Announcements:

- Midterm 2 (next Friday) will cover Chapters 7 to 10, except a few sections. (See Mar 1 on webpage.) Finish new material Monday.
- Two sheets of notes are allowed, same rules as for the one sheet last time.

Homework (due *Wed*, Feb 27):

Chapter 9:

#48, 54 (counts double)

Chapter 9: Sections 4, 5, 9

Sampling Distributions for Proportions:
One proportion or difference in two proportions

Review: Statistics and Parameters

A **statistic** is a numerical value computed from a sample. Its value may differ for different samples. *e.g. sample mean* $\bar{x}$*, sample standard deviation* s*, and sample proportion* $\hat{p}$*.*

A **parameter** is a numerical value associated with a population. Considered fixed and unchanging. *e.g. population mean* μ*, population standard deviation* σ*, and population proportion* p*.*

Review: Sampling Distributions

Statistics as Random Variables: Each new sample taken => sample statistic will change.

The distribution of possible values of a statistic for repeated samples of the same size is called the **sampling distribution** *of the statistic*.
Equivalently: The probability density function (pdf) of a sample statistic is called the **sampling distribution** for that statistic.

Many statistics of interest have sampling distributions that are ***approximately normal*** distributions

Review: Sampling Distribution for a Sample Proportion

Let p = population proportion of interest
or binomial probability of success.

Let $\hat{p}$ = sample proportion or proportion of successes.

If numerous random samples or repetitions of the same size n are taken, the distribution of possible values of $\hat{p}$ is **approximately** a **normal** curve distribution with

- **Mean** = p
- **Standard deviation** = s.d.($\hat{p}$) = $\sqrt{\frac{p(1-p)}{n}}$

This approximate distribution is **sampling distribution of** $\hat{p}$ **.**

Examples for which this applies

- ***Polls*:** to estimate proportion of voters who favor a candidate; population of units = all voters.
- ***Television Ratings*:** to estimate proportion of households watching TV program; population of units = all households with TV.
- ***Genetics*:** to estimate proportion who carry the gene for a disease; population of units = everyone.
- ***Consumer Preferences*:** to estimate proportion of consumers who prefer new recipe compared with old; population of units = all consumers.
- ***Testing ESP*:** to estimate probability a person can successfully guess which of 4 symbols on a hidden card; repeatable situation = a guess.

Example: Sampling Distribution for a Sample Proportion

- Suppose (unknown to us) **40% of a population carry the gene** for a disease (p = 0.40).
- We will take a **random sample of 25** people from this population and count ***X* = number with gene**.
- Although we ***expect*** to find 40% (10 people) with the gene ***on average***, we know the number will ***vary*** for different samples of n = 25.
- X is a **binomial random variable** with

  n = 25 and p = 0.4.
- We are interested in $\hat{p} = \dfrac{X}{n}$

Many Possible Samples, Many $\hat{p}$

Four possible random samples of 25 people:

Sample 1: X =12, proportion with gene =12/25 = 0.48 or 48%.

Sample 2: X = 9, proportion with gene = 9/25 = 0.36 or 36%.

Sample 3: X = 10, proportion with gene = 10/25 = 0.40 or 40%.

Sample 4: X = 7, proportion with gene = 7/25 = 0.28 or 28%.

Note:

- Each sample gave a different answer, which did not always match the population value of p = 0.40 (40%).
- Although we cannot determine whether one sample statistic will accurately estimate the true population parameter, the *sampling distribution* gives probabilities for how far from the truth the sample values could be.

Sampling Distribution for this Sample Proportion

Let p = population proportion of interest
or binomial probability of success = .40

Let $\hat{p}$ = sample proportion or proportion of successes.

If numerous random samples or repetitions of the same size n are taken, the distribution of possible values of $\hat{p}$ is **approximately** a **normal** curve distribution with

- **Mean** = $p = .40$
- **Standard deviation** = s.d.($\hat{p}$) = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{.40(1-.40)}{25}}$

$$= .098 \approx .10$$

This approximate distribution is **sampling distribution of** $\hat{p}$ **.**

Approximately Normal Sampling Distribution for Sample Proportions

Normal Approximation can be applied in ***two situations***:

Situation 1: A random sample is taken from a large population.

Situation 2: A binomial experiment is repeated numerous times.

In each situation, ***three conditions*** must be met:

Condition 1: *The Physical Situation*
There is an actual population or repeatable situation.

Condition 2: *Data Collection*
A random sample is obtained or situation repeated many times.

Condition 3: *The Size of the Sample or Number of Trials*
The size of the sample or number of repetitions is relatively large, np and $n(1-p)$ must be at least 10.

Finishing a few slides from last time….

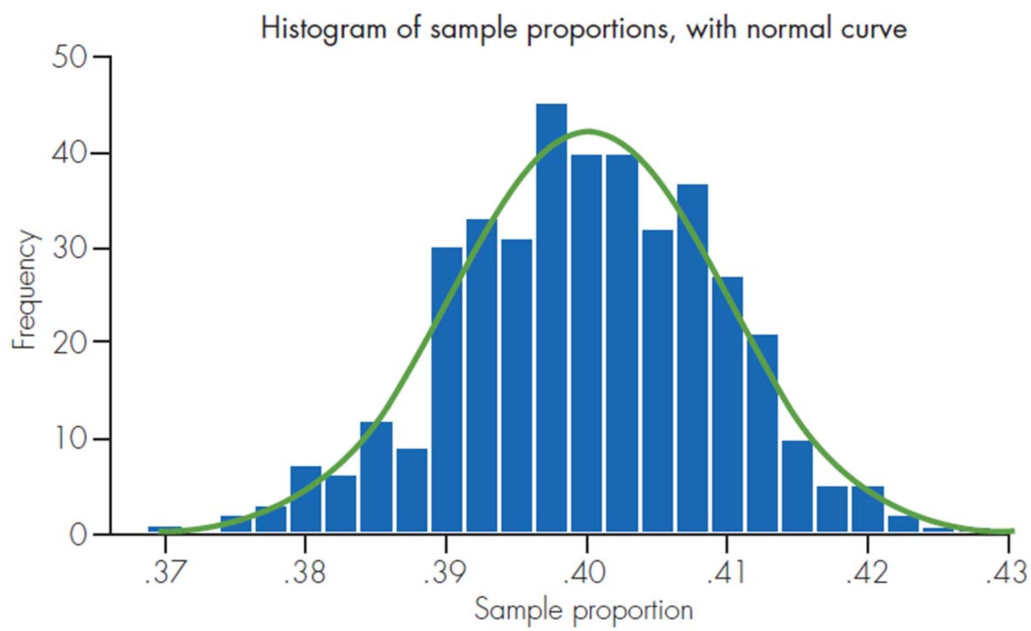
Example 9.4 *Possible Sample Proportions Favoring a Candidate*

Suppose 40% all voters favor Candidate C. Pollsters take a sample of n = 2400 voters. The sample proportion who favor C will have approximately a normal distribution with

$$\text{mean} = p = 0.4 \text{ and s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(1-0.4)}{2400}} = 0.01$$

Histogram at right shows sample proportions resulting from simulating this situation 400 times.

Empirical Rule: Expect
68% from .39 to .41
95% from .38 to .42
99.7% from .37 to .43



A Dilemma and What to Do about It

In practice, we don't know the true population proportion p, so we cannot compute the **standard deviation** of $\hat{p}$,

$$\text{s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} .$$

Replacing p with $\hat{p}$ in the standard deviation expression gives us an estimate that is called the **standard error of** $\hat{p}$.

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} .$$

The *standard error* is an excellent approximation for the *standard deviation*. We will use it to find *confidence intervals,* but will <u>not</u> need it for sampling distribution or hypothesis tests because we <u>assume</u> a specific value for p in those cases.

CI Estimate of the Population Proportion from a Single Sample Proportion

CBS Poll taken this month asked "*In general, do you think gun control laws should be made more strict, less strict, or kept as they are now?*

Poll based on n = 1,148 adults, 53% said "more strict."

Population parameter is p = proportion of *population* that thinks they should be more strict.

Sample statistic is $\hat{p} = .53$

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{.53(.47)}{1148}} = .015$$

If $\hat{p} = 0.53$ and $n = 1148$, the standard error of the sampling distribution of $\hat{p}$ is 0.015. So *two* standard errors is .03.

The sample value of $\hat{p} = .53$ is 95% certain to be within 2 standard errors of population p, so p is probably between .50 and .56.

Another Example

Suppose 60% of seniors who get flu shots remain healthy, independent from one person to the next. A senior apartment complex has 200 residents and they all get flu shots. What proportion will remain healthy?

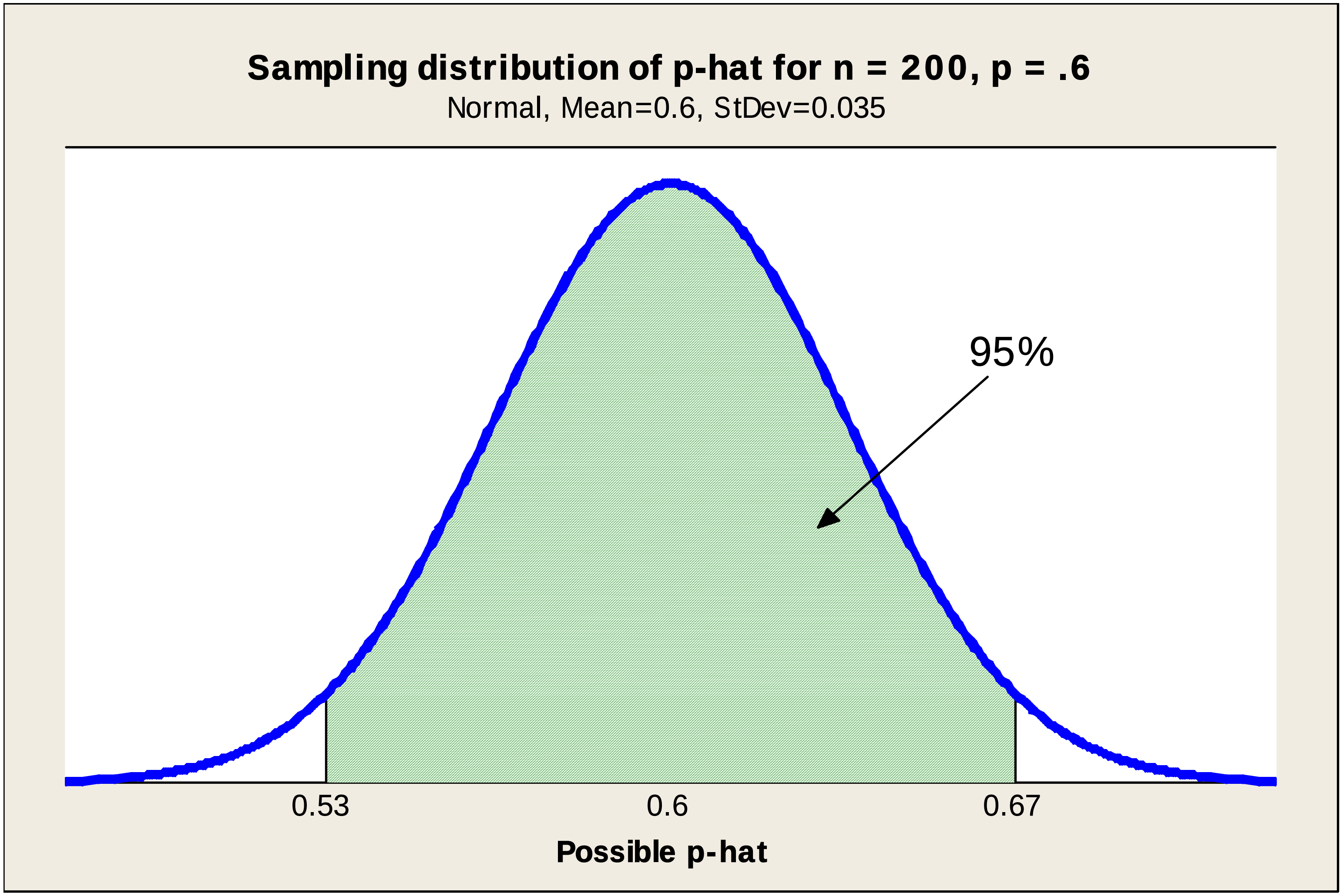
Population of *all* seniors has $p = 0.60$

Sample has $n = 200$ people.

$\hat{p}$ = proportion of *sample* with no flu. Possible values?

Sampling distribution for $\hat{p}$ is:

- Approximately normal
- Mean = p = .60
- Standard deviation of $\hat{p} = \sqrt{\frac{(.4)(.6)}{200}} = .035$



From Empirical Rule, expect 95% of samples to produce $\hat{p}$ to be in the interval mean ± 2s.d. or .60 ± 2(.035) or .60 ± .07 or .53 to .67. So, expect 53% to 67% of residents to stay healthy.

Example: Belief in evolution

Gallup Poll. Feb. 6-7, 2009. N=1,018 adults nationwide. Margin of error given as +/-3%.

"Now, thinking about another historical figure: Can you tell me with which scientific theory Charles Darwin is associated?" Options rotated

Correct response (Evolution, natural selection, etc.)	55%
Incorrect response	10%
Unsure/don't know	34%
No answer	1%

Example, continued

"In fact, Charles Darwin is noted for developing the theory of evolution. Do you, personally, believe in the theory of evolution, do you not believe in evolution, or don't you have an opinion either way?“

(Poll based on n = 1018 adults)

Believe in evolution	**39%**
Do not believe in evolution	**25%**
No opinion either way	**36%**

Example, continued

- Let p = *population proportion* who believe in evolution.
- Our observed $\hat{p}$ = .39, from sample of 1018.
- Based on samples of n = 1018, $\hat{p}$ comes from a distribution of possible values which is:
  - Approximately normal
  - Mean μ = p
  - Standard deviation σ = $\sqrt{\dfrac{p(1-p)}{1018}}$

Based on this, can we use $\hat{p}$ to estimate p?

Estimating the Population Proportion from a Single Sample Proportion

In practice, we don't know the true population proportion p, so we cannot compute the standard deviation of $\hat{p}$,

$$\text{s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}.$$

In practice, we only take one random sample, so we only have one sample proportion $\hat{p}$. Replacing p with $\hat{p}$ in the standard deviation expression gives us an estimate that is called the **standard error of** $\hat{p}$.

$$\text{s.e.}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

If $\hat{p} = 0.39$ and $n = 1018$, then the standard error is 0.0153. So the *true proportion* who believe in evolution is ***almost surely*** between 0.39 – 3(0.0153) = 0.344 and 0.39 + 3(0.0153) = 0.436.

Parameter 2: Difference in two population proportions, based on independent samples

Example research questions:

- How much difference is there between the proportions that would quit smoking if wearing a nicotine patch versus if wearing a placebo patch?
- How much difference is there in the proportion of UCI students and UC Davis students who are an only child?
- Were the proportions believing in evolution the same in 1994 and 2009?

Population parameter:

$\boldsymbol{p_1 - p_2}$ = difference between the two *population* proportions.

Sample estimate:

$\hat{p}_1 - \hat{p}_2$ = difference between the two *sample* proportions.

Review: Independent Samples

Two samples are called **independent samples** when the measurements in one sample are not related to the measurements in the other sample.

- **Random samples** taken separately from two populations and same response variable is recorded.
- **One random sample** taken and a variable recorded, but units are **categorized** to form two populations.
- Participants **randomly assigned** to one of two treatment conditions, and same response variable is recorded.

Sampling distribution for the difference in two proportions $\hat{p}_1 - \hat{p}_2$

- Approximately normal
- Mean is $p_1 - p_2$ = true difference in the *population* proportions
- Standard deviation of $\hat{p}_1 - \hat{p}_2$ is

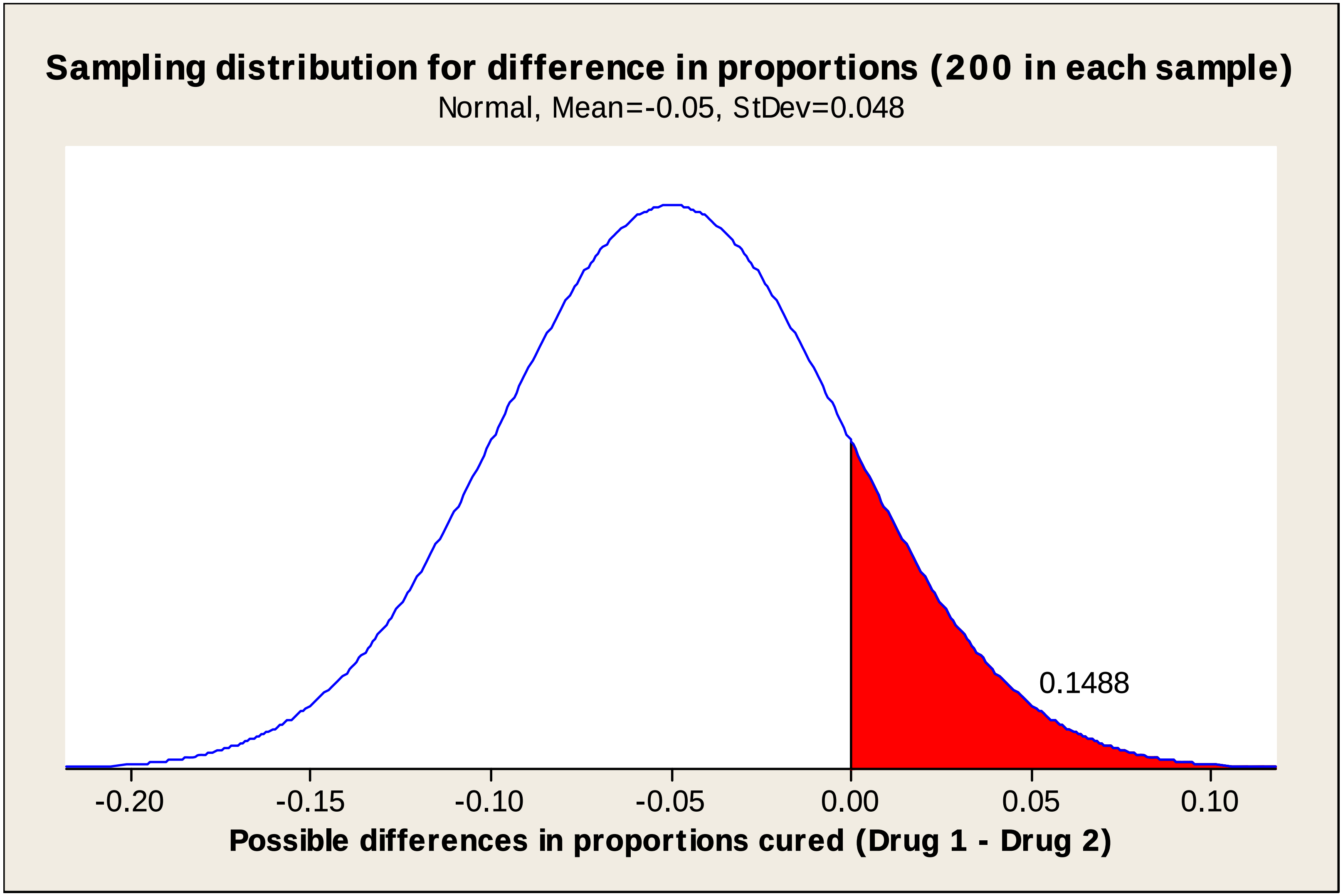
$$s.d.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Ex: 2 drugs, cure rates of 60% and 65%, what is probability that drug 1 will cure more in the *sample* than drug 2 if we sample 200 taking each drug? Want P($\hat{p}_1 - \hat{p}_2 > 0$).

Sampling distribution for $\hat{p}_1 - \hat{p}_2$ is:

- Approximately normal
- Mean $= .60 - .65 = -.05$
- s.d. $= \sqrt{\frac{.6(1-.6)}{200} + \frac{.65(1-.65)}{200}} = .048$

See picture on next slide.


Sampling distribution for difference in proportions (200 in each sample)
Normal, Mean=-0.05, StDev=0.048
0.1488
-0.20
-0.15
-0.10
-0.05
0.00
0.05
0.10
Possible differences in proportions cured (Drug 1 - Drug 2)

General format for all sampling distributions in Chapter 9

The sampling distribution of the sample estimate (the sample statistic) is:

- Approximately normal
- Mean = population parameter
- Standard deviation is called the *standard deviation of* ______, where the blank is filled in with the name of the statistic (p-hat, x-bar, etc.)
- The *estimated* standard deviation is called the *standard error of* _______.

Standard Error of the Difference Between Two Sample Proportions

$$s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Are more UCI than UCD students an only child?

$n_1 = 358$ (UCI, 2 classes combined) $n_2 = 173$ (UCD)

UCI: 40 of the 358 students were an only child = $\hat{p}_1 = .112$

UCD: 14 of the 173 students were an only child = $\hat{p}_2 = .081$

So, $\hat{p}_1 - \hat{p}_2 = .112 - .081 = .031$

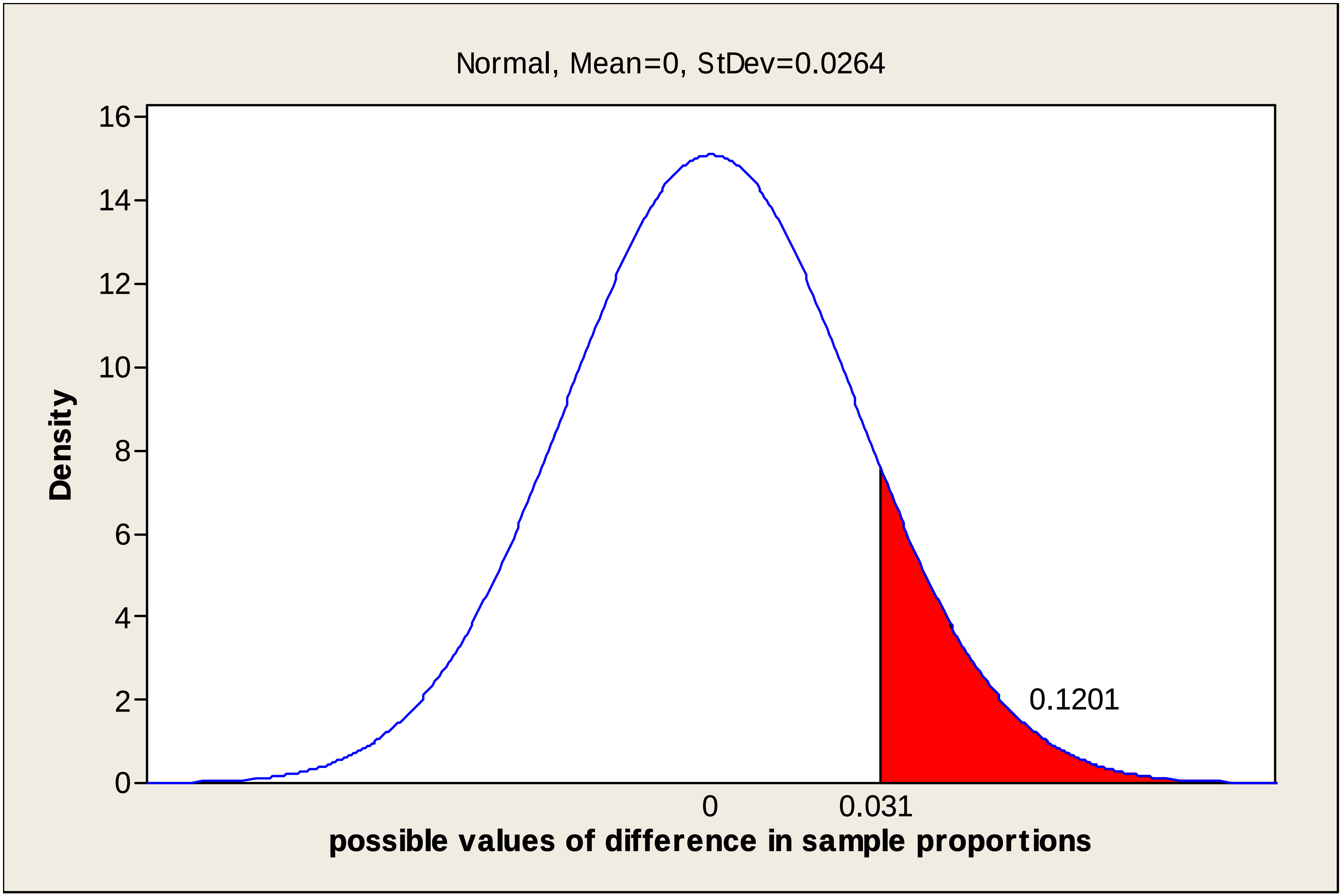
and $s.e.(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{.11(1-.11)}{358} + \frac{.08(1-.08)}{173}} = .0264$

Suppose *population* proportions are the same, so true difference $p_1 - p_2 = 0$

Then the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is:

- Approximately normal
- Mean = population parameter = 0
- The estimated standard deviation is .0264
- *Observed* difference of .031 is z = 1.174 standard errors above the mean of 0.
- So the difference of .031 *could* just be chance variability
- See picture on next slide; area above .031 = .1201

Sampling distribution of $\hat{p}_1 - \hat{p}_2$



Standardized Statistics for Sampling Distributions

Recall the general form for standardizing a value *k* for a random variable with a normal distribution:

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

For all 5 parameters we will consider, we can find where our *observed* sample statistic falls if we hypothesize a specific number for the population parameter:

$$z = \frac{sample\ statistic - population\ \ parameter}{s.d.(sample\ statistic)}$$

Example: *Do college students watch less TV?*

In general, there isn't much correlation between age and hrs/TV per day. In 2008 General Social Survey (very large n), 73% watched ≥2 hours per day. So *assume adult population proportion* is .73.

In a sample of 175 college students (at Penn State), 105 said they watched 2 or more hours per day.
Is it likely that the *population proportion for students* is also .73?

$$\hat{p} = \frac{105}{175} = .6 \qquad s.d.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.73(1-0.73)}{175}} = 0.034$$

$$z = \frac{.6 - .73}{.034} = -3.82$$

This z-score is too small! Area below it is .00007. Students are different from general population.

Case Study 9.1 *Do Americans Really Vote When They Say They Do?*

Election of 1994:

- *Time Magazine Poll*: n = 800 adults (two days after election), **56% reported that they had voted**.
- Info from Committee for the Study of the American Electorate: **only 39% of American adults had voted**.

If true $p = 0.39$ then *sample* proportions for samples of size $n = 800$ should vary approximately normally with …

$$\text{mean} = p = 0.39 \text{ and s.d.}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.39(1-0.39)}{800}} = 0.017$$

Case Study 9.1 *Do Americans Really Vote When They Say They Do?*

If respondents were telling the truth, the sample percent should be no higher than 39% + 3(1.7%) = 44.1%, nowhere near the reported percentage of 56%.

If 39% of the population voted, the **standardized score** for the reported value of 0.56 (56%) is …

$$z = \frac{0.56 - 0.39}{0.017} = 10.0$$

It is virtually ***impossible*** to obtain a standardized score of 10. So most likely, the non-voters lied and said they voted.

Summary (so far)
For one proportion

Sampling distribution for $\hat{p}$

- Approximately normal
- Mean = p
- Standard deviation = $\text{s.d.}(\hat{p}) = \sqrt{\dfrac{p(1-p)}{n}}$
- Standard error = $\text{s.e.}(\hat{p}) = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
- Remember, np and $n(1-p)$ must be at least 10 to use this.

Summary, continued
For difference in two proportions

Sampling distribution for $\hat{p}_1 - \hat{p}_2$

- Approximately normal
- Mean $= p_1 - p_2$
- Standard deviation and standard error:

  See page 329.
- Remember, all of $n_1 p_1$, $n_1(1 - p_1)$, $n_2 p_2$ and $n_2(1 - p_2)$, must be at least 10 to use this. In other words, at least 10 "successes" and 10 "failures" in each group.

Preparing for the Rest of Chapter 9

For all 5 situations we are considering, the sampling distribution of the sample statistic:

- Is approximately normal
- Has mean = the corresponding population parameter
- Has standard deviation that involves the population parameter(s) and thus can't be known without it (them)
- Has standard error that doesn't involve the population parameters and is used to estimate the standard deviation.
- Has standard deviation (and standard error) that get smaller as the sample size(s) n get larger.

Summary table on page 353 will help you with these!