STA 570 Spring 2011

Lecture 11 *Tuesday, Feb 22*

- **Sampling Distribution**

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 samples 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.

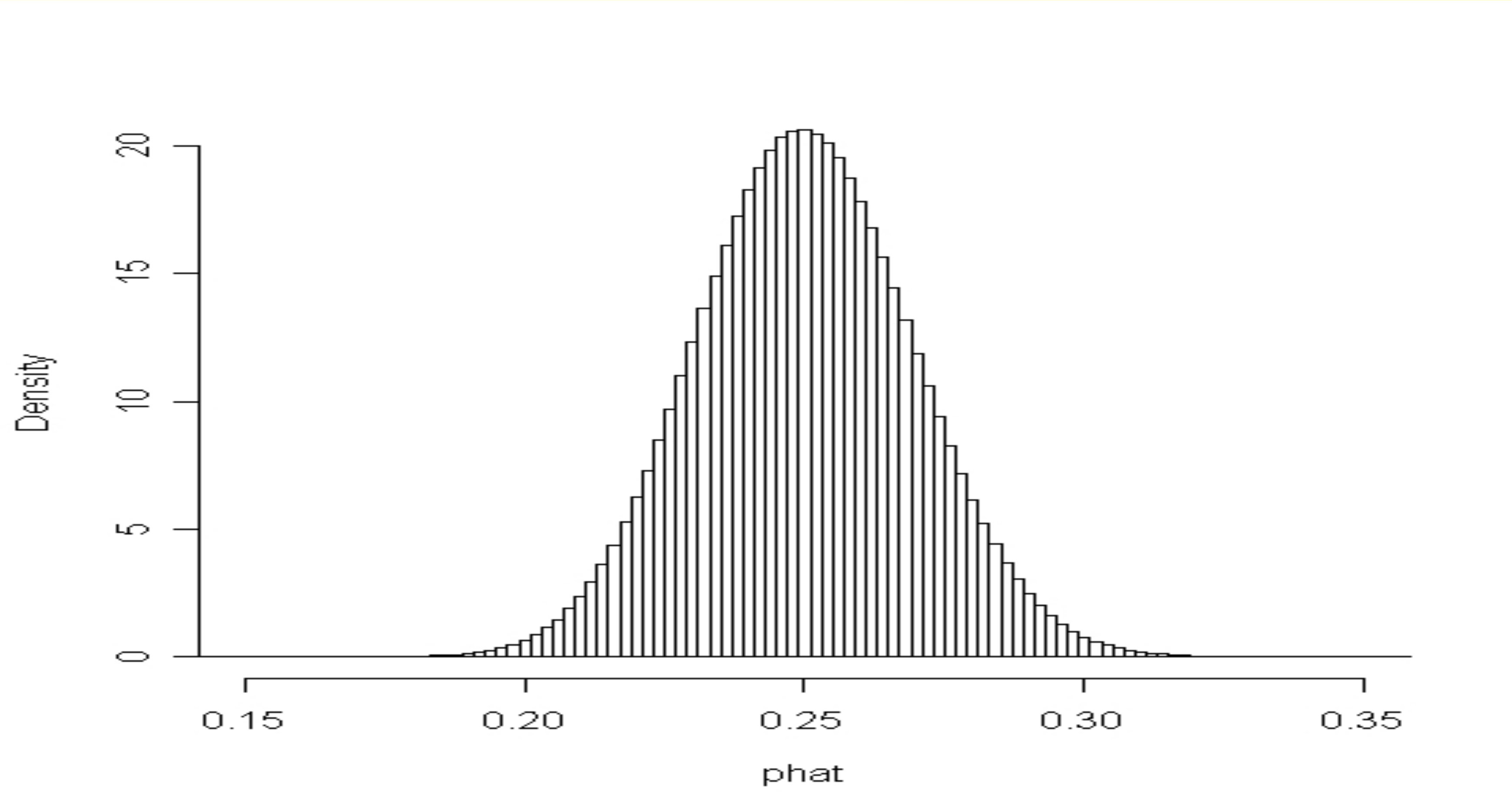
Example 3

- 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 phat=0.260, or our sample of 500 might yield 122 successes, resulting phat=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)*.

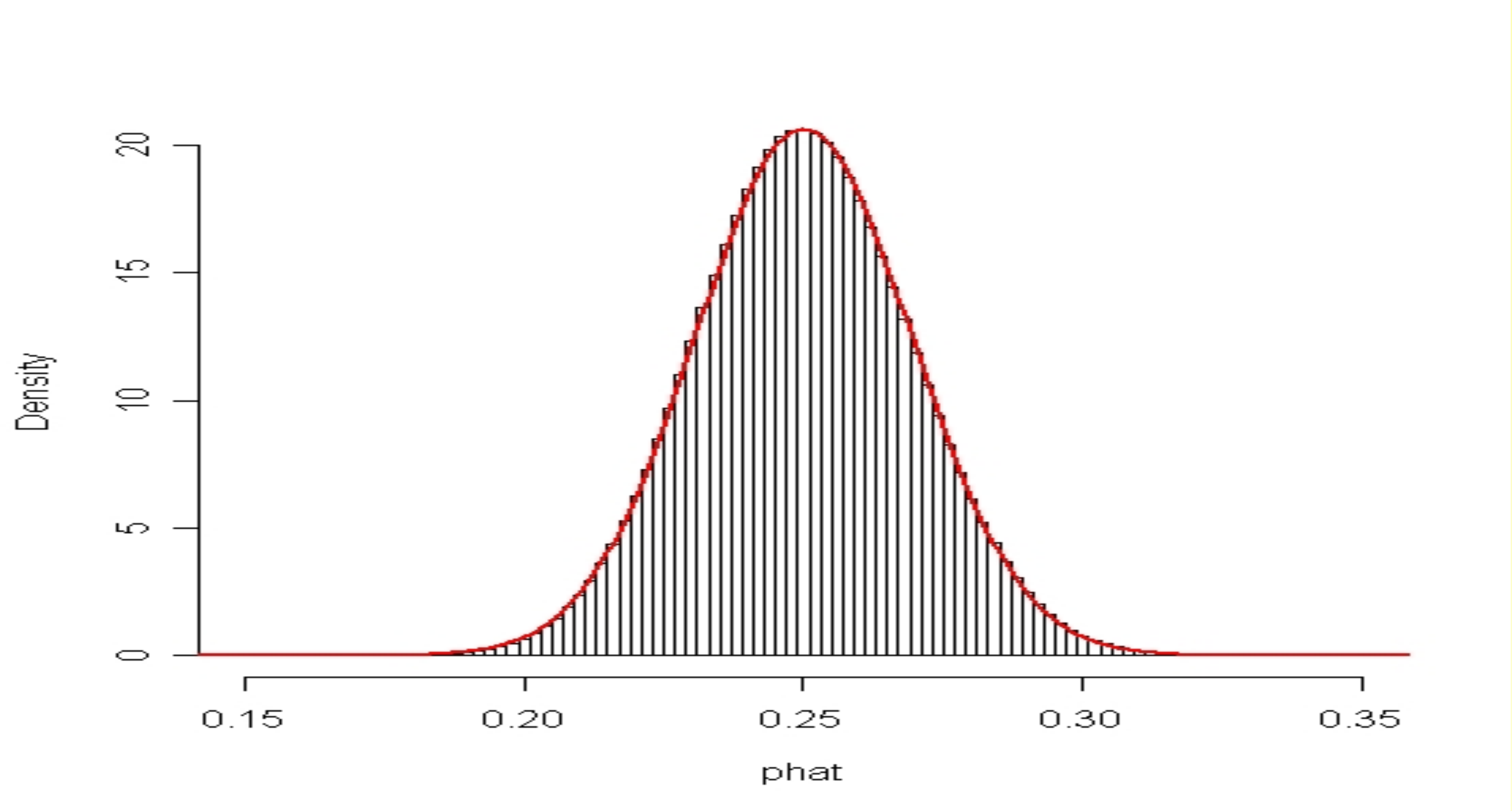


Hm?

- 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, overlayed with a perfect normal curve.



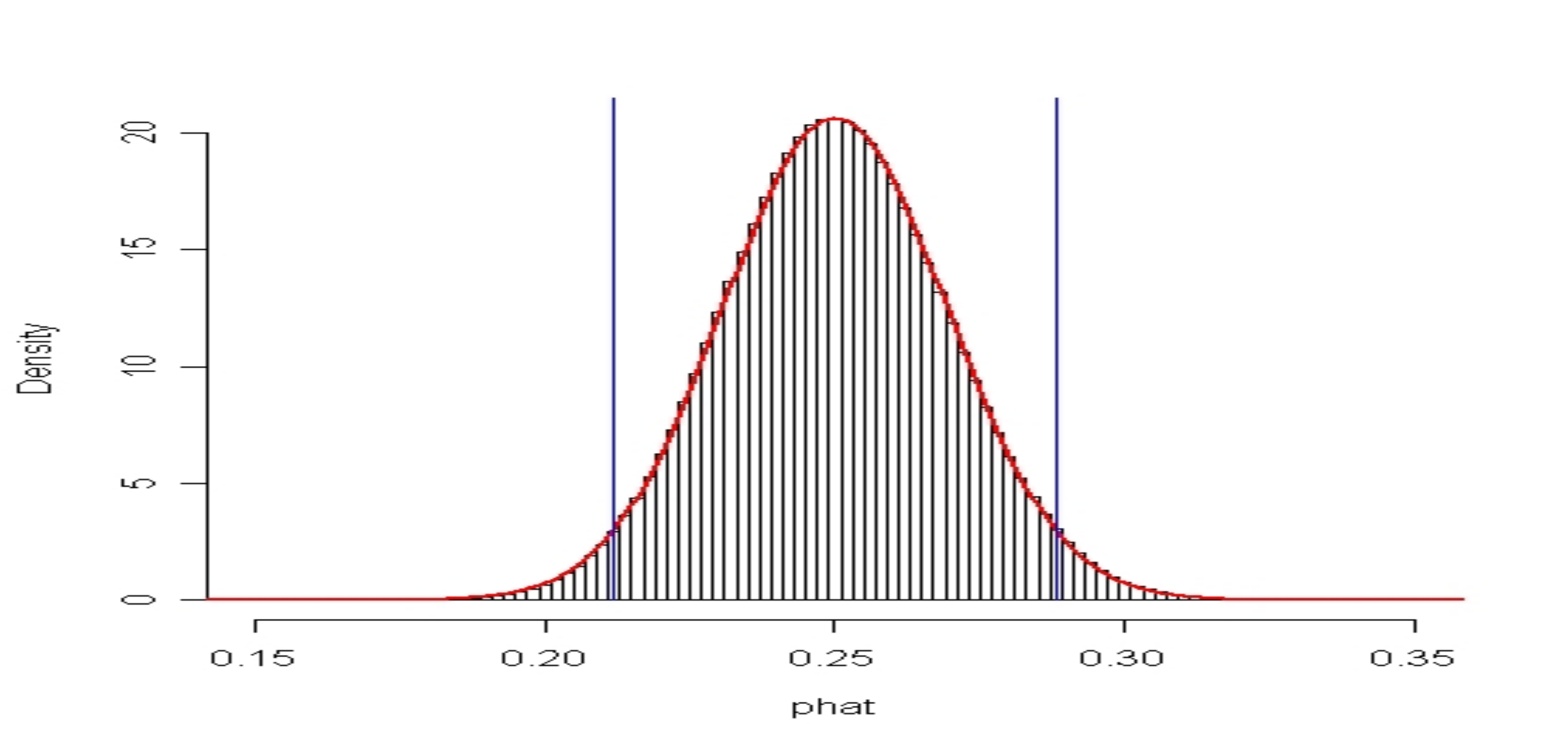
Review

- 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 visits example, we found that the pattern of variation of the sample proportions, called the ***sampling distribution***, followed a normal distribution.

Useful Consequences

- Example 3 (doctor visits): The sampling distribution of the sample proportion of successes is N(0.25,0.0194).
- Recall the 68-95-99.7 rule: About 95% probability that the sample proportion will be between 2 standard deviations (2*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.***

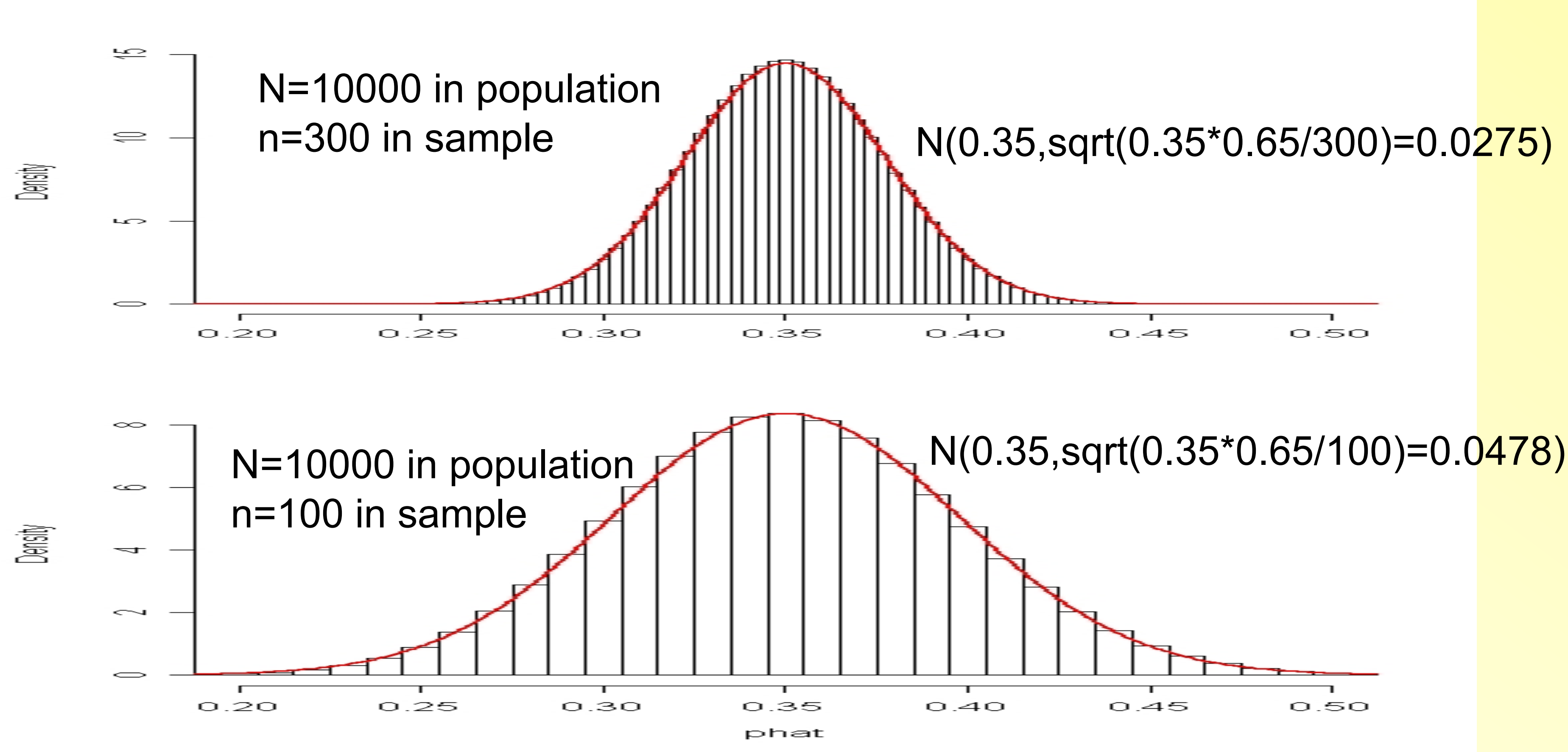
In Practice

- 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.

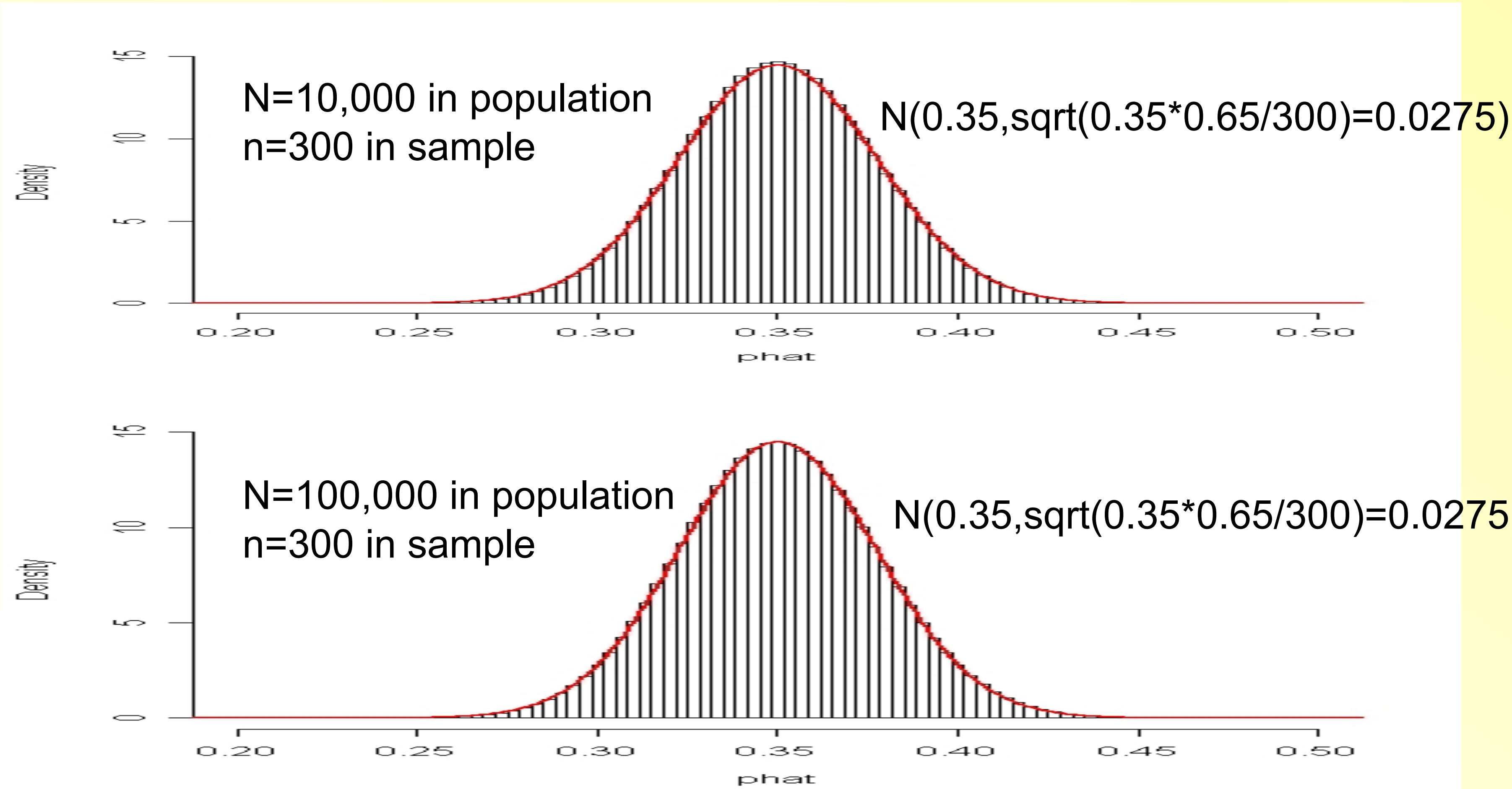
What Matters, What Doesn't

- The center of the sampling distribution is the true proportion p.
- On average, 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.

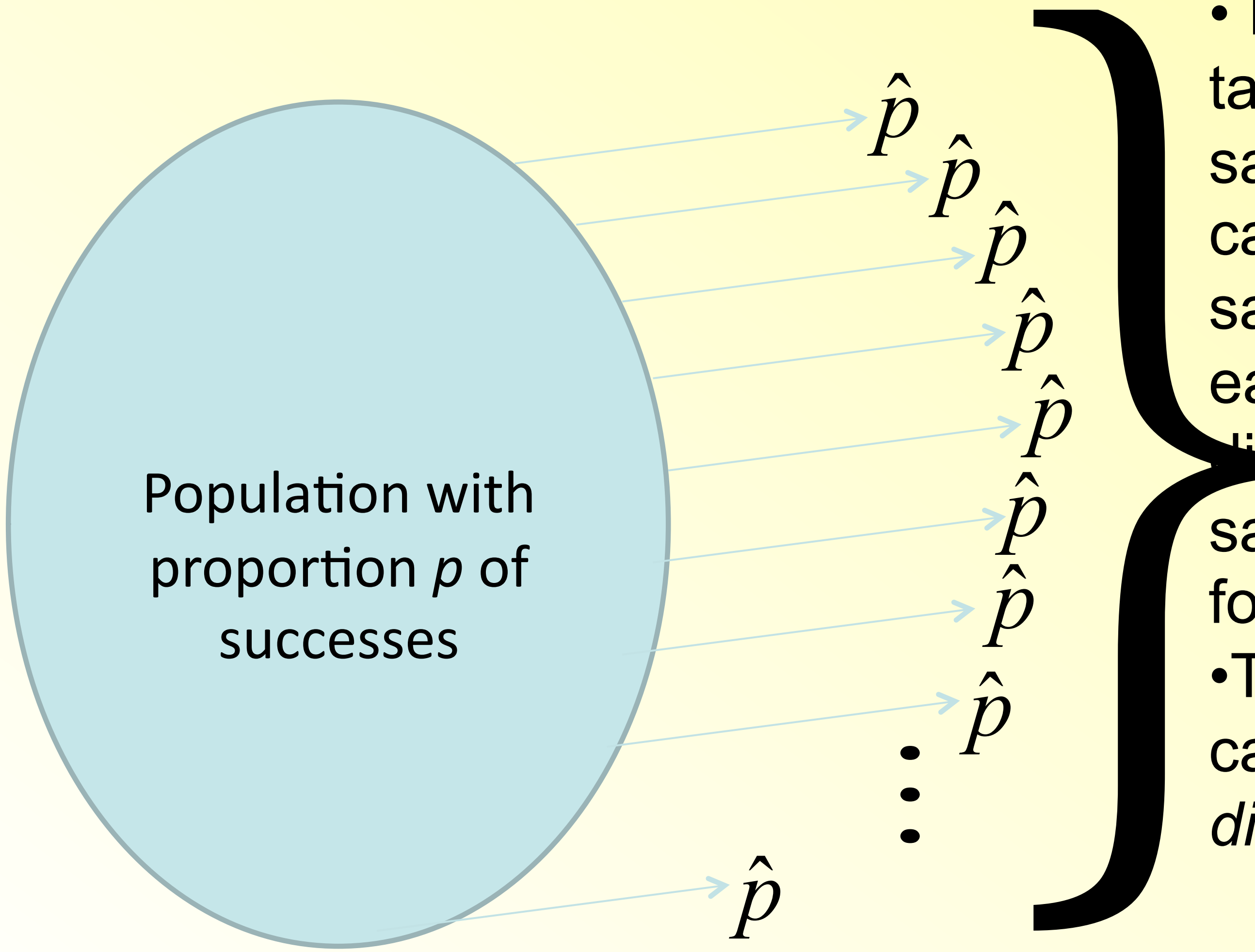
Population Size N=10000, 35% Successes
Comparing n=300 to n=100



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

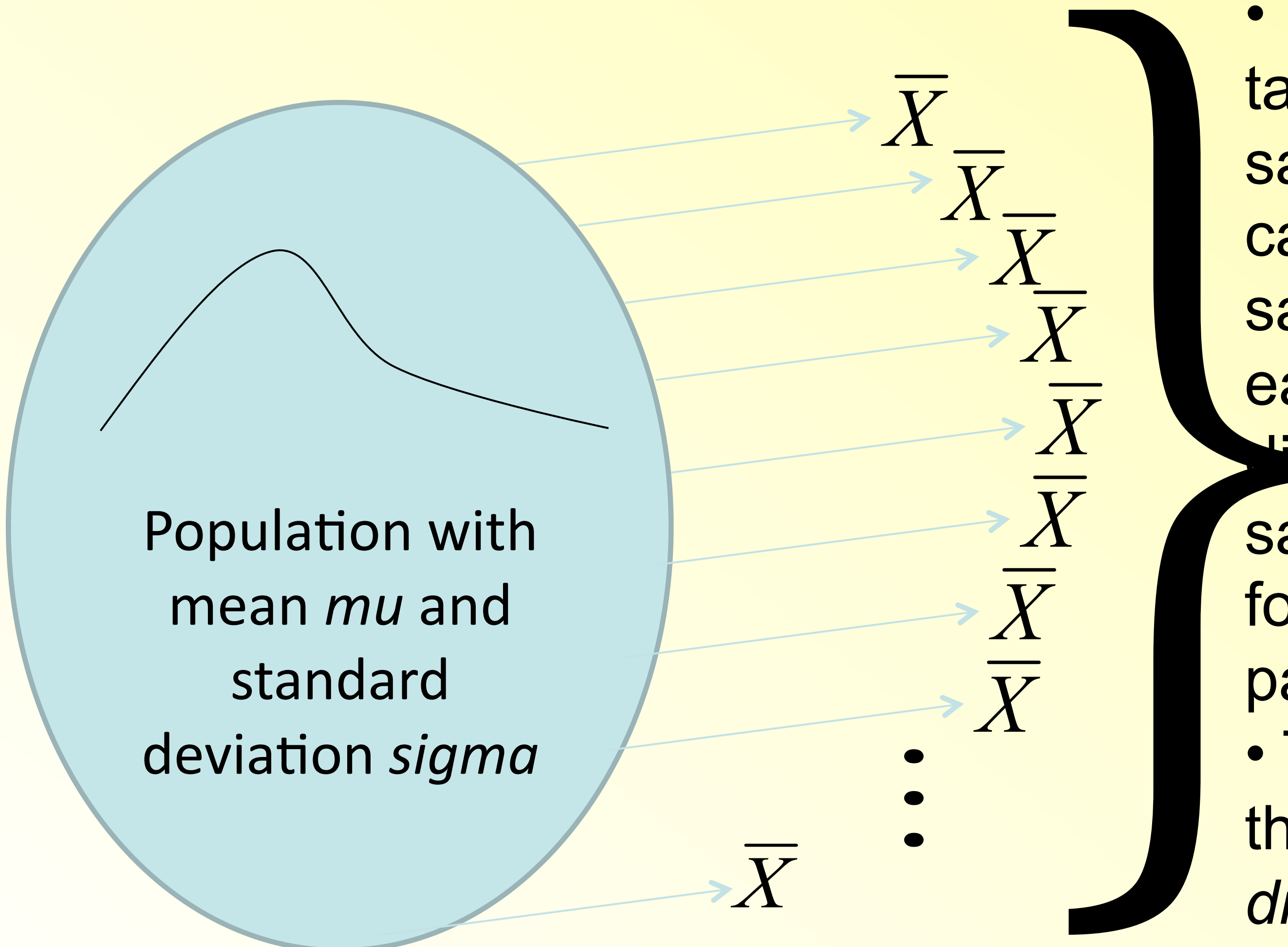
- Expected Value of the $\hat{p}$'s: p.

- Standard deviation of the $\hat{p}$'s: $\sqrt{\dfrac{p(1-p)}{n}}$

  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 $\bar{X}$'s: μ.

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

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

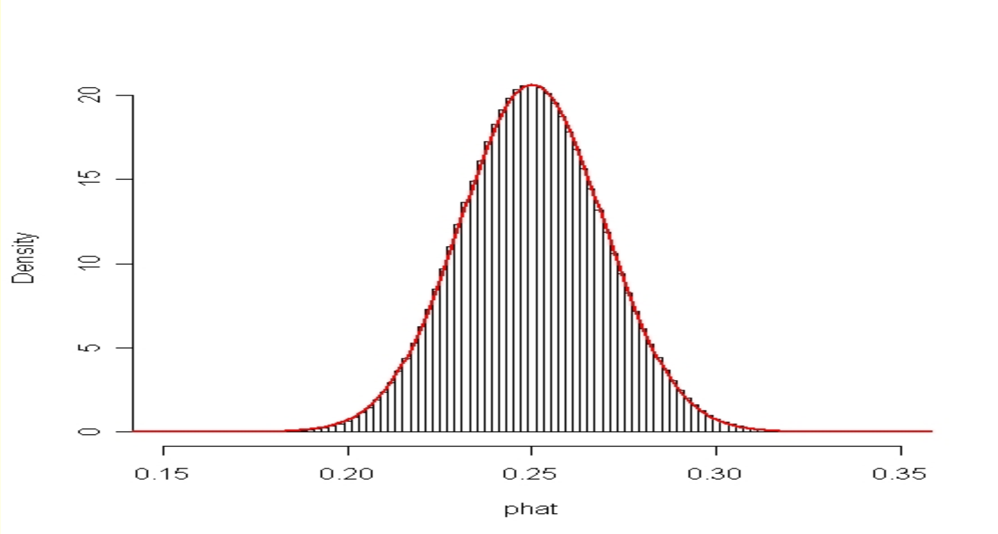
For N/n<20, use a finite population correction factor for the standard deviation: $\sqrt{\dfrac{N-n}{N-1}}$

- ***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 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 $\bar{Y}$ 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

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

(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*

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

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?

Example III: Acceptance Sampling

- Some companies monitor quality by using a method called acceptance sampling.
- An entire batch of items is rejected if a random sample of a particular size includes more than a specified number of defects.
- Assume that a company buys machine bolts in batches of 5000 and rejects the entire batch if, in a sample of 50, at least 2 defects (4% defects) are found.
- If the supplier manufactures bolts with a defect rate of 10%, what is the probability that a random batch will be rejected? How about the rejection rule "4 out of 100"?
- NB: *When we use the continuous normal distribution to approximate a discrete distribution such as "number of defects", a continuity correction should be made. That is, the single value x is represented by the interval from x-0.5 to x+0.5.*

QUIZ!!