KIỂM ĐỊNH GIẢ THUYẾT (2)
(HYPOTHESIS TESTING)

Nguyễn Văn Hạnh

AI Academy Vietnam

November 3, 2020

Nội dung

1. Kiểm định hai giá trị trung bình (tiếp)
2. Kiểm định sự bằng nhau của 2 phương sai
3. Kiểm định Mann-Whitney-Wilcoxon
4. Kiểm định hệ số tương quan
5. Bài tập

Kiểm định t-test cho hai mẫu

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.

Kiểm định t-test cho hai mẫu

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 trung bình mẫu $\bar{X}$ và $\bar{Y}$, ta kiểm định về sự khác biệt giữa trung bình hai tổng thể μ_1 và μ_2.

Kiểm định t-test cho hai mẫu

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 trung bình mẫu $\bar{X}$ và $\bar{Y}$, ta kiểm định về sự khác biệt giữa trung bình hai tổng thể μ_1 và μ_2.
- Giả thuyết đảo: $H_0 : \mu_1 = \mu_2$.
  - 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-test cho hai mẫu

- Trường hợp 1: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn cùng phương sai: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

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: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Giả thiết trong trường hợp này là:
  - Các quan sát trong 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập
  - 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ là độc lập
  - 2 tổng thể có phân bố chuẩn cùng phương sai.

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: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.
- Giả thiết trong trường hợp này là:
  - Các quan sát trong 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập
  - 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ là độc lập
  - 2 tổng thể có phân bố chuẩn cùng phương sai.
- Thống kê của bài toán kiểm định trong trường hợp này là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}}$ có phân bố Student $n+m-2$ bậc tự do, với $S^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$

Kiểm định t-test cho hai mẫu

- Trường hợp 2: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn không cùng phương sai: $\sigma_1^2 \neq \sigma_2^2$.

Kiểm định t-test cho hai mẫu

- Trường hợp 2: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn không cùng phương sai: $\sigma_1^2 \neq \sigma_2^2$.
- Giả thiết trong trường hợp này là:
  - Các quan sát trong 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập
  - 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ là độc lập
  - 2 tổng thể có phân bố chuẩn không cùng phương sai.

Kiểm định t-test cho hai mẫu

- Trường hợp 2: hai mẫu độc lập lấy từ 2 tổng thể có phân bố chuẩn không cùng phương sai: $\sigma_1^2 \neq \sigma_2^2$.
- Giả thiết trong trường hợp này là:
  - Các quan sát trong 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập
  - 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ là độc lập
  - 2 tổng thể có phân bố chuẩn không cùng phương sai.
- Thống kê của bài toán kiểm định trong trường hợp này là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}}$ có phân bố Student ν bậc tự do, với

$$\nu = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}}$$

Kiểm định T cho hai mẫu

- Tính p-giá trị trong trường hợp 2:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_\nu > t_{obs})$

Kiểm định T cho hai mẫu

- Tính p-giá trị trong trường hợp 2:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_\nu > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_\nu < t_{obs})$

Kiểm định T cho hai mẫu

- Tính p-giá trị trong trường hợp 2:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_\nu > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_\nu < t_{obs})$
  - Đối thuyết $H_1 : \mu_1 \neq \mu_2$: p-giá trị $= 2\mathbb{P}(T_\nu > |t_{obs}|)$

Kiểm định T cho hai mẫu

- Tính p-giá trị trong trường hợp 2:
  - Đối thuyết $H_1 : \mu_1 > \mu_2$: p-giá trị $= \mathbb{P}(T_\nu > t_{obs})$
  - Đối thuyết $H_1 : \mu_1 < \mu_2$: p-giá trị $= \mathbb{P}(T_\nu < t_{obs})$
  - Đối thuyết $H_1 : \mu_1 \neq \mu_2$: p-giá trị $= 2\mathbb{P}(T_\nu > |t_{obs}|)$
- Quy tắc KĐ: Ta bác bỏ H_0 ở mức ý nghĩa α nếu p-giá trị $< \alpha$.

Kiểm định sự bằng nhau của 2 phương sai

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.

Kiểm định sự bằng nhau của 2 phương sai

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 phương sai mẫu S_1^2 và S_2^2, ta kiểm định về sự khác biệt giữa phương sai hai tổng thể σ_1^2 và σ_2^2.

Kiểm định sự bằng nhau của 2 phương sai

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 phương sai mẫu S_1^2 và S_2^2, ta kiểm định về sự khác biệt giữa phương sai hai tổng thể σ_1^2 và σ_2^2.
- Giả thuyết đảo: $H_0 : \sigma_1^2 = \sigma_2^2$ với đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$

Kiểm định sự bằng nhau của 2 phương sai

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 phương sai mẫu S_1^2 và S_2^2, ta kiểm định về sự khác biệt giữa phương sai hai tổng thể σ_1^2 và σ_2^2.
- Giả thuyết đảo: $H_0 : \sigma_1^2 = \sigma_2^2$ với đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Thống kê của bài toán kiểm định là $F_{obs} = \frac{S_1^2}{S_2^2}$ có phân bố Fisher với $n-1$ và $m-1$ bậc tự do.

Kiểm định sự bằng nhau của 2 phương sai

- Bài toán: Giả sử có 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 là $\mathcal{N}(\mu_1, \sigma_1^2)$ và $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Từ sự khác biệt giữa 2 phương sai mẫu S_1^2 và S_2^2, ta kiểm định về sự khác biệt giữa phương sai hai tổng thể σ_1^2 và σ_2^2.
- Giả thuyết đảo: $H_0 : \sigma_1^2 = \sigma_2^2$ với đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Thống kê của bài toán kiểm định là $F_{obs} = \frac{S_1^2}{S_2^2}$ có phân bố Fisher với $n-1$ và $m-1$ bậc tự do.
- Nếu $F_{obs} > 1$ thì
  p-giá trị $= \mathbb{P}(F_{n-1,m-1} > F_{obs}) + \mathbb{P}(F_{n-1,m-1} < 1/F_{obs})$
- Nếu $F_{obs} < 1$ thì
  p-giá trị $= \mathbb{P}(F_{n-1,m-1} < F_{obs}) + \mathbb{P}(F_{n-1,m-1} > 1/F_{obs})$

Ví dụ minh hoạ 1

Điều gì xảy ra với doanh nghiệp do gia đình điều hành khi con trai hoặc con gái của ông chủ tiếp quản? Doanh nghiệp có hoạt động tốt hơn sau khi thay đổi nếu ông chủ mới là con đẻ của chủ sở hữu hay doanh nghiệp hoạt động tốt hơn khi có người ngoài làm giám đốc điều hành (CEO)? Để có câu trả lời cho câu hỏi này, các nhà nghiên cứu đã chọn ngẫu nhiên 140 công ty từ năm 1994 đến 2002, 30% trong số đó đã chuyển quyền sở hữu cho con cái và 70% trong số đó bổ nhiệm một người ngoài làm Giám đốc điều hành. Đối với mỗi công ty, các nhà nghiên cứu đã tính toán thu nhập hoạt động theo tỷ lệ tài sản trong năm trước và năm sau khi CEO mới tiếp quản. Sự thay đổi (thu nhập hoạt động sau - thu nhập hoạt động trước đó) đã được ghi lại và cho trong bảng số liệu sau. Những dữ liệu này có cho phép chúng ta suy luận rằng hiệu quả của việc để con đẻ trở thành một CEO khác với hiệu quả của việc thuê người ngoài làm CEO không?

Ví dụ minh hoạ 1

Offspring			Outsider						
−1.95	0.91	−3.15	0.69	−1.05	1.58	−2.46	3.33	−1.32	−0.51
0	−2.16	3.27	−0.95	−4.23	−1.98	1.59	3.2	5.93	8.68
0.56	1.22	−0.67	−2.2	−0.16	4.41	−2.03	0.55	−0.45	1.43
1.44	0.67	2.61	2.65	2.77	4.62	−1.69	−1.4	−3.2	−0.37
1.5	−0.39	1.55	5.39	−0.96	4.5	0.55	2.79	5.08	−0.49
1.41	−1.43	−2.67	4.15	1.01	2.37	0.95	5.62	0.23	−0.08
−0.32	−0.48	−1.91	4.28	0.09	2.44	3.06	−2.69	−2.69	−1.16
−1.7	0.24	1.01	2.97	6.79	1.07	4.83	−2.59	3.76	1.04
−1.66	0.79	−1.62	4.11	1.72	−1.11	5.67	2.45	1.05	1.28
−1.87	−1.19	−5.25	2.66	6.64	0.44	−0.8	3.39	0.53	1.74
−1.38	1.89	0.14	6.31	4.75	1.36	1.37	5.89	3.2	−0.14
0.57	−3.7	2.12	−3.04	2.84	0.88	0.72	−0.71	−3.07	−0.82
3.05	−0.31	2.75	−0.42	−2.1	0.33	4.14	4.22	−4.34	0
2.98	−1.37	0.3	−0.89	2.07	−5.96	3.04	0.46	−1.16	2.68

Ví dụ minh hoạ 1

- Gọi μ_1, μ_2 và σ_1^2, σ_2^2 lần lượt là kỳ vọng và phương sai của sự thay đổi thu nhập sau khi CEO mới tiếp quản là con đẻ hay là thuê người ngoài.
- Ta có $n = 42, m = 98, \bar{x} = -0.10, \bar{y} = 1.24, s_1^2 = 3.79, s_2^2 = 8.03$
- Ta kiểm định trước hết sự bằng nhau của 2 phương sai: giả thuyết đảo: $H_0 : \sigma_1^2 = \sigma_2^2$ với đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Thống kê của bài toán kiểm định là $F_{obs} = \frac{S_1^2}{S_2^2} = 0.472$

Ví dụ minh hoạ 1

- Gọi μ_1, μ_2 và σ_1^2, σ_2^2 lần lượt là kỳ vọng và phương sai của sự thay đổi thu nhập sau khi CEO mới tiếp quản là con đẻ hay là thuê người ngoài.
- Ta có $n = 42, m = 98, \bar{x} = -0.10, \bar{y} = 1.24, s_1^2 = 3.79, s_2^2 = 8.03$
- Ta kiểm định trước hết sự bằng nhau của 2 phương sai: giả thuyết đảo: $H_0 : \sigma_1^2 = \sigma_2^2$ với đối thuyết $H_1 : \sigma_1^2 \neq \sigma_2^2$
- Thống kê của bài toán kiểm định là $F_{obs} = \frac{S_1^2}{S_2^2} = 0.472$
- p-giá trị của bài toán kiểm định là:
  $p = \mathbb{P}(F_{m-1,n-1} < F_{obs}) + \mathbb{P}(F_{m-1,n-1} > 1/F_{obs}) = 0.0055$ nên ta bác bỏ giả thuyết phương sai bằng nhau.

Ví dụ minh hoạ 1

- Tiếp theo ta kiểm định xem có sự khác nhau của 2 kỳ vọng: giả thuyết đảo: $H_0 : \mu_1 = \mu_2$ với đối thuyết $H_1 : \mu_1 \neq \mu_2$
- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}} = -3.22$

Ví dụ minh hoạ 1

- Tiếp theo ta kiểm định xem có sự khác nhau của 2 kỳ vọng: giả thuyết đảo: $H_0 : \mu_1 = \mu_2$ với đối thuyết $H_1 : \mu_1 \neq \mu_2$
- Thống kê của bài toán kiểm định là $t_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_1^2}{n}+\frac{s_2^2}{m}}} = -3.22$
- Bậc tự do của phân bố của thống kê:

$$\nu = \frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}} = 110.69 \approx 111$$

Ví dụ minh hoạ 1

- p-giá trị $= 2\mathbb{P}(T_\nu > |t_{obs}|) = 0.0017$ nên ta bác bỏ H_0.
- Kết luận: Những dữ liệu này cho phép chúng ta suy luận rằng hiệu quả của việc để con đẻ trở thành một CEO khác với hiệu quả của việc thuê người ngoài làm CEO.

Ví dụ minh hoạ 1

Code Python:

```
# ví dụ minh hoạ 1
import numpy as np
import pandas as pd
from scipy.stats import t
from scipy.stats import f
df = pd.read_csv("/content/drive/My Drive/Dataset/dataVD1_lecture9.csv")
x = df.values[0:42,0]
y = df.values[:,1]
n = len(x)
m = len(y)
xbar =np.mean(x)
print(xbar)
ybar =np.mean(y)
print(ybar)
s12 = np.var(x)
print(s12)
s22 = np.var(y)
print(s22)
Fobs = s12/s22
print(Fobs)
p1 = f.cdf(Fobs,n-1,m-1)+1-f.cdf(1/Fobs,n-1,m-1)
print(p1)
tobs =(xbar - ybar)/np.sqrt(s12/n+s22/m)
print(tobs)
nu = round((s12/n+s22/m)**2/((s12/n)**2/(n-1)+(s22/m)**2/(m-1)))
print(nu)
p2 =2*(1-t.cdf(np.abs(tobs),nu))
print(p2)
```

Kiểm định sự bằng nhau của 2 phân bố

- Trong bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không, nếu 2 mẫu độc lập được lấy từ 2 phân bố chuẩn thì ta có thể sử dụng kiểm định T (t-test).

Kiểm định sự bằng nhau của 2 phân bố

- Trong bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không, nếu 2 mẫu độc lập được lấy từ 2 phân bố chuẩn thì ta có thể sử dụng kiểm định T (t-test).
- Hoặc nếu kích thước của 2 mẫu đủ lớn ($n, m > 30$) thì ta có thể xấp xỉ kiểm định Z (Z-test) trên cơ sở của định lý giới hạn trung tâm.

Kiểm định sự bằng nhau của 2 phân bố

- Trong bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không, nếu 2 mẫu độc lập được lấy từ 2 phân bố chuẩn thì ta có thể sử dụng kiểm định T (t-test).
- Hoặc nếu kích thước của 2 mẫu đủ lớn ($n, m > 30$) thì ta có thể xấp xỉ kiểm định Z (Z-test) trên cơ sở của định lý giới hạn trung tâm.
- Nếu hai tổng thể không có phân bố chuẩn hoặc kích thước 2 mẫu nhỏ thì ta cần áp dụng kiểm định phi tham số: kiểm định Mann-Whitney-Wilcoxon.

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không.

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không.
- Đây là bài toán kiểm định phi tham số (giả sử không có thông tin về phân bố của 2 tổng thể)

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không.
- Đây là bài toán kiểm định phi tham số (giả sử không có thông tin về phân bố của 2 tổng thể)
- Được sử dụng trong trường hợp dữ liệu không có phân phối chuẩn, hoặc cho các mẫu kích thước nhỏ (có ít quan sát)

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không.
- Đây là bài toán kiểm định phi tham số (giả sử không có thông tin về phân bố của 2 tổng thể)
- Được sử dụng trong trường hợp dữ liệu không có phân phối chuẩn, hoặc cho các mẫu kích thước nhỏ (có ít quan sát)
- Đây là trường hợp tổng quát cho bài toán kiểm định t (t-test là kiểm định tham số)

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là bài toán kiểm định xem 2 tổng thể có phân bố khác biệt hay không.
- Đây là bài toán kiểm định phi tham số (giả sử không có thông tin về phân bố của 2 tổng thể)
- Được sử dụng trong trường hợp dữ liệu không có phân phối chuẩn, hoặc cho các mẫu kích thước nhỏ (có ít quan sát)
- Đây là trường hợp tổng quát cho bài toán kiểm định t (t-test là kiểm định tham số)
- Có xu hướng sử dụng ít thông tin hơn kiểm định tham số nên khả năng tìm ra được sự sai biệt kém, không mạnh như các phép kiểm định có tham số (t-test, phân tích phương sai...).

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là một trong những kiểm định dựa trên xếp hạng. Các quan sát sẽ được xếp hạng từ giá trị nhỏ nhất tới lớn nhất và sau đó thứ hạng sẽ được sử dụng thay cho các giá trị thực trong tính toán.

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là một trong những kiểm định dựa trên xếp hạng. Các quan sát sẽ được xếp hạng từ giá trị nhỏ nhất tới lớn nhất và sau đó thứ hạng sẽ được sử dụng thay cho các giá trị thực trong tính toán.
- Kiểm định Mann-Whitney-Wilcoxon dùng để kiểm định liệu có tồn tại sự khác biệt giữa hai tổng thể, với điều kiện:
  - Tổng thể không có phân phối chuẩn
  - Dữ liệu ít nhất phải có thang đo thứ bậc
  - Hai mẫu được chọn ngẫu nhiên độc lập với nhau

Kiểm định Mann-Whitney-Wilcoxon

- Kiểm định Mann-Whitney-Wilcoxon là một trong những kiểm định dựa trên xếp hạng. Các quan sát sẽ được xếp hạng từ giá trị nhỏ nhất tới lớn nhất và sau đó thứ hạng sẽ được sử dụng thay cho các giá trị thực trong tính toán.
- Kiểm định Mann-Whitney-Wilcoxon dùng để kiểm định liệu có tồn tại sự khác biệt giữa hai tổng thể, với điều kiện:
  - Tổng thể không có phân phối chuẩn
  - Dữ liệu ít nhất phải có thang đo thứ bậc
  - Hai mẫu được chọn ngẫu nhiên độc lập với nhau
- Cặp giả thuyết cần kiểm định:
  H_0: Phân phối của hai tổng thể là giống hệt nhau
  H_1: Phân phối của hai tổng thể là khác nhau

Kiểm định Mann-Whitney-Wilcoxon

- Kết hợp hai mẫu ngẫu nhiên và xếp hạng tất cả các quan sát từ nhỏ nhất tới lớn nhất. Những giá trị bằng nhau sẽ nhận hạng trung bình của các hạng liên tiếp.

Kiểm định Mann-Whitney-Wilcoxon

- Kết hợp hai mẫu ngẫu nhiên và xếp hạng tất cả các quan sát từ nhỏ nhất tới lớn nhất. Những giá trị bằng nhau sẽ nhận hạng trung bình của các hạng liên tiếp.
- Tính tổng các thứ hạng riêng cho từng mẫu, kí hiệu là R_1 và R_2.

Kiểm định Mann-Whitney-Wilcoxon

- Kết hợp hai mẫu ngẫu nhiên và xếp hạng tất cả các quan sát từ nhỏ nhất tới lớn nhất. Những giá trị bằng nhau sẽ nhận hạng trung bình của các hạng liên tiếp.
- Tính tổng các thứ hạng riêng cho từng mẫu, kí hiệu là R_1 và R_2.
- Tính các thống kê

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; \; U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

và $U = min(U_1, U_2)$

Kiểm định Mann-Whitney-Wilcoxon

- Kết hợp hai mẫu ngẫu nhiên và xếp hạng tất cả các quan sát từ nhỏ nhất tới lớn nhất. Những giá trị bằng nhau sẽ nhận hạng trung bình của các hạng liên tiếp.
- Tính tổng các thứ hạng riêng cho từng mẫu, kí hiệu là R_1 và R_2.
- Tính các thống kê

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; \; U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

và $U = min(U_1, U_2)$

- Tính p-giá trị $= F(U) = F_{n_1,n_2}(U)$

Kiểm định Mann-Whitney-Wilcoxon

- Kết hợp hai mẫu ngẫu nhiên và xếp hạng tất cả các quan sát từ nhỏ nhất tới lớn nhất. Những giá trị bằng nhau sẽ nhận hạng trung bình của các hạng liên tiếp.
- Tính tổng các thứ hạng riêng cho từng mẫu, kí hiệu là R_1 và R_2.
- Tính các thống kê

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1; \; U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

  và $U = min(U_1, U_2)$
- Tính p-giá trị $= F(U) = F_{n_1,n_2}(U)$
- Ta bác bỏ giả thuyết H_0 nếu p-giá trị $< \alpha$.

Ví dụ minh hoạ 2

Một hãng taxi sân bay vừa mới đưa vào khai thác 2 mẫu xe mới trên tuyến đường từ Cầu Giấy đi Nội Bài. Để đánh giá xem liệu hai mẫu này có thời gian đi như nhau hay không, họ tiến hành theo dõi một cách ngẫu nhiên mỗi mẫu 6 chuyến. Thời gian (phút) của các chuyến đi như sau:
Mẫu A: 35, 38, 40, 42, 41, 36
Mẫu B: 29, 27, 30, 33, 39, 37

Ví dụ minh hoạ 2

Một hãng taxi sân bay vừa mới đưa vào khai thác 2 mẫu xe mới trên tuyến đường từ Cầu Giấy đi Nội Bài. Để đánh giá xem liệu hai mẫu này có thời gian đi như nhau hay không, họ tiến hành theo dõi một cách ngẫu nhiên mỗi mẫu 6 chuyến. Thời gian (phút) của các chuyến đi như sau:
Mẫu A: 35, 38, 40, 42, 41, 36
Mẫu B: 29, 27, 30, 33, 39, 37

- Kết hợp 2 mẫu và sắp xếp theo thứ tự từ nhỏ đến lớn

Mẫu A					35	36		38		40	41	42
Mẫu B	27	29	30	33			37		39			
Xếp hạng	1	2	3	4	5	6	7	8	9	10	11	12

Ví dụ minh hoạ 2

- Tổng các thứ hạng riêng cho mẫu A là:
  $R_1 = 5 + 6 + 8 + 10 + 11 + 12 = 52$ và mẫu B là
  $R_2 = 1 + 2 + 3 + 4 + 7 + 9 = 26$.

Ví dụ minh hoạ 2

- Tổng các thứ hạng riêng cho mẫu A là:
  $R_1 = 5 + 6 + 8 + 10 + 11 + 12 = 52$ và mẫu B là
  $R_2 = 1 + 2 + 3 + 4 + 7 + 9 = 26$.
- Tính các thống kê

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 6x6 + \frac{6(6+1)}{2} - 52 = 5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 6x6 + \frac{6(6+1)}{2} - 26 = 31$$

và $U = min(U_1, U_2) = 5$

Kiểm định Mann-Whitney-Wilcoxon

- p-giá trị $= F(U) = F_{6,6}(5) = 0.0206$ (xác suất để U đạt gái trị ≤ 5 là $0,0206$)

Kiểm định Mann-Whitney-Wilcoxon

- p-giá trị $= F(U) = F_{6,6}(5) = 0.0206$ (xác suất để U đạt gái trị ≤ 5 là $0,0206$)
- Với mức ý nghĩa $\alpha = 0,05$ thì $F(U) = 0,0206 < \alpha$ nên ta bác bỏ giả thuyết H_0.

Kiểm định Mann-Whitney-Wilcoxon

- p-giá trị $= F(U) = F_{6,6}(5) = 0.0206$ (xác suất để U đạt gái trị ≤ 5 là $0,0206$)
- Với mức ý nghĩa $\alpha = 0,05$ thì $F(U) = 0,0206 < \alpha$ nên ta bác bỏ giả thuyết H_0.
- Vậy hai mẫu xe này có thời gian đi là khác nhau.

Ví dụ minh hoạ 2

Code Python:

```
# ví dụ minh hoạ 2
from scipy.stats import mannwhitneyu
x=[35, 38, 40, 42, 41, 36]
y=[29, 27, 30, 33, 39, 37]
# compare samples
stat, p = mannwhitneyu(x, y)
print('Statistics=%.3f, p=%.3f' % (stat, p))
# interpret
alpha = 0.05
if p > alpha:
  print('Same distribution (fail to reject H0)')
else:
  print('Different distribution (reject H0)')
```

```
Statistics=5.000, p=0.023
Different distribution (reject H0)
```

Hệ số tương quan

- Hệ số tương quan giữa hai tổng thể X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{var(X)var(Y)}}$$

Hệ số tương quan

- Hệ số tương quan giữa hai tổng thể X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{var(X)var(Y)}}$$

- Nếu $\rho(X,Y) = 0$ thì X và Y không có tương quan.

Hệ số tương quan

- Hệ số tương quan giữa hai tổng thể X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{var(X)var(Y)}}$$

- Nếu $\rho(X,Y) = 0$ thì X và Y không có tương quan.
- Xét mẫu có kích thước n: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Hệ số tương quan

- Hệ số tương quan giữa hai tổng thể X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{var(X)var(Y)}}$$

- Nếu $\rho(X,Y) = 0$ thì X và Y không có tương quan.
- Xét mẫu có kích thước n: $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$
- Ước lượng của hệ số tương quan $\rho(X,Y)$ là hệ số tương quan mẫu

$$r = \frac{\overline{xy} - \bar{x}.\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$$

Kiểm định hệ số tương quan

- Kiểm định giả thuyết: $H_0 : \rho(X, Y) = 0$ (X và Y không có tương quan)
  Với 3 đối thuyết sau:
  - $H_1 : \rho(X, Y) \neq 0$ (X và Y có tương quan)
  - $H_1 : \rho(X, Y) > 0$ (X và Y có tương quan dương)
  - $H_1 : \rho(X, Y) < 0$ (X và Y có tương quan âm)

Kiểm định hệ số tương quan

- Kiểm định giả thuyết: $H_0 : \rho(X, Y) = 0$ (X và Y không có tương quan)
  Với 3 đối thuyết sau:
  - $H_1 : \rho(X, Y) \neq 0$ (X và Y có tương quan)
  - $H_1 : \rho(X, Y) > 0$ (X và Y có tương quan dương)
  - $H_1 : \rho(X, Y) < 0$ (X và Y có tương quan âm)
- Tiêu chuẩn kiểm định là thống kê:

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}}$$

Kiểm định hệ số tương quan

- P-giá trị được tính như sau:
  - $H_1 : \rho(X, Y) \neq 0$: p-giá trị $= 2\mathbb{P}(T_{n-2} > |t_{obs}|)$
  - $H_1 : \rho(X, Y) > 0$: p-giá trị $= \mathbb{P}(T_{n-2} > t_{obs})$
  - $H_1 : \rho(X, Y) < 0$: p-giá trị $= \mathbb{P}(T_{n-2} < t_{obs})$

Kiểm định hệ số tương quan

- P-giá trị được tính như sau:
  - $H_1 : \rho(X, Y) \neq 0$: p-giá trị $= 2\mathbb{P}(T_{n-2} > |t_{obs}|)$
  - $H_1 : \rho(X, Y) > 0$: p-giá trị $= \mathbb{P}(T_{n-2} > t_{obs})$
  - $H_1 : \rho(X, Y) < 0$: p-giá trị $= \mathbb{P}(T_{n-2} < t_{obs})$
- Ta bác bỏ giả thuyết H_0 nếu p-giá trị $< \alpha$

Ví dụ minh hoạ 3

Quan sát độ tuổi X và tỉ trọng cở thể Y (trọng lượng kg chia chiều cao bình phương m2) của 20 người cao tuổi ta thu được số liệu như sau:

Id	age	bmi
1	79	24.72
2	89	25.99
3	70	25.39
4	88	23.22
5	85	24.61
6	68	25.08
7	70	19.88
8	69	25.06
9	74	25.65
10	79	19.95
11	76	22.6
12	76	26.42
13	62	20.32
14	69	19.37
15	72	24.22
16	67	32.11
17	74	25.39
18	69	24.67
19	78	27.13
20	71	23.05

Ví dụ minh hoạ 3

Hãy kiểm định xem độ tuổi X và tỉ trọng cơ thể Y có tương quan nhau không ở mức ý nghĩa 5%.

Id	age	bmi
1	79	24.72
2	89	25.99
3	70	25.39
4	88	23.22
5	85	24.61
6	68	25.08
7	70	19.88
8	69	25.06
9	74	25.65
10	79	19.95
11	76	22.6
12	76	26.42
13	62	20.32
14	69	19.37
15	72	24.22
16	67	32.11
17	74	25.39
18	69	24.67
19	78	27.13
20	71	23.05

Ví dụ minh hoạ 3

- Ta kiểm định giả thuyết: $H_0 : \rho(X, Y) = 0$ với đối thuyết $H_1 : \rho(X, Y) \neq 0$

Ví dụ minh hoạ 3

- Ta kiểm định giả thuyết: $H_0 : \rho(X, Y) = 0$ với đối thuyết $H_1 : \rho(X, Y) \neq 0$
- Ta có $n = 20, \bar{x} = 74.25, \bar{y} = 24.2415, \overline{xy} = 1801.486, \overline{x^2} = 5561.25, \overline{y^2} = 595.966$

Ví dụ minh hoạ 3

- Ta kiểm định giả thuyết: $H_0 : \rho(X, Y) = 0$ với đối thuyết $H_1 : \rho(X, Y) \neq 0$
- Ta có $n = 20, \bar{x} = 74.25, \bar{y} = 24.2415, \overline{xy} = 1801.486, \overline{x^2} = 5561.25, \overline{y^2} = 595.966$
- Suy ra

$$r = \frac{\overline{xy} - \bar{x}.\bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} = 0.0776$$

Ví dụ minh hoạ 3

- Thống kê

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}} = 0.33$$

Ví dụ minh hoạ 3

- Thống kê

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}} = 0.33$$

- p-giá trị $= 2\mathbb{P}(T_{n-2} > |t_{obs}|) = 2\mathbb{P}(T_{18} > 0.33) = 0.745$

Ví dụ minh hoạ 3

- Thống kê

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}} = 0.33$$

- p-giá trị $= 2\mathbb{P}(T_{n-2} > |t_{obs}|) = 2\mathbb{P}(T_{18} > 0.33) = 0.745$
- p-giá trị $= 0.745 >> \alpha = 0.05$ nên ta chấp nhận H_0.

Ví dụ minh hoạ 3

- Thống kê

$$t_{obs} = r\sqrt{\frac{n-2}{1-r^2}} = 0.33$$

- p-giá trị $= 2\mathbb{P}(T_{n-2} > |t_{obs}|) = 2\mathbb{P}(T_{18} > 0.33) = 0.745$
- p-giá trị $= 0.745 >> \alpha = 0.05$ nên ta chấp nhận H_0.
- Ta có thể kết luận rằng độ tuổi X và tỉ trọng cơ thể Y không tương quan nhau.

Ví dụ minh hoạ 3

Code Python:

```
import numpy as np
import pandas as pd
from scipy.stats import t
df = pd.read_csv("/content/drive/My Drive/Dataset/dataVD3_lecture9.csv")
x = df.values[:,1]
y = df.values[:,2]
n = len(x)
print(n)
xbar =np.mean(x)
print(xbar)
ybar =np.mean(y)
print(ybar)
xybar =np.mean(x*y)
print(xybar)
x2bar =np.mean(pow(x,2))
print(x2bar)
y2bar =np.mean(pow(y,2))
print(y2bar)
r =(xybar - xbar*ybar)/np.sqrt((x2bar - pow(xbar,2))*(y2bar - pow(ybar,2)))
print(r)
tobs = r*np.sqrt((n-2)/(1-pow(r,2)))
print(tobs)
p =2*(1-t.cdf(tobs,n-2))
print(p)
```

Bài tập 1

Một quỹ đầu tư có thể được mua trực tiếp (direct) từ các ngân hàng và các tổ chức tài chính khác, hoặc được mua thông qua các nhà môi giới (broker), những người tính phí cho dịch vụ này. Một nhóm các nhà nghiên cứu đã lấy mẫu ngẫu nhiên gồm các quỹ được mua trực tiếp và các quỹ được mua thông qua các nhà môi giới và ghi lại lợi nhuận ròng hàng năm. Số liệu thu được như sau:

Direct					Broker				
9.33	4.68	4.23	14.69	10.29	3.24	3.71	16.4	4.36	9.43
6.94	3.09	10.28	−2.97	4.39	−6.76	13.15	6.39	−11.07	8.31
16.17	7.26	7.1	10.37	−2.06	12.8	11.05	−1.9	9.24	−3.99
16.97	2.05	−3.09	−0.63	7.66	11.1	−3.12	9.49	−2.67	−4.44
5.94	13.07	5.6	−0.15	10.83	2.73	8.94	6.7	8.97	8.63
12.61	0.59	5.27	0.27	14.48	−0.13	2.74	0.19	1.87	7.06
3.33	13.57	8.09	4.59	4.8	18.22	4.07	12.39	−1.53	1.57
16.13	0.35	15.05	6.38	13.12	−0.8	5.6	6.54	5.23	−8.44
11.2	2.69	13.21	−0.24	−6.54	−5.75	−0.85	10.92	6.87	−5.72
1.14	18.45	1.72	10.32	−1.06	2.59	−0.28	−2.15	−1.69	6.95

Bài tập 1

Direct				Broker					
9.33	4.68	4.23	14.69	10.29	3.24	3.71	16.4	4.36	9.43
6.94	3.09	10.28	−2.97	4.39	−6.76	13.15	6.39	−11.07	8.31
16.17	7.26	7.1	10.37	−2.06	12.8	11.05	−1.9	9.24	−3.99
16.97	2.05	−3.09	−0.63	7.66	11.1	−3.12	9.49	−2.67	−4.44
5.94	13.07	5.6	−0.15	10.83	2.73	8.94	6.7	8.97	8.63
12.61	0.59	5.27	0.27	14.48	−0.13	2.74	0.19	1.87	7.06
3.33	13.57	8.09	4.59	4.8	18.22	4.07	12.39	−1.53	1.57
16.13	0.35	15.05	6.38	13.12	−0.8	5.6	6.54	5.23	−8.44
11.2	2.69	13.21	−0.24	−6.54	−5.75	−0.85	10.92	6.87	−5.72
1.14	18.45	1.72	10.32	−1.06	2.59	−0.28	−2.15	−1.69	6.95

Chúng ta có thể kết luận với mức ý nghĩa 5% rằng quỹ mua trực tiếp cho lợi nhuận trung bình cao hơn quỹ mua thông qua nhà môi giới không? Trước hết hãy kiểm định sự bằng nhau của phương sai của 2 tổng thể ở mức ý nghĩa 5%.

Bài tập 2

Để nghiên cứu hiệu quả của một loại thuốc mới để giảm các triệu chứng của bệnh hen suyễn ở trẻ em, người ta tiến hành thực hiện một thí nghiệm lâm sàng Giai đoạn II. Tổng số $n = 10$ người tham gia được chọn ngẫu nhiên để thử thuốc mới hoặc giả dược. Những người tham gia được yêu cầu ghi lại số lần thở gấp trong khoảng thời gian 1 tuần sau khi được điều trị. Dữ liệu được hiển thị bên dưới.
Giả dược: 7, 5, 6, 4, 12
Thuốc mới: 3, 6, 4, 2, 1
Có sự khác biệt nào về số lần khó thở trong khoảng thời gian 1 tuần ở những người tham gia dùng thuốc mới so với những người dùng giả dược không? Qua kiểm tra, có vẻ như những người tham gia sử dụng giả dược có nhiều cơn khó thở hơn, nhưng điều này có ý nghĩa thống kê không?

Bài tập 3

Quan sát độ tuổi X và nồng độ cholesterol trong máu Y của 18 nam thanh niên, người ta thu được số liệu như sau. Hãy kiểm định xem độ tuổi X và nồng độ cholesterol trong máu Y có tương quan dương không ở mức ý nghĩa 5%.

ID	age	chol
1	46	3.5
2	20	1.9
3	52	4
4	30	2.6
5	57	4.5
6	25	3
7	28	2.9
8	36	3.8
9	22	2.1
10	43	3.8
11	57	4.1
12	33	3
13	22	2.5
14	63	4.6
15	40	3.2
16	48	4.2
17	28	2.3
18	49	4