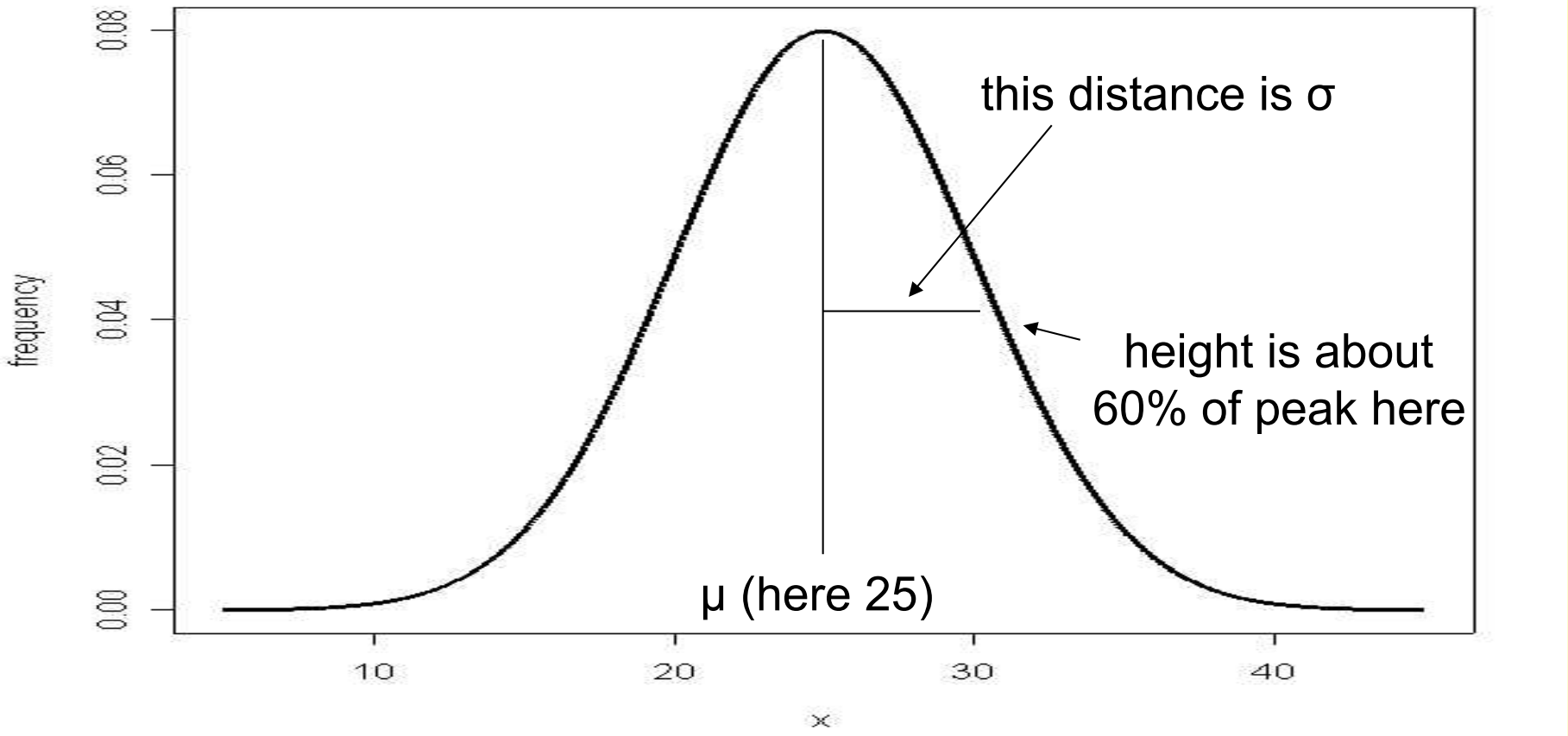
STA 321 Spring 2014

Lecture 11 *Tuesday, March 4*

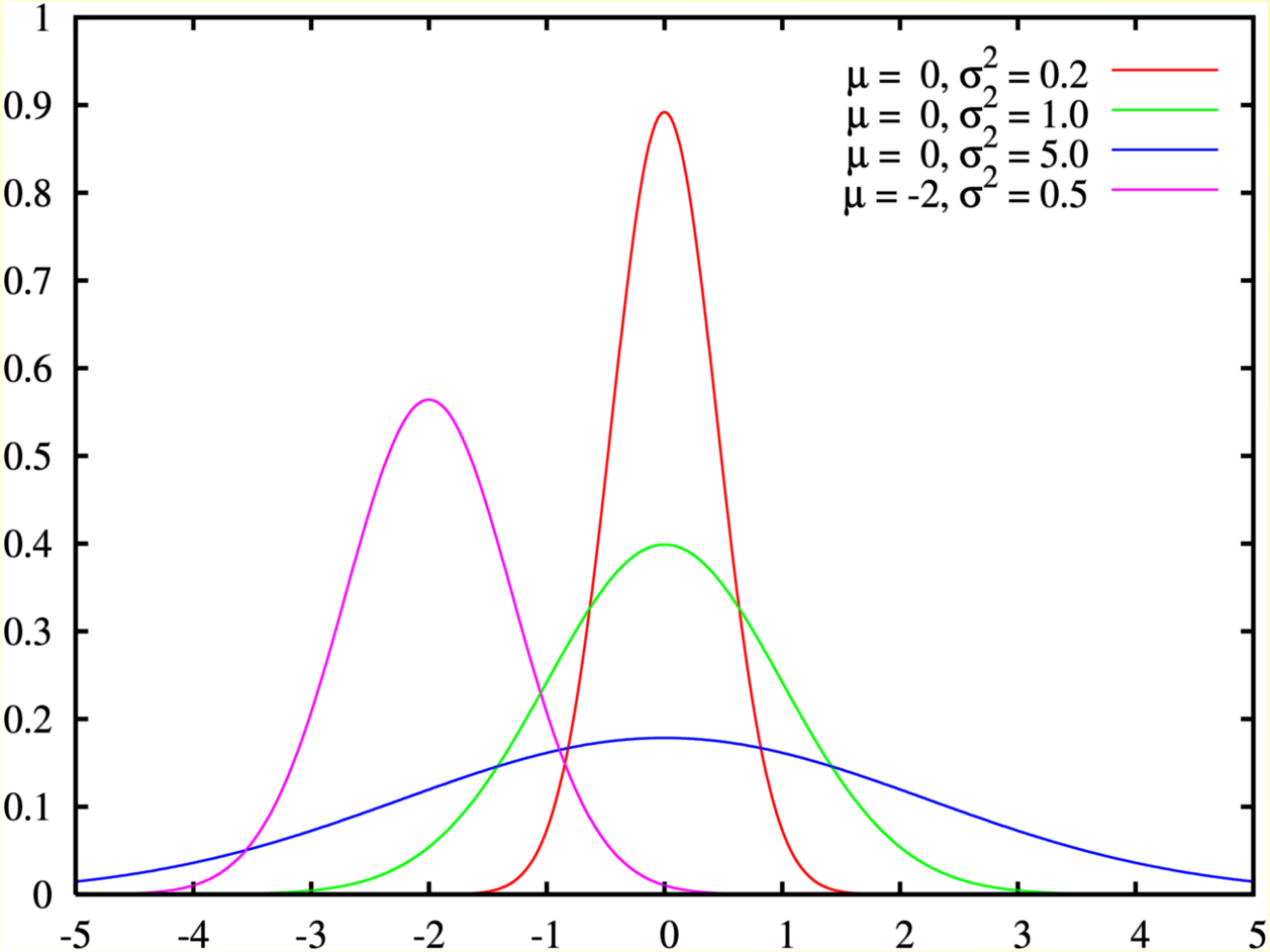
- **Normal Distribution**
- **z-Scores**

- **Sampling Distribution**

Mean (μ) and Standard deviation (σ) for a normal distribution



More on normal distribution



Quartiles of Normal Distributions

- Median: z=0

  (0 standard deviations above the mean)
- Upper Quartile: z = 0.67

  (0.67 standard deviations above the mean)
- Lower Quartile: z= – 0.67

  (0.67 standard deviations below the mean)
- Find the lower and upper quartile of cholesterol levels for men in the US

Another Example

Assume that cholesterol levels of men in the US have an approximately normal distribution with mean 215 (mg/dl) and standard deviation 25 (mg/dl).

http://bcs.whfreeman.com/scc/content/cat_040/spt/normalcurve/normalcurve.html

http://stat.utilities.googlepages.com/tables.htm

http://stattrek.com/Tables/Normal.aspx

z-Scores

- The *z*-score for a value x of a random variable is the number of standard deviations that x is above μ
- If x is below μ, then the *z*-score is negative
- The *z*-score is used to compare values from different normal distributions

Calculating *z*-Scores

- You need to know x, μ, and σ
  to calculate *z*

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

Tail Probabilities

- SAT Scores: Mean=500,
  Standard Deviation =100
- The SAT score 700 has a *z*-score of *z*=2
- The probability that a score is ***beyond*** 700 is the tail probability of *z*=2
- Online tool….
- 2.28% of the SAT scores are ***beyond*** 700 (***above*** 700)

Tail Probabilities

- SAT score 450 has a *z*-score of *z*=-0.5
- The probability that a score is ***beyond*** 450 is the tail probability of *z*=-0.5
- Online tool….
- 30.85% of the SAT scores are ***beyond*** 450 (***below*** 450)

z-Scores

- The *z*-score is used to compare values from different normal distributions
- SAT: μ=500, σ=100
- ACT: μ=21, σ=6
- What is better, 650 in the SAT or 28 in the ACT?

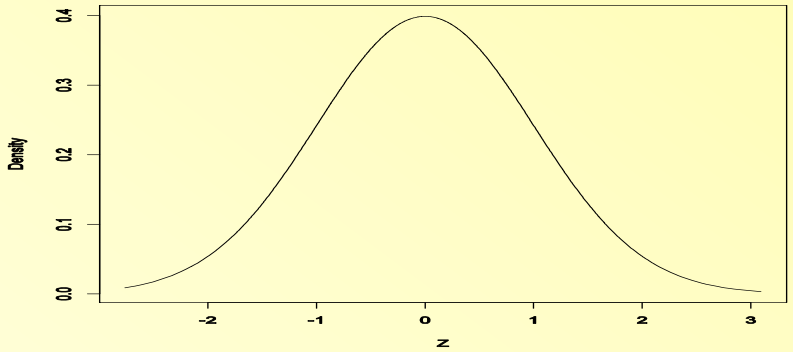
$$z_{SAT} = \frac{x - \mu}{\sigma} = \frac{650 - 500}{100} = 1.5$$

$$z_{ACT} = \frac{x - \mu}{\sigma} = \frac{28 - 21}{6} = 1.17$$

Corresponding tail probabilities? How many percent have better SAT or ACT scores?

Standard Normal Distribution

- The standard normal distribution is the normal distribution with mean $\mu=0$ and standard deviation $\sigma=1$



Standard Normal Distribution

- When values from an arbitrary normal distribution are converted to *z*-scores, then they have a standard normal distribution
- The conversion is done by subtracting the mean μ, and then dividing by the standard deviation σ

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

Example

- The scores on the Psychomotor Development Index (PDI) are approximately normally distributed with mean 100 and standard deviation 15. An infant is selected at random.
- Find the probability that the infant's PDI score is at least 100.
- Find the probability that PDI is between 97 and 103.
- Find the *z*-score for a PDI value of 90. Would you be surprised to observe a value of 90?
- Suppose we convert all the PDI observations to *z*-scores; that is, for each infant, subtract 100 from the value of PDI and divide by 15. Then, what is the distribution of the *z*-scores called? What are the mean and standard deviation of these *z*-scores?

Typical Questions

- One of the following three is given, and you are supposed to calculate one of the remaining
  1. Probability or percentage (right-hand, left-hand, two-sided, middle)
  2. z-score
  3. Observation *x,* original score
- In converting between 1 and 2, you need one of the online tools.
- In transforming between 2 and 3, you need mean and standard deviation and one of the following formulas

$$z = \frac{x - \mu}{\sigma} \qquad x = \mu + z\sigma$$

Note: Most of the time, mu and sigma are provided. If not, things can be a bit more tricky.

Online Tools

http://bcs.whfreeman.com/scc/content/cat_040/spt/normalcurve/normalcurve.html
http://stat.utilities.googlepages.com/tables.htm
http://stattrek.com/Tables/Normal.aspx

- Use these to
  - verify graphically the empirical rule,
  - find probabilities,
  - find percentiles
  - calculate *z*-values for one- and two-tailed probabilities

More Z-Score Examples

- IQ Scores: $\mu=100$ and $\sigma=15$
- An observation $X=125$ is 25 points above the mean, which corresponds to $25/15 = 1.67$ standard deviations above the mean.
- In general, a Z-score for an observation X is $\mathbf{Z=(X-\mu)/\sigma}$
- Observations above the mean get positive Z-scores, observations below the mean get negative Z-scores.

From Percentiles to Z-Scores

- What is the 80^{th} percentile of IQ Scores?
- In other words, for what IQ do 80% of the people fall below it?
- The first step is to find the Z-score corresponding to 80%
- That Z-score is Z=0.84

From Z-scores to IQ Scores

- The Z-score corresponding to 80% is 0.84
- The Z-score formulas are

$Z=(X-\mu)/\sigma$ and $X = \mu + \sigma Z$

- $Z=0.84$, $\mu=100$, and $\sigma=15$, so
- $X = 100+(0.84)(15) = 112.6$
- So 80% of people have an IQ score below 112.6

Middle Percentages

- What about the middle 50% of IQ scores?
- What percentiles does this correspond to? To get the middle 50%, we need to stretch from the 25^{th} percentile to the 75^{th} percentile.
- The 25^{th} percentile corresponds to a Z-score of _____

  the 75^{th} percentile corresponds to a Z-score of _____
- These correspond to IQ's of _____ and _____
- What about the middle 95% of values (removing 2.5% from each tail)
- Finally, the middle 99% of IQ scores?

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**.

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, …, 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 321) 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%
3 (60% yes)	**6160**	**39.73%**
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.
- How close… probability theory will tell us that, too…but we'll have to wait to find out.

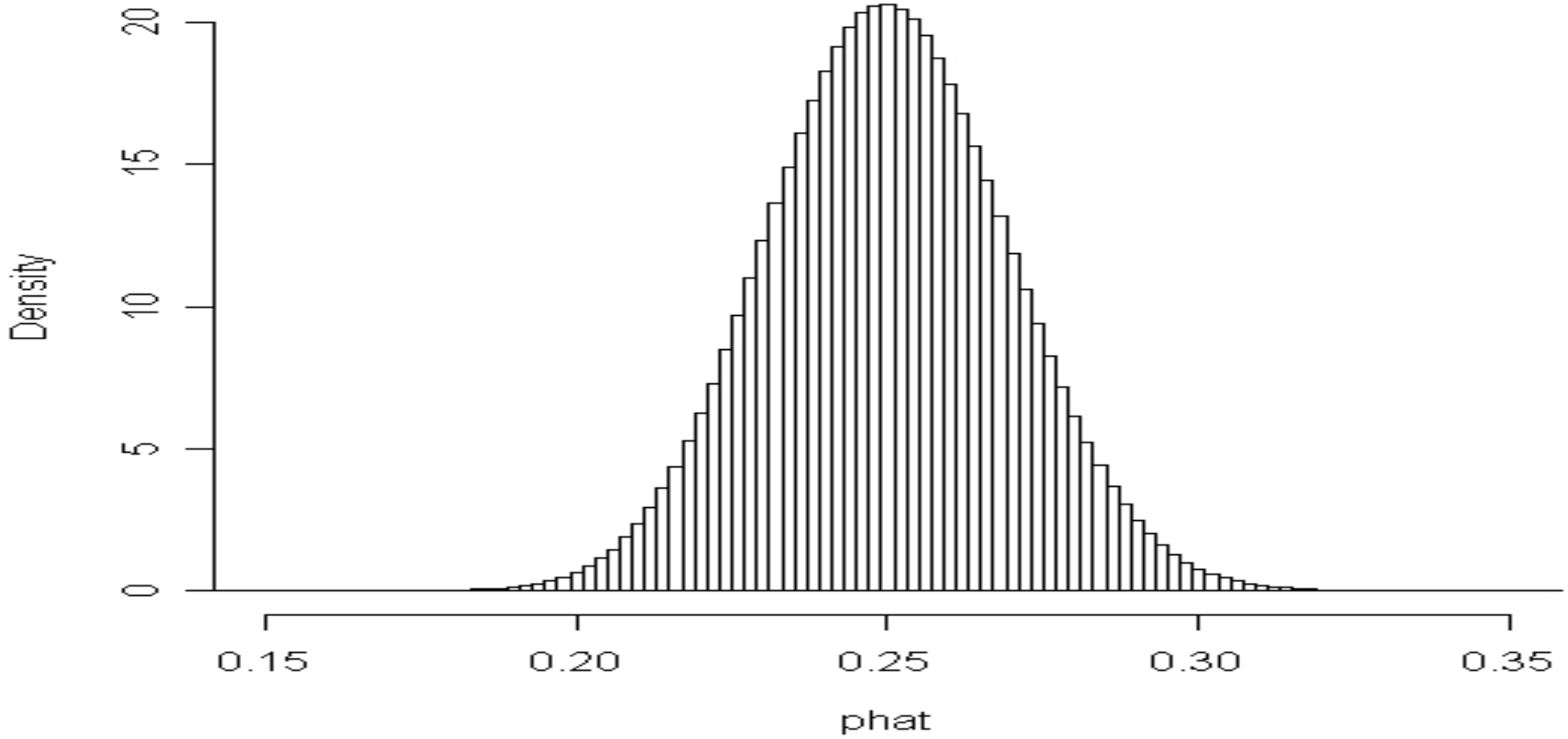
Example 3

- 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)*.

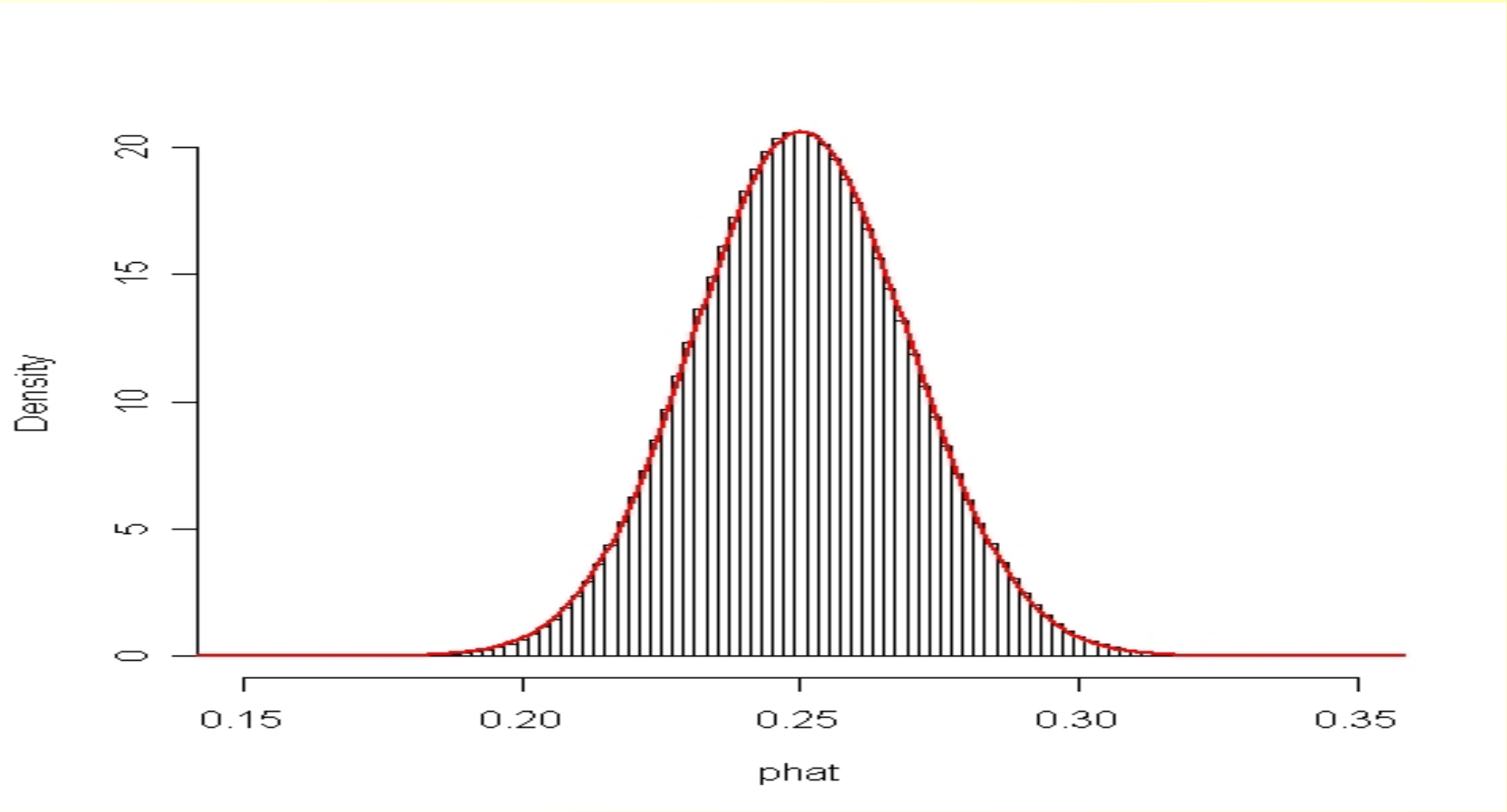


Hm?

- That looks like a bell curve.
- In fact, it looks suspiciously like a bell curve with mean μ=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.



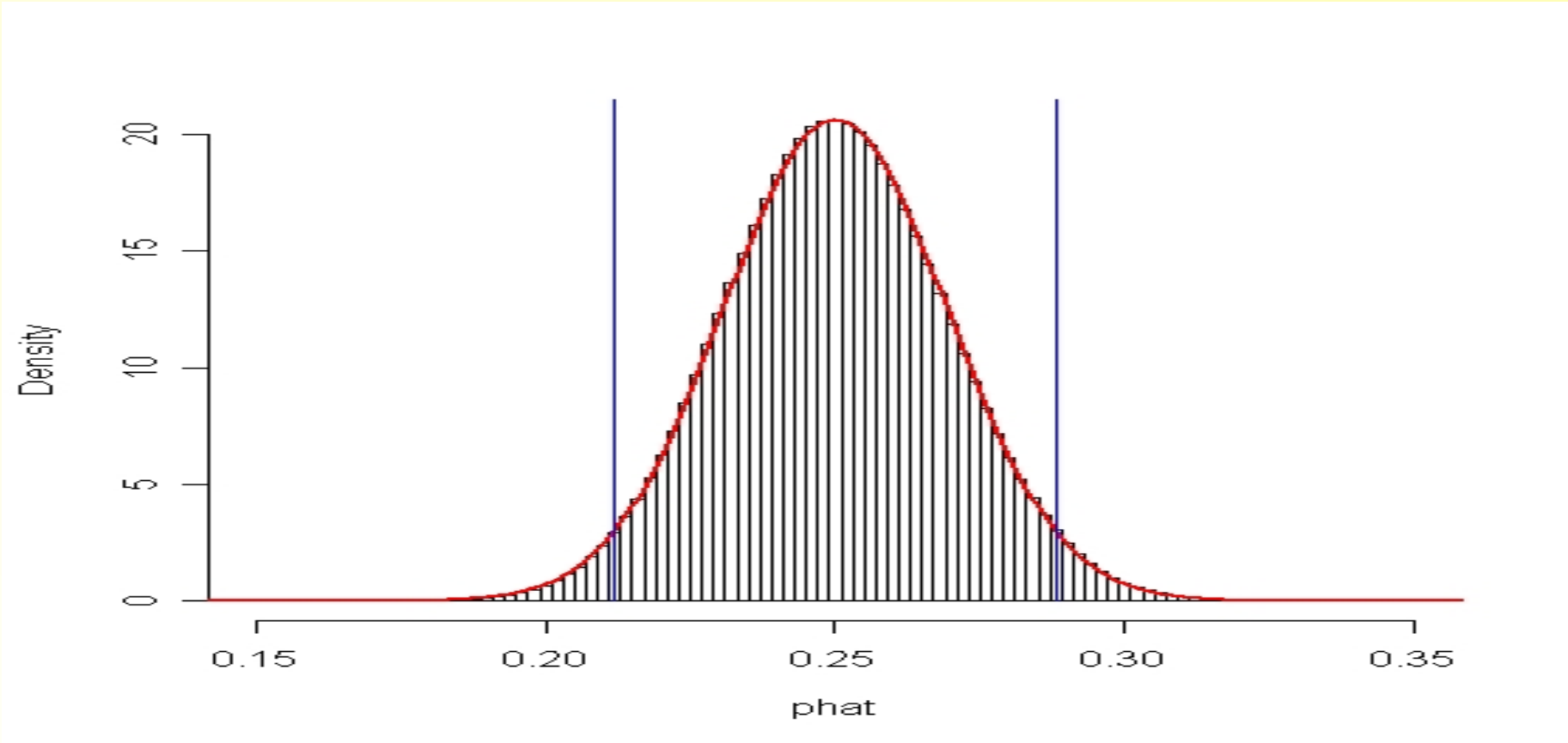
Review

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

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*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 95.5%.

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.***

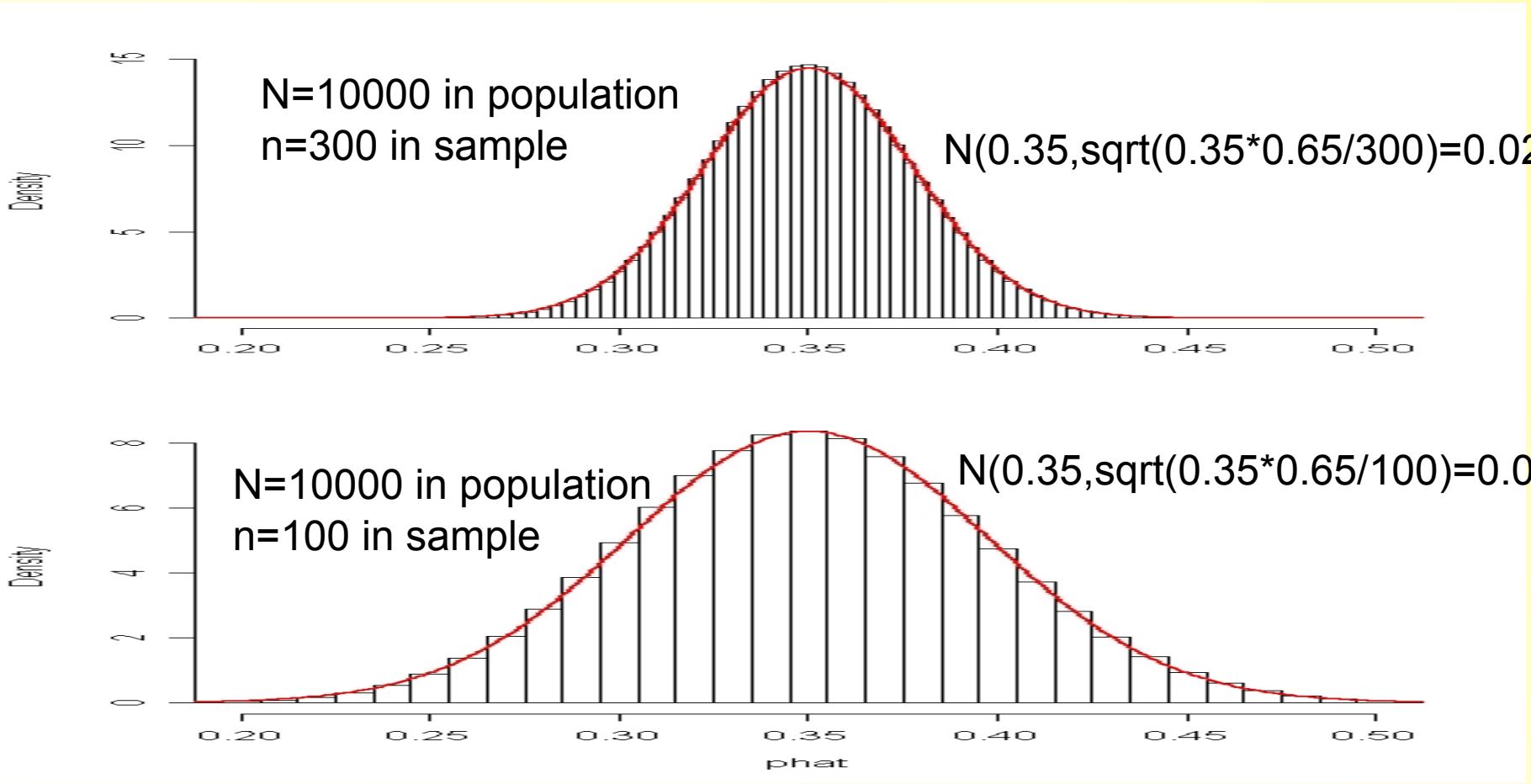
In Practice

- 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.

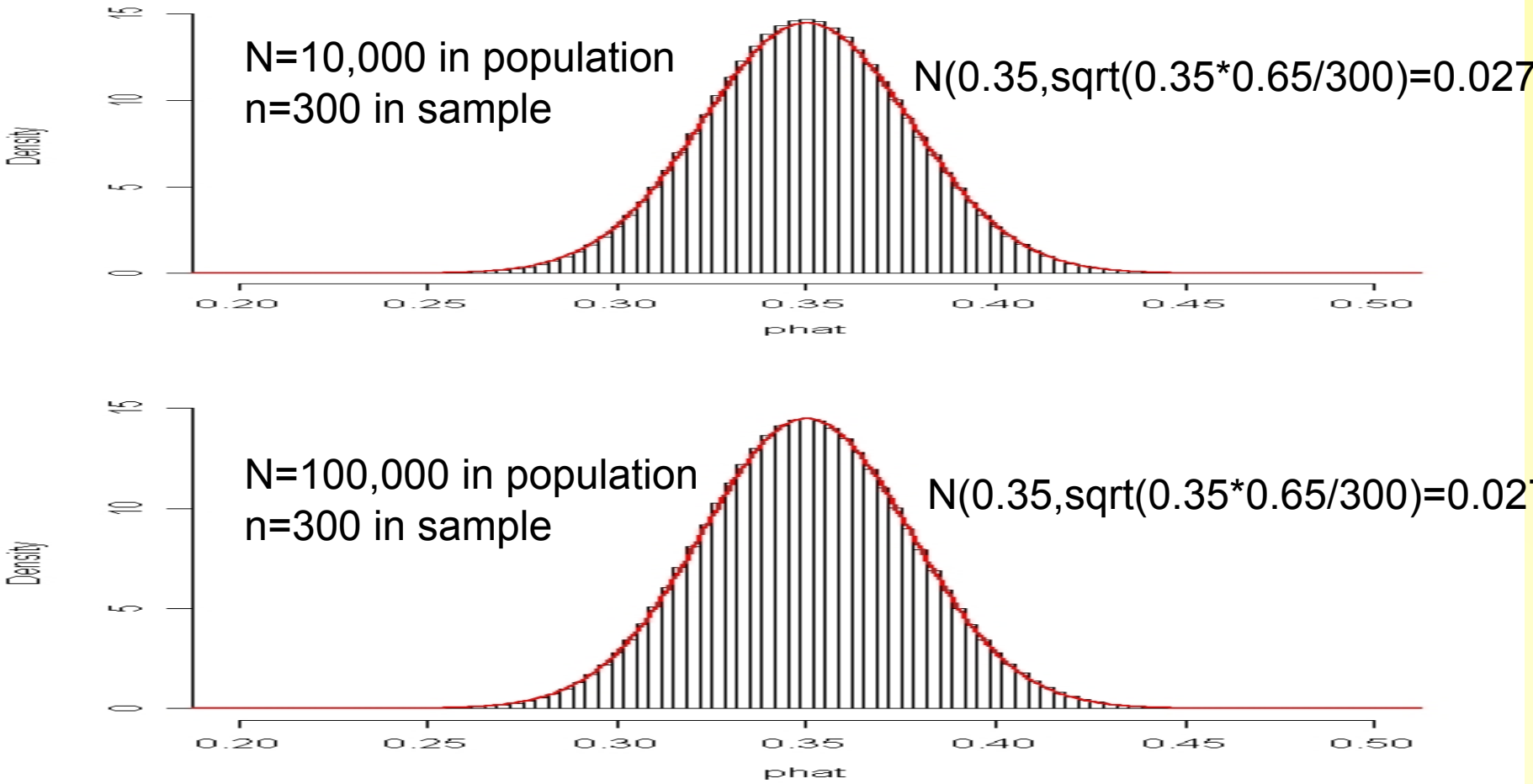
What Matters, What Doesn't

- 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.

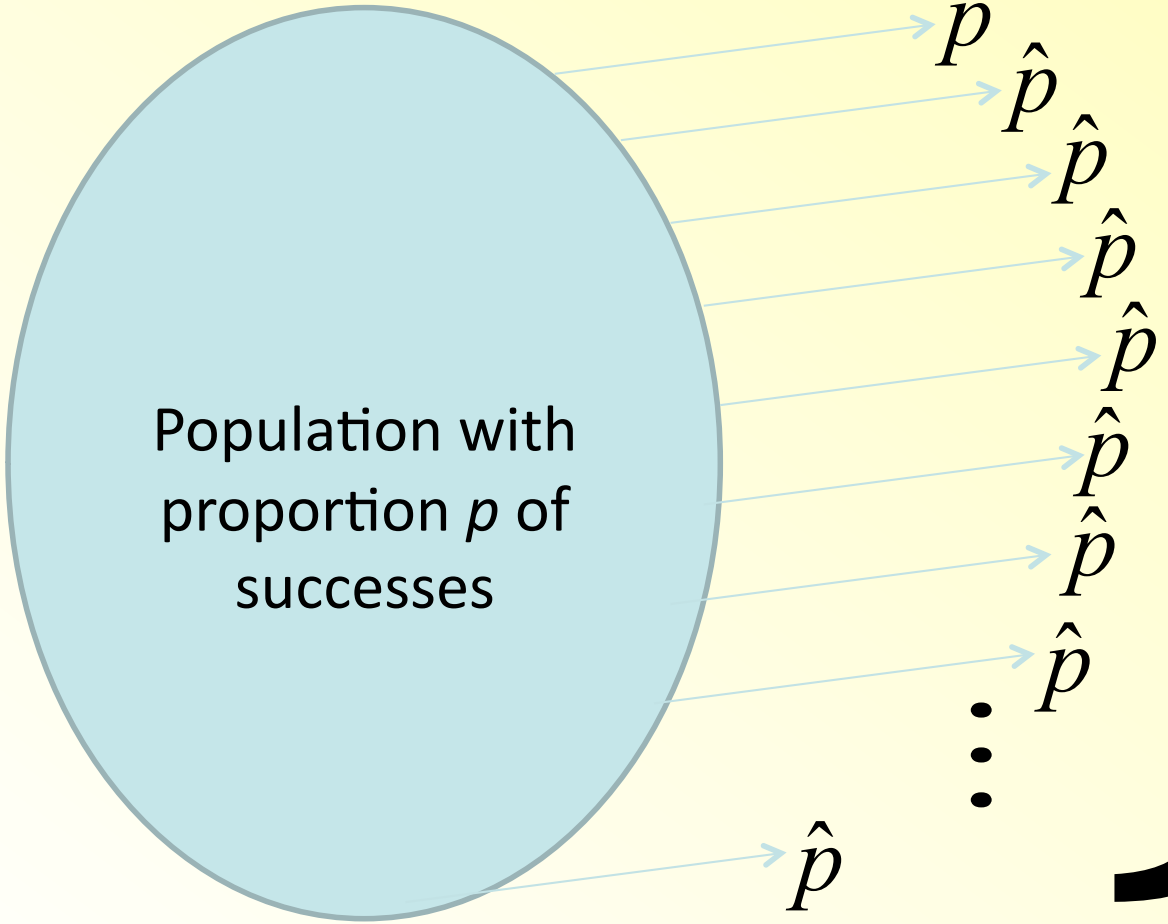
Population Size N=10000, 35% Successes Comparing n=300 to n=100



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



Summary: Sampling Distribution



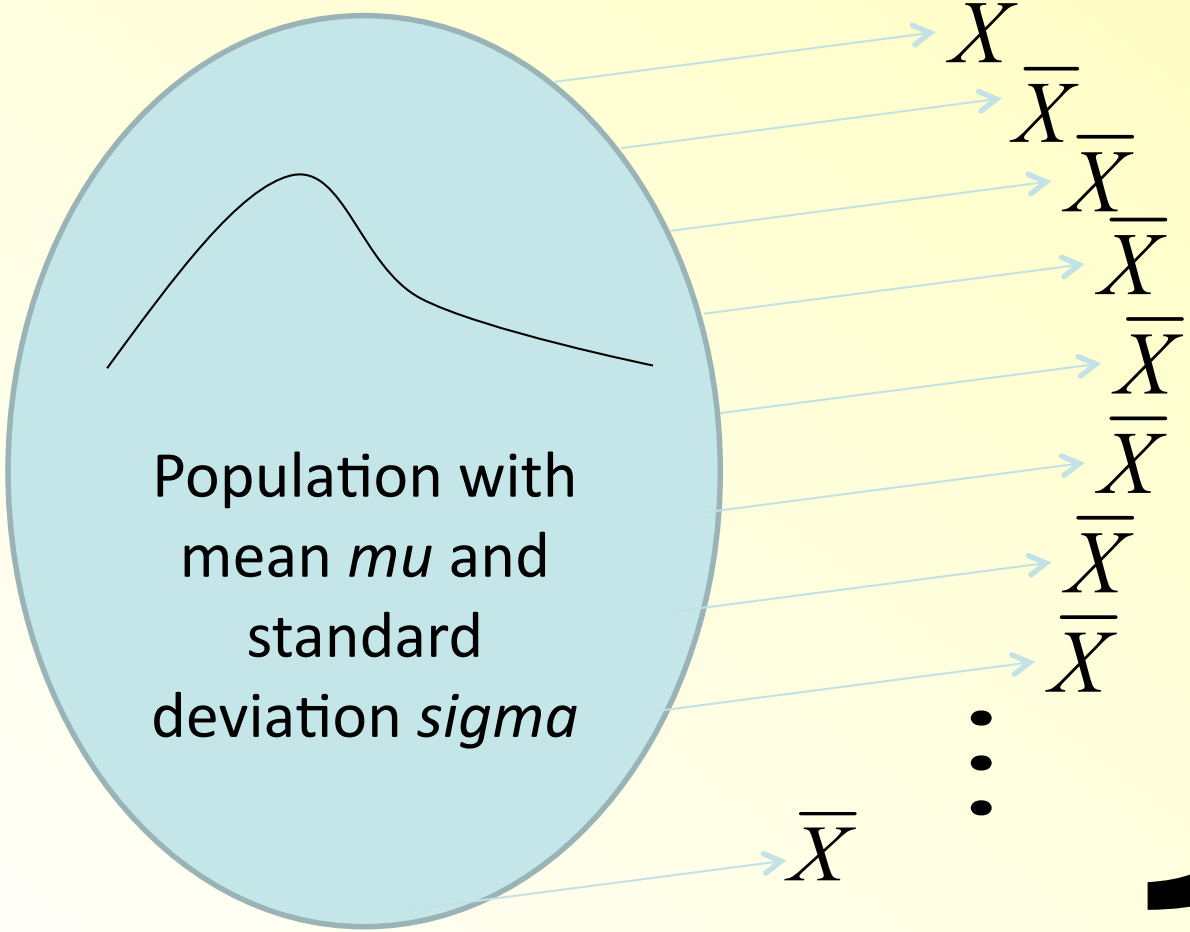
- If you repeatedly take random samples and calculate the sample proportion each time, the distribution of the sample proportions follows a pattern
- This pattern is called the *sampling distribution of p-hat*

Properties of the Sampling Distribution

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

  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.