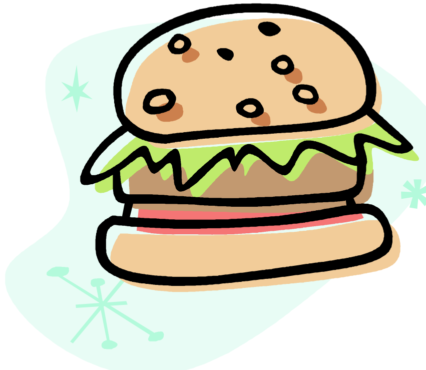
Unit 6

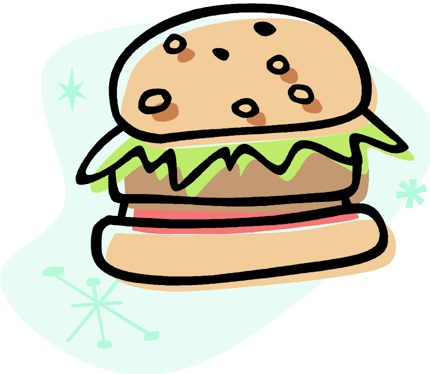
Sampling Variability and Sampling Distributions

Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a national fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.



Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a national fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

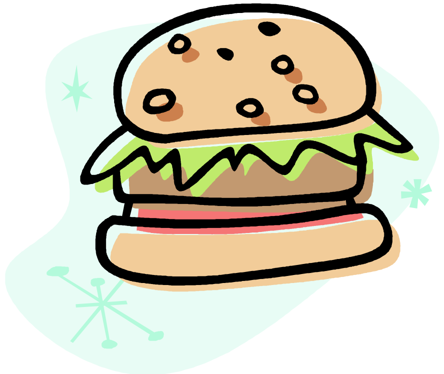
Sample mean good estimate of μ?



Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a national fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

Sample mean good estimate of μ?

How close is the sample mean to μ?

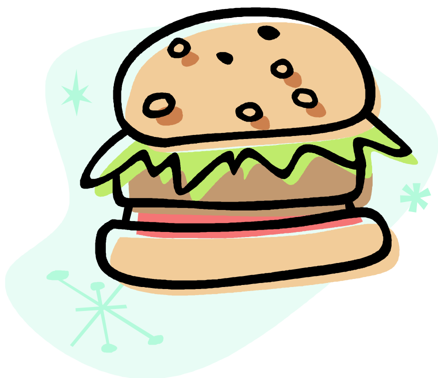


Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a national fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

Sample mean good estimate of μ?

How close is the sample mean to μ?

Sample mean of other samples of n = 50?



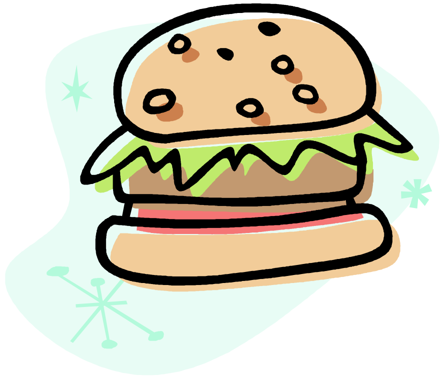
Suppose we are interested in finding the true mean (μ) fat content of quarter-pound hamburgers marketed by a national fast food chain. To learn something about μ, we could obtain a sample of n = 50 hamburgers and determine the fat content of each one.

Sample mean good estimate of μ?

How close is the sample mean to μ?

Sample mean of other samples of n = 50?

Examine **<u>sampling distribution</u>** – long-run behavior of a sample statistic



$\bar{x}$ – sample mean

s – sample standard deviation

$\hat{p}$ – sample proportion

- observed value depends on particular sample selected from population
  - will vary from sample to sample

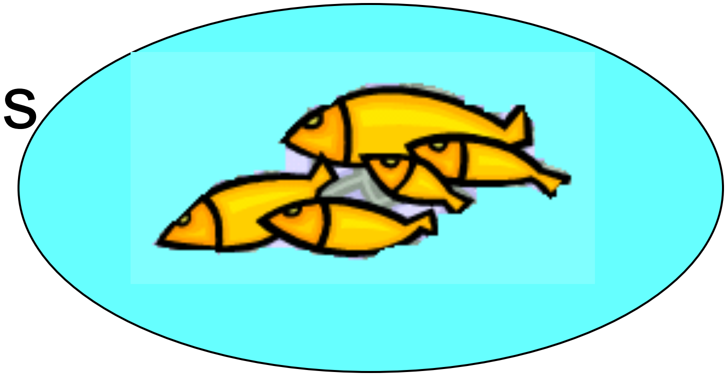
$\bar{x}$ – sample mean

s – standard deviation

$\hat{p}$ – sample proportion

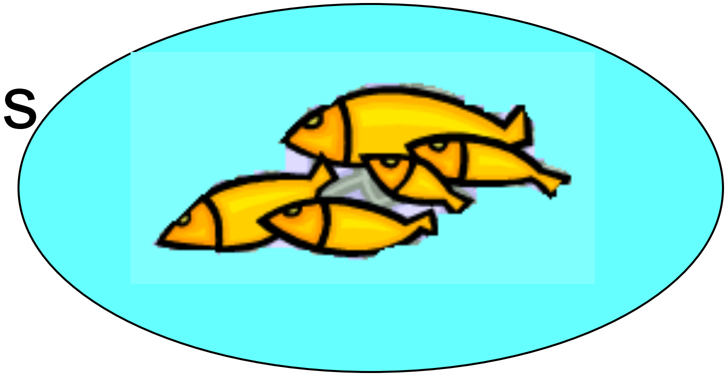
- observed value depends on particular sample selected from population
  - will vary from sample to sample
    - variability called **sampling variability**

Suppose there are 20 fish in the "Newman fish pond." The lengths of the fish (in inches) are given below:



4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

Suppose there are 20 fish in the "Newman fish pond." The lengths of the fish (in inches) are given below:

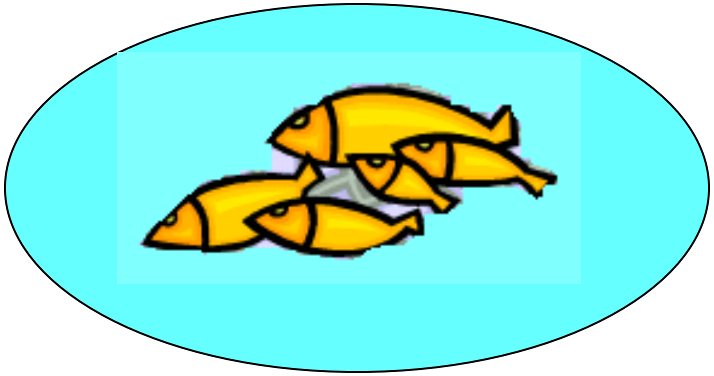


4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.

$\bar{x}$ = 7.27 inches

Newman has a fish pond. Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below:

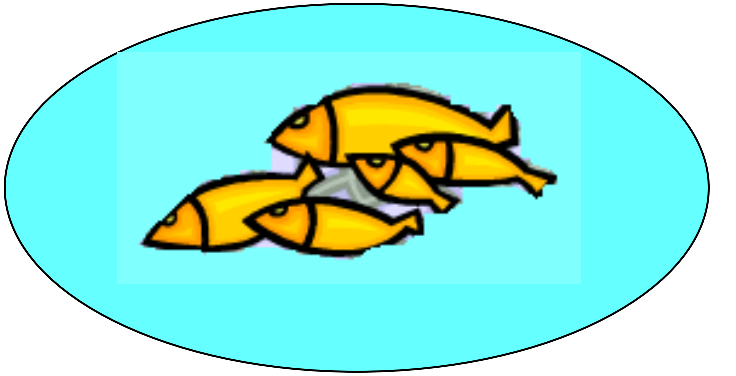


4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.
$\bar{x}$ = 7.27 inches

Let's catch two more samples and look at the sample means.

Newman has a fish pond. Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below:



4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.

$\bar{x}$ = 7.27 inches

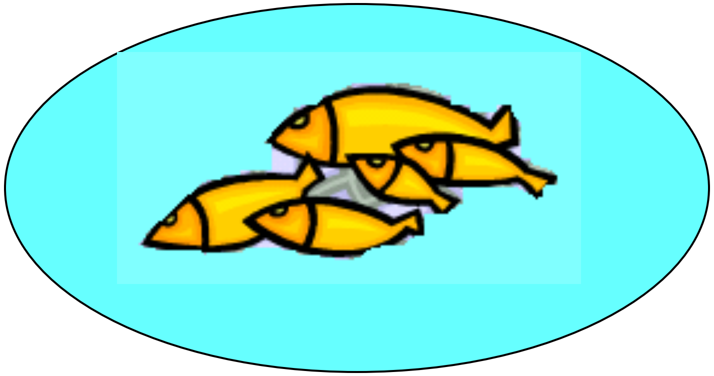
2nd sample - 8.5, 4.6, and 5.6 inches.

$\bar{x}$ = 6.23 inches

3rd sample – 10.3, 8.9, and 13.4 inches.

$\bar{x}$ = 10.87 inches

Newman has a fish pond. Suppose there are 20 fish in the pond. The lengths of the fish (in inches) are given below:



4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

True population mean $\mu = 8$

We caught fish with lengths 6.3 inches, 2.2 inches, and 13.3 inches.

$\bar{x} = 7.27$ inches

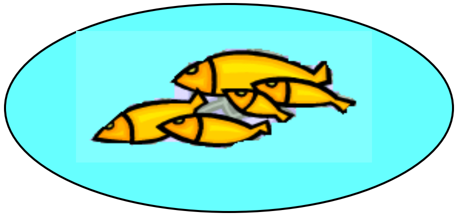
2nd sample - 8.5, 4.6, and 5.6 inches.

$\bar{x} = 6.23$ inches

3rd sample – 10.3, 8.9, and 13.4 inches.

$\bar{x} = 10.87$ inches

Fish Pond Continued . . .

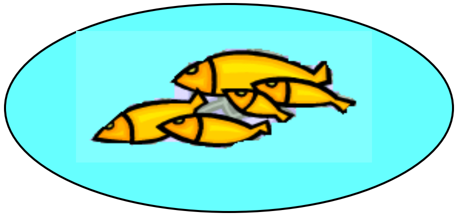


4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

1140 ($_{20}C_3$) different possible samples of size 3 from population.

all those samples, calculate mean length of each sample = distribution of all possible $\bar{x}$

Fish Pond Continued . . .



4.5	5.4	10.3	7.9	8.5	6.6	11.7	8.9	2.2	9.8
6.3	4.3	9.6	8.7	13.3	4.6	10.7	13.4	7.7	5.6

1140 (${}_{20}C_3$) different possible samples of size 3 from population.

all those samples, calculate mean length of each sample = distribution of all possible $\bar{x}$

<u>Sampling distribution</u> of $\bar{x}$

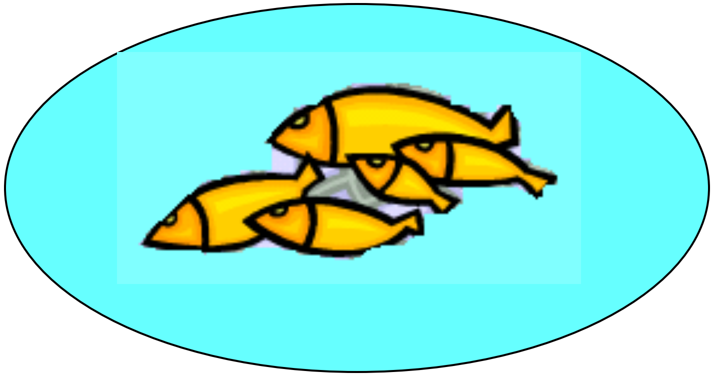
Sampling Distributions

- Distribution considering value for every possible sample of given size

- Could be $\bar{x}$
- Could be s
- Could be the IQR
- Could be any statistic calculated from a given sample

Fish Pond Revisited . . .

Suppose there are only 5 fish in the pond. The lengths of the fish (in inches) are given below:

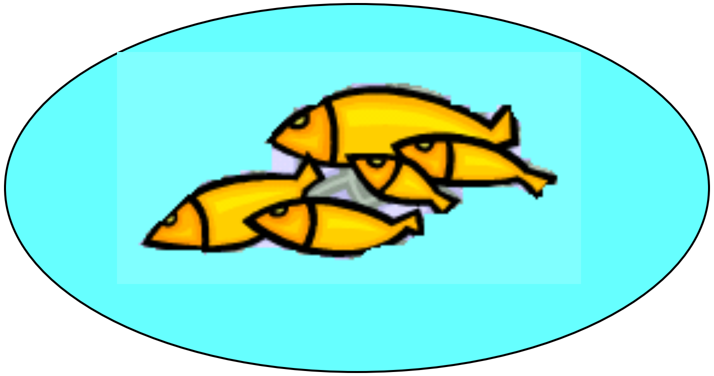


6.6 11.7 8.9 2.2 9.8

$$\mu_x = 7.84$$

$$\sigma_x = 3.262$$

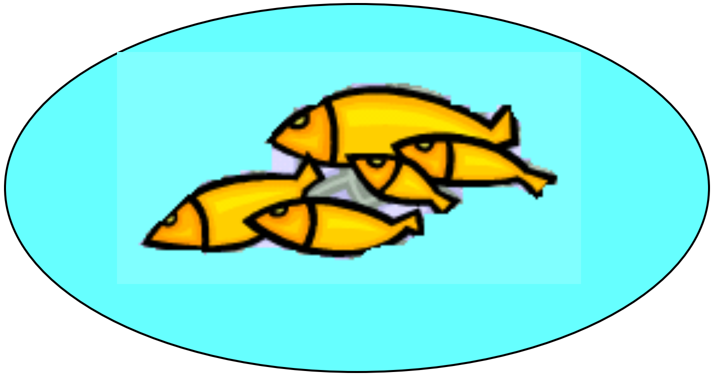
Fish Pond Revisited . . .



6.6 11.7 8.9 2.2 9.8

$\mu_x = 7.84$ and $\sigma_x = 3.262$

Fish Pond Revisited . . .

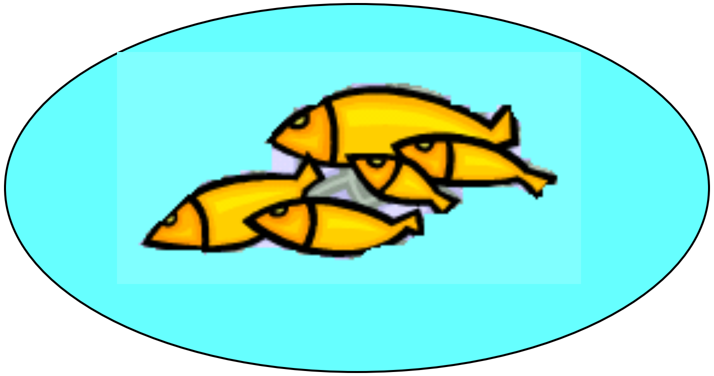


6.6 11.7 8.9 2.2 9.8

$\mu_x = 7.84$ and $\sigma_x = 3.262$

Pairs	6.6 & 11.7	6.6 & 8.9	6.6 & 2.2	6.6 & 9.8	11.7 & 8.9	11.7 & 2.2	11.7 & 9.8	8.9 & 2.2	8.9 & 9.8	2.2 & 9.8
$\bar{x}$	9.15	7.75	4.4	8.2	10.3	6.95	10.75	5.55	9.35	6

Fish Pond Revisited . . .



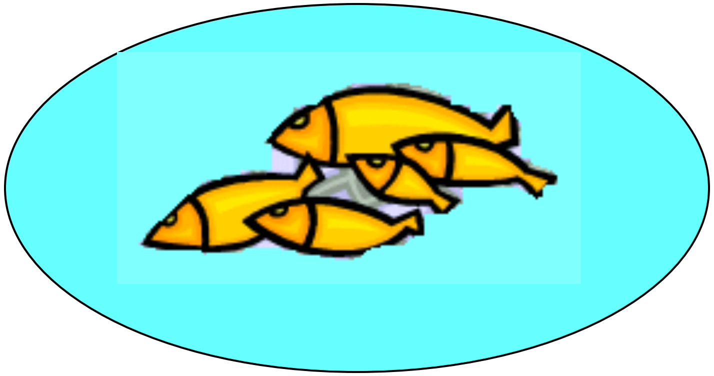
6.6   11.7   8.9   2.2   9.8

$\mu_x = 7.84$ and $\sigma_x = 3.262$

Pairs	6.6 & 11.7	6.6 & 8.9	6.6 & 2.2	6.6 & 9.8	11.7 & 8.9	11.7 & 2.2	11.7 & 9.8	8.9 & 2.2	8.9 & 9.8	2.2 & 9.8
$\bar{x}$	9.15	7.75	4.4	8.2	10.3	6.95	10.75	5.55	9.35	6

sampling distribution of $\bar{x}$ for samples of size 2

Fish Pond Revisited . . .



6.6   11.7   8.9   2.2   9.8

$\mu_x = 7.84$ and $\sigma_x = 3.262$

Pairs	6.6 & 11.7	6.6 & 8.9	6.6 & 2.2	6.6 & 9.8	11.7 & 8.9	11.7 & 2.2	11.7 & 9.8	8.9 & 2.2	8.9 & 9.8	2.2 & 9.8
$\bar{x}$	9.15	7.75	4.4	8.2	10.3	6.95	10.75	5.55	9.35	6

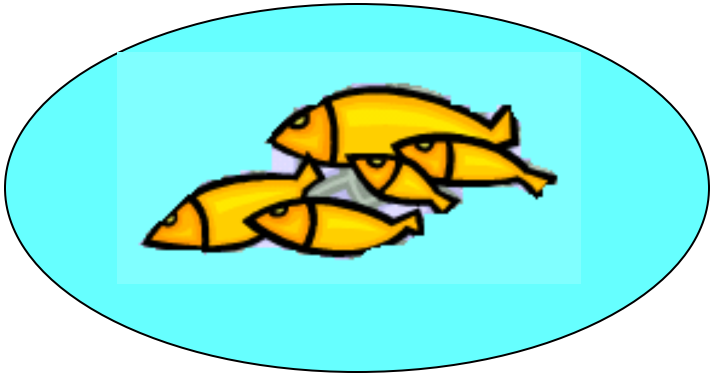
$$\mu_{\bar{x}} = 7.84$$

$$\sigma_{\bar{x}} = 1.998$$

Fish Pond Revisited . . .

6.6   11.7   8.9   2.2   9.8

$\mu_x = 7.84$ and $\sigma_x = 3.262$



Triples	6.6, 11.7, 8.9	6.6, 11.7, 2.2	6.6, 11.7, 9.8	6.6, 8.9, 2.2	6.6, 8.9, 9.8	6.6, 2.2, 9.8	11.7, 8.9, 2.2	11.7, 8.9, 9.8	11.7, 2.2, 9.8	8.9, 2.2, 9.8
$\bar{x}$	9.067	6.833	9.367	5.9	8.433	6.2	7.6	10.133	7.9	6.967

$$\mu_{\bar{x}} = 7.84$$

$$\sigma_{\bar{x}} = 1.332$$

- mean of sampling distribution EQUALS mean of population

$$\mu_{\bar{x}} = \mu$$

- mean of sampling distribution EQUALS mean of population

$$\mu_{\bar{x}} = \mu$$

- As n increases, standard deviation of sampling distribution decreases

$$\text{as } n \uparrow \quad \sigma_{\bar{x}} \downarrow$$

General Properties of Sampling Distributions of $\bar{x}$

$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

– Exact if population is infinite
– approximately correct if population is finite and no more than 10% of the population in sample

General Properties of Sampling Distributions of $\bar{x}$

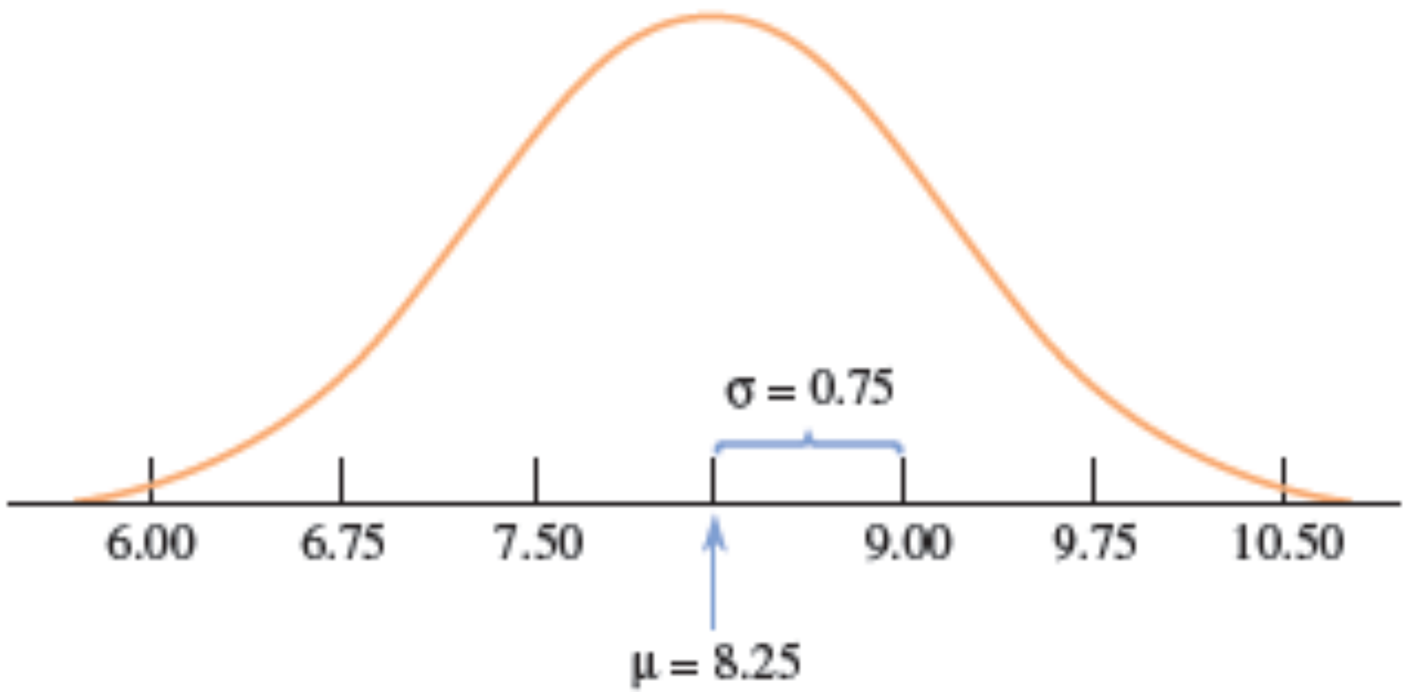
$$\mu_{\bar{x}} = \mu$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Note: fish pond examples standard deviation formula was incorrect because sample sizes > 10% of population

- Exact if population is infinite
- approximately correct if population is finite and no more than 10% of the population in sample

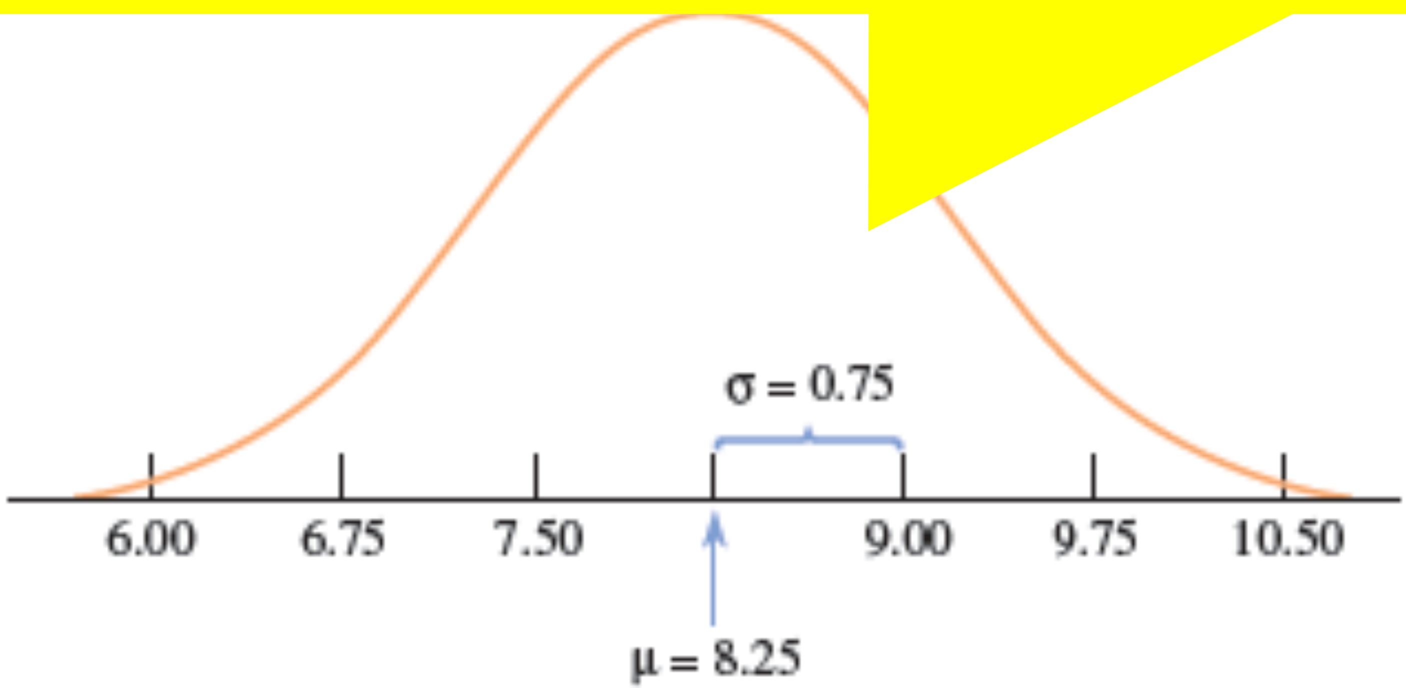
The paper "Mean Platelet Volume in Patients with Metabolic Syndrome and Its Relationship with Coronary Artery Disease" (Thrombosis Research, 2007) includes data that suggests that the distribution of platelet volume of patients who do not have metabolic syndrome is approximately normal with mean $\mu = 8.25$ and standard deviation $\sigma = 0.75$.



The paper “Mean Platelet Volume in Patients with Metabolic Syndrome and Its Relationship with Coronary Artery Disease” (Thrombosis Research, 2007) includes data that suggests that the distribution of platelet volume of patients

Generate 500 random samples of n = 5, n = 10, n = 20, n = 30 from population

Compute sample mean of each

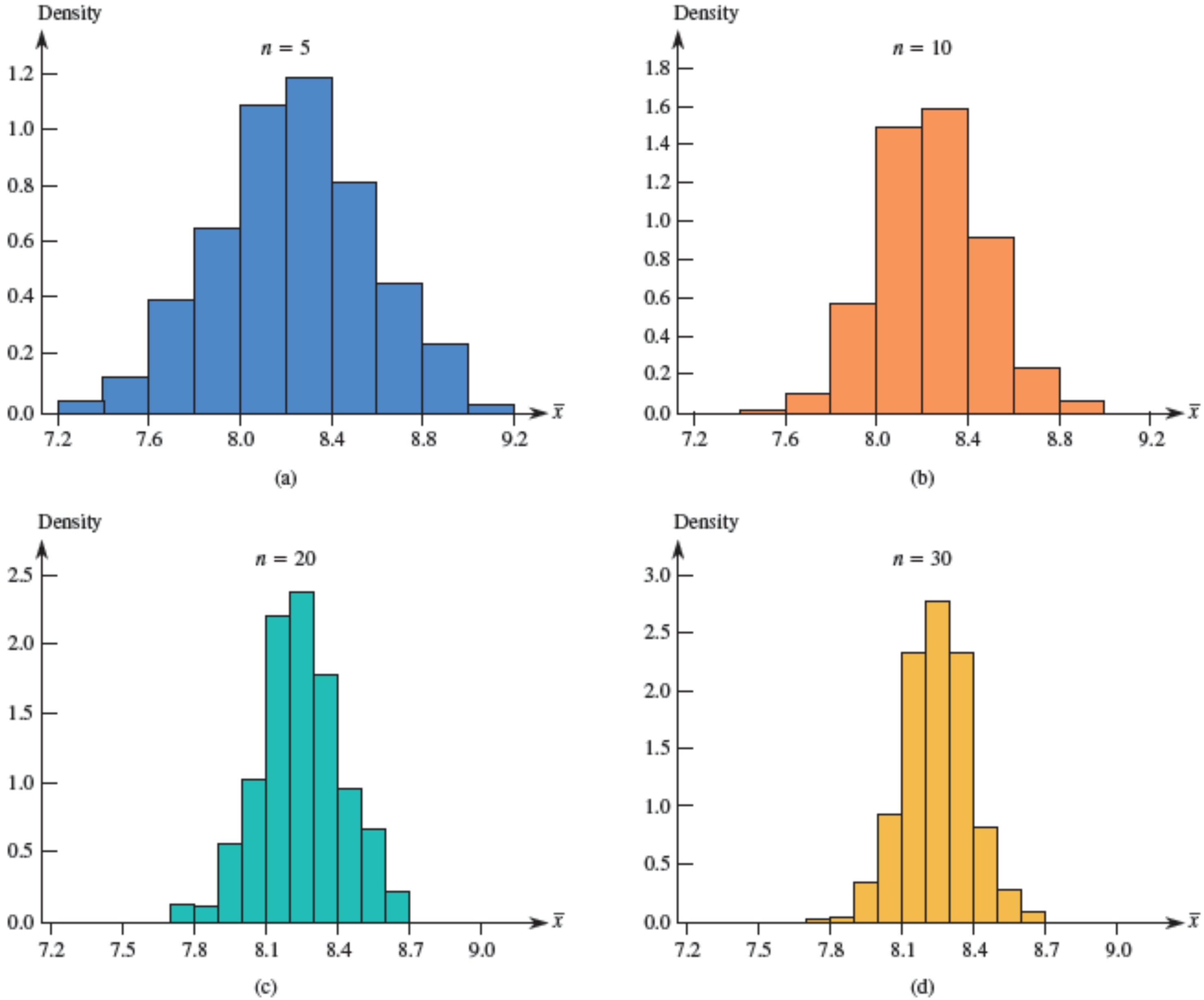


Platelets Continued . . .

500 random samples of $n = 10$, $n = 20$, and $n = 30$

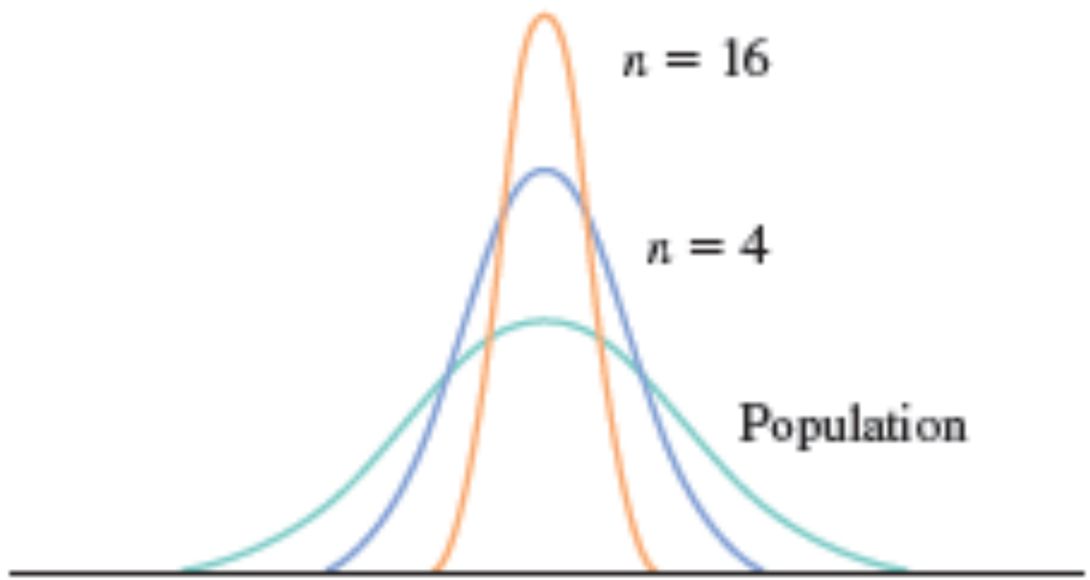
Density histograms display resulting 500 $\bar{x}$ for each given sample size

$\mu = 8.25, \sigma = 0.75$



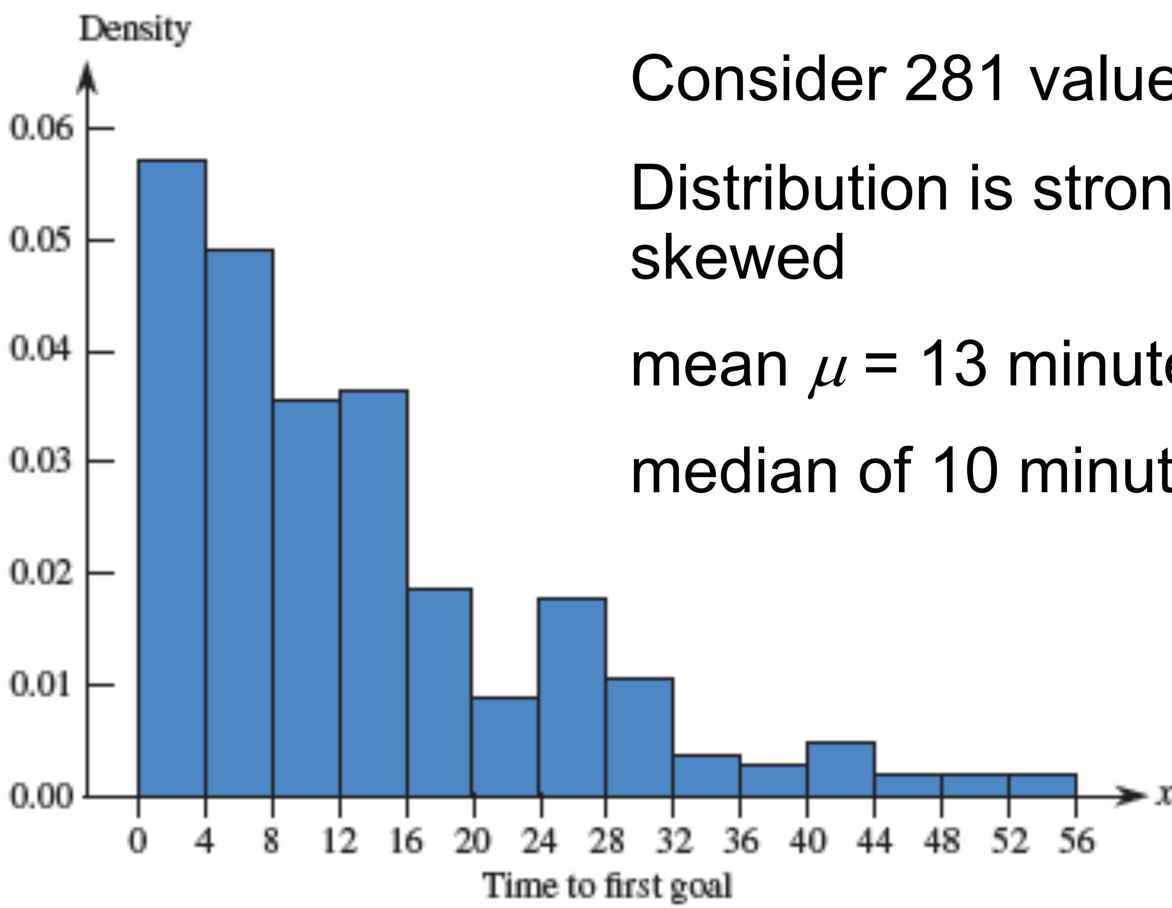
General Properties Continued . . .

When population distribution is normal, sampling distribution of $\bar{x}$ also normal for any sample size n



The paper “Is the Overtime Period in an NHL Game Long Enough?” (American Statistician, 2008) gave data on the time (in minutes) from the start of the game to the first goal scored for the 281 regular season games from the 2005-2006 season that went into overtime. The density histogram for the data is shown below.

The paper “Is the Overtime Period in an NHL Game Long Enough?” (American Statistician, 2008) gave data on the time (in minutes) from the start of the game to the first goal scored for the 281 regular season games from the 2005-2006 season that went into overtime. The density histogram for the data is shown below.



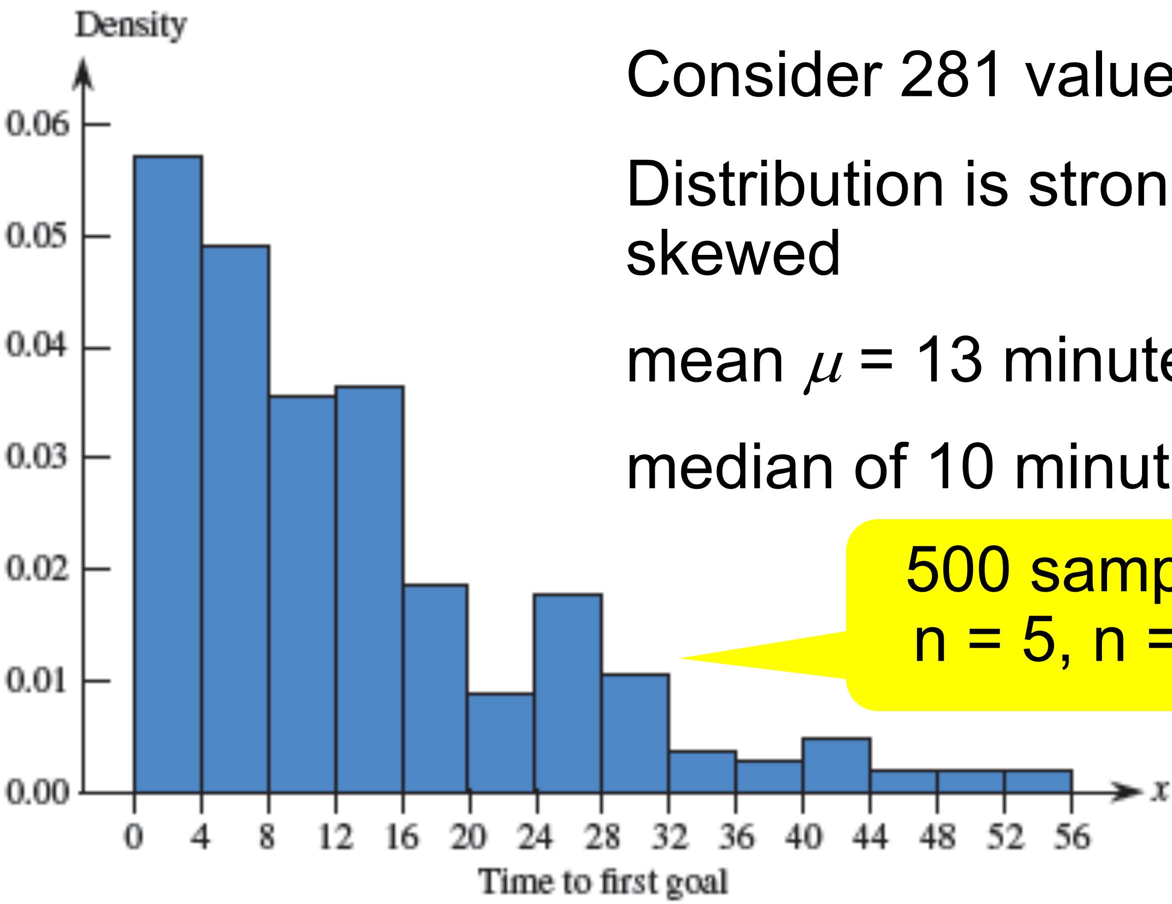
Consider 281 values as population

Distribution is strongly positively skewed

mean μ = 13 minutes

median of 10 minutes

The paper "Is the Overtime Period in an NHL Game Long Enough?" (American Statistician, 2008) gave data on the time (in minutes) from the start of the game to the first goal scored for the 281 regular season games from the 2005-2006 season that went into overtime. The density histogram for the data is shown below.



Consider 281 values as population.

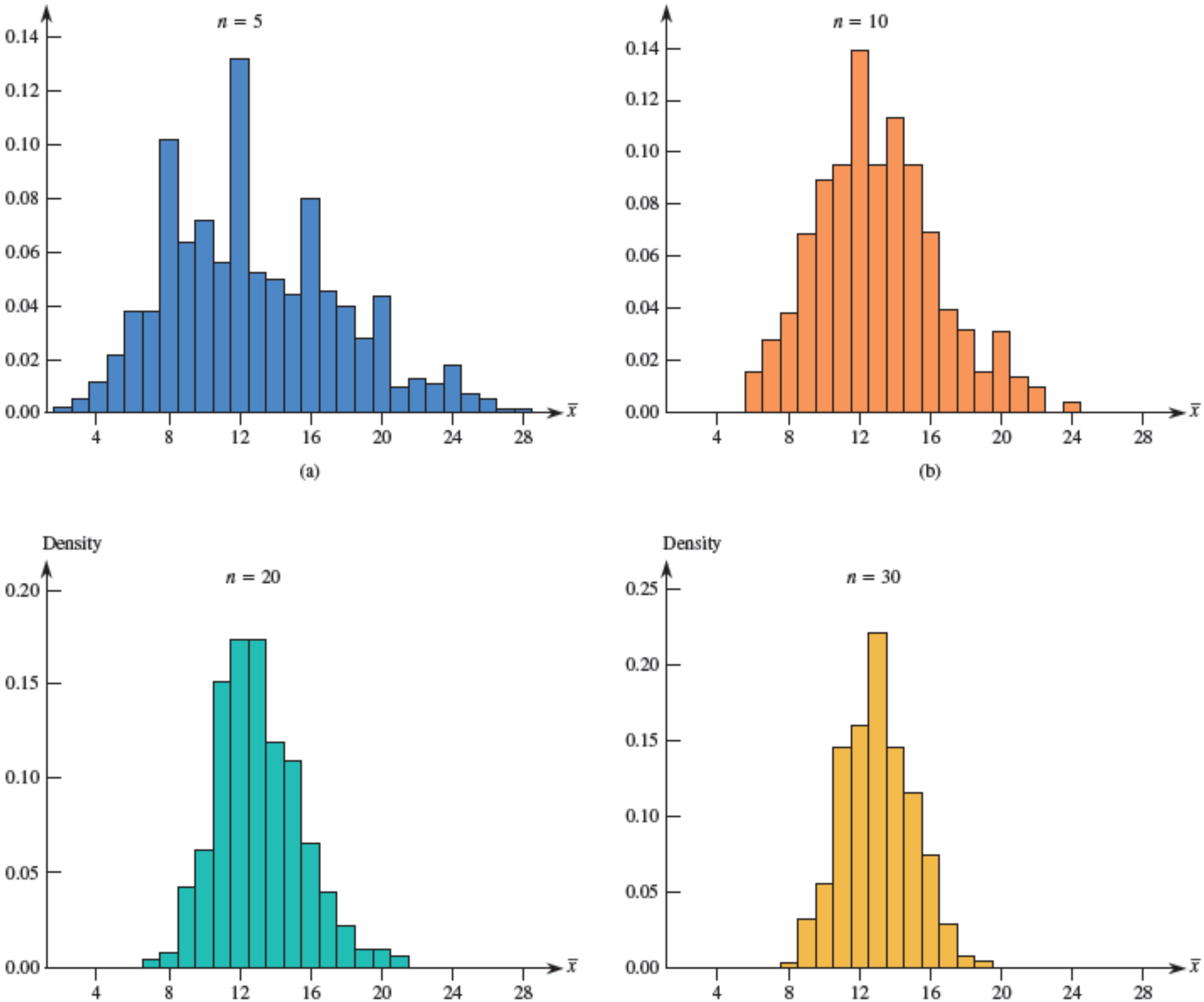
Distribution is strongly positively skewed

mean μ = 13 minutes

median of 10 minutes

500 samples of sample sizes n = 5, n = 10, n = 20, n = 30.

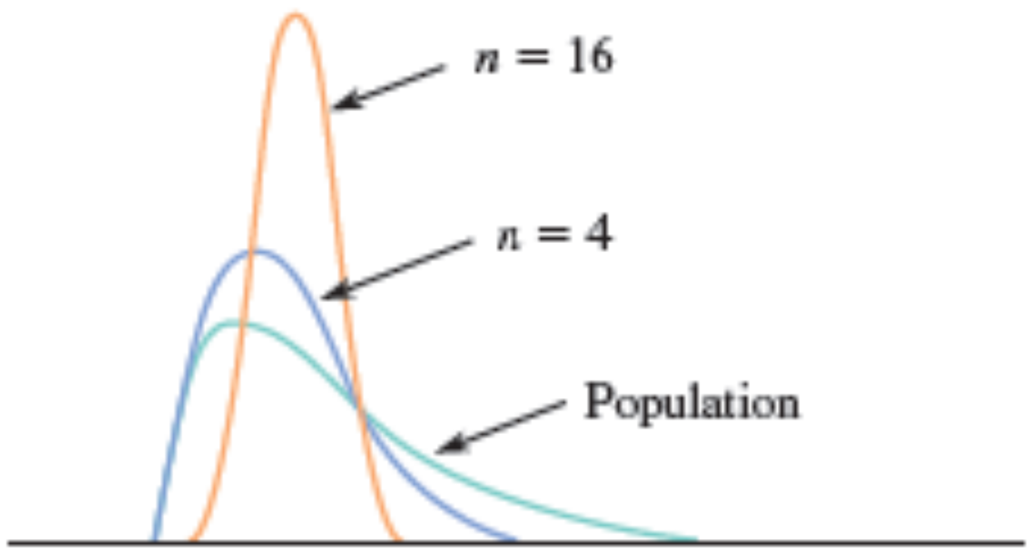
μ = 13 minutes



General Properties Continued . . .

Central Limit Theorem

When n sufficiently large, sampling distribution of $\bar{x}$ is well approximated by normal curve, even when population distribution is not normal



General Properties Continued . . .

Central Limit Theorem

When n sufficiently large, sampling distribution of $\bar{x}$ is well approximated by normal curve, even when population distribution is not normal

How large is "sufficiently large?"

General Properties Continued . . .

Central Limit Theorem

When n sufficiently large, sampling distribution of $\bar{x}$ is well approximated by normal curve, even when population distribution is not normal

CLT can safely be applied if $n \geq 30$.

A soft-drink bottler claims that, on average, cans contain 12 oz of soda. Let x denote the actual volume of soda in a randomly selected can. Suppose that x is normally distributed with σ = .16 oz. Sixteen cans are randomly selected, and the soda volume is determined for each one. Let $\overline{x}$ = the resulting sample mean soda.

If the bottler's claim is correct, then the sampling distribution of $\overline{x}$ is normally distributed with:

A soft-drink bottler claims that, on average, cans contain 12 oz of soda. Let x denote the actual volume of soda in a randomly selected can. Suppose that x is normally distributed with σ = .16 oz. Sixteen cans are randomly selected, and the soda volume is determined for each one. Let $\bar{x}$ = the resulting sample mean soda.

If the bottler's claim is correct, then the sampling distribution of $\bar{x}$ is normally distributed with:

$$\mu_{\bar{x}} = \mu = 12$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.16}{\sqrt{16}} = .04$$

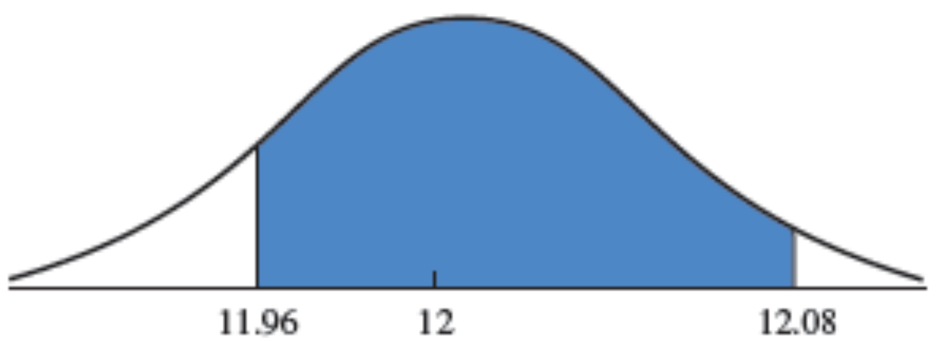
Soda Problem Continued . . .

$$\mu_{\bar{x}} = \mu = 12$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.16}{\sqrt{16}} = .04$$

What is the probability that the sample mean soda volume is between 11.96 ounces and 12.08 ounces?

$P(11.96 < \bar{x} < 12.08) =$



Soda Problem Continued . . .

$$\mu_{\bar{x}} = \mu = 12$$

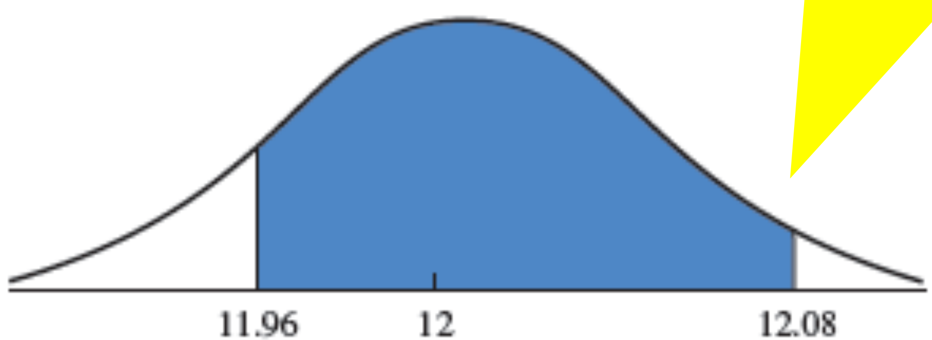
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.16}{\sqrt{16}} =$$

What is the prob
volume is betwe
ounces?

$P(11.96 < \bar{x} < 12.08) =$

To standardized these endpoints, use

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$



Soda Problem Continued . . .

$$\mu_{\bar{x}} = \mu = 12$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.16}{\sqrt{16}} =$$

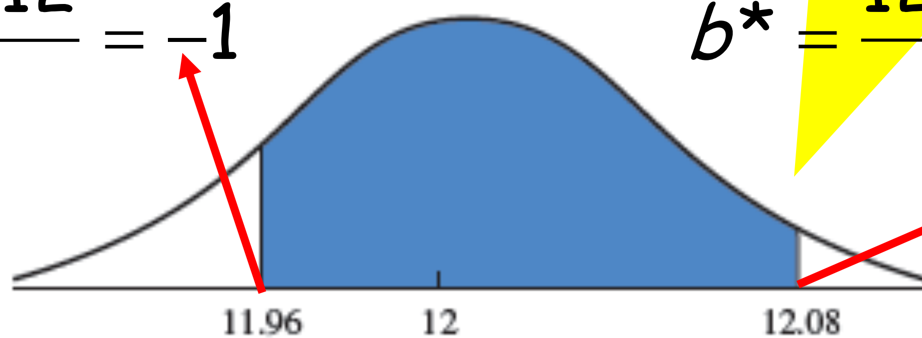
What is the prob
volume is betwe
ounces?

P(11.96 < x̄ < 12.08) =

To standardized these endpoints, use

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$a^* = \frac{11.96 - 12}{.04} = -1 \qquad b^* = \frac{12.08 - 12}{.04} = 2$$



Soda Problem Continued . . .

$$\mu_{\bar{x}} = \mu = 12$$

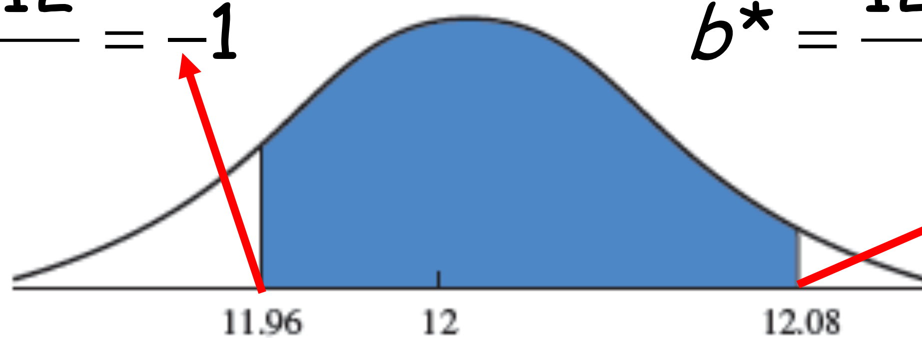
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{.16}{\sqrt{16}} = .04$$

What is the probability that the sample mean soda volume is between 11.96 ounces and 12.08 ounces?

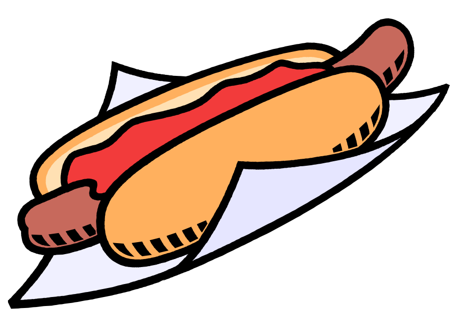
P(11.96 < $\bar{x}$ < 12.08) = .9772 - .1587 = .8185

$$a^* = \frac{11.96 - 12}{.04} = -1$$

$$b^* = \frac{12.08 - 12}{.04} = 2$$

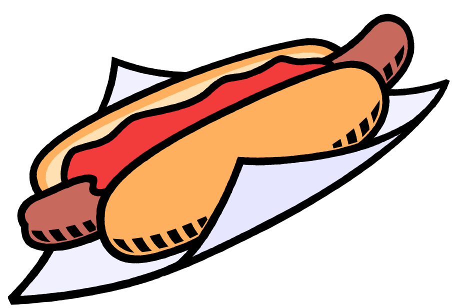


A hot dog manufacturer asserts that one of its brands of hot dogs has a average fat content of 18 grams per hot dog with standard deviation of 1 gram. Consumers of this brand would probably not be disturbed if the mean was less than 18 grams, but would be unhappy if it exceeded 18 grams.



A hot dog manufacturer asserts that one of its brands of hot dogs has a average fat content of 18 grams per hot dog with standard deviation of 1 gram. Consumers of this brand would probably not be disturbed if the mean was less than 18 grams, but would be unhappy if it exceeded 18 grams.

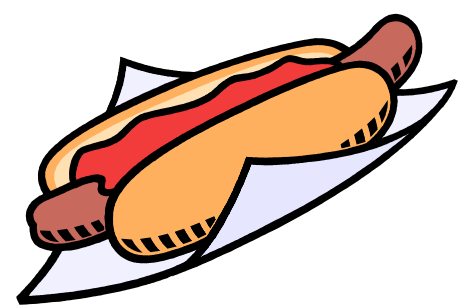
An independent testing organization is asked to analyze a random sample of 36 hot dogs. Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer’s claim is incorrect?



A hot dog manufacturer asserts that one of its brands of hot dogs has a average fat content of 18 grams per hot dog with standard deviation of 1 gram. Consumers of this brand would probably not be disturbed if the mean was less than 18 grams, but would be unhappy if it exceeded 18 grams.

An independent testing organization is asked to analyze a random sample of 36 hot dogs. Suppose the resulting sa s
result suggest th
incorrect?

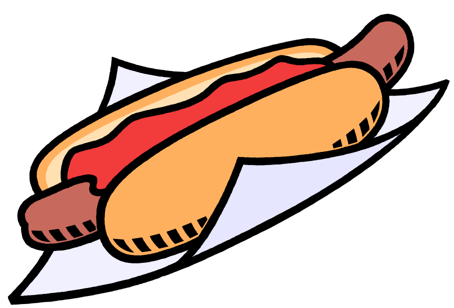
sample size greater than 30,
Central Limit Theorem applies



A hot dog manufacturer asserts that one of its brands of hot dogs has a average fat content of 18 grams per hot dog with standard deviation of 1 gram. Consumers of this brand would probably not be disturbed if the mean was less than 18 grams, but would be unhappy if it exceeded 18 grams.

An independent testing organization is asked to analyze a random sample of 36 hot dogs. Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer's claim is incorrect?

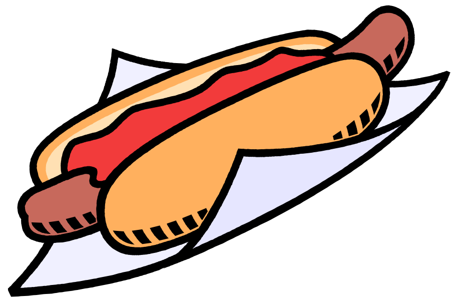
So the distribution of $\bar{x}$ is approximately normal with



A hot dog manufacturer asserts that one of its brands of hot dogs has a average fat content of 18 grams per hot dog with standard deviation of 1 gram. Consumers of this brand would probably not be disturbed if the mean was less than 18 grams, but would be unhappy if it exceeded 18 grams. An independent testing organization is asked to analyze a random sample of 36 hot dogs. Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer’s claim is incorrect?

So the distribution of $\bar{x}$ is approximately normal with

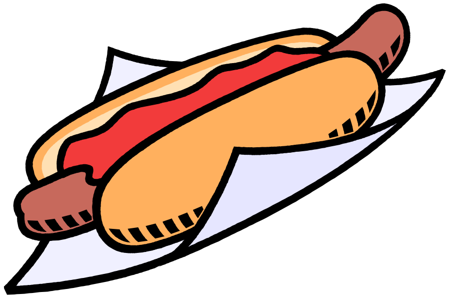
$$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = \frac{1}{\sqrt{36}} = .1667$$



Hot Dogs Continued . . .

$$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = \frac{1}{\sqrt{36}} = .1667$$

Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer's claim is incorrect?

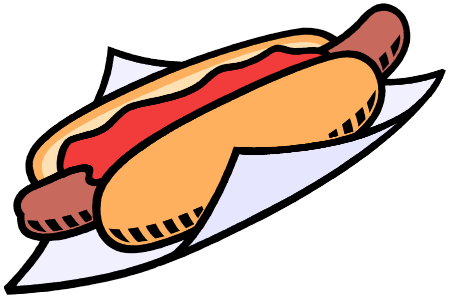


Hot Dogs Continued . . .

$$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = \frac{1}{\sqrt{36}} = .1667$$

Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer's claim is incorrect?

$P(\bar{x} \geq 18.4) =$



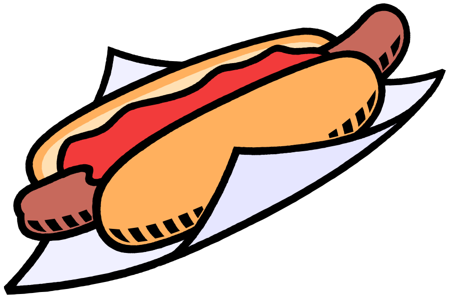
Hot Dogs Continued . . .

$$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = \frac{1}{\sqrt{36}} = .1667$$

Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer's claim is incorrect?

$P(\bar{x} \geq 18.4) =$

$$z = \frac{18.4 - 18}{.1667} = 2.40$$



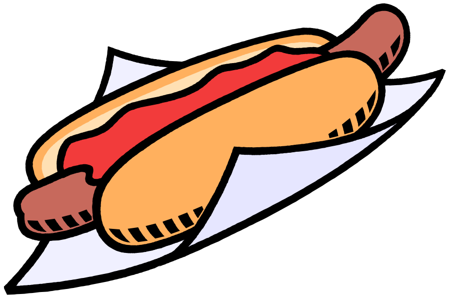
Hot Dogs Continued . . .

$$\mu_{\bar{x}} = 18 \text{ and } \sigma_{\bar{x}} = \frac{1}{\sqrt{36}} = .1667$$

Suppose the resulting sample mean is 18.4 grams. Does this result suggest that the manufacturer's claim is incorrect?

$$P(\bar{x} \geq 18.4) = 1 - .9918 = .0082$$

$$z = \frac{18.4 - 18}{.1667} = 2.40$$



A tire manufacturer designed a new tread pattern for its all weather tires. Repeated tests were conducted on cars of approximately the same weight traveling at 60 miles per hour. The tests showed that the new tread pattern enables the cars to stop completely in an average distance of 125 feet with a standard deviation of 6.5 feet and that the stopping distances are approximately normally distributed.

(a) What is the 70th percentile of the distribution of stopping distances?

(b) What is the probability that at least 2 cars out of 5 randomly selected cars in the study will stop in a distance that is greater than the distance calculated in part (a)?

(c) What is the probability that a randomly selected sample of 5 cars in the study will have a mean stopping distance of at least 130 feet?

Big Town Fisheries recently stocked a new lake in a city park with 2,000 fish of various sizes. The distribution of the lengths of these fish is approximately normal.

(a) Big Town Fisheries claims that the mean length of the fish is 8 inches. If the claim is true, which of the following would be more likely?

- A random sample of 15 fish having a mean length that is greater than 10 inches or
- A random sample of 50 fish having a mean length that is greater than 10 inches Justify your answer.

(b) Suppose the standard deviation of the sampling distribution of the sample mean for random samples of size 50 is 0.3 inch. If the mean length of the fish is 8 inches, use the normal distribution to compute the probability that a random sample of 50 fish will have a mean length less than 7.5 inches.

(c) Suppose the distribution of fish lengths in this lake was nonnormal but had the same mean and standard deviation. Would it still be appropriate to use the normal distribution to compute the probability in part (b) ? Justify your answer.

Sampling Distribution of $\hat{p}$

- p for proportion of successes in population

- p for population proportion and $\hat{p}$ for sample proportion

Sampling Distribution of $\hat{p}$

- Distribution formed by considering value of sample proportion for every possible different sample (of given size)

Sampling Distribution of $\hat{p}$

$$\hat{p} = \frac{\text{number of successes in the sample}}{n}$$

- Distribution formed by consi g value of sample statistic for every possible different sample (of given size)

Suppose we have a population of 6 students:
Alice, Ben, Charles, Denise, Edward, & Frank

Suppose we have a population of 6 students:
Alice, Ben, Charles, Denise, Edward, & Frank

Interested in proportion of girls

Suppose we have a population of 6 students:
Alice, Ben, Charles, Denise, Edward, & Frank

Interested in proportion of girls

Proportion of girls – 1/3

Suppose we have a population of 6 students:
Alice, Ben, Charles, Denise, Edward, & Frank

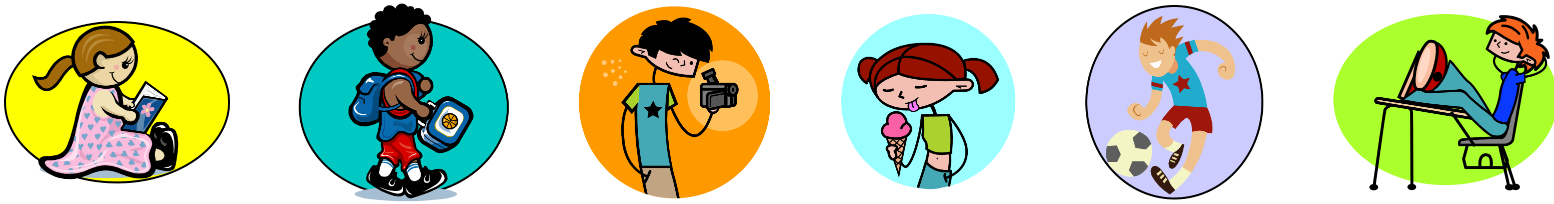
     

Interested in proportion of girls

Proportion of girls – 1/3

Select samples of two from population

Suppose we have a population of 6 students:
Alice, Ben, Charles, Denise, Edward, & Frank



Interested in proportion of girls

Proportion of girls – 1/3

Select samples of two from population

$$_6C_2 = 15$$

15 different samples
sample proportion of the number of girls in each sample.

Alice & Ben	.5
Alice & Charles	.5
Alice & Denise	1
Alice & Edward	.5
Alice & Frank	.5
Ben & Charles	0
Ben & Denise	.5
Ben & Edward	0
Ben & Frank	0
Charles & Denise	.5
Charles & Edward	0
Charles & Frank	0
Denise & Edward	.5
Denise & Frank	.5
Edward & Frank	0

15 different samples
sample proportion of the number of girls in each sample.

Alice & Ben	.5
Alice & Charles	.5
Alice & Denise	1
Alice & Edward	.5
Alice & Frank	.5
Ben & Charles	0
Ben & Denise	.5
Ben & Edward	0
Ben & Frank	0
Charles & Denise	.5
Charles & Edward	0
Charles & Frank	0
Denise & Edward	.5
Denise & Frank	.5
Edward & Frank	0

mean and standard deviation of these sample proportions

15 different samples
sample proportion of the number of females in each sample.

Alice & Ben	.5	Ben & Frank	0
Alice & Charles	.5	Charles & Denise	.5
Alice & Denise	1	Charles & Edward	0
Alice & Edward	.5	Charles & Frank	0
Alice & Frank	.5	Denise & Edward	.5
Ben & Charles	0	Denise & Frank	.5
Ben & Denise	.5	Edward & Frank	0
Ben & Edward	0		

mean and standard deviation of these sample proportions

$$\mu_{\hat{p}} = \frac{1}{3} \quad \text{and} \quad \sigma_{\hat{p}} = 0.29814$$

15 different samples

sample proportion of the number of females in each sample.

Alice & Ben	.5	Ben & Frank	0
Alice & Charles	.5	Charles & Denise	.5
Alice & Denise	1	Charles & Edward	0
Alice & Edward	.5	Charles & Frank	0
Alice & Frank	.5	Denise & Edward	.5
Ben & Charl			.5
Ben & Denis			0
Ben & Edwa			

Compare the mean of the sampling distribution to the population parameter (p)?

mean and stand deviation of these sample proportions

$$\mu_{\hat{p}} = \frac{1}{3} \quad \text{and} \quad \sigma_{\hat{p}} = 0.29814$$

General Properties for Sampling Distributions of $\widehat{p}$

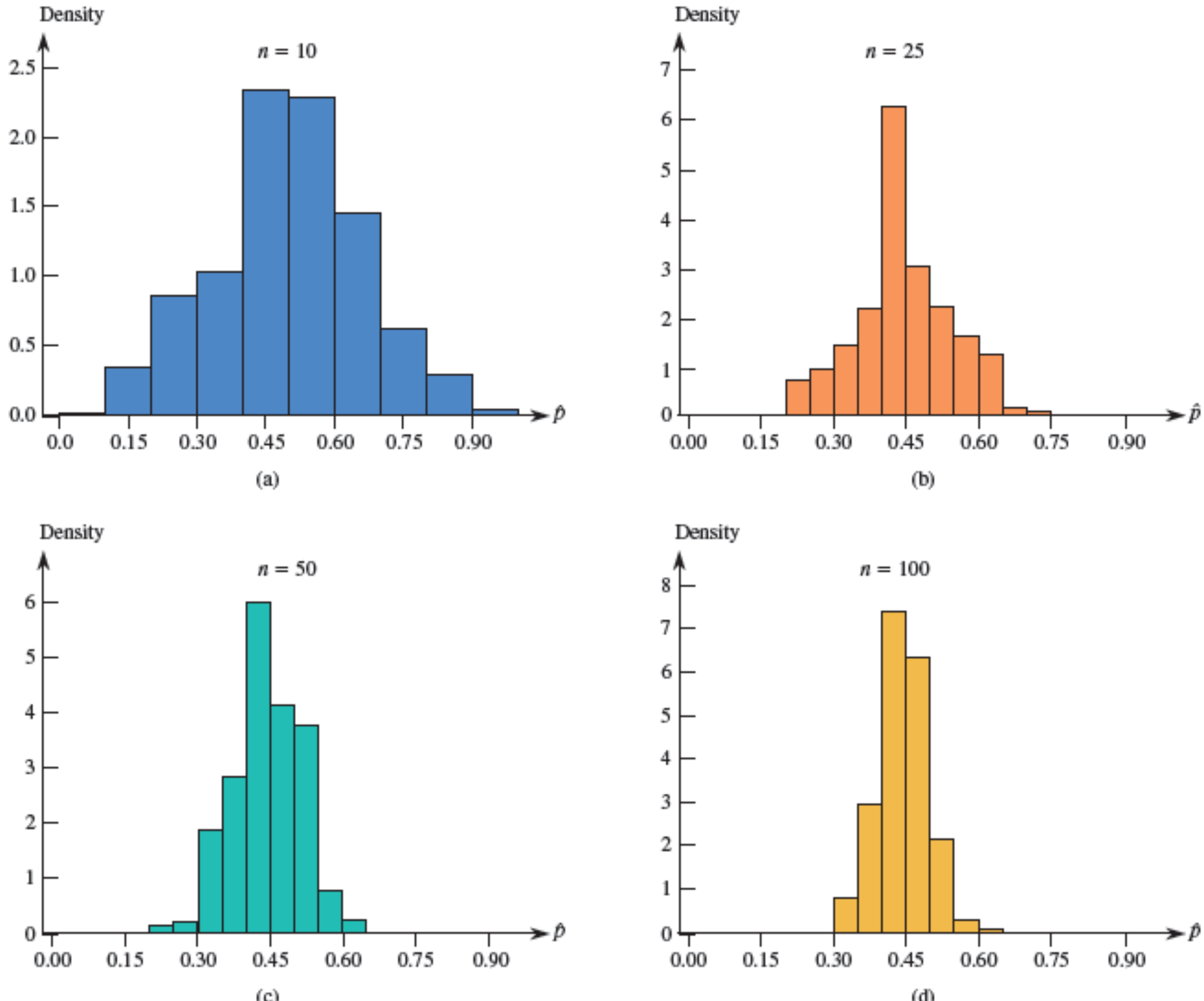
$$\mu_{\hat{p}} = \boldsymbol{p}$$

In the fall of 2008, there were 18,516 students enrolled at California Polytechnic State University, San Luis Obispo. Of these students, 8091 (43.7%) were women. We will use a statistical software package to simulate sampling from this Cal Poly population.

We will generate 500 samples of each of the following sample sizes: n = 10, n = 25, n = 50, n = 100 and compute the proportion of women for each sample.

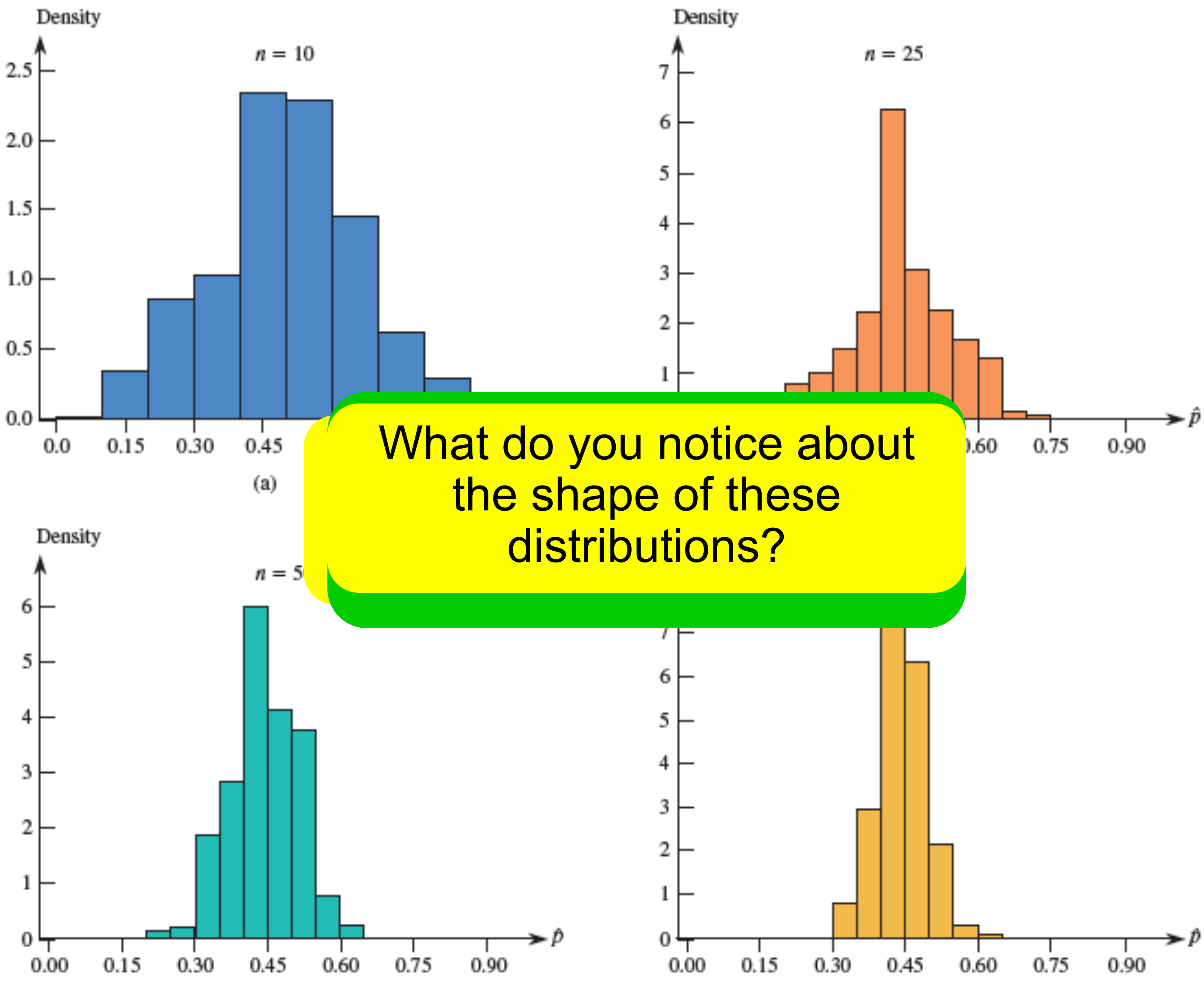
The following histograms display the distributions of the sample proportions for the 500 samples of each sample size.

p = 43.7%



Density
n = 10
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.15
0.30
0.45
0.60
0.75
0.90
p̂
(a)
Density
n = 25
7
6
5
4
3
2
1
0
0.00
0.15
0.30
0.45
0.60
0.75
0.90
p̂
(b)
Density
n = 50
6
5
4
3
2
1
0
0.00
0.15
0.30
0.45
0.60
0.75
0.90
p̂
(c)
Density
n = 100
8
7
6
5
4
3
2
1
0
0.00
0.15
0.30
0.45
0.60
0.75
0.90
p̂
(d)

p = 43.7%



What do you notice about the shape of these distributions?

General Properties for Sampling Distributions of $\hat{p}$

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Rule is exact if population is infinite

Approximately correct if population is finite and no more than 10% of the population is included in sample

General Properties for Sampling Distributions of $\hat{p}$

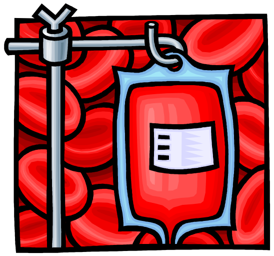
$$\mu_{\hat{p}} = p$$

Note: previous 6 student example this standard deviation formula was not correct because the sample size was more than 10% of population

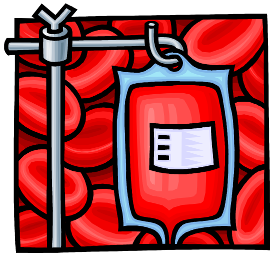
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Rule is exact if population is infinite

Approximately correct if population is finite and no more than 10% of the population is included in sample

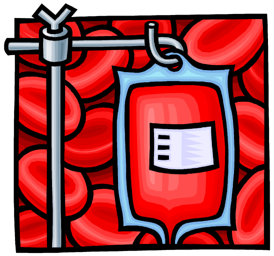


The development of viral hepatitis after a blood transfusion can cause serious complications for a patient. The article "Lack of Awareness Results in Poor Autologous Blood Transfusions" (Health Care Management, May 15, 2003) reported that hepatitis occurs in 7% of patients who receive blood transfusions during heart surgery. We will simulate sampling from a population of blood recipients.



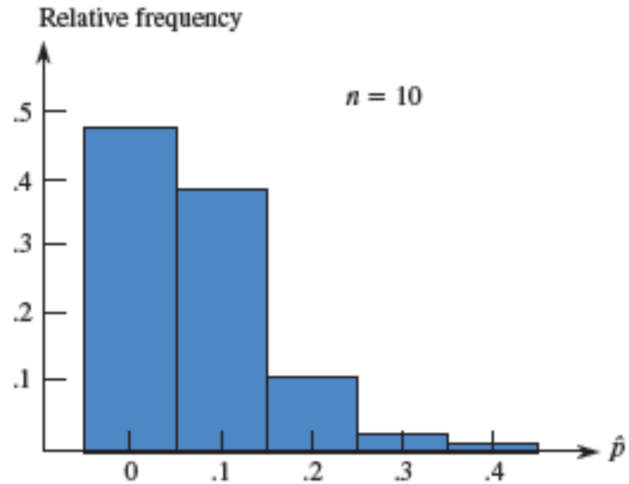
The development of viral hepatitis after a blood transfusion can cause serious complications for a patient. The article "Lack of Awareness Results in Poor Autologous Blood Transfusions" (Health Care Management, May 15, 2003) reported that hepatitis occurs in 7% of patients who receive blood transfusions during heart surgery. We will simulate sampling from a population of blood recipients.

$p = 0.07$

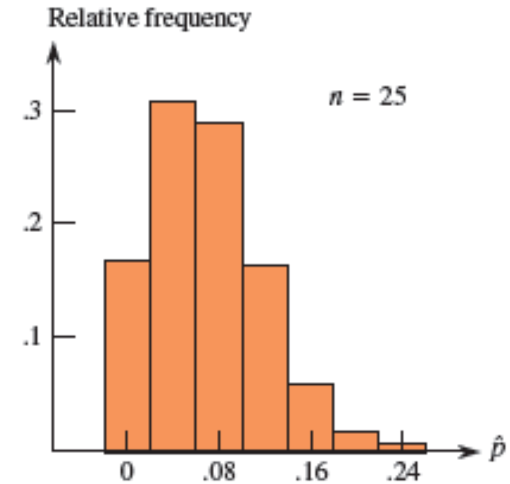


The development of viral hepatitis after a blood transfusion can cause serious complications for a patient. The article "Lack of Awareness Results in Poor Autologous Blood Transfusions" (Health Care Management, May 15, 2003) reported that hepatitis occurs in 7% of patients who receive blood transfusions during heart surgery. We will simulate sampling from a population of blood recipients.

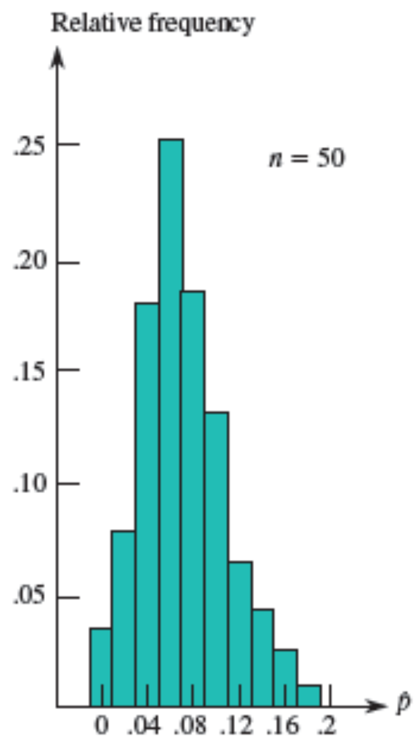
generate 500 samples of each of the following sample sizes: n = 10, n = 25, n = 50, n = 100 and compute the proportion of people who contract hepatitis for each sample


Relative frequency
$n = 10$
.5
.4
.3
.2
.1
0
.1
.2
.3
.4
$\hat{p}$

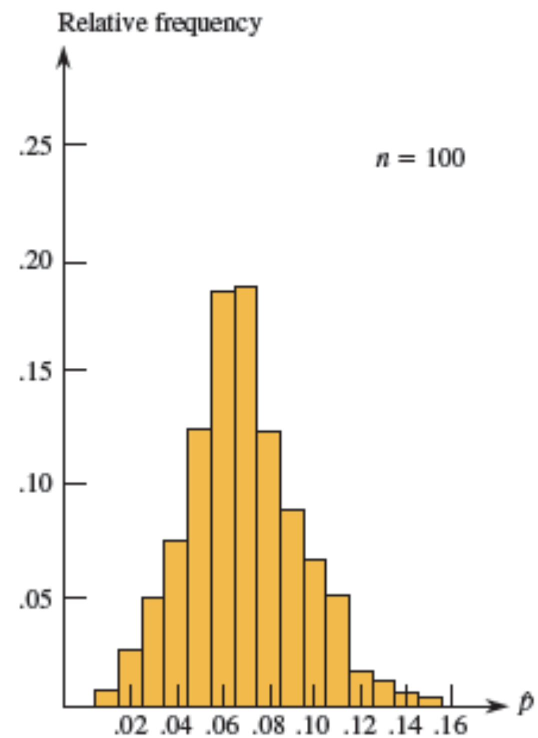
(a)


Relative frequency
$n = 25$
.3
.2
.1
0
.08
.16
.24
$\hat{p}$

(b)

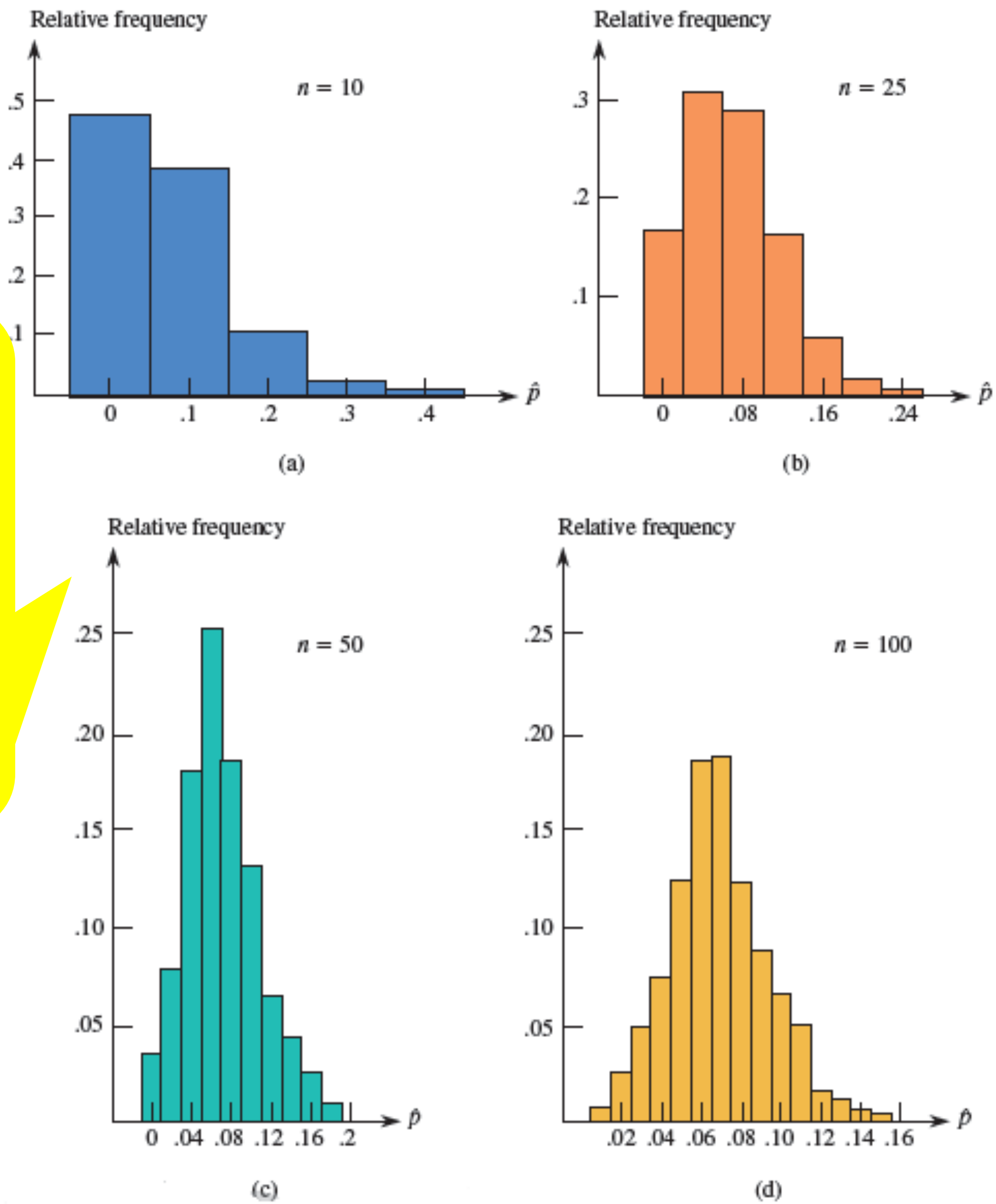

Relative frequency
$n = 50$
.25
.20
.15
.10
.05
0
.04
.08
.12
.16
.2
$\hat{p}$

(c)

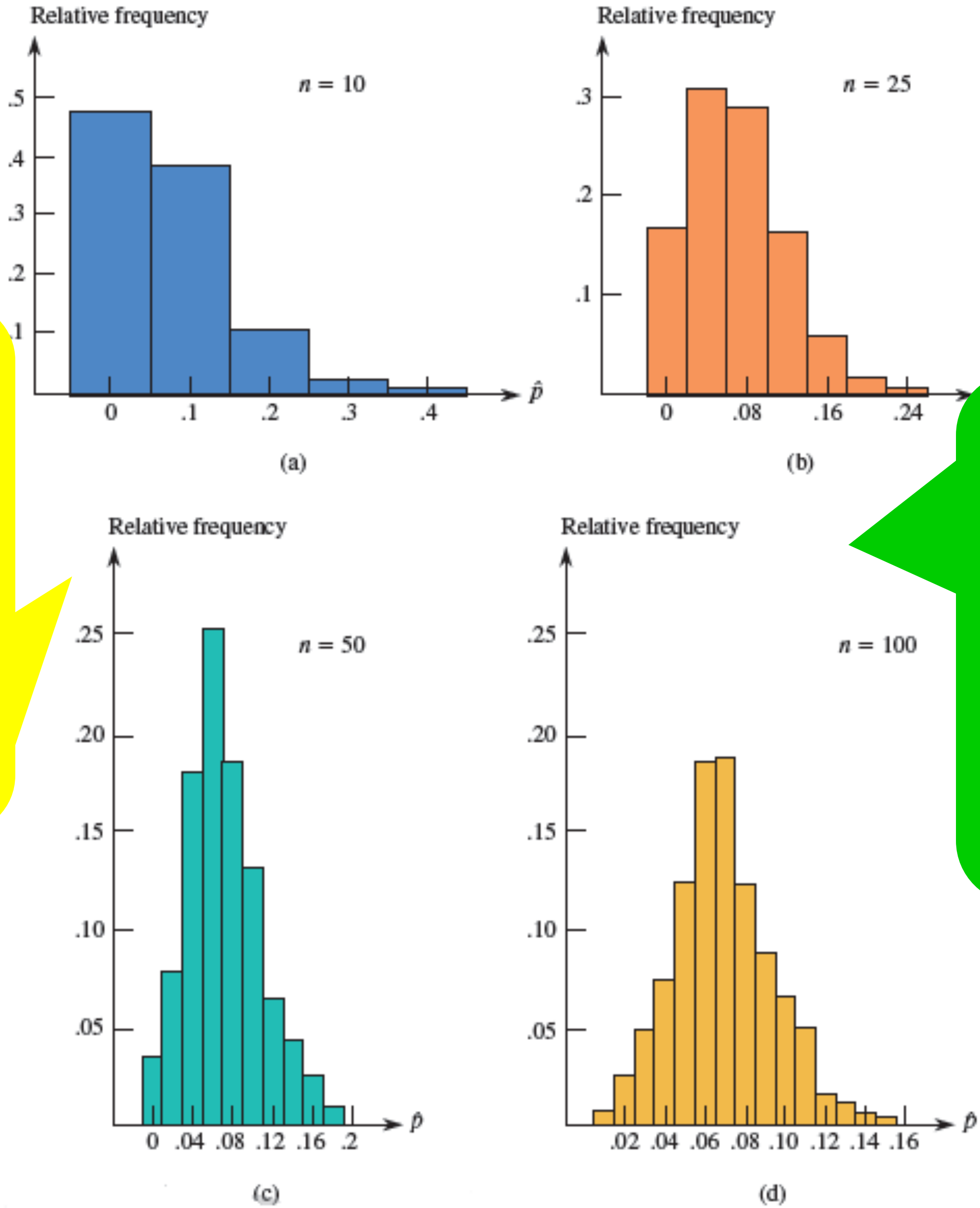

Relative frequency
$n = 100$
.25
.20
.15
.10
.05
.02
.04
.06
.08
.10
.12
.14
.16
$\hat{p}$

(d)

Are these histograms centered around the true proportion p = .07?



Are these histograms centered around the true proportion p = .07?



What happens to the shape of these histograms as the sample size increases?

General Properties Continued . . .

When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal.

General Properties Continued . . .

When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal.

Farther value of p is from 0.5, larger n must be for sampling distribution of $\hat{p}$ to be approximately normal

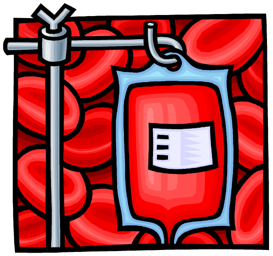
General Properties Continued . . .

When n is large and p is not too near 0 or 1, the sampling distribution of $\hat{p}$ is approximately normal.

Farther value of p is from 0.5, larger n must be for sampling distribution of $\hat{p}$ to be approximately normal

<u>A conservative rule of thumb:</u>

If $np \geq 10$ and $n\,(1 - p) \geq 10$, then a normal distribution provides a reasonable approximation to the sampling distribution of $\hat{p}$.

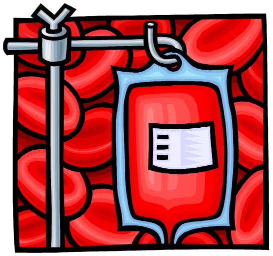


Blood Transfusions Revisited . . .

Let p = proportion of patients who contract hepatitis after a blood transfusion

$p = .07$

Suppose a new blood screening procedure is believed to reduce the incident rate of hepatitis. Blood screened using this procedure is given to $n = 200$ blood recipients. Only 6 of the 200 patients contract hepatitis. Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?



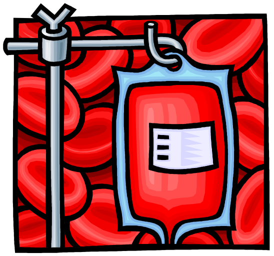
Blood Transfusions Revisited . . .

Let p = proportion of patients who contract hepatitis after a blood transfusion

$p = .07$

consider the sampling distribution of $\hat{p}$

Suppose a new bloo d to reduce the incident rate of hepatitis. Blood screened using this procedure is given to $n = 200$ blood recipients. Only 6 of the 200 patients contract hepatitis. Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

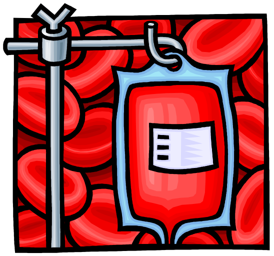


Blood Transfusions Revisited . . .

Let p = .07

$\hat{p}$ = 6/200 = .03

Is the sampling distribution approximately normal?



Blood Transfusions Revisited . . .

Let $p = .07$

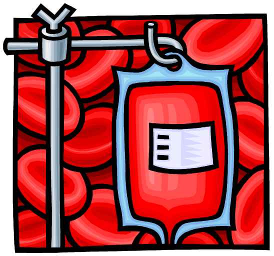
$\hat{p} = 6/200 = .03$

Is the sampling distribution approximately normal?

$np = 200(.07) = 14 \geq 10$

$n(1-p) = 200(.93) = 186 \geq 10$

Yes, we can use a normal approximation.



Blood Transfusions Revisited . . .

Let $p = .07$

$\hat{p} = 6/200 = .03$

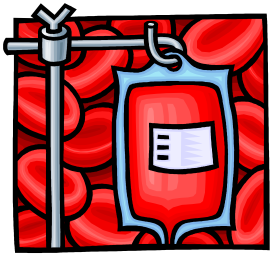
Is the sampling distribution approximately normal?

$np = 200(.07) = 14 \geq 10$

$n(1-p) = 200(.93) = 186 \geq 10$

Yes, we can use a normal approximation.

What is the mean and standard deviation of the sampling distribution?



Blood Transfusions Revisited . . .

Let $p = .07$

$\hat{p} = 6/200 = .03$

Is the sampling distribution approximately normal?

$np = 200(.07) = 14 \geq 10$

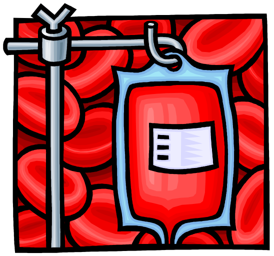
$n(1-p) = 200(.93) = 186 \geq 10$

Yes, we can use a normal approximation.

What is the mean and standard deviation of the sampling distribution?

$$\mu_{\hat{p}} = .07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$



Blood Transfusions Revisited . . .

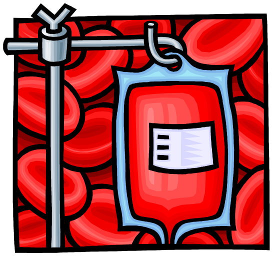
Let p = .07

$\mu_{\hat{p}} = .07$

$\hat{p}$ = 6/200 = .03

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?



Blood Transfusions Revisited . . .

Let $p = .07$

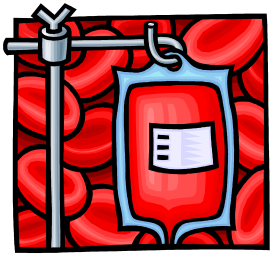
$\hat{p} = 6/200 = .03$

$$\mu_{\hat{p}} = .07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

If not, assume the screening procedure is not effective and $p = .07$.



Blood Transfusions Revisited . . .

Let $p = .07$

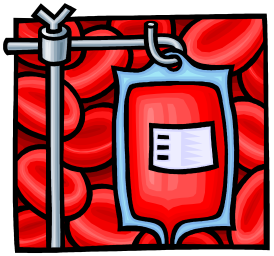
$\hat{p} = 6/200 = .03$

$$\mu_{\hat{p}} = .07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

$P(\hat{p} < .03) =$



Blood Transfusions Revisited . . .

Let p = .07

$\hat{p}$ = 6/200 = .03

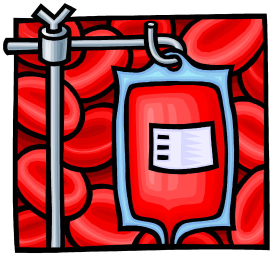
$$\mu_{\hat{p}} = .07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

$P(\hat{p} < .03) =$

$$z = \frac{.03 - .07}{\sqrt{\frac{.07(.93)}{200}}} = -2.22$$



Blood Transfusions Revisited . . .

Let p = .07

$\hat{p}$ = 6/200 = .03

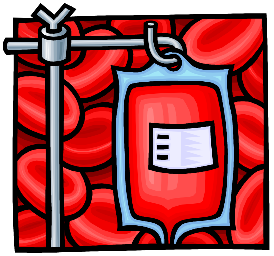
$$\mu_{\hat{p}} = .07$$

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

P($\hat{p}$ < .03) = .0132

$$z = \frac{.03 - .07}{\sqrt{\frac{.07(.93)}{200}}} = -2.22$$



Blood Transfusions Revisited . . .

Let p = .07

$$\mu_{\hat{p}} = .07$$

$\hat{p}$ = 6/200 = .03

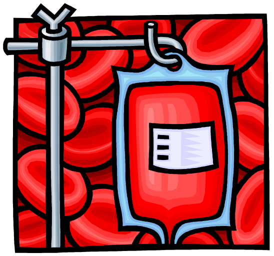
$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7

$P(\hat{p} < .03) = .0132$

$$z = \frac{.03 - .07}{\sqrt{\frac{.07(.93)}{200}}} = -2.22$$

This small probability tells us that it is unlikely that a sample proportion of .03 or smaller would be observed if the screening procedure was ineffective.



Blood Transfusions Revisited . . .

Let p = .07

$$\mu_{\hat{p}} = .07$$

$\hat{p}$ = 6/200 = .03

$$\sigma_{\hat{p}} = \sqrt{\frac{.07(.93)}{200}} = .018$$

Does this result indicate that the true proportion of patients who contract hepatitis when the new screening is used is less than 7%?

$P(\hat{p} < .03) = .0132$

$$z = \frac{.03 - .07}{\sqrt{\frac{.07(.93)}{200}}} = -2.22$$

This new screening procedure appears to yield a smaller incidence rate for hepatitis.

- Suppose that 40% of all U.S. employees contribute to a retirement plan.

- a. In a random sample of 100 employees, what is the approximate probability that at least half of those in the sample contribute to a retirement plan?

- b. Suppose you were told that least 60 of the 100 employees in a sample from Louisiana contribute to a retirement plan. Would you think $p = .40$ for Louisiana? Explain and justify your answer statistically.