

# STA 320

# Fall 2013

*Tuesday, Dec 3*

➤ **Sampling Distribution**

# Populations and Samples

- The goal of **statistical inference** is to determine aspects of a large **population**, much of which is unobserved.
- We determine the aspects of the population by evaluating results of a fraction of the population, known as a **sample**.
- Recall:
  - sample part of the population
  - sample is seen, rest of the population is not

# Examples

- We want to predict an election. To this end, we call 600 people at random and ask them how they will vote. The 600 people we actually observe are the sample, while all voters (likely voters?) are the population.
- We are interested in testing a new drug for migraines. We administer the drug to 500 people and observe the amount of pain reduction. Our sample is the 500 people we observe, while the population is all migraine sufferers.

# Another Example

- We are interested whether job-training actually increases the likelihood of welfare recipients to be employed. We find 200 individuals on welfare, give the job training classes to 100 of them and do nothing with the other 100 (a control group). Later we observe how many of them are employed.
- All 200 individuals comprise the sample (everyone we observe is the sample). The population is all welfare recipients.

# Recall: Parameters

- A numerical aspect of the population is called a **parameter**.
- Typically, we would like to make statements about parameters, but they are unknown.
- Voting example: If we are trying to predict an election, we want to know which way the **entire population of voters** would vote if the election were held today. The population proportion of voters for candidate  $A$  is a parameter.

# Parameters and Statistics

- The proportion of **all** voters who would today vote for  $A$  (however you define  $A$ ) is a parameter.
- A **statistic** is any numerical aspect of the sample.
- We observe the sample, and thus we can calculate the statistic.
- Our goal is to use those known statistics to estimate the unknown population parameters.
- Fortunately, calculated from a good sample or experiment, sample statistics are close to population parameters.
- This forms the basis of **statistical inference**.

# Probability Theory – do we need it?

- Probability theory describes *how far* sample statistics will be from population parameters. If we know this information, we can make educated guesses about the parameters, often quite accurate ones.
- Suppose in the voting example we know there is a high likelihood the sample proportion will be within 4% of the population proportion. If the sample proportion is 58%, we have good reason to conclude the election will go in favor of the A votes.

# Sampling Distributions

- For the probability theory to work, your samples need to be drawn randomly from the population;
- Recall: “Simple random sample” means that every sample has the same probability of being chosen.
- Unfortunately, random sample will give different results each time – because of sampling variation.
- Fortunately, however, probability theory allows us to conclude that there is a **predictable pattern of variation** among the samples.

# Simple Example 1

- Suppose we have a population of 20 people, 12 of which will vote for  $X$  and 8 will vote for  $Y$ . We sample 5 people at random.
- The population will usually be bigger in practice, as will our sample, this is just for illustration.
- Also, in practice, we will obviously not know that 12 (=60%) will vote for  $X$  and 8 (=40%) will vote for  $Y$ .
- Label the people  $A, B, C, D, \dots, T$ . We could sample  $ABCDE$ , or  $ABCDF$ , or  $ABCDG$ , or  $DNORT$ , or any of the other 15,504 possibilities.

# Simple Example 1, contd.

- Each of the 15,504 possible samples of 5 people are equally likely. We don't know which one we will get.
- Probability Theory (not required in STA 320) allows us to determine that 56 of these possible samples have 0 “yes” people, 840 have 1 “yes” person, 3696 have 2 “yes” people, and so forth.

# Simple Example 1, contd.

# (%) "yes" Responses	Number of possible samples	Proportion of possible samples
0 (0% yes)	56	0.36%
1 (20% yes)	840	5.42%
2 (40% yes)	3696	23.84%
<b><u>3 (60% yes)</u></b>	<b><u>6160</u></b>	<b><u>39.73%</u></b>
4 (80% yes)	3960	25.54%
5 (100% yes)	792	5.11%
Total	15504	100%

# Simple Example 1, concluded

- Is your sample proportion guaranteed to be 0.60, exactly equal to your population proportion? No, but there is about a 40% chance it is.
- There is close to a 90% chance that the sample proportion will be within 0.2 of the population proportion.
- Thus, there is a high likelihood that a sample statistic calculated from a random sample will be close to the true (usually unknown) population proportion.
- With more realistic population and sample sizes, there is an even greater chance that the sample statistic will be close to the population parameter.

# Example 2

- Suppose there are 10,000 students on a campus.
- We want to know the average height, but can not measure all 10,000.
- Instead, we sample 100 individuals and measure those.
- Thus, we get to see just one of the *large* number of possible samples of 100 people out of 10,000.
- Fortunately, probability theory still says that our sample average height should be CLOSE to the population average height.

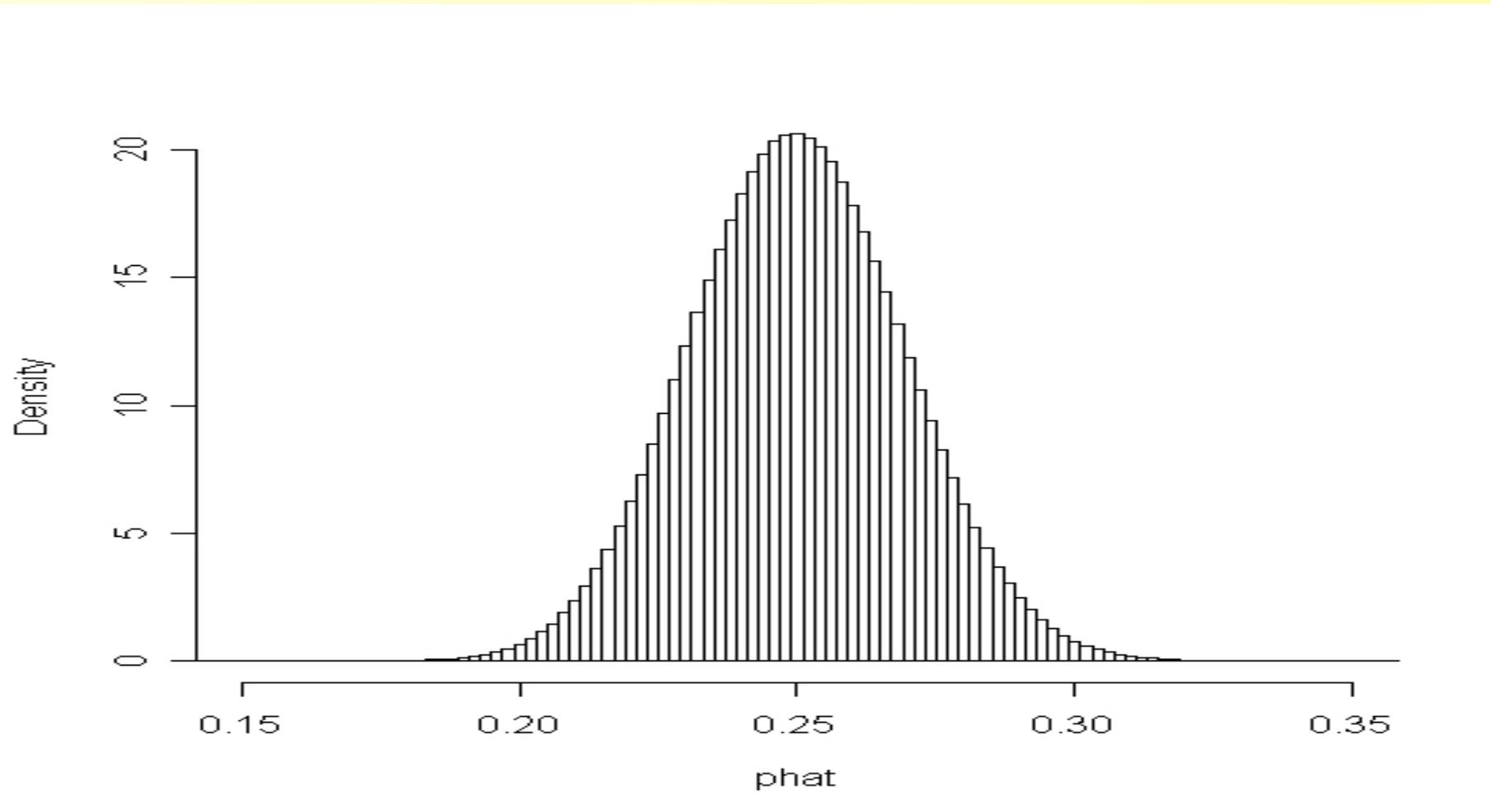
# Example 3

- We are interesting in determining what proportion of a population visits a doctor at least once a year.
- Our population contains 100,000 individuals. Unknown to us, 25,000 visit a doctor at least once a year while 75,000 do not.
- We decide to sample 500 at random and determine whether those individuals visit a doctor at least once a year (termed a success), as opposed to those who do not visit a doctor at least once a year (termed a failure).

- Note our population parameter is  $p=0.25$  (25,000 out of 100,000). This is typically unknown.
- Our sample of 500 might yield 130 successes, resulting in a sample proportion  $\hat{p}=0.260$ , or our sample of 500 might yield 122 successes, resulting  $\hat{p}=0.244$ .
- Because our sample is (and should be!) random, so we are not quite sure what will happen in any *single* sample.
- Again, however, out of the *very many* possible samples, a very large proportion of them have sample proportions close to the true proportion  $p=0.25$ .

- It turns out there are over  $10^{1365}$  (a one with 1365 zeroes after it) ways to pick 500 people out of 100,000 people. Your sample will be ONE of those many possible samples.
- It is still possible to figure out precisely how many of these samples contain 0 (=0%) successes, 1 (=0.2%) success, 2 (=0.4%) successes, and so on up to 500 (=100%) successes.

Graph of sample proportions for all possible samples for selecting 500 people from a population with 25000 successes and 75000 failures (*sampling distribution*).



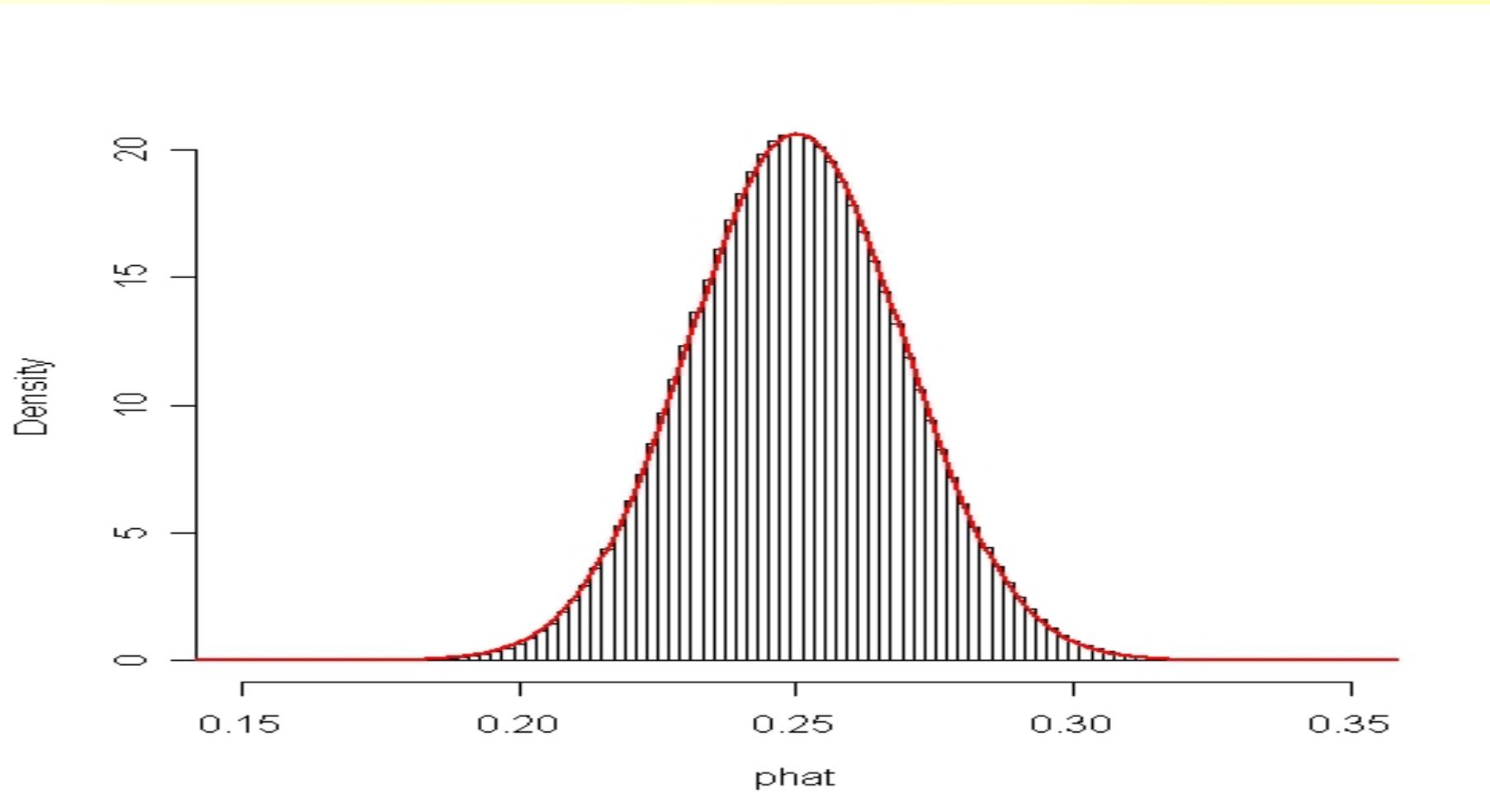
# Hm?

- That looks like a bell curve.
- In fact, it looks suspiciously like a bell curve with mean  $\mu=0.25$  (that is where the peak is).
- And the standard deviation is (less obvious, but true)

$$\sqrt{p(1-p)/n} = \sqrt{0.25*0.75/500} = 0.0194$$

- The next graph combines the histogram of sample proportions with the true bell curve with mean  $=0.25$  and standard deviation  $= 0.0194$ .

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.



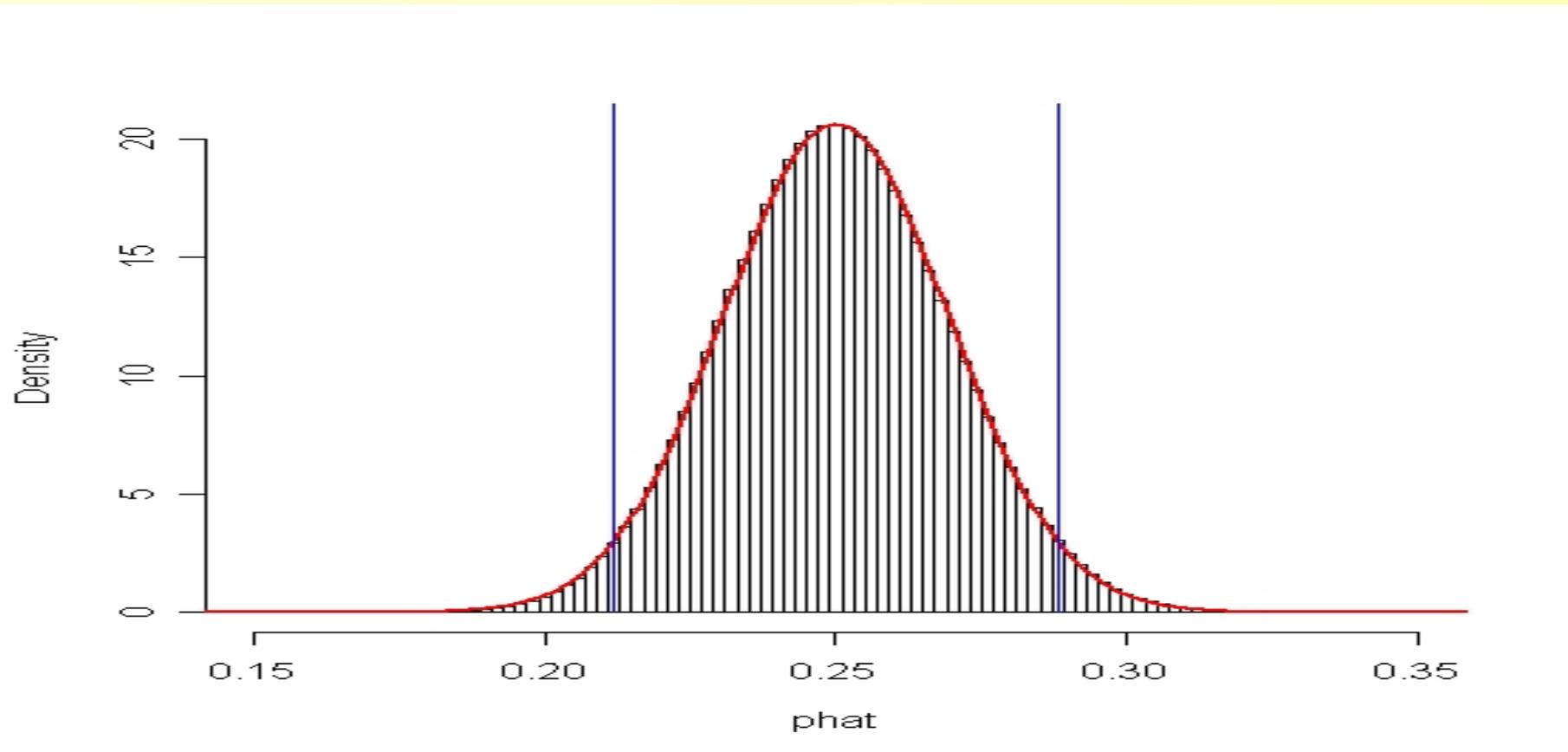
# Review

- 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/clk.html>

# Useful Consequences

- In our Example 3 (doctor visits), we know the sampling distribution of the sample proportion of successes is  $N(0.25, 0.0194)$ .
- Recall the 68-95-99.7 rule. We know there is about 95% probability that the sample proportion will be between 2 standard deviations ( $2 \cdot 0.0194 = 0.0388$ ) of the population proportion.
- There is a 99.7% chance the sample proportion will be within 3 standard deviations ( $0.0582$ ) of the population proportion.

Empirical Rule: About 95% of our observations should fall between the blue lines



- In actuality, we have 94.5%.

# 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.***

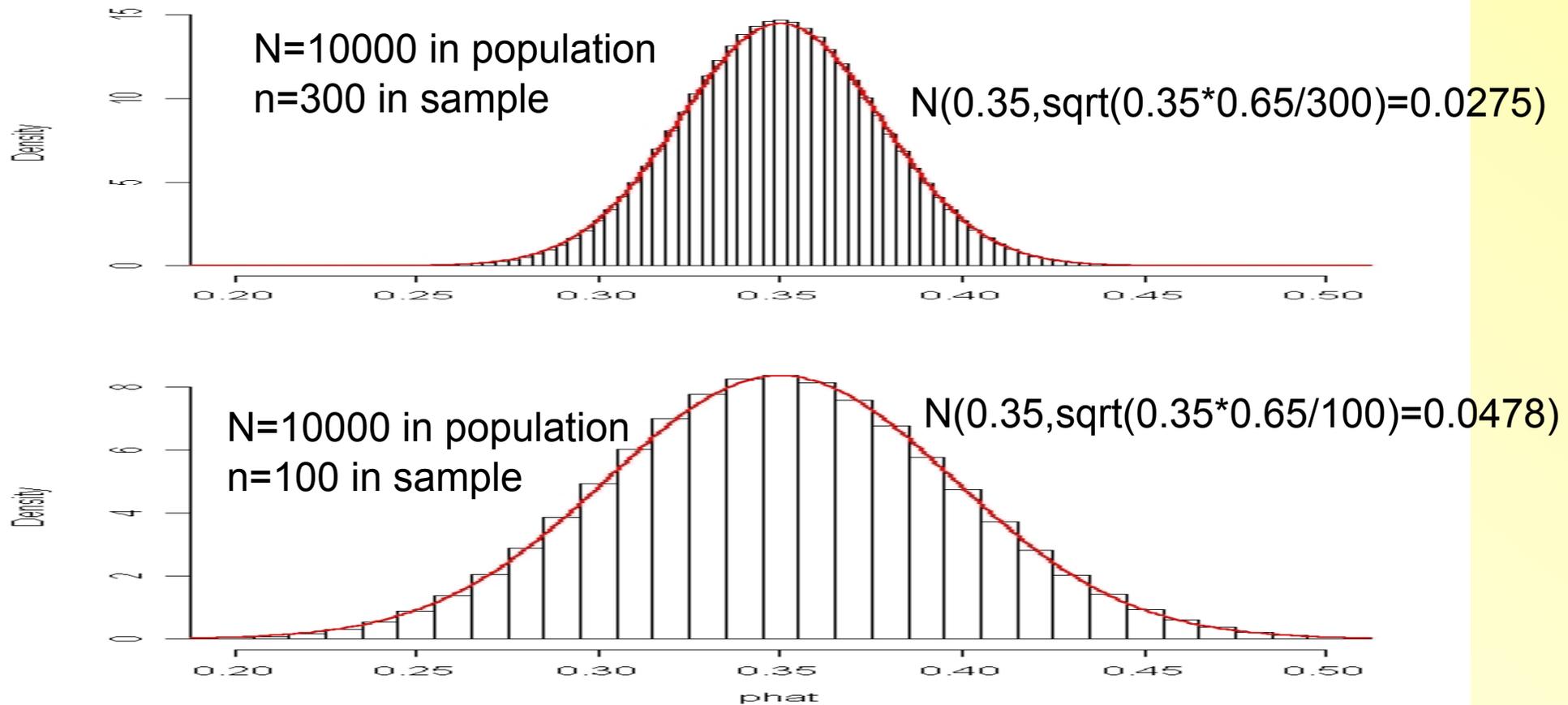
# In Practice

- 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.

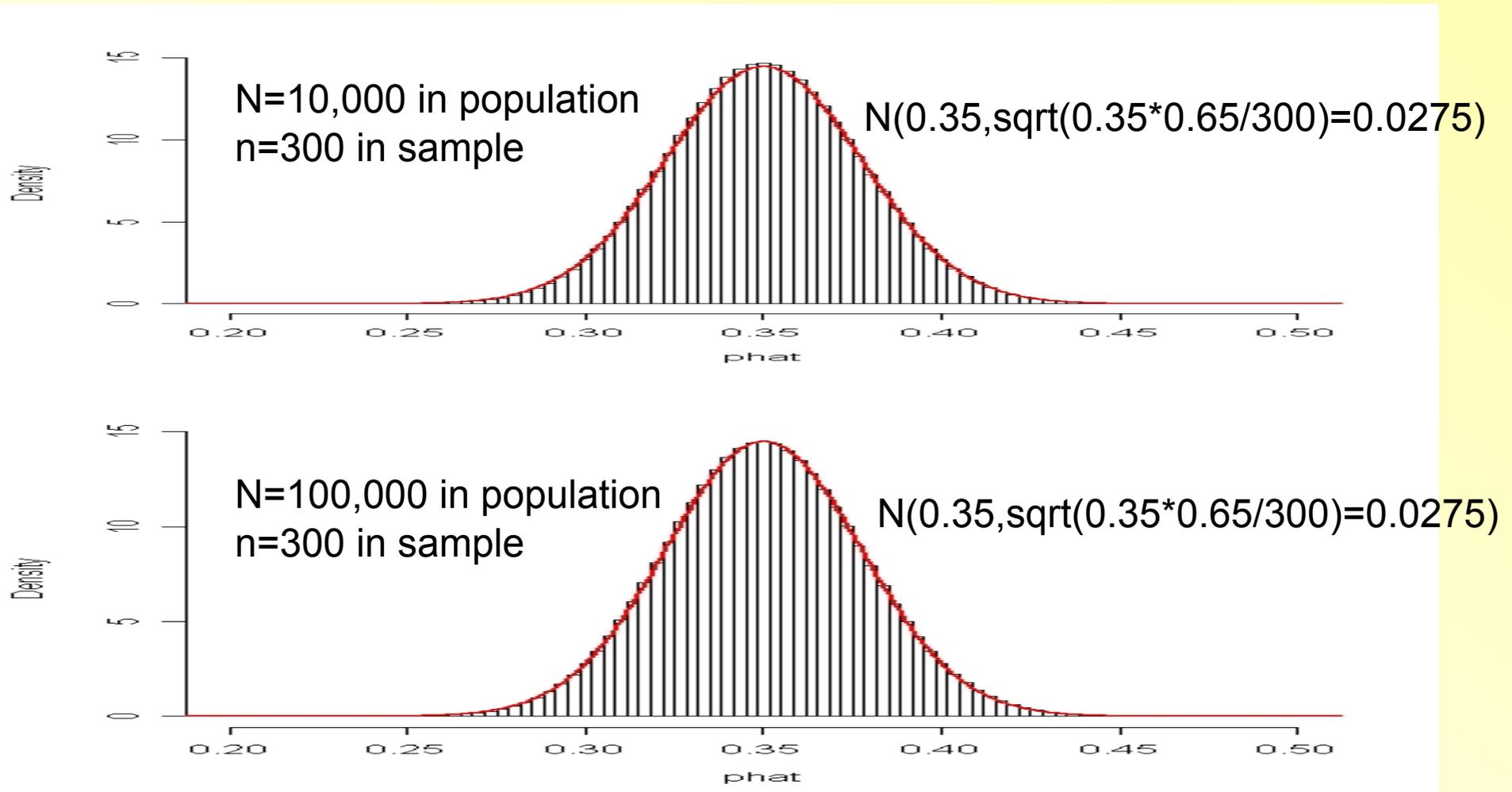
# What Matters, What Doesn't

- The center of the sampling distribution is the true proportion  $p$ .
- On average,  $\hat{p}$  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  $\hat{p}$ . In other words, the closer  $\hat{p}$  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 $N=100000$





# 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.



# Properties of the Sampling Distribution

- Expected Value of the  $\bar{X}$  's:  $\mu$ .

- Standard deviation of the  $\bar{X}$  's:  $\frac{\sigma}{\sqrt{n}}$   
also called the *standard error* of  $\bar{X}$

*For  $N/n < 20$ , use a finite population correction*

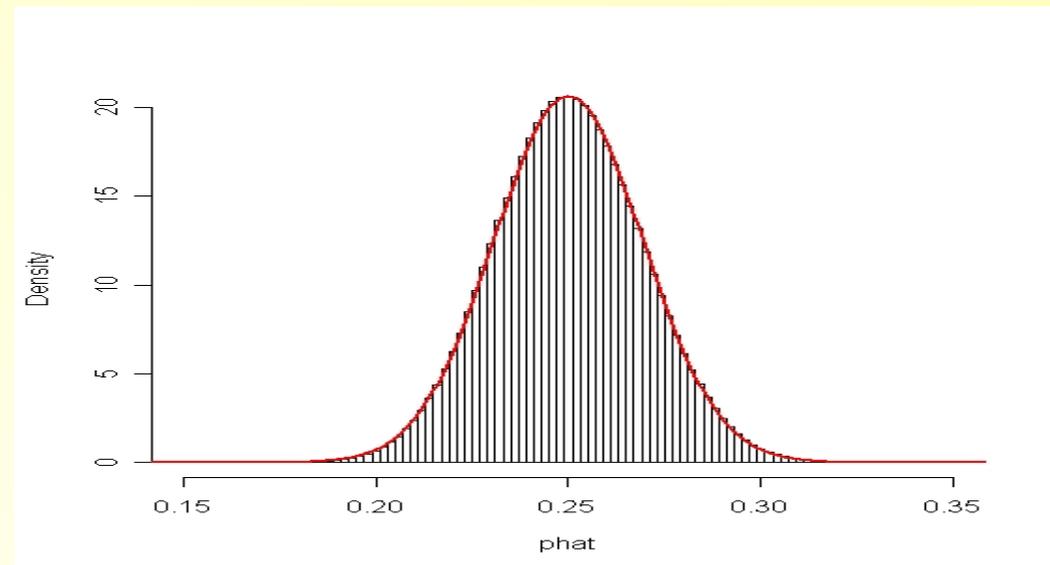
*factor for the standard deviation:*  $\sqrt{\frac{N-n}{N-1}}$

- ***Central Limit Theorem:*** As the sample size increases, the distribution of the  $\bar{X}$  '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  $\bar{Y}$  (and of the sample proportion  $\hat{p}$ ) 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  $\bar{Y}$  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

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

( $\mu$  = population mean,

$\sigma$  = 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  $\bar{Y}$ ,  $\mu$ ,  $\sigma$ , and  $n$  to calculate  $z$

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

## 3. z-Score for a sample proportion

- You need to know  $\hat{p}$ ,  $p$ , and  $n$  to calculate  $z$

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

# Example 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.*

# Example 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  $\bar{Y}$  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