



1. Inferential Statistics
2. Standard Error
3. Confidence Intervals
- 4. Correlation**
5. Hypothesis testing

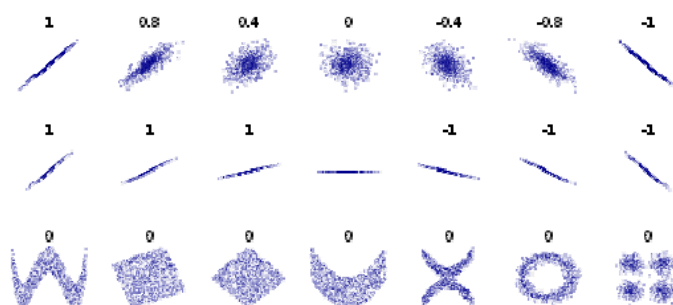
4. Correlation



▣ Pearson correlation

$$\text{correlation}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot (y_i - \bar{y})^2}}$$

- x, y độc lập $\Rightarrow \text{correlation}(x, y) = 0$; điều ngược lại không đúng





4. Correlation (tt.)

❑ Spearman rank correlation

- quan hệ thứ tự của hai dãy: ordinal, interval, ratio
- không có tham số (*nonparametric*): giá trị trung bình, ...

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \in [-1, 1]$$



4. Correlation (tt.)

❑ Spearman rank correlation

VD: 2 dãy số KHÔNG có đồng hạng

i	A(i)	Rank_A(i)	B(i)	Rank_B(i)	d(i)
1	35	3	30	5	-2
2	23	5	33	3	2
3	47	1	45	2	-1
4	17	6	23	6	0
5	10	7	8	8	-1
6	43	2	49	1	1
7	9	8	12	7	1
8	6	9	4	9	0
9	28	4	31	4	0

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 0.9$$

4. Correlation (tt.)



❑ Spearman rank correlation

VD: 2 dãy số CÓ đồng hạng

i	A(i)	Rank_A(i)	B(i)	Rank_B(i)	Rank_B'(i)	d(i)
1	35	3	30	5	5.5	-2.5
2	23	5	33	3	3	2
3	47	1	45	2	2	-1
4	17	6	30	5	5.5	0.5
5	10	7	8	8	8	-1
6	43	2	49	1	1	1
7	9	8	12	7	7	1
8	6	9	4	9	9	0
9	28	4	31	4	4	0

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} = 0.88$$



Nội dung bổ sung



1. Inferential Statistics
2. Standard Error
3. Confidence Intervals
4. Correlation
5. Hypothesis testing

5. Hypothesis Testing



❑ Kiểm định giả thuyết

- một giả thuyết *không chắc chắn* về 1 tham số của tổng thể
- dựa trên dữ liệu mẫu → nên hay không nên *bác bỏ* giả thuyết ?

❑ Giả thuyết thống kê (*statistical hypothesis*)

- giả thuyết về một vấn đề của tổng thể (tham số, phân phối, ...)
- H_0 (*null hypothesis*): giả thuyết 'vô hiệu' cần kiểm định
- H_1, H_a (*alternative hypothesis*): giả thuyết đảo/đối thuyết của H_0

5. Hypothesis Testing (tt.)



❑ Kiểm định hiệu lực của một phát biểu ('đáng nghi ngờ')

VD: Nhà sx cho biết thời gian máy thở cung cấp oxy TB = 75 phút

Đơn vị quản lý chất lượng lấy mẫu ngẫu nhiên để kiểm tra:
'*thời gian cung cấp oxy (trung bình) không ngắn hơn 75 phút*'.

$$\begin{cases} H_0 : \mu \geq 75 \\ H_a : \mu < 75 \end{cases} \leftarrow \text{điều phát biểu (giả định là đúng)}$$

Đơn vị QL không cần quan tâm (ước lượng) g.trị thật sự của μ ,
chủ yếu có đúng với điều phát biểu hay không mà thôi.

- bằng chứng thống kê *cho phép bác bỏ* H_0 hoặc ngược lại,
không cho phép bác bỏ $H_0 \Rightarrow H_0$ có hiệu lực

5. Hypothesis Testing (tt.)



❑ Kiểm định giả thuyết nghiên cứu (“mong đợi”)

VD: Mẫu xe hiện hành tiêu thụ 5 lít xăng / 100 km (1 lít → 20km)

Nhóm kỹ sư tiến hành một số cải tiến trên hệ thống phun xăng
→ hy vọng 1 lít đi được hơn 22km

$$\begin{cases} H_0 : \mu \leq 22 \\ H_a : \mu > 22 \end{cases} \leftarrow \text{điều mong đợi}$$

- bằng chứng thống kê *không cho phép bác bỏ* H_0 , cần thực hiện những nghiên cứu khác, hoặc ngược lại, *cho phép bác bỏ* H_0
⇒ ủng hộ kết quả nghiên cứu

5. Hypothesis Testing (tt.)



❑ Kiểm định giả thuyết cả 2 phía

VD1: Số mặt hàng TB / hóa đơn tại 1 siêu thị trong năm qua là 8.

Bộ phận nghiên cứu thị trường muốn xem xét có hay không có sự thay đổi về số lượng mặt hàng trung bình / hóa đơn

$$\begin{cases} H_0 : \mu = 8 \\ H_a : \mu \neq 8 \end{cases} \leftarrow \begin{array}{l} \text{'hai phía' của ngưỡng} \\ \text{có vai trò như nhau} \end{array}$$

- bằng chứng thống kê *cho phép bác bỏ* H_0 , hoặc ngược lại, *không cho phép bác bỏ* $H_0 \Rightarrow H_0$ có hiệu lực

VD2: Kiểm soát chất lượng sản phẩm (*lot-acceptance sampling*).



5. Hypothesis Testing (tt.)

□ 3 loại kiểm định giả thuyết

- hai phía (*two-tailed/two-sided test*)

$$\begin{cases} H_0 : \theta = \theta_0 \\ H_a : \theta \neq \theta_0 \end{cases}$$

- một phía (*one-tailed/one-sided test*)

$$\begin{cases} H_0 : \theta_0 \leq \theta \\ H_a : \theta_0 > \theta \end{cases} \quad (\text{left-tailed})$$

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_a : \theta > \theta_0 \end{cases} \quad (\text{right-tailed})$$



5. Hypothesis Testing (tt.)

□ Giả thuyết thống kê: *mạnh dạn bác bỏ, miễn cưỡng chấp nhận*

- *tiêu chuẩn kiểm định*: phân phối xs được dùng trong kiểm định
- 2 loại sai sót trong kiểm định

	H_0 ĐÚNG	H_0 SAI
Không bác bỏ H_0	Kết luận đúng	Sai lầm loại II
Bác bỏ H_0	Sai lầm loại I	Kết luận đúng

- *mức ý nghĩa kiểm định* (*level of significance*) α

$$\alpha = P(\text{xs sai lầm loại I})$$

Nếu chi phí cho sai lầm loại I lớn \rightarrow chọn α thấp, và ngược lại.

5. Hypothesis Testing (tt.)



❑ Giả thuyết thống kê: *mạnh dạn bác bỏ, miễn cưỡng chấp nhận*

- *năng lực kiểm định* ($1 - \beta$)

$$\beta = P(\text{xs sai lầm loại II})$$

$$(1 - \beta) = P(\text{bác bỏ } H_0 \mid H_0 \text{ sai})$$

- “không bác bỏ H_0 ” thay vì “chấp nhận H_0 ”: tránh sai lầm loại II

5. Hypothesis Testing (tt.)



❑ Miền bác bỏ (*rejection region*): chứa những giá trị thống kê (không phải xs) làm cơ sở bác bỏ H_0

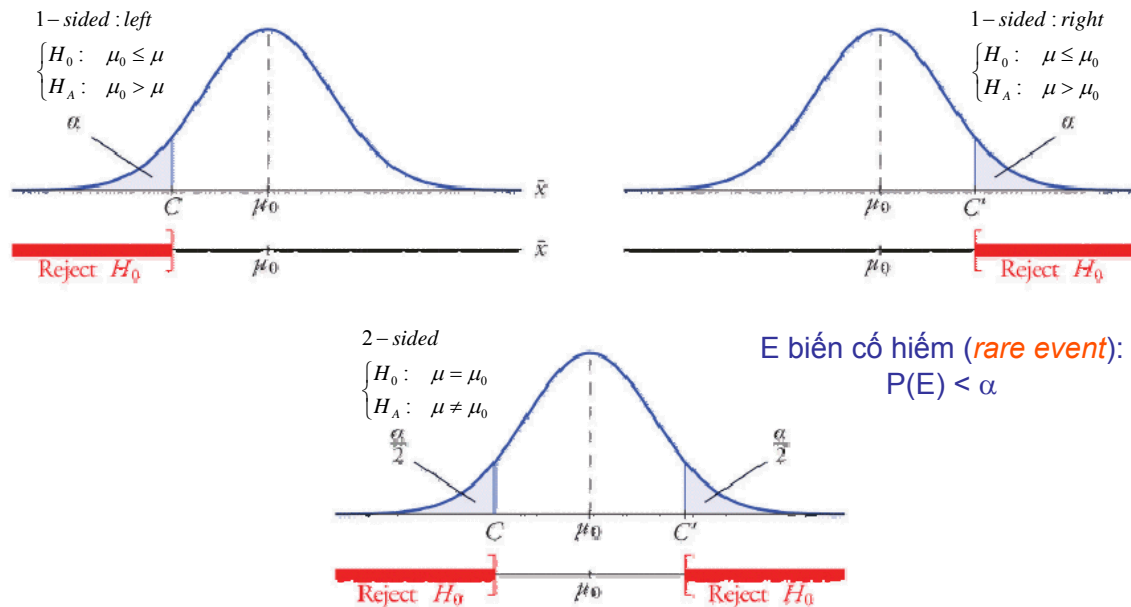
- quy trình kiểm định bắt đầu bằng sự giả định H_0 là đúng
- nếu trị thống kê (dựa trên mẫu) thuộc miền này → **bác bỏ H_0**
- α được sử dụng để xác định (tính toán) miền bác bỏ

5. Hypothesis Testing (tt.)



□ Miền bác bỏ: chứa các giá trị thống kê làm cơ sở bác bỏ H_0

- các giá trị tới hạn (*critical value*) C, C' : tra các bảng phân phối



B7. Inferential Statistics

Bổ sung thêm cho bài giảng

180

5. Hypothesis Testing (tt.)



□ Miền bác bỏ: chứa các giá trị thống kê làm cơ sở bác bỏ H_0

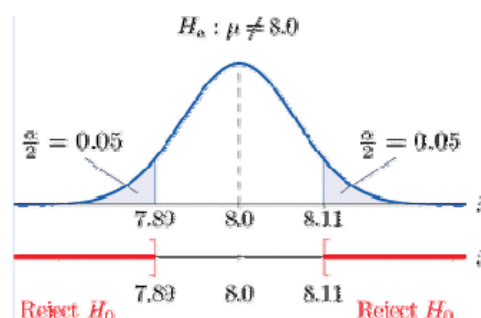
$$\begin{cases} H_0: \mu = 8.0 \\ H_a: \mu \neq 8.0 \end{cases} \quad n = 5 \quad \sigma = 0.15 \quad \alpha = 0.10$$

$$\mu_{\bar{X}} = \mu = 8.0 \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.15}{\sqrt{5}} = 0.067$$

$$\alpha/2 = 0.05 \rightarrow \text{tra Z table: } z_{0.05} = 1.645$$

$$C = 8.0 - (1.645)(0.067) = 7.89$$

$$C' = 8.0 + (1.645)(0.067) = 8.11$$



Nếu trung bình của 1 mẫu < 7.89
hay > 8.11 thì bác bỏ H_0 .

B7. Inferential Statistics

Bổ sung thêm cho bài giảng

181



5. Hypothesis Testing

5.1 Kiểm định trung bình tổng thể μ

- Trường hợp ĐÃ biết σ
- Trường hợp CHƯA biết σ

5.2 Kiểm định (so sánh) trung bình 2 mẫu

5.3 Kiểm định Chi bình phương

5.4 Kiểm định A/B

5.5 Phân tích phương sai (Analysis Of Variance – ANOVA)



5.1 Kiểm định trung bình tổng thể

□ Trường hợp **ĐÃ BIẾT** σ trước khi lấy mẫu

- khai thác dữ liệu trong quá khứ hoặc từ nguồn thông tin khác
- tổng thể có *phân phối chuẩn* hoặc *cỡ mẫu đủ lớn* ($n \geq 30$)

Giả thuyết	1 – sided : right	1 – sided : left	2 – sided
	$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_a : \mu > \mu_0 \end{cases}$	$\begin{cases} H_0 : \mu_0 \leq \mu \\ H_a : \mu_0 > \mu \end{cases}$	$\begin{cases} H_0 : \mu = \mu_0 \\ H_a : \mu \neq \mu_0 \end{cases}$
Thống kê kiểm định	$z = \frac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}} \sim N(0,1)$		

sai số chuẩn SE

5.1 Kiểm định trung bình tổng thể (tt.)



□ Trường hợp ĐÃ BIẾT σ trước khi lấy mẫu

VD1:
$$\begin{cases} H_0: 3 = \mu_0 \leq \mu \\ H_a: \mu < \mu_0 = 3 \end{cases} \quad \sigma = 0.18 \quad n = 36 \quad \alpha = 0.01$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 0.18 / \sqrt{36} = 0.03 \quad z = \frac{(\bar{x} - 3)}{0.03}$$

Lấy mẫu và có: $\bar{x} = 2.92 \Rightarrow z = -2.67$

→ quy tắc kết luận về H_0 dựa trên z ?

5.1 Kiểm định trung bình tổng thể (tt.)



□ Phương pháp bác bỏ dựa trên giá trị tới hạn

- cột mốc cho giá trị thống kê kiểm định
- tạo miền bác bỏ có diện tích = α

Miền bác bỏ H_0	1 – sided : right $z_{\alpha} \leq z$	1 – sided : left $z \leq -z_{\alpha}$	2 – sided $(z \leq -z_{\alpha/2}) \vee (z_{\alpha/2} \leq z)$
----------------------	--	--	--

5.1 Kiểm định trung bình tổng thể (tt.)

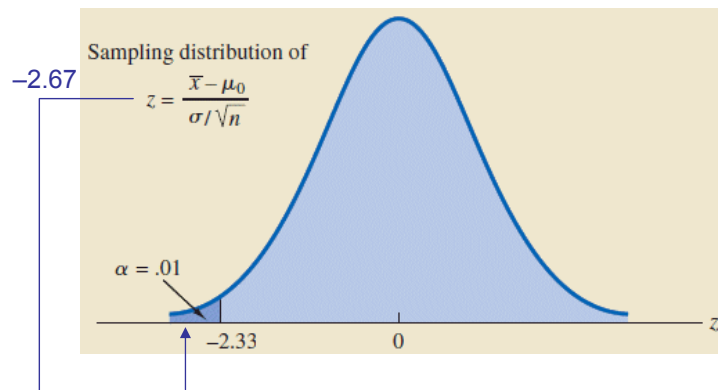


❑ Trường hợp ĐÃ BIẾT σ trước khi lấy mẫu

VD1: Với phần diện tích $\alpha = 0.01$, tra (ngược) bảng Z:

Xs tích lũy 0.0099 (~ 0.01) $\Rightarrow z_{\alpha} = -2.33$ (dòng -2.3 , cột 0.03)

Vì $z = -2.67 < -2.33 = z_{\alpha}$ cho nên có thể bác bỏ H_0 .



5.1 Kiểm định trung bình tổng thể (tt.)



❑ Phương pháp bác bỏ **dựa trên trị số p** (p -value)

- Với mỗi mức ý nghĩa $\alpha \rightarrow$ miền bác bỏ
- Trị số p: ngưỡng xác suất còn có thể chấp nhận giả thuyết H_0
 $\Rightarrow H_0$ luôn luôn bị bác bỏ với mọi $\alpha \geq p$
- Nếu p quá nhỏ thì H_0 gần như bị bác bỏ hoàn toàn
- trị số p có thể được tính toán từ giá trị thống kê kiểm định (z)

5.1 Kiểm định trung bình tổng thể (tt.)



❑ Trường hợp ĐÃ BIẾT σ trước khi lấy mẫu

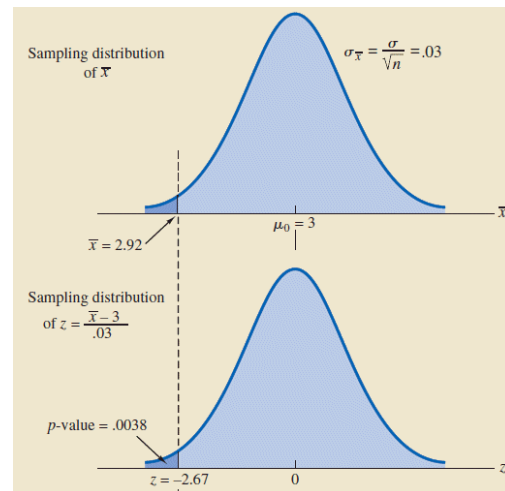
VD: $\bar{x} = 2.92$ có đủ nhỏ (so với $\mu_0 = 3$) để bác bỏ H_0 ?

$$z = -2.67 \Rightarrow p = P(Z \leq -2.67)$$

Tra (xuôi) bảng Z, (dòng -2.6, cột 0.07), ta có: $p = 0.0038$

$\alpha = 0.01$ nghĩa là chấp nhận mức xs (sai lầm) là 0.01 để bác bỏ H_0 .

Vì $p < \alpha$ cho nên có thể bác bỏ H_0 .



5.1 Kiểm định trung bình tổng thể (tt.)



❑ Trường hợp ĐÃ BIẾT σ trước khi lấy mẫu

VD: Mức độ hài lòng hiện tại = 77/100. Lấy ý kiến của 20 khách, \Rightarrow mức độ hài lòng trung bình là 80/100. Mức độ hài lòng của khách có thật sự tăng hay không, với $\alpha = 0.01$ và $\sigma = 8$?

Giả thuyết 1-sided : right

$$\begin{cases} H_0 : \mu \leq \mu_0 = 77 \\ H_a : \mu > \mu_0 = 77 \end{cases}$$

$$\text{Trị thống kê } z = \frac{(80 - 77)}{8 / \sqrt{20}} = 1.68$$

Kết luận: $z = 1.68 < z_{1\%} = 2.33 \Rightarrow \text{KHÔNG BÁC BỎ } H_0$.

5.1 Kiểm định trung bình tổng thể (tt.)



□ Các bước kiểm định giả thuyết dựa trên mẫu dữ liệu thu thập

B1. Xây dựng giả thuyết: H_0 và H_a .

B2. Chọn mức ý nghĩa α : .1, .05, .01 (\rightarrow độ tin cậy .9, .95, .99).

B3. Tính giá trị thống kê kiểm định: chọn phân phối và công thức thống kê kiểm định \rightarrow tính toán giá trị dựa trên mẫu dữ liệu.

A. Phương pháp dựa trên trị số p:

B4a. Tính trị số p: dựa trên giá trị thống kê kiểm định, tra (xuôi) bảng Z.

B5a. Rút ra kết luận về H_0 : bác bỏ H_0 nếu $p \leq \alpha$.

B. Phương pháp dựa trên giá trị tới hạn:

B4b. Xác định miền bác bỏ: ứng với α , tra (ngược) bảng Z $\rightarrow z_\alpha$

B5b. Rút ra kết luận về H_0 : dựa trên miền bác bỏ.

5.1 Kiểm định trung bình tổng thể (tt.)



□ Kiểm định 2 phía dựa trên khoảng tin cậy

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu \neq \mu_0 \end{cases} \quad CI_\alpha = \bar{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Nếu $\mu_0 \in CI_\alpha = \left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right]$ thì không bác bỏ H_0 , và ngược lại.

5.1 Kiểm định trung bình tổng thể (tt.)



❑ Trường hợp CHƯA BIẾT σ trước khi lấy mẫu

- Nếu cỡ mẫu đủ lớn ($n \geq 30$) thì dùng phương sai mẫu (theo CLT, X không cần có phân phối chuẩn)
- Nếu ($n < 30$) thì X cần có *phân phối chuẩn*

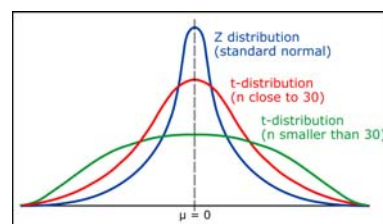
	Lower Tail Test	Upper Tail Test	Two-Tailed Test
Hypotheses	$H_0: \mu \geq \mu_0$ $H_a: \mu < \mu_0$	$H_0: \mu \leq \mu_0$ $H_a: \mu > \mu_0$	$H_0: \mu = \mu_0$ $H_a: \mu \neq \mu_0$
Test Statistic	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$
Rejection Rule: p-Value Approach	Reject H_0 if $p\text{-value} \leq \alpha$	Reject H_0 if $p\text{-value} \leq \alpha$	Reject H_0 if $p\text{-value} \leq \alpha$
Rejection Rule: Critical Value Approach	Reject H_0 if $t \leq -t_\alpha$	Reject H_0 if $t \geq t_\alpha$	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

5.1 Kiểm định trung bình tổng thể (tt.)



❑ Phân phối T (*Student's T distribution*)

$$T \sim \left(0, \frac{df}{df - 2} \right)$$



Khi df đủ lớn thì $\sigma \rightarrow 1$ và T có phân phối chuẩn tắc

Những trường hợp khả dụng:

- tổng thể $X \sim (\mu, \sigma)$
- X phân phối đối xứng, unimodal (1 đỉnh), không outlier, $n \geq 30$
- X phân phối nhọn vừa phải, unimodal, không outlier, $n \geq 40$

5.1 Kiểm định trung bình tổng thể (tt.)



□ Bậc tự do (*degrees of freedom*)

- số lượng tối đa giá trị quan sát có thể thay đổi 1 cách độc lập

Giả sử mẫu S có kích thước là n : $S = \{x_1, x_2, \dots, x_n\}$

a) Giá trị của các quan sát trong S có thể tùy ý
(không có quan hệ ràng buộc nào) $\Rightarrow df = n$

b) Nếu biết giá trị của \bar{S} thì giá trị của x_i sẽ “*bị ràng buộc*”
(không còn tự do) bởi \bar{S} và tập $DF = S - \{x_i\}$
 $\Rightarrow df = (n - 1)$

5.1 Kiểm định trung bình tổng thể (tt.)



□ Kiểm định T (*T-test*)

- kiểm định sự khác biệt về đặc trưng (tham số) giữa 2 nhóm
- kiểm định sự khác biệt bên trong 1 nhóm
- t-score càng lớn \Rightarrow sự khác biệt giữa các nhóm càng lớn
- t-score càng lớn \Rightarrow khả năng mà kết quả lặp lại rất cao

□ Các loại T-test

- *One sample*: kiểm định mean của 1 nhóm với mean đã cho
- *Independent samples*: so sánh means của 2 nhóm
- *Paired sample*: so sánh means của 1 nhóm (nhiều lần/mẫu)

5.1 Kiểm định trung bình tổng thể (tt.)



❑ Trường hợp CHƯA BIẾT σ trước khi lấy mẫu

VD: Thang điểm đánh giá sân bay: 0 đến 10; ngưỡng 7.0 là tốt. Lấy ý kiến 60 hành khách, điểm đánh giá trung bình là 7.25, với độ lệch chuẩn của mẫu $s = 1.052$. Với mức ý nghĩa $\alpha = .05$, sân bay có thật sự tốt ?

1-sided : right

$$\begin{cases} H_0: \mu \leq \mu_0 = 7 \\ H_a: \mu > \mu_0 = 7 \end{cases} \quad t = \frac{(7.25 - 7)}{1.052 / \sqrt{60}} = 1.84$$

Phân phối mẫu có bậc tự do là 59. Kiểm định phía phải nên p là diện tích dưới đường phân phối, bên phải của $t = 1.84$

$$t_{59,0.05} = 1.671 < t = 1.84 < 2.001 = t_{59,0.025}$$

$$\Rightarrow 0.025 < p < 0.05 = \alpha \Rightarrow \text{BÁC BỎ } H_0.$$

5.1 Kiểm định trung bình tổng thể (tt.)



❑ Trường hợp CHƯA BIẾT σ trước khi lấy mẫu

VD: Tuổi thọ trung bình của bóng đèn là 65 giờ. Lấy ngẫu nhiên 21 bóng đèn, có tuổi thọ trung bình là 62,5 giờ. Tuổi thọ của bóng đèn có được cải thiện hay không, với $\alpha = .01$ và $s_x = 3$?

1-sided : left

$$\begin{cases} H_0: 65 = \mu_0 \leq \mu \\ H_a: 65 = \mu_0 > \mu \end{cases}$$

$$t = \frac{(62.5 - 65)}{3 / \sqrt{21}} = -3.82$$

$$t = -3.82 < -t_{20,0.01} = -2.528 \Rightarrow \text{BÁC BỎ } H_0.$$

5. Hypothesis Testing



5.1 Kiểm định trung bình tổng thể μ

5.2 Kiểm định (so sánh) trung bình 2 mẫu

- 2 mẫu độc lập đã biết phương sai
- 2 mẫu độc lập chưa biết phương sai
- 2 mẫu phụ thuộc

5.3 Kiểm định Chi bình phương

5.4 Kiểm định A/B

5.5 Phân tích phương sai (Analysis Of Variance – ANOVA)

5.2 Kiểm định trung bình 2 mẫu



□ Kiểm định 2 mẫu độc lập (*Independent Samples Test*)

- VD: sự khác biệt giữa chiều cao trung bình của SV nam và nữ
- sự khác biệt giữa means của 2 nhóm có phải là do tình cờ ?
- biến độc lập, kiểu categorical (VD: giới tính) \Rightarrow tạo 2 nhóm
- biến phụ thuộc X (VD: chiều cao), liên tục \sim phân phối chuẩn

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định z với 2 mẫu độc lập (*Independent Samples Test*)

- Trường hợp ĐÃ BIẾT các phương sai trước khi lấy mẫu

Giả thuyết	1 – sided : right	1 – sided : left	2 – sided
	$\begin{cases} H_0 : \mu_X - \mu_Y \leq D_0 \\ H_a : \mu_X - \mu_Y > D_0 \end{cases}$	$\begin{cases} H_0 : D_0 \leq \mu_X - \mu_Y \\ H_a : D_0 > \mu_X - \mu_Y \end{cases}$	$\begin{cases} H_0 : \mu_X - \mu_Y = D_0 \\ H_a : \mu_X - \mu_Y \neq D_0 \end{cases}$
Thống kê kiểm định	$z = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}}}$		

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định t với 2 mẫu độc lập (*Independent samples T-test*)

- Trường hợp CHƯA BIẾT các phương sai (*khác nhau*)

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \quad df = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y} \right)^2}{\frac{1}{(n_X - 1)} \left(\frac{s_X^2}{n_X} \right)^2 + \frac{1}{(n_Y - 1)} \left(\frac{s_Y^2}{n_Y} \right)^2}$$

s_X, s_Y : các phương sai mẫu

df: bậc tự do (*degrees of freedom*)

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định t với 2 mẫu độc lập (*Independent samples T-test*)

VD:

$X = \{ 19.7, 20.4, 19.6, 17.8, 18.5, 18.9, 18.3, 18.9, 19.5, 21.95 \}$

$Y = \{ 28.3, 26.7, 20.1, 23.3, 25.2, 22.1, 17.7, 27.6, 20.6, 13.7, 23.2, 17.5, 20.6, 18, 23.9, 21.6, 24.3, 20.4, 23.9, 13.3 \}$

$$\bar{X} = 19.4, s^2_X = 1.4, \bar{Y} = 21.6, s^2_Y = 17.1$$

$$t = 2.24787, df = 24.38 \rightarrow 24$$

$$\text{Tra t-table: } (\alpha = 0.05, df = 24) \rightarrow t_{\alpha/2, df} = 2.064 < t = 2.24787$$

\Rightarrow bác bỏ $H_0 \Rightarrow 2$ means khác nhau

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định t với 2 mẫu độc lập (*Independent samples T-test*)

- Trường hợp CHƯA BIẾT các phương sai (*giống nhau*)

$$t = \frac{(\bar{x} - \bar{y}) - D_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}}$$

s_X, s_Y : các phương sai mẫu

$(n_X + n_Y - 2)$: bậc tự do

5.2 Kiểm định trung bình 2 mẫu (tt.)



❑ Kiểm định t với 2 mẫu độc lập (*Independent samples T-test*)

VD: $X = \{ 1, 2, 2, 3, 3, 4, 4, 5, 5, 6 \}$

$Y = \{ 1, 2, 4, 5, 5, 5, 6, 6, 7, 9 \}$

$t = -1.69, \quad df = (10 + 10 - 2) = 18$

Tra t-table: ($\alpha = 0.05, df = 18$) $\rightarrow t_{\alpha/2, df} = 2.101$

$-2.101 = -t_{\alpha/2, df} < t = -1.69 < t_{\alpha/2, df} = 2.10$

\Rightarrow không thể bác bỏ H_0

(p-value = 0.1073 > α)

5.2 Kiểm định trung bình 2 mẫu (tt.)



❑ Kiểm định mẫu phụ thuộc/liên quan (*Paired / correlated T-test*)

- 2 mẫu có những cặp đôi (matched pairs) cùng đơn vị đo lường
- đo lường nhiều lần trên 1 tổng thể ('đo' bệnh nhân TRƯỚC và SAU khi điều trị)

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định mẫu phụ thuộc/liên quan (*Paired / correlated T-test*)

$$d_i = (x_i - y_i) \quad \bar{d} = \frac{\sum_{i=1}^n d_i}{n} \quad s_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{(n-1)}}$$

$$t = \frac{\bar{d}}{s_d \sqrt{n}}$$

$$df = (n - 1)$$

5.2 Kiểm định trung bình 2 mẫu (tt.)



□ Kiểm định mẫu phụ thuộc/liên quan (*Paired / correlated T-test*)

VD: $X = \{ 6.0, 5.0, 7.0, 6.2, 6.0, 6.4 \}$

$$Y = \{ 5.4, 5.2, 6.5, 5.9, 6.0, 5.8 \}$$

$$\bar{d} = \frac{1.8}{6} = 0.3 \quad s_d = \sqrt{\frac{0.56}{5}} = 0.335$$

$$t = \frac{0.3}{0.335\sqrt{6}} = 2.20$$

$$\text{Tra t-table: } (\alpha = 0.05, df = 5) \rightarrow t_{\alpha/2, df} = 2.571$$

$$-2.571 = -t_{\alpha/2, df} < t = 2.20 < t_{\alpha/2, df} = 2.571$$

\Rightarrow không thể bác bỏ H_0

$$(p\text{-value} = 0.08 > \alpha)$$

5. Hypothesis Testing



5.1 Kiểm định trung bình tổng thể μ

5.2 Kiểm định (so sánh) trung bình 2 mẫu

5.3 Kiểm định Chi bình phương

5.4 Kiểm định A/B

5.5 Phân tích phương sai (Analysis Of Variance – ANOVA)

5.3 Kiểm định Chi bình phương



□ Phân phối Chi bình phương χ^2 (*Chi-squared distribution*)

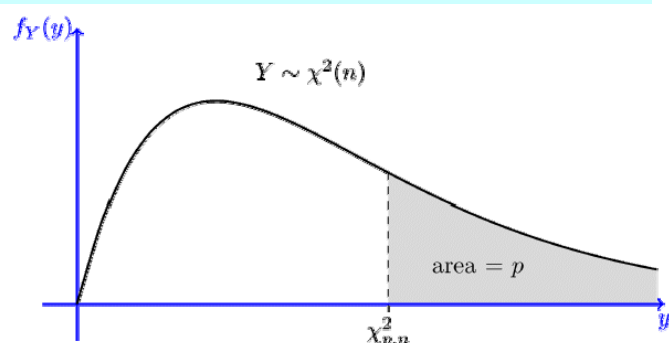
Các biến ngẫu nhiên, độc lập: $Z_1, Z_2, \dots, Z_n \sim N(0, 1)$

$$Y = Z_1^2 + Z_2^2 + \dots + Z_n^2 \sim \chi^2(n)$$

với n bậc tự do

$$f(y) = \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} y^{\frac{n}{2}-1} e^{-\frac{y}{2}}, \quad \forall y > 0 \quad \Gamma(\alpha) = \int_0^{\infty} x^{(\alpha-1)} e^{-x} dx, \quad \forall \alpha > 0$$

- $E[Y] = n$
- $\text{Var}(Y) = 2n$
- n lớn \rightarrow tiệm cận phân phối chuẩn



5.3 Kiểm định Chi bình phương (tt.)



□ Kiểm định tính độc lập của 2 biến kiểu phân loại (*categorical*)

X có thể lấy các giá trị (categories/levels) x_1, x_2, \dots, x_m

Y có thể lấy các giá trị (categories/levels) y_1, y_2, \dots, y_n

Giả thuyết: H_0 : X và Y độc lập

H_a : X và Y không độc lập

5.3 Kiểm định Chi bình phương (tt.)



□ Kiểm định tính độc lập của 2 biến kiểu phân loại (*categorical*)

- Bảng tương quan (*contingency table*): $A \in M_{m,n}(\mathbb{R}^+)$, $a_{ij} \geq 5$

$$\begin{array}{c} x_1 \\ x_2 \\ \vdots \\ x_m \end{array} \begin{array}{c} y_1 \quad y_2 \quad \cdots \quad y_n \\ \left(\begin{array}{cccc} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{array} \right) \end{array} \begin{array}{l} \longrightarrow R_i = \sum_{j=1}^n a_{ij} \\ \downarrow \\ C_j = \sum_{i=1}^m a_{ij} \longrightarrow S = \sum_{i=1}^m R_i = \sum_{j=1}^n C_j \end{array}$$

5.3 Kiểm định Chi bình phương (tt.)



□ Kiểm định tính độc lập của 2 biến kiểu phân loại (categorical)

- giá trị thống kê χ^2

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \quad e_{ij} = \frac{R_i C_j}{S}$$

o_{ij} : observation

e_{ij} : expectation

bậc tự do $df = (m - 1)(n - 1)$

- miền bác bỏ $H_0 = [\chi^2_{\alpha}, \infty)$: nếu giá trị $\chi^2 > \text{critical value } \chi^2_{\alpha}$ (tra bảng) thì có sự phụ thuộc đáng kể giữa 2 biến
- giá trị χ^2 càng nhỏ \rightarrow mối quan hệ càng lớn

5.3 Kiểm định Chi bình phương (tt.)



□ Kiểm định tính độc lập của 2 biến kiểu phân loại (categorical)

- các quan sát (observations) và các kỳ vọng (expectations)

	B1		B2		B3		
	O	E	O	E	O	E	
A1	91	79.55	104	111.8	235	238.7	430
A2	39	29.6	73	41.6	48	88.8	160
A3	18	38.85	31	54.6	161	116.6	210
	148		208		444		800

$$\chi^2 = 86.02$$

Tra t-table: $(\alpha = 0.05, df = 4) \rightarrow \chi_{\alpha, df} = 9.488$

$$9.488 = \chi_{\alpha, df} < \chi^2 = 86.02$$

\Rightarrow bác bỏ $H_0 \Rightarrow$ có mối quan hệ phụ thuộc



5. Hypothesis Testing



5.1 Kiểm định trung bình tổng thể μ

5.2 Kiểm định (so sánh) trung bình 2 mẫu

5.3 Kiểm định Chi bình phương

5.4 Kiểm định A/B

5.5 Phân tích phương sai (Analysis Of Variance – ANOVA)

5.4 Kiểm định A/B



□ Kiểm định A/B (*A/B testing* / *split testing*)

- so sánh hiệu quả của 2 phương án
- hiệu quả dựa trên tỉ lệ chuyển đổi (*conversion rate – CR*):
số lượng tương tác / tổng số thử nghiệm
- áp dụng kiểm định χ^2

A (hay B) có 2 mức: $|\text{DOM}(A)| = 2$

H_0 : Không có sự khác biệt về CR giữa A và B

H_a : Có sự khác biệt về CR giữa A và B





5. Hypothesis Testing

5.1 Kiểm định trung bình tổng thể μ

5.2 Kiểm định (so sánh) trung bình 2 mẫu

5.3 Kiểm định Chi bình phương

5.4 Kiểm định A/B

5.5 Phân tích phương sai (Analysis Of Variance – ANOVA)



5.5 Phân tích ANOVA

□ Kiểm định trung bình của NHIỀU tổng thể (*One-way ANOVA*)

- mỗi quan hệ giữa một biến phụ thuộc (định lượng, liên tục) với một biến độc lập (phân hoạch dữ liệu thành k nhóm)

- $k = 2$: áp dụng t-test để so sánh trung bình của 2 nhóm

VD: Chiều cao trung bình \leftrightarrow giới tính (2 nhóm: Nam, Nữ)

- $k > 2$: áp dụng t-test để so sánh từng cặp đôi

Số lần kiểm định: $n = \binom{k}{2} = \frac{k!}{2!(k-2)!} \rightarrow$ không hiệu quả

Nếu 1 lần t-test có xs sai lầm loại I là α thì sai lầm tích lũy: $(n\alpha)$

\Rightarrow ANOVA (Ronald FISHER)

5.5 Phân tích ANOVA (tt.)



❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

k nhóm từ biến độc lập P (các dân tộc)

Sự khác nhau/biến thiên của biến phụ thuộc X (chiều cao):

- giữa những cá thể trong một nhóm (*within-group variation*) s_W : hoàn toàn do yếu tố ngẫu nhiên
- giữa những cá thể thuộc k nhóm (*between-group variation*) s_B : vừa do yếu tố ngẫu nhiên, vừa do đặc trưng của từng nhóm (do tác động của biến độc lập)

5.5 Phân tích ANOVA (tt.)



❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

Phân tích tỉ số phương sai:

$$F = \frac{s_B}{s_W} = \frac{s_{random} + s_P}{s_{random}}$$

- Nếu $F \approx 1 \Rightarrow s_P$ nhỏ: tác động của P không đáng kể, nghĩa là sự biến thiên của X giữa các nhóm chủ yếu là do ngẫu nhiên \Rightarrow không có sự khác biệt đáng kể giữa các μ_1, \dots, μ_k
- Nếu $F \gg 1 \Rightarrow s_P$ lớn: P có tác động mạnh đến biến thiên của X \Rightarrow có tối thiểu 1 sự khác biệt đáng kể giữa các μ_1, \dots, μ_k



5.5 Phân tích ANOVA (tt.)

□ Kiểm định trung bình của NHIỀU tổng thể (*One-way ANOVA*)

- các quan sát độc lập
- phân phối chuẩn
- σ giống nhau



5.5 Phân tích ANOVA (tt.)

□ Kiểm định trung bình của NHIỀU tổng thể (*One-way ANOVA*)

$H_0: \mu_1, \dots, \mu_k$ (không có sự khác biệt)

H_a : Có tối thiểu 1 sự khác biệt giữa các μ_1, \dots, μ_k

Xét mẫu thứ j gồm n_j quan sát: $\begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{n_j, j} \end{pmatrix} \rightarrow \bar{x}_j$ \bar{x} : TB toàn bộ dữ liệu (k nhóm)

Tổng bình phương (*sum of squares*) biến thiên trong nhóm:

$$SSW_j = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \quad SSW = \sum_{j=1}^k SSW_j \quad dfW = (n - k)$$

Tổng bình phương biến thiên giữa các nhóm:

$$SSB = \sum_{j=1}^k n_j (x_j - \bar{x})^2 \quad dfB = (k - 1)$$



5.5 Phân tích ANOVA (tt.)

❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

ANOVA table

	SS	df	Mean of squares (MS)
<i>Within</i>	<i>SSW</i>	$(n - k)$	$MSW = \frac{SSW}{dfW}$
<i>Between</i>	<i>SSB</i>	$(k - 1)$	$MSB = \frac{SSB}{dfB}$

$$F = \frac{MSB}{MSW}$$

- F lớn hơn giá trị lý thuyết $F_{\alpha, dfB, dfW} \Rightarrow$ bác bỏ H_0 (có khác biệt)
- giá trị F càng lớn \Rightarrow sự khác biệt càng lớn



5.5 Phân tích ANOVA (tt.)

❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

VD: Dữ liệu của 4 nhóm đối tượng

	A	B	C	D	
	8	7	28	28	
	9	17	21	16	
	11	10	26	13	
	4	14	11	12	
	7	12	24	9	
	8	24	19	10	
	5	11		11	
		22		17	
				15	
Mean	7.4286	14.6250	21.5000	14.3333	Overall 14.4717
n _j	7	8	6	9	30
SSB	347.2422	0.1879	298.3798	0.1724	643.9823
SSW	33.7143	247.8750	185.5000	212.0000	679.0893

Source	df	SS	MS
B	3	643.9823	214.6608
W	26	679.0893	26.1188
F statistic		8.2188	

5.5 Phân tích ANOVA (tt.)

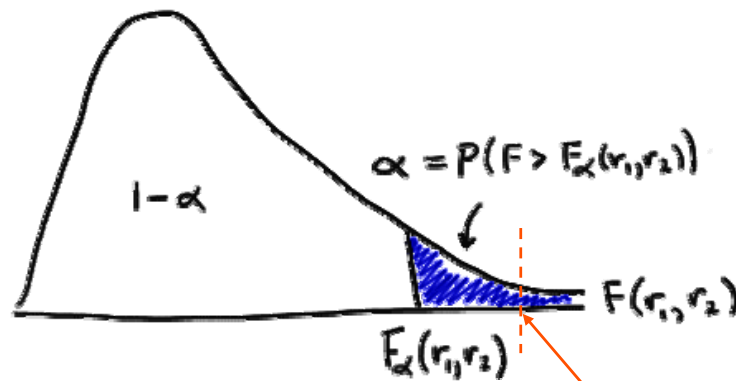


❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

Tra F-table: $\alpha = .05$, $df_B = 3$, $df_W = 26 \rightarrow F_\alpha = 2.9752 < F = 8.2126$

\Rightarrow bác bỏ $H_0 \Rightarrow$ có sự khác biệt đáng kể

(hoặc trị số $p < \alpha \Rightarrow$ bác bỏ H_0)



<https://newonlinecourses.science.psu.edu/stat414/node/294/>
(10/2019)

5.5 Phân tích ANOVA (tt.)



❑ Kiểm định trung bình của NHIỀU tổng thể (One-way ANOVA)

VD: Dữ liệu của 4 nhóm đối tượng

	A	B	C	D	
	25	45	30	54	
	30	55	29	60	
	28	29	33	51	
	36	56	37	62	
	29	40	27	73	Overall
Mean					
n_j					
SSB					
SSW					

5.5 Phân tích ANOVA (tt.)



❑ Kiểm định trung bình của NHIỀU tổng thể (*One-way ANOVA*)

- ANOVA không xác định rõ những (μ_i, μ_j) nào có sự khác biệt
- hậu kiểm (*post-hoc test*) khi cần xác định các (μ_i, μ_j) khác biệt: Least Significant Difference (LSD), Tukey HSD test, Scheffe's test, Bonferroni test, ...

5.5 Phân tích ANOVA (tt.)



❑ Kiểm định trung bình của NHIỀU tổng thể (*One-way ANOVA*)

- kiểm định các quần thể có phân phối chuẩn: Shapiro-Wilk test
- kiểm định các quần thể có cùng σ : Levene test, Bartlett test, ...
→ nếu KHÔNG cùng σ thì áp dụng kiểm định Welch



5.5 Phân tích ANOVA (tt.)



□ Kiểm định trung bình của NHIỀU tổng thể (*Two-way ANOVA*)

- biến độc lập A tạo thành m nhóm (trên DÒNG)
- biến độc lập B tạo thành n nhóm (trên CỘT)

	B_1	B_2	\dots	B_n
A_1	S_{11}	S_{12}	\dots	S_{1n}
A_2	S_{21}	S_{22}	\dots	S_{2n}
\vdots	\vdots	\vdots	\vdots	\vdots
A_m	S_{m1}	S_{m2}	\dots	S_{mn}

S_{ij} : các mẫu
cùng kích thước
 $|S_{ij}| = r$

- tính phương sai theo DÒNG và theo CỘT

5.5 Phân tích ANOVA (tt.)



□ Kiểm định trung bình của NHIỀU tổng thể (*Two-way ANOVA*)

VD:

<i>Genotype</i>	Y1	Y2	Y3	$n = 3$
G1	1.53	4.08	6.69	← mẫu S_{13}
G1	1.83	3.84	5.97	
G1	1.38	3.96	6.33	
G2	3.60	5.70	8.55	← mẫu S_{23}
G2	2.94	5.07	7.95	
G2	4.02	7.20	8.94	
G3	3.99	6.09	10.02	← mẫu S_{33}
G3	3.30	5.88	9.63	
G3	4.41	6.51	10.38	
G4	3.75	5.19	11.40	← mẫu S_{43}
G4	3.63	5.37	9.66	
G4	3.57	5.55	10.53	

$m = 4$

$r = 3$

5.5 Phân tích ANOVA (tt.)



□ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

Bước 1. Tính bình phương mỗi quan sát trong mẫu $S_{ij} \rightarrow S^{(2)}$

		B_1	B_2	\dots	B_n
$S^{(2)} =$	A_1	S_{11}^2	S_{12}^2	\dots	S_{1n}^2
	A_2	S_{21}^2	S_{22}^2	\dots	S_{2n}^2
	\vdots	\vdots	\vdots	\vdots	\vdots
	A_m	S_{m1}^2	S_{m2}^2	\dots	S_{mn}^2

Tổng giá trị ma trận $S^{(2)}$: $|S^{(2)}| = \sum_{i=1}^m \sum_{j=1}^n S_{ij}^{(2)}$

5.5 Phân tích ANOVA (tt.)



□ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

Bước 2. Tính tổng từng DÒNG, từng CỘT của $S \rightarrow S^+$

		B_1	B_2	\dots	B_n	A_i^+
$S^+ =$	A_1	S_{11}^+	S_{12}^+	\dots	S_{1n}^+	$A_1^+ = \sum_{j=1}^n S_{1j}^+$
	A_2	S_{21}^+	S_{22}^+	\dots	S_{2n}^+	$A_2^+ = \sum_{j=1}^n S_{2j}^+$
	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
	A_m	S_{m1}^+	S_{m2}^+	\dots	S_{mn}^+	$A_m^+ = \sum_{j=1}^n S_{mj}^+$
	B_j^+	$B_1^+ = \sum_{i=1}^m S_{i1}^+$	$B_2^+ = \sum_{i=1}^m S_{i2}^+$	\dots	$B_n^+ = \sum_{i=1}^m S_{in}^+$	$ S^+ = \sum_{i=1}^m \sum_{j=1}^n S_{ij}^+$



5.5 Phân tích ANOVA (tt.)

□ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

Bước 3. Tính các tổng bình phương.

$$CM = \frac{|S^+|^2}{rmn} \quad SS_T = |S^{(2)}| - CM$$

$$SS_A = \frac{\sum_{i=1}^m (A_i^+)^2}{rn} - CM \quad SS_B = \frac{\sum_{j=1}^n (B_j^+)^2}{rm} - CM$$

$$SS_{AB} = \frac{\sum_{i=1}^m \sum_{j=1}^n S_{ij}^2}{r} - CM - SS_A - SS_B$$

$$SS_E = SS_T - SS_A - SS_B - SS_{AB}$$



5.5 Phân tích ANOVA (tt.)

□ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

Bước 4. Lập bảng Two-way ANOVA.

Source	SS	df	MS	F
Rows(A)	SS_A	$df_A = (m-1)$	$MS_A = \frac{SS_A}{df_A}$	$F_A = \frac{MS_A}{MS_E}$
Columns(B)	SS_B	$df_B = (n-1)$	$MS_B = \frac{SS_B}{df_B}$	$F_B = \frac{MS_B}{MS_E}$
Interaction(AB)	SS_{AB}	$df_{AB} = (m-1)(n-1)$	$MS_{AB} = \frac{SS_{AB}}{df_{AB}}$	$F_{AB} = \frac{MS_{AB}}{MS_E}$
Error	SS_E	$df_E = mn(r-1)$	$MS_E = \frac{SS_E}{df_E}$	



5.5 Phân tích ANOVA (tt.)

❑ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

VD:

A	B1	B2	B3	Số lớp A (m)	
A	1.53	4.08	6.69	6	
A	1.83	3.84	5.97		Cỡ mẫu r
A	1.38	3.96	6.33	3	
D	3.60	5.70	8.55	Số lớp B (n)	3
B	2.94	5.07	7.95	rab	54
B	4.02	7.20	8.94		
C	3.99	6.09	10.02		
C	3.30	5.88	9.63		
C	4.41	6.51	10.38		
D	3.75	5.19	11.40		
D	3.63	5.37	9.66		
D	3.57	5.55	10.53		
E	1.71	3.60	6.87		
E	2.01	5.10	6.93		
E	2.04	6.99	6.84		
F	3.96	5.25	9.84		
F	4.77	6.28	9.87		
F	4.65	5.07	10.08		



5.5 Phân tích ANOVA (tt.)

❑ Kiểm định trung bình của NHIỀU tổng thể (Two-way ANOVA)

Bước 1. Tính bình phương mỗi quan sát (Y²)

A	B1	B2	B3	
A	2.34	16.65	44.76	
A	3.35	14.75	35.64	
A	1.90	15.68	40.07	
B	12.96	32.49	73.10	
B	8.64	25.70	63.20	
B	16.16	51.84	79.92	
C	15.92	37.09	100.40	
C	10.89	34.57	92.74	
C	19.45	42.38	107.74	
D	14.06	26.94	129.96	
D	13.18	28.84	93.32	
D	12.74	30.80	110.88	
E	2.92	12.96	47.20	
E	4.04	26.01	48.02	
E	4.16	48.86	46.79	
F	15.68	27.56	96.83	
F	22.75	27.88	97.42	
F	21.62	25.70	101.61	2139.08

5.5 Phân tích ANOVA (tt.)



- Kiểm định trung bình của NHIỀU tổng thể (*Two-way ANOVA*)

Bước 2. Tính tổng theo dòng và cột

A	B1	B2	B3	Ai (Total)
A	4.74	11.88	18.99	35.61
B	10.56	17.97	25.44	53.97
C	11.70	18.48	30.03	60.21
D	10.95	16.11	31.59	58.65
E	5.76	15.69	20.64	42.09
F	13.38	15.60	29.79	58.77
Bj (Total)	57.09	95.73	156.48	309.30

Bước 3. Tính các tổng biến thiên

CM 1771.6017

SS(T) 367.4733

5.5 Phân tích ANOVA (tt.)



- Kiểm định trung bình của NHIỀU tổng thể (*Two-way ANOVA*)

Bước 4. Lập bảng ANOVA

Source	SS	df	MS	F statistic
Rows (A)	58.5517	5	11.7103	32.7486
Columns (B)	278.9256	2	139.4628	390.0149
AB	17.1230	10	1.7123	4.7885
Error (Residuals)	12.8730	36	0.3576	



5. Hypothesis Testing (tt.)



❑ Kiểm chứng/xác thực chéo (*cross-validation*)

- Train/Test split: $D = \text{Training_set} \cup \text{Test_set}$

- k-fold

Chia D thành k tập con: $(D_i \cap D_j = \emptyset) \quad |D_i| \approx |D_j|$

k = i: $\text{Test_set} = D_i \quad \text{Training_set} = (D - D_i)$

Error: $\varepsilon_i = f(\text{Training_set}, \text{Test_set}, \theta)$

$$\varepsilon = (\varepsilon_1 + \varepsilon_2 + \dots + \varepsilon_k) / k$$

Tài liệu tham khảo



Anderson et al., *Statistics for Business and Economics*, Cengage, 2016.

Nguyễn Văn Tuấn, *Các bài giảng (youtube)*, 10/2019.