XỬ LÝ THỐNG KÊ BẰNG EXCEL

Các hàm thống kê có thể chia thành 3 nhóm nhỏ sau: Nhóm hàm về Thống Kê, nhóm hàm về Phân Phối Xác Suất, và nhóm hàm về Tương Quan và Hồi Quy TuyếnTính

*NHÓM HÀM VỀ **THỐNG KÊ***

AVEDEV *(number1, number2, ...)*	Tính trung bình độ lệch tuyệt đối các điểm dữ liệu theo trung bình của chúng. Thường dùng làm thước đo về sự biến đổi của tập số liệu
AVERAGE *(number1, number2, ...)*	Tính trung bình cộng
AVERAGEA *(number1, number2, ...)*	Tính trung bình cộng của các giá trị, bao gồm cả những giá trị logic
AVERAGEIF *(range, criteria1)*	Tính trung bình cộng của các giá trị trong một mảng theo một điều kiện
AVERAGEIFS *(range, criteria1, criteria2, ...)*	Tính trung bình cộng của các giá trị trong một mảng theo nhiều điều kiện
COUNT *(value1, value2, ...)*	Đếm số ô trong danh sách.
COUNTA *(value1, value2, ...)*	Đếm số ô có chứa giá trị (không rỗng) trong danh sách.
COUNTBLANK *(range)*	Đếm các ô rỗng trong một vùng.
COUNTIF *(range, criteria)*	Đếm số ô thỏa một điều kiện cho trước bên trong một dãy
COUNTIFS *(range1, criteria1, range2,criteria2,…)*	Đếm số ô thỏa nhiều điều kiện cho trước.
DEVSQ *(number1, number2, ...)*	Tính bình phương độ lệch các điểm dữ liệu từ trung bình mẫu của chúng, rồi cộng các bình phương đó lại.
FREQUENCY *(data_array, bins_array)*	Tính xem có bao nhiêu giá trị thường xuyên xuất hiện bên trong một dãy giá trị, rồi trả về một mảng đứng các số. Luôn sử dụng hàm này ở dạng công thức mảng
GEOMEAN *(number1, number2, ...)*	Trả về trung bình nhân của một dãy các số dương. Thường dùng để tính mức tăng trưởng trung bình, trong đó lãi kép có các lãi biến đổi được cho trước…
HARMEAN *(number1, number2, ...)*	Trả về trung bình điều hòa (nghịch đảo của trung bình cộng) của các số
KURT *(number1, number2, ...)*	Tính độ nhọn của tập số liệu, biểu thị mức nhọn hay mức phẳng tương đối của một phân bố so với phân bố chuẩn
LARGE *(array, k)*	Trả về giá trị lớn nhất thứ k trong một tập số liệu.
MAX *(number1, number2, ...)*	Trả về giá trị lớn nhất của một tập giá trị.

MAXA *(number1, number2, ...)*	Trả về giá trị lớn nhất của một tập giá trị, bao gồm cả các giá trị logic và text
MEDIAN *(number1, number2, ...)*	Tính trung bình vị của các số.
MIN *(number1, number2, ...)*	Trả về giá trị nhỏ nhất của một tập giá trị.
MINA *(number1, number2, ...)*	Trả về giá trị nhỏ nhất của một tập giá trị, bao gồm cả các giá trị logic và text.
MODE *(number1, number2, ...)*	Trả về giá trị xuất hiện nhiều nhất trong một mảng giá trị.
PERCENTILE *(array, k)*	Tìm phân vị thứ k của các giá trị trong một mảng dữ liệu.
PERCENTRANK *(array, x, significance)*	Trả về thứ hạng (vị trí tương đối) của một trị trong một mảng dữ liệu, là số phần trăm của mảng dữ liệu đó
PERMUT *(number, number_chosen)*	Trả về hoán vị của các đối tượng.
QUARTILE *(array, quart)*	Tính điểm tứ phân vị của tập dữ liệu. Thường được dùng trong khảo sát dữ liệu để chia các tập hợp thành nhiều nhóm…
RANK *(number, ref, order)*	Tính thứ hạng của một số trong danh sách các số.
SKEW *(number1, number2, ...)*	Trả về độ lệch của phân phối, mô tả độ không đối xứng của phân phối quanh trị trung bình của nó.
SMALL *(array, k)* :	Trả về giá trị nhỏ nhất thứ k trong một tập số.
STDEV *(number1, number2, ...)*	Ước lượng độ lệch chuẩn trên cơ sở mẫu.
STDEVA *(value1, value2, ...)*	Ước lượng độ lệch chuẩn trên cơ sở mẫu, bao gồm cả những giá trị logic.
STDEVP *(number1, number2, ...)*	Tính độ lệch chuẩn theo toàn thể tập hợp.
STDEVPA *(value1, value2, ...)*	Tính độ lệch chuẩn theo toàn thể tập hợp, kể cả chữ và các giá trị logic.
VAR *(number1, number2, ...)*	Trả về phương sai dựa trên mẫu.
VARA *(value1, value2, …)*	Trả về phương sai dựa trên mẫu, bao gồm cả các trị logic và text.
VARP *(number1, number2, ...)*	Trả về phương sai dựa trên toàn thể tập hợp.
VARPA *(value1, value2, …)*	Trả về phương sai dựa trên toàn thể tập hợp, bao gồm cả các trị logic và text.
TRIMMEAN *(array, percent)*	Tính trung bình phần trong của một tập dữ liệu, bằng cách loại tỷ lệ phần trăm của các điểm dữ liệu ở đầu và ở cuối tập dữ liệu.

NHÓM HÀM VỀ ***PHÂN PHỐI XÁC SUẤT***

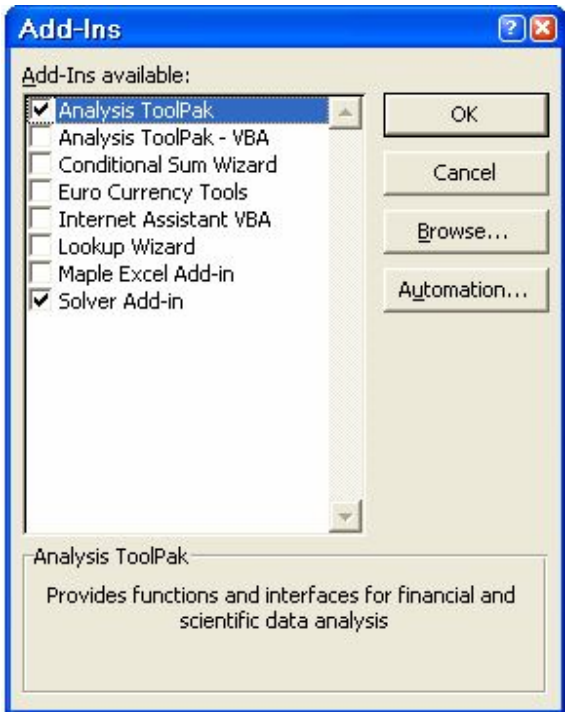
BETADIST *(x, alpha, beta, A, B)*	Trả về giá trị của hàm tính mật độ phân phối xác suất tích lũy *beta*.
BETAINV *(probability, alpha, beta, A, B)*	Trả về nghịch đảo của hàm tính mật độ phân phối xác suất tích lũy *beta*
BINOMDIST *(number_s, trials, probability_s, cumulative)*	Trả về xác suất của những lần thử thành công của phân phối nhị phân.
CHIDIST *(x, degrees_freedom)*	Trả về xác xuất một phía của phân phối *chi-squared*.
CHIINV *(probability, degrees_freedom)*	Trả về nghịch đảo của xác xuất một phía của phân phối *chi-squared*.
CHITEST *(actual_range, expected_range)*	Trả về giá trị của xác xuất từ phân phối *chi-squared* và số bậc tự do tương ứng.
CONFIDENCE *(alpha, standard_dev, size)*	Tính khoảng tin cậy cho một kỳ vọng lý thuyết
CRITBINOM *(trials, probability_s, alpha)*	Trả về giá trị nhỏ nhất sao cho phân phối nhị thức tích lũy lớn hơn hay bằng giá trị tiêu chuẩn. Thường dùng để bảo đảm các ứng dụng đạt chất lượng…
EXPONDIST *(x, lambda, cumulative)* :	Tính phân phối mũ. Thường dùng để mô phỏng thời gian giữa các biến cố…
FDIST *(x, degrees_freedom1, degrees_freedom2)*	Tính phân phối xác suất F. Thường dùng để tìm xem hai tập số liệu có nhiều mức độ khác nhau hay không…
FINV *(probability, degrees_freedom1, degrees_freedom2)*	Tính nghịch đảo của phân phối xác suất F. Thường dùng để so sánh độ biến thiên trong hai tập số liệu.
FTEST *(array1, array2)* :	Trả về kết quả của một phép thử F. Thường dùng để xác định xem hai mẫu có các phương sai khác nhau hay không…
FISHER *(x)*	Trả về phép biến đổi Fisher tại x. Thường dùng để kiểm tra giả thuyết dựa trên hệ số tương quan…
FISHERINV *(y)*	Tính nghịch đảo phép biến đổi Fisher. Thường dùng để phân tích mối tương quan giữa các mảng số liệu…
GAMMADIST *(x, alpha, beta, cumulative)*	Trả về phân phối tích lũy gamma. Có thể dùng để nghiên cứu có phân bố lệch.
GAMMAINV *(probability, alpha, beta)*	Trả về nghịch đảo của phân phối tích lũy gamma.
GAMMLN *(x)*	Tính logarit tự nhiên của hàm gamma.
HYPGEOMDIST *(number1, number2, ...)*	Trả về phân phối siêu bội (xác suất của một số lần thành công nào đó…)

LOGINV *(probability, mean, standard_dev)*	Tính nghịch đảo của hàm phân phối tích lũy lognormal của x (LOGNORMDIST)
LOGNORMDIST *(x, mean, standard_dev)*	Trả về phân phối tích lũy lognormal của x, trong đó logarit tự nhiên của x thường được phân phối với các tham số mean và standard_dev.
NEGBINOMDIST *(number_f, number_s, probability_s)*	Trả về phân phối nhị thức âm (trả về xác suất mà sẽ có number_f lần thất bại trước khi có number_s lần thành công, khi xác suất không đổi của một lần thành công là probability_s)
NORMDIST *(x, mean, standard_dev, cumulative)*	Trả về phân phối chuẩn (normal distribution). Thường được sử dụng trong việc thống kê, gồm cả việc kiểm tra giả thuyết.
NORMINV *(probability, mean, standard_dev)*	Tính nghịch đảo phân phối tích lũy chuẩn.
NORMSDIST *(z)*	Trả về hàm phân phối tích lũy chuẩn tắc (standard normal cumulative distribution function), là phân phối có trị trung bình cộng là zero (0) và độ lệch chuẩn là 1.
NORMSINV *(probability)*	Tính nghịch đảo của hàm phân phối tích lũy chuẩn tắc.
POISSON *(x, mean, cumulative)*	Trả về phân phối poisson. Thường dùng để ước tính số lượng biến cố sẽ xảy ra trong một khoảng thời gian nhất định.
PROB *(x_range, prob_range, lower_limit, upper_limit)*	Tính xác suất của các trị trong dãy nằm giữa hai giới hạn.
STANDARDIZE *(x, mean, standard_dev)*	Trả về trị chuẩn hóa từ phân phối biểu thị bởi mean và standard_dev.
TDIST *(x, degrees_freedom, tails)*	Trả về xác suất của phân phối Student (phân phối t), trong đó x là giá trị tính từ t và được dùng để tính xác suất.
TINV *(probability, degrees_freedom)*	Trả về giá trị t của phân phối Student.
TTEST *(array1, array2, tails, type)*	Tính xác xuất kết hợp với phép thử Student.
WEIBULL *(x, alpha, beta, cumulative)*	Trả về phân phối Weibull. Thường sử dụng trong phân tích độ tin cậy, như tính tuổi thọ trung bình của một thiết bị.
ZTEST *(array, x, sigma)*	Trả về xác suất một phía của phép thử z.

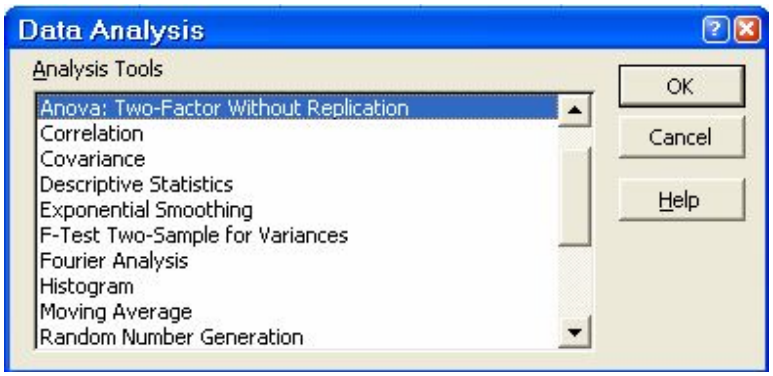
NHÓM HÀM VỀ TƯƠNG QUAN VÀ HỒI QUY TUYẾN TÍNH

CORREL *(array1, array2)*	Tính hệ số tương quan giữa hai mảng để xác định mối quan hệ của hai đặc tính.
COVAR *(array1, array2)*	Tính tích số các độ lệch của mỗi cặp điểm dữ liệu, rồi tính trung bình các tích số đó.
FORECAST *(x, known_y's, known_x's)*	Tính toán hay dự đoán một giá trị tương lai bằng cách sử dụng các giá trị hiện có, bằng phương pháp hồi quy tuyến tính.
GROWTH *(known_y's, known_x's, new_x's, const)*	Tính toán sự tăng trưởng dự kiến theo hàm mũ, bằng cách sử dụng các dữ kiện hiện có.
INTERCEPT *(known_y's, known_x's)*	Tìm điểm giao nhau của một đường thẳng với trục y bằng cách sử dụng các trị x và y cho trước
LINEST *(known_y's, known_x's, const, stats)*	Tính thống kê cho một đường bằng cách dùng phương pháp bình phương tối thiểu (least squares) để tính đường thẳng thích hợp nhất với dữ liệu, rồi trả về mảng mô tả đường thẳng đó. Luôn dùng hàm này ở dạng công thức mảng.
LOGEST *(known_y's, known_x's, const, stats)*	Dùng trong phân tích hồi quy. Hàm sẽ tính đường cong hàm mũ phù hợp với dữ liệu được cung cấp, rồi trả về mảng gía trị mô tả đường cong đó. Luôn dùng hàm này ở dạng công thức mảng.
PEARSON *(array1, array2)*	Tính hệ số tương quan momen tích pearson (r), một chỉ mục không thứ nguyên, trong khoảng từ -1 đến 1, phản ánh sự mở rộng quan hệ tuyến tính giữa hai tập số liệu.
RSQ *(known_y's, known_x's)*	Tính bình phương hệ số tương quan momen tích Pearson (r), thông qua các điểm dữ liệu trong known_y's và known_x's.
SLOPE *(known_y's, known_x's)*	Tính hệ số góc của đường hồi quy tuyến tính thông qua các điểm dữ liệu.
STEYX *(known_y's, known_x's)*	Trả về sai số chuẩn của trị dự đoán y đối với mỗi trị x trong hồi quy.
TREND *(known_y's, known_x's, new_x's, const)*	Trả về các trị theo xu thế tuyến tính

Ngoài cách dùng các hàm trên ta còn dùng menu Analysis ToolPak cài đặt như sau: Trong Excel chọn menu Tools/Add-Ins …/Analysis ToolPak / Ok



Khi chọn menu Tools / Data Analysis …



Chọn các mục cần thiết trong các thực đơn trên để giải các bài toán dưới đây:

I. THỐNG KÊ MÔ TẢ (Descriptive Statistics)

1) Bảng phân phối tần số - Bảng phân phối tần suất

- Nhập dữ liệu
- Dùng hàm: FREQUENCY *(data_array, bins_array)*
  - *data_array : Địa chỉ mảng dữ liệu*
  - *bins_array: Địa chỉ* mảng các giá trị khác nhau của dữ liệu.

Ví dụ : Lập bảng và vẽ biểu đồ dữ liệu sau:

12 13 11 13 15 12 11 10 14 13 12 15

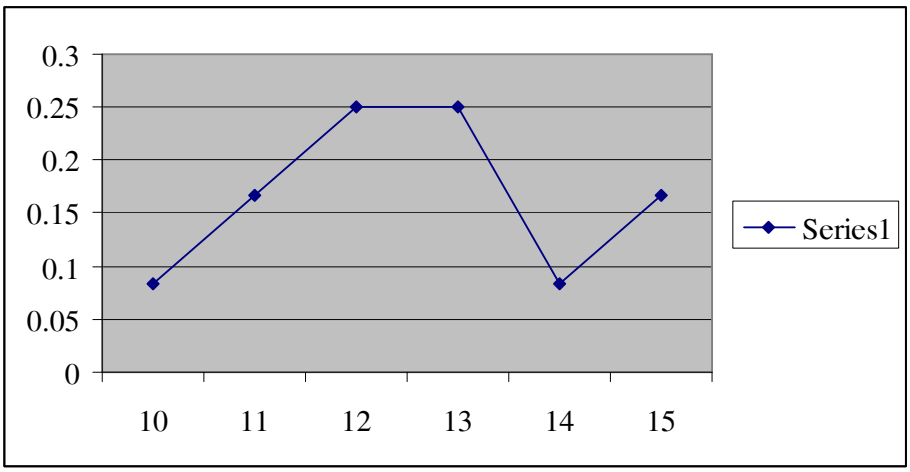
- Lập bảng phân phối tần số:
  - Nhập cột giá trị khác nhau vào C3:C8
  - Đánh dấu khối cột tần số ở D3:D8 , nhấn F2 nhập công thức

    = frequency(A2: A13 , C3:C8) và ấn CTRL+SHIFT +ENTER

- Lập bảng phân phối tần suất:nhập vào G2 công thức =D3/$D$9 ,copy các ô còn lại.

	A	B	C	D	E	F	G	H
1	Dữ liệu		Bảng phân phối tần số			Bảng phân phối tần suất		
2	12		**x**	**n**		**x**	**f**	
3	13		10	1		10	0.083333	
4	11		11	2		11	0.166667	
5	13		12	3		12	0.25	
6	15		13	3		13	0.25	
7	12		14	1		14	0.083333	
8	11		15	2		15	0.166667	
9	10			12			1	
10	14							
11	13							
12	12							
13	15							

- **Vẽ biểu đồ**
  - Chọn menu: Insert/ Chart…/ Line/ Next
  - Nhập vào Data Range : $G$3:$G$8 và chọn mục Column
  - Chọn Tab Series , nhập địa chỉ cột giá trị: $F$3:$F$8 vào Category (X) axis labels
  - Chọn Next , Finish

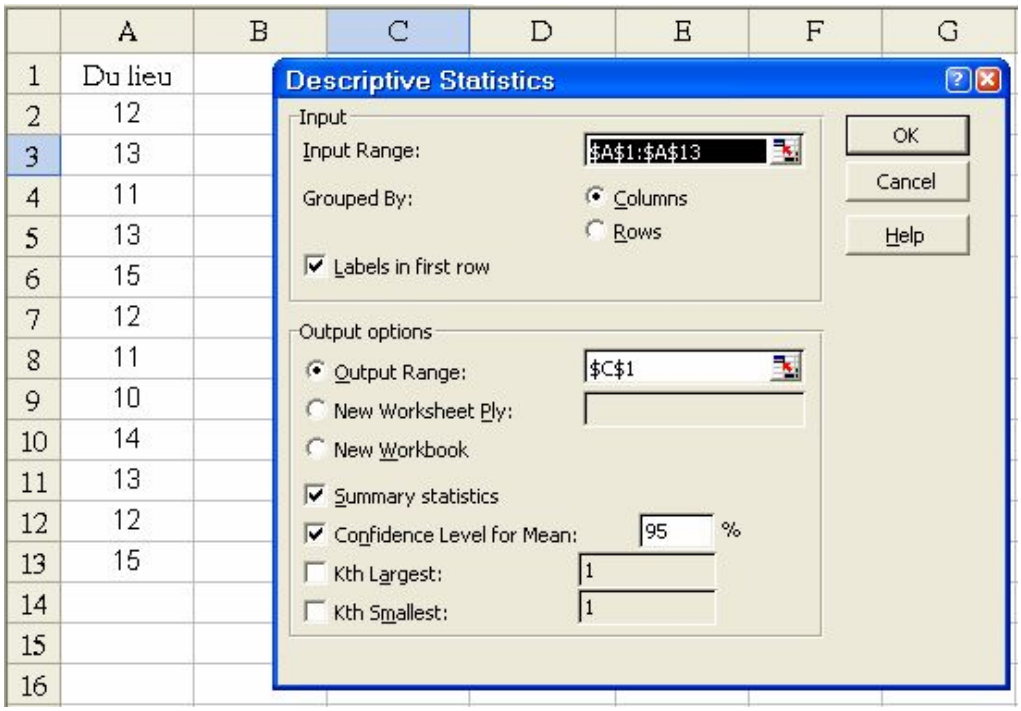


2) Đặc trưng mẫu

Ví dụ: Tính đặc trưng mẫu của dữ liệu sau:

12 13 11 13 15 12 11 10 14 13 12 15

- Nhập dữ liệu trong cột A1:A12
- Chọn menu Tools/Data Analysis…/Descriptive Statistics
- Nhập các mục:
  - Input Range: địa chỉ tuyệt đối chứa dữ liệu $A$1:$A$12
  - Output Range: địa chỉ xuất kết quả
  - Confidence Level for Mean (Độ tin cậy cho trung bình)



- Kết quả bao gồm: Kỳ vọng (trung bình), phương sai, trung vị, mode, độ lệch chuẩn, độ nhọn, độ nghiêng (hệ số bất đối xứng so với phân phối chuẩn), khoảng biến thiên, max, min, sum, số mẫu (count), khoảng tin cậy của trung bình ở mức 95% .

Column1	
Mean	$\bar{x} = 12.58333$
Standard Error	$\frac{S_x}{\sqrt{n}} = 0.451569$
Median	12.5
Mode	12
Standard Deviation	$s_x = 1.564279$
Sample Variance	2.44697
Kurtosis	-0.61768
Skewness	0.157146
Range	5
Minimum	10
Maximum	15
Sum	151
Count	n= 12
Confidence Level(95.0%)	$t_\alpha \frac{S_x}{\sqrt{n}} = 0.993896$

Tính theo các hàm	
Giá trị trung bình	AVERAGE(A1:A12)
Sai số mẫu	
Trung vị	MEDIAN(A1:A12)
Mode	MODE(A1:A12)
Độ lệch chuẩn	STDEV(A1:A12)
Phương sai mẫu	VAR(A1:A12)
Độ nhọn của đỉnh	KURT(A1:A12)
Độ nghiêng	SKEW(A1:A12)
Khoảng biến thiên	MAX()-MIN()
Tối thiểu	MIN(A1:A12)
Tối đa	MAX(A1:A12)
Tổng	SUM(A1:A12)
Số lượng mẫu	COUNT(A1:A12)
Độ chính xác	CONFIDENCE(0,05;S_x;n)

Chú ý : Khi mẫu lớn (n ≥ 30) ta thay $t_\alpha \frac{S_x}{\sqrt{n}}$ bằng $z_\alpha \frac{S_x}{\sqrt{n}}$ trong đó: $Z_\alpha = \text{NORMSINV}(1-\alpha/2)$

II. ƯỚC LƯỢNG THAM SỐ

Để ước lượng trung bình đám đông a ta thực hiện các bước sau:

- Nhập dữ liệu mẫu và xử lý mẫu bằng thống kê mô tả (Descriptive Statistics)
- Tính khoảng ước lượng trung bình a theo: $\bar{x} \pm z_\alpha \frac{S_x}{\sqrt{n}}$; $\bar{x} \pm t_\alpha \frac{S_x}{\sqrt{n}}$

Ví dụ: Khảo sát sức bền chịu lực của mộ loại ống công nghiệp người ta đo 9 ống và thu được các số liệu sau:

4500 6500 5000 5200 4800 4900 5125 6200 5375

	A	B	C	D	E	F	G	H
1								
2	TD1		*TD1*			**Khoảng ước lượng**		
3	4500					công thức	=B4 - B17	=B4 + B18
4	6500		Mean	5288.888889				
5	5000		Standard Error	218.4267894		kết quả	4785.195	5792.582
6	5200		Median	5125				
7	4800		Mode	#N/A				
8	4900		Standard Deviation	655.2803683				
9	5125		Sample Variance	429392.3611				
10	6200		Kurtosis	0.245488664				
11	5375		Skewness	1.049588691				
12			Range	2000				
13			Minimum	4500				
14			Maximum	6500				
15			Sum	47600				
16			Count	9				
17			Confidence Level(95.0%)	503.6934054				
18								

Ví dụ: Tiến hành xem trong một tháng trung bình một sinh viên tiêu hết bao nhiêu tiền gọi điện thoại. Khảo sát ngẫu nhiên 59 sinh viên thu được kết quả:

14	*18*	*22*	*30*	*36*	*28*	*42*	*79*	*36*	*52*	*15*	*47*
95	*16*	*27*	*111*	*37*	*63*	*127*	*23*	*31*	*70*	*27*	*11*
30	*147*	*72*	*37*	*25*	*7*	*33*	*29*	*35*	*41*	*48*	*15*
29	*73*	*26*	*15*	*26*	*31*	*57*	*40*	*18*	*85*	*28*	*32*
22	*36*	*60*	*41*	*35*	*26*	*20*	*58*	*33*	*23*	*35*	

Hãy ước lượng khoảng tin cậy của số tiền gọi điện thoại trung bình hàng tháng của một sinh viên với độ tin cậy 95%.

Đs 33.96481 48.23858

III. KIỂM ĐỊNH GIẢ THIẾT

1) So sánh 2 trung bình với phương sai đã biết hay mẫu lớn (n≥30)

- Dùng menu: *Tools/ Data Analysis... / z-test:Two Sample for Means*
- Tiêu chuẩn kiểm định: $z=\dfrac{\overline{x_1}-\overline{x_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}}$
- Phân vị 2 phía $z_{\alpha/2}$ là: *z Critical two-tail*
- Nếu $|z| > z_{\alpha/2}$ thì bác bỏ H_0 , chấp nhận H_1
  Nếu $|z| \le z_{\alpha/2}$ thì chấp nhận H_0 , bác bỏ H_1

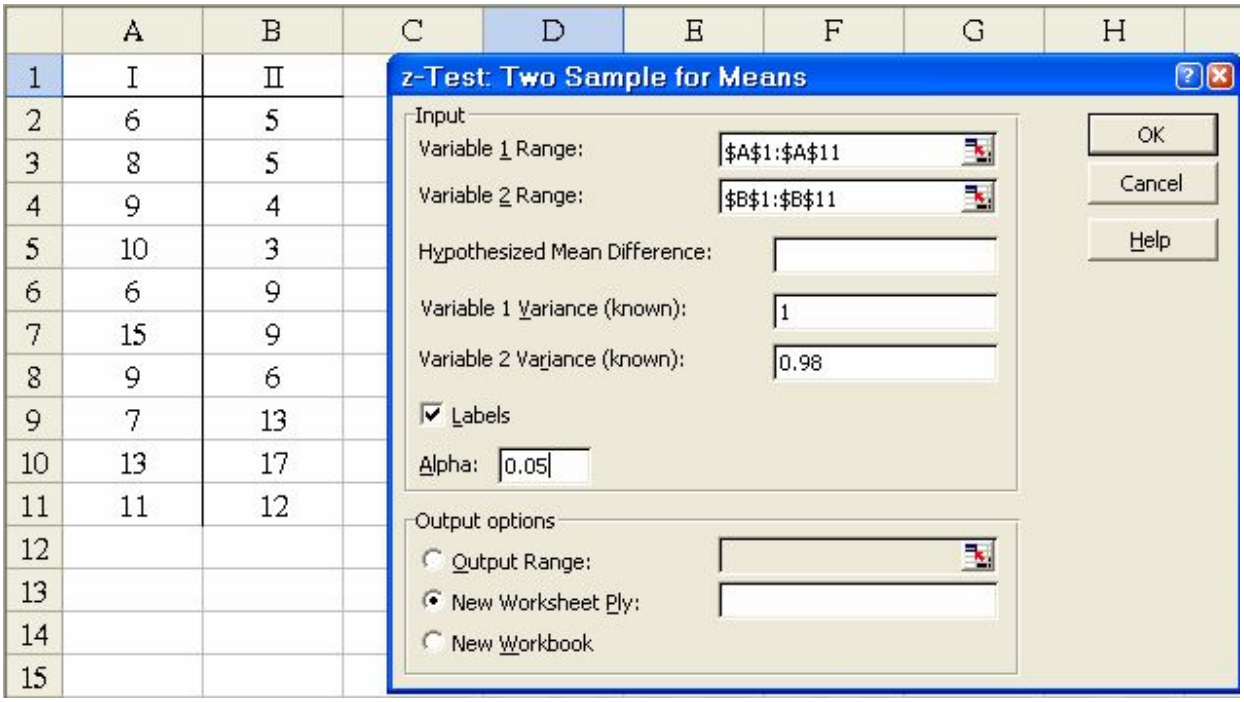
Ví dụ: *Người ta chọn 2 mẫu, mỗi mẫu 10 máy, từ hai lô (I và II được sản xuất với phương sai biết trước tương ứng là 1 và 0,98) để khảo sát thời gian hoàn thành công việc (phút) của chúng:*

I	*6*	*8*	*9*	*10*	*6*	*15*	*9*	*7*	*13*	*11*
II	*5*	*5*	*4*	*3*	*9*	*9*	*6*	*13*	*17*	*12*

Hỏi khả năng hoàn thành công việc của hai máy có khác nhau hay không? α=0,05

Nhập và xử lý dữ liệu

- Variable 1 Range , Variable 2 Range: địa chỉ tuyệt đối của vùng dữ liệu của I, II
- Variable 1 Variance(known), Variable 2 Variance(known): phương sai của I,II
- Labels: chọn khi có tên biến ở đầu cột hoặc hàng
- Alpha : mức ý nghĩa α
- Output options: chọn cách xuất kết quả



	A	B
1	I	II
2	6	5
3	8	5
4	9	4
5	10	3
6	6	9
7	15	9
8	9	6
9	7	13
10	13	17
11	11	12

Kết quả:

H_0: $a_1 = a_2$ *"Khả năng hoàn thành công việc của 2 máy như nhau"*
H_1: $a_1 \neq a_2$ *"Khả năng hoàn thành công việc của 2 máy khác nhau"*

	I	II	
Mean	9.4	8.3	← Trung bình mẫu
Known Variance	1	0.98	← phương sai mẫu đã biết
Observations	10	10	← số quan sát (cỡ mẫu)
Hypothesized Mean Difference	0		
z	2.472066162		← Tiêu chuẩn kiểm định
P(Z<=z) one-tail	0.006716741		← Xác suất 1 phía
z Critical one-tail	1.644853476		← phân vị 1 phía
P(Z<=z) two-tail	0.013433483		← Xác suất 2 phía
z Critical two-tail	1.959962787		← phân vị 2 phía

$\Rightarrow$ $|z| = 2.472066162 > z_{\alpha/2} = 1.959962787$ nên bác bỏ H_0 , chấp nhận H_1
Vậy: "Khả năng hoàn thành công việc của 2 máy khác nhau"

2) So sánh 2 trung bình với dữ liệu từng cặp

- Được dùng khi mẫu bé, phụ thuộc, phương sai 2 mẫu không bằng nhau và mỗi phần tử khảo sát có 2 chỉ tiêu X (trước), Y (sau) khi thay đổi điều kiện thí nghiệm.
- Chọn menu: *Tools/Data Analysis.../ t-test:Paired Two Sample for Means*
- Tiêu chuẩn kiểm định: $t = \dfrac{\overline{D}}{S_D\sqrt{n}}$, $\overline{D} = \dfrac{\sum_{i=1}^{n}(X_i - Y_i)}{n}$, $S_D = \sqrt{\dfrac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1}}$
- Phân vị 2 phía $t_{\alpha/2}$ là: *t Critical two-tail*
- Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0 , chấp nhận H_1
  Nếu $|t| \leq t_{\alpha/2}$ thì chấp nhận H_0 , bác bỏ H_1

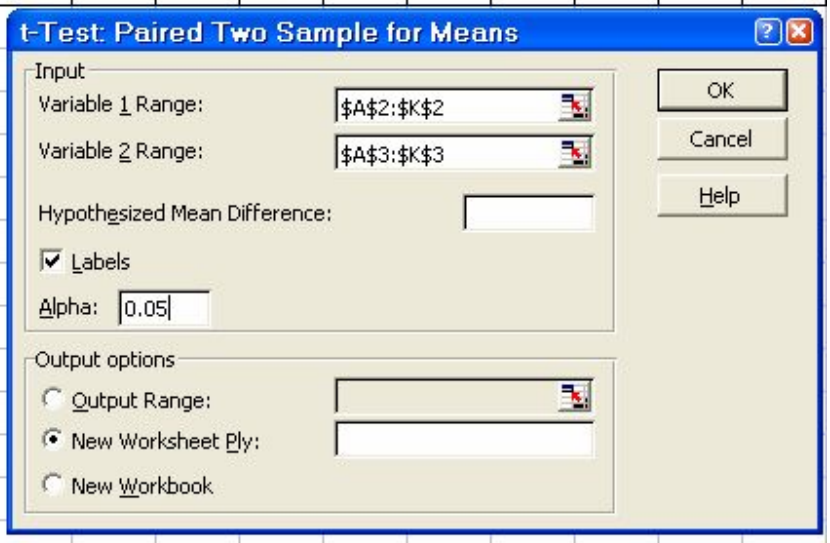
Ví dụ: Để nghiên cứu của một loại thuốc ngủ, người ta cho 10 bệnh nhân uống thuốc. Lần khác họ cũng cho bệnh nhân uống thuốc nhưng là thuốc giả (thuốc không có tác dụng). Kết quả thí nghiệm như sau:

Bệnh nhân	1	2	3	4	5	6	7	8	9	10
Số giờ ngủ có thuốc	6,1	7,0	8,2	7,6	6,5	8,4	6,9	6,7	7,4	5,8
Số giờ ngủ với thuốc giả	5,2	7,9	3,9	4,7	5,3	5,4	4,2	6,1	3,8	6,3

Giả sử số giờ ngủ của các bệnh nhân có qui luật chuẩn. Với mức ý nghĩa $\alpha = 0{,}05$ hãy kết luận về ảnh hưởng của loại thuốc ngủ trên?

- **Nhập và xử lý dữ liệu**

	A	B	C	D	E	F	G	H	I	J	K
1	Bệnh nhân	1	2	3	4	5	6	7	8	9	10
2	Số giờ ngủ có thuốc	6.1	7	8.2	7.6	6.5	8.4	6.9	6.7	7.4	5.8
3	Số giờ ngủ với thuốc giả	5.2	7.9	3.9	4.7	5.3	5.4	4.2	6.1	3.8	6.3



- **Kết quả**

H_0: $a_1 = a_2$ "Thuốc ngủ trên không có tác dụng đến số giờ ngủ"
H_1: $a_1 \neq a_2$ "Thuốc ngủ trên có tác dụng đến số giờ ngủ"

t-Test: Paired Two Sample for Means

	Số giờ ngủ có thuốc	Số giờ ngủ với thuốc giả
Mean	7.06	5.28
Variance	0.720444444	1.577333333
Observations	10	10
Pearson Correlation	-0.388571913	
Hypothesized Mean Difference	0	
df	9	
t Stat	3.183538302	
P(T<=t) one-tail	0.005560693	
t Critical one-tail	1.833113856	
P(T<=t) two-tail	0.011121385	
t Critical two-tail	2.262158887	

$\Rightarrow |t| = 3{,}1835 > t_{\alpha/2} = 2{,}2622$ nên chấp nhận H_1
Vậy loại thuốc ngủ trên có ảnh hưởng làm tăng số giờ ngủ trung bình.

3) So sánh 2 trung bình với phương sai bằng nhau

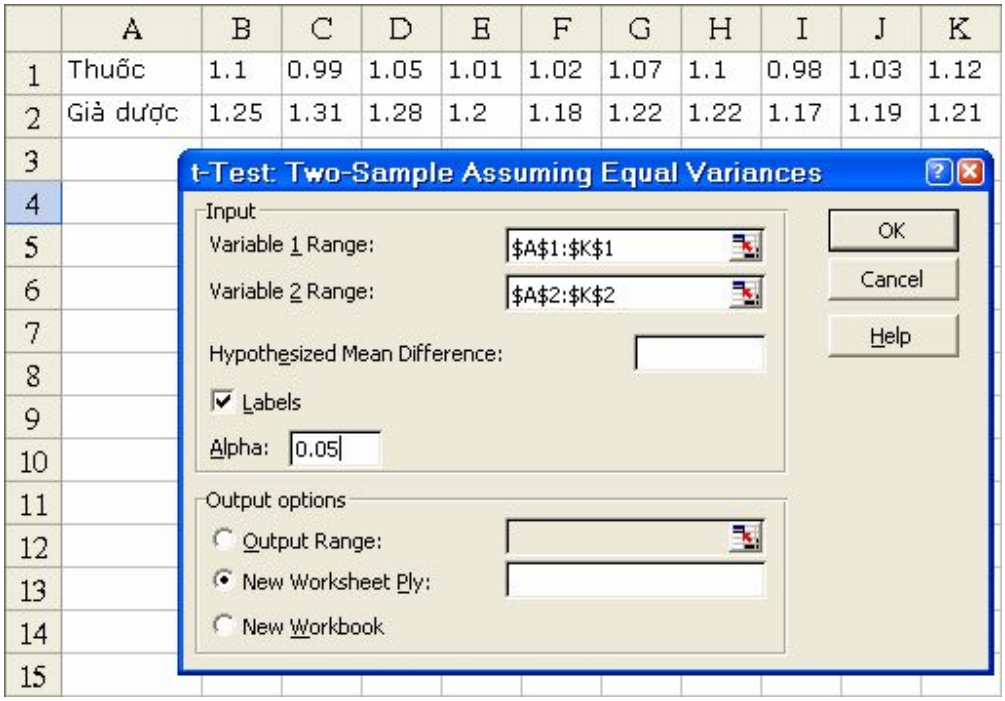
- Được dùng khi 2 mẫu bé , độc lập và phương sai 2 mẫu bằng nhau.
- Chọn menu: *Tools/Data Analysis…/ t-test:Two-Sample Assuming Equal Variances*
- Tiêu chuẩn kiểm định: $t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$, $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
- Phân vị 2 phía $t_{\alpha/2}$ là: *t Critical two-tail*
- Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0 , chấp nhận H_1
  Nếu $|t| \le t_{\alpha/2}$ thì chấp nhận H_0 , bác bỏ H_1

Ví dụ: Người ta cho 10 bệnh nhân uống thuốc hạ cholesterol đồng thời cho 10 bệnh nhân khác uống giả dược, rồi xét nghiệm về nồng độ cholesterol trong máu (g/l)của cả 2 nhóm:

Thuốc	1,10	0,99	1,05	1,01	1,02	1,07	1,10	0,98	1,03	1,12
Giả dược	1,25	1,31	1,28	1,20	1,18	1,22	1,22	1,17	1,19	1,21

Với α=0,05 hãy cho biết thuốc có tác dụng hạ cholesterol trong máu không?

- **Nhập và xử lý dữ liệu**



- **Kết quả**

  H_0: $a_1=a_2$ "Thuốc và giả dược có tác dụng như nhau"
  H_1: $a_1<a_2$ "Thuốc có tác dụng hạ cholesterol trong máu"

t-Test: Two-Sample Assuming Equal Variances

	Thuốc	Giả dược
Mean	1.047	1.223
Variance	0.002401111	0.002001111
Observations	10	10
Pooled Variance	0.002201111	
Hypothesized Mean Difference	0	
df	18	
t Stat	-8.388352782	
P(T<=t) one-tail	6.19807E-08	
t Critical one-tail	1.734063062	
P(T<=t) two-tail	1.23961E-07	
t Critical two-tail	2.100923666	

$\Rightarrow$ t= -8,3884 < $-t_{\alpha}$= -1,7341 nên chấp nhận H_1
Vậy thuốc trên có tác dụng hạ cholesterol trong máu.

4) So sánh 2 trung bình với phương sai khác nhau

- Được dùng khi mẫu bé , độc lập và có phương sai khác nhau (2 mẫu phân biệt)
- Chọnmenu: *Tools/Data Analysis…/ t-test:Two-Sample Assuming Equal Variances*
- Tiêu chuẩn kiểm định: $t=\frac{\overline{X_1}-\overline{X_2}}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$
- Phân vị 2 phía $t_{\alpha/2}$ là: *t Critical two-tail*
- Nếu $|t| > t_{\alpha/2}$ thì bác bỏ H_0 , chấp nhận H_1
  Nếu $|t| \le t_{\alpha/2}$ thì chấp nhận H_0 , bác bỏ H_1

Ví dụ: Thời gian tan rã (phút) của một loại viên bao từ 2 xí nghiệp dược phẩm (XNDP) khác nhau được kiểm nghiệm như sau:

XNDP I	61	71	68	73	71	70	69	74
XNDP II	62	69	65	65	70	71	68	73

Thời gian tan rã của viên bao thuộc hai XNDP có giống nhau không?

- **Nhập, xử lý dữ liệu và kết quả**

H_0 : $a_1=a_2$ "Thời gian tan rã của viên bao 2 XNDP như nhau"
H_1 : $a_1 \neq a_2$ "Thời gian tan rã của viên bao 2 XNDP khác nhau"

	XNDP I	XNDP II
Mean	69.625	67.875
Variance	15.98214286	13.26785714
Observations	8	8
Hypothesized Mean Difference	0	
df	14	
t Stat	0.915208631	
P(T<=t) one-tail	0.187788433	
t Critical one-tail	1.76130925	
P(T<=t) two-tail	0.375576865	
t Critical two-tail	2.144788596	

$\Rightarrow |t| = 0{,}9152 \le 2{,}1448$ nên chấp nhận H_0
Vậy thời gian tan rã của viên bao thuộc 2 XNDP như nhau.

5) So sánh 2 tỉ số

- Đối với thí nghiệm có 2 kết quả, để so sánh 2 tỉ số của 2 kết quả đó, ta dùng kiểm định χ^2 (chi-quared) : $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-np_i)^2}{np_i}$, $np_i = \frac{\text{tổng hàng x tổng cột}}{n}$
  n_{ij}: tần số thực nghiệm, np_{ij}: tần số lý thuyết của ô (i,j) ; r : số hàng ; c : số cột
- Dùng hàm CHITEST(actual_range , expected_range).
  Tính giá trị: $P(X>\chi^2)$ =CHITEST
- Nếu $P(X>\chi^2) > \alpha$ thì chấp nhận H_0 và ngược lại.

Ví dụ: Kết quả điều trị trên 2 nhóm bệnh nhân: một nhóm dùng thuốc và một nhóm dùng giả dược được tóm tắt như sau:

Điều trị	*Số khỏi bệnh*	*Số không khỏi bệnh*
Thuốc	24	15
Giả dược	20	23

Tỉ lệ khỏi bệnh do thuốc và do giả dược có khác nhau không?

- **Nhập và xử lý dữ liệu**

	A	B	C	D	E
1	Thực nghiệm				
2		Điều trị	Khỏi bệnh	Không khỏi	Tổng hàng
3		Thuốc	24	15	=SUM(C3:D3)
4		Giả dược	20	23	=SUM(C4:D4)
5		Tổng cột	=SUM(C3:C4)	=SUM(D3:D4)	=SUM(E3:E4)
6	Lý thuyết				
7		Thuốc	=E3*C5/E5	=E3*D5/E5	
8		Giả dược	=E4*C5/E5	=E4*D5/E5	
9	Giá trị P:	=CHITEST(C3:D4,C7:D8)			
10					

- **Kết quả**

	A	B	C	D	E	F
1	Thực nghiệm					
2		Điều trị	Khỏi bệnh	Không khỏi	Tổng hàng	
3		Thuốc	24	15	39	
4		Giả dược	20	23	43	
5		Tổng cột	44	38	82	
6	Lý thuyết					
7		Thuốc	20.92682927	18.07317073		
8		Giả dược	23.07317073	19.92682927		
9	Giá trị P:	0.172954847				
10						

$\Rightarrow P(X>\chi^2) = 0{,}17295 > \alpha = 0{,}05$, nên chấp nhận H_0

Vậy tỷ lệ khỏi bệnh do thuốc và do giả dược không khác nhau.

6. So sánh 2 phương sai

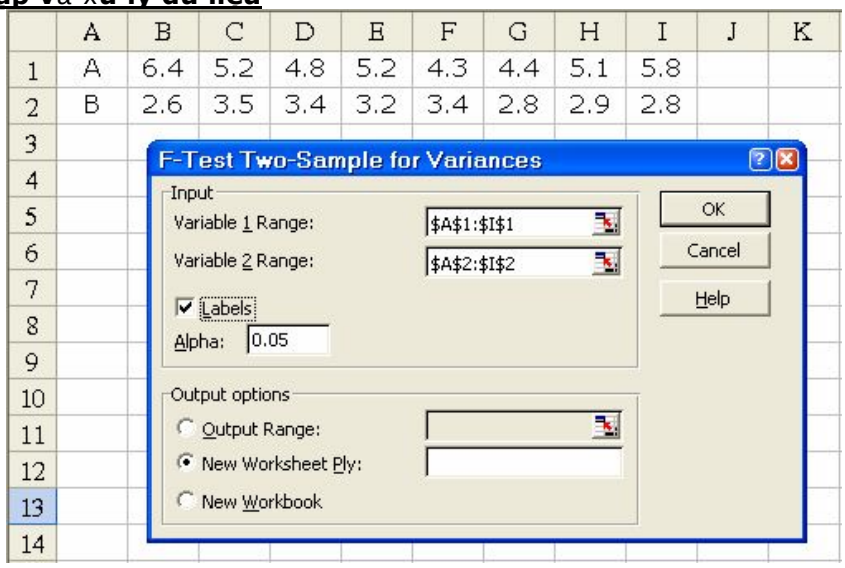
- So sánh 2 phương sai được áp dụng để so sánh độ chính xác của 2 phương pháp định lượng khác nhau.
- Chọn menu: *Tools/Data Analysis…/F-Test Two-Samplefor Variances*
- Tính tiêu chuẩn kiểm định $F=\frac{s_1^2}{s_2^2}$
- Nếu $F < F_\alpha$ thì chấp nhận H_0: $\sigma_1^2 = \sigma_2^2$ và ngược lại.

Ví dụ: Một được phân tích bởi hai phương pháp A và B với kết quả sau:

A	6,4	5,2	4,8	5,2	4,3	4,4	5,1	5,8
B	2,6	3,5	3,4	3,2	3,4	2,8	2,9	2,8

Cho biết phương pháp nào chính xác hơn?

- **Nhập và xử lý dữ liệu**



- **Kết quả**

  H_0: $\sigma_A^2 = \sigma_B^2$ "Hai phương pháp có độ chính xác như nhau"

  H_1: $\sigma_A^2 > \sigma_B^2$ "Độ chính xác của phương pháp B cao hơn"

F-Test Two-Sample for Variances

	A	*B*
Mean	5.15	3.075
Variance	0.485714286	0.116428571
Observations	8	8
df	7	7
F	4.171779141	
P(F<=f) one-tail	0.039514317	
F Critical one-tail	3.787050673	

$\Rightarrow$ F= 4,1718 > 3,7870 nên chấp nhận H_1

Vậy phương pháp B chính xác hơn phương pháp A.

IV. 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 , … , 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 : $a_1 = a_2 = \ldots = a_k$

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: $\bar{x} = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_i} x_{ij} = \frac{T}{n}$ 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		

Ví dụ:

Hàm lượng Alcaloid (mg) trong một loại dược liệu được thu hái từ 3 vùng khác nhau được số liệu sau:

Vùng 1 : 7,5 6,8 7,1 7,5 6,8 6,6 7,8

Vùng 2 : 5,8 5,6 6,1 6,0 5,7

Vùng 3 : 6,1 6,3 6,5 6,4 6,5 6,3

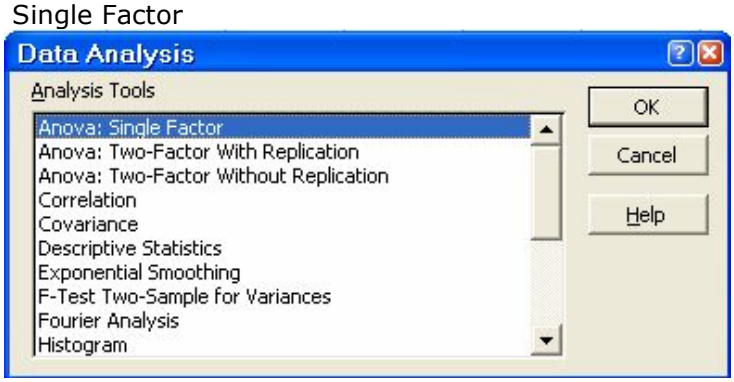
Hỏi hàm lượng Alcaloid có khác nhau theo vùng hay không?

Dùng Excel

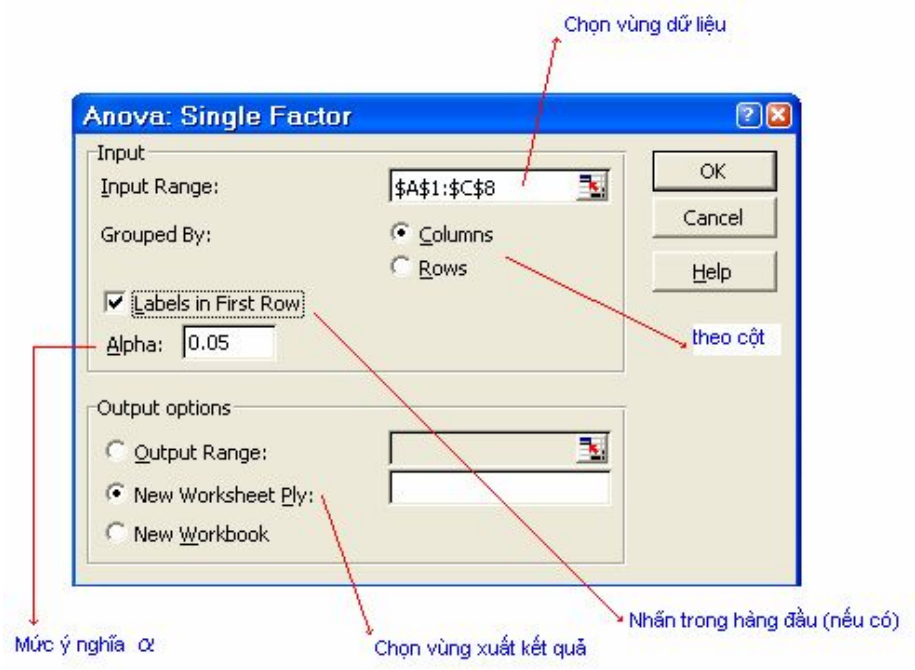
1. Nhập dữ liệu theo cột

	A	B	C
1	Vùng 1	Vùng 2	Vùng 3
2	7.5	5.8	6.1
3	6.8	5.6	6.3
4	7.1	6.1	6.5
5	7.5	6.0	6.4
6	6.8	5.7	6.5
7	6.6		6.3
8	7.8		
9			

2. Chọn mục : Anova: Single Factor



3. Chọn các mục như hình:



4. Kết quả

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Vùng 1	7	50.1	7.157143	0.202857
Vùng 2	5	29.2	5.84	0.043
Vùng 3	6	38.1	6.35	0.023

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	5.326968	2	2.663484	26.56148	1.17756E-05	3.682316674
Within Groups	1.504143	15	0.100276			
Total	6.831111	17				

$\Rightarrow$ F= 26,5615 > $F_{k-1;\ n-k\ ;\ 1-\alpha}$ =3,6823 nên bác bỏ H_0 chấp nhận H_1.

Vậy hàm lượng Alcaloid có sai khác theo vùng.

Bài tập

1. So sánh 3 loại thuốc bổ A, B, C trên 3 nhóm, người ta được kết quả tăng trọng(kg) như sau:

   A: 1,0 1,2 1,4 1,1 0,8 0,6

   B: 2,0 1,8 1,9 1,2 1,4 1,0 1,5 1,8

   C: 0,4 0,6 0,7 0,2 0,3 0,1 0,2

   Hãy so sánh kết quả tăng trọng của 3 loại thuốc bổ trên với α = 0,01

2. Một nghiên cứu được thực hiện nhằm xem xét năng suất lúa trung bình của 3 giống lúa. Kết quả thu thập qua 4 năm như sau:

Năm	A	B	C
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

   Hãy cho biết năng suất lúa trung bình của 3 giống lúa có khác nhau hay không? α=0,01

3. So sánh hiệu quả giảm đau của 4 loại thuốc A, B, C, D bằng cách chia 20 bệnh nhân thành 4 nhóm, mỗi nhóm dùng một loại thuốc giảm đau trên. Kết quả mức độ giảm đau là:

   A: 82 89 77 72 92

   B: 80 70 72 90 68

   C: 77 69 67 65 57

   D: 65 75 67 55 63

   Hỏi hiệu quả giảm đau của 4 loại thuốc có khác nhau không?

   Nếu hiệu quả giảm đau của 4 loại thuốc A, B, C, D khác nhau có ý nghĩa, hãy so sánh từng cặp thuốc với α = 0,05

2. Phân tích phương sai 2 nhân tố không lặp

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

Giả sử nhân tố A có n mức a_1 , a_2 , … , a_n (nhân tố hàng)

B có m mức b_1 , b_2 , … , 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
- Không có sự tương tác giữa nhân tố cột và hàng

* 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 $F_A > F_{n-1\ ;\ (n-1)(m-1)\ ;\ 1-\alpha}$ thì bác bỏ yếu tố A (hàng)
- Nếu $F_B > F_{m-1\ ;\ (n-1)(m-1)\ ;\ 1-\alpha}$ thì bác bỏ yếu tố B (cột)

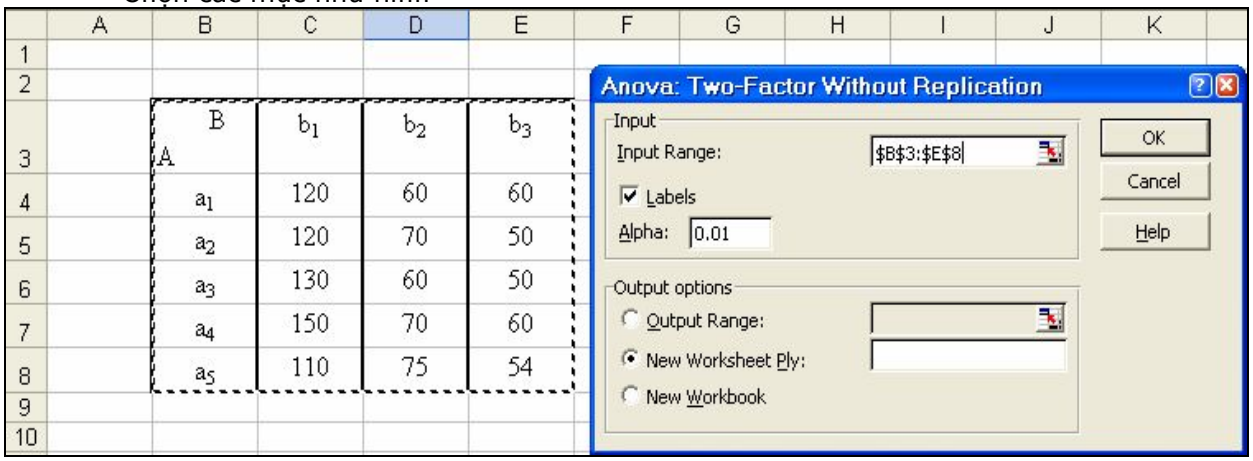
Ví dụ:

Chiết suất chất X từ 1 loại dược liệu bằng 3 phương pháp và 5 loại dung môi, ta có kết quả:

PP Chiết suất (B) / Dung môi (A)	b_1	b_2	b_3
a_1	120	60	60
a_2	120	70	50
a_3	130	60	50
a_4	150	70	60
a_5	110	75	54

Hãy xét ảnh hưởng của phương pháp chiết suất và dung môi đến kết quả chiết suất chất X với α=0,01.

- Giả thiết H_0 : * Trung bình của 3 phương pháp chiết suất bằng nhau
  * Trung bình của 5 dung môi bằng nhau
  * Không có sự tương tác giữa phương pháp chiế suất và dung môi
- Chọn Tools\Data Analysis…\Anova: Two-Factor without replication
- Chọn các mục như hình



- Kết quả

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
a1	3	240	80	1200
a2	3	240	80	1300
a3	3	240	80	1900
a4	3	280	93.33333333	2433.333333
a5	3	239	79.66666667	800.3333333
b1	5	630	126	230
b2	5	335	67	45
b3	5	274	54.8	25.2

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	432.2666667	4	108.0666667	1.124913255	0.409397603	7.006065061
Columns	14498.8	2	7249.4	75.46217904	6.42093E-06	8.64906724
Error	768.5333333	8	96.06666667			
Total	15699.6	14				

$\Rightarrow F_A < F_{4\ ;\ 8\ ;\ 0,99} = 7,006 \Rightarrow$ Dung môi không ảnh hưởng đến kết quả chiết suất.
$F_B > F_{2\ ;\ 8\ ;\ 0,99} = 8,649 \Rightarrow$ Phương pháp ảnh hưởng đến kết quả chiết suất.

Bài tập

1) Nghiên cứu về hiệu quả của 3 loại thuốc A, B, C dùng điều trị chứng suy nhược thần kinh. 12 người bệnh được chia làm 4 nhóm theo mức độ bệnh 1 , 2 , 3 , 4 ; trong mỗi nhóm chia ra để cùng dùng 1 trong 3 loại thuốc trên. Sau 1 tuần điều trị, kết quả đánh giá bằng thang điểm như sau:

Mức độ bệnh / Thuốc	1	2	3	4
A	25	40	25	30
B	30	25	25	25
C	25	20	20	25

Hãy đánh giá hiệu quả của các loại thuốc A, B, C có khác nhau hay không ? với $\alpha = 0,01$

2) Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa đến năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau:

Giống lúa / Loại phân bón	A	B	C
1	65	69	75
2	74	72	70
3	64	68	78
4	83	78	76

Hãy đánh giá sự ảnh hưởng giống lúa, loại phân bón trên năng suất lúa, $\alpha = 0,05$.

3) Để khảo sát ảnh hưởng của 4 loại thuốc trừ sâu (1, 2, 3 và 4) và ba loại giống (B1, B2 và B3) đến sản lượng của cam, các nhà nghiên cứu tiến hành một thí nghiệm loại giai thừa. Trong thí nghiệm này, mỗi giống cam có 4 cây cam được chọn một cách ngẫu nhiên, và 4 loại thuốc trừ sâu áp dụng (cũng ngẫu nhiên) cho mỗi cây cam.
Kết quả nghiên cứu (sản lượng cam) cho từng giống và thuốc trừ sâu như sau:

Thuốc trừ sâu / Giống Cam	1	2	3	4
B1	29	50	43	53
B2	41	58	42	73
B3	66	85	63	85

Hãy cho biết thuốc trừ sâu, giống cam có ảnh h ưởng đến sản lượng cam không? α = 0,05

4) 4 chuyên gia tài chính được yêu cầu dự đoán về tốc độ tăng trưởng (%) trong năm tới của 5 công ty trong ngành nhựa. Dự đoán được ghi nhận như sau:

Công ty	Chuyên gia			
	A	B	C	D
1	8	12	8,5	13
2	14	10	9	11
3	11	9	12	10
4	9	13	10	13
5	12	10	10	10

Hãy lập bảng ANOVA. Có thể nói rằng dự đoán tốc độ tăng trưởng trung bình là như nhau cho cả 5 công ty nhựa được không?

3. Phân tích phương sai 2 nhân tố có lặp

Tương tự như bài toán phân tích phương sai 2 nhân tố không lặp, chỉ khác mỗi mức $((a_i , b_j)$ đều có sự lặp lại r lần thí nghiệm và ta cần khảo sát thêm sự tương tác (interaction term) F_{AB} giữa 2 nhân tố A và B.

* Mẫu điều tra:

A \ B	b_1	b_2	…	b_m
a_1	x_{111} x_{112} : : x_{11r}	x_{121} x_{122} : : x_{12r}	…	x_{1m1} x_{1m2} : : x_{1mr}
a_2	x_{211} x_{212} : : x_{21r}	x_{221} x_{222} : : x_{22r}	…	x_{2m1} x_{2m2} : : x_{2mr}
:	:	:		:
:	:	:		:
a_n	x_{n11} x_{n12} : : x_{n1r}	x_{n21} x_{n22} : : x_{n2r}	…	x_{nm1} x_{nm2} : : x_{nmr}

* Xử lý mẫu: Tính tổng hàng $T_{i**} = \sum_{j,k} x_{ijk}$, tổng cột $T_{*j*} = \sum_{i,k} x_{ijk}$

A \ B	b_1	b_2	…	b_m	$Ti**$
a_1	x_{111} x_{112} : : x_{11r}	x_{121} x_{122} : : x_{12r}	…	x_{1m1} x_{1m2} : : x_{1mr}	$T_{1**} = \sum_{j,k} x_{1jk}$
a_2	x_{211} x_{212} : : x_{21r}	x_{221} x_{222} : : x_{22r}	…	x_{2m1} x_{2m2} : : x_{2mr}	$T_{2**} = \sum_{j,k} x_{2jk}$
:	:	:		:	
:	:	:		:	
a_n	x_{n11} x_{n12} : : x_{n1r}	x_{n21} x_{n22} : : x_{n2r}	…	x_{nm1} x_{nm2} : : x_{nmr}	$T_{n**} = \sum_{j,k} x_{njk}$
T_{*j*}	$T_{*1*} = \sum_{i,k} x_{i1k}$	$T_{*2*} = \sum_{i,k} x_{i2k}$		$T_{*m*} = \sum_{i,k} x_{imk}$	$T = \sum_{i,j,k} x_{ijk}$

Cần tính: $\sum_{i,j,k} x_{ijk}^2$ $\sum_{i} T_{i**}^2$ $\sum_{j} T_{*j*}^2$ $\sum_{i,j} T_{ij*}^2$

Suy ra

$$SST = \sum_{i,j,k}(x_{ijk} - \bar{x})^2 = \sum_{i,j,k} x_{ijk}^2 - \frac{T^2}{nmr}$$

$$SSA = mr\sum_{i}(\overline{x_{i**}} - \bar{x})^2 = \frac{\sum_i T_{i**}^2}{mr} - \frac{T^2}{nmr}$$

$$SSB = nr\sum_{j}(\overline{x_{*j*}} - \bar{x})^2 = \frac{\sum_j T_{*j*}^2}{nr} - \frac{T^2}{nmr}$$

$$SSAB = r\sum_{j,i}(\overline{x_{ij*}} - \overline{x_{i**}} - \overline{x_{*j*}} + \bar{x})^2 = \frac{\sum_{i,j} T_{ij*}^2}{r} - \frac{\sum_j T_{*j*}^2}{nr} - \frac{\sum_i T_{i**}^2}{mr} + \frac{T^2}{nmr}$$

$$SSE = SST - SSA - SSB - SSAB = \sum_{i,j,k} x_{ijk}^2 - \frac{\sum_{i,j} x_{ij*}^2}{r}$$

* **Bảng ANOVA**

Nguồn	SS	df	MS	F
Yếu tố A	SSA	n-1	$MSA = \frac{SSA}{n-1}$	$F_A = \frac{MSA}{MSE}$
Yếu tố B	SSB	m-1	$MSB = \frac{SSB}{m-1}$	$F_B = \frac{MSB}{MSE}$
Tương tác AB	SSAB	(n-1)(m-1)	$MSAB = \frac{SSAB}{(n-1)(m-1)}$	$F_{AB} = \frac{MSAB}{MSE}$
Sai số	SSE	nm(r-1)	$MSE = \frac{SSE}{nm(r-1)}$	
Tổng	SST	nmr-1		

* Kết luận:

- Nếu $F_A > F_{n-1\ ;\ nm(r-1)\ ;\ 1-\alpha}$ thì bác bỏ yếu tố A (h àng)
- Nếu $F_B > F_{m-1\ ;\ nm(r-1)\ ;\ 1-\alpha}$ thì bác bỏ yếu tố B (cột)
- Nếu $F_{AB} > F_{(n-1)(m-1)\ ;\ nm(r-1)\ ;\ 1-\alpha}$ thì có sự tương tác giữa A và B

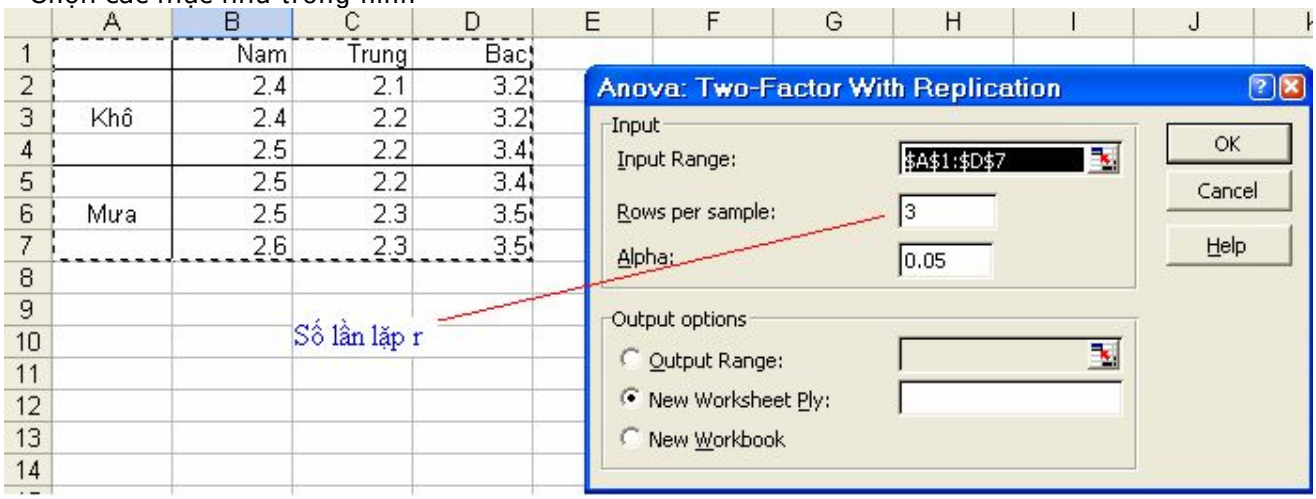
Ví dụ: Hàm lượng saponin (mg) của cùng một loại dược liệu được thu hái trong 2 mùa (khô và mưa: trong mỗi mùa lấy mẫu 3 lần - đầu mùa, giữa mùa, cuối mùa) và từ 3 miền (Nam, Trung, Bắc) thu được kết quả sau:

Mùa	Thời điểm	Miền		
		Nam	Trung	Bắc
Khô	Đầu mùa	2,4	2,1	3,2
	Giữa mùa	2,4	2,2	3,2
	Cuối mùa	2,5	2,2	3,4
Mưa	Đầu mùa	2,5	2,2	3,4
	Giữa mùa	2,5	2,3	3,5
	Cuối mùa	2,6	2,3	3,5

Hãy cho biết hàm lượng saponin có khác nhau theo mùa hay miền không? Nếu có thì 2 yếu tố mùa và miền có sự tương tác với nhau hay không? α = 0,05

Dùng EXCEL

* Chọn Tools\Data Analysis...\Anova: Two Factor With Replication
* Chọn các mục như trong hình



* Bảng ANOVA

SUMMARY	Nam	Trung	Bac	Total
Count	3	3	3	9
Sum	7.3	6.5	9.8	23.6
Average	2.433333	2.166667	3.266667	2.622222222
Variance	0.003333	0.003333	0.013333	0.251944444
Count	3	3	3	9
Sum	7.6	6.8	10.4	24.8
Average	2.533333	2.266667	3.466667	2.755555556
Variance	0.003333	0.003333	0.003333	0.300277778

	Total		
Count	6	6	6
Sum	14.9	13.3	20.2
Average	2.483333	2.216667	3.366667
Variance	0.005667	0.005667	0.018667

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	0.08	1	0.08	16	0.001761696	4.747221283
Columns	4.347778	2	2.173889	434.7777778	6.36194E-12	3.885290312
Interaction	0.01	2	0.005	1	0.396569457	3.885290312
Within	0.06	12	0.005			
Total	4.497778	17				

$\Rightarrow F_A > F_{1;\ 12;\ 0,95} = 4,7472$: Hàm lượng saponin khác nhau theo mùa.

$F_B > F_{2;\ 12\ ;\ 0,95} = 3,8853$: Hàm lượng saponin khác nhau theo miền.

$F_{AB} < F_{2\ ;\ 12\ ;\ 0,95} = 3,8853$: chấp nhận H_0 (không tương tác)

Vậy hàm lượng saponin trong dược liệu khác nhau theo mùa, theo miền và không có sự tương tác giữa mùa và miền trên hàm lượng saponin.

Bài t**ập**

1) Một nghiên cứu được thực hiện nhằm xem xét sự liên hệ giữa loại phân bón, giống lúa và năng suất. Năng suất lúa được ghi nhận từ các thực nghiệm sau:

Giống lúa / Loại phân bón	A	B	C
1	65 68 62	69 71 67	75 75 78
2	74 79 76	72 69 69	70 69 65
3	64 72 65	68 73 75	78 82 80
4	83 82 84	78 78 75	76 77 75

Hãy cho biết sự ảnh hưởng của loại phân bón, giống lúa trên năng suất, α = 0,01

2) Điều tra mức tăng trưởng chiều cao của 1 loại cây trồng theo loại đất trồng và loại phân bón có kết quả:

Loại đất / Loại phân	1	2	3
A	5,5 5,5 6,0	4,5 4,5 4,0	3,5 4,0 3,0
B	5,6 7,0 7,0	5,0 5,5 5,0	4,0 5,0 4,5

Hỏi có sự khác nhau của mức tăng trưởng chiều cao theo loại đất và loại phân bón ?
α=0,05

3) Nghiên cứu sản lượng bông (tạ/ha) theo mật độ trồng A và phân bón B thu được:

Mật độ trồng	Phân bón			
	b1	b2	b3	b4
a1	16 14 21 16	19 20 23 19	19 21 22 20	20 24 21 17
a2	17 15 17 19	19 18 18 20	21 21 22 23	20 20 22 19
a3	18 18 19 17	20 23 21 21	22 18 21 21	25 22 21 23

Hỏi có sự khác nhau của sản lượng bông theo mật độ trồng, theo phân bón với mức α=0,05

V. TƯƠNG QUAN - HỒI QUY

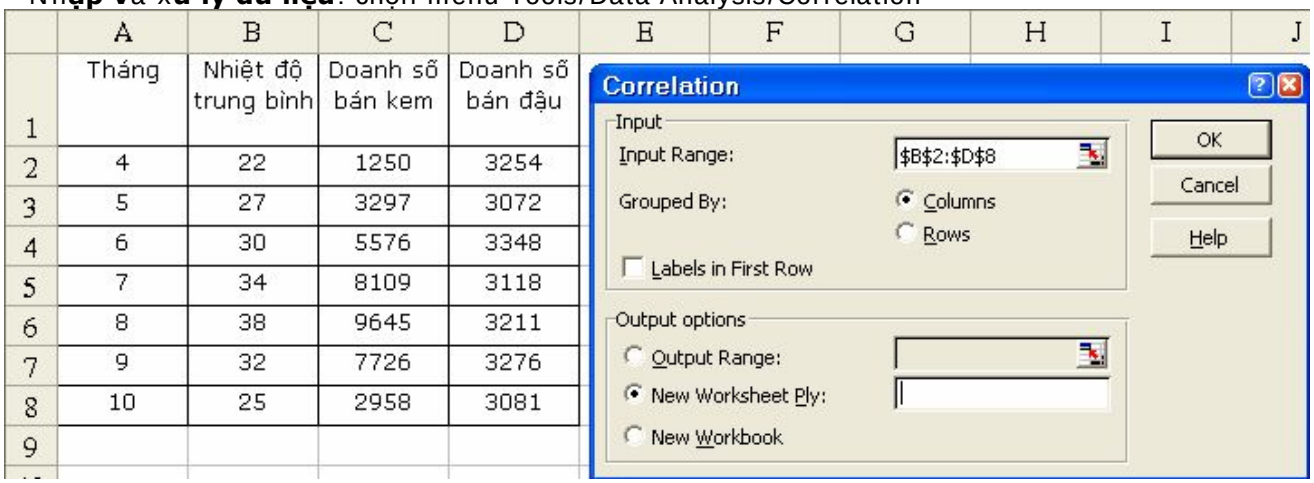
1) Tương quan (Correlation)

- Hệ số tương quan $R = \dfrac{\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}}$
- Nếu R >0 thì X, Y tương quan thuận
  Nếu R <0 thì X, Y tương quan nghịch
- Nếu R=0 thì X , Y không tương quan
- Nếu $|R| = 1$ thì X,Y có quan hệ hàm số bậc nhất.
- Nếu $|R| \to 1$ thì X, Y có tương quan chặt (tương quan mạnh)
- Nếu $|R| \to 0$ thì X, Y có tương quan không chặt (tương quan yếu)

Ví dụ: Khảo sát mới quan hệ giữa nhiệt độ trung bình với doanh số bán kem theo bảng thống kê sau:

Tháng	Nhiệt độ trung bình	Doanh số bán kem	Doanh số bán đậu
4	22	1250	3254
5	27	3297	3072
6	30	5576	3348
7	34	8109	3118
8	38	9645	3211
9	32	7726	3276
10	25	2958	3081

- Nhập và xử lý dữ liệu: chọn menu Tools/Data Analysis/Correlation



- Kết quả

	Column 1	Column 2	Column 3
Column 1	1		
Column 2	0.985572	1	
Column 3	0.127653	0.184818	1

Vì R_{12}=0,9856 chứng tỏ giữa nhiệt độ (*Column 1) và doanh* số bán kem (*Column 2*) có mối quan hệ rất chặt chẽ với nhau và có tương quan thuận.

2) Hồi quy (Regression)

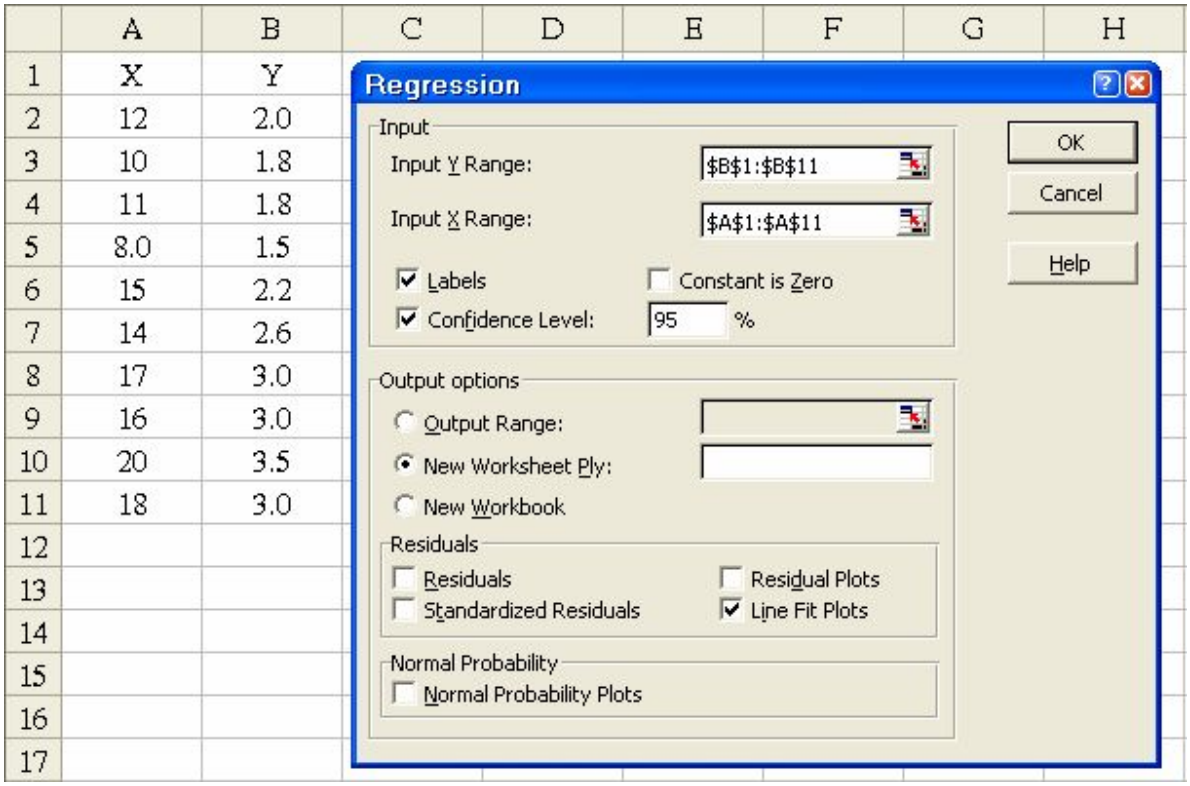
a) Hồi quy đơn tuyến tính

- Phương trình hồi quy tuyến tính: $\overline{y}_x = a + bx$, $a = r\frac{\overline{S_y}}{\overline{S_x}}$, $b = \overline{y} - a\overline{x}$
- 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

Ví dụ: Số liệu về doanh số bán hàng (Y) và chi phí chào hàng (X) của một số công ty, có kết quả sau:

X (triệuđ/năm)	12	10	11	8	15	14	17	16	20	18
Y (tỷ đ/năm)	2	1,8	1,8	1,5	2,2	2,6	3	3	3,5	3

Xác định phương trình hồi quy tuyến tính



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.963150954
R Square	0.927659761
Adjusted R Square	0.918617231
Standard Error	0.191227589
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3.751456073	3.751456	102.5885	7.71522E-06
Residual	8	0.292543927	0.036568		
Total	9	4.044			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	0.053017571	0.243302295	0.217908	0.832956	-0.50803889
X	0.169289534	0.016714013	10.1286	7.72E-06	0.130746927

$$\Rightarrow \bar{y}_x = 0{,}053 + 0{,}1693x$$

- Hệ số hồi quy: 0.832956 > 0,05 : hệ số tự do có ý nghĩa.
  7.72E-06 < 0,05 : hệ số của x không có ý nghĩa.
- Phương trình hồi quy tuyến tính này không thích hợp vì 7.71522E-06 < 0,05.

b) Hồi quy đa tuyến tính

- Phương trình hồi quy đa tuyến tính: $\bar{y}_x = b_0 + b_1x_1 + ... + b_nx_n$
- 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

Ví dụ: Người ta đã dùng ba mức nhiệt độ gồm 105 , 120 và 135 ^{0}C kết hợp với ba khoảng thời gian là 15 , 30 và 60 phút để thực hiện một phản ứng tổng hợp. các hiệu suất của phản ứng (%)được trình bày trong bảng sau đây:

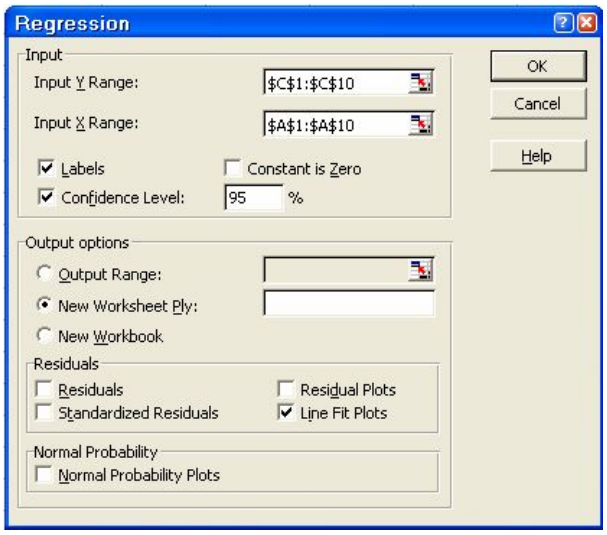
Thời gian (ph) X1	Nhiệt độ (^{0}C) X2	Hiệu suất (%) Y
15	105	1,87
30	105	2,02
60	105	3,28
15	120	3,05
30	120	4,07
60	120	5,54
15	135	5,03
30	135	6,45
60	135	7,26

Hãy cho biết yếu tố nhiệt độ và hoặc yếu tố thời gian có liên quan tuyến tính với hiệu suất của phản ứng tổng hợp? Nếu có thì ở điều kiện nhiệt độ 115 ^{0}C trong 50 phút thì hiệu suất phản ứng sẽ là bao nhiêu?

- Nhập dữ liệu:

	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

- $\overline{Y}_{X1} = b_0 + b_1 X1$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.462512069
R Square	0.213917414
Adjusted R Square	0.101619901
Standard Error	1.811191587
Observations	9

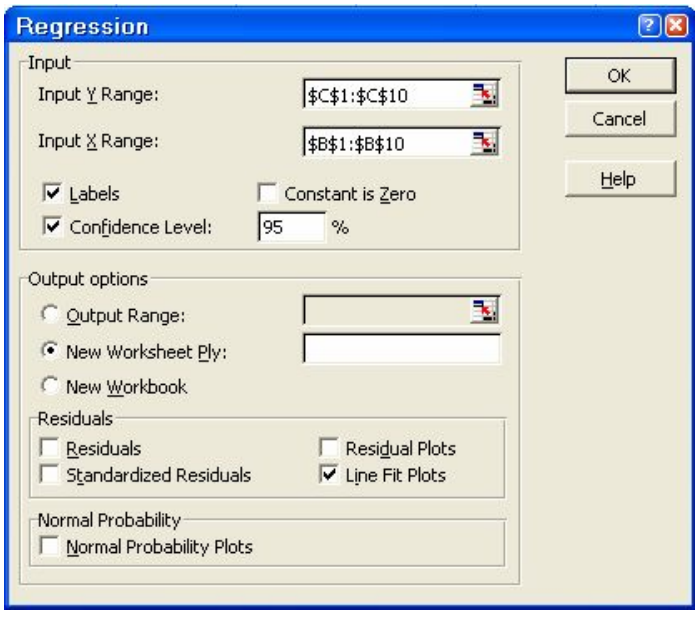
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6.24891746	6.248917	1.904917	0.209994918
Residual	7	22.96290476	3.280415		
Total	8	29.21182222			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	2.726666667	1.280705853	2.129034	0.070771	-0.301719287
X1	0.044539683	0.032270754	1.380187	0.209995	-0.031768471

Phương trình hồi quy: $\overline{Y}_{X1} = 2{,}7267 + 0{,}04454X1$ không thích hợp vì 0.209994918 > 0,05
Nghĩa là : Hiệu suất Y không có liên quan tuyến tính với yếu tố thời gian X1

- $\overline{Y}_{X2} = b_0 + b_2 X2$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.873933544
R Square	0.76375984
Adjusted R Square	0.730011246
Standard Error	0.99290379
Observations	9

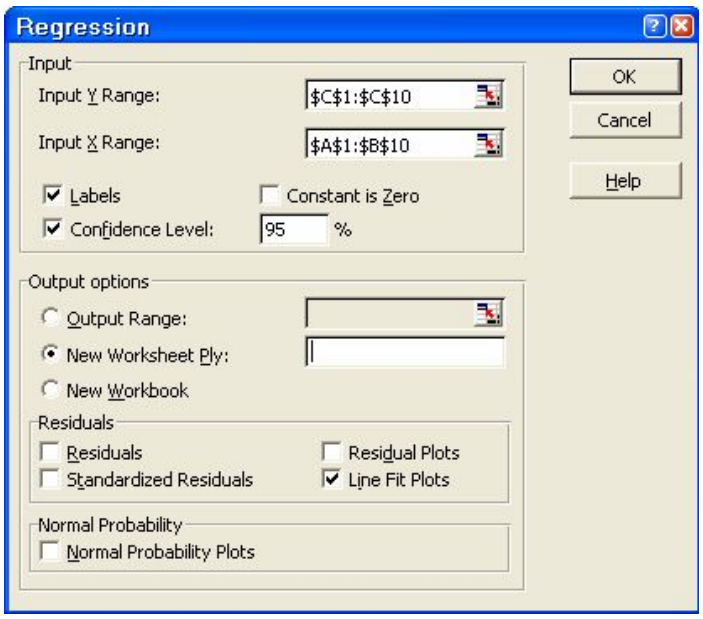
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	22.31081667	22.31082	22.63086	0.002066188
Residual	7	6.901005556	0.985858		
Total	8	29.21182222			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-11.14111111	3.25965608	-3.41788	0.011168	-18.84896742
X2	0.128555556	0.027023418	4.757191	0.002066	0.064655371

Phương trình hồi quy: $\overline{Y}_{X1} = -11{,}1411 + 0{,}1286X1$ này thích hợp vì 0.002066188 < 0,05
Nghĩa là: Hiệu suất Y có liên quan tuyến tính với yếu tố nhiệt độ X2.

- $\overline{Y}_{X1,X2} = b_0 + b_1X1 + b_2X2$



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	1.11235E-05
Residual	6	0.652088095	0.108681		
Total	8	29.21182222			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-12.7	1.101638961	-11.5283	2.56E-05	-15.3956154
X1	0.04454	0.005873842	7.582718	0.000274	0.030166899
X2	0.128556	0.008972441	14.32782	7.23E-06	0.106600767

Phương trình hồi quy: $\overline{Y}_{X1,X2} = -12{,}7 + 0{,}04454X1 + 0{,}1286X2$ này thích hợp vì 1.11235E-05 < 0,05

Nghĩa là:Hiệu suất Y có liên quan tuyến tính với thời gian X1 và nhiệt độ X2.

- khi X1=50 , X2=115 ta dự đoán:

Intercept	-12.7
X1	0.04454
X2	0.128556

Dự đoán hiệu suất Y:

4.31094

Bài t**ập**

1. Cho Y là nhu cầu thịt bò (đơn vị 100 tấn) của 12 tháng liên tiếp (X) trong một khu dân cư :
   X: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
   Y: 15, 18, 18, 16, 14, 18, 20, 21, 19, 20, 24, 26.
   Hãy ước lượng hàm hồi quy tuyến tính đơn, dự báo nhu cầu thịt bò cho 3 tháng tiếp theo.
   Đáp số : y = 0.793706 x + 13.92424.
2. Trong 10 tháng liên tiếp lượng hàng bán ra của một công ty rất thấp, sau đó công ty tung ra thị trường một sản phẩm mới và nhận thấy lượng hàng bán ra tăng theo hàm mũ. Số đơn vị hàng bán ra (Y) trong 6 tháng tiếp theo (X) cho trong bảng sau:

	A	B
1	Y	X
2	33100	11
3	47300	12
4	69000	13
5	102000	14
6	150000	15
7	220000	16

   Hãy ước lượng hàm hồi quy mũ và dự báo lượng hàng bán ra trong các tháng 17, 18, 19, 20 (dùng hàm Growth). *Đáp số : y = 495.3048 +1.463276x* .
3. Tính hàm hồi quy tuyến tính bội với số liệu cho trong bảng dưới

	A	B	C	D	E
1	Y	X1	X2	X3	X4
2	733.300	3.089	76.200	283.500	15.844
3	750.900	3.503	79.400	274.500	19.835
4	747.600	3.817	77.000	268.000	21.797
5	727.600	3.870	74.000	265.700	24.759
6	694.400	3.706	64.400	259.600	28.093
7	702.600	3.851	63.100	256.800	31.121
8	714.000	4.170	66.300	259.300	32.759
9	717.630	4.378	62.900	263.400	34.556
10	750.000	5.000	66.700	273.100	36.788

   trong đó Y là thu nhập quốc dân, X1 là sản lượng điện, X2 là sản lượng than, X3 là sản lượng lương thực, X4 là sản lượng thép. Dùng hai phương pháp: dùng hàm Linest và lệnh Tools / Data Analysis. Dự báo Y với X = (5.2, 65.1, 275.3, 37.8).
   Đáp số: dự báo Y =751.79289.
4. Bảng bên cho số liệu về doanh thu (Y), chi phí cho quảng cáo (X1), tiền lương của nhân viên tiếp thị (X2) của 12 công ty tư nhân, đơn vị là 1 triệu đồng. Xây dựng hàm hồi quy tuyến tính bội Y phụ thuộc vào X1, X2.

	A	B	C	D
1	Y	X1	X2	Trend
2	127	18	10	124.9673
3	149	25	11	147.2661
4	106	19	6	108.4383
5	163	24	16	168.5539
6	102	15	7	103.1741
7	180	26	17	178.324
8	161	25	14	161.5422
9	128	16	12	129.4732
10	139	17	12	131.979
11	144	23	12	147.0134
12	159	22	14	154.025
13	138	15	15	141.2436

Để ước lượng hàm hồi quy ta dùng hàm mảng Linest như sau: đánh dấu khối vùng ô B19: D23, nhập công thức =LINEST(A2 : A13, B2 : C13, True, True), ấn Ctrl + Shift +Enter, kết quả ta được 12 số:

	A	B	C	D
18		m2	m1	b
19		4.75869	2.505729	32.27726
20	SE	0.41038	0.328573	6.253073
21	R^2	0.97566	4.003151	#N/A
22	F	180.355	9	#N/A
23	SS	5780.44	144.2269	#N/A

Tiếp theo, cho các bộ giá trị mới x1, x2 trong khối ô B15 : C17, cần dự báo các giá trị y được tính theo (2) trong khối ô D15 :D17. Thao tác tính: đánh dấu khối vùng ô D15:D17, nhập công thức = Trend(a2: a13,b2: c13, b15: c17, True), ấn Ctrl + Shift +Enter

	A	B	C	D
14				Du bao
15		26	18	183.0827
16		28	19	192.8529
17		30	20	202.623

5. Tính hàm hồi quy của y (sản lượng nông nghiệp) phụ thuộc vào x (lượng phân bón).

	A	B	C	D	E
1	Y	X		m	b
2	13983.800	763.534		4.064894	11456.13
3	14406.400	784.630			
4	15005.300	776.200		X	Du bao Y
5	16829.000	1118.600		1612	18008.740
6	17100.000	1488.000			

Công thức trong ô D2 là = Slope(a2:a6, b2:b6), công thức trong ô E2 là =Intercept(a2:a6, b2:b6), công thức trong ô E5 là =Forecast(d5, a2:a6, b2:b6) để dự báo y với x = 1612.

$y = mx + b$

Do đó tất cả các hàm và lệnh đã trình bày với hồi quy tuyến tính bội cũng đúng với hồi quy tuyến tính đơn. Song đối với hồi quy tuyến tính đơn có thêm ba hàm mới.

– *Hàm Slope(known_y's, known_x's)* ước lượng giá trị m của phương trình (3).
– *Hàm Intercept(known_y's, known_x's)* ước lượng giá trị b của (3).
– *Hàm Forecast(x, known_y's, known_x's)*: dự đoán y theo phương trình (3) với giá trị x biết trước.