Comparing Two Population Proportions

Hypothesis Testing for a difference between Two Population Proportions

Example – Two random samples of **80** and **120** students are selected from two colleges (A and B).

In the random sample of **80** students selected from college A , **20%** of them smoke.

In the random sample of **120** students selected from college B , **10%** of them smoke

From the sample data, we have **no doubt** that

Students in College A are **more likely to smoke** than students in College B.

However, it is **NOT my concern**!

Question - In the **population of all students** in both colleges, are students in College **A** **more likely to smoke** than students in College **B**?

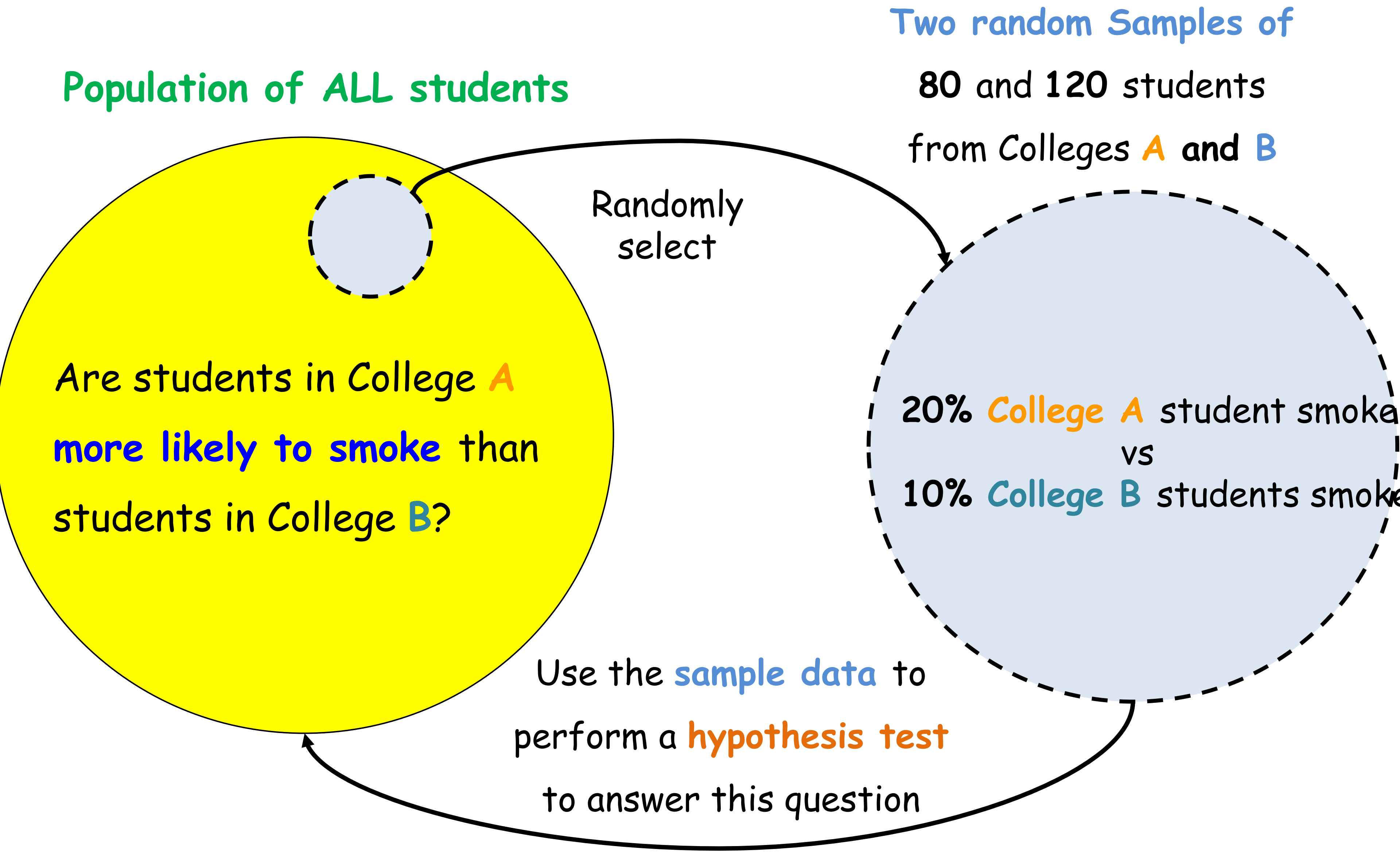
Population of ALL students

Are students in College **A** **more likely to smoke** than students in College **B**?

We don't know the answer to this question because we do not have the **population data**.

Question -In the population of all students in both colleges,

are students in College A more likely to smoke than students in College B?



Step 1 - State the Null and Alternative Hypotheses

Alternative Hypothesis (Ha)

When we want to compare two population proportions,

the **alternative hypothesis** says that

there is **a difference** between two population proportions.

In other words, the **proportion of the 1st population** can be

greater than (>), less than (<) or unequal to (≠)

the **proportion of the 2nd population**

Question of Interest:

Are students in College A **more likely to smoke** than students in College B?

Let's rewrite the question in terms of the proportions

The proportion of ***ALL*** College A students who smoke is **higher than** the proportion of ***ALL*** College B students who smoke

Alternative Hypothesis (Ha)

The proportion of ***ALL*** College A students who smoke is **higher than** the proportion of ***ALL*** College B students who smoke

Null Hypothesis (Ho)

When we want to compare two population proportions,

the **null hypothesis** says that

there is **NO difference** between two population proportions.

In other words, the **proportion of the 1st population**

is the **SAME** as the **proportion of the 2nd population**

Question of Interest:

Are students in College A **more likely to smoke** than students in College B?

Alternative Hypothesis (Ha)

The proportion of ***ALL*** College A students who smoke is **higher than** the proportion of ***ALL*** College B students who smoke

Null Hypothesis (Ho)

The proportion of ***ALL*** College A students who smoke is **the SAME as** the proportion of ***ALL*** College B students who smoke

We need to rewrite the **Null** and **Alternative** hypotheses symbolically

Define

P_A as the proportion of **ALL** College A students who smoke

P_B as the proportion of **ALL** College B students who smoke

Null Hypothesis (Ho)

The proportion of **ALL** College A students who smoke is **the SAME as** the proportion of **ALL** College B students who smoke

$$\boxed{P_A = P_B}$$

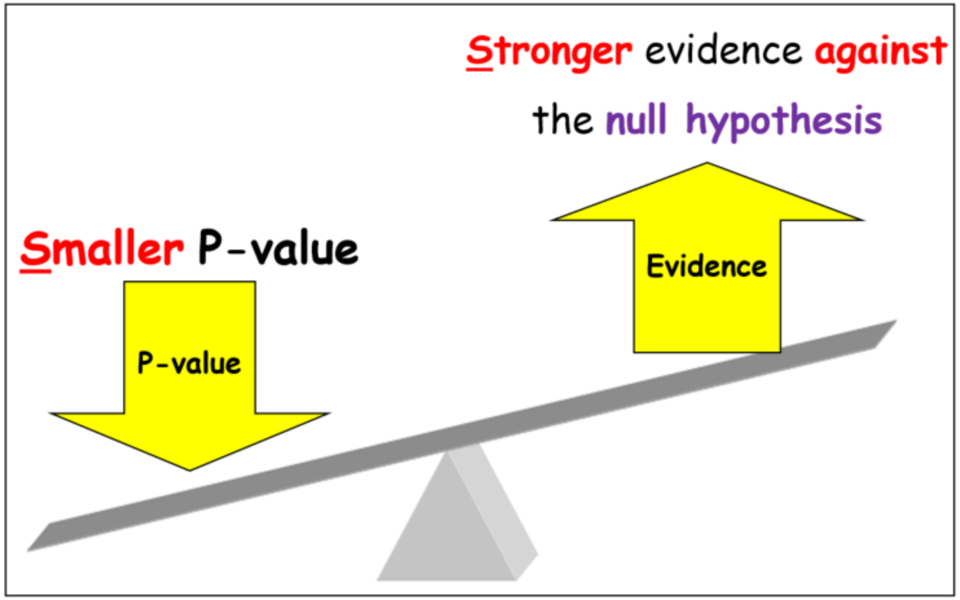
Alternative Hypothesis (Ha)

The proportion of **ALL** College A students who smoke is **higher than** the proportion of **ALL** College B students who smoke

$$\boxed{P_A > P_B}$$

Step 2 - Calculate the Test Statistic and P-value

- First, we **assume that the null hypothesis is true** that
- the **proportion of smokers** is the **SAME** in the **male** and **female populations**.
- Then, we try to find **evidence against** the **null hypothesis**.
- To do it, we need to get the **p-value**.
- To calculate the **p-value**, we need
- a stepping stone, **test statistic**.



For testing a difference between two population proportions,

we use z-statistic. The z-statistic is defined in the following:

Proportion for

1st /2nd sample

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Combined (or pooled)

proportion

Sample Size for

1st /2nd sample

To calculate the **z-statistic**, the first step is to decide

- which is the 1st sample ⟶ College A
- which is the 2nd sample ⟶ College B

To do that, we look at the hypotheses

$$Ho: P_A = P_B$$

$$Ha: \underset{\text{1st}}{P_A} > \underset{\text{2nd}}{P_B}$$

Next, we need to summarize the information.

Sample Size	Proportion of students smoking
$\boldsymbol{n_1} = \mathbf{80}$	$\bar{p}_1 = 20\% \rightarrow \mathbf{0.2}$
$\boldsymbol{n_2} = \mathbf{120}$	$\bar{p}_2 = 10\% \rightarrow \mathbf{0.1}$

Next, we can determine the **combined proportion ($\bar{p}$)**

	Sample Size	Proportion of students smoking	
College A (Sample 1)	$n_1 = 80$	$\bar{p}_1 = 20\% \rightarrow 0.2$	$20\%\ of\ 80$
College B (Sample 2)	$n_2 = 120$	$\bar{p}_2 = 10\% \rightarrow 0.1$	$10\%\ of\ 120$

Total $= 200$

Total students smoking $=$

$$\boxed{\text{Combined proportion of smoking} \quad \bar{p} =}$$

Calculate the test statistic

College A

$n_1 = 80$

$\bar{p}_1 = 0.2$

College B

$n_2 = 120$

$\bar{p}_2 = 0.1$

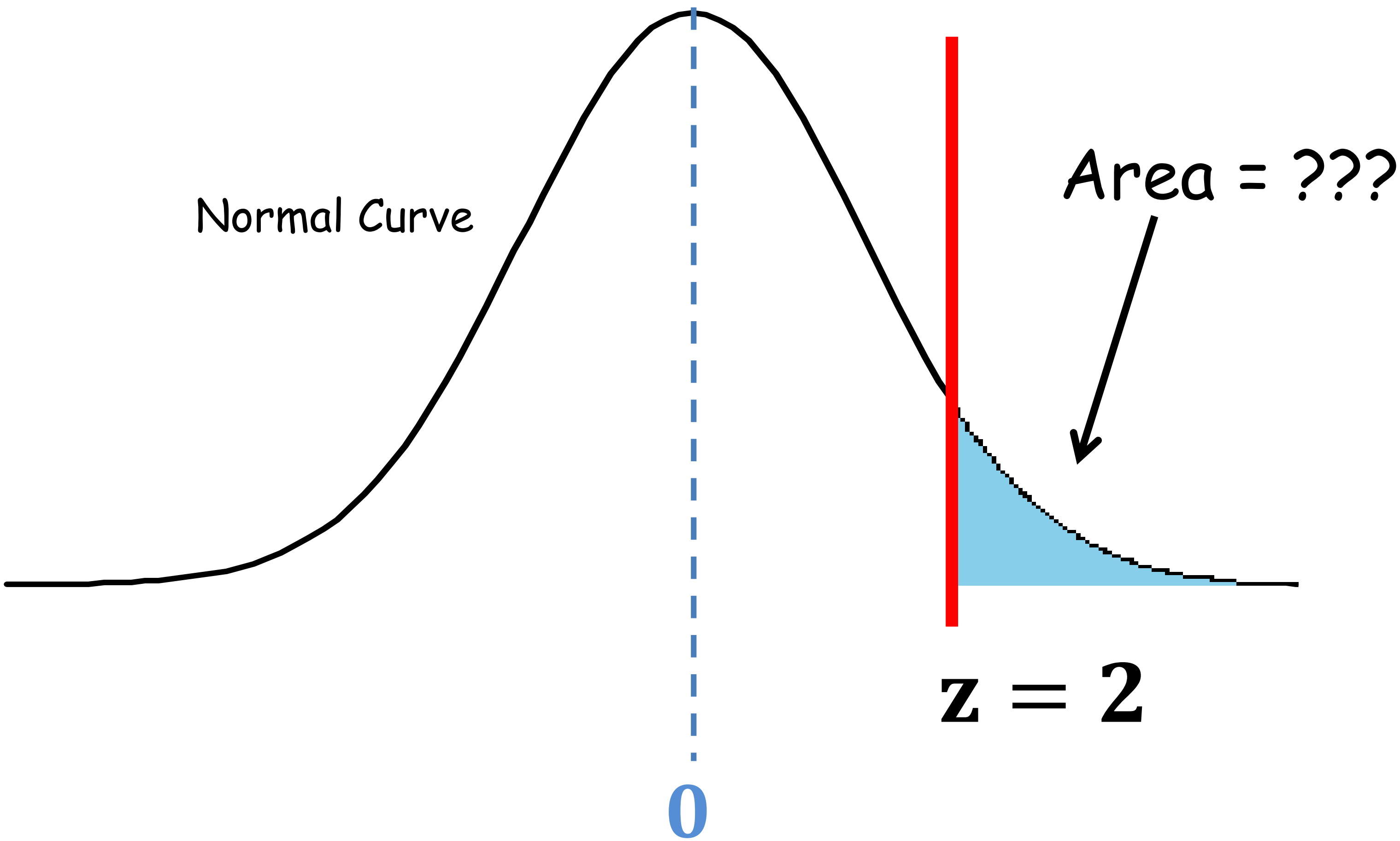
Combine proportion

$\bar{p} = 0.14$

$$z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Determine the P-value

$$Ho: P_A > P_B$$



Step 3 - State the Conclusion

p-value	**<**	Significance level
0.0228	**smaller**	0.05

Since the p-value (0.023) **smaller than** the significance level (0.05),

we **reject Ho**

At 5% significance level, the sample data provide **sufficient evidence** to conclude that the **proportion** of **ALL College A students** who smoke is **higher than** the **proportion** of **ALL College B students** who smoke in this college.

Confidence Interval for a Difference between Two Population Proportions

- In the previous question, we show that
- the **proportion** of all College A students who smoke **is higher than** the **proportion** of all College B students who smoke.
- Question: **How much** is **higher**?
- To answer the question, we need to know
- the **difference** between the proportion of all College A students who smoke and and the proportion of all College B students who smoke.
- Note that the **difference** as

$$\underbrace{p_A}_{\text{the proportion of all College A students who smoke}} - \underbrace{p_B}_{\text{the proportion of all College B students who smoke}}$$

- To estimate any unknown parameter in a population,
- we prefer to use a confidence interval that
- gives a range of possible values of the unknown parameter.

Question: Let's construct a 90% confidence interval to estimate

$$\underbrace{P_A}_{\text{the proportion of all College A students who smoke}} - \underbrace{P_B}_{\text{the proportion of all College B students who smoke}}$$

Here is the formula to calculate the confidence interval for the **difference** between **population proportions**.

Proportion for 1st sample

Proportion for 2nd sample

$$\underbrace{(\bar{p}_1 - \bar{p}_2)}_{\text{Difference between two sample proportions}} \pm \underbrace{\underbrace{z_c}_{\text{z-critical value}} \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}}_{\text{Margin of Error}}$$

Sample Size for 1st/2nd sample

To calculate the **confidence interval**, the first step is to decide

- which is the 1st sample → College A
- which is the 2nd sample → College B

To do that, we need to know how the difference is defined.

$$P_A - P_B$$

1st 2nd

Then, we need to summarize the information.

Sample Size	Proportion of students smoking
$n_1 = 80$	$\bar{p}_1 = 20\% \rightarrow 0.2$
$n_2 = 120$	$\bar{p}_2 = 10\% \rightarrow 0.1$

Next, we need to determine the **z-critical value**

$$(\bar{p}_1 - \bar{p}_2) \quad \pm \; z_c \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

1.645 @90%

Calculate the test statistic

College A

$n_1 = 80$

$\bar{p}_1 = 0.2$

College B

$n_2 = 120$

$\bar{p}_2 = 0.1$

$z_c = 1.645$

$$(\bar{p}_1 - \bar{p}_2) \pm \; z_c \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$$

Interpreting the Confidence Interval

First of all, we don't know the exact **difference** between

- the proportion of **College A** students who smoke and
- the proportion of **College B** students who smoke in the **population**.

From sample data we collect, we are quite confident that the **difference** is between 0.014 and 0.186.

$$\underbrace{P_A - P_B}_{???}$$

0.014 0.186

$$\boxed{P_{Male} - P_{Female}}$$

0.014 0.186

The interval includes **positive number (or numbers larger than 0).**

$$P_A - P_B > 0$$

What can you conclude if the **difference** between two population proportion is ***greater than zero***?

What can you conclude if the **difference** between two population proportion is **greater than zero**?

$$\underbrace{P_A}_{} - \underbrace{P_B}_{} > 0$$

Interpretation:

- In the population of all students at this college,
- we are **90% confident** that
- the proportion of ________________ students who smoke
- is **higher than**
- the proportion of ________________ students who smoke
- by between **1.4%** and **18.6%**

Assumptions and Conditions Required for valid Hypothesis Testing and Confidence Interval

- Not every dataset can be used to test hypotheses and to construct a confidence interval.
- If the dataset does not satisfy certain conditions, then the conclusion may not be value.
- To obtain valid conclusions from a hypothesis test and confidence interval when comparing two proportions, the sample data must satisfy certain conditions. **What are the required conditions?**

Mainly, it requires the observed number of individuals who

- **fall into the category of interest** and
- **do not fall into the category of interest**

in each random sample are both **at least 5**.

Let's get back to the smoking example, There are two groups: College A and College B

So, we need to know

- the number of College A students who smoke / do not
- the number of College B students who smoke / do not

	Sample Size	Proportion of students smoking	Number of students smoking
College A	**80**	**20%**	$20\% \; of \; 80 = 0.2 \times 80 = 16$
College B	**120**	**10%**	$10\% \; of \; 120 = 0.1 \times 120 = 12$

	Sample Size	Number of students smoking	Number of students not smoking
College A	80		
College B	120		

Since there are **at least 5** students who smoke / do not smoke in each group

any conclusions drawn from the hypothesis test and confidence interval are valid

Thank You