Chương 7: PHÂN TÍCH PHƯƠNG SAI (ANOVA)

Phân tích phương sai là một mô hình dùng để xem xét sự biến động của một biến ngẫu nhiên định lượng X chịu tác động trực tiếp của một hay nhiều yếu tố nguyên nhân (định tính).

- *Dạng 1: Phân tích phương sai 1 yếu tố*
  (One-Way Analysis of Variance)
- *Dạng 2: Phân tích phương sai 2 yếu tố không lặp (chỉ BTL)*
- *Dạng 3: Phân tích phương sai 2 yếu tố có lặp (chỉ BTL).*

Trong mô hình phân tích phương sai 1 yếu tố, chúng ta kiểm định so sánh trung bình của biến ngẫu nhiên X ở những tổng thể (còn gọi là nhóm) khác nhau dựa vào các mẫu quan sát lấy từ những tổng thể này. Các tổng thể được phân biệt bởi các mức độ khác nhau của yếu tố đang xem xét.

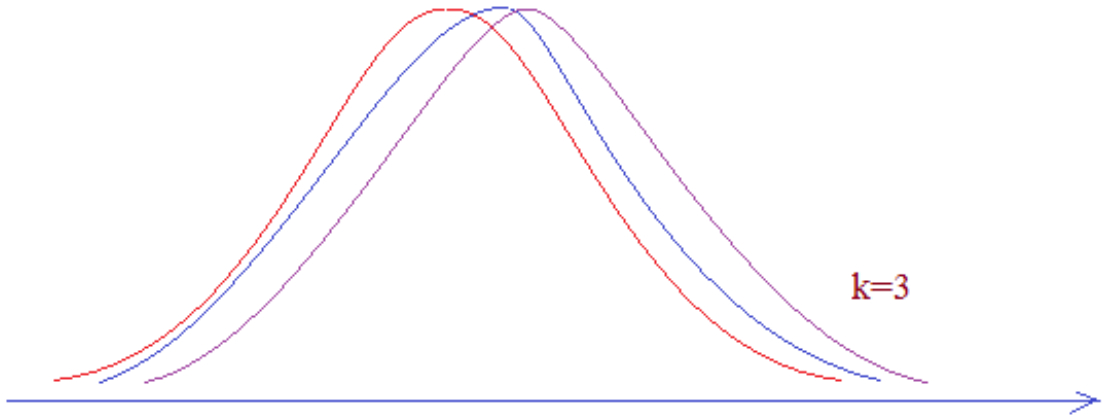
7.1 Giả thiết của bài toán ANOVA MỘT YẾU TỐ:

(Điều kiện bài toán hay là giả thiết mô hình)

- Các tổng thể có phân phối chuẩn $N(\mu_i ; \sigma_i^2)$;

  i = 1; 2; …; k. k là số tổng thể (thông thường $k \geq 3$).

- Phương sai các tổng thể bằng nhau $(\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2)$.
- Các mẫu quan sát (từ các tổng thể) được lấy độc lập.



7.2 Các bước thực hiện bài toán:

* ĐẶT GIẢ THIẾT KIỂM ĐỊNH:

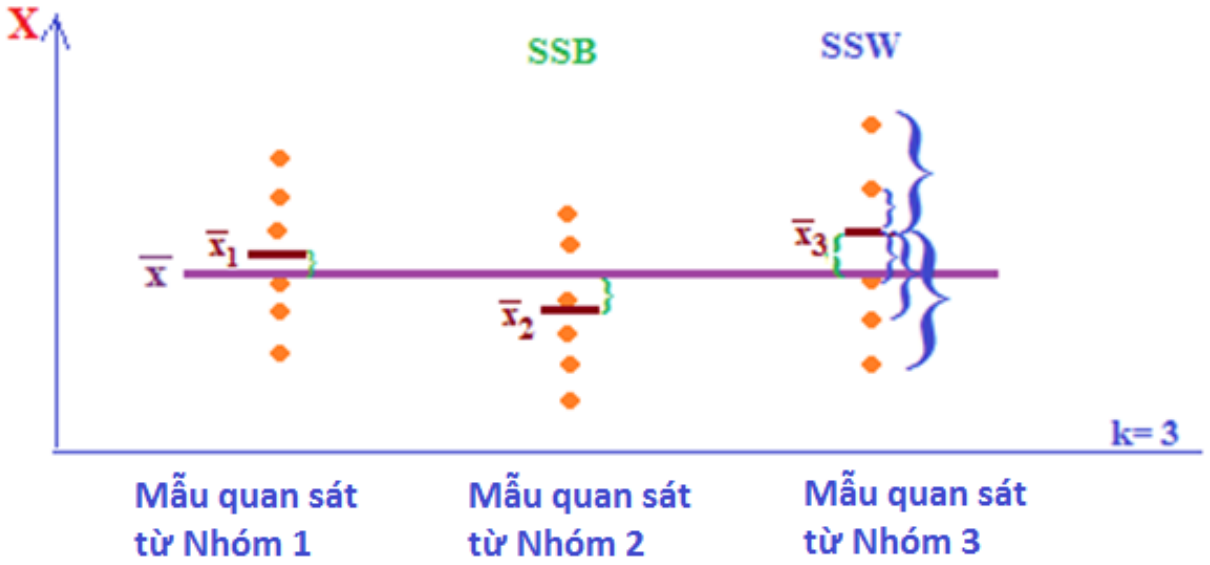
Giả thiết không H_0 : $\mu_1 = \mu_2 = \ldots = \mu_k$.

Giả thiết đối H_1: $\exists\ \mu_i \neq \mu_j$; với $i \neq j$

* TÍNH GIÁ TRỊ KIỂM ĐỊNH THỐNG KÊ:

	Nhóm 1	Nhóm 2	...	Nhóm k
Các mẫ quan sát	x_{11} x_{21} ... $x_{n1;1}$	x_{12} x_{22} ... $x_{n2;2}$	...	x_{1k} x_{2k} ... $x_{n1;k}$
Kích thước từng mẫu	n_1	n_2		n_k
Trung bình từng mẫu	$\overline{x_1}$	$\overline{x_2}$		$\overline{x_n}$
Kích thước mẫu gộp	$N = n_1 + n_2 + \ldots + n_k$			
Trung bình mẫu gộp	$\overline{x} = \sum_j \sum_i x_{ij} / N = \left(n_1 \overline{x_1} + n_2 \overline{x_2} + \ldots + n_k \overline{x_k}\right) / N$			

SSB (SSTr)	Sum of squares between group	$SSB = \sum_{j=1}^{k} n_j \times \left(\overline{x}_j - \overline{x}\right)^2$
SSW (SSE)	Sum of squares within group	$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(x_{ij} - \overline{x}_j\right)^2$
SST	Total sum of squares	$SST = \sum_{j=1}^{k} \sum_{i=1}^{n_i} \left(x_{ij} - \overline{x}\right)^2$

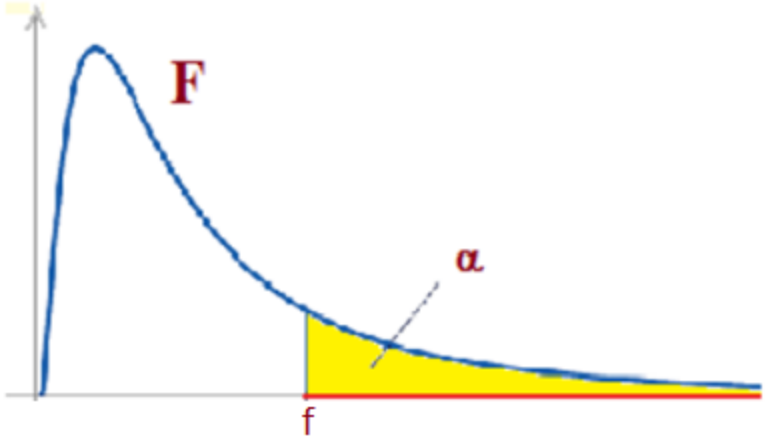


	Mẫu 1	Mẫu 2	....	Mẫu k	Mẫu gộp
Các quan sát	(x_{i1})	(x_{i2})		(x_{ik})	
Kích thước mẫu	n_1	n_2	...	n_k	$N = n_1 + n_2 + ... + n_k$ (1)
Trung bình mẫu của từng nhóm	$\overline{x}_1$	$\overline{x}_2$	...	$\overline{x}_k$	$\overline{x}$ (2)
Tổng bình phương chênh lệch giữa các nhóm	$n_1 \times (\overline{x}_1 - \overline{x})^2$ SSB_1	$n_2 \times (\overline{x}_2 - \overline{x})^2$ SSB_2	...	$n_k \times (\overline{x}_k - \overline{x})^2$ SSB_K	$SSB = SSB_1 + ... + SSB_K$ (3)
Tổng bình phương chênh lệch trong nội bộ nhóm	$\sum_{i=1}^{n_1} (x_{i1} - \overline{x}_1)^2$ SSW_1	Bấm $(n-1).s^2$ hoặc $n.\sigma_X^2$	...	$\sum_{i=1}^{n_k} (x_{ik} - \overline{x}_k)^2$ SSW_K	$SSW = SSW_1 + ... + SSW_K$ (4)
$\sum_{i=1}^{n_i} x_{ij}^2$	$\sum_{i=1}^{n_1} x_{i1}^2$ T_1	Bấm . ..SUM - $\sum x^2$	...	$\sum_{i=1}^{n_k} x_{ik}^2$ T_K	$T = \sum_{i=1}^{n_i}\sum_{j=1}^{k} x_{ij}^2$ $T = T_1 + ... + T_K$ (5)
Tổng bình phương chênh lệch toàn bộ	$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_i} (x_{ij} - \overline{x})^2$ Bấm $T - N \times (\overline{x})^2$ hoặc $SST = SSB + SSW$ (Định lý) (6)				
	* Nếu nhập đủ N giá trị của mẫu gộp thì SST được đọc từ MTBT bằng cách bấm $(n-1).s^2$				

Source of Variation	Tổng Bình phương chênh lệch	Bậc tự do	Phương sai (Trung bình BPCL)	Tiêu chuẩn kiểm định F
Between Groups	**SSB** (SSTr) **(3)**	k -1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups	**SSW** (SSE) **(4)**	N - k	$MSW = \frac{SSW}{N-k}$	
Total	**SST** **(6)**	N –1		

* MIỀN BÁC BỎ:

RR = (f_α (k-1; N - k); +∞)



Nhận xét:

- **SSB** (hay SSTr)**:** Phần biến thiên của giá trị X do các mức độ của yếu tố đang xem xét tạo ra.
- **SSW** (hay SSE)**:** Phần biến thiên của giá trị X do các yếu tố nào đó không được đề cập đến tạo ra.
- **SST:** Tổng các biến thiên của X do tất cả các yếu tố tạo ra.

*** HỆ SỐ XÁC ĐỊNH:** $R^2 = \frac{SSB}{SST} \times 100\%$

Hệ số xác định R^2 của mô hình Phân tích phương sai được sử dụng để đo mức độ ảnh hưởng của yếu tố được xem xét trong mô hình đối với sự biến động của các giá trị của biến ngẫu nhiên X quanh giá trị trung bình của nó.

R^2 càng lớn thì mô hình càng gọi là thích hợp.

7.3 Phân tích sâu Anova một yếu tố:

Khi kết luận cho bài toán Anova, có 2 trường hợp xảy ra:

+ Chưa bác bỏ được giả thiết H_0, hay là chưa có bằng chứng về sự khác biệt của các trung bình.

+ Bác bỏ H_0, chấp nhận $H_1 \Rightarrow$ Trung bình của các nhóm không bằng nhau (hay là sự khác biệt có ý nghĩa thống kê). Nói chung thì chúng ta không biết được sự khác biệt đó là từ một hay từ những nhóm nào. Do đó ta có thể muốn phân tích thêm: nhóm nào đó có trung bình lớn hơn, bằng, hay nhỏ hơn so với những nhóm khác?

Có nhiều phương pháp đưa đến kết quả mong muốn. Chúng ta còn gọi đó là các phương pháp so sánh bội (Multiple comparison methods).

Phương pháp được trình bày ở đây là Fisher's LSD (Least Significant Difference).

- Dùng LSD test: So sánh lần lượt tất cả các cặp trung bình của 2 nhóm khác nhau với các giả thiết tương ứng:

  $H_0: \mu_i = \mu_j$; $H_1: \mu_i \neq \mu_j$; $i \neq j$.

  Tính $LSD = t_{\frac{\alpha}{2}}(N\text{-}k) \times \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

  Giả thiết H_0 bị bác bỏ khi : $\left|\overline{x_i} - \overline{x_j}\right| > LSD$

  (Bài giảng này không đề cập chi tiết hơn)

- Dùng các khoảng tin cậy (LSD confidence intervals) để ước lượng các chênh lệch của trung bình 2 nhóm bất kỳ. Từ đó tìm ra các cặp nhóm có trung bình khác biệt.

* Khoảng ước lượng LSD với độ tin cậy 1-α cho độ chênh lệch ($\mu_i - \mu_j$) là:

$$\left(\overline{x_i} - \overline{x_j}\right) \pm t_{\frac{\alpha}{2}}(N-k) \times \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

$i; j \in \{1; 2; \ldots; k\}. i \neq j .$

* Số khoảng tin cậy cần tìm: C^2_k .

* Nếu khoảng tin cậy không chứa số 0 thì ta nói có sự khác biệt giữa hai giá trị trung bình μ_i và μ_j có ý nghĩa thống kê. Cụ thể hơn, nếu khoảng tin cậy chỉ gồm các số dương, xem như $\mu_i > \mu_j$. Ngược lại, ta nói $\mu_i < \mu_j$ nếu khoảng tin cậy nằm toàn bộ ở phần giá trị âm trên trục số thực.

* Nếu khoảng tin cậy chứa số 0 thì ta không kết luận được sự khác biệt giữa μ_i và μ_j .

Ví dụ 26:

Khi theo dõi tác động của các điều kiện ngoại cảnh đến sự sinh trưởng của 1 loại cây non, người ta gieo trồng cùng 1 loại hạt giống trong 3 điều kiện ngoại cảnh A, B, C khác nhau và thu được số liệu mẫu sau:

Điều kiện ngoại cảnh	Chiều cao của cây (cm)					
A	48	51	57	62	59	55
B	46	42	45	50	47	51
C	44	55	53	56	54	

Hãy dùng phương pháp Anova để so sánh chiều cao trung bình của các cây con trong 3 điều kiện ngoại cảnh trên với mức ý nghĩa 5%. (Lưu ý bổ sung thêm các giả thiết cần có để thực hiện được yêu cầu bài toán). Tính hệ số xác định R^2.

Hướng dẫn:

* Gọi μ_1; μ_2 ; μ_3 lần lượt là chiều cao trung bình của các cây con được trồng trong các điều kiện ngoại cảnh A; B; C.

Giả thiết kiểm định H_0: $\mu_1 = \mu_2 = \mu_3$

Giả thiết đối H_1: $\exists\ \mu_i \neq \mu_j$ với $i \neq j$

Các giả thiết cần có: *Xem điều kiện bài toán .*

* Miền bác bỏ RR = ($f_{0.05}(2; 14); +\infty$) = (3.7389; $+\infty$)

Tra bảng Fisher $\alpha = 0.05$; *bậc tử: $n_1 = 2$;*

bậc mẫu: $n_2 = 14$

* Tính tiêu chuẩn kiểm định:

Xem 2 bảng phía sau.

	A	B	C	Mẫu gộp
	48 62 51 59 57 55	**46 50 42 47 45 51**	**44 56 55 54 53**	
(1) Kích thước mẫu	$n_1 = 6$	$n_2 = 6$	$n_3 = 5$	N = 17
(2) Trung bình mẫu	$\bar{x}_1 = 55.3333$	$\bar{x}_2 = 46.8333$	$\bar{x}_3 = 52.4$	$\bar{x} = 51.4706$
(3) **SSB** (SSTr)	$SSB_1 = 89.5248$			**222.8686**
(4) **SSW** (SSE)	$SSW_1 = 133.3333$			**281.3667**
(5) Σx_i^2	$T_1 = 18504$			
(6) **SST**	**504.2353**			

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	222.869	2	111.434	5.54465	0.01685	3.73889
Within Groups	281.367	14	20.0976			
Total	504.235	16				

$f_{0.05}(2;14)$

* Kết luận:

Do tiêu chuẩn kiểm định $F_{qs} = 5.5447 \in RR$ nên bác bỏ H_0; chấp nhận $H_1 \Rightarrow$ Chiều cao trung bình của cây non sinh trưởng ở các điều kiện A; B; C là không bằng nhau.

Cách nói khác: Chiều cao (trung bình) của cây phụ thuộc vào điều kiện ngoại cảnh.

* Hệ số xác định: $R^2 = (SSB/SST)*100\% \approx 44.2\ \%$

Yếu tố điều kiện ngoại cảnh giải thích 44.2% sự chênh lệch về chiều cao của các cây non trong 3 vùng.

So sánh bội qua các khoảng tin cậy 95% LSD:

Khoảng tin cậy 95% dành cho	$\overline{x_i} - \overline{x_j}$	$t_{\frac{0.05}{2}}(17-3)$	$SE = \sqrt{MSW\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$	Cận dưới	Cận trên
$\mu_1 - \mu_2$	8.5	2.1448	2.5883	2.9487	14.0514
$\mu_1 - \mu_3$	2.9333	2.1448	2.7146	-2.8890	8.7556
$\mu_2 - \mu_3$	-5.5667	2.1448	2.7146	-11.389	0.2556

- Hỗ trợ tra bảng Student trong Excel: =TINV.2T(0.05; 14)
- Trong 3 khoảng tin cậy trên thì khoảng tin cậy cho μ_1-μ_2 chỉ gồm các số dương. Có thể nói rằng chiều cao trung bình của các cây con ở điều kiện ngoại cảnh A là lớn hơn hẳn so với chiều cao trung bình của các cây con ở điều kiện ngoại cảnh B.

Ví dụ 27:

Khi đo mức độ bụi trong không khí tại 3 khu vực trong thành phố, người ta được số liệu sau (đơn vị mg/m^3):

Số thứ tự quan sát	Khu vực 1	Khu vực 2	Khu vực 3
1	0,54	0,48	0,56
2	0,60	0,49	0,62
3	0,72	0,55	0,60
4	0,67	0,62	
5	0,83	0,57	

Với mức ý nghĩa 5%, có thể coi như mức độ bụi trung bình ở các khu vực trên là như nhau không? Lưu ý bổ sung thêm các giả thiết cần có để thực hiện được yêu cầu bài toán.

Tìm hệ số R^2 và nêu ý nghĩa.

	Khu vực 1	Khu vực 2	Khu vực 3	Mẫu gộp
	0.54 0.67 0.6 0.83 0.72	0.48 0.62 0.49 0.57 0.55	0.56 0.62 0.6	
(1) Kích thước mẫu	$n_1 = 5$	$n_2 = 5$	$n_3 = 3$	N = 13
(2) Trung bình mẫu	$\overline{x_1}$ =	$\overline{x_2}$ =	$\overline{x_3}$ =	$\overline{x}$ =
(3) (SSG) **SSB**				**0.0427**
(4) **SSW**				**0.0652**
(5) Σx_i^2				
(6) **SST**				

Ví dụ 28:

Dưới đây là mẫu thống kê về số buổi tham gia công tác xã hội trong năm của sinh viên các khóa. Hãy sử dụng mô hình Anova để xét xem thời gian tham gia CTXH của sinh viên có bị ảnh hưởng bởi tiến độ học trong trường của sinh viên hay không, kết luận với mức ý nghĩa 1%.

Năm 1	Năm 2	Năm 3	Năm 4
8	7	8	6
7.5	9	8	7
6	8.5	5	4
7	7	7	6
5	6	8	5

Hãy tìm khoảng tin cậy 99% cho số buổi tham gia CTXH chênh lệch trung bình giữa sinh viên năm 2 và năm 4.

Bài tập tham khảo:

Từ giáo trình Xác suất – thống kê & Phân tích số liệu; tài liệu (3).

Năm 1992, trong một nghiên cứu về ảnh hưởng của trục cuốn ép lên cường độ chịu nén của các loại thùng carton tiêu chuẩn RSC được sản xuất, Burgess đã tiến hành đo cường độ chịu nén của bốn loại thùng carton khác nhau. Dưới đây là dữ liệu thu được

Hộp	Lực nén					
1	655.5	788.3	734.3	721.4	679.1	699.4
2	789.2	772.5	786.9	686.1	732.1	774.8
3	737.1	639.0	727.1	671.7	717.2	727.1
4	535.1	628.7	542.4	559.0	586.9	520.0

Hãy so sánh cường độ chịu nén của bốn loại thùng carton với mức ý nghĩa $\alpha = 0.01$.

Một công ty dược phẩm so sánh ba công thức thuốc giảm đau cho chứng đau nửa đầu. Trong thí nghiệm, 27 người tình nguyện được chia ngẫu nhiên thành ba nhóm, mỗi nhóm tương ứng với một công thức thuốc. Những người này sẽ uống thuốc khi bị đau nửa đầu và và phản hồi mức đau đầu sau khi uống thuốc (10 là mức đau nhất). Dưới đây là dữ liệu thu được

Drug A	4	5	4	3	2	4	3	4	4
Drug B	6	8	4	5	4	6	5	8	6
Drug C	6	7	6	6	7	5	6	5	5

Hãy so sánh tác dụng của ba công thức thuốc trên với mức ý nghĩa 0.05.

Một nhà sản xuất các túi giấy dùng để đựng hoa quả muốn tăng độ chịu kéo của sản phẩm. Các kỹ sư tin rằng độ chịu kéo phụ thuộc vào tỉ lệ gỗ cứng có trong bột giấy và tỉ lệ này có giá trị từ 5% đến 20%. Nhóm kỹ sư phụ trách nghiên cứu này đã quyết định thử nghiệm ở bốn mức tỉ lệ: 5%, 10%, 15%, và 20%. Họ kiểm tra sáu mẫu ở mỗi mức tỉ lệ. Tất cả 24 mẫu được kiểm tra độ chịu kéo với cùng một thiết bị và theo thứ tự ngẫu nhiên. Dưới đây là dữ liệu thu được

Tỉ lệ gỗ cứng	Độ chịu kéo						Tổng	Trung bình
5%	7	8	15	11	9	10	60	10.00
10%	12	17	13	18	19	15	94	15.67
15%	14	18	19	17	16	18	102	17.00
20%	19	25	22	23	18	20	127	21.17

Hãy tính các tổng bình phương, trung bình bình phương và giá trị thống kê F.

Bài giải. Ta có $I = 4, J = 6, \overline{x} = 15.96$. Suy ra

$$SSTr = 6[(10.00 - 15.96)^2 + (15.67 - 15.96)^2 + (17.00 - 15.96)^2 + (21.17 - 15.96)^2] = 382.7917$$

$$SSE = (7 - 10.00)^2 + (8 - 10.00)^2 + \cdots + (20 - 21.17)^2 = 130.1667$$

$$SST = (7 - 15.96)^2 + (8 - 15.96)^2 + \cdots + (20 - 15.96)^2 = 512.9583$$

Các đại lượng này có bậc tự do lần lượt là

$$\mathrm{df}(SSTr) = 4 - 1 = 3, \quad \mathrm{df}(SSE) = (4)(6 - 1) = 20, \quad \mathrm{df}(SST) = 24 - 1 = 23.$$

Do đó

$$MSTr = \frac{382.7917}{3} = 127.5972 \quad \text{và} \quad MSE = \frac{SSE}{20} = 6.5083$$

Cuối cùng

$$F = \frac{MSTr}{MSE} = \frac{127.5972}{6.5083} = 19.605.$$

Từ giáo trình Xác suất – thống kê & Phân tích số liệu; tài liệu (3).