

Phân Tích ANOVA (Analysis of Variance)

Khái Quát

- So sánh giá trị trung bình từ hai mẫu.
- So sánh phương sai từ hai mẫu.
- Phân tích ANOVA một chiều và hai chiều.

So Sánh Giá Trị Trung Bình Từ Hai Mẫu

Ví Dụ:

- Chúng ta có nhiều quần thể con (sub-populations) khác nhau và muốn so sánh trung bình của chúng.
- Ví dụ:
 - So sánh hiệu quả của một loại thuốc với nam và nữ.
 - So sánh chiều cao của nam thanh niên thành thị và nông thôn.
 - So sánh tỷ suất lợi nhuận giữa hai nhóm doanh nghiệp niêm yết và không niêm yết.
- Thu thập hai mẫu từ hai quần thể tiến hành phân tích ANOVA (analysis of variance).

Giả Định

- $\mathbf{X} = \{X_1, X_2, \dots, X_n\}$ là mẫu ngẫu nhiên có cùng **phân phối chuẩn** với giá trị trung bình μ_1 và phương sai σ_1^2 .
- $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_m\}$ là mẫu ngẫu nhiên có cùng **phân phối chuẩn** với giá trị trung bình μ_2 và phương sai σ_2^2 .
- Các biến ngẫu nhiên X_i và Y_j độc lập với nhau.
- $\mathbf{X}$ và $\mathbf{Y}$ có phương sai bằng nhau $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Giả Thuyết H_0 và H_1

- Giả thuyết hai phía:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

- Giả thuyết một phía:

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

hoặc:

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Biến Kiểm Định Thống Kê

- Với:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad \bar{Y} = \frac{1}{m} \sum_{i=1}^m Y_i$$

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2, \quad S_Y^2 = \frac{1}{m-1} \sum_{i=1}^m (Y_i - \bar{Y})^2$$

- Ta có:

$$t = \frac{(n+m-2)^{\frac{1}{2}}(\bar{X} - \bar{Y})}{\left(\frac{1}{n} + \frac{1}{m}\right)^{\frac{1}{2}} ((n-1)S_X^2 + (m-1)S_Y^2)^{1/2}} \sim t(m+n-2)$$

Ví Dụ:

$$t = \frac{(n + m - 2)^{\frac{1}{2}}(\bar{X} - \bar{Y})}{\left(\frac{1}{n} + \frac{1}{m}\right)^{\frac{1}{2}} \left((n - 1)S_X^2 + (m - 1)S_Y^2\right)^{1/2}} \sim t(m + n - 2)$$

- Đánh giá ảnh hưởng hai loại phân bón (Bình Điền và Phú Mỹ) với năng suất lúa, nhà nghiên cứu nông nghiệp thử nghiệm trên nhiều mẫu đất ngẫu nhiên giống nhau. Từ dữ liệu, nhà nghiên cứu tính ra được:

Phân bón Bình Điền: $\bar{X}_n = 52, S_X^2 = 4.2, n = 15$

Phân bón Phú Mỹ: $\bar{Y}_m = 51, S_Y^2 = 3.4, m = 18$

- Hãy kiểm định giả thuyết xem có sự khác biệt về năng suất lúa khi sử dụng hai loại phân bón?

Khoảng Bác Bỏ

Giả thuyết H_1	Khoảng bác bỏ với giá trị kiểm định α
$H_1: \mu_1 \neq \mu_2$	$t \geq t_{\alpha/2, m+n-2}$ hoặc $t \leq -t_{\alpha/2, m+n-2}$
$H_1: \mu_1 < \mu_2$	$t \leq -t_{\alpha, m+n-2}$
$H_1: \mu_1 > \mu_2$	$t \geq t_{\alpha, m+n-2}$

- Ghi chú: chúng ta có thể khái quát giả thuyết H_0 với giá trị t dùng để kiểm định tương ứng.

$$H_0: \mu_1 - \mu_2 = \Delta$$

$$t = \frac{(n + m - 2)^{\frac{1}{2}}(\bar{X} - \bar{Y} - \Delta)}{\left(\frac{1}{n} + \frac{1}{m}\right)^{\frac{1}{2}} ((n - 1)S_X^2 + (m - 1)S_Y^2)^{1/2}}$$

So Sánh Giá Trị Phương Sai Từ Hai Mẫu

Giả Định

- $X_1, X_2, \dots, X_n$ là mẫu ngẫu nhiên có cùng phân phối chuẩn $N(\mu_1, \sigma_1^2)$. $Y_1, Y_2, \dots, Y_m$ là mẫu ngẫu nhiên có cùng phân phối chuẩn $N(\mu_2, \sigma_2^2)$
- Các biến ngẫu nhiên X_i và Y_j độc lập với nhau từng đôi một.
- Chúng ta có thể chứng minh:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n-1, m-1)$$

Kiểm Định F (F-test)

- Giả thuyết hai phía H_0 và H_1 :

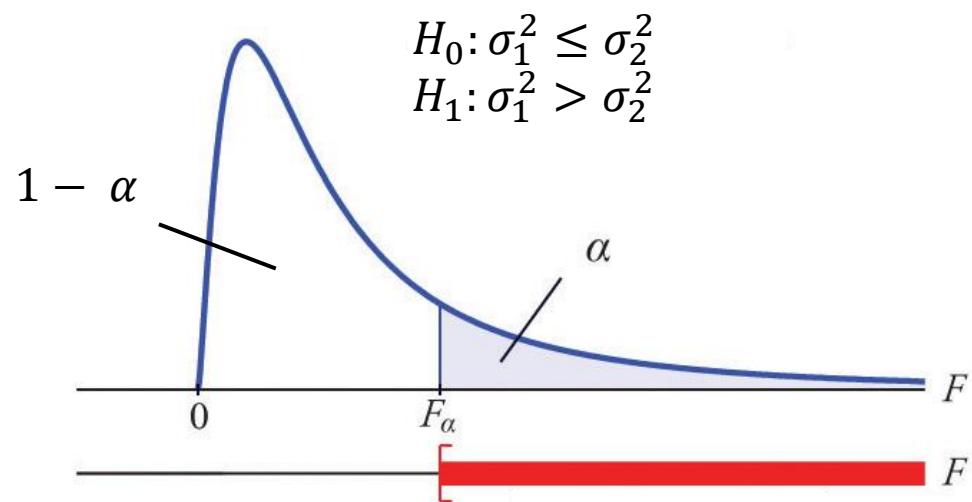
$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

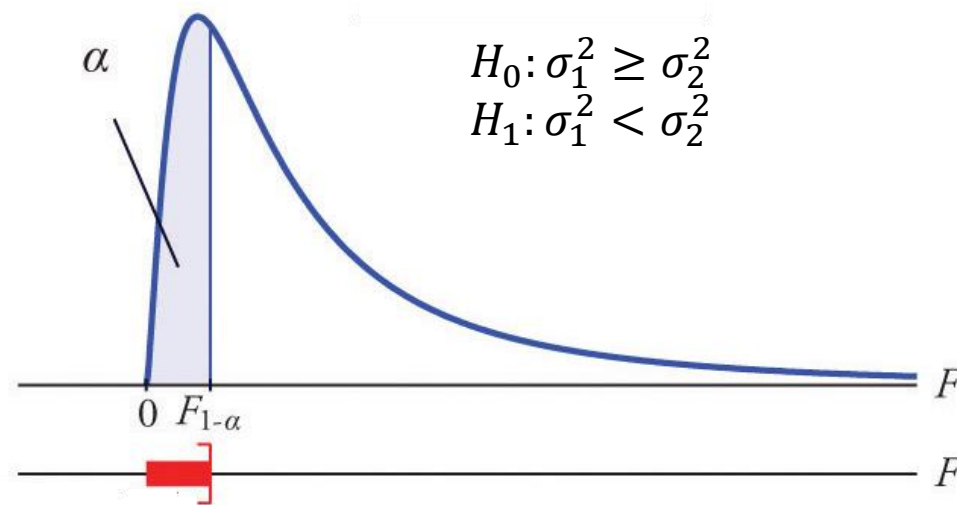
- Sử dụng phương sai mẫu, giá trị kiểm định:

$$F = \frac{S_1^2}{S_2^2} \sim F(n-1, m-1)$$

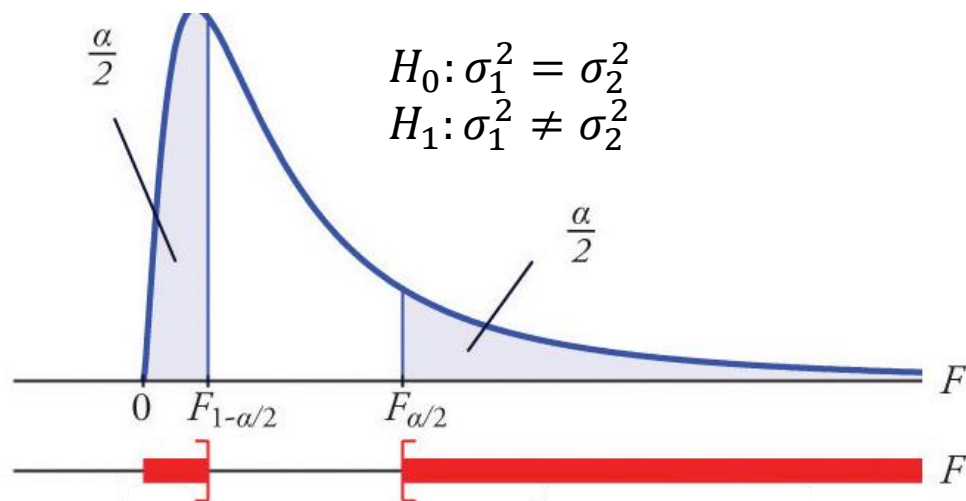
Giả thuyết H_1	Khoảng bác bỏ với giá trị kiểm định α
$H_1: \sigma_1^2 \neq \sigma_2^2$	$F_{\text{kiểm định}} \geq F_{\alpha/2, n-1, m-1}$ hoặc $F_{\text{kiểm định}} \leq F_{1-\alpha/2, n-1, m-1}$
$H_1: \sigma_1^2 < \sigma_2^2$	$F_{\text{kiểm định}} \leq F_{1-\alpha, n-1, m-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	$F_{\text{kiểm định}} \geq F_{\alpha, n-1, m-1}$



Khoảng bác bỏ: $F_{\text{kiểm định}} \geq F_{\alpha, n-1, m-1}$



Khoảng bác bỏ: $F_{\text{kiểm định}} \leq F_{1-\alpha, n-1, m-1}$



Khoảng bác bỏ: $F_{\text{kiểm định}} \geq F_{\alpha/2, n-1, m-1}$ hoặc $F_{\text{kiểm định}} \leq F_{1-\alpha/2, n-1, m-1}$

Cách Tra Cứu Phân Phối F

- Sử dụng bảng.
- Sử dụng hai hàm sau đây cho phân phối để tính giá trị tới hạn hoặc p -value.

`F.DIST()`

`F.INV()`

Ví Dụ:

- Ví dụ: Hai mẫu có hai giá trị phương sai mẫu là $S_1^2 = 9.2$, $S_2^2 = 4.3$ với cỡ mẫu là $n = 25$ và $m = 20$. Hãy kiểm định xem phương sai của phân phối tổng thể này có bằng nhau hay không với mức $\alpha = 0.05$.

Chú Ý:

- Giả định phân phối chuẩn là rất quan trọng để kiểm định cho kết quả tin cậy.
- Nếu nghi ngờ mẫu không đến từ phân phối chuẩn, chúng ta có thể kiểm định tính chuẩn của phân phối (tuần tiếp theo).

Phân Tích ANOVA Một Chiều

So Sánh Giá Trị Trung Bình Nhiều Mẫu (1)

- Giả sử chúng ta có k nhóm biến ngẫu nhiên độc lập $X_1, X_2, \dots, X_k$ có **phân phối chuẩn**, với k giá trị kì vọng khác nhau $\mu_1, \mu_2, \dots, \mu_k$ và phương sai bằng nhau σ^2 .
- Mỗi X_i ($i = 1, 2, \dots, k$) nhóm có cỡ mẫu $n_1, n_2, \dots, n_k$.
 - Ví dụ: $X_1 = \{X_{11}, X_{12}, \dots, X_{1n_1}\}$
- Các tham số $\mu_1, \mu_2, \dots, \mu_k$ và σ^2 đều là ẩn số.

So Sánh Giá Trị Trung Bình Nhiều Mẫu (2)

- Phân tích ANOVA một chiều (one-way ANOVA) có thể được thực hiện dựa vào giả thuyết H_0 và H_1 như sau:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1: H_0 \text{ không đúng}$$

Ví dụ: Hiện thực hóa cho mẫu ngẫu nhiên chiều cao giữa 3 tỉnh (cm)

- Mẫu này gồm:

$$X_1 = \{X_{1;1}, X_{1;2}, \dots, X_{1;20}\}$$

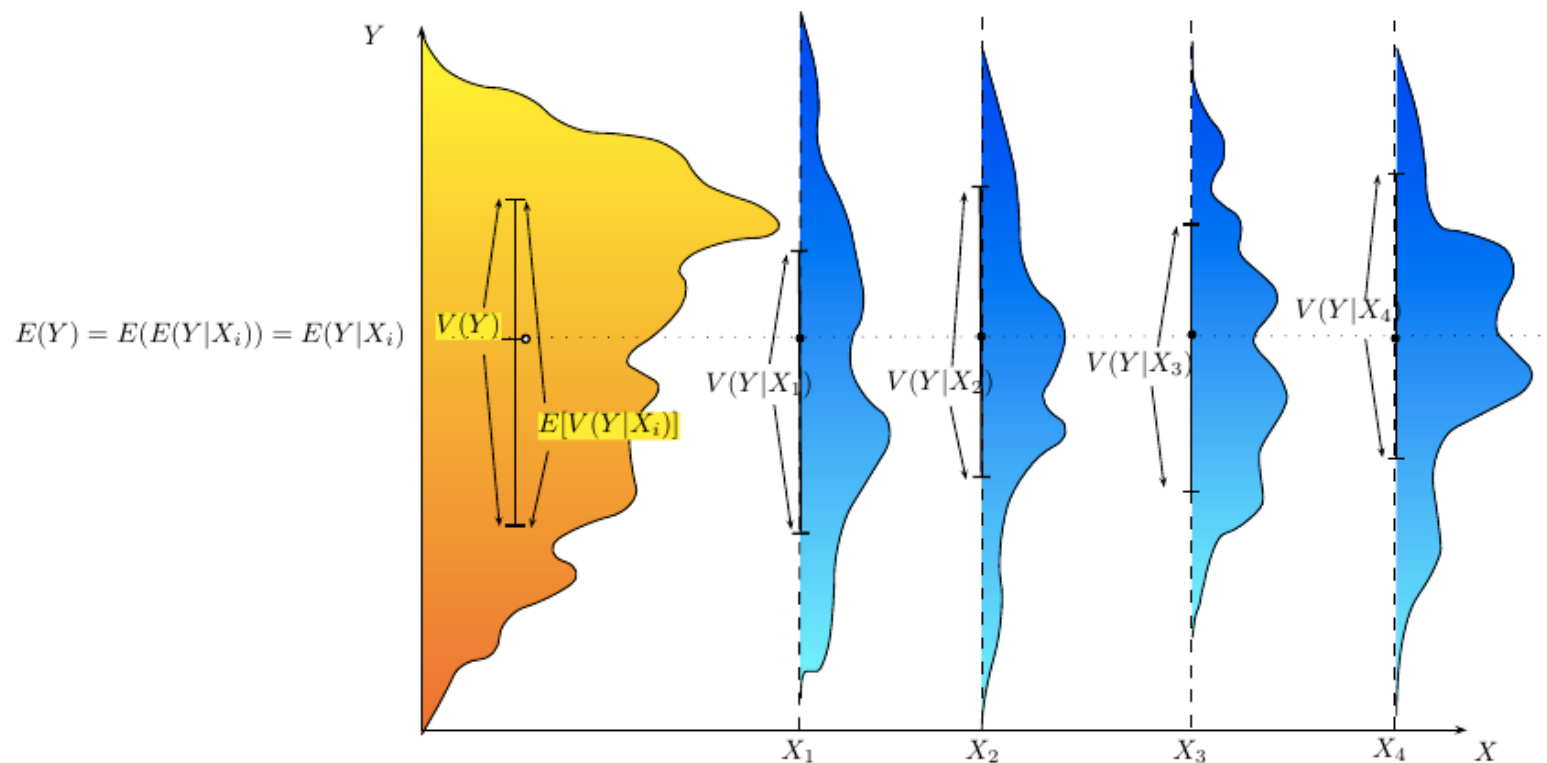
$$X_2 = \{X_{2;1}, X_{2;2}, \dots, X_{2;15}\}$$

$$X_3 = \{X_{3;1}, X_{3;2}, \dots, X_{3;18}\}$$

- Cỡ mẫu lần lượt là 20, 15 và 18.

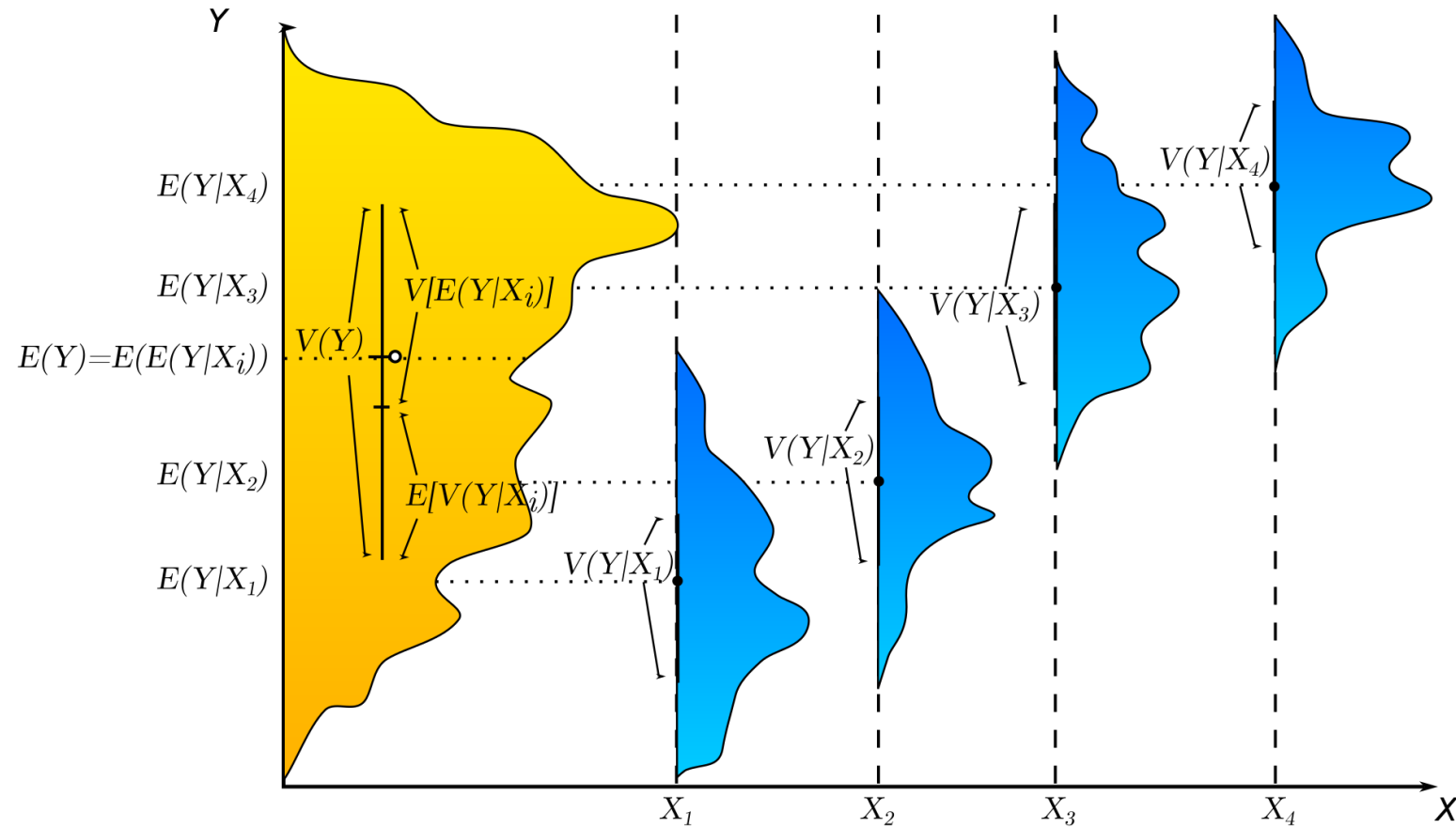
stt	TP HCM	TT Huế	Điện Biên
1	164.5	163.1	171.5
2	168.9	173.5	157.2
3	158.6	172	163.5
4	158.4	165.1	164.5
5	162	167	165.9
6	165.3	176.2	164.2
7	171.9	173.4	159.7
8	169.8	159	169.3
9	160.9	166.7	163.2
10	170.2	180.6	165
11	179.7	177.1	156.3
12	178.5	170.3	161.5
13	168.2	170.6	160.6
14	162	168.4	160.3
15	171	172.1	166.7
16	166.6		168.3
17	173.1		164
18	175.4		164.2
19	168.7		
20	181.5		

Biến Đồng Nhất



Nguồn: https://en.wikipedia.org/wiki/Analysis_of_variance

Biến Không Đồng Nhất Giá Trị μ_i



Nguồn: https://en.wikipedia.org/wiki/Analysis_of_variance

Tổng Bình Phương (Sum of Squares)

- Tổng bình phương giữa k nhóm (among-group variation):

$$SS_A = \sum_{i=1}^k n_i (\bar{X}_k - \bar{\bar{X}})^2$$

- Tổng bình phương nội nhóm (within-group variation):

$$SS_W = \sum_{j=1}^k \sum_{i=1}^{n_i} (X_{ij} - \bar{X}_j)^2$$

- Tổng toàn bộ bình phương (total variation):

$$SS_T = \sum_{j=1}^k \sum_{i=1}^{n_i} (X_{ij} - \bar{\bar{X}})^2$$

- Chúng ta có thể chứng minh được:

$$SS_T = SS_A + SS_W$$

với:

- $\bar{\bar{X}} = \frac{\sum_{j=1}^k \sum_{i=1}^{n_i} X_{ij}}{n}$: trung bình lớn.
- $\bar{X}_j$: giá trị trung bình của biến X_j ($j = 1, 2, \dots, k$).
- X_{ij} : giá trị i trong biến X_j ($j = 1, 2, \dots, k$).
- $n = n_1 + n_2 + \dots + n_k$ tổng số biến ngẫu nhiên.
- k số lượng nhóm biến ngẫu nhiên.

Tổng Bình Phương (Sum of Squares) (1)

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_1: H_0 \text{ không đúng}$$

$$F_{\text{kiểm định}} \sim F(k - 1, n - k)$$

	Bậc tự do phân phối χ^2 của SS/σ^2	Tổng bình phương	Trung bình của tổng bình phương (Mean Square)	Giá trị $F_{\text{kiểm định}}$
Giữa các nhóm (Among groups)	$k - 1$	SS_A	$MSA = \frac{SS_A}{k - 1}$	$F_{\text{kiểm định}} = \frac{MSA}{MSW}$
Nội nhóm (Within groups)	$n - k$	SS_W	$MSW = \frac{SS_W}{n - k}$	
Tổng toàn bộ	$n - 1$	SS_T		

Giá Trị Kiểm Định và Khoảng Bác Bỏ

- Giả thuyết

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

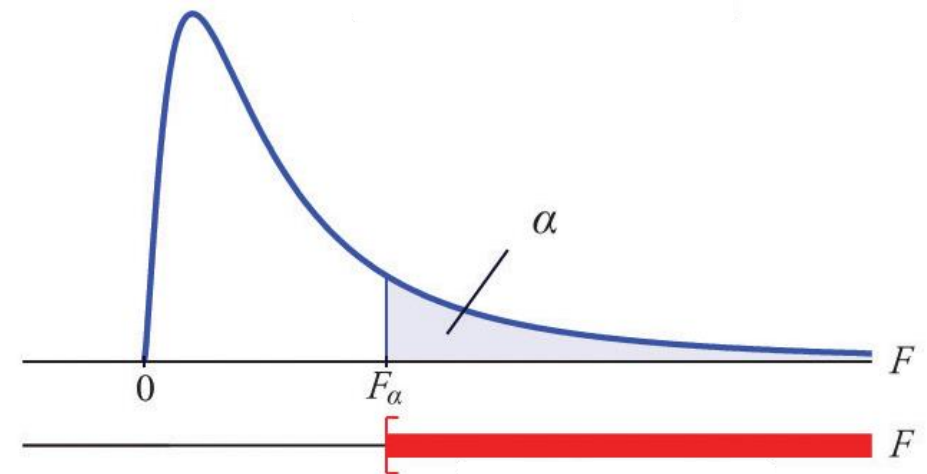
$$H_1: H_0 \text{ không đúng}$$

- Giá trị kiểm định:

$$F_{\text{kiểm định}} \sim F(k - 1, n - k)$$

- Khoảng bác bỏ với mức độ tin cậy α :

$$F_{\text{kiểm định}} \geq F_{\alpha, k-1, n-k}$$



Phân Tích ANOVA Một Chiều

- Mô hình ANOVA một chiều có thể được viết dưới dạng:

$$Y_{i,j} = \mu_j + \varepsilon_{i,j}$$

với

- $i = 1, 2, 3, \dots, n_j$ là số thứ tự các quan sát trong cùng 1 nhóm.
 - $j = 1, 2, 3, \dots, k$ là số thứ tự đến các nhóm bị ảnh hưởng (treatment groups).
 - μ_j là trung bình cho từng nhóm bị ảnh hưởng
 - $\varepsilon \sim N(0, \sigma^2)$ là sai số có phân phối chuẩn và độc lập với nhau cho tất cả i và j .
- Giá trị trung bình và phương sai của $Y_{i,j}$:

$$E(Y_{i,j}) = \mu_j, \quad \text{Var}(Y_{i,j}) = \sigma^2$$

Ví dụ: Mẫu cho chiều cao giữa 3 tỉnh (cm)

stt	TP HCM	TT Huế	Điện Biên
1	164.5	163.1	171.5
2	168.9	173.5	157.2
3	158.6	172	163.5
4	158.4	165.1	164.5
5	162	167	165.9
6	165.3	176.2	164.2
7	171.9	173.4	159.7
8	169.8	159	169.3
9	160.9	166.7	163.2
10	170.2	180.6	165
11	179.7	177.1	156.3
12	178.5	170.3	161.5
13	168.2	170.6	160.6
14	162	168.4	160.3
15	171	172.1	166.7
16	166.6		168.3
17	173.1		164
18	175.4		164.2
19	168.7		
20	181.5		

- Biến ngẫu nhiên cho từng giá trị quan sát viết chung là $Y_{i,j}$
- Có 3 nhóm (bị ảnh hưởng) nên $j = 3$.
- Trung bình từng nhóm là $E(Y_{i,1}) = \mu_1$, $E(Y_{i,2}) = \mu_2$, $E(Y_{i,3}) = \mu_3$.
- Ba nhóm đều có sai số xung quanh giá trị trung bình như nhau, độ lệch chuẩn σ .
- Ví dụ: giá trị hiện thực hóa (mẫu quan sát): $Y_{3,1} = 158.6$, $Y_{6,2} = 176.2$

Phân Tích ANOVA Hai Chiều (two-way ANOVA)

- Mô hình ANOVA hai chiều có thể được viết dưới dạng:

$$Y_{i,j} = \mu_j + \gamma_j + (\mu\gamma)_{i,j} + \varepsilon_{i,j,k}$$

- Giả định về biến ngẫu nhiên tương tự như phân tích ANOVA một chiều.

So sánh chiều cao của trẻ 2 tuổi			
		Nhóm B	
		Thành thị	Nông thôn
Nhóm A	Sử dụng sữa mẹ	Các nhóm giả thuyết: - Không có sự khác biệt giữa nhóm hai nhóm sử dụng sữa. - Không có sự khác biệt giữa nhóm nông thôn/ thành thị. - Không có sự tương tác giữa các nhóm A và nhóm B.	
	Sử dụng sữa công thức		

Giới Thiệu Về Mô Hình Hồi Quy

- Mô hình hồi quy đơn giản một biến phụ thuộc (simple regression model):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Kết quả hồi quy là trung bình có điều kiện:

$$E(Y_i | X_i = x_i) = \beta_0 + \beta_1 X_i$$

- Khi biến X_i là biến định tính, kiểm định từ phân tích ANOVA tương tự kiểm định như mô hình hồi quy đơn giản.

- Ví dụ: biến giả (dummy variables) nam/ nữ.

- Biến định tính còn có thể khái quát thành một hay nhiều biến định lượng:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots \beta_k X_{ki} + \varepsilon_i$$

So Sánh Mô Hình ANOVA và Hồi Quy Đơn Giản

- Mô hình ANOVA một chiều

$$Y_{i,j} = \mu_j + \varepsilon_{i,j}$$

- Cỡ mẫu có thể khác nhau cho từng j , nên phải viết thành $Y_{i,j}$.
- μ_j là một (ẩn) số cố định, không phải là biến ngẫu nhiên.
- $E(Y_{i,j}) = \mu_j$ hoặc $E(Y_j) = \mu_j$

- Mô hình hồi quy đơn giản:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- Cỡ mẫu chung là n : $i = 1, 2, 3, \dots, n$
- β_0, β_1 là tham (ẩn) số cần ước lượng.
- Y_i và X_i là (thường) biến liên tục nên không cần phân thành các nhóm j khác nhau.
- $E(Y_i | X_i = x_i) = \beta_0 + \beta_1 X_i$ hoặc
 $E(Y | X = x) = \beta_0 + \beta_1 X$

Chú Ý Về Thuật Ngữ Và Kí Hiệu

- Phần kinh tế lượng (econometrics), thuật ngữ được sử dụng như sau:

$$y = \beta_0 + \beta_1 x + u$$

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- y, x là viết chung cho hiện tượng cần nghiên cứu (lương, thu nhập, số năm đi học,...); và cũng hiểu như là biến ngẫu nhiên để thực hiện nghiên cứu định lượng dùng lý thuyết xác suất, thống kê.
- y_i, x_i là biến ngẫu nhiên, được hiểu là một quan sát thứ i trong mẫu n quan sát cũng được thu thập từ hiện tượng y, x .
- Viết y_i, x_i thường để tiện cho việc giải thích công thức tính toán.
- Giả định căn bản nhất là từng cặp (y_i, x_i) là các cặp biến I.I.D.