KIỂM ĐỊNH GIẢ THUYẾT (1) (HYPOTHESIS TESTING)

Nguyễn Văn Hạnh

AI Academy Vietnam

Tháng 08 năm 2022

Nội dung

1. Bài toán kiểm định giả thuyết
2. Kiểm định giá trị trung bình
3. Ví dụ minh hoạ và Thực hành trên Python

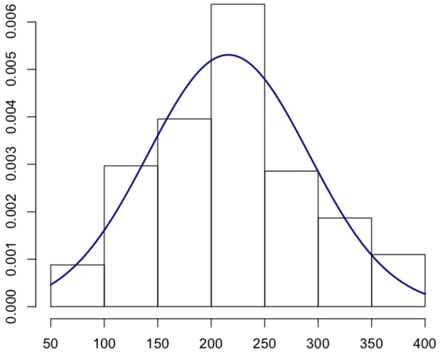
Đặt vấn đề

Biết rằng tiền điện trung bình của các hộ gia đình tại một khu vực (gồm khoảng 1 triệu hộ) trong tháng 5/2020 là 200 nghìn đồng. Khảo sát tiền điện tháng 6/2020 của 200 hộ, ta thu được bảng số liệu như sau:

196.65	468.75	320.50	300.50	213.05	140.60	290.00	216.95	360.50	317.95	195.55
220.50	255.60	289.00	194.55	374.25	382.05	185.55	219.10	215.60	220.00	186.75
97.80	340.50	88.50	209.50	234.04	333.00	291.10	108.50	245.00	184.00	153.50
219.50	214.15	155.20	140.40	108.50	410.00	125.50	220.30	160.00	300.50	310.20
244.40	194.50	210.20	360.00	456.50	237.40	235.00	203.25	109.20	240.15	260.50
275.50	101.55	455.50	246.25	291.55	262.00	378.65	194.50	248.00	262.92	85.75
248.00	204.75	310.70	213.10	320.50	125.60	110.25	77.35	119.50	313.50	222.00
388.10	110.50	160.00	210.00	310.30	380.10	281.00	105.35	280.15	188.80	272.50
103.40	213.50	280.50	119.50	166.10	180.50	212.00	154.75	100.50	452.60	436.35
225.00	124.30	170.00	127.35	107.90	140.00	195.00	315.10	241.05	168.00	120.50
223.95	237.05	285.45	100.50	228.55	248.70	175.80	466.05	219.00	216.00	425.50
390.00	176.85	240.50	226.00	108.70	160.00	470.50	225.00	440.00	265.00	162.80
260.50	175.80	73.05	460.50	263.60	59.50	198.00	416.50	315.50	155.00	190.00
158.50	225.00	266.70	153.60	238.00	297.60	201.75	240.50	270.90	196.65	299.20
70.50	125.60	100.40	240.00	240.00	224.05	194.00	247.00	325.40	102.20	166.10
361.00	430.00	240.00	250.50	470.00	157.75	98.40	236.50	230.85	317.65	200.70
165.00	350.50	319.15	275.88	203.05	234.50	220.75	180.50	436.50	403.00	460.50
220.00	103.50	222.15	170.50	224.15	460.00	260.40	200.50	311.40	260.00	251.55
100.60	212.20									

Đặt vấn đề

Phân bố của dữ liệu có thể xấp xỉ bởi phân bố chuẩn:



Đặt vấn đề

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tháng 6/2020 tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$.

Đặt vấn đề

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tháng 6/2020 tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$.
- Tham số μ là trung bình tổng thể (là tiền điện trung bình tháng 6/2020 của tất cả các hộ trong khu vực).

Đặt vấn đề

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tháng 6/2020 tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$.
- Tham số μ là trung bình tổng thể (là tiền điện trung bình tháng 6/2020 của tất cả các hộ trong khu vực).
- Trung bình mẫu $\bar{x} = 236.78$ (tiền điện trung bình của 200 hộ được quan sát là 236.78 nghìn đồng cao hơn trung bình trong tháng 5).

Đặt vấn đề

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tháng 6/2020 tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$.
- Tham số μ là trung bình tổng thể (là tiền điện trung bình tháng 6/2020 của tất cả các hộ trong khu vực).
- Trung bình mẫu $\bar{x} = 236.78$ (tiền điện trung bình của 200 hộ được quan sát là 236.78 nghìn đồng cao hơn trung bình trong tháng 5).
- Câu hỏi: Ta có thể kết luận rằng tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 cao hơn tháng 5 hay không?

Đặt vấn đề

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tháng 6/2020 tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$.
- Tham số μ là trung bình tổng thể (là tiền điện trung bình tháng 6/2020 của tất cả các hộ trong khu vực).
- Trung bình mẫu $\bar{x} = 236.78$ (tiền điện trung bình của 200 hộ được quan sát là 236.78 nghìn đồng cao hơn trung bình trong tháng 5).
- Câu hỏi: Ta có thể kết luận rằng tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 cao hơn tháng 5 hay không?
- Tức là ta cần đưa ra quyết định có chấp nhận hay không chấp nhận giả thuyết "$\mu > 200$".

Kiểm định giả thuyết: Các khái niệm

- Giả thuyết nghiên cứu (research hypothesis): "$\mu > 200$" là giả thuyết cần đưa ra quyết định có chấp nhận hay bác bỏ.

Kiểm định giả thuyết: Các khái niệm

- Giả thuyết nghiên cứu (research hypothesis): "$\mu > 200$" là giả thuyết cần đưa ra quyết định có chấp nhận hay bác bỏ.
- Ta xây dựng giả thuyết trái với giả thuyết nghiên cứu gọi là giả thuyết không (hay giả thuyết đảo: null hypothesis), kí hiệu là $H_0 : \mu \leq 200$ (tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 không cao hơn tháng 5).

Kiểm định giả thuyết: Các khái niệm

- Giả thuyết nghiên cứu (research hypothesis): "$\mu > 200$" là giả thuyết cần đưa ra quyết định có chấp nhận hay bác bỏ.
- Ta xây dựng giả thuyết trái với giả thuyết nghiên cứu gọi là giả thuyết không (hay giả thuyết đảo: null hypothesis), kí hiệu là $H_0 : \mu \leq 200$ (tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 không cao hơn tháng 5).
- Để đưa ra quyết định chấp nhận giả thuyết nghiên cứu, ta đi kiểm định xem thông tin từ mẫu là "bằng chứng" có đủ "mạnh" để bác bỏ giả thuyết đảo H_0 hay không.

Kiểm định giả thuyết: Các khái niệm

- Giả thuyết nghiên cứu (research hypothesis): "$\mu > 200$" là giả thuyết cần đưa ra quyết định có chấp nhận hay bác bỏ.
- Ta xây dựng giả thuyết trái với giả thuyết nghiên cứu gọi là giả thuyết không (hay giả thuyết đảo: null hypothesis), kí hiệu là $H_0 : \mu \leq 200$ (tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 không cao hơn tháng 5).
- Để đưa ra quyết định chấp nhận giả thuyết nghiên cứu, ta đi kiểm định xem thông tin từ mẫu là "bằng chứng" có đủ "mạnh" để bác bỏ giả thuyết đảo H_0 hay không.
- Đối thuyết (alternative hypothesis) H_1: là giả thuyết nghiên cứu ($\mu > 200$).

Kiểm định giả thuyết: Các khái niệm

- Giả thuyết nghiên cứu (research hypothesis): "$\mu > 200$" là giả thuyết cần đưa ra quyết định có chấp nhận hay bác bỏ.
- Ta xây dựng giả thuyết trái với giả thuyết nghiên cứu gọi là giả thuyết không (hay giả thuyết đảo: null hypothesis), kí hiệu là $H_0 : \mu \leq 200$ (tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 không cao hơn tháng 5).
- Để đưa ra quyết định chấp nhận giả thuyết nghiên cứu, ta đi kiểm định xem thông tin từ mẫu là "bằng chứng" có đủ "mạnh" để bác bỏ giả thuyết đảo H_0 hay không.
- Đối thuyết (alternative hypothesis) H_1: là giả thuyết nghiên cứu ($\mu > 200$).
- Bài toán kiểm định: kiểm định giả thuyết $H_0 : \mu \leq 200$ với đối thuyết $H_1 : \mu > 200$.

Kiểm định giả thuyết: Các khái niệm

- Bài toán kiểm định: kiểm định giả thuyết $H_0 : \mu \leq 200$ với đối thuyết $H_1 : \mu > 200$.
- Các khả năng của bài toán kiểm định:

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o
H_o is true	Correct Decision (No error)	Type I Error
	Probability = 1 - α	Probability = α
H_o is false	Type II Error	Correct Decision (No error)
	Probability = β	Probability = 1 - β

Kiểm định giả thuyết: Các khái niệm

- Bài toán kiểm định: kiểm định giả thuyết $H_0 : \mu \leq 200$ với đối thuyết $H_1 : \mu > 200$.
- Các khả năng của bài toán kiểm định:

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o
H_o is true	Correct Decision (No error)	Type I Error
	Probability = $1 - \alpha$	Probability = α
H_o is false	Type II Error	Correct Decision (No error)
	Probability = β	Probability = $1 - \beta$

- Sai lầm loại I (Type I error): Ta kết luận tiền điện tháng 6 cao hơn tháng 5 nhưng thực tế không phải vậy.

Kiểm định giả thuyết: Các khái niệm

- Bài toán kiểm định: kiểm định giả thuyết $H_0 : \mu \leq 200$ với đối thuyết $H_1 : \mu > 200$.
- Các khả năng của bài toán kiểm định:

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o
H_o is true	Correct Decision (No error)	Type I Error
	Probability = 1 - α	Probability = α
H_o is false	Type II Error	Correct Decision (No error)
	Probability = β	Probability = 1 - β

- Sai lầm loại II (Type II error): Ta kết luận tiền điện tháng 6 không cao hơn tháng 5 nhưng thực tế không phải vậy.

Kiểm định giả thuyết: Các khái niệm

- Bài toán kiểm định: kiểm định giả thuyết $H_0 : \mu \leq 200$ với đối thuyết $H_1 : \mu > 200$.
- Các khả năng của bài toán kiểm định:

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o
H_o is true	Correct Decision (No error)	Type I Error
	Probability = 1 - α	Probability = α
H_o is false	Type II Error	Correct Decision (No error)
	Probability = β	Probability = 1 - β

- Kiểm định mức ý nghĩa (test of significance) ở mức $\alpha(= 5\%)$ là bài toán kiểm định sao cho xác suất sai lầm loại I không lớn hơn α (và α gọi là mức ý nghĩa).

Các bước thực hiện bài toán kiểm định

- Bước 1: Xét giả thiết trong bài toán kiểm định.

Các bước thực hiện bài toán kiểm định

- Bước 1: Xét giả thiết trong bài toán kiểm định.
- Bước 2: Xây dựng cặp giả thuyết - đối thuyết: H_0, H_1

Các bước thực hiện bài toán kiểm định

- Bước 1: Xét giả thiết trong bài toán kiểm định.
- Bước 2: Xây dựng cặp giả thuyết - đối thuyết: H_0, H_1
- Bước 3: Tính toán giá trị thống kê (test statistic) của bài toán kiểm định (gọi là tiêu chuẩn kiểm định) và p-giá trị (p-value).

Các bước thực hiện bài toán kiểm định

- Bước 1: Xét giả thiết trong bài toán kiểm định.
- Bước 2: Xây dựng cặp giả thuyết - đối thuyết: H_0, H_1
- Bước 3: Tính toán giá trị thống kê (test statistic) của bài toán kiểm định (gọi là tiêu chuẩn kiểm định) và p-giá trị (p-value).
- Bước 4: Xây dựng quy tắc kiểm định
- Bước 5: Đưa ra kết luận

Ví dụ: Các bước thực hiện bài toán kiểm định

- Bước 1: Giả thiết trong bài toán kiểm định: dữ liệu có phân bố chuẩn.

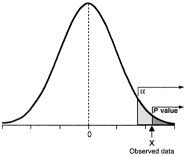
Ví dụ: Các bước thực hiện bài toán kiểm định

- Bước 1: Giả thiết trong bài toán kiểm định: dữ liệu có phân bố chuẩn.
- Bước 2: Xây dựng cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 200$ với $H_1 : \mu > 200$

Ví dụ: Các bước thực hiện bài toán kiểm định

- Bước 1: Giả thiết trong bài toán kiểm định: dữ liệu có phân bố chuẩn.
- Bước 2: Xây dựng cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 200$ với $H_1 : \mu > 200$
- Bước 3: Tính toán giá trị thống kê và p-giá trị:

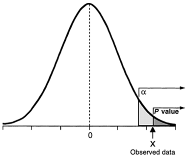
$$t_{obs} = \frac{\bar{x} - 200}{s/\sqrt{n}} = 5.35 \text{ và } \text{p-giá trị} = \mathbb{P}(T_{n-1} > 5.35) = 0.00000012$$



Ví dụ: Các bước thực hiện bài toán kiểm định

- Bước 3: Tính toán giá trị thống kê và p-giá trị:

$$\text{p-giá trị} = \mathbb{P}(T_{n-1} > 3.53) = 0.00000012$$

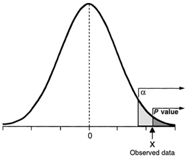


- Bước 4: Quy tắc kiểm định: Ta bác bỏ H_0 vì
$\Leftrightarrow$ p-giá trị $= 0.00000012 = \mathbb{P}(T_{n-1} > t_{obs}) < \alpha = 0.05$

Ví dụ: Các bước thực hiện bài toán kiểm định

- Bước 3: Tính toán giá trị thống kê và p-giá trị:

$$\text{p-giá trị} = \mathbb{P}(T_{n-1} > 3.53) = 0.00000012$$



- Bước 4: Quy tắc kiểm định: Ta bác bỏ H_0 vì
  $\Leftrightarrow$ p-giá trị $= 0.00000012 = \mathbb{P}(T_{n-1} > t_{obs}) < \alpha = 0.05$
- Bước 5: Kết luận: Có rất nhiều bằng chứng để nói rằng tiền điện trung bình của tất cả các hộ trong khu vực trên trong tháng 6 cao hơn tháng 5 ở mức ý nghĩa $1.2 * 10^{-7}$.

Ý nghĩa của p-giá trị

- Nếu p-giá trị nhỏ hơn 1%, thì có rất nhiều bằng chứng để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê rất cao.

Ý nghĩa của p-giá trị

- Nếu p-giá trị nhỏ hơn 1%, thì có rất nhiều bằng chứng để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê rất cao.
- Nếu p-giá trị nằm giữa 1% và 5%, thì có nhiều bằng chứng mạnh để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê cao.

Ý nghĩa của p-giá trị

- Nếu p-giá trị nhỏ hơn 1%, thì có rất nhiều bằng chứng để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê rất cao.
- Nếu p-giá trị nằm giữa 1% và 5%, thì có nhiều bằng chứng mạnh để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê cao.
- Nếu p-giá trị nằm giữa 5% và 10%, thì có bằng chứng yếu để nói rằng đối thuyết H_1 là đúng. Khi p-giá trị lớn hơn 5%, chúng ta nói rằng kết quả không có ý nghĩa thống kê.

Ý nghĩa của p-giá trị

- Nếu p-giá trị nhỏ hơn 1%, thì có rất nhiều bằng chứng để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê rất cao.
- Nếu p-giá trị nằm giữa 1% và 5%, thì có nhiều bằng chứng mạnh để nói rằng đối thuyết H_1 là đúng hay có ý nghĩa thống kê cao.
- Nếu p-giá trị nằm giữa 5% và 10%, thì có bằng chứng yếu để nói rằng đối thuyết H_1 là đúng. Khi p-giá trị lớn hơn 5%, chúng ta nói rằng kết quả không có ý nghĩa thống kê.
- Khi p-giá trị vượt quá 10%, thì có rất ít hoặc không có bằng chứng để nói rằng đối thuyết H_1 là đúng.

Kiểm định tham số và phi tham số

- Mô hình tham số: xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể X có phân bố $f(x, \theta)$ với tham số θ thuộc một không gian hữu hạn chiều, ta biết thông tin về dạng phân bố f của tổng thể X. Kiểm định về tham số θ hay về phân bố f là kiểm định tham số (parametric test).

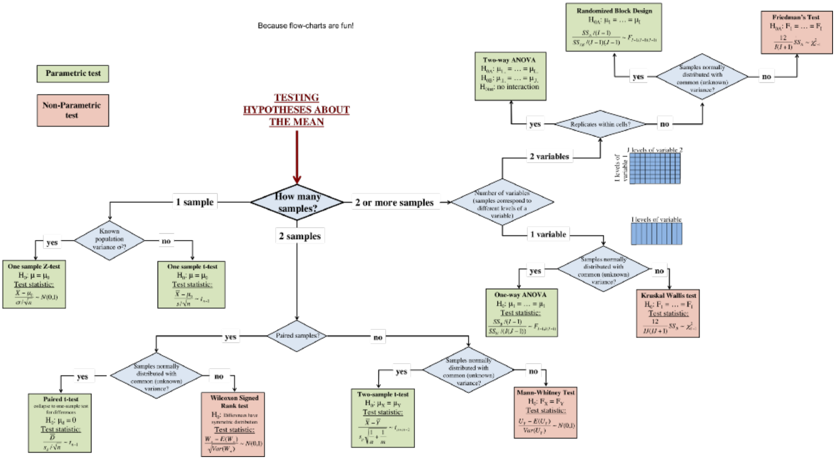
Kiểm định tham số và phi tham số

- Mô hình tham số: xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể X có phân bố $f(x, \theta)$ với tham số θ thuộc một không gian hữu hạn chiều, ta biết thông tin về dạng phân bố f của tổng thể X. Kiểm định về tham số θ hay về phân bố f là kiểm định tham số (parametric test).
- Mô hình phi tham số: xét mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể X có phân bố $f(x)$ thuộc một lớp không gian các hàm có số chiều vô hạn (ta không biết thông tin về dạng phân bố f của tổng thể X). Kiểm định về phân bố f là kiểm định phi tham số (nonparametric test).

Kiểm định giá trị trung bình

- Kiểm định giá trị trung bình của một tổng thể (ta quan sát được một mẫu): t-test (kiểm định tham số: phân bố chuẩn)
- Kiểm định (so sánh) giá trị trung bình của hai một tổng thể (ta quan sát được hai mẫu): t-test (kiểm định tham số: phân bố chuẩn), mann-whitney-wilcoxon test (kiểm định phi tham số)
- Kiểm định (so sánh) giá trị trung bình của nhiều hơn hai một tổng thể (ta quan sát được nhiều hơn mẫu): mô hình ANOVA (kiểm định tham số), kruskal wallis test (kiểm định phi tham số)

Kiểm định giá trị trung bình



Kiểm định t-test cho một mẫu (1)

- Giả thiết trong bài toán kiểm định T một mẫu: mẫu $(X_1, X_2, \dots, X_n)$ lấy từ tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$.

Kiểm định t-test cho một mẫu (1)

- Giả thiết trong bài toán kiểm định T một mẫu: mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$.
- Giả thuyết đảo: Quy tắc kiểm định cho giả thuyết đơn $H_0 : \mu = \mu_0$ hay giả thuyết hợp ($H_0 : \mu \leq \mu_0$; hoặc $H_0 : \mu \geq \mu_0$) là giống nhau nên chúng ta chỉ xét giả thuyết đơn $H_0 : \mu = \mu_0$.

Kiểm định t-test cho một mẫu (1)

- Giả thiết trong bài toán kiểm định T một mẫu: mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$.
- Giả thuyết đảo: Quy tắc kiểm định cho giả thuyết đơn $H_0 : \mu = \mu_0$ hay giả thuyết hợp ($H_0 : \mu \leq \mu_0$; hoặc $H_0 : \mu \geq \mu_0$) là giống nhau nên chúng ta chỉ xét giả thuyết đơn $H_0 : \mu = \mu_0$.
- Ba đối thuyết được xét là:
  - Đối thuyết một phía về bên phải $H_1 : \mu > \mu_0$.

Kiểm định t-test cho một mẫu (1)

- Giả thiết trong bài toán kiểm định T một mẫu: mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$.
- Giả thuyết đảo: Quy tắc kiểm định cho giả thuyết đơn $H_0 : \mu = \mu_0$ hay giả thuyết hợp ($H_0 : \mu \leq \mu_0$; hoặc $H_0 : \mu \geq \mu_0$) là giống nhau nên chúng ta chỉ xét giả thuyết đơn $H_0 : \mu = \mu_0$.
- Ba đối thuyết được xét là:
    - Đối thuyết một phía về bên phải $H_1 : \mu > \mu_0$.
    - Đối thuyết một phía về bên trái $H_1 : \mu < \mu_0$.

Kiểm định t-test cho một mẫu (1)

- Giả thiết trong bài toán kiểm định T một mẫu: mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$.
- Giả thuyết đảo: Quy tắc kiểm định cho giả thuyết đơn $H_0 : \mu = \mu_0$ hay giả thuyết hợp ($H_0 : \mu \leq \mu_0$; hoặc $H_0 : \mu \geq \mu_0$) là giống nhau nên chúng ta chỉ xét giả thuyết đơn $H_0 : \mu = \mu_0$.
- Ba đối thuyết được xét là:
  - Đối thuyết một phía về bên phải $H_1 : \mu > \mu_0$.
  - Đối thuyết một phía về bên trái $H_1 : \mu < \mu_0$.
  - Đối thuyết hai phía $H_1 : \mu \neq \mu_0$.

Kiểm định t-test cho một mẫu (2)

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$

Kiểm định t-test cho một mẫu (2)

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu > \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} > t_{obs})$

Kiểm định t-test cho một mẫu (2)

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu > \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} > t_{obs})$
  - Đối thuyết $H_1 : \mu < \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} < t_{obs})$

Kiểm định t-test cho một mẫu (2)

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu > \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} > t_{obs})$
  - Đối thuyết $H_1 : \mu < \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} < t_{obs})$
  - Đối thuyết $H_1 : \mu \neq \mu_0$: p-giá trị $= 2\mathbb{P}(T_{n-1} > |t_{obs}|)$

Kiểm định t-test cho một mẫu (2)

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu > \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} > t_{obs})$
  - Đối thuyết $H_1 : \mu < \mu_0$: p-giá trị $= \mathbb{P}(T_{n-1} < t_{obs})$
  - Đối thuyết $H_1 : \mu \neq \mu_0$: p-giá trị $= 2\mathbb{P}(T_{n-1} > |t_{obs}|)$
- Quy tắc KĐ: Ta bác bỏ H_0 ở mức ý nghĩa α nếu p-giá trị $< \alpha$.

Kiểm định t-test cho hai mẫu

Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai

- Giả thiết trong bài toán kiểm định T hai mẫu: 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ lấy từ 2 tổng thể có phân bố chuẩn với cùng phương sai là $\mathcal{N}(\mu_1, \sigma^2)$ và $\mathcal{N}(\mu_2, \sigma^2)$.

Kiểm định t-test cho hai mẫu

Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai

- Giả thiết trong bài toán kiểm định T hai mẫu: 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ lấy từ 2 tổng thể có phân bố chuẩn với cùng phương sai là $\mathcal{N}(\mu_1, \sigma^2)$ và $\mathcal{N}(\mu_2, \sigma^2)$.
- Giả thuyết đảo: $H_0 : \mu_1 = \mu_2$

Kiểm định t-test cho hai mẫu

Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai

- Giả thiết trong bài toán kiểm định T hai mẫu: 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ lấy từ 2 tổng thể có phân bố chuẩn với cùng phương sai là $\mathcal{N}(\mu_1, \sigma^2)$ và $\mathcal{N}(\mu_2, \sigma^2)$.
- Giả thuyết đảo: $H_0 : \mu_1 = \mu_2$
- Ba đối thuyết được xét là:
    - Đối thuyết một phía về bên phải $H_1 : \mu_1 > \mu_2$.

Kiểm định t-test cho hai mẫu

Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai

- Giả thiết trong bài toán kiểm định T hai mẫu: 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ lấy từ 2 tổng thể có phân bố chuẩn với cùng phương sai là $\mathcal{N}(\mu_1, \sigma^2)$ và $\mathcal{N}(\mu_2, \sigma^2)$.
- Giả thuyết đảo: $H_0 : \mu_1 = \mu_2$
- Ba đối thuyết được xét là:
  - Đối thuyết một phía về bên phải $H_1 : \mu_1 > \mu_2$.
  - Đối thuyết một phía về bên trái $H_1 : \mu_1 < \mu_2$.

Kiểm định t-test cho hai mẫu

Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai

- Giả thiết trong bài toán kiểm định T hai mẫu: 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ lấy từ 2 tổng thể có phân bố chuẩn với cùng phương sai là $\mathcal{N}(\mu_1, \sigma^2)$ và $\mathcal{N}(\mu_2, \sigma^2)$.
- Giả thuyết đảo: $H_0 : \mu_1 = \mu_2$
- Ba đối thuyết được xét là:
  - Đối thuyết một phía về bên phải $H_1 : \mu_1 > \mu_2$.
  - Đối thuyết một phía về bên trái $H_1 : \mu_1 < \mu_2$.
  - Đối thuyết hai phía $H_1 : \mu_1 \neq \mu_2$.

Kiểm định T cho hai mẫu

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, với $S^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$

Kiểm định T cho hai mẫu

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x} - \bar{y}}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$, với

  $S^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} > t_{obs})$

Kiểm định T cho hai mẫu

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, với $S^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} < t_{obs})$

Kiểm định T cho hai mẫu

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, với $S^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} < t_{obs})$
  - Đối thuyết $H_1 : \mu_1 \neq \mu_2$: p-giá trị $= 2\mathbb{P}(T_{n+m-2} > |t_{obs}|)$

Kiểm định T cho hai mẫu

- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$, với
  $S^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$
- Tính p-giá trị trong trường hợp:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_{n+m-2} < t_{obs})$
  - Đối thuyết $H_1 : \mu_1 \neq \mu_2$: p-giá trị $= 2\mathbb{P}(T_{n+m-2} > |t_{obs}|)$
- Quy tắc KĐ: Ta bác bỏ H_0 ở mức ý nghĩa α nếu p-giá trị $< \alpha$.

Ví dụ minh hoạ

Một công ty tiến hành thu gom báo cũ từ các hộ gia đình và tái chế chúng. Một nhà phân tích tài chính của công ty đó đã tính toán rằng công ty sẽ kiếm được lợi nhuận nếu khối lượng báo cũ trung bình thu gom từ của các hộ gia đình trong tuần vượt quá 2 kg. Một mẫu ngẫu nhiên gồm 148 hộ gia đình đã được lấy từ một cộng đồng, và khối lượng báo cũ trong tuần của các hộ đó được ghi lại và liệt kê trong bảng sau:

2.5	0.7	3.4	1.8	1.9	2.0	1.3	1.2	2.2	0.9	2.7	2.9	1.5	1.5	2.2
3.2	0.7	2.3	3.1	1.3	4.2	3.4	1.5	2.1	1.0	2.4	1.8	0.9	1.3	2.6
3.6	0.8	3.0	2.8	3.6	3.1	2.4	3.2	4.4	4.1	1.5	1.9	3.2	1.9	1.6
3.0	3.7	1.7	3.1	2.4	3.0	1.5	3.1	2.4	2.1	2.1	2.3	0.7	0.9	2.7
1.2	2.2	1.3	3.0	3.0	2.2	1.5	2.7	0.9	2.5	3.2	3.7	1.9	2.0	3.7
2.3	0.6	0.0	1.0	1.4	0.9	2.6	2.1	3.4	0.5	4.1	2.2	3.4	3.3	0.0
2.2	4.2	1.1	2.3	3.1	1.7	2.8	2.5	1.8	1.7	0.6	3.6	1.4	2.2	2.2
1.3	1.7	3.0	0.8	1.6	1.8	1.4	3.0	1.9	2.7	0.8	3.3	2.5	1.5	2.2
2.6	3.2	1.0	3.2	1.6	3.4	1.7	2.3	2.6	1.4	3.3	1.3	2.4	2.0	
1.3	1.8	3.3	2.2	1.4	3.2	4.3	0.0	2.0	1.8	0.0	1.7	2.6	3.1	

Ví dụ minh hoạ

2.5	0.7	3.4	1.8	1.9	2.0	1.3	1.2	2.2	0.9	2.7	2.9	1.5	1.5	2.2
3.2	0.7	2.3	3.1	1.3	4.2	3.4	1.5	2.1	1.0	2.4	1.8	0.9	1.3	2.6
3.6	0.8	3.0	2.8	3.6	3.1	2.4	3.2	4.4	4.1	1.5	1.9	3.2	1.9	1.6
3.0	3.7	1.7	3.1	2.4	3.0	1.5	3.1	2.4	2.1	2.1	2.3	0.7	0.9	2.7
1.2	2.2	1.3	3.0	3.0	2.2	1.5	2.7	0.9	2.5	3.2	3.7	1.9	2.0	3.7
2.3	0.6	0.0	1.0	1.4	0.9	2.6	2.1	3.4	0.5	4.1	2.2	3.4	3.3	0.0
2.2	4.2	1.1	2.3	3.1	1.7	2.8	2.5	1.8	1.7	0.6	3.6	1.4	2.2	2.2
1.3	1.7	3.0	0.8	1.6	1.8	1.4	3.0	1.9	2.7	0.8	3.3	2.5	1.5	2.2
2.6	3.2	1.0	3.2	1.6	3.4	1.7	2.3	2.6	1.4	3.3	1.3	2.4	2.0	
1.3	1.8	3.3	2.2	1.4	3.2	4.3	0.0	2.0	1.8	0.0	1.7	2.6	3.1	

Câu hỏi: Những dữ liệu này có cung cấp đủ cơ sở để cho phép nhà phân tích kết luận rằng một nhà máy tái chế sẽ có lãi không ở mức ý nghĩa 5%?

Ví dụ minh hoạ

- Gọi μ là khối lượng báo cũ trung bình thu gom từ của các hộ gia đình trong tuần tại ku vực trên. Ta kiểm định cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 2$ với $H_1 : \mu > 2$

Ví dụ minh hoạ

- Gọi μ là khối lượng báo cũ trung bình thu gom từ của các hộ gia đình trong tuần tại ku vực trên. Ta kiểm định cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 2$ với $H_1 : \mu > 2$
- Ta có $n = 148; \bar{x} = 2.18; s = 0.962$, nên giá trị thống kê và p-giá trị là

$$Z_{obs} = \frac{\bar{x} - 2}{s/\sqrt{n}} = 2.24 \text{ và } \text{ p-giá trị} = \mathbb{P}(Z > 2.24) = 0.013$$

Ví dụ minh hoạ

- Gọi μ là khối lượng báo cũ trung bình thu gom từ của các hộ gia đình trong tuần tại ku vực trên. Ta kiểm định cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 2$ với $H_1 : \mu > 2$
- Ta có $n = 148; \bar{x} = 2.18; s = 0.962$, nên giá trị thống kê và p-giá trị là

$$Z_{obs} = \frac{\bar{x} - 2}{s/\sqrt{n}} = 2.24 \text{ và } \text{p-giá trị} = \mathbb{P}(Z > 2.24) = 0.013$$

- Ta bác bỏ H_0 ở mức ý nghĩa 0.05 vì
p-giá trị $= 0.013 = \mathbb{P}(T_{n-1} > t_{obs}) < \alpha = 0.05$

Ví dụ minh hoạ

- Gọi μ là khối lượng báo cũ trung bình thu gom từ của các hộ gia đình trong tuần tại ku vực trên. Ta kiểm định cặp giả thuyết - đối thuyết: $H_0 : \mu \leq 2$ với $H_1 : \mu > 2$
- Ta có $n = 148; \bar{x} = 2.18; s = 0.962$, nên giá trị thống kê và p-giá trị là

$$Z_{obs} = \frac{\bar{x} - 2}{s/\sqrt{n}} = 2.24 \text{ và } \text{p-giá trị} = \mathbb{P}(Z > 2.24) = 0.013$$

- Ta bác bỏ H_0 ở mức ý nghĩa 0.05 vì
  p-giá trị $= 0.013 = \mathbb{P}(T_{n-1} > t_{obs}) < \alpha = 0.05$
- Kết luận: Những dữ liệu này cung cấp đủ cơ sở để cho phép nhà phân tích kết luận rằng một nhà máy tái chế sẽ có lãi ở mức ý nghĩa 5%.
- Hãy xây dựng công thức tính xác suất sai lầm loại II cho bài toán kiểm định trên và tính với $\mu = 3$.

Ví dụ minh hoạ

Code Python:

```
import numpy as np
import pandas as pd
from scipy.stats import t
df = pd.read_csv("/content/drive/My Drive/Dataset/data4_lecture8.csv")
x = df.values[:,0]
n = len(x)
print(n)
xbar =np.mean(x)
print(xbar)
s = np.std(x)
print(s)
mu0 = 2
tobs =(xbar - mu0)/(s/np.sqrt(n))
print(tobs)
pvalue = 1- t.cdf(tobs,n-1)
print(pvalue)
```

Bài tập 1

Người ta muốn nghiên cứu xem cách thiết kế công việc (đề cập đến các chuyển động của công nhân) có ảnh hưởng đến năng suất của công nhân hay không. Hai thiết kế công việc đang được xem xét để sản xuất bàn máy tính mới. Hai mẫu được chọn ngẫu nhiên và độc lập: mẫu gồm 25 công nhân lắp ráp bàn theo thiết kế A và mẫu gồm 25 công nhân lắp ráp bàn theo thiết kế B. Thời gian lắp ráp được ghi lại. Thời gian lắp ráp trung bình theo hai thiết kế có khác nhau ở mức ý nghĩa $\alpha = 5\%$ không?

Design A							Design B						
6.8	5.0	7.9	5.2	7.6	5.0	5.9	5.2	6.7	5.7	6.6	8.5	6.5	5.9
5.2	6.5	7.4	6.1	6.2	7.1	4.6	6.7	6.6	4.2	4.2	4.5	5.3	7.9
6.0	7.1	6.1	5.0	6.3	7.0	6.4	7.0	5.9	7.1	5.8	7.0	5.7	5.9
6.1	6.6	7.7	6.4				4.9	5.3	4.2	7.1			

Bài tập 2

Quan sát thời gian sống sót X (tính bằng ngày) của 72 con chuột lang bị nhiễm trực khuẩn lao, ta thu được bảng số liệu như sau:

12	15	22	24	24	32	32	33	34	38	38	43	44
48	52	53	54	54	55	56	57	58	58	59	60	60
60	60	61	62	63	65	65	67	68	70	70	72	73
75	76	76	81	83	84	85	87	91	95	96	98	99
109	110	121	127	129	131	143	146	146	175	175	211	233
258	258	263	297	341	341	376						

Giả sử X có phân phối mũ với tham số $\theta > 0$ và hàm mật độ xác suất như sau: $f(x;\theta) = \theta e^{-\theta x}, x > 0$. Ở mức ý nghĩa 5%, có thể cho rằng tham số θ lớn hơn 0.5% hay không? Hãy tính xác suất sai lầm loại 2 khi $\theta = 1\%$