

Hypothesis testing

Định nghĩa

Sử dụng dữ liệu mẫu để đưa ra suy luận về một tổng thể - hay liệu có đủ bằng chứng trong mẫu của mình để bác bỏ một tuyên bố (giả thuyết gốc) về toàn bộ tổng thể hay không.

Các thành phần

1. Giả thuyết:

- **Giả thuyết Gốc (H_0 - Null Hypothesis):** Giả định ban đầu của bài toán, rằng không có khác biệt nào giữa các quan hệ.
- **Ví dụ:**
 - "Chiều cao trung bình của nam và nữ là như nhau."
 - "Không có mối quan hệ giữa tập thể dục và giảm cân."
 - "Thuốc mới không có tác dụng đối với huyết áp."
- **Giả thuyết Đối (H_1):** Mâu thuẫn với H_0 , rằng có khác biệt trong các quan hệ, gồm 2 loại:
 - **Hai phía (\neq):** Không rõ hướng.
 - Ví dụ: "Chiều cao trung bình của nam và nữ là *khác nhau*."
 - **Một phía (trái phải):** Có rõ hướng.
 - Ví dụ trái: "Chiều cao trung bình của nữ *thấp hơn* chiều cao trung bình của nam."
 - Ví dụ phải: "Chiều cao trung bình của nam *cao hơn* chiều cao trung bình của nữ."

2. Mức Ý nghĩa - ngưỡng sai(α):

- Thường là 5% hoặc 1%
- Nó thể hiện **xác suất mắc phải Lỗi Loại I** - xác suất kết luận sai H_0
- Có thể hiểu là ngưỡng sai sót nếu kết luận sai H_0
 - Nếu $\alpha = 0.05$, chấp nhận 5% khả năng bác bỏ sai giả thuyết gốc khi nó thực sự đúng.

3. Các phương pháp thống kê Kiểm định:

- Các phương pháp thống kê kiểm định phụ thuộc vào loại dữ liệu và kiểm định giả thuyết bạn đang thực hiện (ví dụ: thống kê z, thống kê t, thống kê F, thống kê khi bình phương).
- Các phương pháp này sẽ cho ra tóm tắt mức độ dữ liệu mẫu ban đầu khác biệt so với những gì mong đợi H_0 đúng.

4. Giá trị P (P-value):

- Giả sử H_0 đúng - P là xác suất trong thực tế có thể thu thập được dữ liệu có mức độ cực đoan bằng hoặc hơn so với mức độ cực đoan của bộ dữ liệu đã thu thập được để phân tích.
- Được tính toán dựa trên kết quả từ một trong các phương pháp thống kê và phân phối của thống kê đó theo giả thuyết gốc.
- P nhỏ \rightarrow khả năng tìm được dữ liệu cực đoan bằng hoặc hơn dữ liệu hiện tại là thấp \rightarrow góp phần chống lại H_0
- P nhỏ \rightarrow khả năng tìm được dữ liệu cực đoan bằng hoặc hơn dữ liệu hiện tại là cao \rightarrow không góp phần chống lại H_0

5. Quy tắc Quyết định (So sánh giá trị p với α):

- Nếu giá trị $p \leq \alpha$: Bác bỏ H_0 hay đủ bằng chứng để ủng hộ H_1 với mức sai số α
- Nếu giá trị $p > \alpha$: Không bác bỏ H_0

Các Loại Lỗi

- **Lỗi Loại I (False positive):** Bác bỏ H_0 khi H_0 đúng trong thực tế
 - **Xác suất:** Xác suất mắc Lỗi Loại I bằng với mức ý nghĩa (α), thường là giá trị nhỏ.
- **Lỗi Loại II (False negative):** Không bác bỏ H_0 khi H_0 sai trong thực tế
 - **Xác suất:** Xác suất mắc Lỗi Loại II được ký hiệu là β (beta). β thường không xác định và khó tính toán trực tiếp hơn.

Các Bước trong Kiểm định Giả thuyết

1. **Phát biểu Giả thuyết:**
 - Xác định rõ ràng H_0 và giả thuyết đối H_1 . Quyết định xem đó là kiểm định một phía hay hai phía.
2. **Đặt Mức Ý nghĩa (α):**
 - Thường là 1% hoặc 5%
3. **Chọn phương pháp và tính Thống kê Kiểm định:**
 - Chọn thống kê kiểm định phù hợp dựa trên dữ liệu và giả thuyết.
 - Tính thống kê kiểm định từ dữ liệu mẫu của bạn.
4. **Xác định Giá trị P:**
 - Tính giá trị p liên quan đến thống kê kiểm định.
5. **Đưa ra Quyết định:**
 - So sánh giá trị p với mức ý nghĩa (α).
 - Nếu giá trị $p \leq \alpha$, bác bỏ H_0 .
 - Nếu giá trị $p > \alpha$, không bác bỏ H_0 .
6. **Rút ra Kết luận:**
 - Phát biểu kết luận của bạn trong bối cảnh câu hỏi nghiên cứu ban đầu. Giải thích ý nghĩa thực tế của quyết định của bạn. Tránh nói rằng bạn "chấp nhận" H_0 ; thay vào đó, hãy nói rằng bạn "không bác bỏ" H_0 .

Các phương pháp thống kê kiểm định

1. Thống kê Z (Z-statistic)

- **Công thức

$$Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- \bar{x} : Trung bình mẫu (sample mean)
- μ_0 : Giá trị trung bình tổng thể được giả định trong giả thuyết gốc (population mean under H_0)
- σ : Độ lệch chuẩn của tổng thể (population standard deviation)
- n : Kích thước mẫu (sample size)
- Thống kê Z đo lường số lượng độ lệch chuẩn mà trung bình mẫu (\bar{x}) cách xa giá trị trung bình tổng thể giả định (μ_0) theo giả thuyết gốc.

2. Thống kê T (T-statistic)

- Công thức LaTeX:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- \bar{x} : Trung bình mẫu (sample mean)
 - μ_0 : Giá trị trung bình tổng thể được giả định trong giả thuyết gốc (population mean under H_0)
 - s : Độ lệch chuẩn mẫu (sample standard deviation)
 - n : Kích thước mẫu (sample size)
- Tương tự như thống kê Z, thống kê T cũng đo lường số lượng "sai số chuẩn" (standard errors) mà trung bình mẫu (\bar{x}) cách xa giá trị trung bình tổng thể giả định (μ_0) theo giả thuyết gốc. Tuy nhiên, nó sử dụng độ lệch chuẩn mẫu (s) thay vì độ lệch chuẩn tổng thể (σ).

3. Thống kê Chi bình phương (χ^2)

- Công thức

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

- χ^2 : Thống kê Khi bình phương
 - \sum : Ký hiệu tổng
 - k : Số lượng danh mục hoặc nhóm (number of categories)
 - O_i : Tần số quan sát được (observed frequency) cho danh mục thứ i
 - E_i : Tần số mong đợi (expected frequency) cho danh mục thứ i theo giả thuyết gốc
- Kiểm định tính phù hợp của Khi bình phương đo lường sự khác biệt giữa tần số quan sát được trong các danh mục và tần số mong đợi nếu một phân phối hoặc tỷ lệ cụ thể (được nêu trong giả thuyết gốc) là đúng cho tổng thể. Thống kê χ^2 lớn cho thấy sự khác biệt lớn giữa quan sát và mong đợi.