Bài tập lớn xác suất thống kê

Ví dụ 3.4 trang 161:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1														
2		B1	B2	B3	B4									
3	A1	9	14	16	12									
4	A2	12	15	12	10									
5	A3	13	14	11	14									
6	A4	10	11	13	13									
7														
8	Ti…	51	49	52	47		SUMSQTi..	9915	SSR	3.6875	MSR	1.2292	FR	3.1055
9	Tj…	44	54	52	49		SUMSQTj…	9957	SSC	14.1875	MSC	4.7292	FC	11.9484
10	T…k	41	53	57	48		SUMSQT..k	10043	SSF	35.6875	MSF	11.896	F	30.0552
11	T..	199					SQT….	39601	SSE	2.375	MSE	0.3958		
12							SUMSQ Yij.	2531	SST	55.9375				
13														
14														

Nhấn lần lượt đơn lệnh thiết lặp các biểu thức và tính giá trị thống kê:

- **<u>Tính các giá trị Ti…,Tj.. và T..k ,T…</u>**

- **Các giá trị Ti..**

  Chọn ô B8 và nhập biểu thức =SUM(B3:E3)
  Chọn ô C8 và nhập biểu thức =SUM(B4:E4)
  Chọn ô D8 và nhập biểu thức =SUM(B5:E5)
  Chọn ô E8 và nhập biểu thức =SUM(B6:E6)

- **Các giá trị T.j.**

  Chọn ô B9 và nhập =SUM(B3:B6)
  Dùng con trỏ kéo kí tự điền từ ô B9 tới ô E9.

- **Các giá trị T..k**

  Chọn ô B10 và nhập =SUM(B3,C6,D5,E4)
  Chọn ô C10 và nhập =SUM(B4,C3,D6,E5)
  Chọn ô D10 và nhập =SUM(B5,C4,D3,E6)
  Chọn ô È10 và nhập =SUM(B6,C5,D4,E3)

- **Giá trị T..**

  Chọn ô B11 nhập =SUM(B3:E6)

- **<u>Tính các giá trị G và G</u>**

  -Các giá trị G và G

Chọn ô H8 nhập =SUMSQ(B8:E8)

Dùng con trỏ kéo kí hiệu điền từ ô H8 tới H10

-giá trị G

Chọn ô H11 nhập =SUMSQ(B11)

-giá trị G

Chọn ô H12 nhập =SUMSQ(B3:E6)

- **Tính các giá trị SSR ,SSC,SSF,SST và SSE**

-Các giá trị SSR,SSC và SSF

Chọn ô J8 nhập =H8/4-39601/SUMSQ(4)

Dùng con trỏ kéo kí tự điền từ ô J8 tới J10

-Giá trị SST

Chọn ô J12 nhập =H12-H11/SUMSQ(4)

-Giá trị SSE

Chọn ô J11 và nhập =J12-(J10+J9+J8)

- **Tính các giá trị MSR, MSC,MSF và MSE**

-Các giá trị MSR,MSC và MSF

Chọn ô L8 nhập =J8/(4-1)

Dùng con trỏ kí tự kéo từ ô L8 tới ô L10

-Giá trị MSE

Chọn ô L11nhập =J11/(3*2)

- **tính giá trị G và F**

chọn ô N8 nhập =L8/$L11

dùng con trỏ kéo kí tự từ ô N8 tới ô N10

⇨ **kết quả và biện luận**

F_r =3.1055 < $F_{0.05(3,6)}$ =4.76 => chấp nhận Ho(pH)

F_c=11.95 >$F_{0.05(3,6)}$=4.76 =>bác bỏ Ho(nhiệt độ)

F=30.05 > $F_{0.05 (3,6)}$=4.76 =>bác bỏ Ho(chất xúc tác)

Vậy chỉ có nhiệt và chất xúc tác gây ảnh hưởng đến hiệu suất.

Ví dụ 4.2 (trang 170)

Hồi quy (Regression)

a) Hồi quy đơn tuyến tính

- Phương trình hồi quy tuyến tính: $\bar{y}_x = a + bx \quad , a = r\frac{\overline{S_y}}{\overline{S_x}} \quad , b = \bar{y} - a\bar{x}$
- Kiểm định hệ số a,b
  * Giả thiết H_0: Hệ số hồi quy không có ý nghĩa (= 0)
    H_1: Hệ số hồi quy có ý nghĩa (≠ 0)
  * Trắc nghiệm $t < t_{\alpha,n-2}$: chấp nhận H_0
- Kiểm định phương trình hồi quy
  * Giả thiết H_0:"Phương trình hồi quy tuyến tính không thích hợp"
    H_1: "Phương trình hồi quy tuyến tính thích hợp"
  * Trắc nghiệm $F < F_{\alpha,1,n-2}$: chấp nhận H_0

Nhập số liệu vào bảng excel:

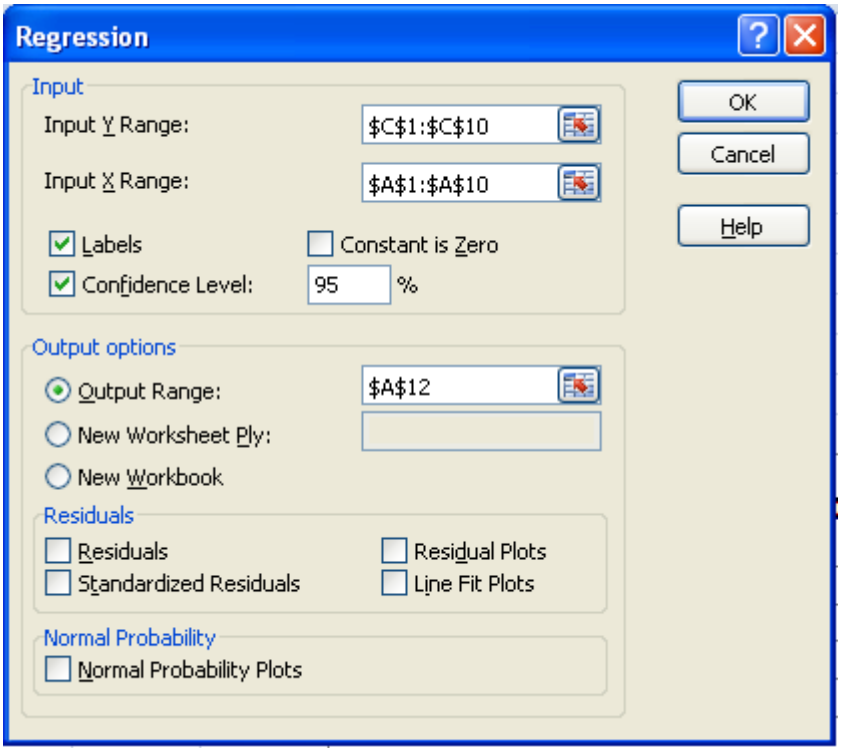
	A	B	C
1	X1	X2	Y
2	15	105	1.87
3	30	105	2.02
4	60	105	3.28
5	15	120	3.05
6	30	120	4.07
7	60	120	5.54
8	15	135	5.03
9	30	135	6.45
10	60	135	7.26

Dùng lệnh Tools và lệnh Data Analysis

Chọn chương trình Regression, lần lượt ấn định các chi tiết:

- phạm vi của biến số Y (input Y range)
- Phạm vi của biến số X (input X range)
- Nhãn dữ liệu (Labels)

- Mức tin cậy (Confidence Level)
- Tọa độ đầu ra (output Range)
- Và tùy chọn khác như đường hồi quy (Line Fit Plots), biểu thức sai số (residuals plots...)



Phương trình hồi quy $\hat{Y}_{X1}=f(x1)$

$$\hat{Y}_{X1}=2.73+0.04X1 \quad (R^2=0.21;\ s=1.81)$$

Regression Statistics								
Multiple R	0.462512							
R Square	0.213917							
Adjusted R Square	0.10162							
Standard Error	1.811192							
Observations	9							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	6.24891746	6.248917	1.904917	0.209995			
Residual	7	22.96290476	3.280415					
Total	8	29.21182222						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	2.726667	1.280705853	2.129034	0.070771	-0.30172	5.755055	-0.30172	5.755055
X1	0.04454	0.032270754	1.380187	0.209995	-0.03177	0.120848	-0.03177	0.120848

T0=2.19<T0.05=2.365 (hay P^2v =0.071 >α=0.05

⇨ Chấp nhận giả thuyết Ho.

T1 =1.38 < T0.05=2.365 (hay Pv =0.209 >α=0.05)

⇨ Chấp nhận giả thuyết Ho.

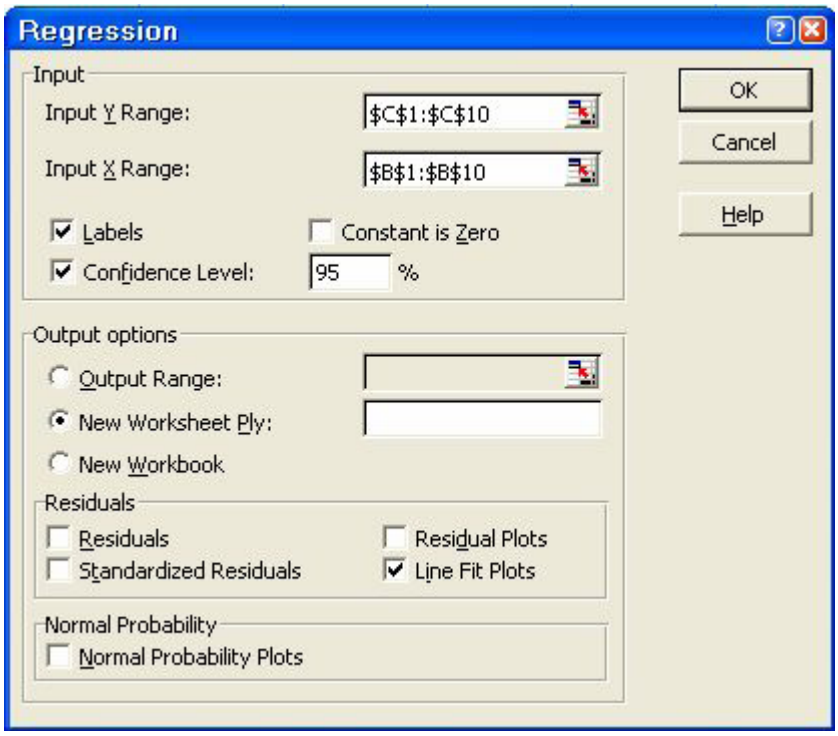
F=1.905 < $F^3$0.05=5.590 (hay F^4s=0.209 >α=0.05) =>Chấp nhận giả thuyết Ho.

Vậy cả 2 hệ số 2.37 (Bo) và 0.04(B1) của phương trình hồi quy

$$\hat{Y}xi = 2.73 + 0.04\,X1$$

Đều không có ý nghĩa thống kê. Nói cách khác hồi quy này không thích hợp.

Kết Luận: yếu tố thời gian không liên quan tuyến tính tới hiệu suất của phả ứng tổng hợp.



Phương trình hồi quy $\hat{Y}\,X2 = f(X2)$

$$\hat{Y}\,x2 = 2.73 + 0.04\,X2\ (R^2\text{=0.76; s=0.99})$$

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.873934							

R Square	0.76376							
Adjusted R Square	0.730011							
Standard Error	0.992904							
Observations	9							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	22.31081667	22.31082	22.63086	0.002066			
Residual	7	6.901005556	0.985858					
Total	8	29.21182222						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11.1411	3.25965608	-3.41788	0.011168	-18.849	-3.43325	-18.849	-3.43325
X2	0.128556	0.027023418	4.757191	0.002066	0.064655	0.192456	0.064655	0.192456

T0=3.418>T0.05=2.365 (hay P^2v =0.011 <α=0.05)

⇨ Bác bỏ giả thuyết Ho.

T2 =4.757 >T0.05=2.365 (hay Pv =0.00206 <α=0.05)

⇨ Bác bỏ giả thuyết Ho.

F=22.631 > $F^3$0.05=5.590 (hay F^4s=0.00206 <α=0.05)

⇨ Bác bỏ giả thuyết Ho.

Vậy cả 2 hệ số -11.14 (Bo) và 0.13(B2) của phương trình hồi quy

$$\hat{\mathbf{Y}}x2 = -11.14 + 0.13\ X2$$

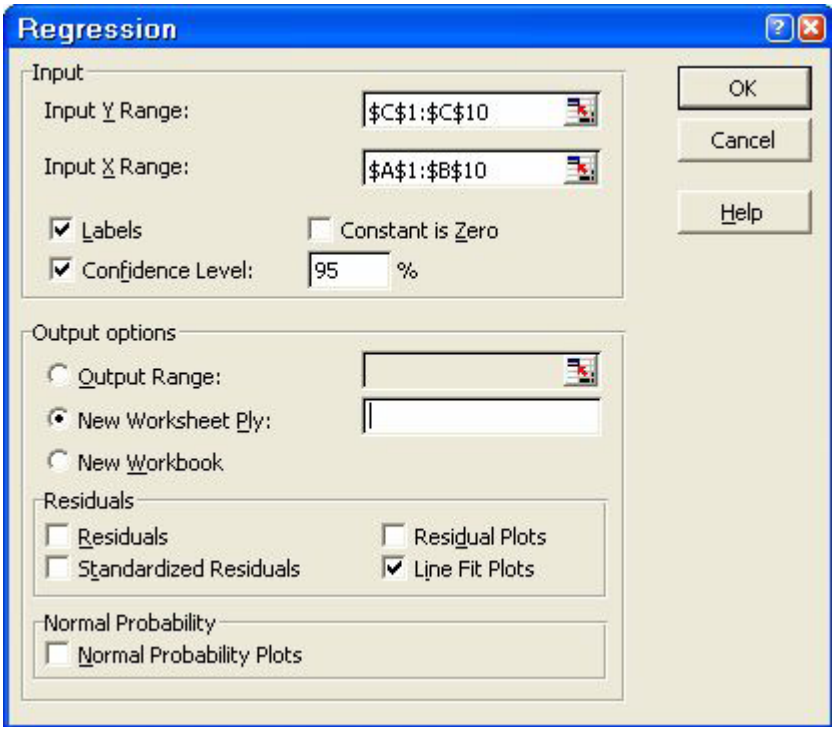
Đều có ý nghĩa thống kê. Nói cách khác hồi quy này thích hợp.

Kết Luận: yếu tố nhiệt độ có liên quan tuyến tính với hiệu suất của phản ứng tổng hợp.

Hồi quy đa tuyến tính

- Phương trình hồi quy đa tuyến tính: $\overline{y}_x = b_0 + b_1 x_1 + \ldots + b_n x_n$
- Kiểm định hệ số b_j
  * Giả thiết H_0: Các hệ số hồi quy không có ý nghĩa ($b_j = 0$)
    H_1: Có ít nhất vài hệ số hồi quy có ý nghĩa ($b_j \neq 0$)
  * Trắc nghiệm $t < t_{\alpha,n-2}$: chấp nhận H_0

- Kiểm định phương trình hồi quy
  * Giả thiết H_0:"Phương trình hồi quy không thích hợp"
    H_1: "Phương trình hồi quy thích hợp với ít nhất vài b_j"
  * Trắc nghiệm $F < F_{\alpha,1,n-2}$: chấp nhận H_0



Phương trình hồi quy $\hat{Y}x1,x2 = f(X1, X2)$

$$\hat{Y}x1,x2 = -12.70 + 0.04\,X1 + 0.13X2$$ (R^2=0.97; s=0.33)

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.988776							
R Square	0.977677							
Adjusted R Square	0.970236							
Standard Error	0.329669							
Observations	9							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	28.55973413	14.27987	131.3921	1.11E-05			
Residual	6	0.652088095	0.108681					
Total	8	29.21182222						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12.7	1.101638961	-11.5283	2.56E-05	-15.3956	-10.0044	-15.3956	-10.0044
X1	0.04454	0.005873842	7.582718	0.000274	0.030167	0.058912	0.030167	0.058912
X2	0.128556	0.008972441	14.32782	7.23E-06	0.106601	0.15051	0.106601	0.15051

T0=11.528>T0.05=2.365 (hay P^2v =2.260 *10^{-5} >α=0.05)

⇨ Bác bỏ giả thuyết Ho.

T2 =7.583 >T0.05=2.365 (hay Pv =0.00207 <α=0.05)

⇨ Bác bỏ giả thuyết Ho.

T2 =14.328> T0.05=2.365 (hay Pv =7.233 * 10^{-6} >α=0.05)

⇨ Bác bỏ giả thuyết Ho.

F=131.329 < F 0.05=5.140 (hay Fs=1.112 * 10^{-5} >α=0.05)

⇨ Bác bỏ giả thuyết Ho.

Vậy cả 2 hệ số -12.14 (Bo) và 0.13(B2) của phương trình hồi quy

$$\hat{Y}\ x2 = -11.14 + 0.13\ X2$$

KếtLuận: Hiệu suất của phản ứng tổng hợp có liên quan tuyến tính với cả

2 yếu tố là thời gian và nhiệt độ.

Sự tuyến tính của phương trình **Ŷ** x1,x2 = -12,70 + 0,04X1 + 0.13X2 có thể được trình bày trên biểu đồ phân tán. (*scatterplots*)

Muốn dự đoán hiệu suất của phản ứng bằng phương trình hồi quy :

Ŷx1,x2 = -12,70 + 0,04X1 + 0.13X2 bạn chỉ cần chọn 1 ô, vídụ B21, sau đó nhập hàm và được kết quả như sau:

	B21		= B17 + B18*50 + B19*115	
	A	B	C	D
17	Interrcept	-12.7	1.1.1638961	-11.52827782
18	X1	0.044539683	0.005873842	7.582717621
19	X2	0.128555556	0.008972441	14.32782351
20				
21	Dựđoán 4,310873016			

<u>Câu 2</u>: bệnh đau mắt hột được chia làm 4 thời kì T_1, T_2, T_3 và T_4.Kết quả kiểm tra mắt hột ở 3 tỉnh A, B, C được cho trong bảng sau đây:

Địa phương	Mức độ đau mắt hột			
	T_1	T_2	T_3	T_4
A	**47**	**189**	**807**	**1768**
B	**53**	**746**	**1387**	**946**
C	**16**	**228**	**438**	**115**

Hãy nhận định xem tình hình đau mắt hột(cơ cấu phân bố các mức độ T1,T2,T3,T4) Ở 3 tỉnh trên có giống nhau không? Mức ý nghĩa 1%)

<u>CƠ SỞ LÍ THUYẾT</u> Giaû söû ta coù k taäp hôïp chính $H_1, H_2, ... H_k$. Moãi caù theå cuûa chuùng coù theå mang hay khoâng mang ñaëc tính A.

Goïi p_1 laø tyû leä coù theå mang ñaëc tính A trong taäp hôïp chính H_i (i = 1, 2, ...k).

Caùc tyû leä naøy ñöôïc goïi laø caùc tyû leä lyù thuyeát maø chuùng ta chöa bieát.

Ta muoán kieåm ñònh giaû thieát sau:

H_o: $p_1 = p_2 = ... = p_k$ (taát caû caùc tyû leä naøy baèng nhau).

Töø moãi taäp hôïp chính H_i ta ruùt ra moät ngaãu nhieân coù kích thöôùc n_i, trong ñoù chuùng ta thaáy coù m_i caù theå mang ñaëc tính A. caùc döõ lieäu naøy ñöôïc trình baøy trong baûng sau ñaây:

Maãu	1	2	...	k	Toång
Coù A	**m_1**	**m_2**	**...**	**m_k**	**m**
Khoâng A	**l_1**	**l_2**	**...**	**l_k**	**l**
Toång	**n_1**	**n_2**	**...**	**n_k**	**$N = m + l = \sum n_i$**

Neáu giaû thieát

$$H_o: p_1 = p_2 = ... = p_k = p$$

Laø ñuùng thì tyû leä chung p ñöôïc öôùc löôïng baèng tyû soá giöõa soá caù theå ñaëc tính A cuûa toaøn boä k maãu goäp laïi treân toång soá caù theå cuûa k maãu goäp laïi.

$$\hat{p} = \frac{m}{N}$$

Tyû leä caù theå khoâng coù ñaëc tính A ñöôïc öôùc löôïng bôûi

$$\hat{q} = 1 - \hat{p} = \frac{l}{N}$$

Khi ñoù soá caù theå coù ñaëc tính A trong maãu thöù i (maãu ruùt töø taäp hôïp chính H_i) seõ xaáp xæ baèng

$$\mu_i = n_i p = \frac{n_i m}{N}$$

vaø soá caù theå khoâng coù ñaëc tính A trong maãu thöù i seõ xaáp xæ baèng

$$\hat{\mu}_i = n_i q = n_i \frac{l}{N}$$

Caùc soá μ_i vaø $\hat{\mu}_i$ ñöôïc goïi laø caùc taàn soá lyù thuyeát (TSLT), coøn caùc soá m_i, l_i ñöôïc goïi laø caùc taàn soá quan saùt (TSQS).

Ta quyeát ñònh baùc boû H_o khi TSLT caùch xa TSQS moät caùch "baát thöôøng". Khoaûng caùch giöõa TSQS vaø TSLT ñöôïc ño baèng test thoáng keâ sau ñaây:

$$T = \sum_{i=1}^{k} \frac{(m_i - \mu_i)^2}{\mu_i} + \sum_{i=1}^{k} \frac{(l_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Ngöôøi ta chöùng minh ñöôïc raèng neáu H_o ñuùng vaø caùc taàn soá lyù thuyeát khoâng nhoû thua 5 thì T seõ coù phaân boá xaáp xæ phaân boá χ^2 vôùi k – 1 baäc töï do. Thaønh thöû mieàn baùc boû H_o coù daïng {T > c}, ôû ñoù c ñöôïc tìm töø ñieàu kieän P{T > c} = α. Vaäy c chính laø phaân vò möùc α cuûa phaân boá χ^2 vôùi k – 1 baäc töï do

Nhập bảng số liệu vào Excel:

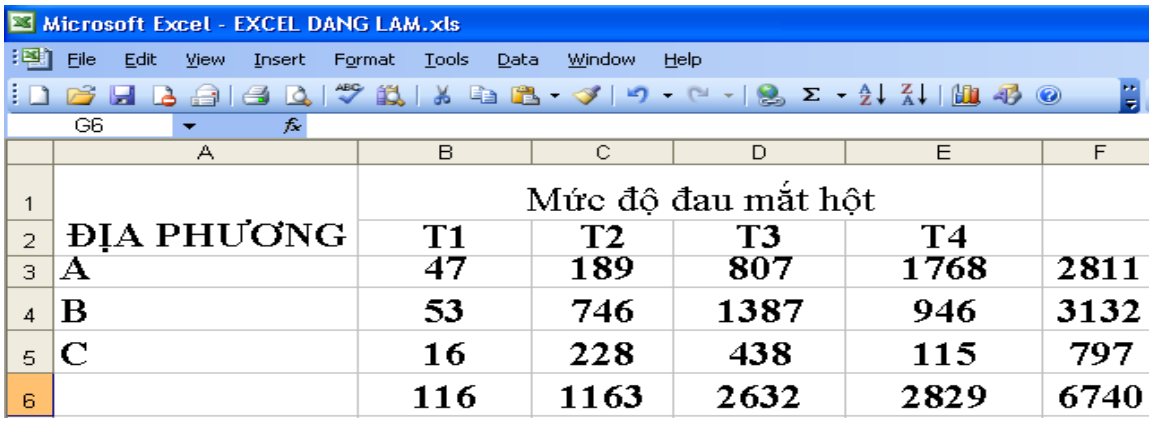
	A	B	C	D	E
1		Mức độ đau mắt hột			
2	ĐỊA PHƯƠNG	T1	T2	T3	T4
3	A	47	189	807	1768
4	B	53	746	1387	946
5	C	16	228	438	115

Tính các giá trị :

Tổng hàng :

- Chọn ô F3 và nhập biểu thức =SUM(B3:E3)
- Dùng con trỏ kéo kí hiệu tự điền từ ô F3 đến ô F5
- Tổng cột :
- Chọn ô B6 và nhập vào biểu thức =SUM(B3:B5)
- Dùng con trỏ kéo kí tự điền từ ô B6 đến ô E6
- Chọn ô F6 =sum(B6:E6)

Ta được kết quả:



	A	B	C	D	E	F
1		Mức độ đau mắt hột				
2	ĐỊA PHƯƠNG	T1	T2	T3	T4	
3	A	47	189	807	1768	2811
4	B	53	746	1387	946	3132
5	C	16	228	438	115	797
6		116	1163	2632	2829	6740

Tính TSLT :

- Chọn ô B8 và nhập vào biểu thức =B$6*$F3/$F$6
- Dùng con trỏ kéo kí tự điền từ ô B8 đến ô D10

Ta được kết quả :

	A	B	C	D	E	F
1		Mức độ đau mắt hột				
2	ĐỊA PHƯƠNG	T1	T2	T3	T4	
3	A	47	189	807	1768	2811
4	B	53	746	1387	946	3132
5	C	16	228	438	115	797
6		116	1163	2632	2829	6740
7		48.3792	485.04	1097.71	1179.869	
8		53.9039	540.43	1223.06	1314.604	
9		13.7169	137.52	311.232	334.5272	
10						

Tất cả các TSLT đều lớn hơn 5 :

<u>Ta tính T</u>

ta có công thức :

$$T = \sum \frac{n_{ij}^2}{n_{ij}} - 2n + n = \sum \frac{n_{ij}^2}{n_{ij}} = n\sum \frac{n_{ij}^2}{n_{io}n_{oj}} - n = n\left\{\sum \frac{n_{ij}^2}{n_{io}n_{oj}} - 1\right\}$$

- Chọn ô B12 và nhập vào biểu thức =B3*B3/(B$6*$F3)
- Dùng con trỏ kèo kí tự điền từ ô B11 đến ô E13
- Chọn tiếp ô B14 và nhập vào biểu thức =F6*(SUM(B11:E13)-1)
- Ta đươc kết quả : T ≈ 1010

	A	B	C	D	E	F
1		Mức độ đau mắt hột				
2	ĐỊA PHƯƠNG	T1	T2	T3	T4	
3	A	47	189	807	1768	2811
4	B	53	746	1387	946	3132
5	C	16	228	438	115	797
6		116	1163	2632	2829	6740
7		48.3792	485.04	1097.71	1179.869	
8		53.9039	540.43	1223.06	1314.604	
9		13.7169	137.52	311.232	334.5272	
10						
11		0.00677	0.0109	0.08802	0.393071	
12		0.00773	0.1528	0.23337	0.101001	
13		0.00277	0.0561	0.09145	0.005865	
14		1010.02				
15						

Tra bảng phân phối χ^2 (α=1%) với bậc tự do (3-1)*(4-1)=6 ta được : 16,81
Vì T>16,81 => bác bỏ H_0

Vậy đau mắt hột ở 3 tỉnh trên khác nhau.

Câu 3: **Bảng sau đây cho số liệu người chết về ung thư ở 3 nước Mỹ,Anh,Nhật.Người chết được phân loại theo cơ quan bị ung thư.**

Chỗ ung thư	Nước		
	Mỹ	**Nhật**	**Anh**
Ruột	**11**	**5**	**5**
Ngực	**15**	**3**	**7**
Dạ dày	**3**	**22**	**3**
Bộ phận khác	**41**	**30**	**15**

a) Hãy tính tần số lí thuyết của bảng số liệu trên.

b) Có thể áp dụng tiêu chuẩn χ^2 được không?

c) Với mức ý nghĩa α = 1% hãy so sánh phân bố tỉ lệ chết về ung thư của 3 nước nói trên.

CƠ SỞ LÍ THUYẾT

Maãu / Tính traïng	1	2		J		K	Toång soá
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1k}	n_{10}
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2k}	n_{20}
...	...	...	...	...	...	...	...
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ik}	n_{i0}

...	**...**	**...**	**...**	**...**	**...**	**...**	**...**
$\mathbf{A_r}$	$\mathbf{n_{r1}}$	$\mathbf{n_{r2}}$	**...**	$\mathbf{n_{rj}}$	**...**	$\mathbf{n_{rk}}$	$\mathbf{n_{r0}}$
Toång soá	$\mathbf{n_{o1}}$	$\mathbf{n_{o2}}$	**...**	$\mathbf{n_{oj}}$	**...**	$\mathbf{n_{ok}}$	**n**

Kyù hieäu $n_{io} = \sum_{j=1}^{k} n_{ij}$

$$n_{oj} = \sum_{i=1}^{r} n_{ij}$$

Nhö vaäy n_{oj} laø kích thöôùc cuûa maãu thöù j, coøn n_{io} laø toång soá caù theå coù tính traïng A_i trong toaøn boä k maãu ñang xeùt

$$n = \sum_{i=1}^{r} n_{io} = \sum_{j=1}^{k} n_{oj}$$

Laø toång soá taát caû caùc caù theå cuûa k maãu ñang xeùt.

Neáu giaû thieát H_o laø ñuùng nghóa laø

$$\begin{cases} p_1^1 = p_1^2 = K = p_1^k = p_1 \\ p_2^1 = p_2^2 = K = p_2^k = p_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_i^1 = p_i^2 = K = p_i^k = p_i \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ p_r^1 = p_r^2 = K = p_r^k = p_r \end{cases}$$

thì caùc tyû leä chung $p_1, p_2,...p_r$ ñöôïc öôùc löôïng bôûi:

$$\hat{p}_i = \frac{n_{io}}{n}$$

Ñoù öôùc löôïng cho xaùc suaát ñeå moät caù theå coù mang tính traïng A_i. khi ñoù soá caù theå coù tính traïng A_i trong maãu thöù j seõ xaáp xæ baèng

$$\hat{n}_{ij} = n_{oj}\hat{p}_i = \frac{n_{oj}n_{io}}{n}$$

Caùc soá $\hat{n}_{ij} (i = 1,2,...r;\ j = 1,2,...k)$

ñöôïc goïi laø caùc taàn soá lyù thuyeát (TSLT), caùc soá n_{ij} ñöôïc goïi laø caùc taàn soá quan saùt (TSQS).

Ta quyeát ñònh baùc boû H_o khi caùc TSLT caùch xa TSQS moät caùch baát thöôøng. Khoaûng caùch giöõa TSQS vaø TSLT ñöôïc ño baèng test thoáng keâ sau ñaây

$$T = \sum_{f=1}^{k}\sum_{i=1}^{r} \frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}} = \sum \frac{(TSQS - TSLT)^2}{TSLT}$$

Ngöôøi ta chöùng minh ñöôïc raèng neáu H_o ñuùng vaø caùc TSLT khoâng nhoû hôn 5 thì T seõ coù phaân boá xaáp xæ phaân boá χ^2 vôùi (k-1)(r-1) baäc töï do. Thaønh thöû mieàn baùc boû coù daïng {T > c} ôû ñoù c ñöôïc tìm töø ñieàu kieän P{T > c} = α. Vaäy c laø phaân vò möùc α cuûa phaân boá χ^2 vôùi (k-1)(r-1) baäc töï do.

***Chuù yù.* T coù theå bieán ñoåi thaønh caùc daïng sau ñaây.**

Ta coù $\dfrac{\left(n_{ij}-\hat{n}_{ij}\right)^2}{\hat{n}_{ij}}=\dfrac{n_{ij}^2}{\hat{n}_{ij}}-2n_{ij}+\hat{n}_{ij}$

Ñeå yù raèng: $\sum\sum n_{ij}=\sum\sum\hat{n}_{ij}=n$

Vaäy $T=\sum\dfrac{n_{ij}^2}{\hat{n}_{ij}}-2n+n=\sum\dfrac{n_{ij}^2}{\hat{n}_{ij}}=n\sum\dfrac{n_{ij}^2}{n_{io}n_{oj}}-n=n\left\{\sum\dfrac{n_{ij}^2}{n_{io}n_{oj}}-1\right\}$

Nhập bảng số liệu vào Excel , thiết lập các biểu thức và tính các giá trị thống kê.

Tính các giá trị :

- Tổng hàng :
  - Chọn ô E2 và nhập biểu thức =SUM(B2:D2)
  - Dùng con trỏ kéo kí hiệu tự điền từ ô E2 đến ô E5
- Tổng cột :
  - Chọn ô B6 và nhập vào biểu thức =SUM(B2:B5)
  - Dùng con trỏ kéo kí tự điền từ ô B6 đến ô E6

Ta được kết quả như hình sau :

	A	B	C	D	E
1		**Mỹ**	**Nhật**	**Anh**	**Tổng**
2	**Ruột**	**11**	**5**	**5**	**21**
3	**Ngực**	**15**	**3**	**7**	**25**
4	**Dạ dày**	**3**	**22**	**3**	**28**
5	**Bộ phận khác**	**41**	**30**	**15**	**86**
6	**Tổng**	**70**	**60**	**30**	**160**
7					

a) Ta có công thức tần số lý thuyết(TSLT) :

$$\hat{n}_{ij}=n_{oj}\hat{p}_i=\frac{n_{oj}n_{io}}{n}$$

Thao tác trên Excel :

- Chọn ô B8 và nhập vòa biểu thức =B$6*$E2/$E$6
- Dùng con trỏ kéo kí tự điền từ ô B8 đến ô D8
- Dùng con trỏ kéo kí tự điền từ ô D8 xuống ô D11

Ta được kết quả sau :

	A	B	C	D	E	F	G
1		Mỹ	Nhật	Anh	Tổng		
2	Ruột	11	5	5	21		
3	Ngực	15	3	7	25		
4	Dạ dày	3	22	3	28		
5	Bộ phận khác	41	30	15	86		
6	Tổng	70	60	30	160		
7							
8		9.1875	7.875	3.9375			
9		10.9375	9.375	4.6875			
10		12.25	10.5	5.25			
11		37.625	32.25	16.125			
12							
13							

b)không thể áp dụng chuẩn χ^2 vì không phải TSLT đều không nhỏ hơn 5.
c.)vì có TSLT nhỏ hơn 5.
Nên ta cần ghép 2 dòng đầu tiên :

	Mỹ	Nhật	Anh
Ruột+Ngực	26	8	12
Dạ dày	3	22	3
Bộ phận khác	41	30	15

Tính các giá trị :
Nhập các giá trị vào bảng Excel :

- Chọn ô E2 và nhập vào biểu thức =SUM(B2:D2)
- Dùng con trỏ kéo kí tự điền từ ô E2 đến ô E4
- Chọn ô B5 và nhập vào biểu thức =SUM(B2:B4)
- Dùng con trỏ kéo kí tự điền từ ô B5 đến ô E5

Ta được kết quả :

	A	B	C	D	E	F
1		Mỹ	Nhật	Anh	Tổng	
2	Ruột+Ngực	26	8	12	46	
3	Dạ dày	3	22	3	28	
4	Bộ phận khác	41	30	15	86	
5	Tổng	70	60	30	160	
6						

Tính TSLT :

- Chọn ô B7 và nhập vào biểu thức =B$5*$E2/$E$5
- Dùng con trỏ kéo kí tự điền từ ô B7 đến ô D9

Ta được kết quả :

	A	B	C	D	E	F
1		**Mỹ**	**Nhật**	**Anh**	**Tổng**	
2	**Ruột+Ngực**	**26**	**8**	**12**	**46**	
3	**Dạ dày**	**3**	**22**	**3**	**28**	
4	**Bộ phận khác**	**41**	**30**	**15**	**86**	
5	**Tổng**	**70**	**60**	**30**	**160**	
6						
7		**20.125**	**17.25**	**8.625**		
8		**12.25**	**10.5**	**5.25**		
9		**37.625**	**32.25**	**16.125**		
10						

Tất cả các TSLT đều lớn hơn 5 :
Ta tính T
Ta có công thức :

$$T = \sum \frac{n_{ij}^2}{\hat{n}_{ij}} - 2n + n = \sum \frac{n_{ij}^2}{\hat{n}_{ij}} = n\sum \frac{n_{ij}^2}{n_{io}n_{oj}} - n = n\left\{\sum \frac{n_{ij}^2}{n_{io}n_{oj}} - 1\right\}$$

Từ đó ta được :

$$T=160\left[\frac{26^2}{70.46}+\frac{8^2}{60.46}+\frac{12^2}{30.46}+\frac{3^2}{70.28}+\frac{22^2}{60.28}+\frac{3^2}{30.28}+\frac{41^2}{70.86}+\frac{30^2}{60.86}+\frac{15^2}{30.86}-1\right]\approx 29{,}07$$

Dùng Excel

- Chọn ô B11 và nhập vào biểu thức =B2*B2/(B$5*$E2)
- Dùng Dùng con trỏ kèo kí tự điền từ ô B11 đến ô D13
- Chọn tiếp ô B18 và nhập vào biểu thức =E5*(SUM(B11:D13)-1)

Ta đươc kết quả : T ≈ 29,07

	A	B	C	D	E	F	G
1		Mỹ	Nhật	Anh	Tổng		
2	Ruột+Ngực	26	8	12	46		
3	Dạ dày	3	22	3	28		
4	Bộ phận khác	41	30	15	86		
5	Tổng	70	60	30	160		
6							
7		20.125	17.25	8.625			
8		12.25	10.5	5.25			
9		37.625	32.25	16.125			
10							
11		0.209938	0.0231884	0.104348			
12		0.004592	0.2880952	0.010714			
13		0.279236	0.1744186	0.087209			
14							
15		29.07828					
16							
17							

Tra bảng phân phối χ^2 (α=1%) với bậc tự do (3-1)*(3-1)=4ta được : 13,28
Vì T>13,28 => **bác bỏ H_0 . Vậy tỉ lệ chết về ung thư của 3 nước là khác nhau** .

Câu 4:

Theo dõi doanh thu của 4 cửa hàng của 1 công ty(triệu đ/tháng)người ta được số liệu như sau:

Tháng kinh doanh	Cửa hàng			
	1	2	3	4
1	*12,3*	*14,2*	*15,6*	*17,2*
2	*12,6*	*12,4*	*17,1*	*15,8*
3	*11,6*	*11,5*	*18,2*	*12,2*
4	*15,2*	*11,6*	*12,5*	
5	*18,6*		*11,8*	
6	*17,1*			

Hãy so sánh doanh thu trung bình/tháng của các cửa hàng thuộc công ty nói trên. **Mức ý nghĩa α = 5%.**

CƠ SỞ LÍ THUYẾT

PHÂN TÍCH PHƯƠNG SAI (ANOVA)

1. Phân tích phương sai 1 nhân tố

Giả sử nhân tố A có k mức $X_1, X_2, \dots, X_k$ với X_j có phân phối chuẩn $N(a,\sigma^2)$ có mẫu điều tra

X_1	X_2	---	X_k
x_{11}	x_{12}		x_{1k}
x_{21}	x_{22}		x_{2k}
⋮	⋮		⋮
⋮	⋮	...	⋮
$x_{n_1 1}$	⋮		$x_{n_k k}$
	$x_{n_2 2}$		

Với mức ý nghĩa α , hãy kiểm định giả thiết :

$H_0 : a_1 = a_2 = \dots = a_k$

H_1 : "Tồn tại $j_1 \neq j_2$ sao cho $a_{j1} \neq a_{j2}$ "

- Đặt:
  - Tổng số quan sát: $n = \sum_{j=1}^{k} n_j$
  - Trung bình mẫu nhóm j (j =1, .. , k): $\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} = \frac{T_j}{n_j}$ với $T_j = \sum_{i=1}^{n_j} x_{ij}$
  - Trung bình mẫu chung: $\bar{x} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} = \frac{T}{n}$ v ới $T = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij} = \sum_{j=1}^{k} T_j$
  - Phương sai hiệu chỉnh nhóm j: $S_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)^2$

- SST = $\sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2$ Tổng bình phương các độ lệch.
- SSA = $\sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$ Tổng bình phương độ lệch riêng của các nhóm so với $\bar{x}$

$$SST = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{T^2}{n} \qquad SSA = \sum_{j=1}^{k}\frac{T_j^2}{n_j} - \frac{T^2}{n} \qquad SSE = SST - SSA$$

$$MSA = \frac{SSA}{k-1} \qquad MSE = \frac{SSE}{n-k}$$

- Nếu H_0 đúng thì $F = \frac{MSA}{MSE}$ có phân phối Fisher bậc tự do k-1; n-k
- Miền B_α : $F > F_{k-1;\, n-k\,;\, 1-\alpha}$

Bảng ANOVA

Nguồn sai số	Tổng bình phương SS	Bậc tự do df	Bình phương trung bình MS	Giá trị thống kê F
Yếu t ố (Between Group)	SSA	k-1	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$
Sai số (Within Group)	SSE = SST - SSA	n-k	$MSE = \frac{SSE}{n-k}$	
Tổng cộng	SST	n-1		

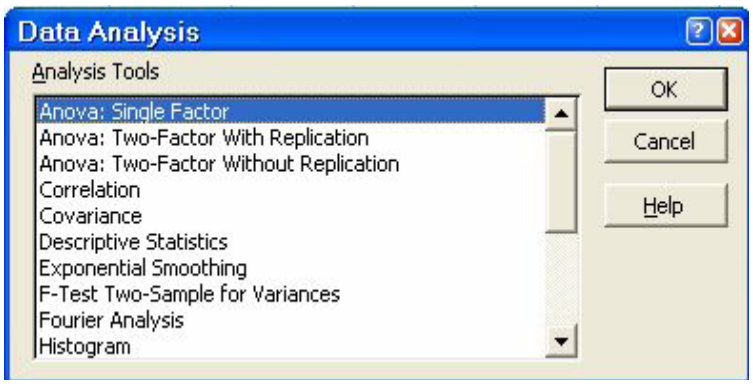
Dùng Excel

- Áp dụng "Anova: Single Factor"

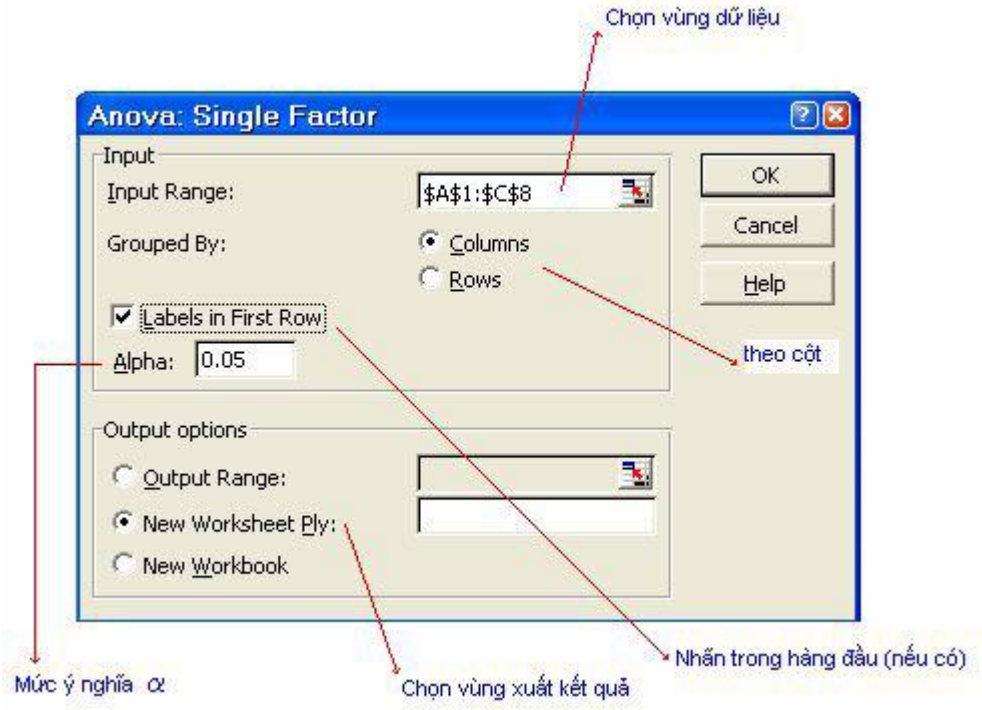
1. Nhập dữ liệu theo cột

	A	B	C	D	E
1		cửa hàng			
2	tháng	1	2	3	4
3	1	12.3	14.2	15.6	17.2
4	2	12.6	12.4	17.1	15.8
5	3	11.6	11.5	18.2	12.2
6	4	15.2	11	12.5	
7	5	18.6	6	11.8	
8	6	17.1			
9					

2. Chọn mục : Anova: Single Factor



3. Chọn các mục như hình:



4. Kết quả

5.

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Column 1	6	87.4	14.5667	8.1787
Column 2	4	49.7	12.425	1.5625
Column 3	5	75.2	15.04	7.873
Column 4	3	45.2	15.0666	6.6533

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*

Between Groups	18.9899	3	6.3300	0.9805	0.4299	3.3439
Within Groups	90.3795	14	6.4557			
Total	109.3694	17				

- F=MSF/MSE=0.9805<F crit=3.3439 => Chấp Nhận giả thiết H_0
- Vậy doanh thu trung bình/tháng của các cửa hàng bằng nhau

Câu 5: . Mức ý nghĩa α = 5%. Hãy phân tích tình hình kinh doanh của 1 số ngành nghề ở 4 quận nội thành trên cơ sở số liệu về doanh thu của 1 số mặt hàng như sau::

Ngành nghề kinhdoanh	Khu vực kinh doanh			
	Q1	Q2	Q3	Q4
Điện lạnh	2.5:2.7:2.0:3.0	3.1:3.5:2.7	2.0:2.4	5.0:5.4
VLXD	0.6:10.4	15.0	9.5:9.3:9.1	19.5:17.5
Dịch vụ tin học	1.2:1.0:9.8:1.8	2.0:2.2:1.8	1.2:1.3:1.2	5.0:4.8:5.2

CƠ SỞ LÍ THUYẾT:

Phân tích nhằm đánh giá sự ảnh hưởng của 2 nhân tố (yếu tố) A và B trên các giá trị quan sát x_{ij}

Giả sử nhân tố A có n mức $a_1, a_2, \dots, a_n$ (nhân tố hàng)

B có m mức $b_1, b_2, \dots, b_m$ (nhân tố cột)

* Mẫu điều tra:

A \ B	b_1	b_2	…	b_m
a_1	x_{11}	x_{12}	…	x_{1m}
a_2	x_{21}	x_{22}	…	x_{2m}
:	:	:		:
:	:	:		:
a_n	x_{n1}	x_{n2}	…	x_{nm}

* Giả thiết H_0:

- Trung bình nhân tố cột bằng nhau
- Trung bình nhân tố hàng bằng nhau

Tiến hành tính toán theo bảng dưới đây:

A \ B	b_1	b_2	...	b_m	$T_{i*}=\sum_j x_{ij}$	$\sum_j x_{ij}^2$
a_1	x_{11}	x_{12}	...	x_{1m}	T_{1*}	$\sum_j x_{1j}^2$
a_2	x_{21}	x_{22}	...	x_{2m}	T_{2*}	$\sum_j x_{2j}^2$
:	:	:		:	:	
:	:	:		:	:	
a_n	x_{n1}	x_{n2}	...	x_{nm}	T_{n*}	$\sum_j x_{nj}^2$
$T_{*j}=\sum_i x_{ij}$	T_{*1}	T_{*2}	...	T_{*m}	$T=\sum_{i,j} x_{ij}$	
$\sum_i x_{ij}^2$	$\sum_i x_{i1}^2$	$\sum_i x_{i2}^2$		$\sum_i x_{im}^2$		$\sum_{i,j} x_{ij}^2$

* Bảng ANOVA

Nguồn	SS	df	MS	F
Yếu tố A	$SSA=\frac{\sum_i T_{i*}^2}{m}-\frac{T^2}{m.n}$	n-1	$MS\ (A = \frac{SSA}{n-1}$	$F_A=\frac{SSA}{SSE}$
Yếu tố B	$SSB=\frac{\sum_j T_{*j}^2}{n}-\frac{T^2}{m.n}$	m-1	$MSB=\frac{SSB}{m-1}$	$F_B=\frac{SSB}{SSE}$
Sai số	SSE=SST-SSA-SSB	(n-1)(m-1)	$MSE=\frac{SSE}{(n-1)(m-1)}$	
Tổng	$SST=\sum_{i,j} x_{ij}^2-\frac{T^2}{m.n}$	nm-1		

* Kết luận :

· Nếu FA > F n-1 ; (n-1)(m-1) ; 1-a thì bá c bỏ yếu tố A (h àng)

· Nếu FB > F m-1 ; (n-1)(m-1) ; 1-a thì bá c bỏ yếu tố B (cột)

Nhập số liệu cho bảng:

	A	B	C	D	E	F	G	H
1								
2			Khu vực kinh doanh					
3			Q1	Q2	Q3	Q4		
4			2.5	3.1	2	5		
5			2.7	3.5	2.4	5.4		
6			2	2.7				
7		Điện lạnh	3					
8			0.6	15	9.5	19.5		
9			10.4		9.3	17.5		
10		VLXD			9.1			
11			1.2	2	1.2	5		
12			1	2.2	1.3	4.8		
13			9.8	1.8	1.2	5.2		
14		Dịch vụ tin học	1.8					
15								

Ta sử dụng hàm Average để tình trung bình các số liệu cho các cột:

ô J5 =SUM (C4:C7). Kéo điền vào các ô từ J5 đến M5

ô J6 và nhập biểu thức = SUM (C8:C10). Kéo điền từ ô J6 đến ô M6

ô J7 và nhập biểu thức = SUM (C11:C14). Kéo điền từ ô J7 đến ô M7

Ta có bảng kết quả như sau:

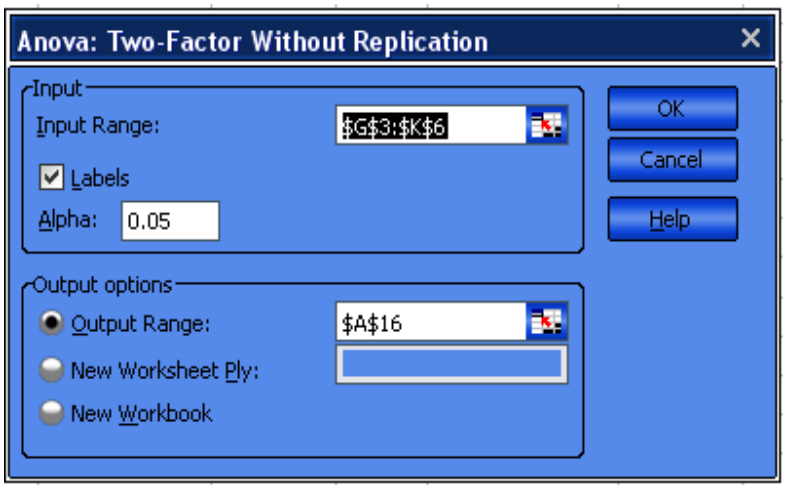
J	K	L	M
10.2	9.3	4.4	10.4
11	15	27.9	37
13.8	6	3.7	15

Dùng lệnh “Anova: Two-Factor Without Replication”

a) Nhấp lần lượt đơn lệnh Tools và lệnh Data Analysis.

b) Chọn chương trình Anova: Two-Factor Without Replication trong hộp thoại Data Analysis rồi nhấp nút OK.

c) Trong hộp thoại Anova: Two-Factor Without Replication lần lượt nhập vào như hình sau:



Anova: Two-Factor Without Replication

SUMMARY	Count	Sum	Average	Variance
điện lạnh	4	34.3	8.575	7.9758333
vlxd	4	90.9	22.725	142.56917
dvth	4	38.5	9.625	31.5225
q1	3	35	11.66666667	3.5733333
q2	3	30.3	10.1	20.73
q3	3	36	12	189.73
q4	3	62.4	20.8	202.12

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	497.2	2	248.6233333	4.4521578	0.065241	5.14325285
Columns	211.1	3	70.38083333	1.2603265	0.368924	4.75706266
Error	335.1	6	55.84333333			
Total	1043	11				

Kết quả biện luận:

⇨ $F_R = 4.4522 < F_{2,6,1-0.05} = 5,14325285$ → chấp nhận giả thiết H_0.

Doanh thu của các cửa hàngkhông phụ thuộc vào ngành nghề.

⇨ $F_C = 1.26033 < F_{3,6,1-0.05} = 4,757062664$ → Chấp nhận giả thiết H_0 .

Doanh thu của các cửa hàng không phụ thuộc vào khu vực kinh doanh.