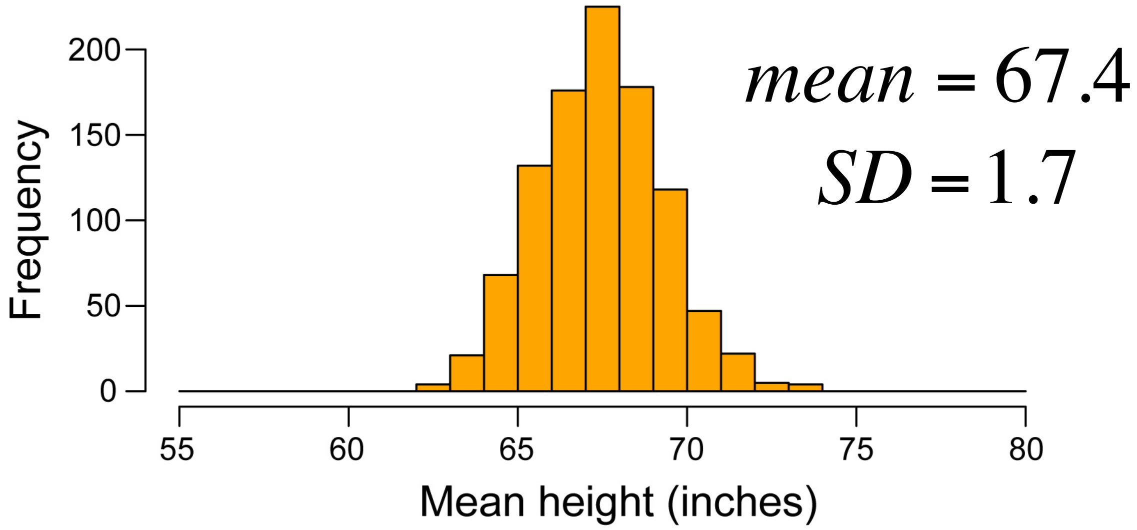
Heights of BIOL300 students (*N* = 157)



$\mu = 67.4$

$\sigma = 3.9$

Mean heights of samples of size 5
(1000 samples)



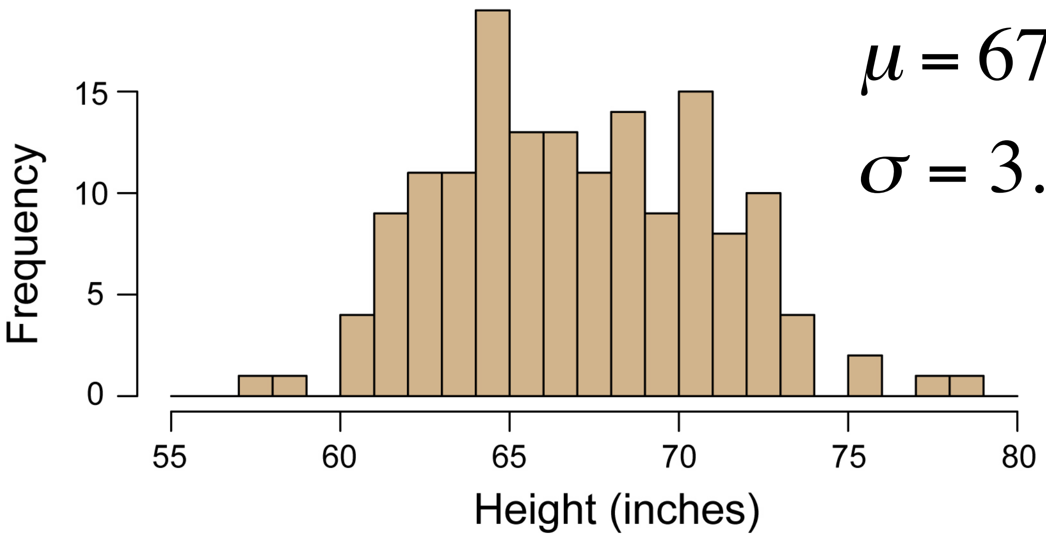
$mean = 67.4$

$SD = 1.7$

Standard error of the mean

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

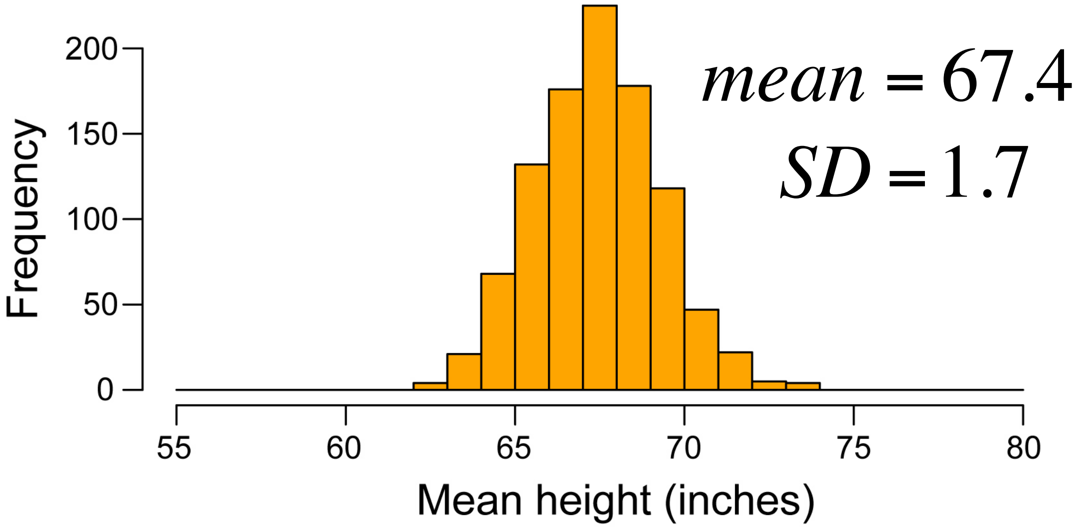
Heights of BIOL300 students (*N* = 157)



$\mu = 67.4$

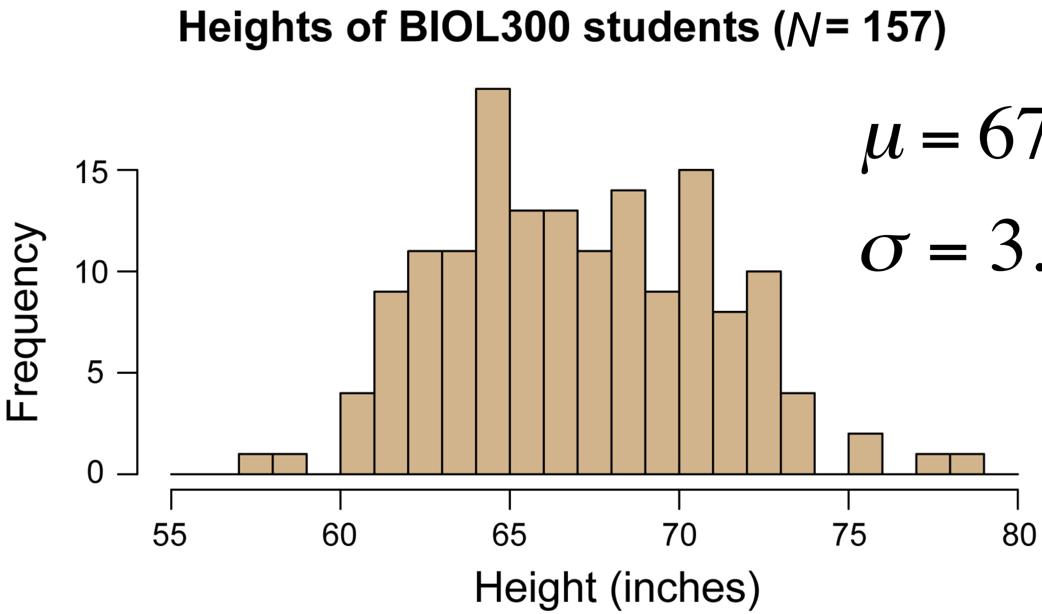
$\sigma = 3.9$

Mean heights of samples of size 5
(1000 samples)



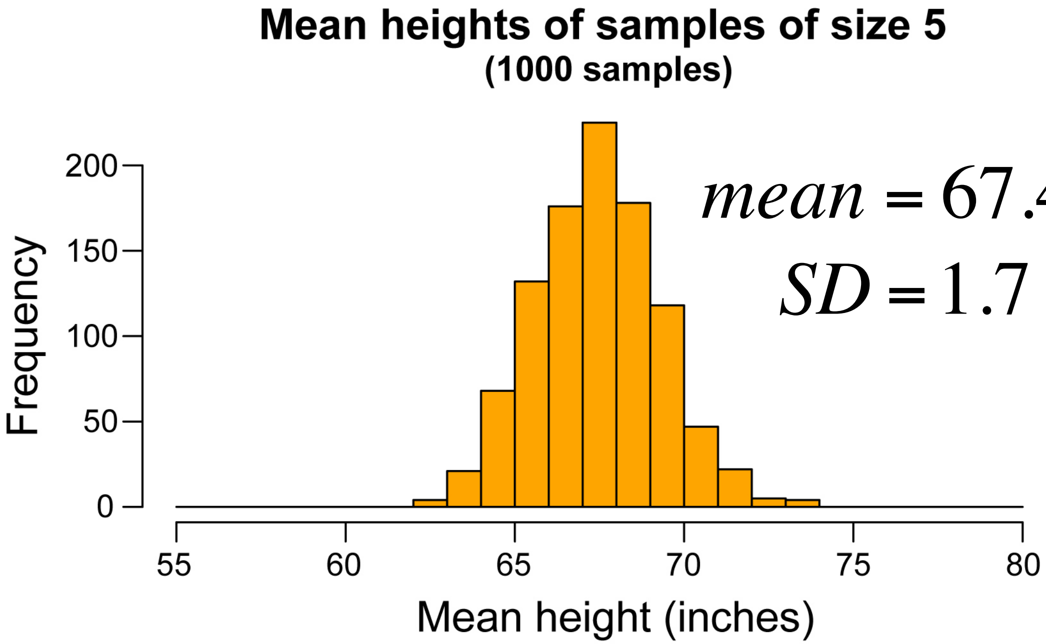
$mean = 67.4$

$SD = 1.7$



$$\mu = 67.4$$

$$\sigma = 3.9$$



$$mean = 67.4$$

$$SD = 1.7$$

$$\mu_{\bar{Y}} = \mu = 67.4$$

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{3.9}{\sqrt{5}} = 1.7$$

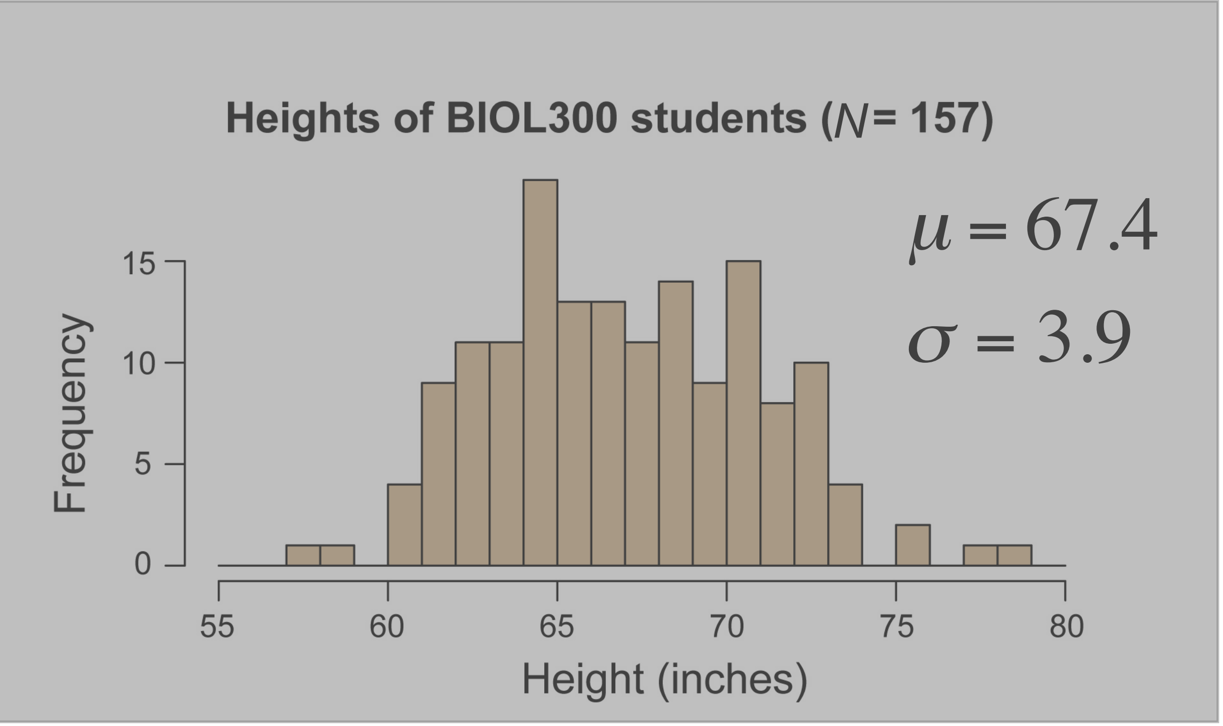
The math works!

The problem is, we rarely know σ.

Estimate of the standard error of the mean

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$$

This gives us some knowledge of the likely difference between our sample mean and the true population mean.


Heights of BIOL300 students (*N*= 157)

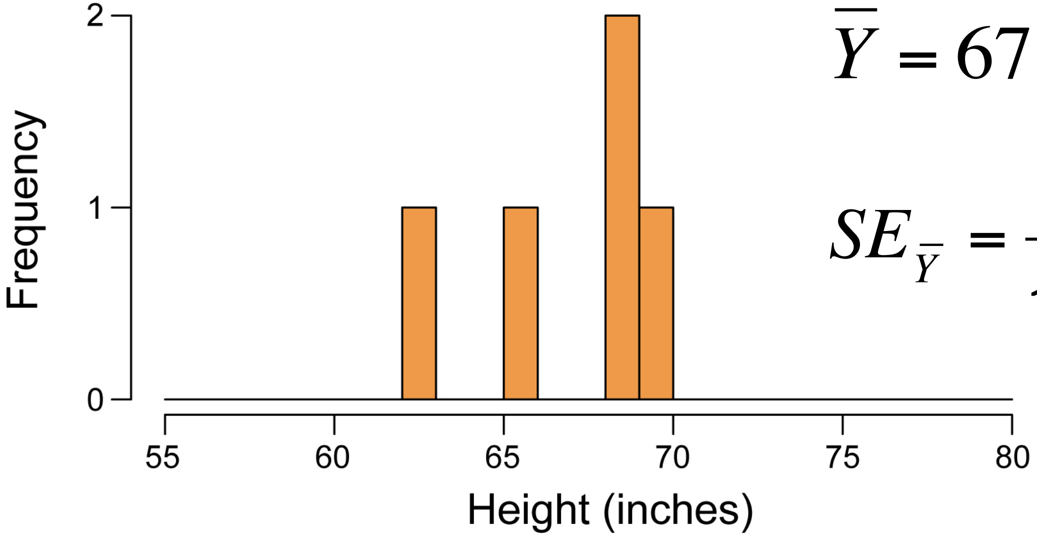
$\mu = 67.4$

$\sigma = 3.9$

In most cases, we don't know the real population distribution.

We only have a sample.

Heights of a sample of students (*n* = 5)



$\bar{Y} = 67.1 \qquad s = 3.1$

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{5}} = 1.4$$

We use this as an estimate of $\sigma_{\bar{Y}}$

95% Confidence Interval

The 95% confidence interval provides a plausible range for a parameter. All values for the parameter lying within the interval are plausible, given the data, whereas those outside are unlikely.

95% Confidence Interval

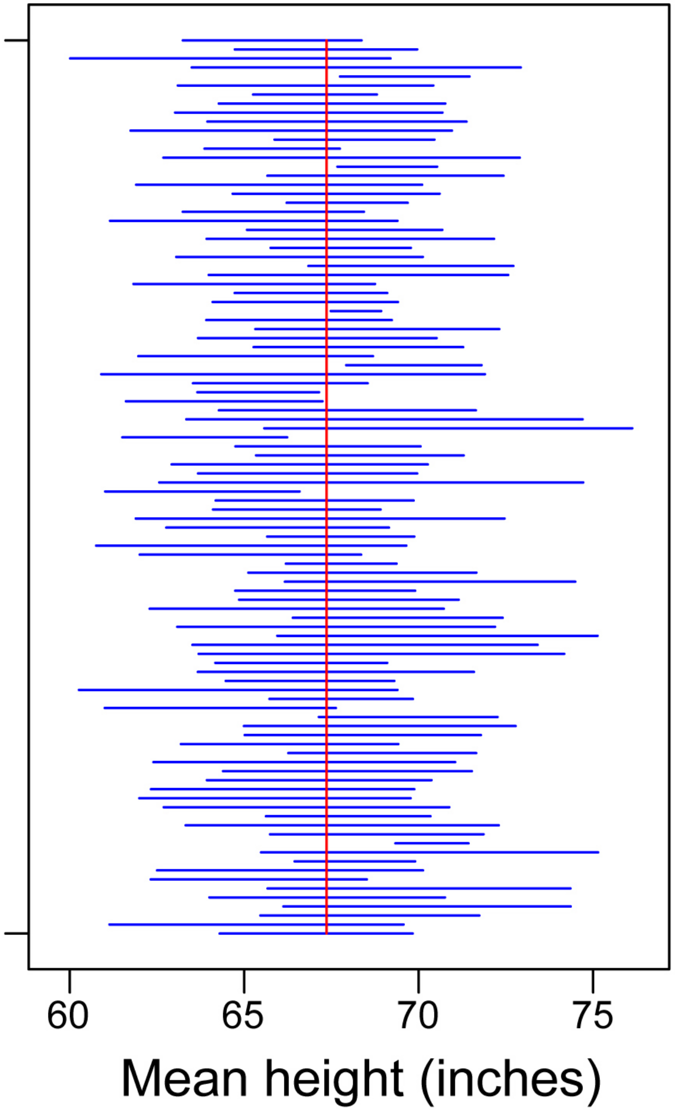
If you took 1000 independent samples and calculated the 95% confidence interval of your estimate from each, ~950 of them would contain the true population parameter

The 2SE rule-of-thumb

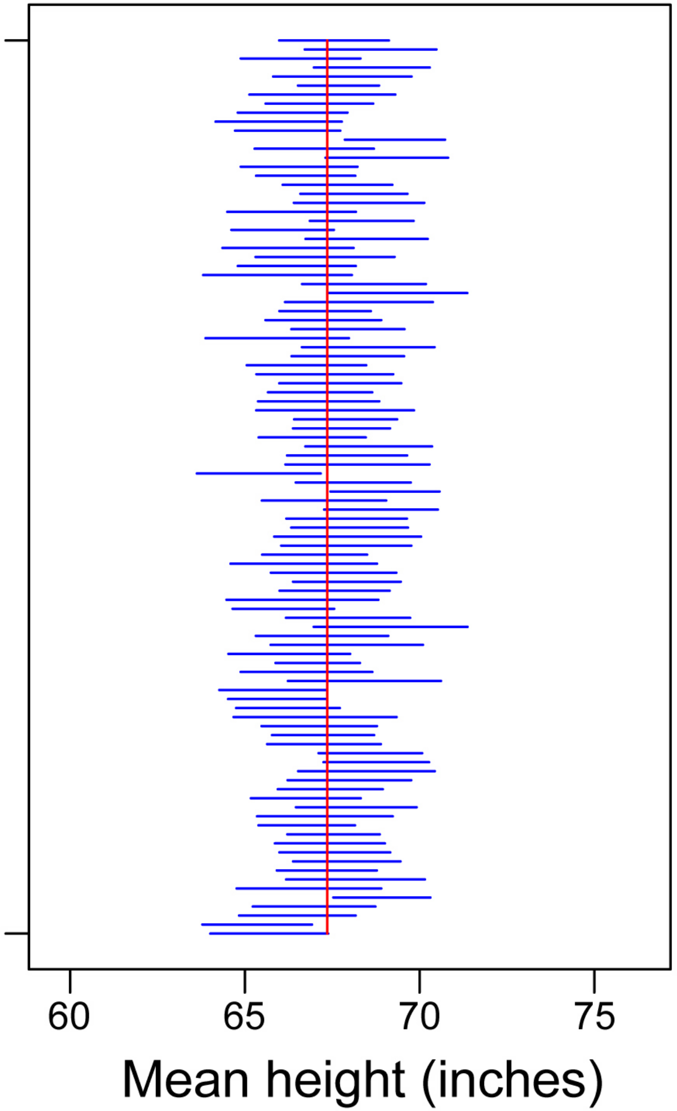
The interval from $\bar{Y} - 2SE_{\bar{Y}}$ to $\bar{Y} + 2SE_{\bar{Y}}$ provides a rough estimate of the 95% confidence interval for the mean.

(*Assuming normally distributed population and/or sufficiently large sample size.*)

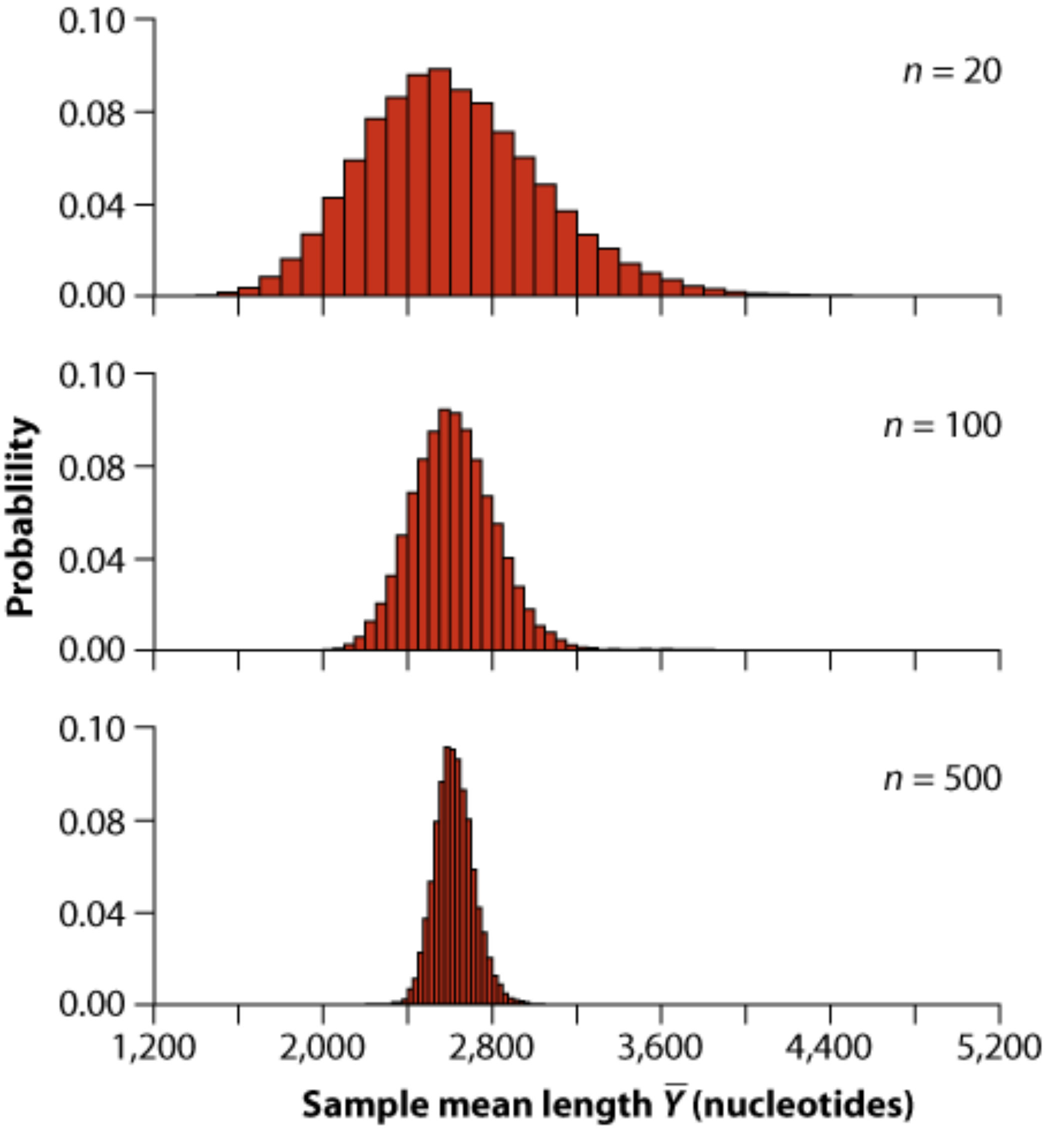
Means ± 2 SE of samples of size 5
(100 samples)



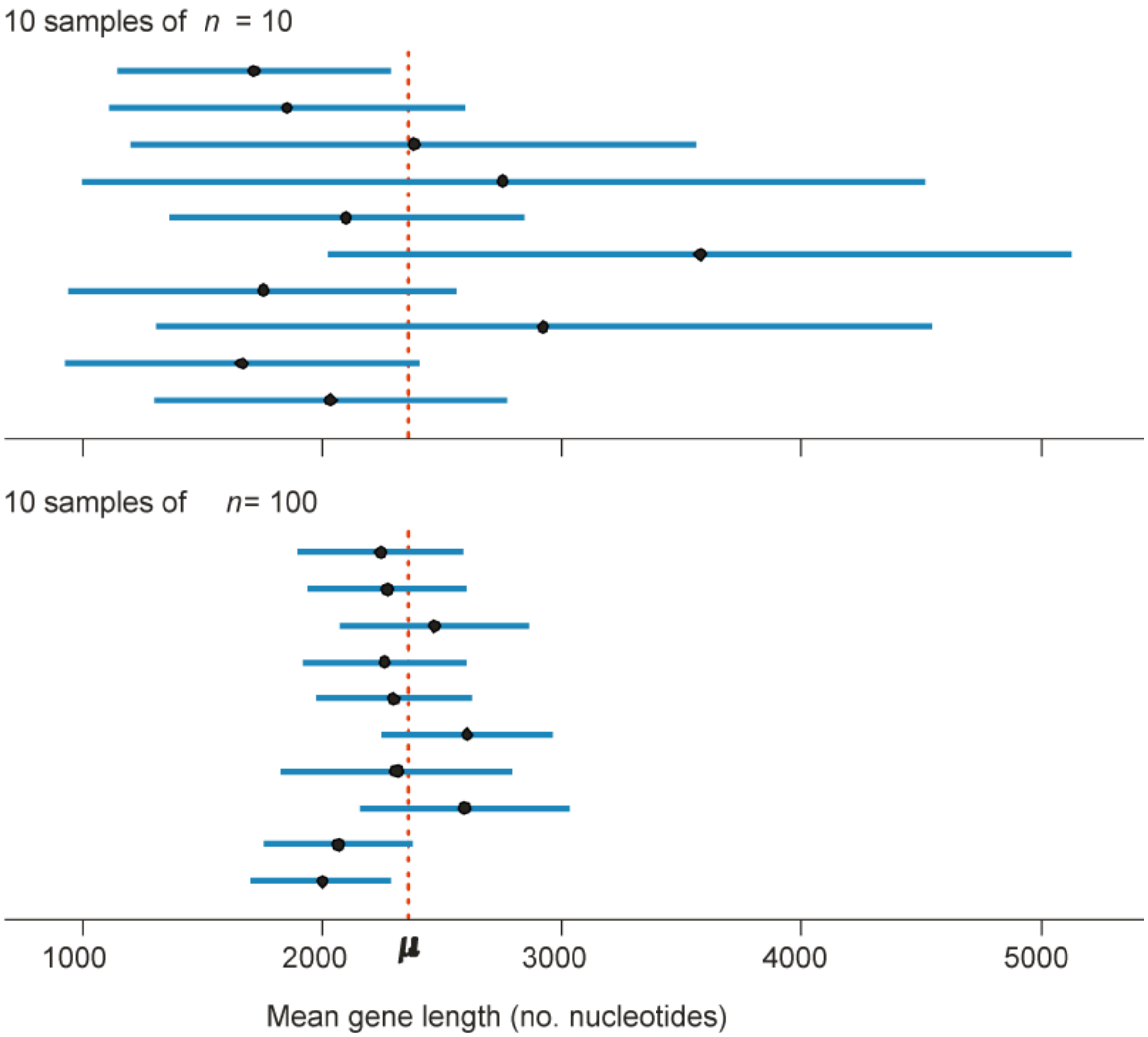
Means ± 2 SE of samples of size 20
(100 samples)



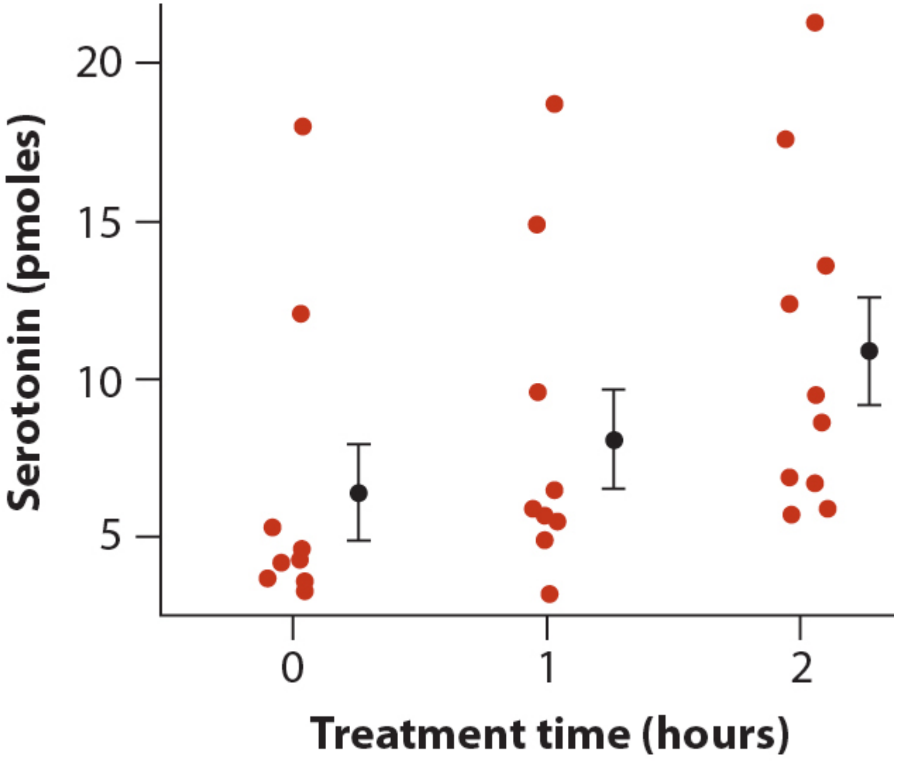
Sampling distributions of gene sizes



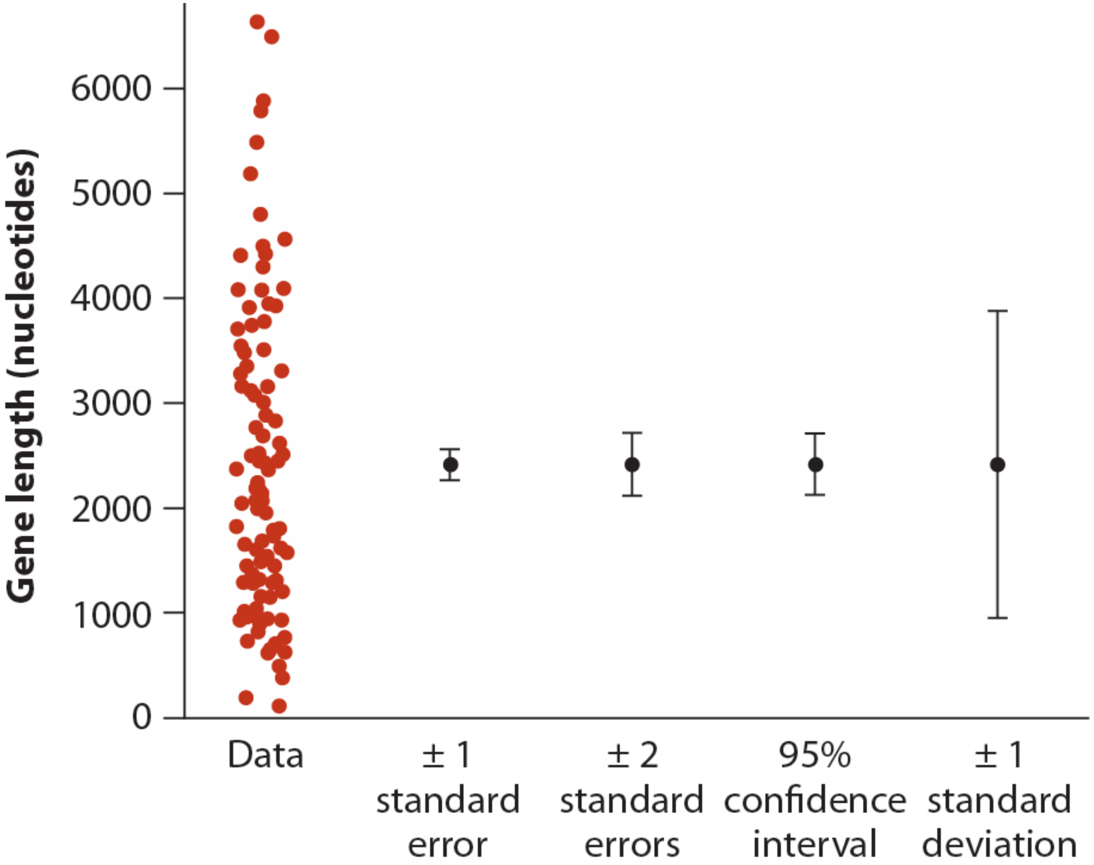
Confidence interval



Error bars



Error bars



Gene length (nucleotides)
6000
5000
4000
3000
2000
1000
0
Data
± 1 standard error
± 2 standard errors
95% confidence interval
± 1 standard deviation

Pseudoreplication

The error that occurs when samples are not independent, but they are treated as though they are.

Example: Pseudoreplication

You are interested in average pulse rate of mountain climbers. Since they are hard to find, you decide to take 10 measurements from each climber. You study 6 climbers, so you have 60 measurements.

What is your sample size (*n*)?

Avoiding pseudoreplication

You are interested in average pulse rate of mountain climbers. Since they are hard to find, you decide to take 10 measurements from each climber. You study 6 climbers, so you have 60 measurements.

Take the mean blood pressure for each climber, so that you have 6 pulse rates, one for each climber ($n = 6$).

As in other vertebrates, individual zebrafish differ from one another along the shy-bold behavioral spectrum. In addition to other behavioral differences, bolder individuals tend to be more aggressive, whereas shy individuals tend to be less aggressive. Norton et al. (2011) compared several behaviors associated with this syndrome between zebrafish that had the *spd* mutant at the *Fgfr1a* gene and the "wild type" lacking the mutation. The data below are measurements of the amount of time, in seconds, that individual zebrafish with and without this mutation spent in aggressive activity over 5 minutes when presented with a mirror image.

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

1)What is the mean and standard deviation of seconds in aggressive activity for each genotype?

2)What are the standard errors of these estimates of the means?

3)Give approximate 95% confidence intervals of the means. Provide upper and lower limits.

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

1) What is the mean and standard deviation of seconds in aggressive activity for each genotype?

Mean:

Wild type: 0 + 21 + 22 + 28 + 60 + 80 + 99 + 101 + 106 + 129 + 168 = 814

814 / 11 = 74.0

Spd mutant: 96 + 97 + 100 + 127 + 128 + 156 + 162 + 170 + 190 + 195 = 1421

1412/10 = 142.1

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

1) What is the mean and standard deviation of seconds in aggressive activity for each genotype?

Y_i	$Y_i - \bar{Y}$	$(Y_i - \bar{Y})^2$
0	-74	5476
21	-53	2809
22	-52	2704
28	-46	2116
60	-14	196
80	6	36
99	25	625
101	27	729
106	32	1024
129	55	3025
168	94	8836

Sums: 814 *27576*

Wild type:

$s^2 = 27576 / 10 = 2757.6$

s = 52.5

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

1) What is the mean and standard deviation of seconds in aggressive activity for each genotype?

Y_i	$Y_i - \overline{Y}$	$(Y_i - \overline{Y})^2$
0	-74	5476
21	-53	2809
22	-52	2704
28	-46	2116
60	-14	196
80	6	36
99	25	625
101	27	729
106	32	1024
129	55	3025
168	94	8836

Sums: 814 *27576*

Wild type:

s^2 = 27576 / 10 = 2757.6

s = 52.5

Spd mutant:

s^2 = 12818.9 / 9 = 1424.3

s = 37.7

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

2)What are the standard errors of these estimates of the means?

Wild type: 52.5 / $\sqrt{11}$ = 15.8

Spd mutant: 37.7 / $\sqrt{10}$ = 11.9

Wild type: 0, 21, 22, 28, 60, 80, 99, 101, 106, 129, 168

Spd mutant: 96, 97, 100, 127, 128, 156, 162, 170, 190, 195

3)Give approximate 95% confidence intervals of the means. Provide upper and lower limits.

Wild type:

74 + 2(15.8) = 105.6
74 – 2(15.8) = 42.4
$42.4 < \mu < 105.6$

Spd mutant:

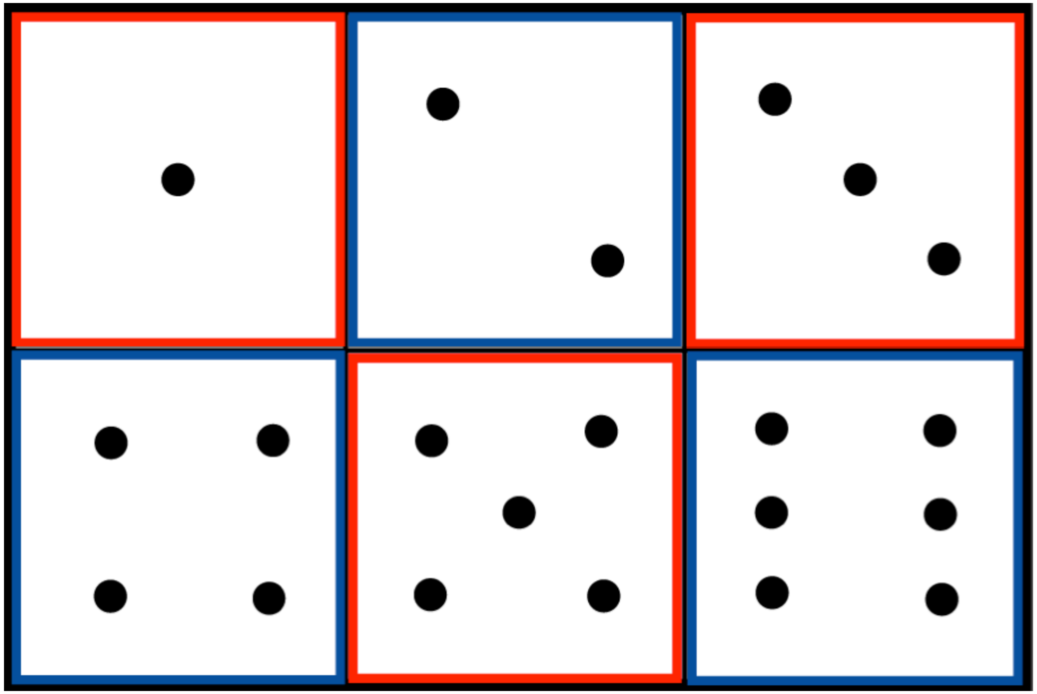
142.1 + 2(11.9) = 165.9
142.1 – 2(11.9) = 118.3
$118.3 < \mu < 165.9$

Probability

The *probability* of an event is its true relative frequency; the proportion of times the event would occur if we repeated the same process over and over again.

Two events are *mutually exclusive* if they cannot both be true.

Mutually exclusive



Mutually exclusive

Heads, Tails

Boy, Girl

Ace, King

Apple, Orange

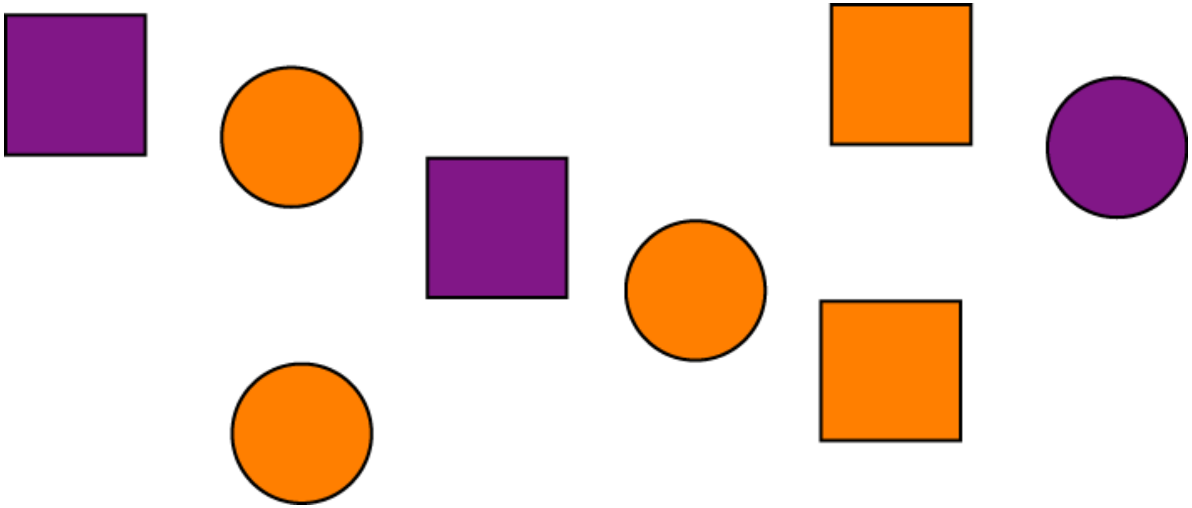
Mutually exclusive

$$\Pr(A \text{ and } B) = 0$$

Not mutually exclusive

Pr(*A* and *B*) ≠ 0

Pr(*purple* AND *square*) ≠ 0



Not mutually exclusive

Heads 1st flip, Tails 2nd flip

Boy, Green eyes

Ace, Hearts

Apple, Red

For example

Event A: First child is female

Event B: Second child is female

P(A) = 0.48

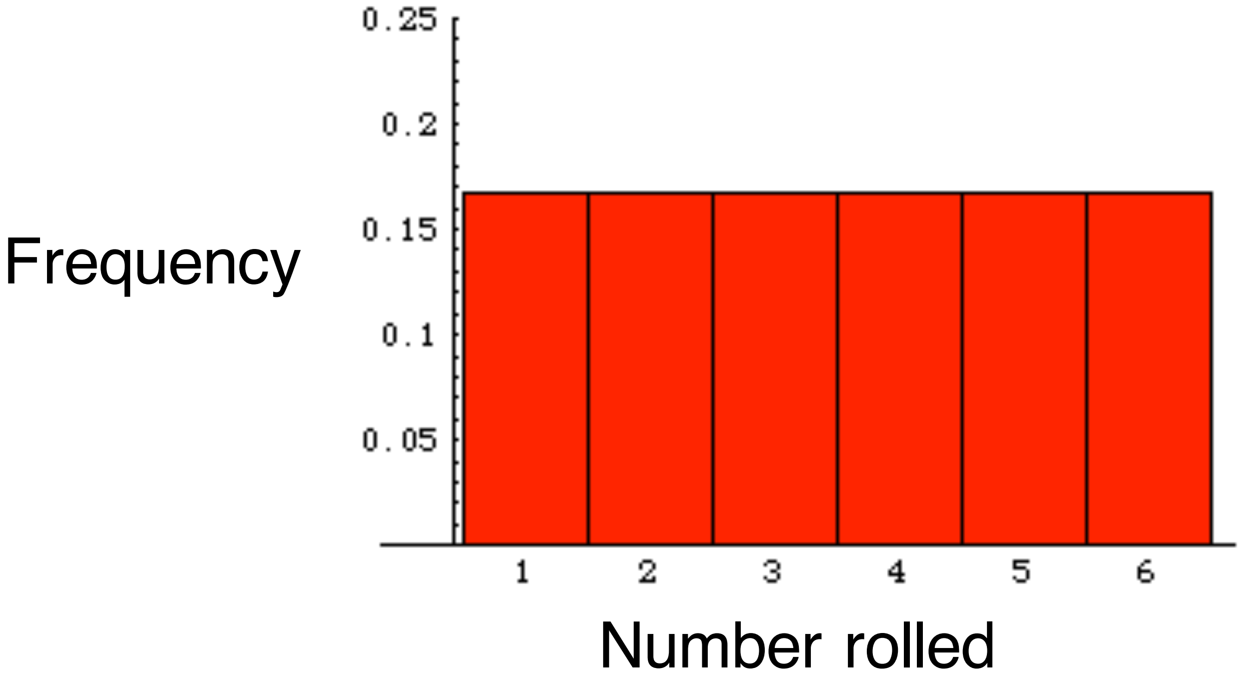
P(B) = 0.48

But P(*A* and *B*) ≠ 0, so these events are NOT mutually exclusive.

Probability distribution

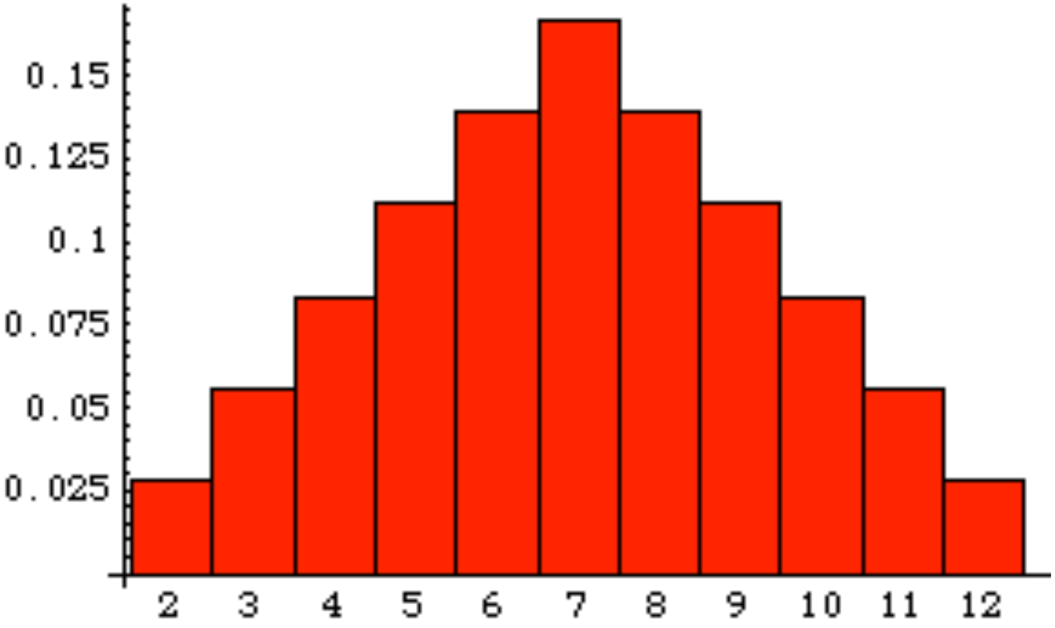
A *probability distribution* describes the true relative frequency (a.k.a. the probability) of all possible values of a random variable.

Probability distribution for the outcome of a roll of a die



Probability distribution for the sum of a roll of two dice

Frequency



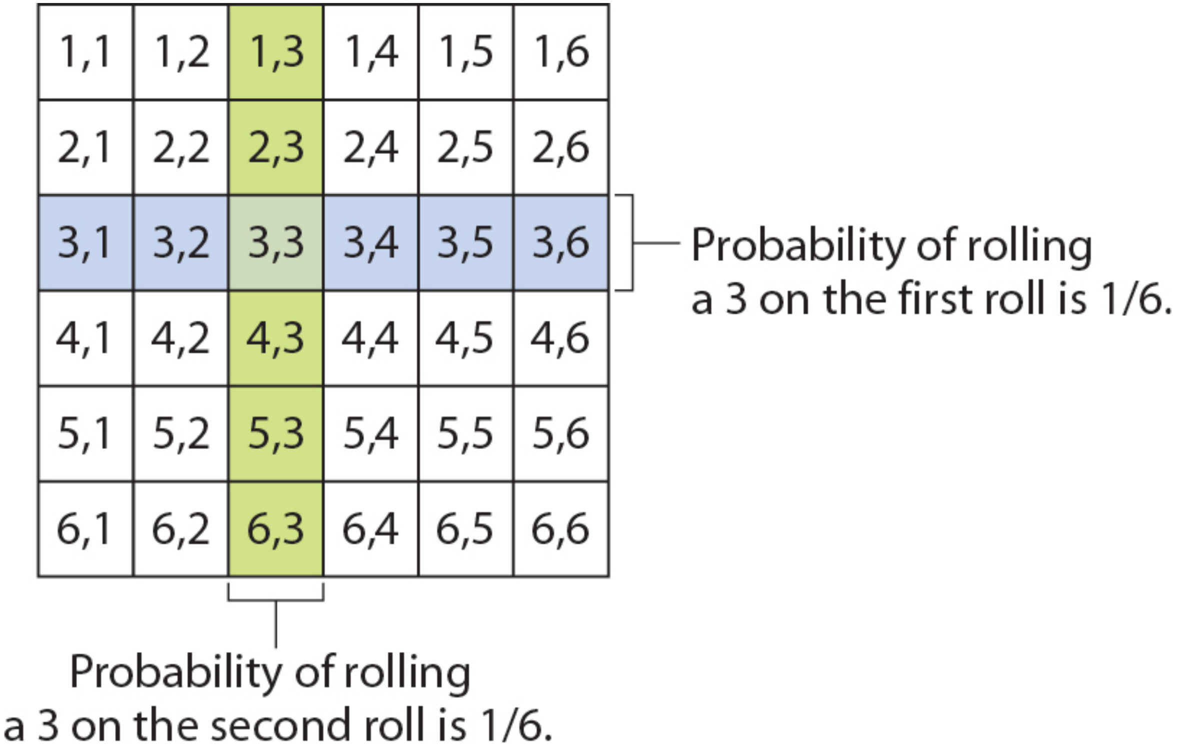
Sum of two dice

The addition principle

If two events *A* and *B* are mutually exclusive, then

$$\Pr[A \text{ OR } B] = \Pr[A] + \Pr[B]$$

The addition principle



1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Probability of rolling
a 3 on the first roll is 1/6.

Probability of rolling
a 3 on the second roll is 1/6.

P[1st or 2nd roll is 3] = 1/6 + 1/6 = 1/3

The probability of a range

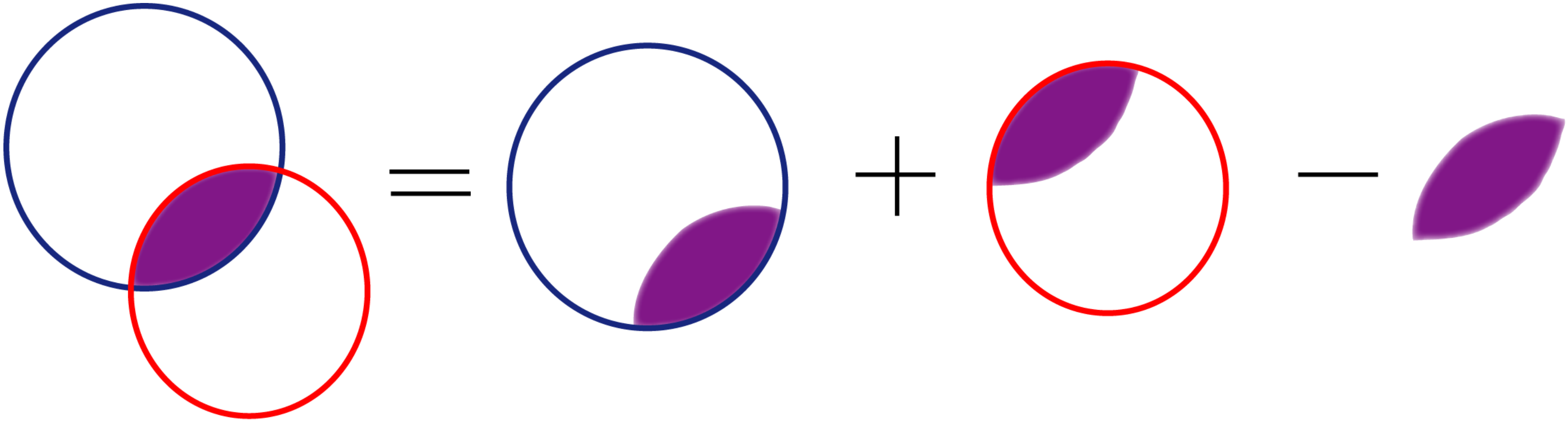
Pr[*Number of boys* ≥ 6] = Pr[6] + Pr[7] + Pr[8]....

The probabilities of all possibilities add to 1.

Probability of *Not*

Pr[NOT *rolling a 2*] = 1 – Pr[*Rolling a 2*] = 5/6

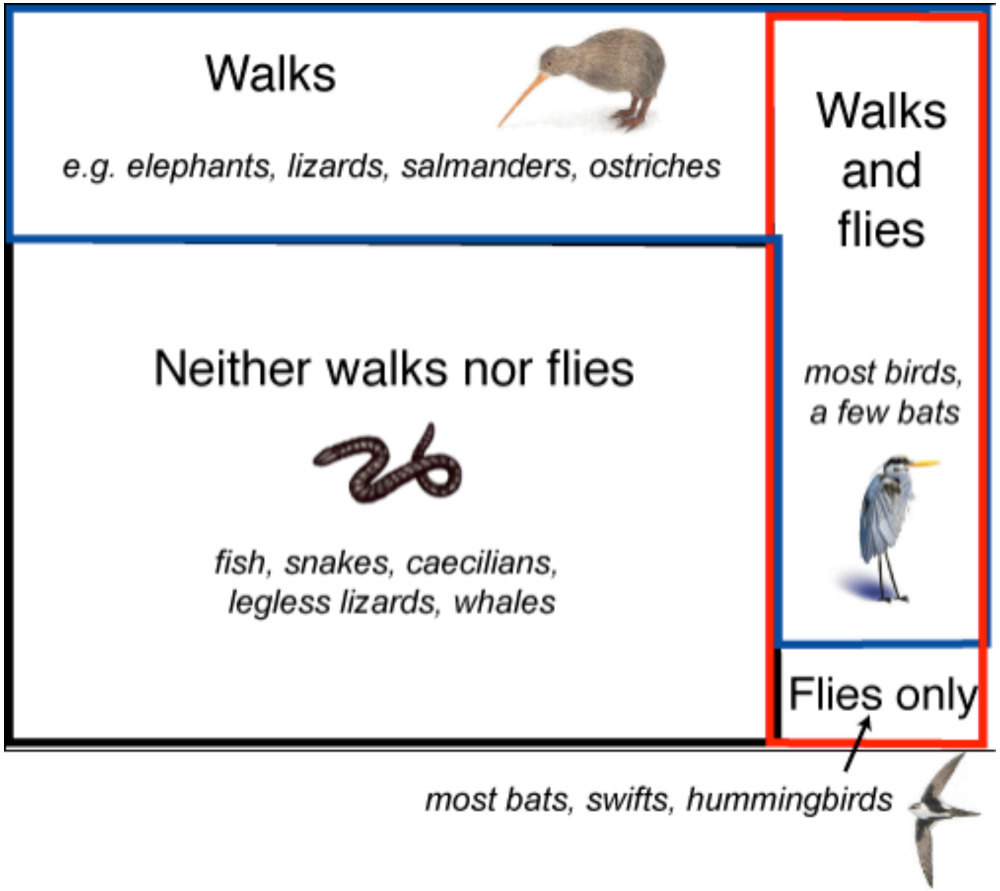
General Addition Principle



Pr[*A* OR *B*] = Pr[*A*] + Pr[*B*] - Pr[*A* AND *B*]

General Addition Principle

$$\Pr[A \text{ OR } B] = \Pr[A] + \Pr[B] - \Pr[A \text{ AND } B].$$


Walks
e.g. elephants, lizards, salmanders, ostriches
Walks and flies
most birds, a few bats
Neither walks nor flies
fish, snakes, caecilians, legless lizards, whales
Flies only
most bats, swifts, hummingbirds

General addition principle

Pr[*A* OR *B*] = Pr[*A*] + Pr[*B*] - Pr[*A* AND *B*].

If two events *A* and *B* are mutually exclusive, then Pr[A AND B] = 0, therefore:

Pr[*A* OR *B*] = Pr[*A*] + Pr[*B*]

Independence

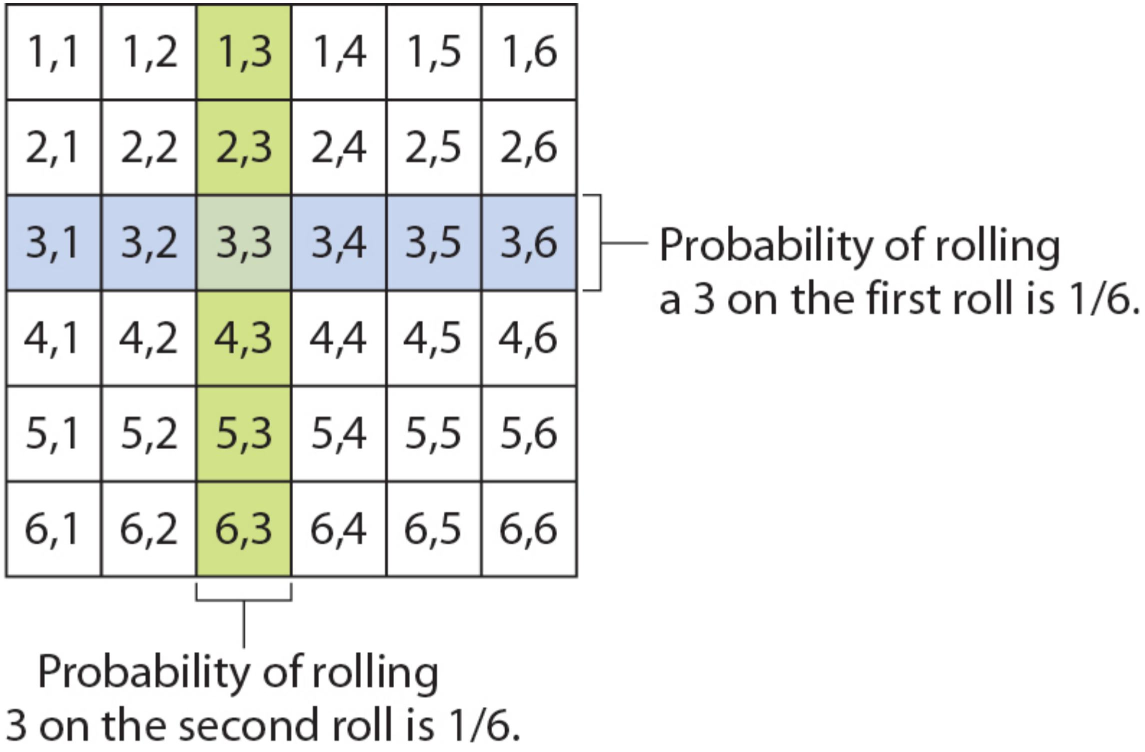
Two events are *independent* if the occurrence of one gives no information about whether the second will occur.

The multiplication principle

The *multiplication principle*: If two events A and B are independent, then

$$\Pr[A \text{ AND } B] = \Pr[A] \times \Pr[B]$$

The multiplication principle



1,1	1,2	1,3	1,4	1,5	1,6
2,1	2,2	2,3	2,4	2,5	2,6
3,1	3,2	3,3	3,4	3,5	3,6
4,1	4,2	4,3	4,4	4,5	4,6
5,1	5,2	5,3	5,4	5,5	5,6
6,1	6,2	6,3	6,4	6,5	6,6

Probability of rolling
a 3 on the first roll is 1/6.

Probability of rolling
a 3 on the second roll is 1/6.

P[1st and 2nd roll is 3] = 1/6 x 1/6 = 1/36

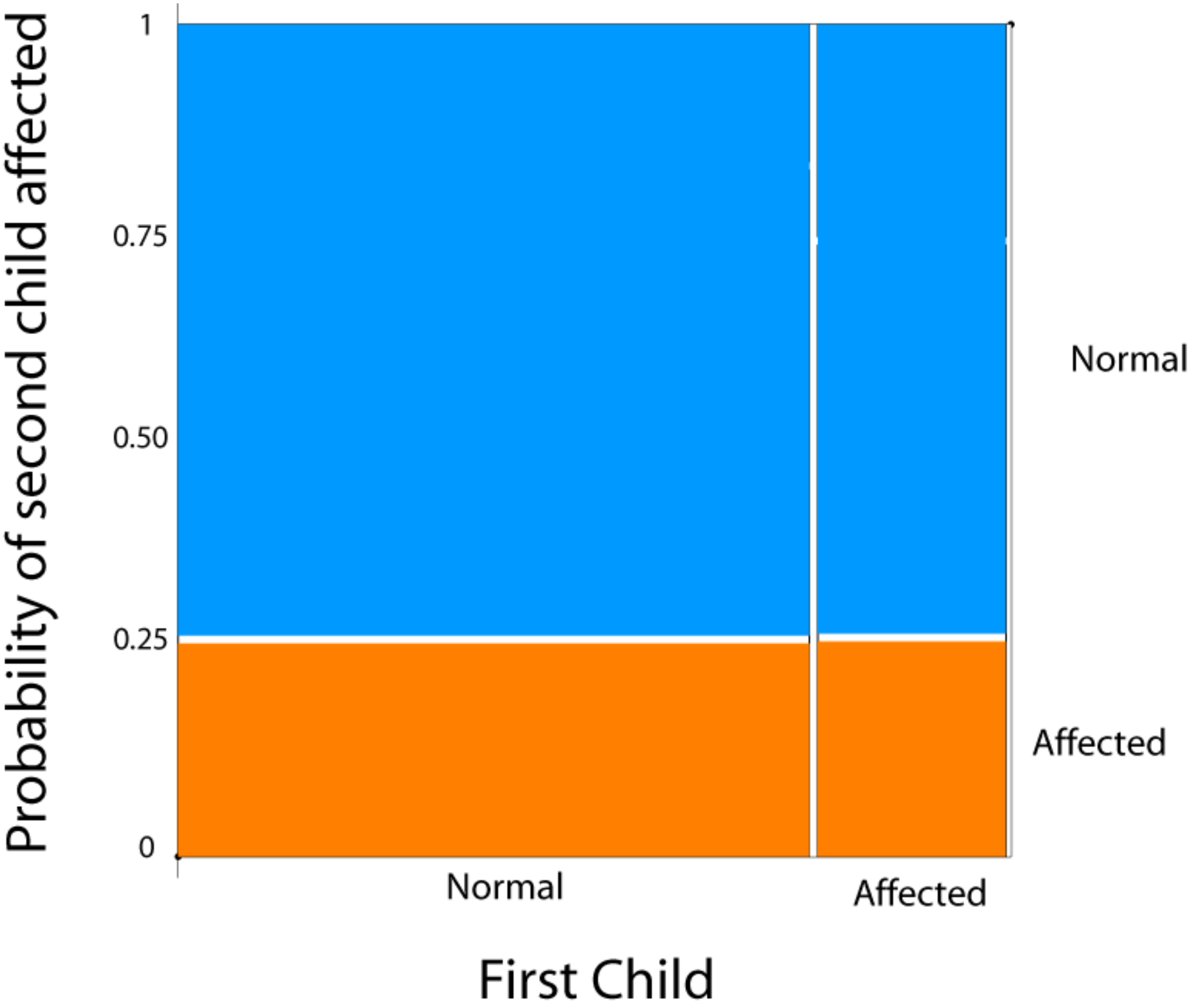
Offspring of two "carriers":

Pr[congenital nightblindness]=0.25

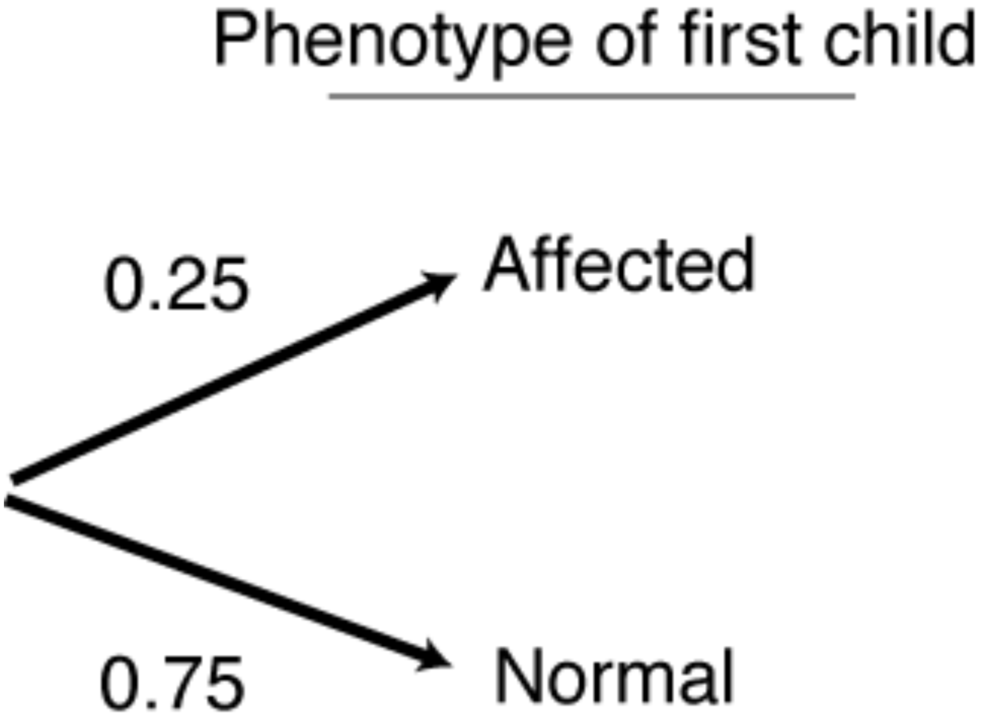
What is the probability that two kids from this family both have nightblindedness?

Pr[(*first child has nightblindness*) AND

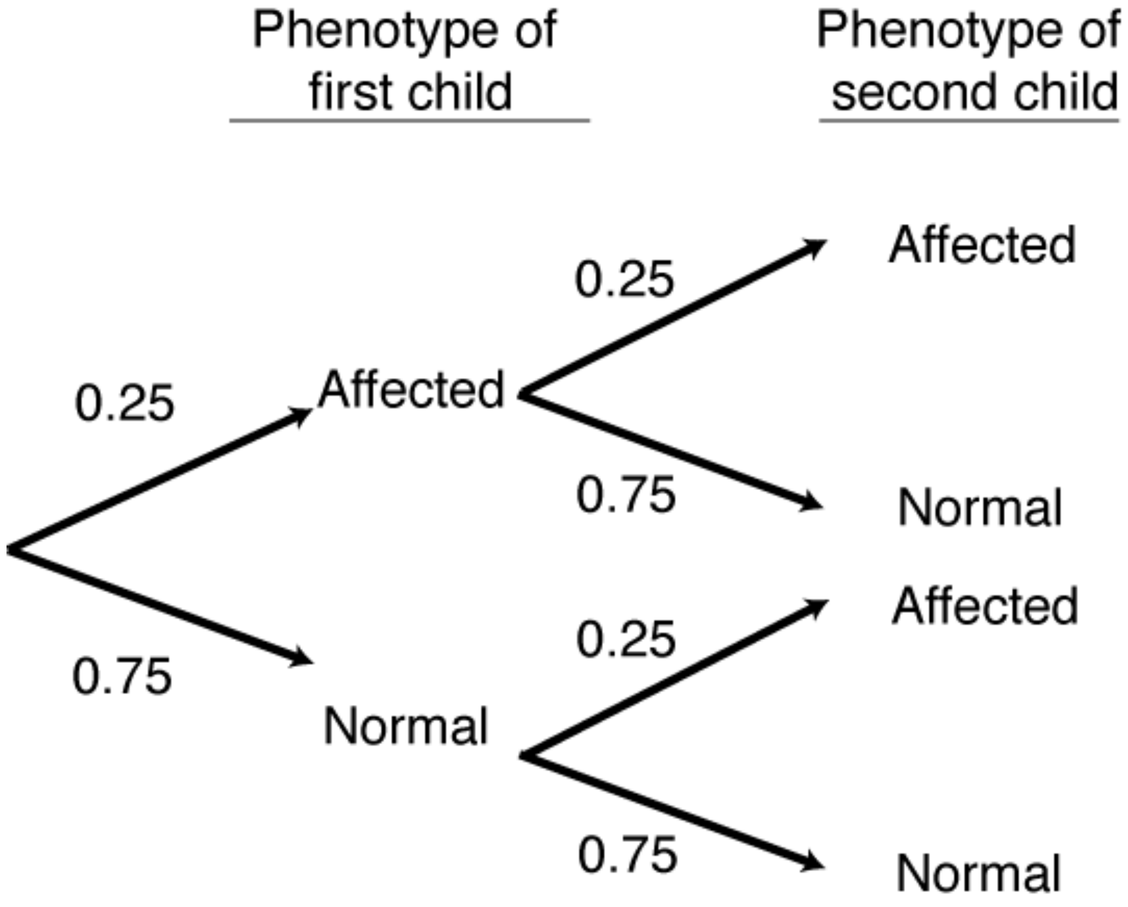
(*second child has nightblindness*)] = 0.25 × 0.25 = 0.0625.


Probability of second child affected
1
0.75
0.50
0.25
0
Normal
Affected
Normal
Affected
First Child

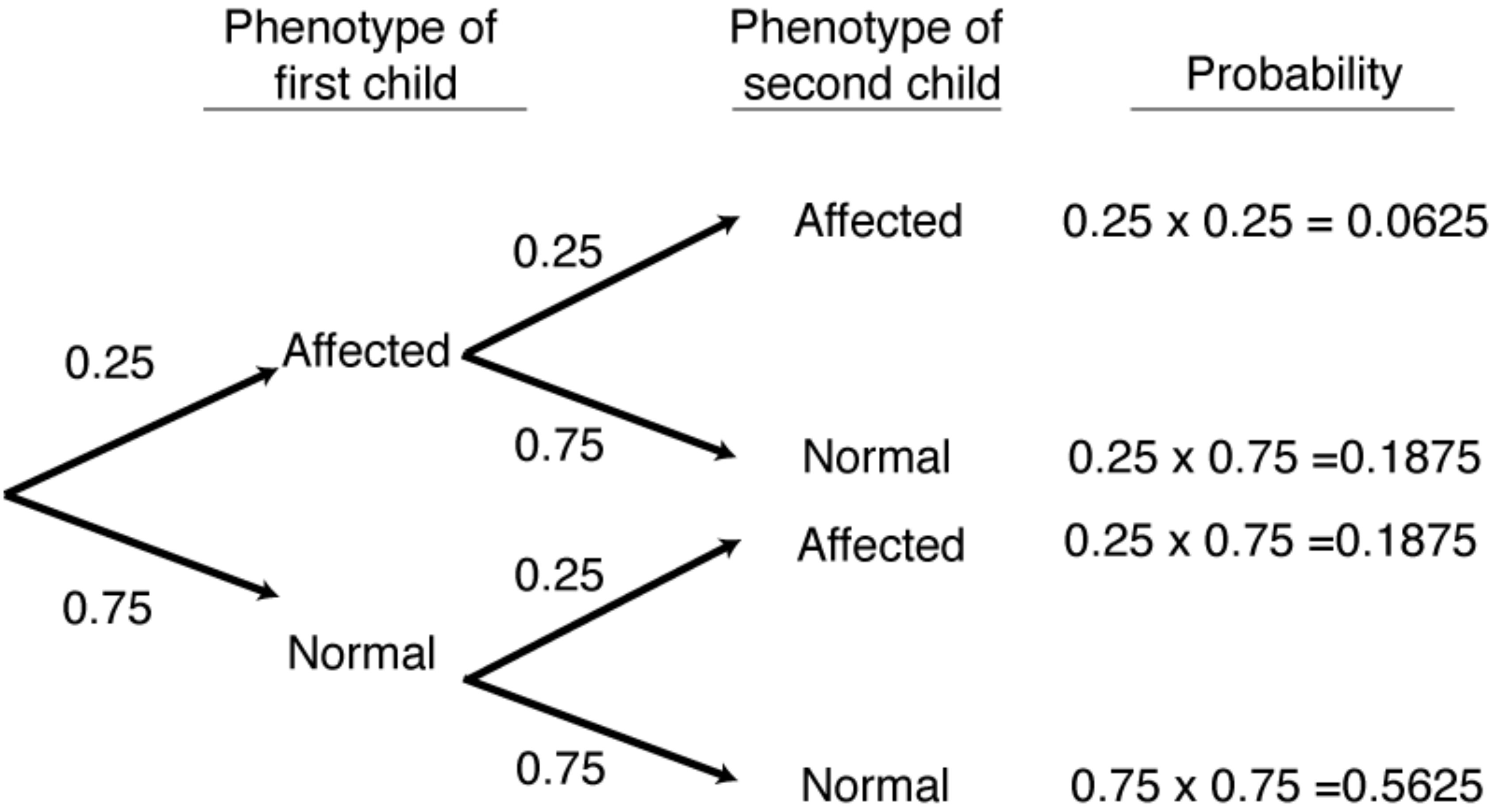
Probability trees



Phenotypes in two-child family



Phenotypes in two-child family



Phenotype of first child

Phenotype of second child

Probability

0.25

Affected

0.25

Affected

0.25 x 0.25 = 0.0625

0.75

Normal

0.25 x 0.75 =0.1875

0.75

Normal

0.25

Affected

0.25 x 0.75 =0.1875

0.75

Normal

0.75 x 0.75 =0.5625

Short summary

The probability of A *OR* B involves addition.

$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ if the two are mutually exclusive.

The probability of A *AND* B involves multiplication

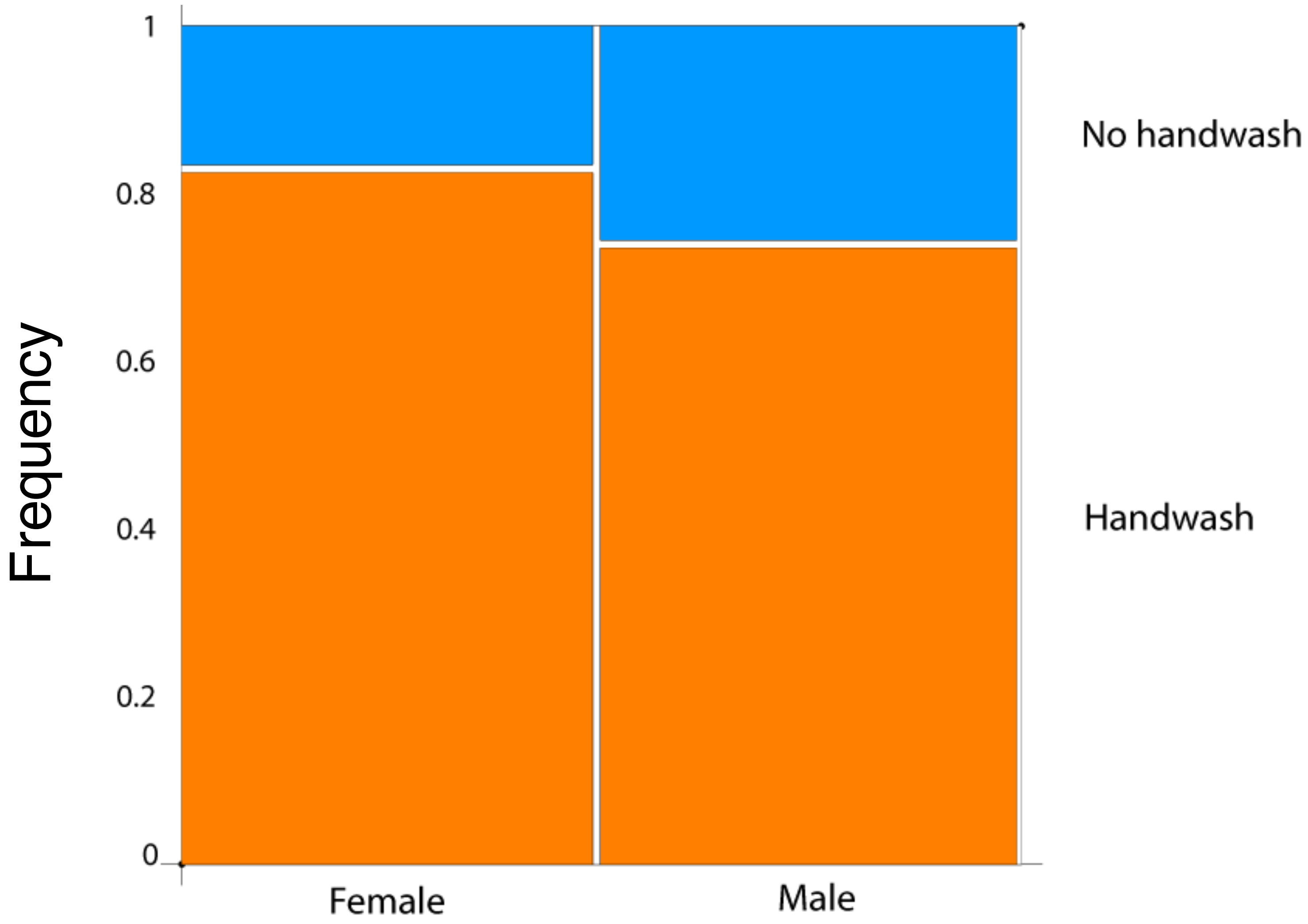
$\Pr(A \text{ and } B) = \Pr(A)\,\Pr(B)$ if the two are independent

Dependent events

Variables are not always independent.

The probability of one event may depend on the outcome of another event

Washing hands


Frequency
1
0.8
0.6
0.4
0.2
0
Female
Male
No handwash
Handwash

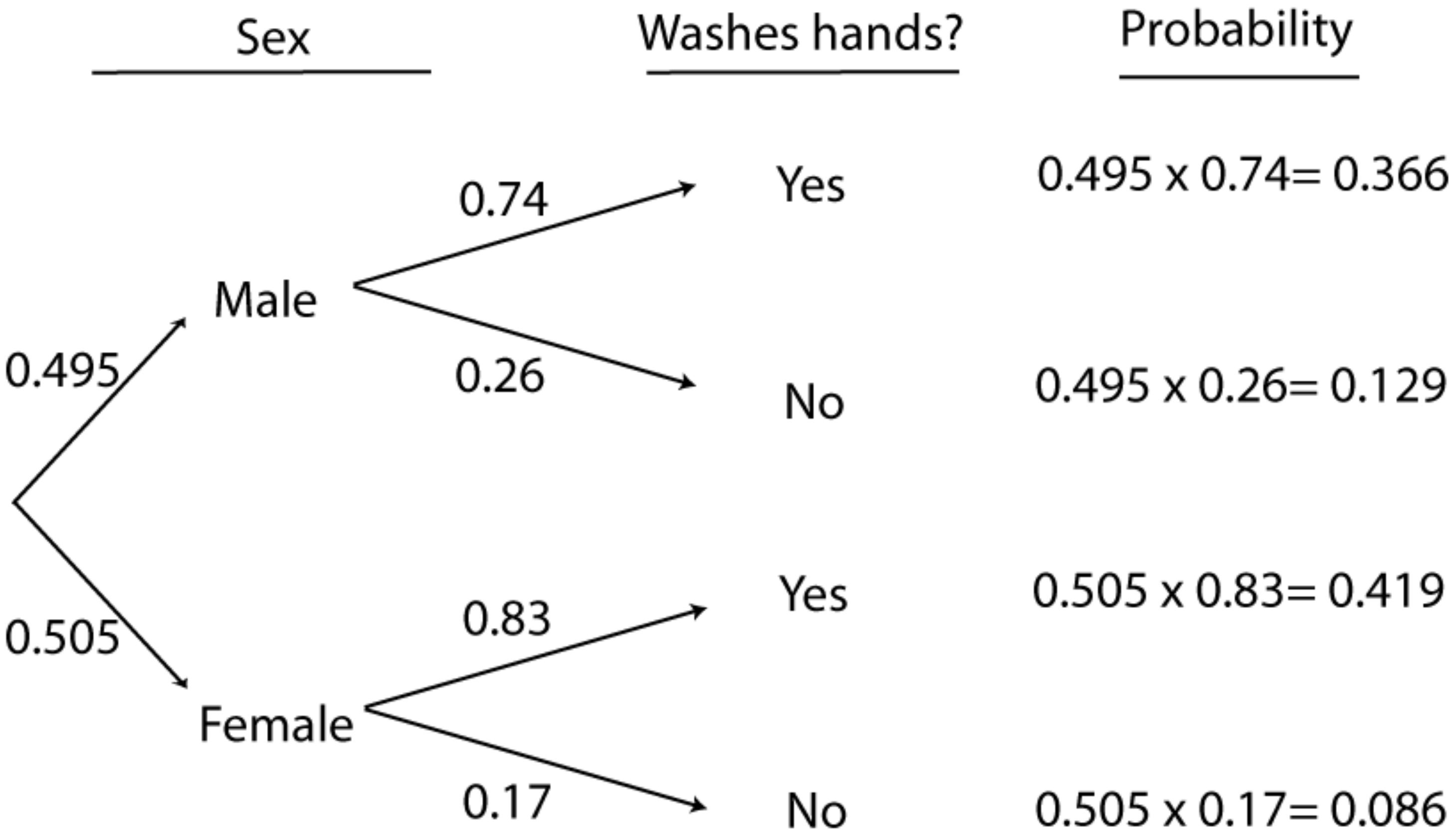
Hand washing after using the restroom

- Pr[male] = 0.495

- Pr[male washes his hands] = 0.74

- Pr[female washes her hands] = 0.83

Hand washing



Sex

Washes hands?

Probability

0.495

Male

0.74

Yes

0.495 x 0.74= 0.366

0.26

No

0.495 x 0.26= 0.129

0.505

Female

0.83

Yes

0.505 x 0.83= 0.419

0.17

No

0.505 x 0.17= 0.086

Conditional probability

The conditional probability of an event is the probability of that event occurring *given that* a condition is met.

Pr[AlB]

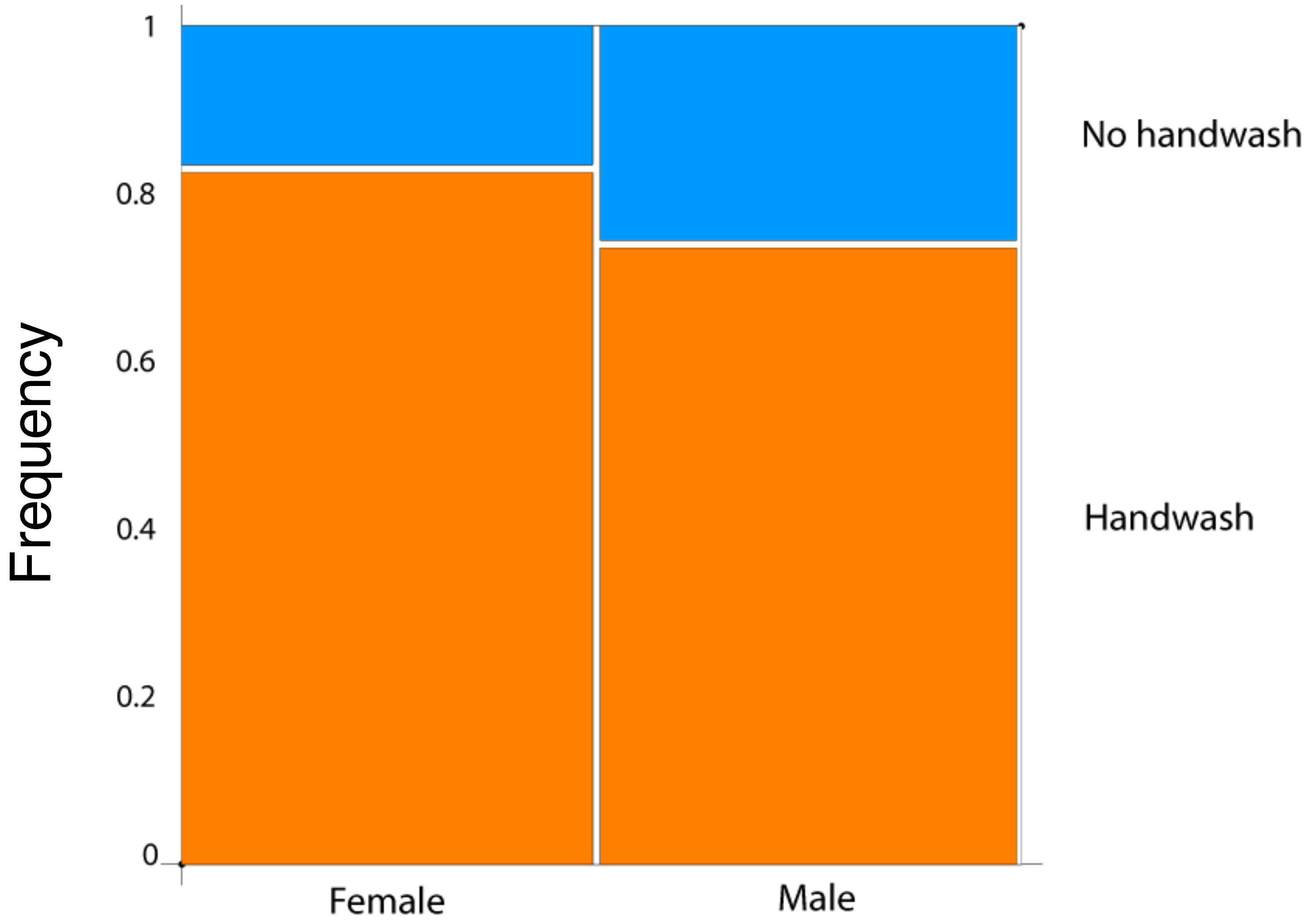
Pr(A | B) means the probability of A if B is true.

It is read as "the probability of A given B."

Pr(hand washing | male) = 0.74.

Law of total probability

$$\Pr[A] = \sum_{All\ values\ of\ B} \Pr[A \mid B]\Pr[B]$$


Frequency
1
0.8
0.6
0.4
0.2
0
Female
Male
No handwash
Handwash

The probability of hand washing is

$$\Pr[\text{hand washing}] =$$

$$\Pr(\text{hand washing} \mid \text{male}) \Pr(\text{male}) +$$

$$\Pr(\text{hand washing} \mid \text{female}) \Pr(\text{female})$$

$$= 0.74\ (0.495) + 0.83\ (0.505) = 0.785$$

The general multiplication rule

$$\Pr[A \text{ AND } B] = \Pr[A] \Pr[B \mid A]$$

The general multiplication rule

$$\Pr[A \text{ AND } B] = \Pr[A]\,\Pr[B \mid A]$$

If two events *A* and *B* are independent, then Pr[B|A] = Pr[B], therefore:

$$\Pr[A \text{ AND } B] = \Pr[A] \times \Pr[B]$$

The general multiplication rule

$\Pr[A \text{ AND } B] = \Pr[A] \Pr[B \mid A]$

$\Pr[A \text{ AND } B] = \Pr[B] \Pr[A \mid B]$

Therefore

$\Pr[B] \Pr[A \mid B] = \Pr[A] \Pr[B \mid A]$

Bayes' theorem

$$\Pr\left[A \mid B\right] = \frac{\Pr\left[B \mid A\right]\Pr[A]}{\Pr[B]}$$