Today: Finish Chapter 9 (Sections 9.6 to 9.8 and 9.9 Lesson 3)

ANNOUNCEMENTS:
• Quiz #7 begins after class today, ends Friday at noon.
• Jason Kramer will give the lecture on Friday.
• Please check your grades on eee and let me know if there are problems.

HOMEWORK: (Due Friday)
Chapter 9: #66, 90, 91
Update on the five situations we will cover for the rest of this quarter:

<table>
<thead>
<tr>
<th>Parameter name and description</th>
<th>Population parameter</th>
<th>Sample statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>For Categorical Variables:</strong> [Done!]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One population proportion (or probability)</td>
<td>$p$</td>
<td>$\hat{p}$</td>
</tr>
<tr>
<td>Difference in two population proportions</td>
<td>$p_1 - p_2$</td>
<td>$\hat{p}_1 - \hat{p}_2$</td>
</tr>
<tr>
<td><strong>For Quantitative Variables:</strong> [Today, Fr, M]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>One population mean</td>
<td>$\mu$</td>
<td>$\bar{x}$</td>
</tr>
<tr>
<td>Population mean of paired differences (dependent samples, paired)</td>
<td>$\mu_d$</td>
<td>$\bar{d}$</td>
</tr>
<tr>
<td>Difference in two population means (independent samples)</td>
<td>$\mu_1 - \mu_2$</td>
<td>$\bar{x}_1 - \bar{x}_2$</td>
</tr>
</tbody>
</table>

For each situation will we:

- Learn about the **sampling distribution** for the sample statistic
- Learn how to find a **confidence interval** for the true value of the parameter
- **Test hypotheses** about the true value of the parameter
Recall, general format for all sampling distributions in Ch. 9:

The sampling distribution of the sample statistic is approximately normal, with:

- **Mean** = population parameter ($p$, $p_1 - p_2$, $\mu$, etc.)

- **Standard deviation** = *standard deviation of _______;* the blank is filled in with the statistic ($\hat{p}$, $\hat{p}_1 - \hat{p}_2$, $\bar{x}$ etc.)

- Often the standard deviation must be estimated, and then it is called the *standard error of _______.*

See summary table on pages 382-383 for all details!
Today: Sampling distributions for:
- one mean
- mean difference for paired data
- difference between means for independent samples

Remember, two samples are called **independent samples** when the measurements in one sample are not related to the measurements in the other sample. Could come from:
- Separate samples
- One sample, divided into two groups by a categorical variable (such as male or female)
- Randomization into two groups where each unit goes into only one group

**Paired data** occur when two measurements are taken on the same individuals, or individuals are paired in some way.
Sampling Distribution for a Sample Mean (Section 9.6)

Suppose we take a random sample of size $n$ from a population and measure a quantitative variable.

Notation:

$\mu =$ mean for the population of measurements.

$\sigma =$ standard deviation for the population of measurements.

$\overline{x} =$ sample mean for a random sample of $n$ individuals.

$s =$ sample standard deviation for the random sample
The sampling distribution of the sample mean \( \bar{x} \) is approximately normal, with:

- Mean = population parameter = \( \mu \)

- Standard deviation = *standard deviation of \( \bar{x} = \)

\[
s.d.(\bar{x}) = \frac{\sigma}{\sqrt{n}}
\]

- Often the standard deviation must be estimated, and then it is called the *standard error of \( \bar{x} \). Replace \( \sigma \) with the sample standard deviation, so

\[
s.e.(\bar{x}) = \frac{s}{\sqrt{n}}
\]
Suppose we want to estimate the mean weight loss for the population of people who attend weight loss clinics for 10 weeks. *Suppose* the distribution of weight losses is approximately normal, \( \mu = 8 \) pounds, \( \sigma = 5 \) pounds. (Empirical rule: see picture)

*Population of individual weight losses*
• We plan to take a random sample of 25 people from this population and record weight loss for each person, then find sample mean \( \bar{x} \).

• We know the value of the sample mean will vary for different samples of \( n = 25 \). How much will they vary? Where is the center of the distribution of possibilities?

Results for four possible random samples of 25 people, with the corresponding sample mean \( \bar{x} \) and sample standard deviation \( s \):

Sample 1: \( \bar{x} = 8.32 \) pounds, \( s = 4.74 \) pounds.
Sample 2: \( \bar{x} = 6.76 \) pounds, \( s = 4.73 \) pounds.
Sample 3: \( \bar{x} = 8.48 \) pounds, \( s = 5.27 \) pounds.
Sample 4: \( \bar{x} = 7.16 \) pounds, \( s = 5.93 \) pounds.
Note:

• Each sample had a different sample mean, which did not always match the population mean of 8 pounds.

• Although we cannot determine whether one sample mean will accurately reflect the population mean, statisticians have determined what to expect for all possible sample means.

\[ \mu = \text{mean for population of interest} = 8 \text{ pounds} \]

\[ \sigma = \text{standard deviation for population of interest} = 5 \text{ pounds.} \]

\[ \bar{x} = \text{sample mean for a random sample of} \ n \text{ individuals.} \]

Then the sampling distribution of \( \bar{x} \) is approximately normal, with

• Mean = \( \mu \)

• Standard deviation = \( s.d.(\ \bar{x} \ ) = \frac{\sigma}{\sqrt{n}} \)
Example: Mean of 25 weight losses, the distribution of possible values is approximately normal with:

- mean = 8 pounds
- standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{25}} = 1$ pound

Compare: individual weight loss, $\bar{x}$ for $n = 25$, $\bar{x}$ for $n = 100$

<table>
<thead>
<tr>
<th></th>
<th>Individual weight loss</th>
<th>Mean of 25</th>
<th>Mean of 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>8 pounds</td>
<td>8 pounds</td>
<td>8 pounds</td>
</tr>
<tr>
<td>St. Dev.</td>
<td>5 pounds</td>
<td>1 pound</td>
<td>$\frac{1}{2}$ pound</td>
</tr>
</tbody>
</table>

Conditions for sampling distribution of $\bar{x}$ to be approximately normal:

- Population (individual values) are approx. bell-shaped OR
- Sample size is large (at least 30, more if outliers)
Comparing original population with sampling distribution of $\bar{x}$:

From the empirical rule:

<table>
<thead>
<tr>
<th></th>
<th>68%</th>
<th>95%</th>
<th>99.7%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individuals</td>
<td>3 to 13 pounds</td>
<td>-2 to 18 pounds</td>
<td>-7 to 23 pounds</td>
</tr>
<tr>
<td>Mean of $n = 25$</td>
<td>7 to 9 pounds</td>
<td>6 to 10 pounds</td>
<td>5 to 11 pounds</td>
</tr>
</tbody>
</table>
Note that larger sample size will result in smaller s.d.($\bar{x}$)

Compare sampling distribution for $n = 25$ and $n = 100$:

In other words, for larger samples, $\bar{x}$ will be closer to $\mu$ in general, and thus will be a better estimate for $\mu$. 
Example where the original population is *not* bell-shaped:

A bus runs every 10 minutes. When you show up at the bus stop, it could come immediately, or anytime up to 10 minutes. So the time you wait for it is *uniform*, from 0 to 10 minutes, and independent from day to day.

Population mean $= \mu = 5$ minutes, population s.d. $= \sigma = \sqrt{\frac{10^2}{12}} = 2.9$

What is the sampling distribution for $\bar{x}$ for $n = 40$ days?

Even though the original times are *uniform* (flat shape), the possible values of the *sample mean* $\bar{x}$ are:

- Approximately **normal**
- Mean $= 5$ minutes
- Standard deviation $= \frac{\sigma}{\sqrt{n}} = \frac{2.9}{\sqrt{40}} = 0.46$ minutes
Examples of possible samples:

<table>
<thead>
<tr>
<th>Original values and sampling distribution of mean for n = 40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution</td>
</tr>
<tr>
<td>---------------</td>
</tr>
<tr>
<td>Normal</td>
</tr>
<tr>
<td>Distribution</td>
</tr>
<tr>
<td>Uniform</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>6.9</th>
<th>6.2</th>
<th>8.8</th>
<th>8.3</th>
<th>7.1</th>
<th>6.5</th>
<th>7.3</th>
<th>9.5</th>
<th>3.4</th>
<th>9.9</th>
<th>5.8</th>
<th>1.4</th>
<th>3</th>
<th>4.1</th>
<th>4.4</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.7</td>
<td>1.2</td>
<td>7.4</td>
<td>0.7</td>
<td>6.8</td>
<td>7.7</td>
<td>6.2</td>
<td>6.1</td>
<td>3.3</td>
<td>1</td>
<td>5.3</td>
<td>9.4</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>9.4</td>
</tr>
<tr>
<td>8.1</td>
<td>3.9</td>
<td>7.2</td>
<td>8.6</td>
<td>1</td>
<td>0.4</td>
<td>9.9</td>
<td>9.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.6</td>
<td>3.2</td>
<td>0.8</td>
<td>0.2</td>
<td>8.5</td>
<td>1.4</td>
<td>4.7</td>
<td>0.5</td>
<td>9.7</td>
<td>8.9</td>
<td>6.3</td>
<td>3.3</td>
<td>0.8</td>
<td>4.1</td>
<td>2.6</td>
<td>3.7</td>
</tr>
<tr>
<td>5.7</td>
<td>3.2</td>
<td>8.9</td>
<td>2.3</td>
<td>1.1</td>
<td>9.9</td>
<td>3</td>
<td>0.8</td>
<td>7.9</td>
<td>0.8</td>
<td>5.9</td>
<td>2.5</td>
<td>7.9</td>
<td>7.6</td>
<td>4</td>
<td>2.2</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
<td>6.1</td>
<td>6.9</td>
<td>8.1</td>
<td>2.6</td>
<td>9.6</td>
<td>5.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ \bar{x} = 5.29 \]
\[ \bar{x} = 4.3 \]
Sections 9.7 and 9.8: Sampling distributions for mean of paired differences, and for differences in means for independent samples. Need to learn to distinguish between these two situations.

Notation for paired differences:
- $d_i =$ difference in the two measurements for individual $i = 1, 2, \ldots, n$
- $\mu_d =$ mean for the population of differences, if all possible pairs were to be measured
- $\sigma_d =$ the standard deviation for the population of differences
- $\bar{d} =$ the mean for the sample of differences
- $s_d =$ the standard deviation for the sample of differences

Example: IQ measured after listening to Mozart and to silence
- $d_i =$ difference in IQ for student $i$ for the two conditions
- $\mu_d =$ population mean difference, if all students measured (unknown)
- $\bar{d} =$ the mean for the sample of differences = 9 IQ points

Based on sample, we want to estimate mean population difference
Notation for difference in means for independent samples:

\[ \mu_1 = \text{population mean for the first population} \]
\[ \mu_2 = \text{population mean for the second population} \]

Parameter of interest is \( \mu_1 - \mu_2 = \text{the difference in population means} \)

\[ \bar{x}_1 = \text{sample mean for the sample from the first population} \]
\[ \bar{x}_2 = \text{sample mean for the sample from the second population} \]

The sample statistic is \( \bar{x}_1 - \bar{x}_2 = \text{the difference in sample means} \)

\[ \sigma_1 = \text{population standard deviation for the first population} \]
\[ \sigma_2 = \text{population standard deviation for the second population} \]

\[ s_1 = \text{sample standard deviation for the sample from the 1}^{\text{st}} \text{ population} \]
\[ s_2 = \text{sample standard deviation for the sample from the 2}^{\text{nd}} \text{ population} \]

\[ n_1 = \text{size of the sample from the 1}^{\text{st}} \text{ population} \]
\[ n_2 = \text{size of the sample from the 2}^{\text{nd}} \text{ population} \]
Examples where **independent samples** might be used:
- Compare weight loss for men and women at the weight loss clinic.
- Compare UCI students with students from another campus on quantitative measures like hours spent studying per week, income, etc.
- Compare number of sick days off from work for people who had a flu shot and people who didn’t.
- Compare change in blood pressure for people randomly assigned to a meditation program or an exercise program for 3 months.

Examples where **paired data** might be used:
- Estimate average difference in income for husbands and wives.
- Compare SAT scores before and after a training program.
- Weight loss example can be thought of as paired difference, with weight before and weight after the program.

Note that **paired differences** are similar to the “one mean” situation, except special notation tells us that the means are for differences.
Conditions for the sampling distributions for these two situations are the same as for a single mean, with a slight twist:

- For paired differences, population of differences must be bell-shaped OR sample must be large.
- For difference in means for independent samples, both populations must be bell-shaped OR both sample sizes must be large.

In both cases, the sampling distribution for the sample statistic is approximately normal, with mean = population parameter of interest.

For paired differences: 
\[ \text{s.d.}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}} \] (same as one mean, but with d’s)

For difference in two means: 
\[ \text{s.d.}(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \]
Standardized Statistics:

For all 5 cases in Chapter 9, as long as the conditions are satisfied for the sampling distribution to be approximately normal, the **standardized statistic** for a sample statistic is:

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

Note that the denominator has s.d., *not* s.e.

**For one mean:**

\[
    z = \frac{\bar{x} - \mu}{\text{s.d.} (\bar{x})} = \frac{\bar{x} - \mu}{\sigma} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}
\]
Example of weight loss clinic. Suppose in a sample of 25 clients the average weight loss is 0.

If population mean weight loss is really 8 pounds with $\sigma = 5$ pounds, how unlikely is a sample mean of 0 pounds for $n = 25$?

Probability of x-bar being 0 or less is $6.2210\times10^{-16}$. 

![Sampling distribution for x-bar when mu=8, sigma=5, n=25](image)
How to compute this answer:

Sample means for $n = 25$ are approximately normal with mean of 8 pounds and s.d. of 1 pound. So, the standardized score for 0 is:

$$z = \frac{\sqrt{25}(0 - 8)}{5} = -8$$

A z-score of –8 is not very likely! So if we saw a 0 mean weight loss, we would not believe that the population mean is 8 pounds!

**Note:** When $\sigma$ is not known, we must use the sample standard deviation $s$ instead. **Standard error** of $\bar{x}$ is

$$s.e.(\bar{x}) = \frac{s}{\sqrt{n}}$$

In that case, *standardized statistic* has a *t-distribution*, also called *Student’s t distribution*.
Student’s $t$ distribution

In 1908 William Sealy Gossett figured out the formula for the $t$ distribution. Called Student’s $t$ because… explained in class!
Standardized Statistic Using Standard Error

Usually we don’t know \( \sigma \) (population standard deviation), so we need to use \( s \) (sample standard deviation). In that case, the standardized statistic for \( \bar{x} \) is

\[
t = \frac{\bar{x} - \mu}{s.e.(\bar{x})} = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{\sqrt{n}(\bar{x} - \mu)}{s}
\]

This has a *Student’s t distribution* with degrees of freedom \( = n - 1 \)

- It looks almost exactly like the normal distribution
- It is completely specified by knowing the “df”
- It gets closer and closer to the normal distribution, and when degrees of freedom \( = \) infinity, it is exactly the normal distribution.
Comparison of \( t \) distribution with \( df = 5 \) and standard normal distribution

For example, middle 95\% for \( t \) with \( df = 5 \) is \(-2.57\) to \(+2.57\)
For standard normal, it is about \(-2\) to \(+2\)
In Chapter 11 (Friday) we will learn how to find probabilities.
Summary of sampling distributions for the 5 parameters (p. 382):
• The statistic has a sampling distribution.
• It is approximately normal if the sample(s) is (are) large enough.
• The mean of the sampling distribution = the parameter.
• The standard deviation of the sampling distribution is in the table below, in the column “standard deviation of the statistic.”
• Sometimes it needs to be estimated, then “standard error” is used.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Statistic</th>
<th>Standard Deviation of the Statistic</th>
<th>Standard Error of the Statistic</th>
<th>Standardized Statistic with s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>One proportion</td>
<td>( p )</td>
<td>( \hat{p} )</td>
<td>( \sqrt{\frac{p(1-p)}{n}} )</td>
<td>( \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} )</td>
</tr>
<tr>
<td>Difference Between Proportions</td>
<td>( p_1 - p_2 )</td>
<td>( \hat{p}_1 - \hat{p}_2 )</td>
<td>( \sqrt{\frac{p_1(1-p_1)+p_2(1-p_2)}{n_1+n_2}} )</td>
<td>( \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)+\hat{p}_2(1-\hat{p}_2)}{n_1+n_2}} )</td>
</tr>
<tr>
<td>One Mean</td>
<td>( \mu )</td>
<td>( \bar{x} )</td>
<td>( \frac{\sigma}{\sqrt{n}} )</td>
<td>( \frac{s}{\sqrt{n}} )</td>
</tr>
<tr>
<td>Mean Difference, Paired Data</td>
<td>( \mu_d )</td>
<td>( \bar{d} )</td>
<td>( \frac{\sigma_d}{\sqrt{n}} )</td>
<td>( \frac{s_d}{\sqrt{n}} )</td>
</tr>
<tr>
<td>Difference Between Means</td>
<td>( \mu_1 - \mu_2 )</td>
<td>( \bar{x}_1 - \bar{x}_2 )</td>
<td>( \sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1+n_2}} )</td>
<td>( \sqrt{\frac{s_1^2 + s_2^2}{n_1+n_2}} )</td>
</tr>
<tr>
<td>Parameter</td>
<td>Statistic</td>
<td>Standard Deviation of the Statistic</td>
<td>Standard Error of the Statistic</td>
<td>$z$ or $t$? (with s.e.)</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>----------------------------------</td>
<td>-------------------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>One proportion</td>
<td>$p$</td>
<td>$\hat{p}$</td>
<td>$\sqrt{\frac{p(1-p)}{n}}$</td>
<td>$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$</td>
</tr>
<tr>
<td>Difference Between Proportions</td>
<td>$p_1 - p_2$</td>
<td>$\hat{p}_1 - \hat{p}_2$</td>
<td>$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$</td>
<td>$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$</td>
</tr>
<tr>
<td>One Mean</td>
<td>$\mu$</td>
<td>$\bar{x}$</td>
<td>$\frac{\sigma}{\sqrt{n}}$</td>
<td>$\frac{s}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Mean Difference, Paired Data</td>
<td>$\mu_d$</td>
<td>$\bar{d}$</td>
<td>$\frac{\sigma_d}{\sqrt{n}}$</td>
<td>$\frac{s_d}{\sqrt{n}}$</td>
</tr>
<tr>
<td>Difference Between Means</td>
<td>$\mu_1 - \mu_2$</td>
<td>$\bar{x}_1 - \bar{x}_2$</td>
<td>$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$</td>
<td>$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$</td>
</tr>
</tbody>
</table>