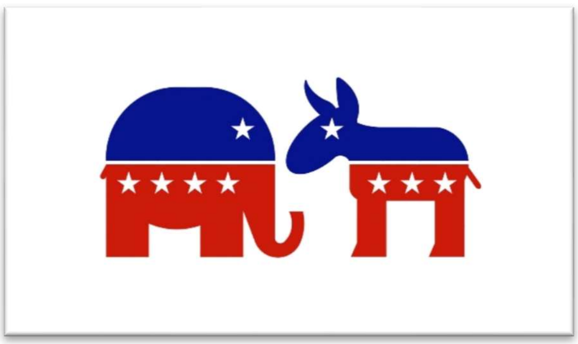
# Unit 11 Slides: Inference for $\mu_1 - \mu_2$ and $p_1 - p_2$

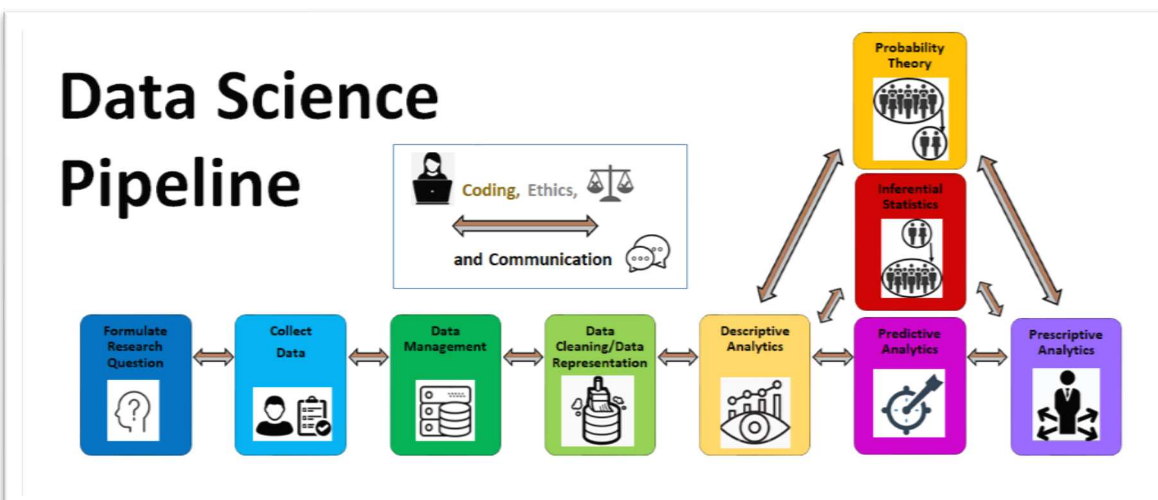



***Case Studies:***

- **Is there an association between childhood lead exposure and IQ?**
  - *What is a plausible range of values for* $\mu_{lo} - \mu_{hi}$, *the* ***difference*** *in the* ***average*** *IQ score of children with low lead level exposure and the* ***average*** *IQ score of children with high lead level exposure?*
  - *Is there sufficient evidence to suggest* $\mu_{lo} - \mu_{hi} \neq 0$, ie. that *there is a* ***difference*** *in the* ***average*** *IQ score of children with low lead level exposure and the* ***average*** *IQ score of children with high lead level exposure?*

- **Is there an association between political party and approval for the direction the country is going in (in 2017)?**
  - *What is a plausible range of values for* $p_{dem} - p_{rep}$, *the* ***difference*** *in the* ***proportion*** *of democrats that approve vs. the* ***proportion*** *of republicans that approve?*
  - *Is there sufficient evidence to suggest* $p_{dem} - p_{rep} \neq 0$, ie. that *there is a* ***difference*** *in the* ***proportion*** *of democrats that approve vs. the* ***proportion*** *of republicans that approve?*

## Purpose of this Lecture:



In this lecture we will cover the following topics in **inference** and **probability.**

1. Two main types of inference for unknown population parameters.
2. How to better account for the additional uncertainty introduced by having to estimate additional parameters in probability and inference?
   2.1. Issues with plugging in s for $\sigma$
   2.2. t-score of a sample mean
   2.3. distribution of t-scores (under certain conditions)
   **2.4. t-distribution**
3. Properties of the **Sampling Distribution of <u>Sample Mean Differences</u>**
   3.1. Mean
   3.2. Standard Deviation
   3.3. When is it normal?
4. Conducting **Inference on a Population Mean Difference ($\boldsymbol{\mu_1 - \mu_2}$)**
   4.1. Creating a confidence interval for $\boldsymbol{\mu_1 - \mu_2}$
   4.2. Conducting a hypothesis test to test the claim: $\boldsymbol{\mu_1 - \mu_2 \neq 0}$
   4.3. Conducting a hypothesis test to test the claim: $\mu_1 - \mu_2 \neq 0$ -with a p-value – if you know $\sigma_1$ and $\sigma_2$
   4.4. Conducting a hypothesis test to test the claim: $\mu_1 - \mu_2 \neq 0$ -with a p-value – if you DON'T know $\sigma_1$ and $\sigma_2$
   4.5. Conducting a hypothesis test to test the claim: $\mu_1 - \mu_2 \neq 0$ -with a confidence interval
5. Properties of **Sampling Distribution of <u>Sample Proportion Differences</u>**
   5.1. Mean
   5.2. Standard Deviation
   5.3. When is it normal?
6. Conducting **Inference on a Population Proportion Difference ($\boldsymbol{p_1 - p_2}$)**
   6.1. Creating a confidence interval for $\boldsymbol{p_1 - p_2}$
   6.2. Conducting a hypothesis test to test the claim: $\boldsymbol{p_1 - p_2 \neq 0}$

<u>Additional resources:</u>

Sections 5.1, 5.3, 6.2 in Diez, Barr, and Cetinkaya-Rundel, (2015), *OpenIntro Statistics* https://www.openintro.org/download.php?file=os3&redirect=/stat/textbook/os3.php

# 1. Two Main Types of Inference for Unknown Population Parameters

Suppose we were interested in a population parameter of a large, unknown population. But all we can collect is a random sample from this population.

## What we wish we could know:


Is there an association between lead exposure and childhood IQ for ALL children?

Is $\mu_{lo} - \mu_{hi} \neq 0$?,

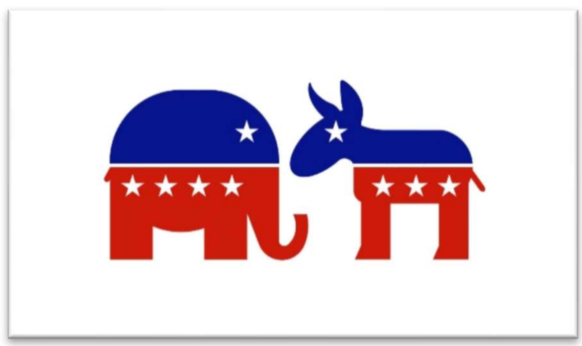
- $\mu_{lo}$: **average** IQ score of ALL children with low lead level exposure
- $\mu_{hi}$: **average** IQ score of ALL children with high lead level exposure

## What we *can* answer:

1. **Confidence Interval:** What is a plausible range of values for the $\mu_{lo} - \mu_{hi}$?
2. **Hypothesis Test:** Is there sufficient evidence to suggest $\mu_{lo} - \mu_{hi} \neq 0$?

---

## What we wish we could know:



Is there an association between political party and opinion on the direction the country is going in (in 2017) for ALL adults living in the U.S.?

Is $p_{dem} - p_{rep} \neq 0$?,

- $p_{dem}$: the **proportion** of ALL democrats living in the U.S. (in 2017) that approve of the direction the party is going in
- $p_{rep}$: the **proportion** of ALL republicans living in the U.S. (in 2017) that approve of the direction the party is going in

## What we *can* answer:

1. **Confidence Interval:** What is a plausible range of values for the $p_{dem} - p_{rep}$?
2. **Hypothesis Test:** Is there sufficient evidence to suggest $p_{dem} - p_{rep} \neq 0$?

# 2. How to better account for the additional uncertainty introduced by having to estimate additional parameters in probability and inference?

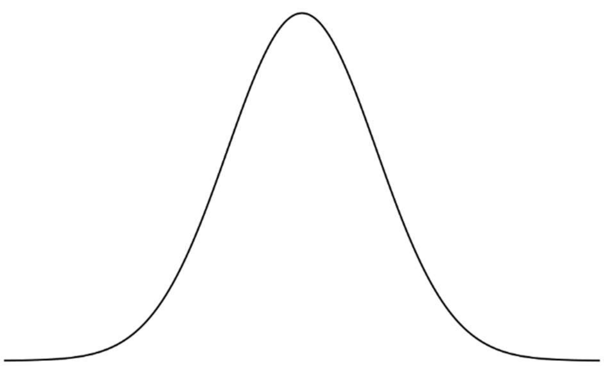
## 2.1. Issues with Plugging in *s* for $\sigma$ (RECAP)

**If Central Limit Theorem Conditions for sample means hold (for a given sample size n), then the sampling distribution of sample means (drawn from samples of this size n) is**

**____________________.**

$$\bar{X} \sim N(mean = \_\_\_\_, std = \_\_\_\_\_\_)$$

**Sampling Distribution of Sample Means**

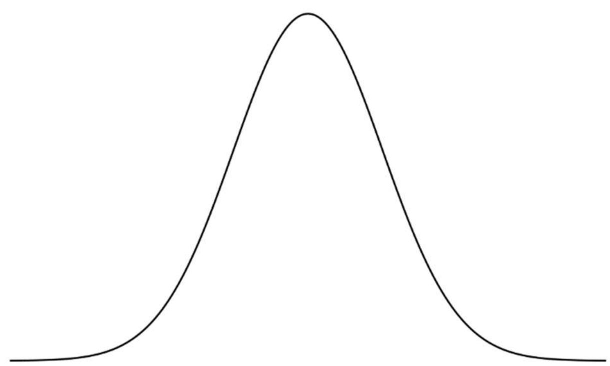


**Thus if the sampling distribution of sample means is __________________, then the distribution of z-scores of sample means is the**

**______________________________.**

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(mean = \_\_\_, std = \_\_\_\_)$$

**Distribution of Sample Mean z-scores**



# BUT….. What if we don't know σ?

- **Previous "Work Around":**

  In the past we said ____________________________________________.

- **Why is this "work around" not the best?**
  - We can't conduct inference on μ when ____________________.
  - ALL approximations where we plug in _______ for _______ are

    ________________________________.

- **How do we make our "work around" for situations like these better?**
  - Use the **t-scores of sample means.**
  - Use the **t-distribution** to calculate probabilities involving these sample means.

## 2.2. t-score of a sample mean

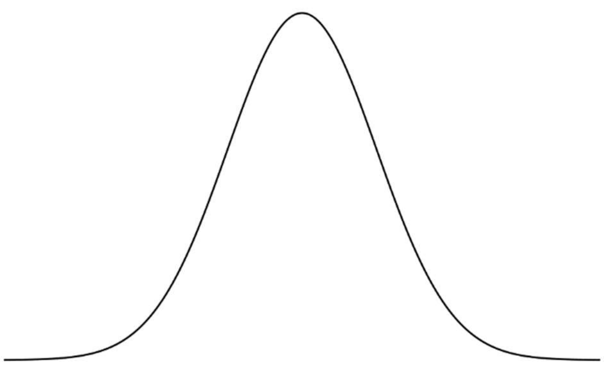
- We define the **t-score of a sample mean** $\bar{x}_0$ as $t = \frac{\bar{x}_0 - \mu}{\frac{s}{\sqrt{n}}}$
- We define the **t-score of a sample mean random variable** $\bar{X}$ as $T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$

# 2.3. Distribution of t-scores (under certain conditions)

**If Central Limit Theorem Conditions for sample means hold (for a given sample size n), then the sampling distribution of sample means (drawn from samples of this size n) is ____________________.**

$$\bar{X} \sim N(mean = \_\_\_\_, std = \_\_\_\_\_\_)$$

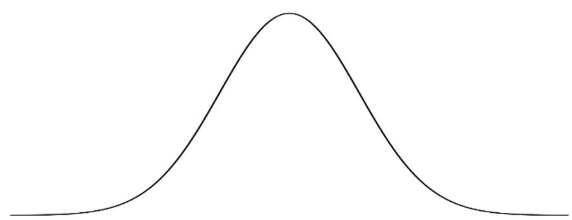
**Sampling Distribution of Sample Means**



**Thus if the sampling distribution of sample means is __________________, then the distribution of t-scores of sample means is the ______________________________________.**

$$T = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

**Distribution of Sample Mean t-scores**



# 2.4. t-Distribution

## ***Random Variable that Follows the t-Distribution*:**

**Definition**: A continuous random variable is said to follow the **t-distribution** with $\nu$ degrees of freedom if it has the following probability density function (pdf).

**Short-Hand: ______________________**

**Probability Density Function**:

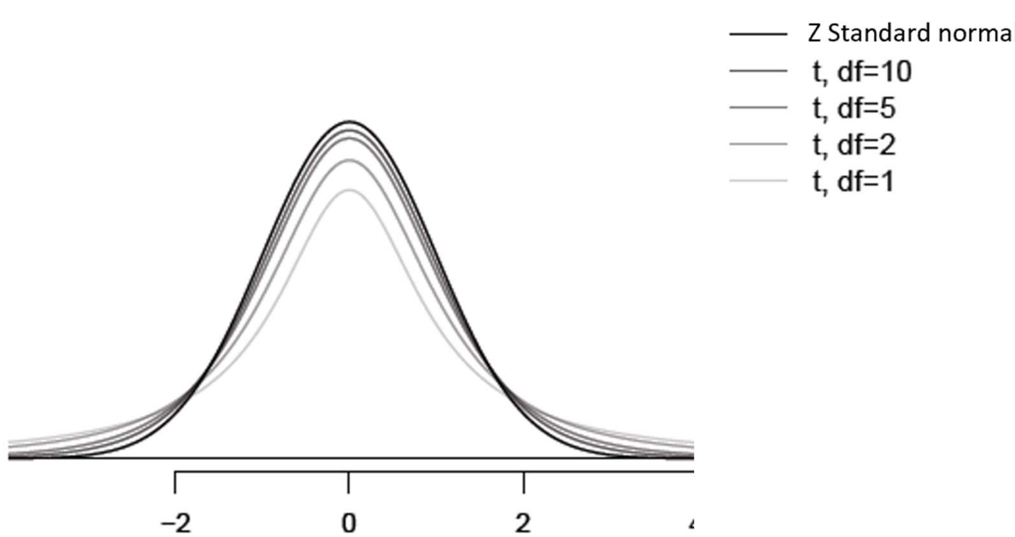
$$f(x) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, for -\infty < x < \infty$$

**Parameter that Dictates Shape**:___________________

**Properties**:

- Always centered at:____________
- As the degrees of freedom increases, the _____________ of the distribution _______________ and the peak of the distribution ___________________.

**Comparison to Standard Normal Distribution**:



**Go to Unit 11 slides (Section 2.4) for these exercises.**

**Ex:** Calculate the probability that a t-score (that is an observation from the t-distribution with 20 degrees of freedom) is greater than 1.96.

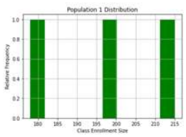
**Ex:** Calculate the t-score that creates a right tail area of 0.025 under the t-distribution with 20 degrees of freedom.

**Ex:** Suppose we know that the average GPA of ALL UIUC students is 3.3. We then randomly select 20 UIUC students and find that they have a sample mean GPA of 3.5 and a standard deviation of 0.3. Suppose that the distribution of all UIUC student GPAs is approximately normal. Calculate the probability (the most accurate one) of randomly selecting a sample mean that is greater than or equal to the sample mean that we collected.

# 3. Properties of the Sampling Distribution of Sample Mean Differences

## Population 1 of Numerical Data

| | course | section | enrolled |
|---|---|---|---|
| 4 | badm210 | A | 215 |
| 5 | badm210 | C | 197 |
| 6 | badm210 | B | 178 |

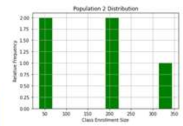



Mean of Population Distribution= __________ = $E[X_1]$

Standard Deviation of Population Distribution= __________ = $SD[X_1]$

## Population 2 of Numerical Data

| | course | section | enrolled |
|---|---|---|---|
| 0 | cs105 | B | 345 |
| 1 | cs105 | A | 201 |
| 2 | stat107 | A | 197 |
| 3 | stat207 | A | 53 |
| 7 | adv307 | A | 37 |

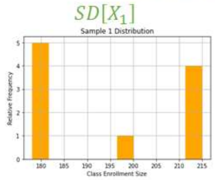



Mean of Population Distribution= __________ = $E[X_2]$

Standard Deviation of Population Distribution= __________ = $SD[X_2]$

## Sample 1 of Numerical Data

| | course | section | enrolled |
|---|---|---|---|
| 4 | badm210 | A | 215 |
| 4 | badm210 | A | 215 |
| 4 | badm210 | A | 215 |
| 6 | badm210 | B | 178 |
| 6 | badm210 | B | 178 |
| 4 | badm210 | A | 215 |
| 6 | badm210 | B | 178 |
| 5 | badm210 | C | 197 |
| 6 | badm210 | B | 178 |
| 6 | badm210 | B | 178 |



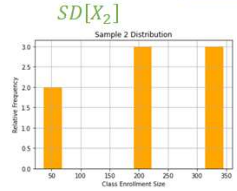

Mean of Sample 1 Distribution= __________

Standard Deviation of Sample 1 Distribution= __________

Size of Sample 1 Distribution= __________

## Sample 2 of Numerical Data

| | course | section | enrolled |
|---|---|---|---|
| 0 | cs105 | B | 345 |
| 0 | cs105 | B | 345 |
| 3 | stat207 | A | 53 |
| 0 | cs105 | B | 345 |
| 2 | stat107 | A | 197 |
| 7 | adv307 | A | 37 |
| 2 | stat107 | A | 197 |
| 2 | stat107 | A | 197 |




Mean of Sample 2 Distribution= __________

Standard Deviation of Sample 2 Distribution= __________

Size of Sample 2 Distribution= __________

Collect Many Random Samples (all of size **n1=10**) drawn with replacement.

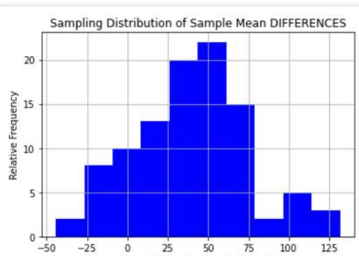
| Random Sample of n1=10 Course Enrollments *(drawn with replacement from population 1)* | Random Sample of n1=10 Course Enrollments *(drawn with replacement from population 1)* | ... | Random Sample of n1=10 Course Enrollments *(drawn with replacement from population 1)* |
|---|---|---|---|
| 197 | 197 | ... | 178 |
| 178 | 215 | ... | 197 |
| 178 | 215 | ... | 178 |
| 197 | 197 | ... | 178 |
| 215 | 215 | ... | 178 |
| 178 | 197 | ... | 178 |
| 215 | 197 | ... | 197 |
| 197 | 197 | ... | 178 |
| 178 | 197 | ... | 215 |
| 178 | 215 | ... | 197 |

| Random Sample of n2=8 Course Enrollments *(drawn with replacement from population 2)* | Random Sample of n2=8 Course Enrollments *(drawn with replacement from population 2)* | ... | Random Sample of n2=8 Course Enrollments *(drawn with replacement from population 2)* |
|---|---|---|---|
| 201 | 37 | ... | 53 |
| 201 | 53 | ... | 53 |
| 37 | 197 | ... | 37 |
| 345 | 201 | ... | 197 |
| 197 | 53 | ... | 53 |
| 197 | 197 | ... | 197 |
| 53 | 201 | ... | 345 |
| 53 | 53 | ... | 201 |

Collect Many Random Samples (all of size **n2=8**) drawn with replacement.

## Sampling Distribution of Sample Mean Differences

| *Sample Means* (from population 1) | *Sample Means* (from population 2) | Sampling Distribution of Sample Mean Differences |
|---|---|---|
| 191.1 | 160.5 | 30.6 |
| 204.2 | 124 | 80.2 |
| ... | ... | ... |
| 187.4 | 142 | 45.4 |




**Need to know the following to make an inference about *Unknown* Population Mean Difference $\mu_1 - \mu_2$:**

Mean of **Sampling** Distribution $\approx E[\bar{X}_1 - \bar{X}_2] =$ __________

Standard deviation of **Sampling** Distribution $\approx SD[\bar{X}_1 - \bar{X}_2] =$ __________

Shape of **Sampling** Distribution is ____________________ when the 5 conditions below are met:

1. ____________________ and ____________________
2. ____________________ and ____________________
3. ____________________ and ____________________
4. ____________________ and ____________________
5. ____________________ is independent of ____________________.

## 3.1. Mean of the Sampling Distribution of Sample Mean Differences

Knowns:

- $\boldsymbol{E[\bar{X}_1] = \mu_1}$
- $\boldsymbol{E[\bar{X}_2] = \mu_2}$
- Property: $\boldsymbol{E[aY + bZ] = aE[Y] + bE[Z]}$

Proof:

- $\boldsymbol{E[\bar{X}_1 - \bar{X}_2] =}$ ________________
  $\boldsymbol{= \mu_1 - \mu_2}$

## 3.2. Standard Deviation of the Sampling Distribution of Sample Mean Differences

Knowns:

- $\boldsymbol{V[\bar{X}_1] = \frac{\sigma_1^2}{n_1}}$
- $\boldsymbol{V[\bar{X}_2] = \frac{\sigma_2^2}{n_2}}$
- Property: $\boldsymbol{V[aY + bZ] = a^2V[Y] + b^2V[Z]}$

Proof:

- $\boldsymbol{V[\bar{X}_1 - \bar{X}_2] =}$ ________________
  $$\boldsymbol{= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
- $\boldsymbol{SD[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

## 3.3. When is the Sampling Distribution of Sample Mean Differences normal?

Knowns:

- Central Limit Theorem (for sample means):
  - $\bar{X}_1 \sim N(mean = \mu_1, std = \frac{\sigma_1}{\sqrt{n_1}})$ when:
    - Sample 1 is randomly collected
    - $n_1 < 10\%$ of population 1
    - $n_1 > 30$ or population 1 distribution is normal
- Central Limit Theorem (for sample means):
  - $\bar{X}_2 \sim N(mean = \mu_2, std = \frac{\sigma_2}{\sqrt{n_2}})$ when:
    - Sample 2 is randomly collected
    - $n_2 < 10\%$ of population 2
    - $n_2 > 30$ or population 2 distribution is normal

- Property: If two random variables are ___________________ and normal, then the sum (difference) of these random variables is ________________________.

Central Limit Theorem (for sample mean DIFFERENCES):

If the following conditions hold, then the sampling distribution of sample mean differences will be approximately normal.

$$\bar{X}_1 - \bar{X}_2 \sim N(mean = \mu_1 - \mu_2, std = \sqrt{\boldsymbol{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}})$$

a. $n_1 > 30$ OR population 1 distribution (equivalently sample 1 distribution) is normal.
b. $n_2 > 30$ OR population 2 distribution (equivalently sample 2 distribution) is normal.
c. Sample 1 is collected randomly and $n_1 < 10\%$ of population 1 size
d. Sample 2 is collected randomly and $n_2 < 10\%$ of population 2 size
e. Observations in sample 1 and sample 2 are independent (between samples).

# 4. Conducting Inference on a Population Mean Difference $\mu_1 - \mu_2$

## 4.1. Creating a (1-α)100% Confidence Interval for $\mu_1 - \mu_2$

2. **Check the Central Limit Theorem conditions for sample mean differences.**
   a. $n_1 > 30$ OR population 1 distribution (equivalently sample 1 distribution) is normal.
   b. $n_2 > 30$ OR population 2 distribution (equivalently sample 2 distribution) is normal.
   c. Sample 1 is collected randomly and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is collected randomly and $n_2 < 10\%$ of population 2 size
   e. Observations in sample 1 and sample 2 are independent (between samples).

If they are met, then your confidence interval interpretations will not be invalid.

3. **If you know $\sigma_1$ and $\sigma_2$** the confidence interval for $\mu_1 - \mu_2$ is calculated by:

$$(point\ estimate) \pm (critical\ value)(standard\ error)$$

$$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

4. **<u>If you DON'T know both $\sigma_1$ and $\sigma_2$</u>** the confidence interval for $\boldsymbol{\mu_1 - \mu_2}$ is calculated by:

$$(point\ estimate) \pm (critical\ value)(standard\ error)$$

$$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\min\{n_1-1, n_2-1\}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

**<u>Go to Unit 11 slides (Section 4.1) for exercise in confidence interval building.</u>**

## 4.2. Conducting Hypothesis Testing for the claim $\mu_1 - \mu_2 = 0$

As with the other population parameters that we discussed in Unit 10 (ie. $\mu$ and *p)* we can make a conclusion about a **null** and **alternative hypothesis**

$H_0: (population\ parameter) = null\ value$

$H_A: (population\ parameter) \neq null\ value$

by using either a:

1. p-value,
2. test statistic, or
3. confidence interval.

In this case,

- our population parameter is $\boldsymbol{\mu_1 - \mu_2}$, and
- our null value is 0.

Like with $\mu$ and *p*, the way that we calculate the confidence interval or p-value is different, but the way that we use them to make a conclusion about our hypotheses is the same.

## 4.3. Conducting Hypothesis Testing for the claim $\mu_1 - \mu_2 \neq 0$ with a p-value – if you know $\sigma_1$ and $\sigma_2$.

1. **Set up two hypotheses.**

$H_0: \mu_1 - \mu_2 = 0$
$H_A: \mu_1 - \mu_2 \neq 0$

2. **Check the CLT Conditions (for Sample Means Differences)**
   a. $n_1 > 30$ OR population 1 distribution (equivalently sample 1 distribution) is normal.
   b. $n_2 > 30$ OR population 2 distribution (equivalently sample 2 distribution) is normal.
   c. Sample 1 is collected randomly and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is collected randomly and $n_2 < 10\%$ of population 2 size
   e. Observations in sample 1 and sample 2 are independent (between samples).

If these conditions hold, then the claims that you make with this analysis will be valid.

3. **Collect an Observed Sample Statistic:**

Collect a random sample from population 1 and population 2 and calculate the sample mean difference

$$\bar{x}_1 - \bar{x}_2$$

4. **Calculate the p-value**

One way: $p-value = \begin{Bmatrix} 2P(\bar{X} \geq (\bar{x}_1 - \bar{x}_2)), & if\ (\bar{x}_1 - \bar{x}_2) \geq 0 \\ 2P(\bar{X} \leq (\bar{x}_1 - \bar{x}_2)), & if\ (\bar{x}_1 - \bar{x}_2) \leq 0 \end{Bmatrix}$,

assuming $\bar{X} \sim N(mean =$ ____________________, $std\ deviation =$ ___________________$)$

We call $z = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ the **test statistic** for this hypothesis test.

Another way: $p-value = 2P\left(Z \geq \left| \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right|\right)$,

assuming $Z \sim N(mean =$ _______, $std\ deviation =$ _________$)$

## 5. Make a Decision

a. If $\mathbf{p-value} < \boldsymbol{\alpha}$, then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."

b. If $\mathbf{p-value} \geq \boldsymbol{\alpha}$, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

## 4.4. Conducting Hypothesis Testing for the claim $\boldsymbol{\mu_1 - \mu_2 \neq 0}$ with a p-value – if you DON'T know both $\sigma_1$ and $\sigma_2$.

1. **Set up two hypotheses.**

   $H_0: \mu_1 - \mu_2 = 0$
   $H_A: \mu_1 - \mu_2 \neq 0$

2. **Check the CLT Conditions (for Sample Means Differences)**
   a. $n_1 > 30$ OR population 1 distribution (equivalently sample 1 distribution) is normal.
   b. $n_2 > 30$ OR population 2 distribution (equivalently sample 2 distribution) is normal.
   c. Sample 1 is collected randomly and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is collected randomly and $n_2 < 10\%$ of population 2 size
   e. Observations in sample 1 and sample 2 are independent (between samples).

3. **Collect an Observed Sample Statistic:**

   Collect a random sample from population 1 and population 2 and calculate the sample mean difference

$$\bar{x}_1 - \bar{x}_2$$

4. **Calculate the p-value**

$$p-value = 2P\left(T_{\min\{n_1-1,n_2-1\}} \geq \left|\frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}\right|\right),$$

assuming $T \sim t-distribution\ with$ $\min\{n_1-1, n_2-1\}$degrees of freedom

We call $t = \frac{(\bar{x}_1-\bar{x}_2)-0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$ the **test statistic** for this hypothesis test.

5. **Make a Decision**

a. If $\mathbf{p - value < \alpha,}$ then we “reject the null hypothesis.” And we say that “there IS sufficient evidence to suggest the alternative hypothesis.”

b. If $\mathbf{p - value \geq \alpha}$, then we “fail to reject the null hypothesis.” And we say that “there IS NOT sufficient evidence to suggest the alternative hypothesis.”

## 4.5. Conducting Hypothesis Testing for the claim $\boldsymbol{\mu_1 - \mu_2 \neq 0}$ with a confidence interval.

1. **Set up two hypotheses.**
   $H_0: \mu_1 - \mu_2 = 0$
   $H_A: \mu_1 - \mu_2 \neq 0$
2. **Check the CLT Conditions (for Sample Means Differences)**
   a. $n_1 > 30$ OR population 1 distribution (equivalently sample 1 distribution) is normal.
   b. $n_2 > 30$ OR population 2 distribution (equivalently sample 2 distribution) is normal.
   c. Sample 1 is collected randomly and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is collected randomly and $n_2 < 10\%$ of population 2 size
   e. Observations in sample 1 and sample 2 are independent (between samples).

If they are met, then your confidence interval interpretations will not be invalid.

3. **Collect an Observed Sample Proportion Difference $\bar{x}_1 - \bar{x}_2$ and Create a Confidence Interval Around it:**

4. **Make a decision**
   a. If the null value (0 in this case) is ____________________________, then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."
   b. If the null value (0 in this case) is ____________________________, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

**Go to Unit 11 slides (Section 4.5) for exercise in confidence interval building.**

# 5. Properties of the Sampling Distribution of Sample Proportion Differences

**Population 1 of Categorical Data**

| | toss | value |
|---|---|---|
| 0 | head | 1 |
| 1 | tail | 0 |

Population 1 Proportion= ________

**Population 2 of Categorical Data**

| | toss | value |
|---|---|---|
| 0 | head | 1 |
| 1 | head | 1 |
| 2 | tail | 0 |

Population 2 Proportion= ________

**Sample 1 of Categorical Data**

| toss | value | |
|---|---|---|
| 0 | head | 1 |
| 0 | head | 1 |
| 1 | tail | 0 |
| 1 | tail | 0 |
| 1 | tail | 0 |
| 1 | tail | 0 |
| 0 | head | 1 |
| 0 | head | 1 |
| 0 | head | 1 |
| 0 | head | 1 |

Sample 1 Proportion= ________

**Sample 2 of Categorical Data**

| toss | value | |
|---|---|---|
| 0 | head | 1 |
| 0 | head | 1 |
| 0 | head | 1 |
| 2 | tail | 0 |
| 2 | tail | 0 |
| 0 | head | 1 |
| 2 | tail | 0 |
| 1 | head | 1 |

Sample 2 Proportion= ________

Collect Many Random Samples (all of size **n1=10**) drawn with replacement.

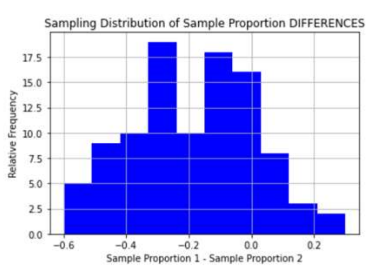
| Random Sample of n1=10 Values *(drawn with replacement from population 1)* | Random Sample of n1=10 Values *(drawn with replacement from population 1)* | ... | Random Sample of n1=10 Values *(drawn with replacement from population 1)* |
|---|---|---|---|
| 1 | 1 | ... | 1 |
| 1 | 0 | ... | 1 |
| 0 | 0 | ... | 1 |
| 0 | 0 | ... | 0 |
| 1 | 1 | ... | 0 |
| 0 | 1 | ... | 1 |
| 0 | 1 | ... | 0 |
| 1 | 0 | ... | 0 |
| 1 | 0 | ... | 0 |
| 1 | 0 | ... | 0 |

| Random Sample of n2=8 Values *(drawn with replacement from population 2)* | Random Sample of n2=8 Values *(drawn with replacement from population 2)* | ... | Random Sample of n2=8 Values *(drawn with replacement from population 2)* |
|---|---|---|---|
| 1 | 1 | ... | 1 |
| 1 | 0 | ... | 1 |
| 0 | 0 | ... | 1 |
| 0 | 0 | ... | 0 |
| 0 | 1 | ... | 1 |
| 1 | 1 | ... | 1 |
| 1 | 1 | ... | 1 |
| 1 | 0 | ... | 0 |

Collect Many Random Samples (all of size **n2=8**) drawn with replacement.

**Sampling Distribution of Sample Proportion Differences**

| *Sample Proportions (from population 1)* | *Sample Proportions (from population 2)* | Sampling Distribution of Sample Proportion Differences |
|---|---|---|
| 0.6 | 0.625 | -0.025 |
| 0.4 | 0.5 | -0.1 |
| ... | ... | |
| 0.4 | 0.75 | -0.35 |




**Need to know the following to make an inference about *Unknown* Population Proportion Difference $p_1 - p_2$:**

Mean of **Sampling** Distribution $\approx E[\hat{P}_1 - \hat{P}_2] =$ ________

Standard deviation of **Sampling** Distribution $\approx SD[\hat{P}_1 - \hat{P}_2] =$ ________

Shape of **Sampling** Distribution is ________________ when the 5 conditions below are met:

1. ______________________ OR ______________________
2. ______________________ OR ______________________
3. ______________________ and ______________________
4. ______________________ and ______________________
5. ________________ is independent of ________________.

## 5.1. Mean of the Sampling Distribution of Sample Proportion Differences

Knowns:

- $E[\widehat{P}_1] = p_1$
- $E[\widehat{P}_2] = p_2$
- Property: $E[aY + bZ] = aE[Y] + bE[Z]$

Proof:

- $E[\widehat{P}_1 - \widehat{P}_2] =$ ________________

  $= p_1 - p_2$

## 5.2. Standard Deviation of the Sampling Distribution of Sample Proportion Differences

Knowns:

- $V[\widehat{P}_1] = \frac{p_1(1-p_1)}{n_1}$
- $V[\widehat{P}_2] = \frac{p_2(1-p_2)}{n_2}$
- Property: $V[aY + bZ] = a^2V[Y] + b^2V[Z]$

Proof:

- $V[\widehat{P}_1 - \widehat{P}_2] =$ ________________

  $$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- $SD[\widehat{P}_1 - \widehat{P}_2] = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

## 3.3. When is the Sampling Distribution of Sample Proportion Differences normal?

Knowns:

- Central Limit Theorem (for sample proportions):
  - $\hat{P}_1 \sim N(mean = p_1, std = \sqrt{\frac{p_1(1-p_1)}{n_1}})$ when:
    - Sample 1 is randomly collected
    - $n_1 < 10\%$ of population 1
    - $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
- Central Limit Theorem (for sample proportions):
  - $\hat{P}_2 \sim N(mean = p_2, std = \sqrt{\frac{p_2(1-p_2)}{n_2}})$ when:
    - Sample 2 is randomly collected
    - $n_2 < 10\%$ of population 2
    - $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$

- Property: If two random variables are ___________________ and normal, then the sum/difference of these random variables is ________________________.

Central Limit Theorem (for sample proportion DIFFERENCES):

If the following conditions hold, then the sampling distribution of sample proportion differences will be approximately normal

(ie. $\hat{P}_1 - \hat{P}_2 \sim N(mean = p_1 - p_2, std = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}})$

a. $n_1 p_1 \geq 10$ and $n_1(1-p_1) \geq 10$
b. $n_2 p_2 \geq 10$ and $n_2(1-p_2) \geq 10$
c. Sample 1 is randomly selected and $n_1 < 10\%$ of population 1 size
d. Sample 2 is randomly selected and $n_2 < 10\%$ of population 2 size
e. Sample 1 is independent of sample 2.

# 6. Conducting Inference on a Population Mean Difference $p_1 - p_2$

## 6.1. Creating a (1-α)100% Confidence Interval for $p_1 - p_2$

2. **Check the Central Limit Theorem conditions for sample mean differences.**
   a. $n_1 p_1 \geq 10$ and $n_1(1 - p_1) \geq 10$
   b. $n_2 p_2 \geq 10$ and $n_2(1 - p_2) \geq 10$
   c. Sample 1 is randomly selected and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is randomly selected and $n_2 < 10\%$ of population 2 size
   e. Sample 1 is independent of sample 2.

If they are met, then your confidence interval interpretations will not be invalid.

3. The confidence interval for $p_1 - p_2$ is calculated by:

$$(point\ estimate) \pm (critical\ value)(standard\ error)$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}$$

**Note**: Usually we don't know $p_1$ and $p_2$ to plug into the conditions and standard error. So what we can do is plug in $\hat{p}_1$ for $p_1$ and $\hat{p}_2$ for $p_2$.

## 6.2. Conducting Hypothesis Testing for the claim $p_1 - p_2 \neq 0$

As with the other population parameters that we discussed in Unit 10 (ie. $\mu$ and *p)* we can make a conclusion about a **null** and **alternative hypothesis**

$H_0: (population\ parameter) = null\ value$

$H_A: (population\ parameter) \neq null\ value$

by using either a:

1. p-value,
2. test statistic, or
3. confidence interval.

In this case,

- our population parameter is $\boldsymbol{p_1 - p_2}$, and
- our null value is 0.

Like with $\mu$ and *p*, the way that we calculate the confidence interval or p-value is different, but the way that we use them to make a conclusion about our hypotheses is the same.

## 6.3. Conducting Hypothesis Testing for the claim $p_1 - p_2 \neq 0$ with a p-value

1. **Set up two hypotheses.**

   $H_0: p_1 - p_2 = 0$

   $H_A: p_1 - p_2 \neq 0$

2. **Check the CLT Conditions (for Sample Prportion Differences)**
   a. $n_1 p_1 \geq 10$ and $n_1(1 - p_1) \geq 10$
   b. $n_2 p_2 \geq 10$ and $n_2(1 - p_2) \geq 10$
   c. Sample 1 is randomly selected and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is randomly selected and $n_2 < 10\%$ of population 2 size
   e. Sample 1 is independent of sample 2.

   If they are met, then your confidence interval interpretations will not be invalid.

3. **Collect an Observed Sample Statistic:**

   Collect a random sample from population 1 and population 2 and calculate the sample proportion difference

   $$\hat{p}_1 - \hat{p}_2$$

4. **Calculate the p-value**

   One way: $p - value = \begin{Bmatrix} 2P\left(\hat{P}_1 - \hat{P}_2 \geq (\hat{p}_1 - \hat{p}_2)\right), & if\ (\hat{p}_1 - \hat{p}_2) \geq 0 \\ 2P\left(\hat{P}_1 - \hat{P}_2 \leq (\hat{p}_1 - \hat{p}_2)\right), & if\ (\hat{p}_1 - \hat{p}_2) \leq 0 \end{Bmatrix},$

   assuming $\hat{P}_1 - \hat{P}_2 \sim N(mean =$ ____________________, $std\ deviation =$ ______________________)

   We call $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$ the **test statistic** for this hypothesis test.

   Another way: $p - value = 2P\left(Z \geq \left| \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \right|\right),$

   assuming $Z \sim N(mean =$ _______, $std\ deviation =$ _________)

**Note**: Usually we don't know $p_1$ and $p_2$ to plug into the conditions and standard error. So what we can do is plug in $\hat{p}_1$ for $p_1$ and $\hat{p}_2$ for $p_2$.

## 5. Make a Decision

a. If $\mathbf{p-value < \alpha,}$ then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."

b. If $\mathbf{p-value \geq \alpha}$, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

## 6.4. Conducting Hypothesis Testing for the claim $p_1 - p_2 \neq 0$ with a confidence interval.

1. **Set up two hypotheses.**
   $H_0: \mu_1 - \mu_2 = 0$
   $H_A: \mu_1 - \mu_2 \neq 0$
2. **Check the CLT Conditions (for Sample Proportion Differences)**
   a. $n_1 p_1 \geq 10$ and $n_1(1 - p_1) \geq 10$
   b. $n_2 p_2 \geq 10$ and $n_2(1 - p_2) \geq 10$
   c. Sample 1 is randomly selected and $n_1 < 10\%$ of population 1 size
   d. Sample 2 is randomly selected and $n_2 < 10\%$ of population 2 size
   e. Sample 1 is independent of sample 2.

If they are met, then your confidence interval interpretations will not be invalid.

3. **Collect an Observed Sample Proportion Difference $\hat{p}_1 - \hat{p}_2$ and Create a Confidence Interval Around it:**

**Note**: Usually we don't know $p_1$ and $p_2$ to plug into the conditions and standard error. So what we can do is plug in $\hat{p}_1$ for $p_1$ and $\hat{p}_2$ for $p_2$.

4. **Make a decision**
   a. If the null value (0 in this case) is ____________________________, then we "reject the null hypothesis." And we say that "there IS sufficient evidence to suggest the alternative hypothesis."
   b. If the null value (0 in this case) is ____________________________, then we "fail to reject the null hypothesis." And we say that "there IS NOT sufficient evidence to suggest the alternative hypothesis."

**Go to Unit 11 slides (Section 6.4) for exercise in hypothesis testing and confidence interval building.**