STA 320 Fall 2013

Thursday, Dec 5

- **Sampling Distribution**

Review

- We cannot tell what will happen in any given individual sample (just as we can not predict a single coin flip in advance).
- We CAN tell a lot about the pattern of variation amongst many samples (just as we can predict that if you flip the coin a lot, you will get about 50% heads and 50% tails).
- In our doctor example, we found that the pattern of variation of the sample proportions, called the ***sampling distribution***, followed a normal distribution.
- http://www.amstat.org/publications/jse/v6n3/applets/clt.html

Sampling Distributions for Proportions

- Suppose we have a population of size N consisting of M successes and N-M failures.
- We sample a group of n people at random.
- Suppose further that
  - n/N is small (rule of thumb: less than 5%)
  - n is not small (rule of thumb: n>25)
  - M/N=p is not too close to 0 or 1 (rule of thumb: 0.05<p<0.95).
- Then the sampling distribution of the sample proportion is
  - **normal**
  - with **mean M/N=p** (the population proportion)
  - and **standard deviation sqrt(p(1-p)/n**).
- *Why this is true is beyond the scope of this course. It is because of a beautiful mathematical theorem:* ***Central Limit Theorem.***

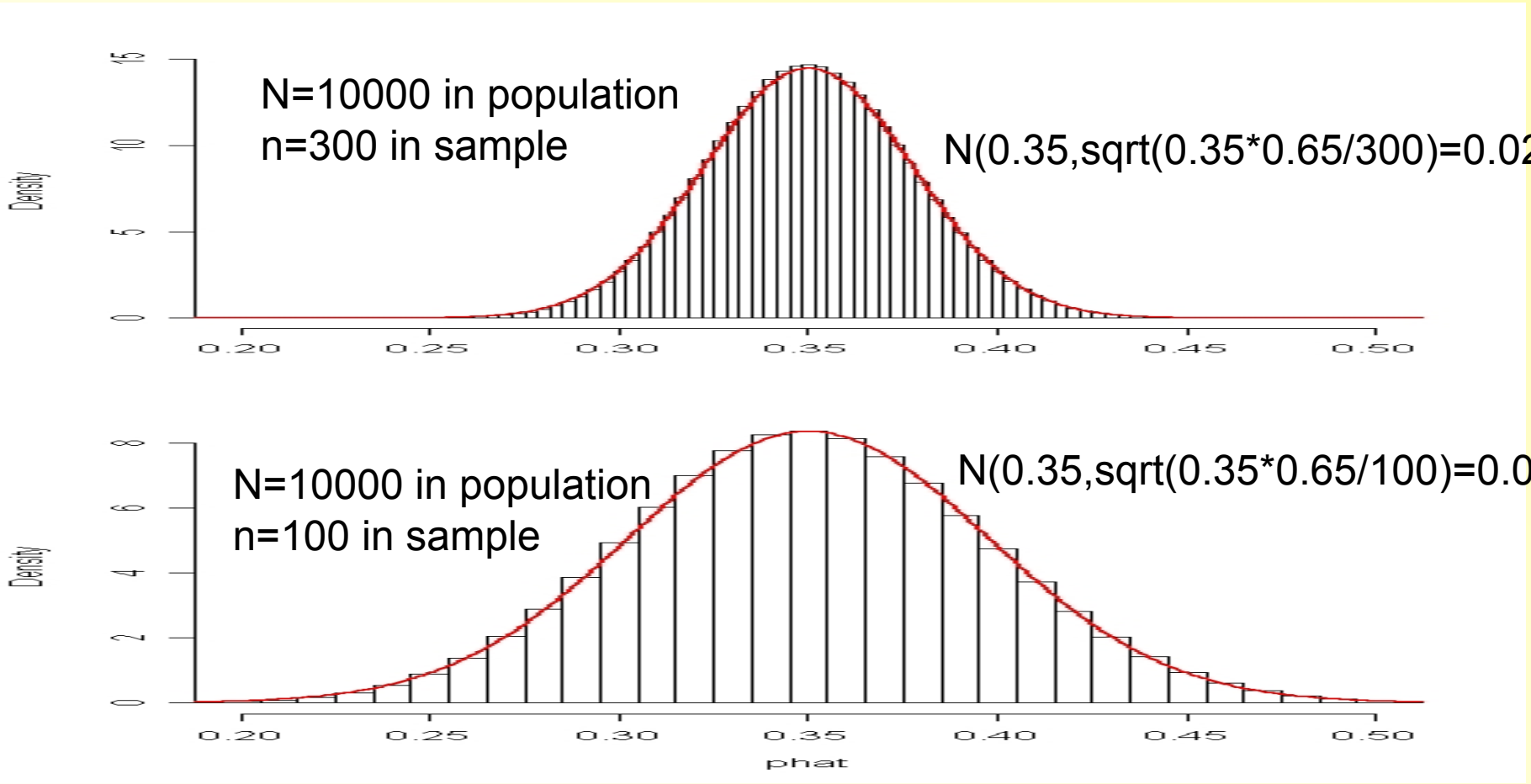
In Practice

- Unfortunately, we typically only get to draw one sample. How do you know if you got one of the samples that fall in the middle 95% (closer to the true proportion) as opposed to the outer 5% (farther from the true proportion)?
- Answer – really, you don't.
- But it's more likely you're in the 95% group than the 5% group.
- Want to be more sure?
- Construct a 99% group instead of a 1% group, then the odds are even more in your favor.

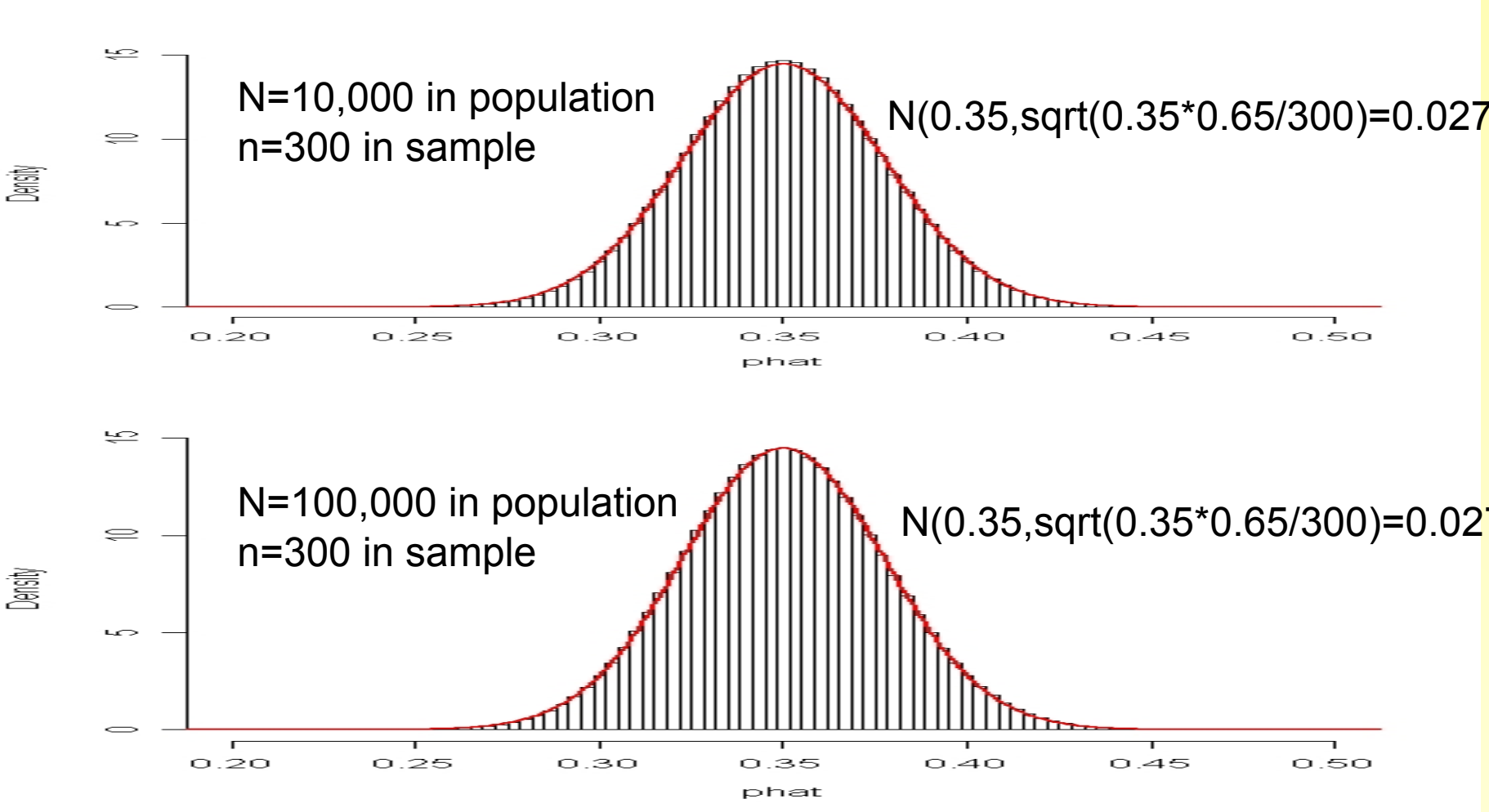
What Matters, What Doesn't

- The center of the sampling distribution is the true proportion p.
- On average, p-hat is centered around p.
- The sample size appears in the standard deviation sqrt(p (1-p)/n).
- The bigger the sample size, the smaller the standard deviation of p-hat. In other words, the closer p-hat tends to be to p.
- The population size does NOT matter.
- As long as you are sampling less than 1 in 20 people, it does not matter whether it is 1 of every 2000 or 1 of every 2 million.

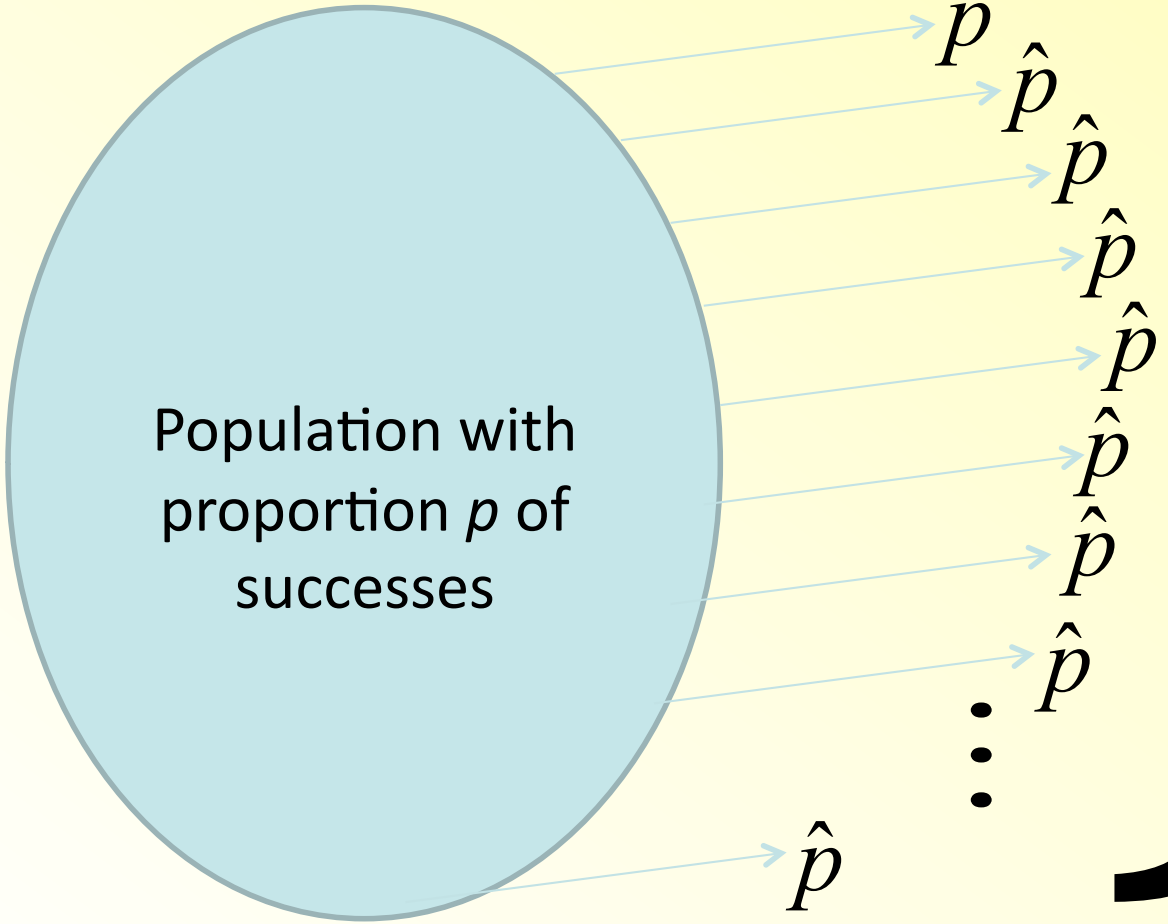
Population Size N=10000, 35% Successes Comparing n=300 to n=100



Sample Size n=300, 35% Successes Comparing N=10000 to N=100000



Summary: Sampling Distribution



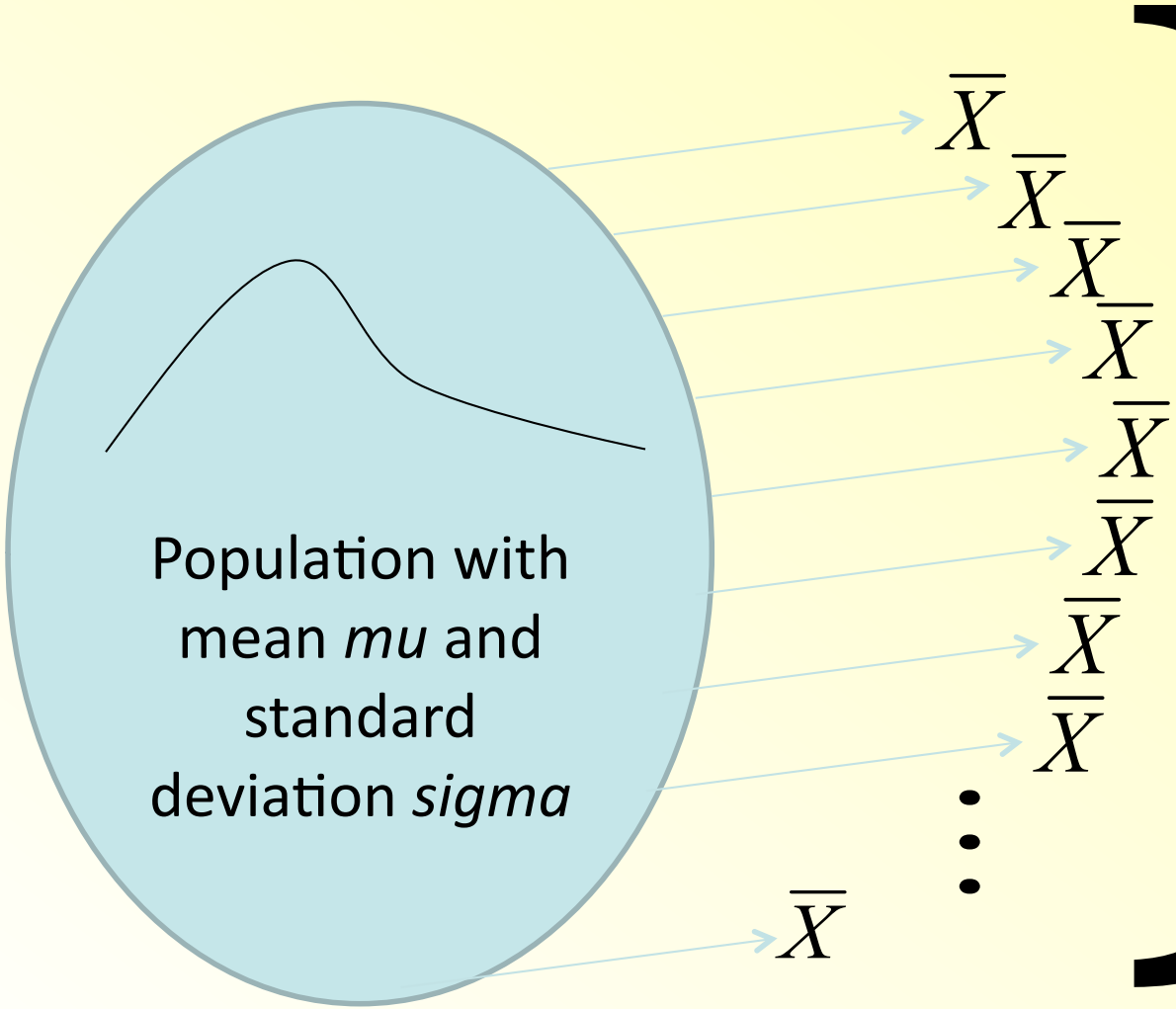
- If you repeatedly take random samples and calculate the sample proportion each time, the distribution of the sample proportions follows a pattern
- This pattern is called the *sampling distribution of p-hat*

Properties of the Sampling Distribution

- Expected Value of the $\hat{p}$'s: p.
- Standard deviation of the $\hat{p}$'s: $\sqrt{\frac{p(1-p)}{n}}$

  also called the *standard error* of $\hat{p}$
- ***Central Limit Theorem***: As the sample size increases, the distribution of the $\hat{p}$'s gets closer and closer to the normal.

Sampling Distribution of Means



- If you repeatedly take random samples and calculate the sample mean each time, the distribution of the sample means follows a pattern
- This pattern is the *sampling distribution of X-bar*

Properties of the Sampling Distribution

- Expected Value of the $\overline{X}$'s: μ.

- Standard deviation of the $\overline{X}$'s: $\frac{\sigma}{\sqrt{n}}$

also called the *standard error* of $\overline{X}$

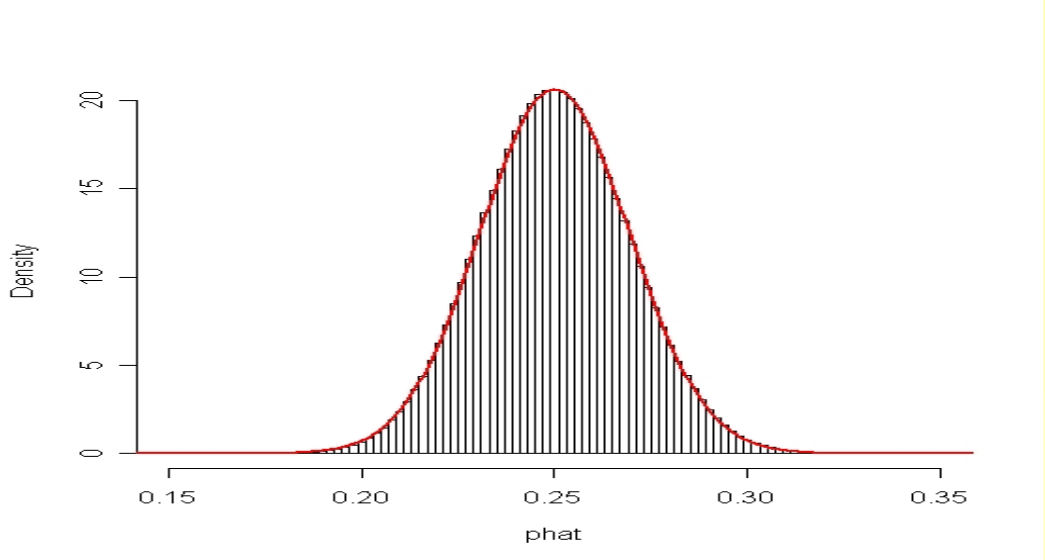
For N/n<20, use a finite population correction factor for the standard deviation: $\sqrt{\frac{N-n}{N-1}}$

- ***Central Limit Theorem***: As the sample size increases, the distribution of the $\overline{X}$'s gets closer and closer to a normal curve.

Summary: Sampling Distribution

- We cannot tell what will happen in any given individual sample.
- We CAN tell a lot about the pattern of variation amongst many samples.

Graph of sample proportions for all possible samples for selecting 500 people from a population with 25000 successes and 75000 failures, overlaid with a perfect normal curve.



Summary: Population, Sample, and Sampling Distribution

- Population
  - Total set of all subjects of interest
  - Can be described by (unknown) parameters
  - Want to make inference about its parameters
- Sample
  - Data that we observe
  - We describe it, using descriptive statistics
  - For large *n*, the sample resembles the population
- Sampling Distribution
  - Probability distribution of a statistic (for example, sample mean, sample proportion)
  - Used to determine the probability that a statistic falls within a certain distance of the population parameter
  - For large *n*, the sampling distribution (of sample mean, sample proportion) looks more and more like a normal distribution

Summary: Central Limit Theorem

- The most important theorem in statistics
- For random sampling, as the sample size *n* grows, the sampling distribution of the sample mean $\overline{Y}$ (and of the sample proportion p-hat) approaches a normal distribution
- Amazing: This is the case even if the population distribution is discrete or highly skewed
  - Online applet 1
  - Online applet 2
- The Central Limit Theorem can be proved mathematically (STA 524)

Central Limit Theorem

- Usually, the sampling distribution of $\overline{Y}$ is approximately normal for sample sizes of at least *n*=25 (rule of thumb)
- In addition, we know that the parameters of the sampling distribution are mean=mu and standard error= $\frac{\sigma}{\sqrt{n}}$
- For example:

  If the sample size is at least *n*=25, then with 95% probability, the sample mean falls between

$$\mu - 1.96\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 1.96\frac{\sigma}{\sqrt{n}}$$

(μ = population mean,

σ = population standard deviation)

Calculating *z*-Scores

1. **z-Score for an individual observation**
   - You need to know *Y*, *mu*, and sigma to calculate *z*

$$z = \frac{Y - \mu}{\sigma}$$

2. **z-Score for a sample mean**
   - You need to know *Y-bar*, *mu*, *sigma*, and *n* to calculate *z*

$$z = \frac{\bar{Y} - \mu}{\sigma / \sqrt{n}}$$

3. **z-Score for a sample proportion**
   - You need to know *p-hat*, *p*, and *n* to calculate *z*

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

Population Parameters and Sample Statistics

Population parameter	*Value*	*Sample statistic used to estimate*
p *proportion of population with a certain characteristic*	*Unknown*	$\hat{p}$
μ *mean value of a population variable*	*Unknown*	$\overline{x}$

- The value of a population parameter is a **fixed** number, it is NOT random; its value is **not known.**
- The value of a sample statistic is calculated from sample data
- The value of a sample statistic will vary from sample to sample (sampling distributions)

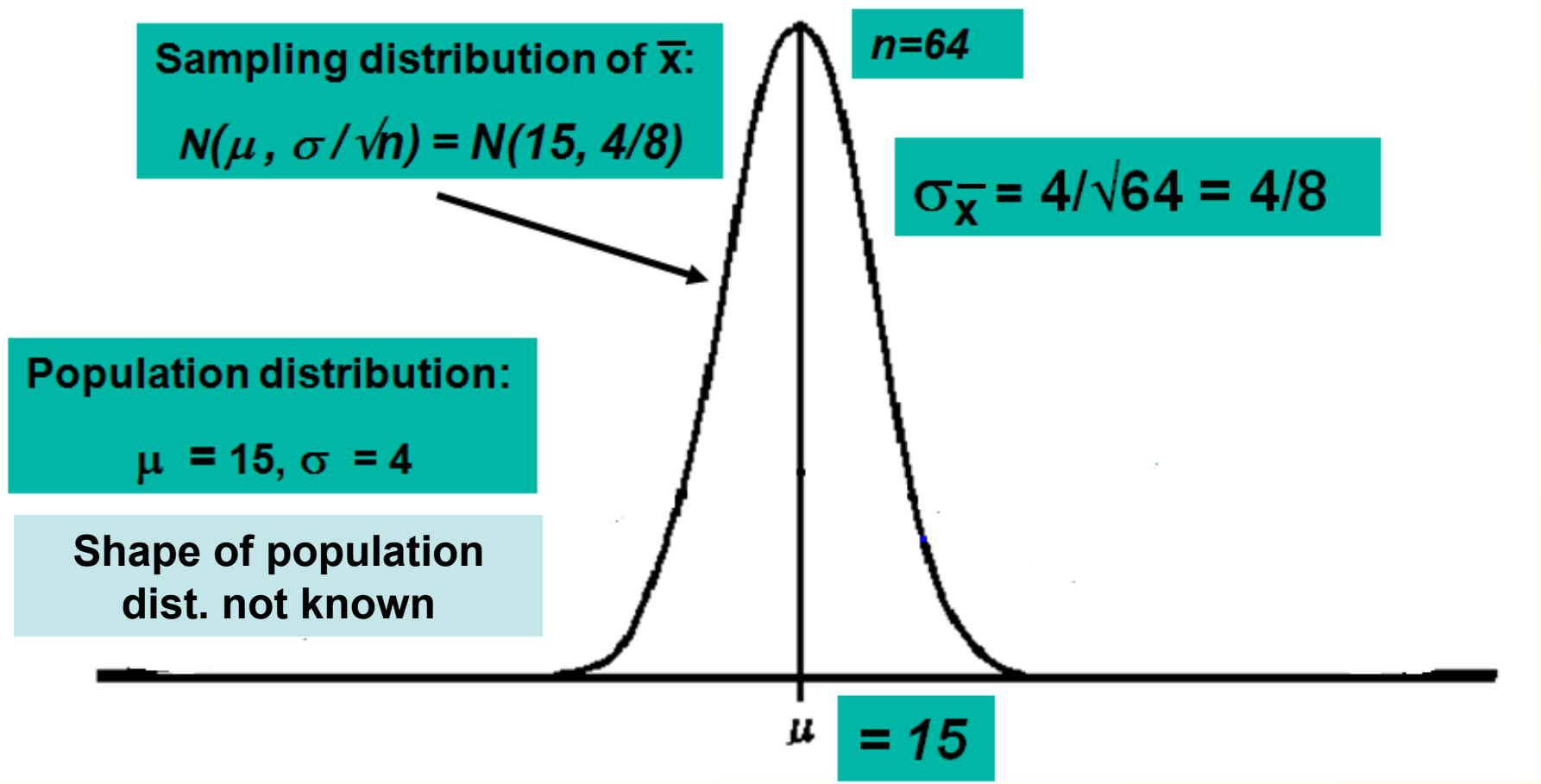
More Example

A random sample of n=64 observations is drawn from a population with mean μ=15 and standard deviation σ=4.

a. $E(\bar{X}) = \mu = 15; SD(\bar{X}) = \frac{SD(X)}{\sqrt{n}} = \frac{4}{8} = .5$

b. The shape of the sampling distribution model for $\bar{x}$ is approx. normal (by the CLT) with mean $E(\bar{X}) = 15$ and $SD(\bar{X}) = .5$. The answer depends on the sample size since $SD(\bar{X}) = \frac{SD(X)}{\sqrt{n}}$.

Graphically



More Example (cont.)

c. $\bar{x} = 15.5;$

$$z = \frac{\bar{x} - \mu}{SD(\bar{X})} = \frac{15.5 - 15}{.5} = \frac{.5}{.5} = 1$$

This means that $\bar{x}$=15.5 is one standard deviation above the mean $E(\bar{X}) = 15$

More Example 2

- The probability distribution of 6-month incomes of account executives has mean $20,000 and standard deviation $5,000.
- a) A single executive's income is $20,000. Can it be said that this executive's income exceeds 50% of all account executive incomes?

More Example 2

- The probability distribution of 6-month incomes of account executives has mean $20,000 and standard deviation $5,000.
- a) A single executive's income is $20,000. Can it be said that this executive's income exceeds 50% of all account executive incomes?

ANSWER No. P(X<$20,000)=? No information given about shape of distribution of X; we do not know the median of 6-mo incomes.

More Example 2(cont.)

- b) n=64 account executives are randomly selected. What is the probability that the sample mean exceeds $20,500?

More Example 2(cont.)

- b) n=64 account executives are randomly selected. What is the probability that the sample mean exceeds $20,500?

answer $\mathbf{E(x) = \$20,000, SD(x) = \$5,000}$

$$E(\bar{x}) = \$20,000, SD(\bar{x}) = \frac{SD(x)}{\sqrt{n}} = \frac{5,000}{\sqrt{64}} = 625$$

$$\text{By CLT}, \bar{X} \sim N(20,000, 625)$$

$$P(\bar{X} > 20,500) = P\left(\frac{\bar{X} - 20,000}{625} > \frac{20,500 - 20,000}{625}\right) =$$

$$P(z > .8) = 1 - .7881 = .2119$$

More Example 3

- A sample of size n=16 is drawn from a normally distributed population with mean E(x)=20 and SD(x)=8.

More Example 3

- A sample of size n=16 is drawn from a normally distributed population with mean E(x)=20 and SD(x)=8.

$$X \sim N(20,8);\ \bar{X} \sim N(20, \frac{8}{\sqrt{16}})$$

$$a)\ P(\bar{X} \geq 24) = P(\frac{\bar{X}-20}{2} \geq \frac{24-20}{2}) = P(z \geq 2) =$$

$$1 - .9772 = .0228$$

$$b)\ P(16 \leq \bar{X} \leq 24) = P\left(\frac{16-20}{2} \leq z \leq \frac{24-20}{2}\right) =$$

$$P(-2 \leq z \leq 2) = .9772 - .0228 = .9544$$

More Example 3 (cont.)

- c. Do we need the Central Limit Theorem to solve part a or part b?

More Example 3 (cont.)

- c. Do we need the Central Limit Theorem to solve part a or part b?

- NO. We are given that the population is normal, so the sampling distribution of the mean will also be normal for any sample size n. The CLT is not needed.

More Example 4

- Battery life X~N(20, 10). Guarantee: avg. battery life in a case of 24 exceeds 16 hrs. Find the probability that a randomly selected case meets the guarantee.

More Example 4

- Battery life X~N(20, 10). Guarantee: avg. battery life in a case of 24 exceeds 16 hrs. Find the probability that a randomly selected case meets the guarantee.

$$E(\overline{x}) = 20; SD(\overline{x}) = \frac{10}{\sqrt{24}} = 2.04.\ \overline{X} \sim N(20, 2.04)$$

$$P(\overline{X} > 16) = P(\frac{\overline{X} - 20}{2.04} > \frac{16 - 20}{2.04}) = P(z > -1.96) =$$

$$.1 - .0250 = .9750$$

More Example 5

Cans of salmon are supposed to have a net weight of 6 oz. The canner says that the net weight is a random variable with mean μ=6.05 oz. and stand. dev. σ=.18 oz.

Suppose you take a random sample of 36 cans and calculate the sample mean weight to be 5.97 oz.

- Find the probability that the mean weight of the sample is less than or equal to 5.97 oz.

Population X: amount of salmon in a can

E(x)=6.05 oz, SD(x) = .18 oz

- $\overline{X}$ sampling dist: $E(\overline{x})=6.05$ $SD(\overline{x})=.18/6=.03$
- By the CLT, $\overline{X}$ sampling dist is approx. normal
- $P(\overline{X} \leq 5.97) = P(z \leq [5.97-6.05]/.03)$
  $=P(z \leq -.08/.03)=P(z \leq -2.67)= .0038$

- How could you use this answer?

- **Suppose you work for a "consumer watchdog" group**
- **If you sampled the weights of 36 cans and obtained a sample mean $\overline{x} \leq 5.97$ oz., what would you think?**
- **Since $P(\overline{x} \leq 5.97) = .0038$, either**
  - **you observed a "rare" event (recall: 5.97 oz is 2.67 stand. dev. below the mean) and the mean fill E(x) is in fact 6.05 oz. (the value claimed by the canner)**
  - **the true mean fill is less than 6.05 oz., (the canner is lying).**

More Example 6

- X: weekly income. E(x)=600, SD(x) = 100
- n=25; $\bar{X}$ sampling dist: E($\bar{x}$)=600 SD(x) =100/5=20
- P($\bar{X} \leq 550$)=P($z \leq$ [550-600]/20)
  =P($z \leq$ -50/20)=P($z \leq$ -2.50) = .0062

Suspicious of claim that average is $600; evidence is that average income is less.

More Example 7

- 12% of students at UK are left-handed. What is the probability that in a sample of 50 students, the sample proportion that are left-handed is less than 11%?

More Example 7

- 12% of students at UK are left-handed. What is the probability that in a sample of 50 students, the sample proportion that are left-handed is less than 11%?

$$E(\hat{p}) = p = .12; SD(\hat{p}) = \sqrt{\frac{.12 * .88}{50}} = .046$$

By the CLT, $\hat{p} \sim N(.12, .046)$

$$P(\hat{p} < .11) = P\left(\frac{\hat{p} - .12}{.046} < \frac{.11 - .12}{.046}\right)$$

$$= P(z < -.22) = .4129$$

Quiz I

- For women aged 18-24, systolic blood pressures are normally distributed with mean 114.8 [mm Hg] and standard deviation 13.1 [mm Hg]
- Hypertension is commonly defined as a value above 140. If a woman between 18 and 24 is randomly selected, find the probability that her systolic blood pressure is above 140
- For a sample of 4 women, find the probability that their mean systolic blood pressure is above 140
- *Note that for this problem, we don't actually need the central limit theorem because the variable "blood pressure" has a normal distribution – we don't need to rely on averages.*

Quiz II

- Analysts think that the length of time people work at a job has a mean of 6.1 years and a standard deviation of 4.3 years.
- Do you expect this distribution to be left-skewed or right-skewed or symmetric? Why?
- Can you calculate the probability that a randomly chosen person spends less than 5 years on his/ her job?
- What is the probability that 100 people selected at random spend an average of less than 5 years on their job?

Review: Multiple Choice Question

The Central Limit Theorem implies that

1. All variables have approximately bell-shaped sample distributions if a random sample contains at least 30 observations
2. Population distributions are normal whenever the population size is large
3. For large random samples, the sampling distribution of $\overline{Y}$ is approximately normal, regardless of the shape of the population distribution
4. The sampling distribution looks more like the population distribution as the sample size increases
5. All of the above