A researcher collects data on **smoking status (smoker/non-smoker)** and **incidence of lung disease (yes/no)** for 500 adults. They want to test whether smoking and lung disease are related. Which procedure is correct?

A) One-sample z-test for proportion

B) Two-sample z-test for proportions

C) Chi-square goodness-of-fit test

D) Chi-square test for independence

D

An agronomist is an expert in soil management and crop production. A certain state hires an agronomist to investigate whether there is a linear relationship between a wheat stalk's height and the yield of wheat. The agronomist collected data and used the data to test the claim that there is a linear relationship at a significance level of $\alpha = 0.05$. The agronomist tested the following hypotheses.

$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$

The test yielded a $p$-value of 0.25. Which of the following is a correct conclusion about the claim?

Ⓐ The null hypothesis is rejected because $0.25 > 0.05$. There is sufficient evidence to suggest that there is a linear relationship between a wheat stalk's height and its yield.

Ⓑ The null hypothesis is not rejected because $0.25 > 0.05$. There is sufficient evidence to suggest that there is a linear relationship between a wheat stalk's height and its yield.

Ⓒ The null hypothesis is rejected because $0.25 > 0.05$. There is not sufficient evidence to suggest that there is a linear relationship between a wheat stalk's height and its yield.

Ⓓ The null hypothesis is not rejected because $0.25 > 0.05$. There is not sufficient evidence to suggest that there is a linear relationship between a wheat stalk's height and its yield.

Ⓔ The null hypothesis is accepted because $0.25 > 0.05$. There is sufficient evidence to suggest that there is not a linear relationship between a wheat stalk's height and its yield.

D

3 Mark for Review

A scientist is investigating whether percent concentration can be used to predict density in apple juice. A scientist selected a random sample of 12 apple juice varieties and recorded the density, in pounds per cubic inch, and the percent concentration of each apple juice variety. The scientist wants to estimate the mean change in the density, in pounds per cubic inch, for each increase of 1 percent concentration of apple juice.

Assuming the conditions for inference have been met, which of the following inference procedures is most appropriate for this investigation?

(A) A linear regression $t$-interval for slope

(B) A matched-pairs $t$-interval for a mean difference

(C) A two-sample $t$-interval for a difference between means

(D) A one-sample $t$-test for means

(E) A two-sample $z$-interval for a difference between proportions

A

4 Mark for Review

Researchers are examining the relationship between hours of sleep and athletic performance among college athletes. Athletic performance will be measured on a numeric scale, with greater numbers indicating better performance. The researchers expect that the more hours the athletes sleep, the better they will perform. Assuming all conditions for inference are met, the researchers will create a 95 percent confidence interval for the slope of the regression line for predicting athletic performance from amount of sleep.

For which of the following would the confidence interval support the researchers' expectations?

(A) The confidence interval includes only positive values. ✓

(B) The confidence interval includes only negative values.

(C) The confidence interval has a width less than 1.

(D) The confidence interval has a width greater than 1.

(E) The confidence interval includes the value 0.

A

1. **Identify the research question:** Determine whether you are estimating or testing, and whether the focus is on a mean, proportion, or relationship.
2. **Determine the type of data:** Identify if the data are quantitative (numeric) or categorical (counts/proportions).
3. **Consider the design:** Check whether the sample is independent, paired, or from multiple groups/populations.
4. **Select the procedure:** Match the scenario to the correct inference method:
   - 1-mean or 2-mean t-tests/intervals for quantitative data
   - 1-proportion or 2-proportion z-tests/intervals for categorical data
   - Chi-square tests for categorical counts
   - Regression t-tests/intervals for slopes in quantitative relationships
5. **Check conditions and interpret:** Verify assumptions (normality, sample size, independence, expected counts) and plan to interpret results in context.

- **Population** – 总体
- **Sample** – 样本
- **Bias** – 偏差
- **Random sampling** – 随机抽样
- **Control group** – 对照组

A teacher wants to investigate whether **the number of hours students study per week** predicts their **statistics exam scores**. She collects data from 30 students.

| Student | Hours Studied (x) | Exam Score (y) |
|---|---|---|
| 1 | 2 | 68 |
| 2 | 5 | 78 |
| ... | ... | ... |
| 30 | 10 | 92 |

## Tasks

### (a) Procedure Selection

Which inference procedure is appropriate to test whether study hours **significantly predict exam scores**? Justify your choice.

### (b) Hypotheses

State the null and alternative hypotheses in terms of the **population slope** **($\beta_1$)**.

### (c) Conditions

Check the conditions for performing a **linear regression inference**.

### (d) Test Statistic & p-value

Given $\hat{\beta}_1 = 2.1$ and $SE_{\hat{\beta}_1} = 0.8$, calculate the **t-statistic** and approximate the **p-value**.

### (e) Conclusion

Write a clear conclusion in context about whether **hours studied predicts exam scores**.

**(a) Procedure Selection:**

- **t-test for slope in linear regression** — tests if $\beta_1 \neq 0$, i.e., if x predicts y.

**(b) Hypotheses:**

- $H_0 : \beta_1 = 0$ (no linear relationship)
- $H_a : \beta_1 \neq 0$ (hours studied predicts exam score)

**(c) Conditions:**

1. **Linearity** — scatterplot looks roughly linear
2. **Independence** — observations are independent
3. **Normality of residuals** — residuals roughly normal
4. **Constant variance** — residuals have roughly equal spread

**(d) Test Statistic & p-value:**

$$t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} = \frac{2.1}{0.8} \approx 2.625$$

- df = n − 2 = 28
- Two-tailed p-value ≈ 0.013

**(e) Conclusion:**

- p < 0.05 → reject $H_0$
- There is evidence that **more hours studied is associated with higher exam scores**.

1. Slope ($\beta_1$) — 斜率
2. Intercept ($\beta_0$) — 截距
3. Least-squares regression line — 最小二乘回
4. Predicted value ($\hat{y}$) — 预测值
5. Null hypothesis ($H_0$) — 原假设
6. Alternative hypothesis ($H_a$) — 备择假设
7. t-test / Test statistic (t) — t检验 / t值
8. p-value — p值
9. Confidence interval — 置信区间
10. Linearity — 线性关系

### 1. Linear Relationship

- The slope ($\beta_1$) measures **how much $y$ changes for a one-unit increase in $x$**.
- The regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

---

### 2. Hypothesis Testing

- **Null hypothesis:** $H_0 : \beta_1 = 0$ → no linear association.
- **Alternative hypothesis:** $H_a : \beta_1 \neq 0$ (two-tailed) or directional ($>$ or $<$).
- **t-test** is used to test if the slope is significantly different from 0.

---

### 3. Confidence Intervals

- A **confidence interval for $\beta_1$** gives a range of plausible slopes in the population.
- If the CI does **not include 0**, the slope is statistically significant.

---

### 4. Conditions for Inference

- **Linearity:** Relationship between $x$ and $y$ is roughly linear.
- **Independence:** Observations are independent.
- **Normality of residuals:** Residuals follow an approximately normal distribution.
- **Constant variance (homoscedasticity):** Residuals have roughly equal spread for all $x$.

---

### 5. Interpretation

- Slope describes the **direction (positive/negative) and strength** of the linear relationship.
- Always report results **in context** of the problem.
- Statistically significant slope → evidence of an association between $x$ and $y$.

## FRQ: Study Hours and Exam Scores

A teacher wants to investigate whether the **number of hours students study per week** predicts their **statistics exam scores**. She collects data from 25 students.

| Student | Hours Studied (x) | Exam Score (y) |
|---|---|---|
| 1 | 2 | 65 |
| 2 | 4 | 70 |
| ... | ... | ... |
| 25 | 12 | 95 |

### Tasks

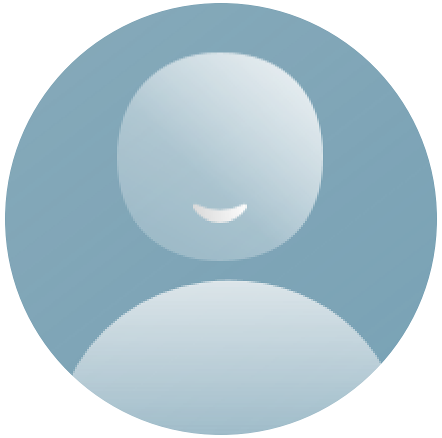
**(a) Procedure Selection**
Which inference procedure is appropriate to test whether study hours significantly predict exam scores? Justify your choice.

**(b) Hypotheses**
State the null and alternative hypotheses in terms of the population slope $\beta_1$.

**(c) Conditions**
List and explain the conditions required for inference on the slope.

**(d) Test Statistic & p-value**
Given $\hat{\beta}_1 = 2.5$ and $SE_{\hat{\beta}_1} = 1.0$, calculate the **t-statistic** and approximate the **p-value**.

**(e) Conclusion**
Write a clear conclusion in context about whether **study hours predict exam scores**.



**(a) Procedure Selection:**

- **t-test for slope in linear regression** — tests if $\beta_1 \neq 0$.

**(b) Hypotheses:**

- $H_0 : \beta_1 = 0$ (no linear relationship)
- $H_a : \beta_1 \neq 0$ (positive linear relationship)

**(c) Conditions:**

- **Linearity:** $x$ and $y$ roughly linear
- **Independence:** observations are independent
- **Normality of residuals:** residuals approximately normal
- **Constant variance:** residuals have roughly equal spread

**(d) Test Statistic & p-value:**

$$t = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} = \frac{2.5}{1.0} = 2.5$$

- df = n – 2 = 23
- p-value ≈ 0.02 (two-tailed)

**(e) Conclusion:**

- $p < 0.05$ → reject $H_0$
- There is evidence that **more study hours are associated with higher exam scores.**