PHÂN TÍCH PHƯƠNG SAI (ANOVA)

Mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các số trung bình của các mẫu quan sát từ các nhóm này và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các số trung bình này.

Trong nghiên cứu, phân tích phương sai được dùng như là một công cụ để xem xét ảnh hưởng của một hay một số yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng).

PHÂN TÍCH PHƯƠNG SAI

Ví dụ:

- Nghiên cứu ảnh hưởng của phương pháp đánh giá của giáo viên đến kết quả học tập của sinh viên.
- Nghiên cứu ảnh hưởng của bậc thợ tới năng suất lao động.
- Nghiên cứu ảnh hưởng của phương pháp bán hàng, trình độ (kinh nghiệm) của nhân viên bán hàng đến doanh số

PHÂN TÍCH PHƯƠNG SAI

- **Phân tích phương sai một yếu tố**
- **Phân tích phương sai hai yếu tố**

Phân tích phương sai một yếu tố

Phân tích phương sai một yếu tố là phân tích ảnh hưởng của một yếu tố nguyên nhân (dạng biến định tính định tính) đến một yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

Phân tích phương sai một yếu tố

Giả sử cần so sánh số trung bình của k tổng thể độc lập. Ta lấy k mẫu có số quan sát là n_1, n_2… n_k; tuân theo phân phối chuẩn. Trung bình của các tổng thể được ký hiệu là μ_1; μ_2 ….μ_k thì mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định giả thuyết như sau:

H_o: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_1: Tồn tại ít nhất 1 cặp có $\mu_i \neq \mu_j$; $i \neq j$

Phân tích phương sai một yếu tố

Để kiểm định ta đưa ra 3 giả thiết sau:

1) Mỗi mẫu tuân theo phân phối chuẩn $N(\mu, \sigma 2)$
2) Các phương sai tổng thể bằng nhau
3) Ta lấy k mẫu độc lập từ k tổng thể. Mỗi mẫu được quan sát n_j lần.

Các bước tiến hành:

Bước 1: Tính các trung bình mẫu và trung bình chung của k mẫu

- Ta lập bảng tính toán như sau:

TT	**k mẫu quan sát**				
	1	2	3	…	k
1	X_{11}	X_{12}	X_{13}		X_{1k}
2	X_{21}	X_{22}	X_{23}		X_{2k}
3	X_{31}	X_{32}	X_{33}		X_{3k}
…					
…					
j	X_{j1}	X_{j2}	X_{j3}		X_{jk}
Trung bình	$\bar{x}$	$\bar{x}$	$\bar{x}$		$\bar{x}$

Bước 1: Tính các trung bình mẫu và trung bình chung của k mẫu

Trung bình mẫu $\bar{x}_1$ $\bar{x}_2$ $\bar{x}_k$ được tính theo công thức:

$$\bar{x}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} (i = 1, 2, ..k)$$

Trung bình chung của k mẫu được tính theo công thức:

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i \bar{x}_i}{\sum_{i=1}^{k} n_i} (i = 1, 2, ..k)$$

Bước 2: Tính các tổng độ lệch bình phương

Tổng các độ lệch bình phương trong nội bộ nhóm (nội bộ từng mẫu - SSW) được tính theo công thức sau:

Nhóm 1	Nhóm 2	Nhóm k
$SS_1 = \sum_{j=1}^{n_1} (X_{j1} - \bar{x}_1)^2$	$SS_2 = \sum_{j=1}^{n_2} (X_{j2} - \bar{x}_2)^2$	$SS_k = \sum_{j=1}^{n_k} (X_{jk} - \bar{x}_k)^2$
$SSW = SS_1 + SS_2 + \ldots + SS_k = \sum_{i=1}^{k} \sum_{ij=11}^{n_i} (X_{ij} - \bar{x}_i)^2$		

Bước 2: Tính các tổng độ lệch bình phương

Tổng các độ lệch bình phương giữa các nhóm(SSB)

$$SSB = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})^2$$

Tổng các độ lệch bình phương của toàn bộ tổng thể(SST)

$$SST = SSW + SSB = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \overline{x})^2$$

Bước 3: Tính các phương sai (phương sai của nội bộ nhóm và phương sai giữa các nhóm)

Ta ký hiệu k là số nhóm (mẫu); n là tổng số quan sát của các nhóm thì các phương sai được tính theo công thức sau:

$MSW = \frac{SSW}{n - k}$	$MSB = \frac{SSB}{k - 1}$

MSW: Là phương sai nội bộ nhóm

SSB: Là phương sai giữa các nhóm

Bước 4: Kiểm định giả thuyết

- Tính tiêu chuẩn kiểm định F (F thực nghiệm)

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

- $F > F((k-1; n-k); \alpha)$

  Ta bác bỏ giả thuyết H_0 cho rằng trị trung bình của k tổng thể bằng nhau

Bước 4: Kiểm định giả thuyết

- Tìm F lý thuyết (F tiêu chuẩn = F (k-1; n-k; α)):
- F lý thuyết là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và ; n-k bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.
- F lý thuyết có thể tra qua hàm FINV(α, k-1, n-1) trong EXCEL.
- Nếu F thực nghiệm > F lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể không bằng nhau

Bảng phân tích phương sai 1 yếu tố khi sử dụng máy tính (phần mềm EXCEL hoặcSPSS) tóm tắt như sau:

- Bảng gốc bằng tiếng Anh

Source of variation	*Sum of squares (SS)*	*Degree of freedom (df)*	*Mean squares (MS)*	*F- ratio*
Between - groups	SSB	(k-1)	MSB	$F = \frac{MSB}{MSW}$
Within - groups	SSW	(n-k)	MSW	
Total	SST	(n-1)		

Bảng phân tích phương sai 1 yếu tố khi sử dụng máy tính (phần mềm EXCEL hoặcSPSS) tóm tắt như sau:

Bảng phân tích phương sai tổng quát dịch ra tiếng việt – ANOVA

Nguồn biến động	Tổng độ lệch bình phương (SS)	Bậc tự do (df)	Phương sai (MS)	F- Tỷ số
Giữa các mẫu	SSB	(k-1)	MSB	$F = \frac{MSB}{MSW}$
Trong nội bộ các mẫu	SSW	(n-k)	MSW	
Tổng số	SST	(n-1)		

Ví dụ 1:

Có tài liệu về cách cho điểm môn Nguyên lý thống kê của 3 giáo viên như sau (điểm tối đa là 100). Hãy cho biết cách chấm điểm của 3 giáo viên có sai khác nhau không?

TT	A	B	C
1	82	74	79
2	86	82	79
3	79	78	77
4	83	75	78
5	85	76	82
6	84	77	79

Ví dụ 1:

Đặt giả thuyết

Ho: Cách chấm điểm của 3 giáo viên không sai khác nhau

H1: Cách chấm điểm của 3 giáo viên có sai khác nhau

Ho: $\mu 1 = \mu 2 = \mu 3$;

H1: Tồn tại ít nhất 1 cặp có $\mu_i \neq \mu_j$; $i \neq j$

- Từ kết quả lấy mẫu của 3 nhóm ta tính các độ lệch bình phương thể hiện qua bảng sau:

					SS_1	SS_2	SS_3	
TT	A	B	C	Chung (Xbq)	$(X_{1j} - \overline{x1})^2$	$(X_{2j} - \overline{x2})^2$	$(X_{3j} - \overline{x3})^2$	**Cộng**
1	82	74	79		1,36	9,00	0,00	
2	86	82	79		8,03	25,00	0,00	
3	79	78	77		17,36	1,00	4,00	
4	83	75	78		0,03	4,00	1,00	
5	85	76	82		3,36	1,00	9,00	
6	84	77	79		0,69	0,00	0,00	
Trung bình	$\overline{x1}$ = **83,17**	$\overline{x2}$ = **77,00**	$\overline{x3}$ = **79,00**	$\overline{x}$ = **79,72**				
P.sai (δ_i^2)	6,17	8,00	2,80	11,98				
Cộng					30,83	40,00	14,00	SSW=84,83
$(\overline{xi} - \overline{x})^2 nj$	71,185	44,463	3,130					SSB=118,78

Ví dụ 1:

$SSW = SS_1 + SS_2 + SS_3 = 84,83$

$$SSB = \sum_{i=1}^{k} n_i (\overline{x_i} - \overline{x})^2 = 118,78$$

- Tính các phương sai:

$$MSW = \frac{SSW}{n - k} = \frac{84,83}{15} = 5,66$$

$$MSB = \frac{SSB}{k - 1} = \frac{118,78}{3 - 1} = 59,39$$

Ví dụ 1:

- Tính F thực nghiệm:

$$F = \frac{MSB}{MSW} = \frac{59,39}{5,66} = 10,5$$

- Tra bảng F lý thuyết (F (0.05; 2; 15)) = 3,68

  So sánh F thực nghiệm với F lý thuyết ta thấy: F thực nghiệm > F lý thuyết

  Bác bỏ Ho, nghĩa là cách cho điểm của 3 giáo viên có khác nhau.

Sử dụng kết quả của máy tính, phần mềm Excel chúng ta cũng có kết quả tương tự (bảng sau)

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	6	499	**83,17**	6,17
B	6	462	**77,00**	8,0
C	6	474	**79,00**	2,8

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	118,78	2	59,39	10,50	0,00	3,68
Within Groups	84,83	15	5,66			
Total	203,61	17				

Phân tích phương sai 2 yếu tố

Phân tích phương sai 2 yếu tố nhằm xem xét cùng lúc hai yếu tố nguyên nhân (dướidạng dữ liệu định tính) ảnh hưởng đến yếu tố kết quả (dưới dạng dữ liệu định lượng) đang nghiên cứu.

Ví dụ: Nghiên cứu ảnh hưởng của loại chất đốt và loại lò sấy đến tỷ lệ vải loại 1 sấy khô.

Phân tích phương sai 2 yếu tố giúp chúng ta đưa thêm yếu tố nguyên nhân vào phân tích làm cho kết quả nghiên cứu càng có giá trị.

Phân tích phương sai 2 yếu tố

Giả sử ta nghiên cứu ảnh hưởng của 2 yếu tố nguyên nhân định tính đến một yếu tố kết quả định lượng nào đó.

Ta lấy mẫu không lặp lại, sau đó các đơn vị mẫu của yếu tố nguyên nhân thứ nhất sắp xếp thành K nhóm (cột), các đơn vị mẫu của yếu tố nguyên nhân thứ hai sắp xếp thành H khối (hàng). Như vậy, ta có bảng kết hợp 2 yếu tố nguyên nhân gồm K cột và H hàng và (K x H) ô dữ liệu. Tổng số mẫu quan sát là $n = (K \times H)$.

Phân tích phương sai 2 yếu tố
Dạng tổng quát

Hàng (Khối)	Cột (nhóm)			
	1	2	…	K
1	X_{11}	X_{21}		X_{K1}
2	X_{12}	X_{22}		X_{K2}
…				
H	X_{1K}	X_{2K}		X_{KH}

Các bước tiến hành

Để kiểm định ta đưa ra 2 giả thiết sau:

1) Mỗi mẫu tuân theo phân phối chuẩn $N(\mu, \sigma^2)$

2) Ta lấy K mẫu độc lập từ K tổng thể, H mẫu độc lập từ H tổng thể. Mỗi mẫu được quan sát 1 lần không lặp.

Bước 1: Tính các số trung bình

Trung bình riêng của từng nhóm (K cột)	**Trung bình riêng của từng khối (H hàng)**
$\overline{X}_i = \frac{\sum_{j=1}^{H} X_{ij}}{H}$ $i = 1, 2 \ldots K$	$\overline{X}_j = \frac{\sum_{i=1}^{K} X_{ij}}{K}$ $j = 1, 2 \ldots H$

Trung bình chung của toàn bộ mẫu quan sát

$$\overline{X} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{H} X_{ij}}{n} = \frac{\sum_{i=1}^{K} \overline{X}_i}{K} = \frac{\sum_{j=1}^{H} \overline{X}_j}{H}$$

Bước 2. Tính tổng các độ lệch bình phương

Diễn giải	Công thức
1. Tổng các độ lệch bình phương chung (SST) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của tất cả các yếu tố*	$SST = \sum_{i=1}^{K} \sum_{j=1}^{H} (X_{ij} - \overline{X})^2$
2. Tổng các độ lệch bình phương giữa các nhóm (SSK) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ nhất (xếp theo cột)*	$SSK = H \sum_{i=1}^{K} (\overline{X_i} - \overline{X})^2$

Bước 2. Tính tổng các độ lệch bình phương

Diễn giải	Công thức
3.Tổng các độ lệch bình phương giữa các nhóm (SSH) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân thứ hai (xếp theo hàng)*	$SSH = K\sum_{j=1}^{H}(\overline{X_j} - \overline{X})^2$
4. Tổng các độ lệch bình phương phần dư (ERROR) *Phản ánh biến động của yếu tố kết quả do ảnh hưởng của yếu tố nguyên nhân khác không nghiên cứu*	$SSE = SST - SSK - SSH$

Bước 3. Tính các phương sai

Diễn giải	**Công thức**
1. Phương sai giữa các nhóm (cột) (MSK)	$MSK = \frac{SSK}{K - 1}$
2. Phương sai giữa các khối (hàng) (MSH)	$MSH = \frac{SSH}{H - 1}$
3. Phương sai phần dư (MSE)	$MSE = \frac{SSE}{(K - 1)(H - 1)}$

Bước 4. Kiểm định giả thuyết
Tính tiêu chuẩn kiểm định F (F thực nghiệm)

$F_1 = \frac{MSK}{MSE}$	Trong đó: MSK là phương sai giữa các nhóm (cột) MSE là phương sai phần dư F1 dùng kiểm định cho yếu tố nguyên nhân thứ nhất
$F_2 = \frac{MSH}{MSE}$	Trong đó: MSH là phương sai giữa các khối (hàng) MSE là phương sai phần dư F_2 dùng kiểm định cho yếu tố nguyên nhân thứ hai

Bước 4. Kiểm định giả thuyết

Tìm F lý thuyết cho 2 yếu tố nguyên nhân

- Yếu tố nguyên nhân thứ nhất:

  F tiêu chuẩn = F (k-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với k-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

  F lý thuyết có thể tra qua hàm FINV(α, k-1, (k-1)(h-1)) trong EXCEL

Bước 4. Kiểm định giả thuyết

Tìm F lý thuyết cho 2 yếu tố nguyên nhân

- Yếu tố nguyên nhân thứ hai:

  F tiêu chuẩn = F (h-1; (k-1)(h-1), α) là giá trị giới hạn tra từ bảng phân phối F với h-1 bậc tự do của phương sai ở tử số và (k-1)(h-1) bậc tự do của phương sai ở mẫu số với mức ý nghĩa α.

  F lý thuyết có thể tra qua hàm FINV(α, h-1, (k-1)(h-1)) trong EXCEL.

Bước 4. Kiểm định giả thuyết

Nếu F1 thực nghiệm > F1 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể nhóm (cột) không bằng nhau.

Nếu F2 thực nghiệm > F2 lý thuyết, bác bỏ Ho, nghĩa là các số trung bình của k tổng thể khối (hàng) không bằng nhau.

Bảng phân tích phương sai 2 yếu tố khi sử dụng máy tính (phần mềm EXCEL hoặcSPSS) tóm tắt như sau:

Bảng gốc bằng tiếng Anh

Source of variation	*Sum of squares(SS)*	*Degree of freedom(df)*	*Mean squares(MS)*	*F- ratio*
Rows	SSH	(h-1)	MSH	F_1
Columns	SSK	(k-1))	MSK	F_2
Error	SSE	(k-1))(h-1)	MSE	
Total	SST	(n-1)		

Bảng phân tích phương sai tổng quát dịch ra tiếng Việt – ANOVA

Nguồn biến động	*Tổng độ lệch bình phương (SS)*	*Bậc tự do (df)*	*Phương sai (MS)*	*F- Tỷ số*
Giữa các hàng	SSH	(h-1)	MSH	F_1
Giữa các cột	SSK	(k -1)	MSK	F_2
Phần dư	SSE	(k -1) (h-1)	MSE	
Tổng số	SST	(n-1)		

Ví dụ 2:

Có tài liệu về giá bán đậu tương của các tỉnh qua 2 năm như sau (đồng/kg)

Tỉnh	**2003**	**2004**
Sơn La	4440	4247,7
Hà Tây	4850	4294,3
Đắc Lắc	4400	4284,3
Đồng Nai	4500	4314,3

Yêu cầu: Sử dụng kết quả phân tích phương sai so sánh giá bán đậu tương qua 2năm và giữa 4 tỉnh?

Sử dụng phân tích phương sai (ANOVA) 2 yếu tố lấy mẫu không lặp trong Excel cho kết quả sau:

ANOVA: Two-Factor Without Replication

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Sơn La	2	8687,7	4343,85	18489,645
Hà Tây	2	9144,3	4572,15	154401,245
Đắc Lắc	2	8684,3	4342,15	6693,245
Đồng Nai	2	8814,3	4407,15	17242,245
2003	4	18190,0	4547,50	42358,333
2004	4	17140,6	4285,15	778,89

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F thực nghiệm*	*P-value*	*F crit*
Rows	70240,34	3	23413,45	1,1871	0,4456	9,2766
Columns	137655	1	137655,04	6,9791	0,0775	10,1280
Error	59171,34	3	19723,78			
Total	**267066,7**	**7**				

Từ kết quả phân tích ANOVA ở bảng trên cho thấy:

Xét theo hàng:

So sánh giá bán đậu tương bình quân giữa các tỉnh với giả thuyết là

Ho: Giá bán trung bình đậu tương giữa các tỉnh không sai khác nhau

F thực nghiệm =1,18; F lý thuyết = 9,27.

Như vậy, F thực nghiệm < F lý thuyết, ta chấp nhận Ho với xác suất có ý nghĩa là 55, 44%.

Từ kết quả phân tích ANOVA ở bảng trên cho thấy:

Xét theo cột:

So sánh giá bán đậu tương bình quân giữa các năm với giả thuyết là

Ho: Giá bán trung bình đậu tương giữa các năm không sai khác nhau

F thực nghiệm =6,97; F lý thuyết = 10,12.

Như vậy, F thực nghiệm < F lý thuyết, ta chấp nhận Ho vớixác suất có ý nghĩa là 92,25%.