A researcher compares two groups:

- Group 1: 120 people, 72 support a policy
- Group 2: 100 people, 50 support the same policy

Let $\hat{p}_1 - \hat{p}_2$ represent the difference in sample proportions.

Which of the following best describes the **sampling distribution** of $\hat{p}_1 - \hat{p}_2$?

A) Approximately normal, mean = 0.60 – 0.50, standard deviation = $\sqrt{\frac{0.6(0.4)}{120} + \frac{0.5(0.5)}{100}}$

B) Approximately normal, mean = 0, standard deviation = $\sqrt{\frac{0.6(0.4)}{120} + \frac{0.5(0.5)}{100}}$

C) Not approximately normal because sample sizes are different

A

D) Approximately normal, mean = 0.60 – 0.50, standard deviation = $\frac{0.6-0.5}{\sqrt{120+100}}$

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

- $\hat{p}_1 - \hat{p}_2$ estimates the difference between two population proportions $p_1 - p_2$.
- The mean of the sampling distribution is the true difference:

  $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$.
- The standard deviation is

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

- The distribution is approximately normal if

  $n_1 p_1 \geq 10,\ n_1(1-p_1) \geq 10,\ n_2 p_2 \geq 10,\ n_2(1-p_2) \geq 10.$
- Samples must be **independent** (10% condition for each group if sampling without replacement).

Sample Proportion	样本比例
Difference in Sample Proportions	样本比例差
Population Proportion	总体比例
Sampling Distribution	抽样分布
Standard Error (of difference)	差值标准误

A researcher wants to compare the proportion of students who prefer online learning at two schools.

- School A: 120 out of 200 students prefer online learning.
- School B: 90 out of 150 students prefer online learning.

a) Find the **sample proportions** for each school.

b) Calculate the **difference in sample proportions** $(\hat{p}_A - \hat{p}_B)$.

c) Explain what the **sampling distribution of the difference in sample proportions** represents.

d) If both schools′ samples are **random and independent**, why can we use a **normal approximation** for the sampling distribution?

a) $\hat{p}_A = 120/200 = 0.6$, $\hat{p}_B = 90/150 = 0.6$

b) $\hat{p}_A - \hat{p}_B = 0.6 - 0.6 = 0$

c) Sampling distribution = all possible differences in sample proportions; shows variability around $p_A - p_B$

d) Normal approximation applies because samples are random, independent, and success/failure conditions are met ($np \geq 10$, $n(1-p) \geq 10$)

STRETCH and CHALLENGE

Question 3: Linking CI and Test

Two surveys report:

- Group 1: $\hat{p}_1 = 0.62, n_1 = 200$
- Group 2: $\hat{p}_2 = 0.55, n_2 = 180$

a) Construct a 95% confidence interval

b) Use the interval to test $H_0 : p_1 = p_2$

c) Explain why the CI and test give consistent conclusions

Question 4: Advanced Reasoning

A student claims:

> "If 0 is in the confidence interval for $p_1 - p_2$, then the two proportions are equal."

a) Evaluate the claim

b) Explain the correct statistical interpretation

c) Describe a scenario where the interval includes 0 but the proportions differ in reality

Question 3: Linking CI and Test

a) 95% CI:

$$0.62 - 0.55 \pm 1.96\sqrt{\frac{0.62(0.38)}{200} + \frac{0.55(0.45)}{180}} \approx -0.04 \text{ to } 0.18$$

b) Since 0 is inside CI → fail to reject H_0

c) The CI gives the same conclusion as a 2-sided test at $\alpha = 0.05$

Question 4: Advanced Reasoning

a) Claim is **false**

b) Correct interpretation: If 0 is in the confidence interval, the data do not provide strong evidence of a difference. It does **not prove equality**.

c) Scenario: small sample sizes → wide CI including 0, even though true proportions differ slightly

1. **Sampling Distribution** — 抽样分布
2. **Sample Proportion** — 样本比例
3. **Difference of Proportions** — 比例差
4. **Population Proportion** — 总体比例
5. **Mean of Sampling Distribution** — 抽样分布的均值
6. **Standard Error (SE)** — 标准误
7. **Approximate Normality** — 近似正态性
8. **Central Limit Theorem (CLT)** — 中心极限定理
9. **Independence Condition** — 独立条件
10. **Success/Failure Condition** — 成功/失败条件（$np \geq 10$, $n(1-p) \geq 10$）
11. **Large Sample Condition** — 大样本条件
12. **Z-Statistic** — Z统计量
13. **Confidence Interval** — 置信区间
14. **Hypothesis Test** — 假设检验
15. **Pooled Proportion** — 合并比例

1. **Sampling Distribution of $\hat{p}_1 - \hat{p}_2$** — The distribution of differences between sample proportions over many samples.
2. **Mean of the Distribution** — Equal to the difference in population proportions: $\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$.
3. **Standard Error (SE)** — Measures variability:

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

4. **Approximate Normality** — With large enough samples, the sampling distribution is nearly normal (Central Limit Theorem).
5. **Independence Condition** — Samples must be independent.
6. **Success/Failure Condition** — Each group should have at least 10 successes and 10 failures ($np \geq 10$, $n(1-p) \geq 10$).
7. **Confidence Intervals** — Use the sampling distribution to estimate $p_1 - p_2$ with a range of plausible values.
8. **Hypothesis Tests** — Use the distribution to test if $p_1 = p_2$ or another specified difference.
9. **Pooled vs Unpooled** — For hypothesis testing under $H_0 : p_1 = p_2$, a pooled proportion is used to calculate SE.
10. **Interpretation in Context** — Always relate statistical results back to the real-world scenario.

A university wants to compare the proportion of students who prefer online classes between two departments.

- **Department A:** 75 out of 120 students prefer online classes
- **Department B:** 60 out of 100 students prefer online classes

Let $\hat{p}_A$ and $\hat{p}_B$ represent the sample proportions.

a) Calculate $\hat{p}_A$, $\hat{p}_B$, and $\hat{p}_A - \hat{p}_B$.

b) Check whether the conditions for using a normal model for $\hat{p}_A - \hat{p}_B$ are satisfied.

c) Construct a 95% confidence interval for $p_A - p_B$.

d) Interpret your interval in context.

e) Based on the interval, is there convincing evidence that the proportions differ? Explain.

f) If the sample sizes were smaller (e.g., 12 and 10 students), explain how this would affect the validity of the interval.

a) $\hat{p}_A = 0.625, \hat{p}_B = 0.60, \hat{p}_A - \hat{p}_B = 0.025$

b) Conditions satisfied (independence + success/failure ≥ 10)

c) 95% CI: –0.104 to 0.154

d) We are 95% confident the true difference is between –10.4% and 15.4%

e) No convincing evidence of a difference (CI includes 0)

f) Smaller sample sizes → normal approximation may not hold; interval may be invalid

True or False

- Using small, non-independent samples can lead to misleading conclusions for the difference of proportions.

True

True or False

- When using a pooled proportion for a hypothesis test, the standard error may be smaller than the standard error for a confidence interval.

True

True or False

. The normal approximation for $\hat{p}_1 - \hat{p}_2$ is valid even if one sample has 8 successes and 12 failures.

False

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

1. Define sample proportions:

$$\hat{p}_1 = \frac{X_1}{\_\_\_\_}, \quad \hat{p}_2 = \frac{X_2}{\_\_\_\_}$$

Where $X_1 \sim Binomial(n_1, p_1)$, $X_2 \sim Binomial(n_2, p_2)$.

---

2. Variance of a sample proportion:

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X) = \frac{1}{n^2}(n\_\_\_\_(1 - \_\_\_\_)) = \frac{\_\_\_\_(1 - \_\_\_\_)}{n}$$

So:

$$Var(\hat{p}_1) = \frac{p_1(1 - \_\_\_\_)}{n_1}, \quad Var(\hat{p}_2) = \frac{p_2(1 - \_\_\_\_)}{n_2}$$

---

3. Variance of the difference of independent variables:

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) \quad \text{because the samples are } \_\_\_\_$$

---

4. Combine formulas:

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{\_\_\_\_} + \frac{p_2(1-p_2)}{\_\_\_\_}$$

Step 1: Define sample proportions

For independent samples:

$$\hat{p}_1 = \frac{X_1}{n_1}, \quad \hat{p}_2 = \frac{X_2}{n_2}$$

Where $X_1 \sim Binomial(n_1, p_1), X_2 \sim Binomial(n_2, p_2)$.

Step 2: Recall variance of a sample proportion

For a binomial variable:

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{1}{n^2}(np(1-p)) = \frac{p(1-p)}{n}$$

So:

$$Var(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}, \quad Var(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

Step 3: Variance of the difference of independent variables

If two random variables are independent:

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2)$$

- This works because independence ⇒ covariance = 0.

Step 4: Combine the formulas

$$Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$