

Validating Data Models with Statistical Analysis

Part 1: Office Roleplay Dialogue

Scenario: A Data Analyst, Hiroshi, is discussing statistical methods for validating a data model with his colleague, Sophia.

Sophia: Hey Hiroshi, have you started working on the statistical tests for our new sales forecast model?

Hiroshi: Yes! I'm using **hypothesis testing** to check if our new model performs better than the previous one.

Sophia: That makes sense. What's the **p-value** showing? Is the difference statistically significant?

Hiroshi: The **p-value** is 0.03, so it's below the standard 0.05 threshold. That means we can reject the null hypothesis.

Sophia: That's a good sign. Did you calculate the **confidence interval** to see the possible range of error?

Hiroshi: Yes, the **confidence interval** shows that the model's predictions are stable within a 95% range.

Sophia: Great! Are you also running a **regression analysis** to identify key factors affecting sales?

Hiroshi: Exactly! The regression results show that advertising spend has the strongest impact on sales growth.

Sophia: Interesting. Would it make sense to run an **ANOVA (Analysis of Variance)** to compare regional sales trends?

Hiroshi: Yes! ANOVA will help us determine if there are significant differences in sales performance across different regions.

Sophia: Perfect. Let's finalize these tests, document the findings, and present them at the next meeting.

Hiroshi: Sounds good! I'll finish up the report and share it with you later today.

Part 2: Comprehension Questions

1. What is Hiroshi using hypothesis testing for?

- (A) To change the company's branding strategy
- (B) To increase the sample size
- (C) To adjust employee salaries
- (D) To check if the new model performs better than the old one

2. What does a p-value of 0.03 indicate?

- (A) The data is inaccurate
- (B) The result is statistically significant
- (C) The hypothesis is always correct
- (D) The confidence interval is invalid

3. Why does Hiroshi use regression analysis?

- (A) To identify key factors affecting sales
- (B) To replace all missing values
- (C) To determine employee satisfaction levels
- (D) To merge multiple datasets into one file

4. What is ANOVA used for in this analysis?

- (A) To remove duplicate records

- (B) To check for missing values
 - (C) To compare regional sales trends
 - (D) To create a new data visualization
-

Part 3: Key Vocabulary Definitions in Japanese

1. **Hypothesis Testing (仮説検定)** – データを分析して、ある仮説が統計的に正しいかどうかを検証する手法。
 2. **P-Value (P 値・有意確率)** – 仮説検定の結果を示す数値で、通常 0.05 未満であれば統計的に有意とされる。
 3. **Confidence Interval (信頼区間)** – 推定値の誤差範囲を示し、どの程度の確率で真の値がその範囲内にあるかを示す指標。
 4. **Regression Analysis (回帰分析)** – 変数間の関係进行分析し、ある要因がどの程度結果に影響を与えるかを測定する方法。
 5. **ANOVA (Analysis of Variance) (分散分析)** – グループ間の平均値の違いが統計的に有意かどうかを判定する分析手法。
-

Part 4: Questions & Correct Answers

1. **What is Hiroshi using hypothesis testing for?**



(D) To check if the new model performs better than the old one

2. What does a p-value of 0.03 indicate?

☒ (B) The result is statistically significant

3. Why does Hiroshi use regression analysis?

☒ (A) To identify key factors affecting sales

4. What is ANOVA used for in this analysis?

☒ (C) To compare regional sales trends