Bài 3 PHÂN TíCH PHƯƠNG SAI MỘT NHÂN TỐ

Muốn so sánh nhiều trung bình của nhiều biến chuẩn thì phải bố trí thí nghiệm, thông thường là thí nghiệm một nhân tố và hai nhân tố sau đó phân tích phương sai. Excel không đề cập đến các kiểu bố trí thí nghiệm và cũng không đề cập đến việc so sánh các trung bình sau khi phân tích phương sai mà chỉ phân tích phương sai theo 3 mô hình: Một nhân tố, hai nhân tố không lặp lại quan sát và hai nhân tố có lặp lại quan sát với số lần lặp bằng nhau.

1/ Phân tích phương sai một nhân tố.

Mô hình này dùng khi bố trí thí nghiệm một nhân tố kiểu hoàn toàn ngẫu nhiên (Completely randomized design - CRD). Mô hình toán học tương ứng là:

$$x_{ij} = m + a_i + e_{ij} \qquad i = 1, k \qquad j = 1, n_i$$

x_{ij} quan sát thứ j ở mức thứ i của nhân tố, tất cả có k mức, mức i có n_i quan sát

m - trung bình toàn bộ a_i - chênh lệch giữa trung bình của mức i với trung bình toàn bộ

e_{ij} - sai số ngẫu nhiên của lần quan sát thứ j ở mức i của nhân tố

Với giả thiết: Các e_{ij} độc lập và phân phối chuẩn N $(0,\sigma^2)$ ta có thể tiến hành việc phân tích phương sai nhằm kiểm định giả thiết H_0 : tất cả các $a_i = 0$ (hay các trung bình của các mức bằng nhauh) với đối thiết H_1 : ít nhất có một $a_i \neq 0$ (hay các trung bình của các mức không bằng nhauh).

Để cụ thể ta xét thí dụ về năng suất của 4 giống khoai (đơn vị tạ / ha). Hai giống A và B mỗi giống có 4 quan sát, 2 giống C và D mỗi giống có 6 quan sát, mỗi giống là một mức.

Giống							Số quan sát
A	160	172	144	158			4
B	294	304	303	281			4
C	260	292	267	271	260	281	6
D	253	243	261	232	257	240	6

Việc tính toán bao gồm:

Tổng số quan sát N = $\sum_{i=1}^{k} n_i$ Trung bình toàn bộ: $\bar{x} = \dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}}{n}$

Các trung bình ở các mức $\bar{x}_i = \dfrac{\sum_{j=1}^{n_i} x_{ij}}{n_i}$

Tổng bình phương toàn bộ: SST = $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$ với N -1 bậc tự do

Tổng bình phương do nhân tố T: SSA = $\sum\sum(\bar{x}_i-\bar{x})^2$ với k - 1 bậc tự do

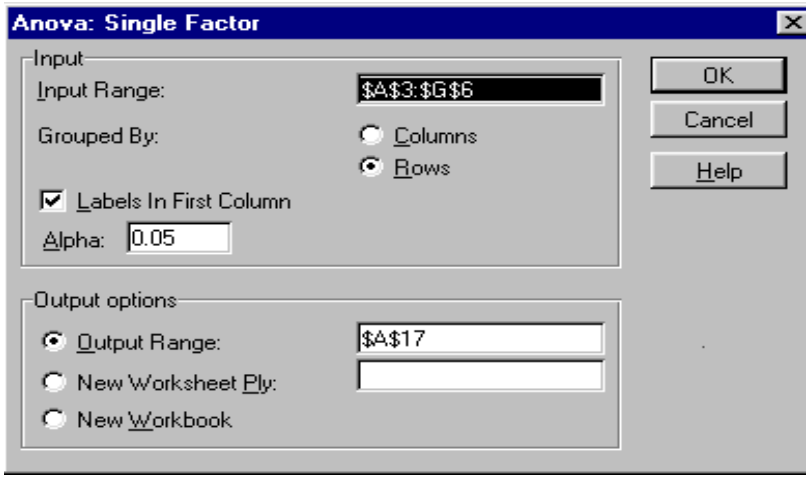
Tổng bình phương do sai số: SSE = SST - SSA với N - k bậc tự do

Sau khi tính xong tất cả các thông tin được tóm tắt vào trong một bảng gọi là bảng phân tích phương sai (ANOVA)

Nguồn	BTd	Tổng BF	BF tbình	Ftn	Flt
Nhân tố	dfa =3	44438.38	s^2a =14812.79	110.2262	3.238867
Sai số	dfe = 16	2150.167	s^2e =134.3854		
Toàn bộ	dft = 19	46588.55			

Bình phương trung bình (Mean squares) bằng tổng bình phương (Sum squares) chia cho bậc tự do (Degree of freedom) tương ứng. Giá trị F_{tn} bằng s^2a / s^2e , giá trị F_{lt} bằng giá trị tra cứu ở bảng Fisher Snedecor với mức ý nghĩa α, bậc tự do của tử số dfa và bậc tự do của mẫu số dfe, hoặc dùng hàm Finv (α,dfa,dfe) là hàm định sẵn trong Excel.

Nếu dùng Data Analysis thì vào Anova single factor



Kết quả được bảng các thống kê cơ bản sau cho từng mức của nhân tốK

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
A	4	634	158.5	131.6667
B	4	1182	295.5	113.6667
C	6	1631	271.8333	158.9667
D	6	1486	247.6667	123.8667

Tiếp theo là bảng ANOVA

Source of Variation	*SS*	*df*	*MS*	F_{tn}	*P-value*	F_{lt}
Between Groups	44438.38	3	14812.79	110.2262	6.73E-11	3.238867
Within Groups	2150.167	16	s^2e =134.3854			
Total	46588.55	19				

P- value là xác suất p ($F > F_{tn}$) để biến F có phân phối Fisher lấy giá trị lớn hơn F_{tn}

Nếu $F_{tn} > F_{lt}$ (hay P- value < 0,05) thì kết luận: Bác bỏ H_0 ở mức ý nghĩa $\alpha = 0,05$

Khi so sánh trung bình của 4 giống có thể dùng các kiểm định Student, Newman - Keuls, Duncan , Tukey, Scheffe, v. v . . .

Phương pháp kinh điển của Student, còn gọi là LSD (Least significance difference), như sau: Muốn so 2 trung bình $\bar{x}_i$ và $\bar{x}_j$ ta tính LSD = t(α, dfe) * $\sqrt{s^2_e(\frac{1}{n_i}+\frac{1}{n_j})}$,

trong đó s^2_e lấy ở trong bảng ANOVA còn n_i và n_j là số quan sát của 2 mức.

Nếu giá trị tuyệt đối của hiệu giữa 2 trung bình nhỏ hơn hay bằng LSD thì chấp nhận H_0, ngược lại thì bác bỏ H_0.

Thí dụ so giống B và C ta có hiệu 2 trung bình là 295,5 - 271,83 = 23,67

LSD = 2,12 x $\sqrt{134,3854*(\frac{1}{4}+\frac{1}{6)}}$ = 15, 863 kết luận trung bình 2 giống khác nhau

Nếu so A và B phải lấy LSD = 17.38 còn nếu so C và D phải lấy LSD = 14,19

2/ Phân tích phương sai hai nhân tố không lặp lại quan sát

Bố trí thí nghiệm với 2 nhân tố rất ít khi không lặp lại quan sát, nhưng phần này của Excel có thể dùng để phân tích thí nghiệm một nhân tố bố trí kiểu khối ngẫu nhiên đủ (Randomized complete block design), khi đó khối được coi là nhân tố thứ hai. Nhân tố chính để ở hàng, khối để ở cột, tất cả có a mức của nhân tố và b khối

Mô hình toán học như sau:

$$x_{ij} = m + a_i + b_j + e_{ij}$$

m là trung bình chungm, a_i là chênh lệch giữa trung bình ở mức i của nhân tố và trung bình chung, b_j là chênh lệch giữa trung bình của khối j với trung bình chung còn e_{ij} là sai số ngẫu nhiên với giả thiết độc lập, phân phối chuẩn N (0, σ^2).

Khi phân tích ta làm như phần trên đối với một nhân tố, tính tổng quan sát N = ab, trung bình toàn bộ $\bar{x}$, trung bình theo hàng $\bar{x}_{i.}$, trung bình theo cột $\bar{x}_{.j}$ sau đó lần lượt tính

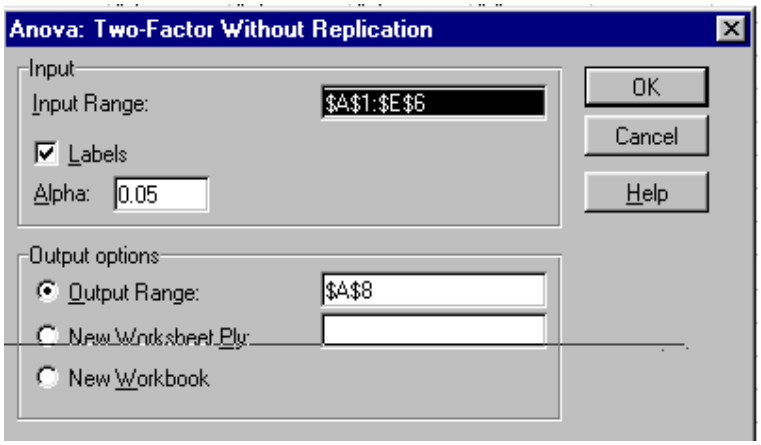
Tổng bình phương toàn bộ SST = $\sum_{i=1}^{a}\sum_{j=1}^{b}(x_{ij}-\bar{x})^2$ với N - 1 bậc tự do

Tổng bình phương do nhân tố SSA = $\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{i.}-\bar{x})^2$ với a - 1 bậc tự do

Tổng bình phương theo khối SSB = $\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{x}_{.j}-\bar{x})^2$ với b - 1 bậc tự do

Tổng bình phương do sai số SSE = SST - SSA - SSB với (a - 1)(b - 1) bậc tự do

Vào Data Analysis ta có đối thoại sau:



Nghiên cứu 5 giống, bố trí theo 4 khối
Ta có bảng số liệu sau:

	K1	K2	K3	K4
G1	47.8	46.9	45.4	44.1
G2	53.7	50.3	50.6	48
G3	46.7	42	42.4	40.7
G4	48	47	45.9	45.7
G5	41.8	40	43	41.6

Bảng tóm tắt các thống kê

	Count	Sum	Average	Variance
Giống				
G1	4	184.2	46.05	2.67
G2	4	202.6	50.65	5.483333333
G3	4	171.8	42.95	6.776666667
G4	4	186.6	46.65	1.136666667
G5	4	166.4	41.6	1.52
Khối				
K1	5	238	47.6	17.965
K2	5	226.2	45.24	17.353
K3	5	227.3	45.46	10.508
K4	5	220.1	44.02	8.887

Bảng phân tích phương sai

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	199.312	4	49.828	30.60061	3.27E-06	3.25916
Columns	33.22	3	11.07333	6.800409	0.006249	3.4903
Error	19.54	dfe=12	s^2e = 1.628333			
Total	252.072	19				

So sánh F_{tn} và F_{lt} ta có thể kết luận về 2 kiểm định:

Kiểm định giả thiết H_0 đối với các a_i : " các a_i đều bằng 0" Đối thiết H_1: " có $a_i \neq 0$"

Kiểm định giả thiết H_0 đối với các b_j : " các b_j đều bằng 0" Đối thiết H_1: " có $b_j \neq 0$"

Nếu $F_{tn} > F_{lt}$ thì bác bỏ H_0 (hoặc Ph - value <0,05) ở mức ý nghĩa $\alpha = 0,05$

So sánh các trung bình của các mức của nhân tố được làm tương tự như phần một nhân tố, ở đây

$$LSD = t(\alpha, dfe) * \sqrt{2 \times \frac{se^2}{b}}$$

các ký hiệu dfe, s^2e lấy ở bảng Anova còn b là số khối

3/ Phân tích phương sai hai nhân tố có lặp lại quan sát

Khi bố trí thí nghiệm hai nhân tố kiểu hoàn toàn ngẫu nhiên ta sắp xếp nhân tố A có a mức ở hàng, nhân tố B có b mức ở cột, mỗi công thức (mức ai của nhân tố A kết hợp với mức bm của nhân tố B) được lặp lại r lần. Mô hình toán học như sau:

$$x_{i\,j\,k} = m + a_i + b_j + (ab)_{i\,j} + e_{i\,j\,k}$$

m là trung bình chung, a_i là chênh lệch giữa trung bình ở mức i của nhân tố A so với trung bình chung, b_j là chênh lệch giữa trung bình ở mức j của nhân tố B so với trung bình chung, $(ab)_{i\,j}$ là chênh lệch giữa trung bình của công thức (a_i, b_j) với $a_i + b_j + m$, $e_{i\,j\,k}$ là sai số độc lập, phân phối chuẩn N ($0,\sigma^2$).

Tính tổng quan sát N = abr, trung bình toàn bộ $\bar{x}$, trung bình theo hàng $\bar{x}_{i..}$, trung bình theo cột $\bar{x}_{.j.}$, trung bình theo công thức $\bar{x}_{ij.}$ sau đó lần lượt tính

Tổng bình phương toàn bộ SST = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ijk} - \bar{x})^2$ với N - 1 bậc tự do

Tổng bình phương do nhân tốA SSA = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(\bar{x}_{i..} - \bar{x})^2$ với a - 1 bậc tự do

Tổng bình phương do nhân tố B SSB = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(\bar{x}_{.j.} - \bar{x})^2$ với b - 1 bậc tự do

Tổng bình phương do tương tác SSAB = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(x_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x})^2$

với (a - 1)(b - 1) bậc tự do

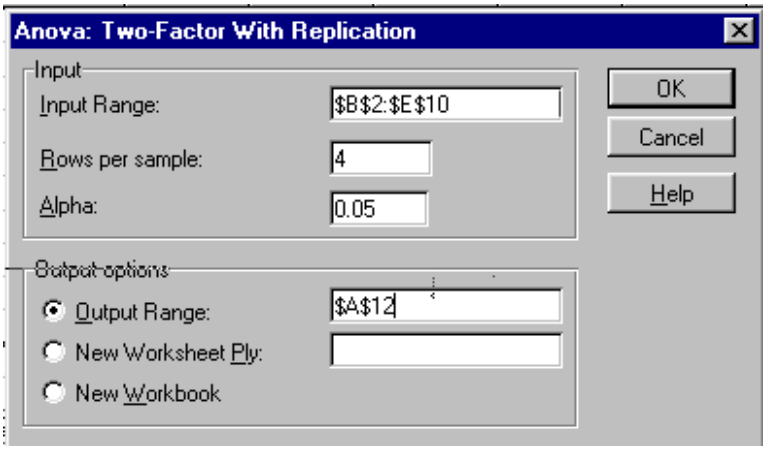
Tổng bình phương do sai số SSE = SST - SSA - SSB - SSAB với ab (r-1) bậc tự do

Toàn bộ thông tin được ghi vào trong bảng phân tích phương sai (ANOVA).

Thí dụ ta có 2 nhân tố: Đạm (2 mức) ghi ở hàng, Lân (3 mức) ghi ở cột, mỗi công thức lặp lại 4 lần. Sắp xếp số liệu như ở bảng dưới sau đó vào Data Analysis, kết qủa chúng ta được bảng các thống kê cơ bản, bảng phân tích phương sai, dựa vào đây có thể kiểm định 3 giả thiết H_0 đối với đạm, lân và tương tác, tiếp theo có thể so sánh các trung bình ứng với các mức đạm khác nhau, các trung bình ứng với các mức lân khác nhau và còn có thể so sánh các trung bình ứng với các công thức (tổ hợp đam x lân) khác nhau.

Lân

Đạm	B1	B2	B3
A1	24.1 25.8 23 27	28.4 29.7 30.1 27.4	28.7 30.4 32 27
A2	30.7 34.4 34 31	46.7 45.4 47.1 46.3	59.4 50.7 64.5 60.1



Khai báo input range phải bao gồm cả cột đầu ghi các mức đạm và hàng đầu ghi các mức lân, rows per sample là số lần lặp r

SUMMARY	B1	B2	B3	Total	
Count	4	4	4	12	Bốn dòng này tính các thống
Sum	99.9	115.6	118.1	333.6	kê cho từng công thức k, lần
Average	24.975	28.9	29.525	27.8	lượt:A1B1, A1B2,A1B3,A1B4
Variance	3.149167	1.526667	4.649167	6.967273	sau cùng là thống kê chung cho mức đạm A1
Count	4	4	4	12	Bốn dòng này tính các thống
Sum	130.1	185.5	234.7	550.3	kê cho từng công thức, lần
Average	32.525	46.375	58.675	45.85833	lượt:A2B1, A2B2,A2B3,A2B4
Variance	3.7825	0.529167	33.3625	134.7517	sau cùng là thống kê chung cho mức đạm A2

Total			
Count	8	8	8
Sum	230	301.1	352.8
Average	28.75	37.6375	44.1
Variance	19.25714	88.13125	259.0686

Bốn dòng này tính các thống kê chung cho các mức lân lần lượt: B1, B2, B3

Bảng phân tích phương sai

ANOVA

Source of Variation	***SS***	***df***	***MS***	F_{tn}	***P-value***	F_{lt}
Sample	1956.62	1	1956.62	249.7858	5.36E-12	4.413863
Columns	950.3308	2	475.1654	60.66049	1E-08	3.554561
Interaction	467.5808	2	233.7904	29.84611	1.92E-06	3.554561
Within	140.9975	dfe=18	s^2e=7.833194			
Total	3515.53	23				