

## TRƯỜNG ĐẠI HỌC BÁCH KHOA THÀNH PHỐ HỒ CHÍ MINH KHOA KHOA HỌC ỨNG DỤNG

**BỘ MÔN TOÁN ỨNG DỤNG**

**BÁO CÁO BÀI TẬP LỚN**

**XÁC SUẤT THỐNG KÊ**

# ĐỀ TÀI 5

**GVHD: NGUYỄN KIỀU DUNG THỰC HIỆN: NHÓM 3 – L02**

### 21300205 – Nguyễn Văn Ba

1. 21300382 – Trần Minh Chiến (NT)
2. 21300730 – Vương Phạm Phi Dương
3. 21300814 – Phùng Tiến Đạt
4. 21301031 – Chế Minh Hải
5. 21301189 – Phan Đức Hiếu
6. 81301915 – Hoàng Văn Khương

Thành phố Hồ Chí Minh, tháng 11 năm 2014

**BÀI 1:**

Tìm một dữ liệu định lượng (A) và một dữ liệu định tính (B) thích hợp, sử dụng các dữ liệu đó cho các yêu cầu sau:

1. Thực hiện phương pháp phân tổ dữ liệu (A).
2. Vẽ đồ thị phân phối tần số và đa giác tần số (A).
3. Tính các đặc trưng mẫu và ước lượng giá trị trung bình của dấu hiệu quan sát với độ tin cậy 95% (A).
4. Trình bày dữ liệu định tính (B) dạng phân loại bằng các đồ thị.

#### BÀI LÀM:

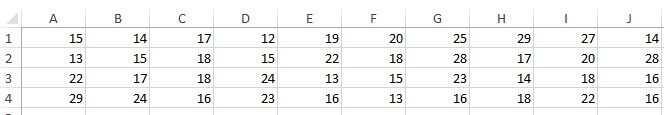
* **Dạng bài:** Thống kê mô tả.
* **Dữ liệu (A):** Khảo sát thời gian hoàn thành một sản phẩm tiện của 40 sinh viên khi học thực tập Cơ khí ta có bảng số liệu: Thời gian (phút)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 15 | 14 | 17 | 12 | 19 | 20 | 25 | 29 | 27 | 14 |
| 13 | 15 | 18 | 15 | 22 | 18 | 28 | 17 | 20 | 28 |
| 22 | 17 | 18 | 24 | 13 | 15 | 23 | 14 | 18 | 16 |
| 29 | 24 | 16 | 23 | 16 | 13 | 16 | 18 | 22 | 16 |

* **Dữ liệu (B):** Phân ngành sinh viên khoa cơ khí khóa 2013 của trường Đại học Bách Khoa thành phố Hồ Chí Minh.

|  |  |
| --- | --- |
| **Ngành học** | **Số sinh viên** |
| KSTN Cơ điện tử | 30 |
| KSTN Chế tạo máy | 30 |
| Cơ điện tử | 100 |
| Kỹ thuật nhiệt lạnh | 83 |
| Kỹ thuật chế tạo | 158 |
| Kỹ thuật thiết kế | 50 |
| Kỹ thuật máy xây dựng và nâng chuyển | 45 |

1. **Thực hiện phương pháp phân tổ dữ liệu A:**
   * Nhập dữ liệu (A) vào Excel:



1

+ Xác định số tổ cần chia: k = (2 × n)3

Chọn ô A6 nhập vào biểu thức =(2\*Count(A1:J4))^(1/3) Kết quả 4.31

Chọn k = 4

+ Xác định trị số khoảng cách h theo công thức: ℎ =

(Xmax –Xmin ) k

Chọn ô A7 nhập vào biểu thức =(Max(A1:J4)-Min(A1:J4))/4 Kết quả 4.25

Chọn h =4

+ Ta xác định được các cận trên và cận dưới các tổ lần lượt là:

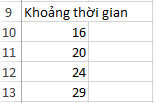
 Tổ 1: 12 – 16

 Tổ 2: 16 – 20

 Tổ 3: 20 – 24

 Tổ 4: 24 – 29

 Nhập vào các ô từ A9 đến A13 lần lượt các giá trị:



* + Chọn chức năng Data/ Data Analysis/Histogram.

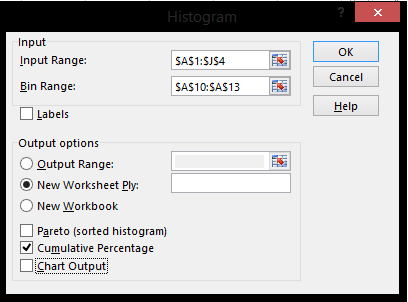
+ Input Range: địa chỉ tuyệt đối chứa dư liệu.

+ Bin Range: địa chỉ chứa bảng phân nhóm.

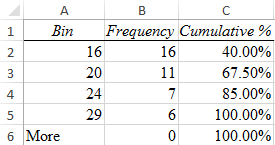
+ Output options: vị trí xuất kết quả.

+ Confidence Level for Mean: độ tin cậy cho trung bình.

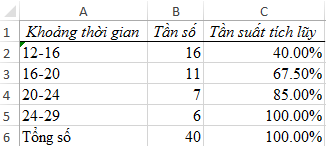
+ Chọn Cumulative Percentage để tính tần suất tích lũy nếu không Excel chỉ tính tần số.



* + Kết quả:



* + - Có thể chỉnh sửa lại như sau:

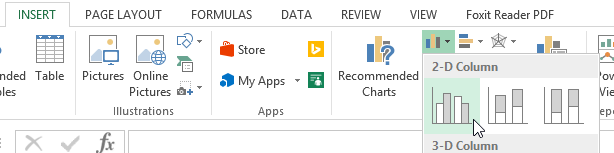


## Vẽ đồ thị phân phối tần số và đa giác tần số (A):

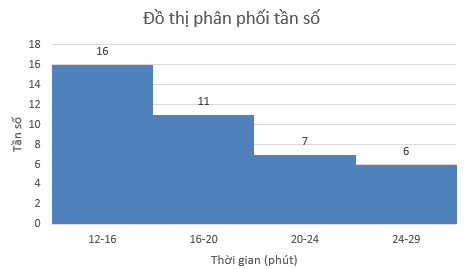
* + Vẽ đồ thị phân phối tần số:

+ Quét chọn bảng tần số B2:B5

+ Dùng chức năng Insert Column Chart trên menu Insert.



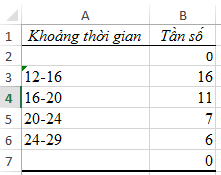
* + - Kết quả sau khi chỉnh sửa:



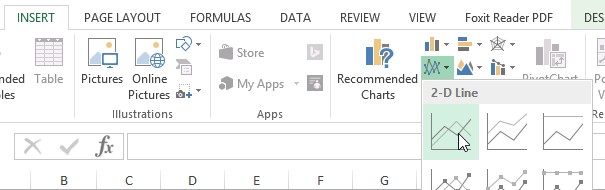
* + Vẽ đa giác tần số:

+ Sử dụng bảng phân phối tần số của dữ liệu (A):

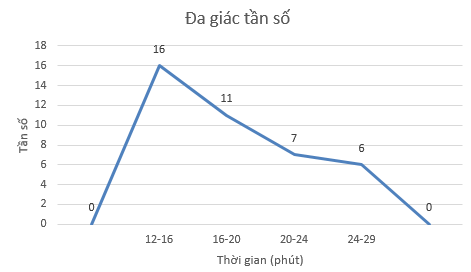
+ Thêm giá trị 0 vào đầu và cuối bảng phân phối tần số:



+ Quét chọn B2:B7, dùng chức năng Insert Line Chart trên menu Insert

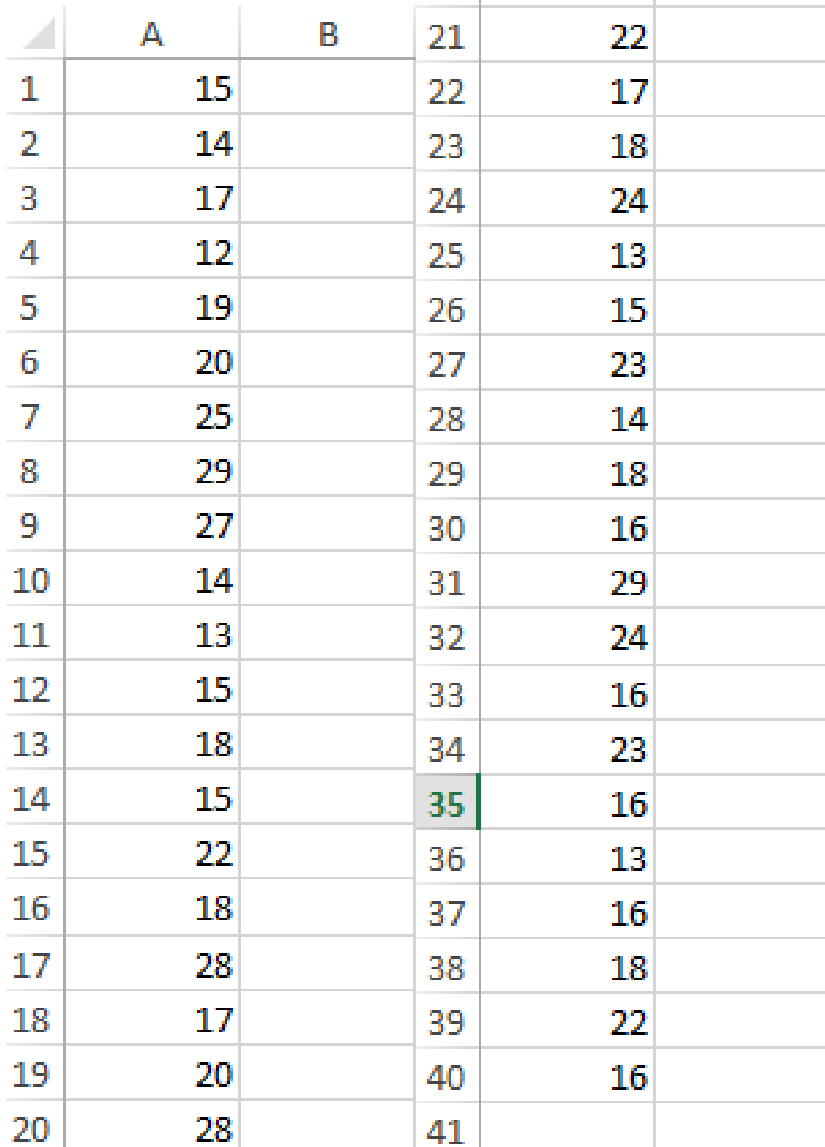


* + - Kết quả sau chỉnh sửa:



## Tính các đặc trưng mẫu và ước lượng giá trị trung bình của dấu hiệu quan sát với độ tin cậy 95% (A).

* + Nhập dữ liệu vào bảng tính:

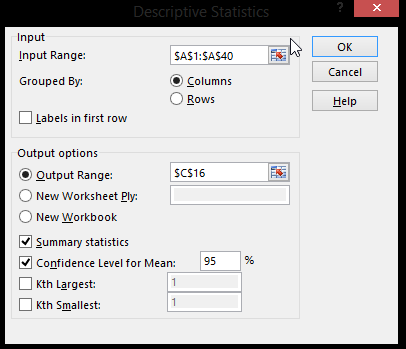


* + Chọn chức năng Data/Data Analysis/Descriptive Statistics.

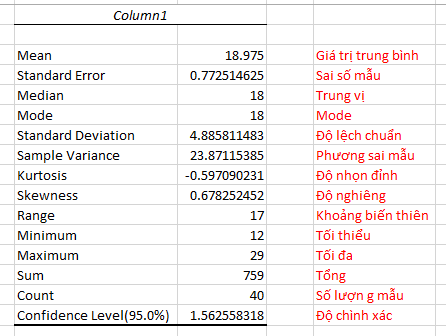
+ Input Range: địa chỉ tuyệt đối chứa dư liệu.

+ Output options: vị trí xuất kết quả.

+ Confidence Level for Mean: độ tin cậy cho trung bình.

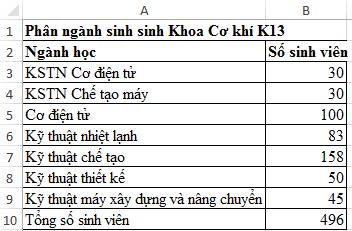


* + Kết quả nhận được:



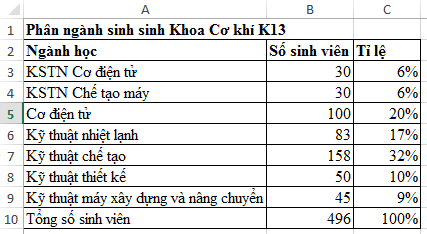
## Trình bày dữ liệu định tính (B) dạng phân loại bằng các đồ thị.

* + Nhập dữ liệu và bảng tính:



* + Tính tỉ lệ sinh viên cho các ngành:

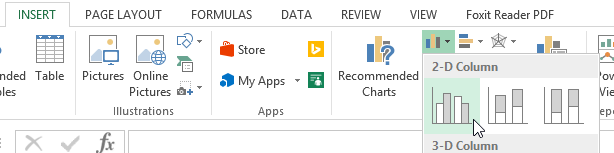
Nhập vào C3: =B3/$B$10, copy cho các ô còn lại.



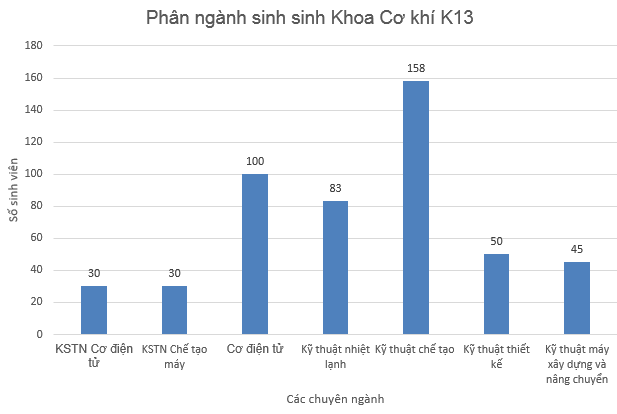
* + Vẽ biểu đồ đứng thể hiện số lượng sinh viên ở các chuyên ngành.

+ Quét chọn cột Số sinh viên (B3:B9)

+ Dùng chức năng Insert /Insert Column Chart/2-D Column trên menu Insert.



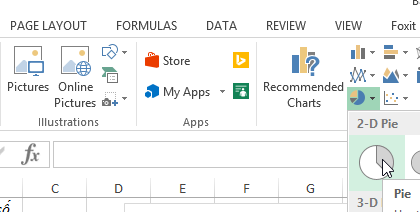
* + - Kết quả thu được:



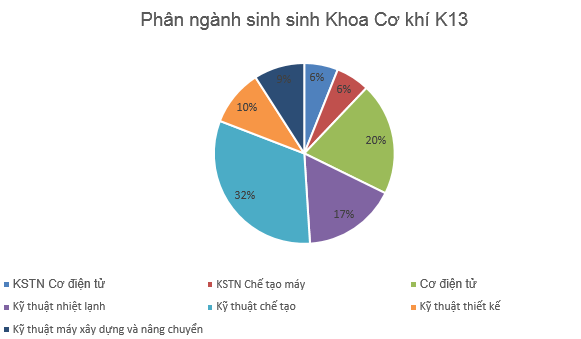
* + Vẽ biểu đồ tròn thể hiện tỉ lệ sinh viên ở các chuyên ngành.

+ Quét chọn cột Số sinh viên (C3:C9)

+ Dùng chức năng Insert/Insert Pie/2-D Pie trên menu Insert.



* + - Kết quả thu được:



# BÀI 2:

Theo dõi doanh số bán hàng của một 2 cửa hàng, người ta thu được kết quả sau:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cửa hàng 1 | 10.2 | 9.2 | 4.4 | 3.2 | 5.6 | 6.3 | 7.4 | 8.4 | 3.9 | 7.2 | 6.5 | 6.2 | 7.4 | 7.5 |
| Cửa hàng 2 | 8.8 | 9 | 5.1 | 4.2 | 4.1 | 5.8 | 6.3 | 6.7 | 5.6 | 6.7 | 6.7 | 7.6 |  |  |

Với mức ý nghĩa 3%, có thể cho rằng doanh số bán hàng của 2 cửa hàng có sự phân tán như nhau hay không? Giả thiết doanh số bán hàng mỗi ngày của các cửa hàng tuân theo quy luật chuẩn.

#### BÀI LÀM:

* **Dạng bài:** Kiểm định giả thuyết cho phương sai hai tổng thể.
* **Công cụ:** F-Test Two-Sample for Variances

#### Cơ sở lý thuyết:

- Khi cần kiểm định hai tổng thể có biến động như nhau hay không chúng ta dùng phương pháp kiểm định phương sai của hai tổng thể độc lập dựa trên một đại lượng F như sau:

2

s

1

F = 2

s

2

Trong đó: s2 là phương sai của mẫu thứ nhất, mẫu này có cỡ n1.

1

s2 là phương sai của mẫu thứ hai, mẫu này có cỡ n2.

2

* Thông thường để xác địnhmaẫu nào là mẫu thứ nhất và mẫu nào là mẫu thứ hai ta làm như sau,trong khi tính đại lượng F thì giá trị phương sai lớn hơn sẽ được đặt ở tử số, và như vậy mẫu tương ứng với phương sai đó là mẫu thứ nhất.
* Giả thiết đặt ra là kiểm định hai bên:

HO:o2 H1:o2

1

1

= o2

≠ o2

2

2

* Nếu tỉ số F rất lớn hoặc rất nhỏ ta có thể suy diễn bằng hai phương sai tổng thể khó mà bằng nhau, ngược lại nếu tỉ số này gần đến 1 ta sẽ có bằng chứng ủng hộ giả thuyết H0. Như vậy tỉ lệ F lớn đến đâu thì xem như là đủ bằng chứng bác bỏ H0 và ngược lại.
* Nếu tổng thể lấy mẫu được giả định có phân phối bình thường thì tỉ lệ F có phân phối xác suất gọi tên là phân phối Fisher. Các giá trị tới hạn của phân phối F phụ thuộc và hai giá trị bậc tự do, bậc tự do tử số (df1 = n1 − 1) gắn liền với mậu thứ nhất và bậctuự do mẫu số gắn liền với mẫu thứ hai (df2 = n2 − 1).
* Quy tắc thực sự để bác bỏ H0 với kiểm định hai bên khi df1 = n1 − 1 và df2 = n2 −

1, mức ý nghĩa  là: giả thiết H0 bị bác bỏ nếu giá trị kiểm định F lớn hơn giá trị tới hạn trên FU = Fdf1;df2;α /2 của phân phối F hoặc bé hơn giá trị tới hạn dưới FL = Fdf1;df2;1–α /2 tức là Ftt < Fdf1;df2;1–α /2 hoặc Ftt > Fdf1;df2;α /2.

* Nếu chúng ta kiểm định bên phải:

HO:o2 H1:o2

1

1

= o2

> o2

2

2

Quy tắc bác bỏ H0 là khi Ftt > FU (n1–1;n2–1;α ).

* Giả thiết:

HO:o2 = o2

1 2

H1:o2 > o2

1 2

2 2

2



Giá trị thống kê: F =

o s

2 1

s

1

o2s2 s2

=

1 2 2





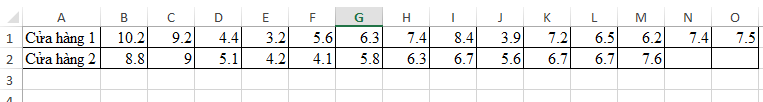
Phân phối Fischer: y1 = N1 − 1; y2 = N2 − 1

Biện luận:

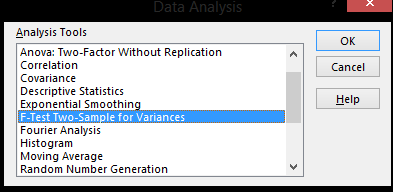
Nếu F < Fα (y1,y2) Chấp nhận giả thuyết H0 với xác xuất (1-)100%.

#### Thực hiện bài toán bằng excel:

* Nhập dữ liệu vào bảng tính:



* Vào Data/ Data Analysis/ F-Test Two-Sample for Variances.

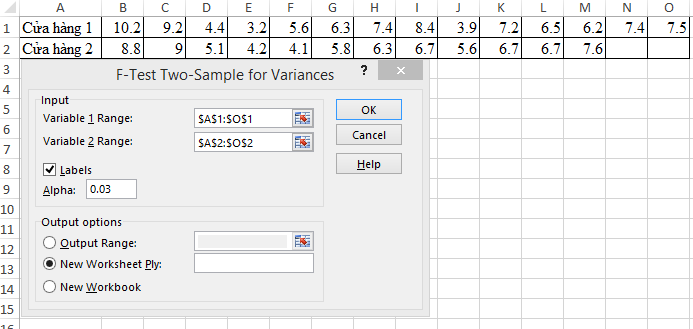


* Chọn các mục như hình:

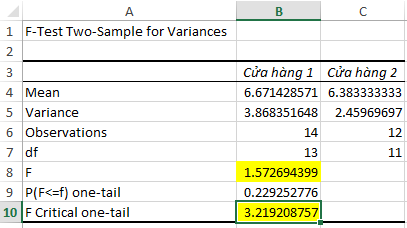
+ Input: địa chỉ tuyệt đối chứa dư liệu tương ứng của mẫu 1 và 2

+ Output options: vị trí xuất kết quả.

+ Apha: mức ý nghĩa 



#### Kết quả:



* + **Biện luận:**

Giả thiết HO:o2

1

= o2

“Doanh số bán hàng của 2 cửa hàng có sự phân tán như nhau”.

H1:o2

2

1

> o2: “Doanh số bán hàng của 2 cửa hàng không phân tán như nhau”.

F = 1.5727 < F 0.03 = 3.2192  Chấp nhận giả thuyết H0.

2

#### Vậy: Doanh số bán hàng của 2 cửa hàng có sự phân tán như nhau.

**BÀI 3:**

Đo mức độ bụi trong không khí tại các khu vực trong thành phố tại cùng một thời điểm, người ta được số liệu sau (đơn vị mg/m3):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Số thứ tự quan sát | Các khu vực | | | |
| KV1 | KV2 | KV3 | KV4 |
| 1 | 0.54 | 0.48 | 0.56 | 0.47 |
| 2 | 0.60 | 0.49 | 0.62 | 0.52 |
| 3 | 0.72 | 0.55 | 0.60 | 0.56 |
| 4 | 0.67 | 0.62 | 0.71 | 0.53 |
| 5 | 0.83 | 0.57 | 0.73 |  |
| 6 | 0.63 | 0.59 |  |  |

Mức độ nhiễm bụi của các khu vực trên có được coi là như nhau hay không? Hãy kết luận bằng giá trị P. Tìm hệ số xác định R2 của bài toán.

#### BÀI LÀM:

* + **Dạng bài**: Kiểm định giá trị trung bình một nhân tố.
  + **Phương pháp giải**: Phân tích phương sai một nhân tố.
  + **Công cụ giải**: Anova single factor.

#### Cơ sở lý thuyết:

Giả sử nhân tố A có k mức X1, X2, …, Xk với Xj có phân phối chuẩn N(a,2) có mẩu điều tra:

|  |  |  |  |
| --- | --- | --- | --- |
| X1 | X2 | … | Xk |
| x11 | x12 |  | x1k |
| x21 | x22 |  | x2k |
| . | . |  | . |
| . | . | … | . |
| . | . |  | . |
| xn1 | xm2 |  | xpk |

Với mức ý nghĩa  ta kiểm định giả thiết:

o H0: a1=a2= … = ak

o H1: “Tồn tại j1≠j2 sao cho aj1≠aj2

Đặt:

* Tổng quan sát:

k

n = ) nj

j= 1

* Trung bình nhóm j (j=1, …, k):

nj nj

1

xj̅= n

Tj

) xij= n

rớiTj = ) xij

j j= 1 j

i= 1

* Trung bình mẫu chung:

k nj

### 1 T

k nj k

### x̅= ) ) xij= rới T = ) ) xij= ) Tj n n

j= 1 i= 1

* Phương sai hiệu chỉnh nhóm j:

nj

j= 1 i= 1

i= 1

2 1 2

Sj = nj

) (xij− x¯j)

− 1

i= 1

* Tổng bình phương các độ lệch:

k nj

2

### STT = ) ) (xij− x¯y)

j= 1 i= 1

* Tổng bình phương độ lệch riêng của các nhóm so với x̅:

k

2

SSA = ) nj(x¯j− ¯x)

j=1

k

nj

SST = ) ) x

2

ij

k

SSA = )

j= 1

2

−

nj n

Tj

T2

SSE = SST − SSA

j= 1 i= 1

SSA

MSA =

k − 1

SSE

MSE =

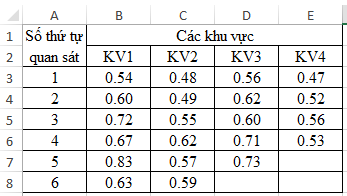
n − k

* + Bảng ANOVA:

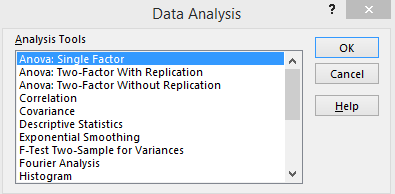
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nguồn sai số | Tổng bình phương SS | Bậc tự do df | Bình phương trung bình MS | Giá trị thống kê F |
| Yếu tố | SSA | k-1 | SSA  MSA = k − 1 | MSA  F = MSE |
| Sai số | SSE = SST - SSA | n-k | SSE  MSE =  n − k |  |
| Tổng cộng | SST | n-1 |  |  |

#### Thực hiện bài toán bằng excel:

* Nhập dữ liệu vào bảng tính:



* Vào Data/ Data Analysis/Anova: Single Factor.

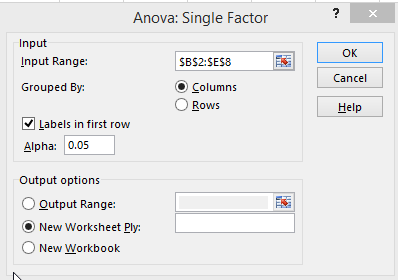


* Chọn các mục như hình:

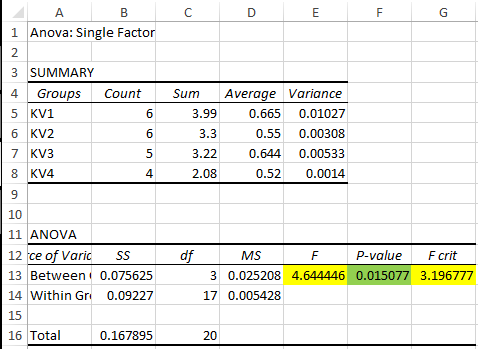
+ Input Range: địa chỉ tuyệt đối chứa dư liệu.

+ Output options: vị trí xuất kết quả.

+ Apha: mức ý nghĩa .



* + Kết quả:



#### Biện luận:

Giả thiết H0: Mức độ nhiễm bụi của các khu vực là như nhau.

H1: Mức độ nhiễm bụi của các khu vực không như nhau.

Ta có: **P-value = 0.015077**  Có ý nghĩa thống kê. F = 4.644446 > F3,3,0.95 = 3.196777

 Bác bỏ giả thiết H0

#### Vậy: Mức độ nhiễm bụi của các khu vực không như nhau.

* **Tính hệ số xác định R2 (R Square):**

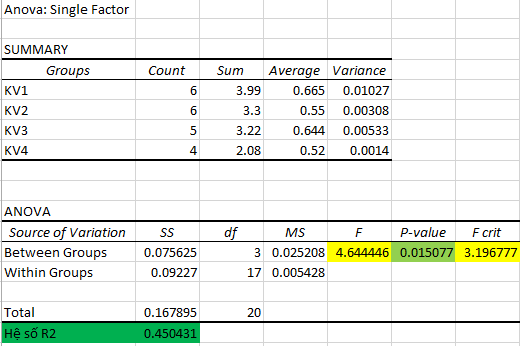
Từ bẳng “Anova: Single Factor” ta tính được hệ số tương quan R 2 theo công thức:

#### Thực hiện trên Excel:

R 2 = SSF

SST

* Chọn ô B17 nhập công thức: =B13/B16
* Ta được kết quả như sau:



#### Vậy hệ số xác định R2 = 0.450431

**BÀI 4:**

Bảng số liệu sau cho biết số người chết về bệnh ung thư ở 3 nước Mỹ, Nhật, Anh trong thời gian khảo sát. Người chết được phân loại theo cơ quan bị ung thư.

|  |  |  |  |
| --- | --- | --- | --- |
| Bộ phận bị 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 |

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

#### BÀI LÀM:

* + **Dạng bài**: Kiểm định giả thiết về tỉ lệ.
  + **Phương pháp giải**: Áp dụng Kiểm định chi bình phương 2.
  + **Công cụ giải**: hàm CHITEST trên Excel.

#### Cơ sở lý thuyết:

* Trong thống kê, kiểm định chi bình phương hay kiểm tra 2. (đôi khi đọc là "khi bình phương") là một họ các phương pháp kiểm định giả thiết thống kê trong đó thống kê kiểm định tuân theo phân bố 2 nếu giả thuyết không là đúng. Chúng gồm:
* Kiểm định chi bình phương Pearson
* Kiểm định chi bình phương Yates
* Kiểm định chi bình phương Mantel-Haenszel

 ...

* Dạng thống kê kiểm định thông dụng nhất là:

(o − e)2

y2 =

e

Với o là dữ liệu đo đạc, e là giá trị dự đoán chính xác.

Xét một bộ A gồm r tính trạng, A = (A1, A2, ...Ar), trong đó mỗi cá thể của tập hợp chính H có và chỉ có một trong các tính trạng (hay phạm trù) Ai.

Gọi pi (i = 1, 2, ... r) là tỷ lệ cá thể tính trạng Ai trong tập hợp chính H. Khi đó véctơ

=(p1, p2, ...pr) được gọi là phân bố của A trong tập hợp chính H.

Giả sử (p1, p2,...pr) là phân bố của (A1, A2,...Ar) trong tập hợp chính H và (q1, q2,...qr) là phân bố của A = (A1, A2,...Ar) trong tập hợp chính Y. Ta nói (A1, A2...Ar) có phân bố như nhau trong X và Y nếu (p1, p2,...pr) = (q1, q2,...qr)  p1 = q1,...pr = qr.

Chúng ta muốn kiểm định xem A = (A1, A2,...Ar) có cùng phân số trong X và Y hay không dựa trên các mẫu ngẫu nhiên rút từ X và Y.

Tổng quát hơn, giả sử ta có k tập hợp chính H1, H2,...Hk. Gọi i  pi , pi ,pi  là phân bố của A = (A1, A2,...Ar) trong tập hợp chính Hi.

1 2 r

Ta muốn kiểm định giả thuyết sau:

Ho:       

1 2 k

(Các phân bố này là như nhau trên các tập hợp chính Hi).

Chú ý rằng H0 tương đương với hệ đẳng thức sau:

p1  p2   pk

1 1 1

p1  p2   pk

 2 2 2

p1  p2   pk

 i i i

p1  p2   pk

r r r

Từ mỗi tập hợp chính chúng ta chọn ra một mẫu ngẫu nhiên. Mẫu ngẫu nhiên chọn từ tập hợp chính Hi được gọi là mẫu ngẫu nhiên thứ i (i = 1, 2,... k).

Giả sử trong mẫu ngẫu nhiên thứ i:

Có n1i cá thể có tính trạng A1 n2i cá thể có tính trạng A2

..............................

nri cá thể có tính trạng Ar

Ký hiệu:

k

nio nij

j1

r

; noj  nij

i1

Như vậy n0j là kích thước của mẫu thứ j, còn nio là tổng số cá thể có tính trạng Ai trong toàn bộ k mẫu đang xét:

r k

n  nio  noj

i1 j1

là tổng số tất cả các cá thể của k mẫu đang xét.

* Nếu giả thiết H0 là đúng nghĩa là:

p1  p2   pk  p

1 1 1 1



p1  p2   pk  p



 2 2 2 2

p1  p2   pk  p

 i i i i



p1  p2   pk  p

 r r r r

thì các tỷ lệ chung p1, p2,...pr được ước lượng bởi:

p  nio

i n

Đó ước lượng cho xác suất để một cá thể có mang tính trạng Ai. khi đó số cá thể có tính trạng Ai trong mẫu thứ j sẽ xấp xỉ bằng:

nij  noj pi

n n

 oj io

n

Các số nij(i  1,2,...r; j  1,2,...k) được gọi là các tần số lý thuyết (TSLT), các số nij được gọi là các tần số quan sát (TSQS).

Ta quyết định bác bỏ Ho khi các TSLT cách xa TSQS một cách bất thường. Khoảng cách giữa TSQS và TSLT được đo bằng test thống kê sau đây:

k r

T  

f1 i1

nij  nij 

nij

2

(TSQS TSLT)2





TSLT

Người ta chứng minh được rằng nếu H0 đúng và các TSLT không nhỏ hơn 5 thì T sẽ có phân bố xấp xỉ phân bố với (k-1)(r-1) bậc tự do. Thành thử miền bác bỏ có dạng {T > c} ở đó c

được tìm từ điều kiện P{T > c} = . Vậy c là phân vị mức  của phân bố 2 với (k-1)(r-1) bậc tự do.

Đối với thí nghiệm có các kết quả, để so sánh các tỉ số của các kết quả đó, ta

dùng kiểm định 2 (chi-quared):

r c 2

y2 = ) )

i= 1 j= 1

(nij− npi) tổng ℎàng × tổng cột rớinpi =

npi n

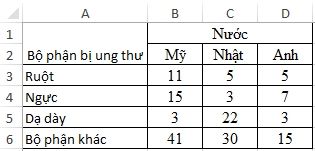
nij: tần số thực nghiệm; npij: tần số lý thuyết của ô (i,j); r: số hàng; c: số cột

Dùng hàm CHITEST (actual\_range, expected\_range). Tính giá trị: P(X > y2) = CHITEST

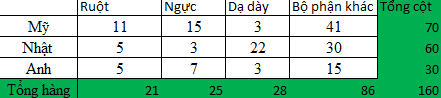
Nếu: P(X > y2) > α thì chấp nhận H0 và ngược lại.

#### Thực hiện bài toán bằng excel:

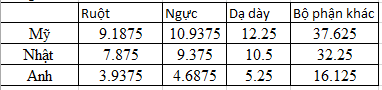
* Nhập dữ liệu vào bảng tính:



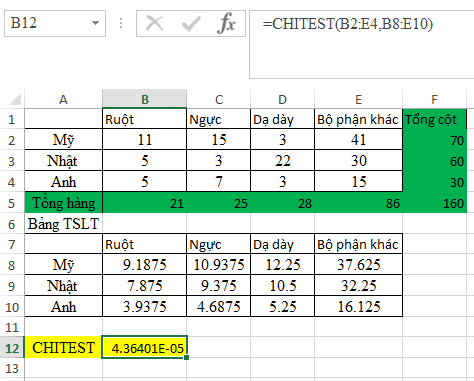
* Tính tổng các hàng và các cột.



* Tính các tần số lý thuyết: tần số lý thuyết = (Tổng hàng ×Tổng cột)/(Tổng cộng):



* + Sử dụng hàm CHITEST tính xác suất P(X> 2 ):



* + - Kết quả và biện luận:

Giả thiết H0: Phân bố tỉ lệ chết về ung thư của 3 nước như nhau. H1: Phân bố tỉ lệ chết về ung thư của 3 nước khác nhau.

Ta có: P(X> 2 ) = 4.36401E-05 < α=0.05

 Bác bỏ giả thiết H0, chấp nhận giả thiết H1

#### Vậy: phân bố tỉ lệ chết về ung thư của 3 nước khác nhau.

**BÀI 5:**

1. Tìm một dữ liệu ngẫu nhiên 2 chiều (X, Y) có kích thước n >10 để sử dụng mô hình hồi quy tuyến tính đơn. Thực hiện các yêu cầu:
   1. Tìm hệ số tương quan giữa X,Y.
   2. Quan hệ giữa X,Y có được coi như quan hệ tuyến tính hay không? Nếu có, hãy ước lượng đường hồi quy tuyến tính Y theo X.
   3. Tìm hệ số xác định R2.
   4. Tìm sai số chuẩn của ước lượng.
2. Xem Ví dụ 4.2 trong Chương 4, phần Phụ lục trong sách XSTK của trường. (Đề bài bắt đầu từ: Người ta đã dùng 3 mức nhiệt độ…). Thực hiện lại các yêu cầu của bài.

#### BÀI LÀM:

1. Bảng số liệu về thời gian thao tác trên máy tiện so với khối lượng phoi thải của gia công tiện được lấy ngẫu nhiên ở một số máy tiện của xưởng C1:

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Thời gian (phút) | 40 | 30 | 45 | 15 | 30 | 60 | 50 | 25 | 20 | 45 | 10 | 15 |
| Khối lượng phoi thải(gam) | 500 | 300 | 600 | 20 | 250 | 800 | 1200 | 150 | 50 | 200 | 150 | 200 |

* 1. **Tìm hệ số tương quan giữa X và Y:**

#### Cơ sở lý thuyết:

* Hệ số tương quan:

R =

∑ xiyi − ∑ xi∑ yi

J [n ∑ x2 − (∑ x )2][n ∑ y2 − (∑ y )2]

i

i

i

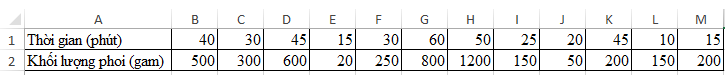
i

* 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 bậc nhất.
* Nếu |R |→ 1 thì X, Y có tương quan chặt (tương quan mạnh).

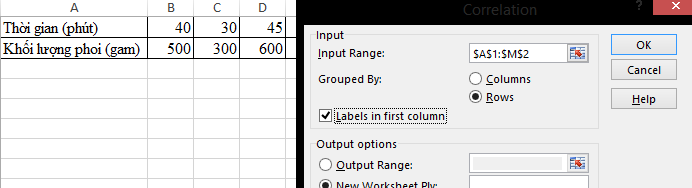
* Nếu |R |→ 0 thì X, Y có tương quan không chặt (tương quan yếu).

#### Thực hiện trên Excel:

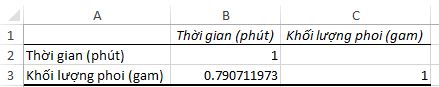
* Nhập số liệu vào bảng tính:



* Chọn chức năng Data/Data Analysis/Correlation.



* + Kết quả:



 Ta có hệ số tương quan là **R = 0.790711973** chứng tỏ giữa thời gian và khối lương phoi thu được có quan hệ khá chặc chẻ và có tương quan thuận.

## Quan hệ giữa X,Y có được coi như quan hệ tuyến tính hay không? Nếu có, hãy ước lượng đường hồi quy tuyến tính Y theo X.

#### Cơ sở lý thuyết:

Giả thiết H0: X và Y không có tương quan tuyến tính:

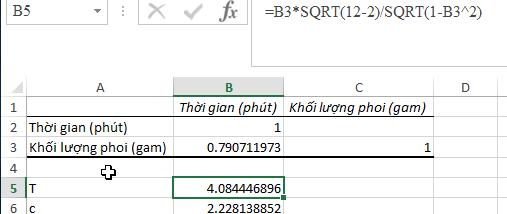
r√n − 2

T =

√ 1 − r2

#### Thực hiện trên Excel:

* Tính T : chọn ô B5 và nhập biểu thức =B3\*SQRT(12-2)/SQRT(1-B3^2).
* Tính c: chọn ô B6 và nhập biểu thức =TINV(0.05,10) (c là phân vị mức α/2=0.025 của phân bố Student với n-2=10 bậc tự do).



Vì |T| > c nên bác bỏ giả thiết H0.

Vậy: X và Y có tương quan tuyến tính.

## Ước lượng đường hồi quy tuyến tính Y theo X.

#### Cơ sở lý thuyết:

* Phương trình hồi quy tuyến tính:

y¯s¯ = a + bx , a = rS¯¯

S¯y¯

, b = y¯− ax̅

s

* Kiểm định hệ số a, b:

+

+

+

Giả thiết H0: Hệ số hồi quy không có ý nghĩa (=0).

H1: Hệ số hồi quy có ý nghĩa (≠0).

Trắc nghiệm t < t ,n-2: chấp nhận H0

* Kiểm định phương trình hồi quy:

+

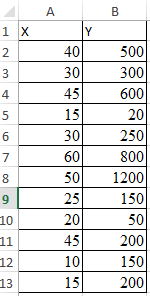
+

Giả thiết H0: “Phương trình hồi quy tuyến tính không thích hợp”.

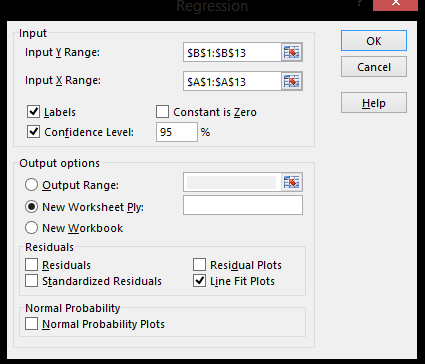
H1: “Phương trình hồi quy tuyến tính thích hợp”.

Trắc nghiệm F < F ,1,n-2: chấp nhận H0

* + **Thực hiện trên Excel:**
* Nhập số liệu vào bảng tính:

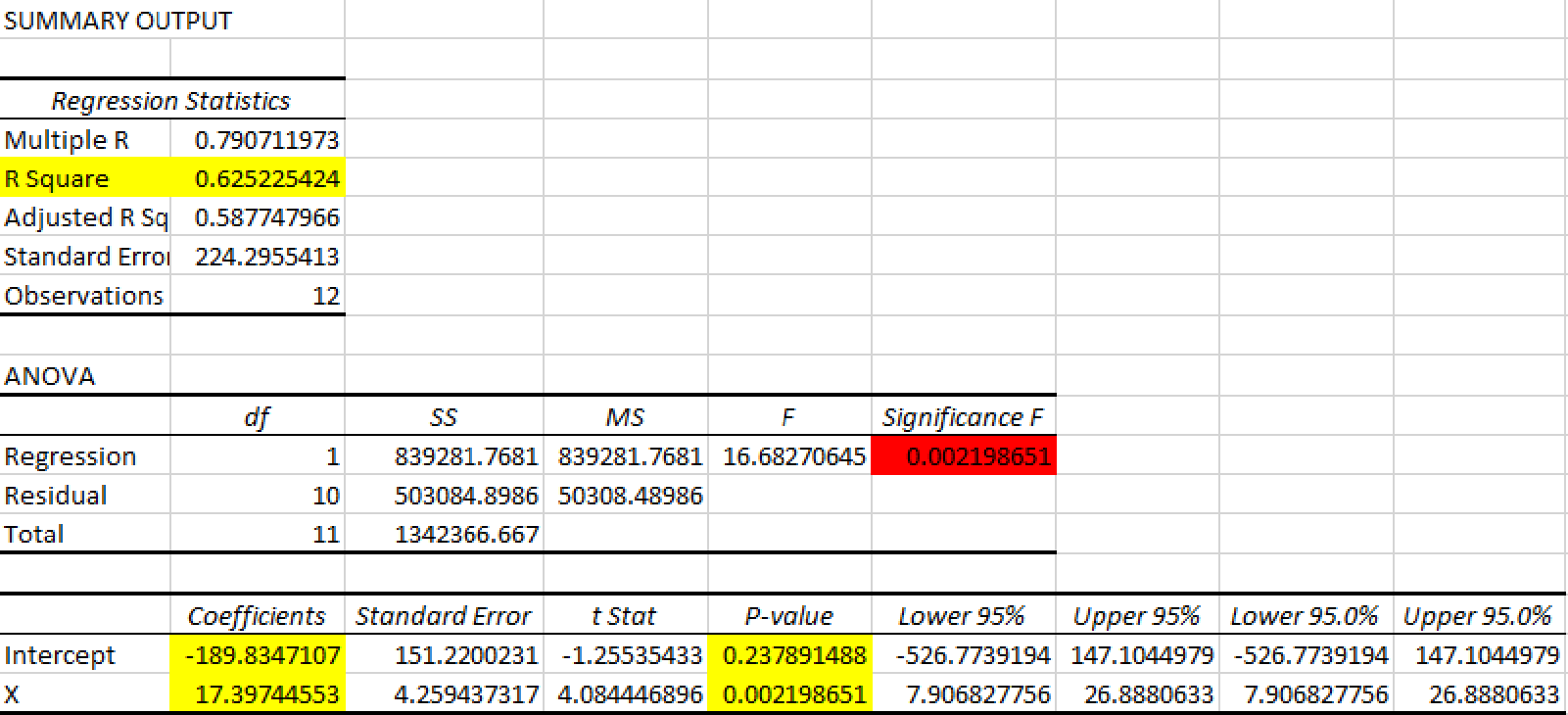


* Dùng chức năng Data/Data Analysis/Regression.



* Kết quả:

**Y**



#### Biện luận:



1400

1200

1000

800

Y

Predicted Y Linear (Y)

600

400

200

0

0

20

40

**X**

60

80

-200

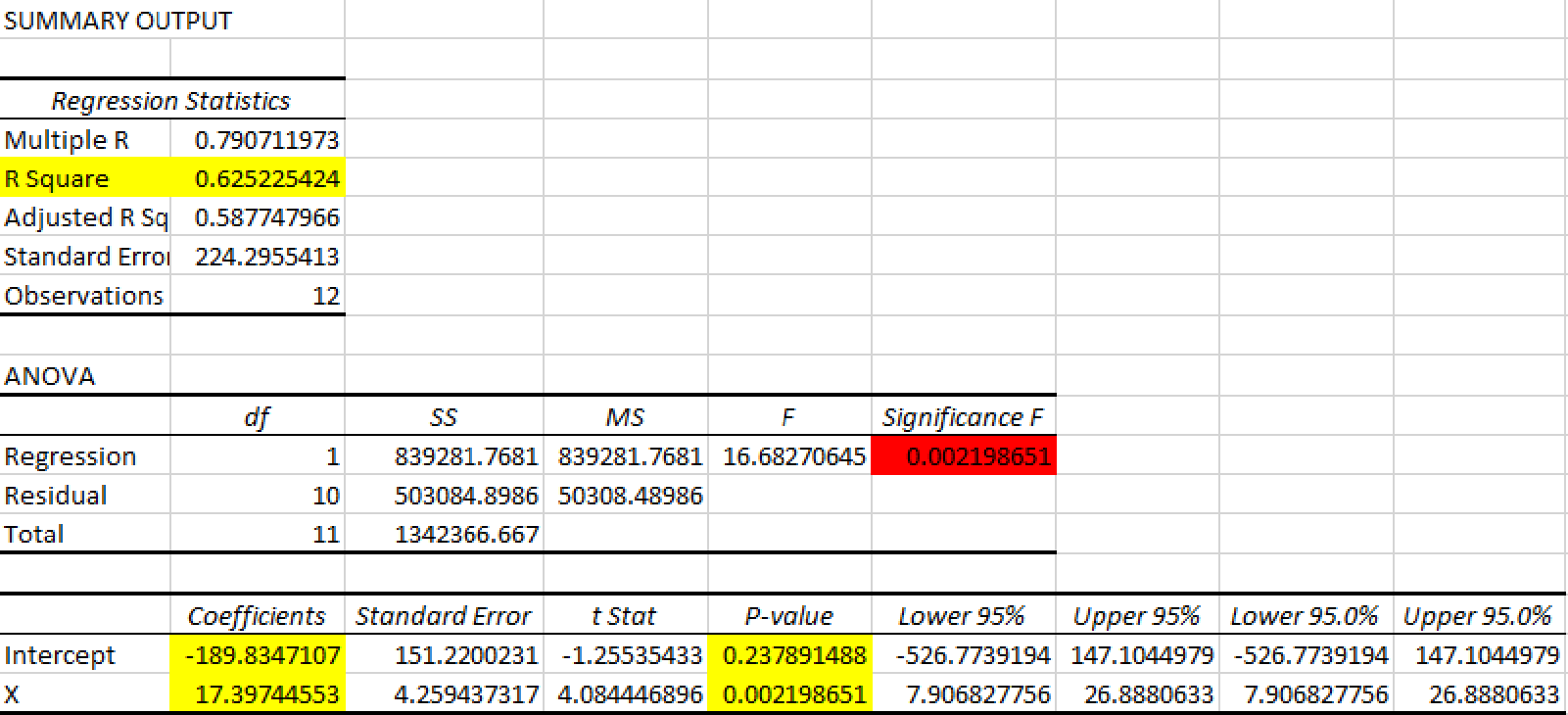
Phương trình hồi quy: y¯s¯ = − 1898347+ 17.3974x

Hệ số hồi quy: 0.2379 > 0.05  Hệ số tự do có ý nghĩa.

0.00219 < 0.05  Hệ số của x không có ý nghĩa.

 Phương trình hồi quy tuyến tính này không thích hợp vì 0.002198 < 0.05

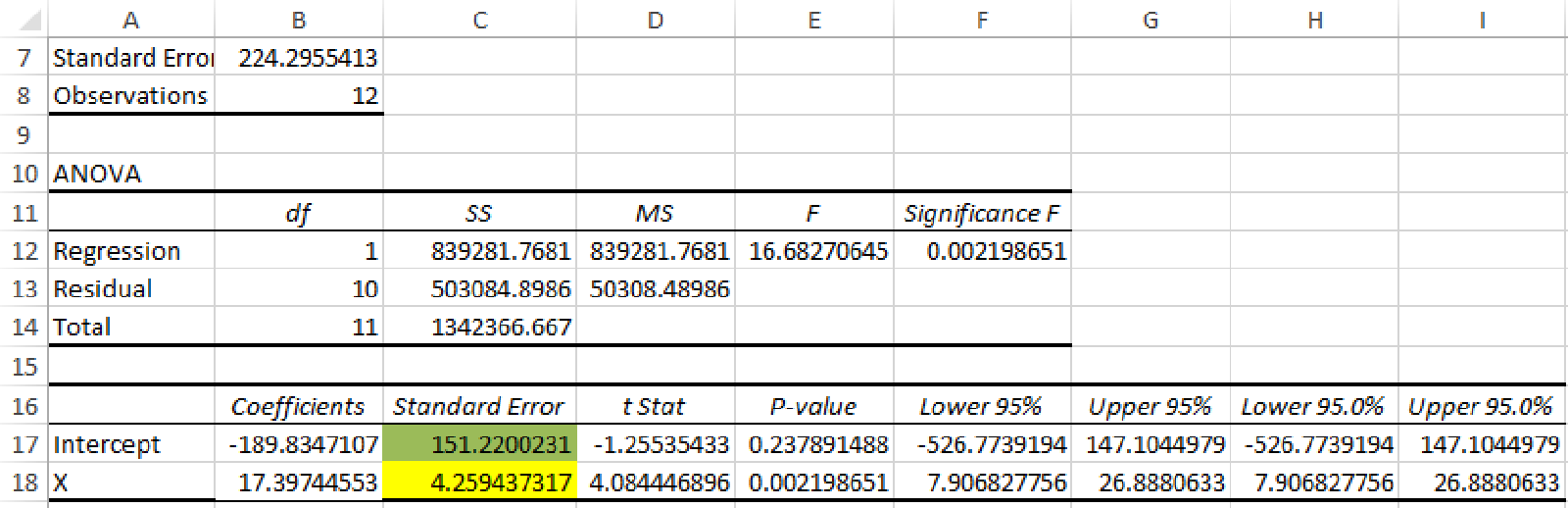
## Tìm hệ số xác định R2:



 Dùng kết quả từ bảng SUMMARY OUTPUT từ câu trên ta xác định được hệ số hồi quy: **R2 = 0.6252**

## Tìm sai số chuẩn của ước lượng:

 Dùng kết quả từ bảng SUMMARY OUTPUT từ câu trên ta xác định được sai số chuẩn của ước lượng:



* Đối với biến tự do: SE = 151,2200
* Đối với biến X: SE = 4.2594

#### Ví dụ 4.2:

Người ta dùng ba mức nhiệt độ gồm 105, 120 và 135oC 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 (phút) X1 | Nhiệt độ (oC)  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 hoặc yếu tố thời gian có liên quan tính tuyến với hiệu suất của phản ứng tổng hợp? Nếu có thì điều kiện nhiệt độ 115oC trong vòng 50 phút thì hiệu suất phản ứng sẽ là bao nhiêu?

#### BÀI LÀM:

* **Dạng bài:** Hồi quy tuyến tính đa tham số.

#### Cơ sở lý thuyết:

Phương trình tổng quát cho biến phụ thuộc Y có liên quan đến k biến số độc lập Xi (i=1,2,...,k):

YˆX1,X2,… ,Xk = B0 + B1X1 + B2X2 + … + BkXk

* Bảng ANOVA:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Nguồn sai số | Bậc tự do | Tổng số bình phương | Bình phương trung bình | Giá trị thống kê |
| Hồi quy | k | SSR | MSR = SSR  k | F = MSR  MSR |
| Sai số | N - k - 1 | SSE | MSE = SSE  (N – k – 1) |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tổng cộng | N - 1 | SST = SSR + SSE |  |  |

* Giá trị thống kê:
  + Giá trị R-bình phương: Giá trị R2:

R 2 = SSR = kF

(R3 ≤ 0.81 là khá tốt)

SST (N –k–1)+ kF

Giá trị R2 được hiệu chỉnh (Adjusted R Square)

2 = (N − 1)R

2

− k k(1 − R 2)

= R 2 −

Rii

(N − k − 1)

(N − k − 1)

(R 2 sẽ trở nên âm hay không xác định nếu R2

ii

hay N nhỏ)

o Độ lệch chuẩn:

S = J

SSN (N –k–1)

(S ≤ 0.30 là khá tốt)

* Trắc nghiệm thống kê**:**
  + Trắc nghiệm t:

H O : Bi = 0 “Các hệ số hồi quy không có ý nghĩa”

HO : Bi ≠ 0 “Có ít nhất vài hệ số hồi quy có ý nghĩa”

Bậc tự do của t: y = N - k - 1

|Bi–þi| S2

i

t=

J S2

n

; Sn = ∑ (X –X¯)2

* + Trắc nghiệm F:

H O : Bi = 0 “phương trình hồi quy không thích hợp”

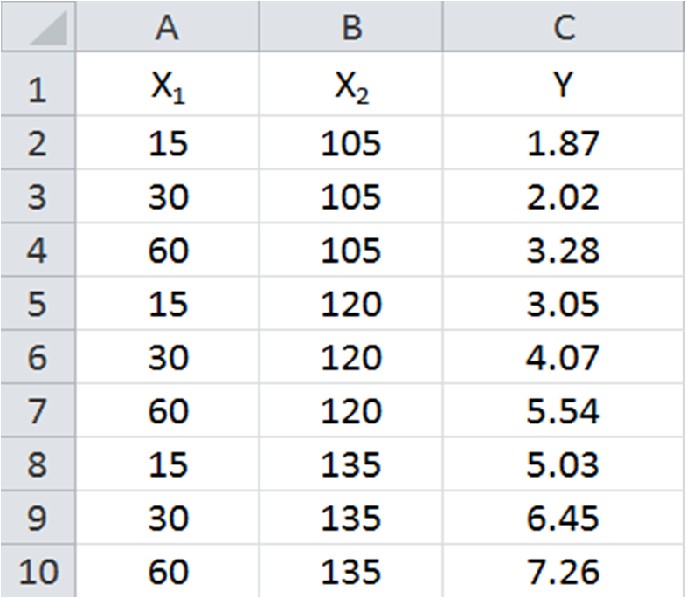
HO : Bi ≠ 0 “phương trình hồi quy thích hợp” với ít nhất vài Bi. Bậc tự do của giá trị F: v1 = 1, v2 = N -k – 1

#### Thực hiện trên Excel:

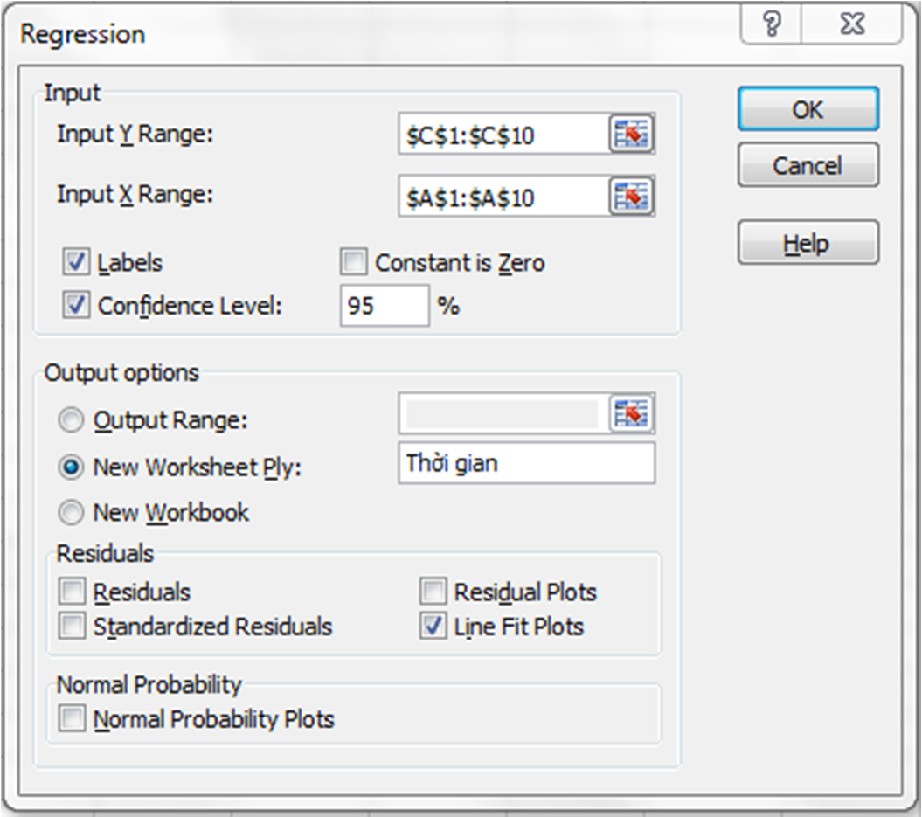
Giả thiết H0: Phương trình hồi quy không thích hợp.

Ta tìm phương trình hồi quy tính tuyến đa tham số để chỉ ra sự phụ thuộc hoặc không phụ thuộc giữa yếu tố thời gian (X1) và nhiệt độ (X2) với hiệu suất phản ứng tổng hợp (Y).

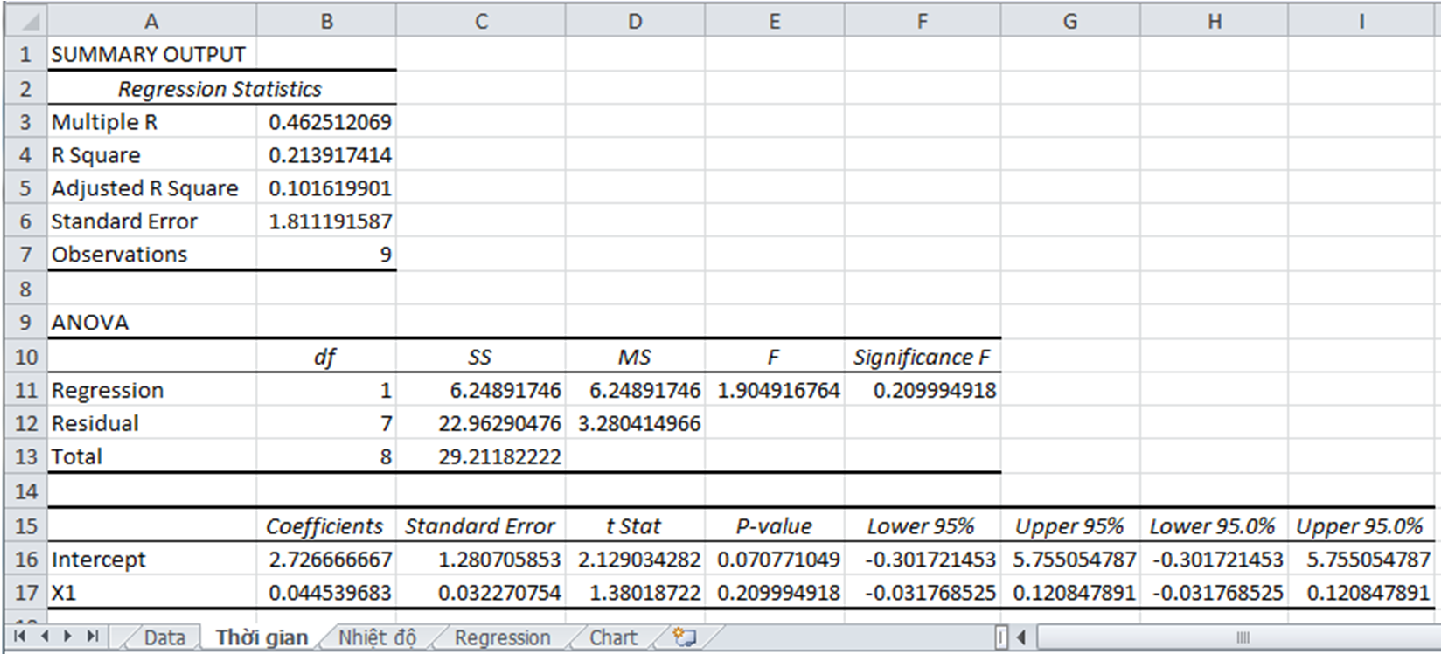
* + Nhập dữ liệu theo cột:



* + Chọn chức năng Data/ Data Analysis/Regression.
* Hồi quy theo Thời gian (X1):
  + Input Y Range: *Phạm vi biến số Y*
  + Input X Range: *Phạm vi biến số X*
  + Labels: *Dữ liệu bao gồm nhãn*
  + Confidence Level: *Mức tin cậy*
  + Output options: Nơi xuất kết quả.



* + - Kết quả:



* + - Phương trình hồi quy: ŶX1 = f(X1) = 2.7267 + 0.0445X1 với R2 = 0.2139 và S = 1.8112

t0 = 2.1290 < t0.05 = 2.365 (tra bảng VII với n = 7, α = 0.025) hay P2 = 0.0708 > α = 0.05

V

 Chấp nhận giả thiết H0.

t1 = 1.3802 < t0.05 = 2.365 hay PV = 0.209 > α = 0.05

 Chấp nhận giả thiết H0.

3 4

F = 1.9049 < FO.O5 = 5.590 (tra bảng VIII với n1 = 1 và n2 = 7) hay FS

= 0.209 > α = 0.05

 Chấp nhận giả thiết H0.

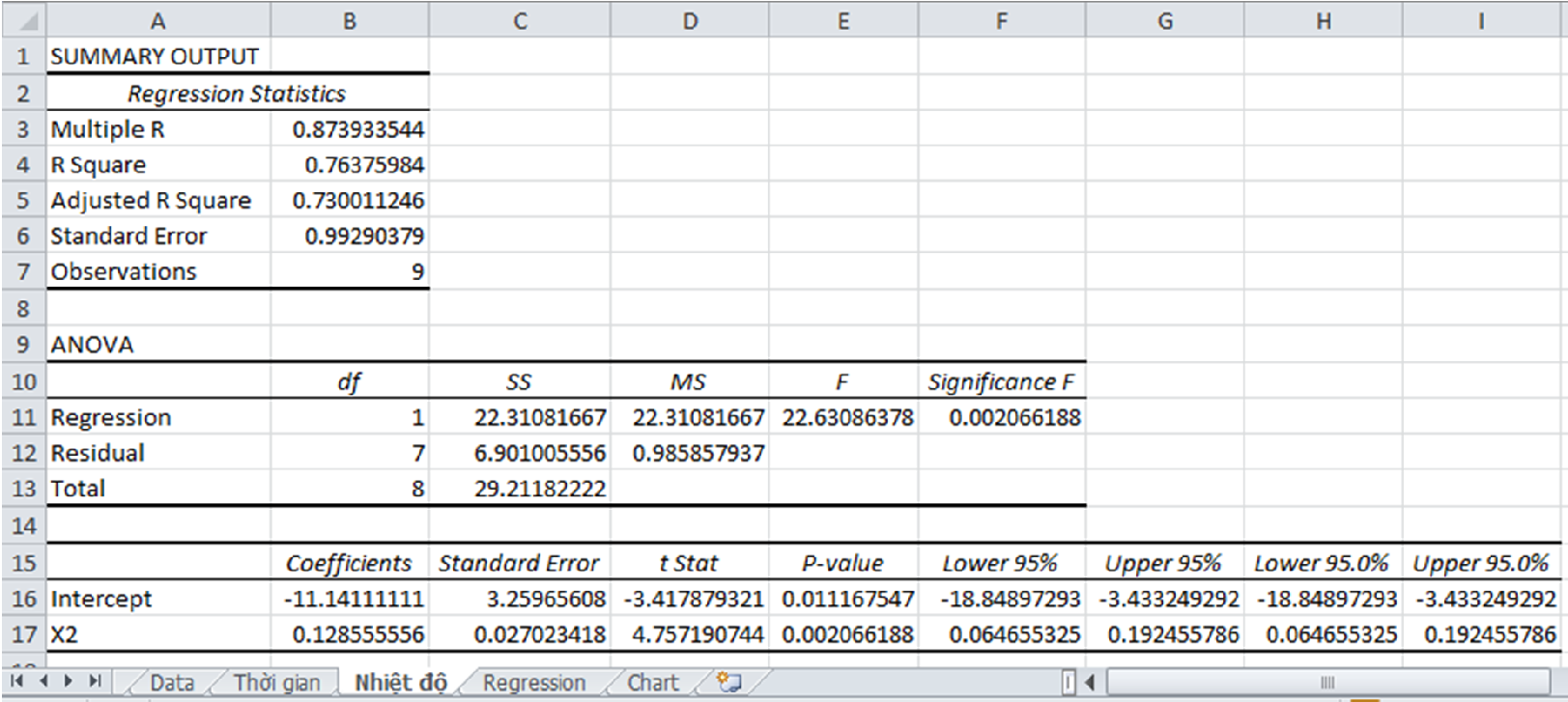
#### Vậy: Phương trình hồi quy trên không có ý nghĩa thống kê. Nói 1 cách khác, phương trình hồi quy này không thích hợp.

**Kết luận: Yếu tố thời gian không có liên quan tính tuyến với hiệu suất của phản ứng tổng hợp.**

* **Hồi quy theo Nhiệt độ (X2):**

Các thông số ở cửa sổ Nhiệt độ như Hồi quy theo X1, trừ Input X Range là ***$B$1:$B$10***

* + Kết quả:



Phương trình hồi quy:

ŶX2 = f(X2) = -11.1411 + 0.1286X2 với R2 = 0.7638 và S = 0.9929 t0 = 3.4179 > t0.05 = 2.365 hay P2 = 0.0112 < α = 0.05

V

 Bác bỏ giả thiết H0.

t1 = 4.7572 > t0.05 = 2.365 hay PV = 0.0021 < α = 0.05

 Bác bỏ giả thiết H0.

3 4

F = 22.6309 > FO.O5 = 5.590 hay FS

= 0.0021 < α = 0.05

 Bác bỏ giả thiết H0.

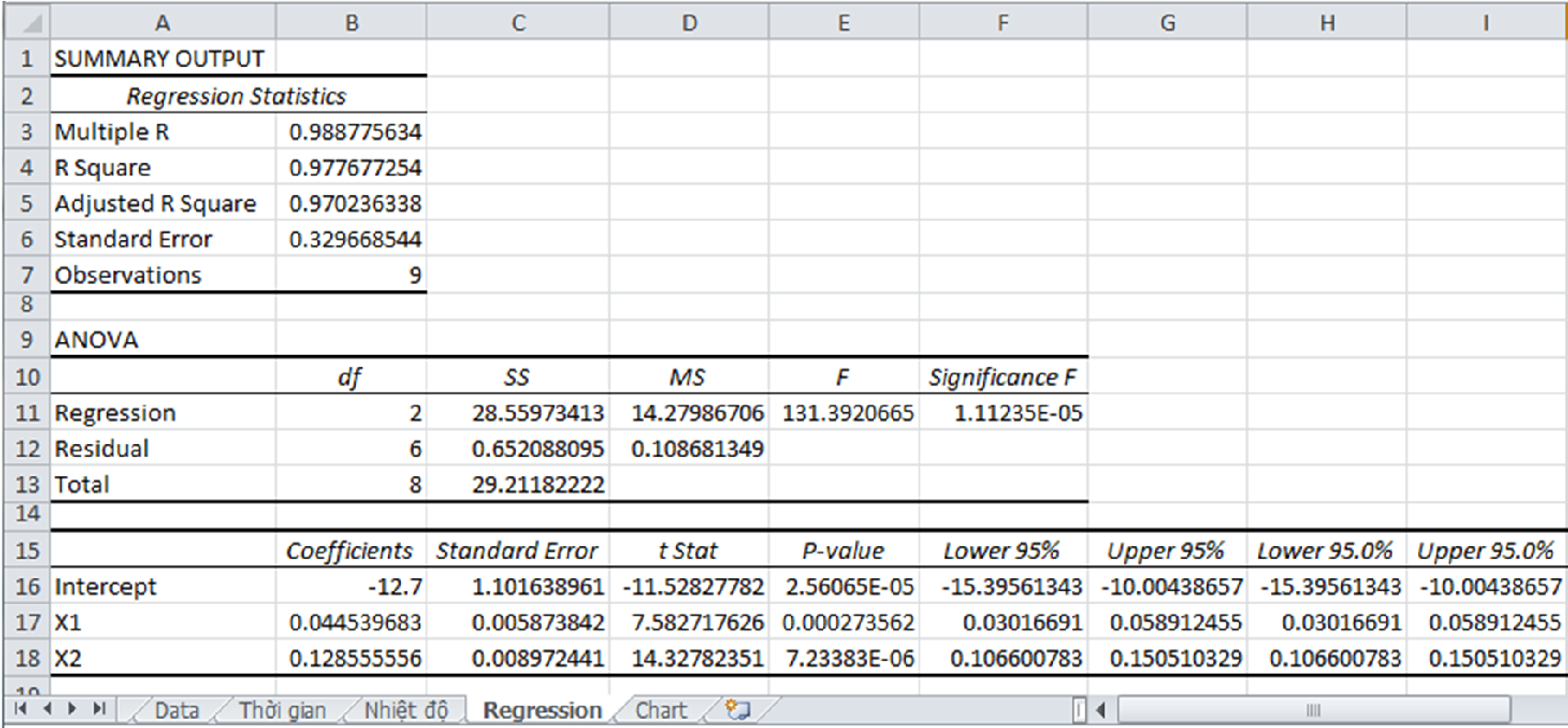
#### Vậy: Phương trình hồi quy trên có ý nghĩa thống kê. Nói 1 cách khác, phương trình hồi quy này thích hợp.

**Kết luận: Yếu tố nhiệt độ có liên quan tính tuyến với hiệu suất của phản ứng tổng hợp.**

* **Hồi quy theo Thời gian (X1) và Nhiệt độ (X2):**

Các thông số ở cửa sổ Regression như Hồi quy theo X1, trừ Input X Range là ***$A$1:$B$10***

* + Kết quả:



* + Phương trình hồi quy:

ŶX1, X2 = f(X1,X2) = -12.7000 + 0.0445X1 + 0.1286X2 với R2 = 0.9777 và S = 0.3297

2 –5

t0 = 11.5283 > t0.05 = 2.365 hay PV = 2.5607\*10

< α = 0.05

 Bác bỏ giả thiết H0.

t1 = 7.5827 > t0.05 = 2.365 hay PV = 0.0003 < α = 0.05

 Bác bỏ giả thiết H0.

t2 = 14.3278 > t0.05 = 2.365 hay PV = 7.2338\*10–6 < α = 0.05

 Bác bỏ giả thiết H0.

F = 131.3921 > F0.05 = 5.140 (tra bảng VII với n1 = 2 và n2 = 6) hay FS = 1,1123\*10–5 < α = 0.05

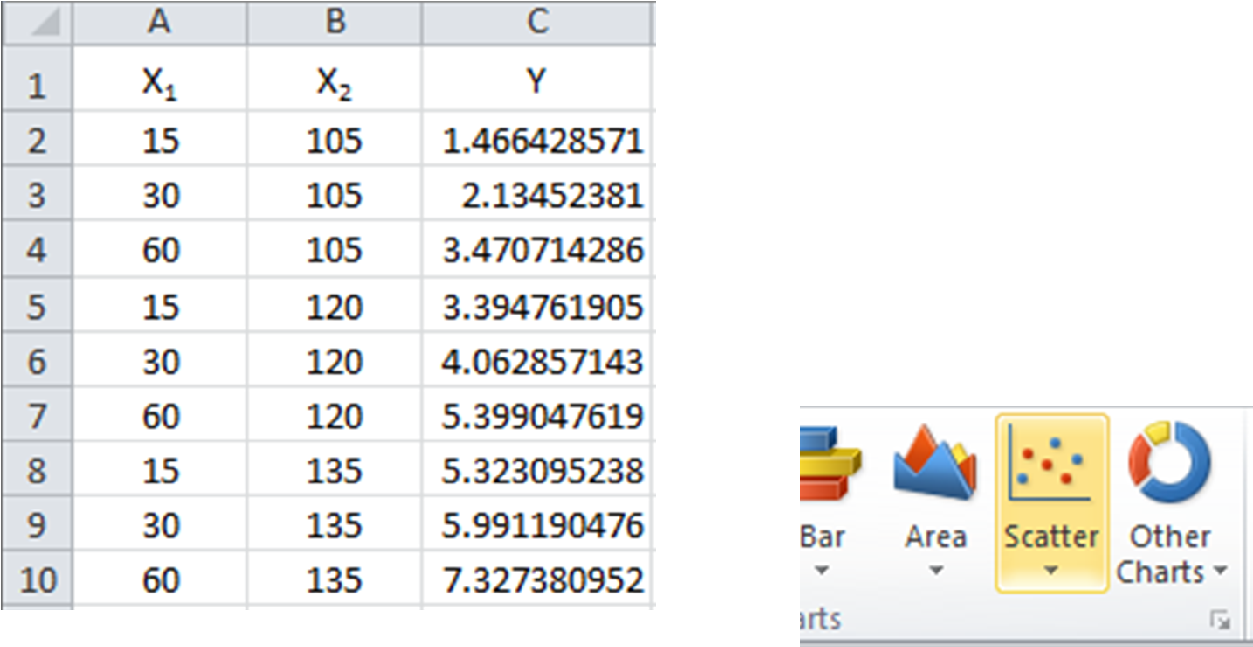
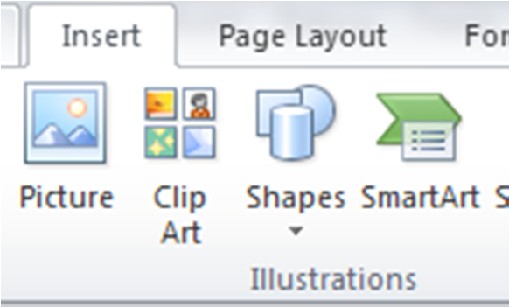
 Bác bỏ giả thiết H0.

#### Vậy: Phương trình hồi quy trên có ý nghĩa thống kê. Nói 1 cách khác, phương trình hồi quy này thích hợp.

**Kết luận: Hiệu suất phản ứng có liên quan tính tuyến với cả hai yếu tố là thời gian và nhiệt độ.**

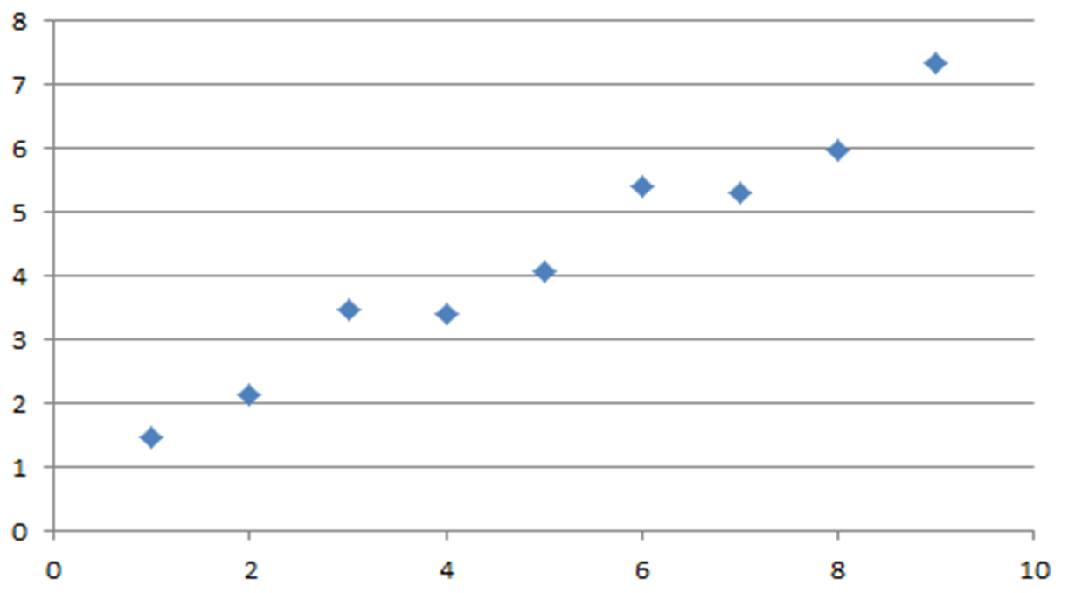
Dữ liệu với hàm hồi quy YˆX ,X **= -12.7000 + 0.0445X1 + 0.1286X2**:

1 2



* Vẽ biểu đồ: chọn ô C2, vào Insert -> Scatter -> Scatter with only Maker

Sự tính tuyến của phương trình hồi quy YX1, X2 = -12.7000 + 0.0445X1 + 0.1286X2 có thể được trình bày trên biểu đồ phân tán:



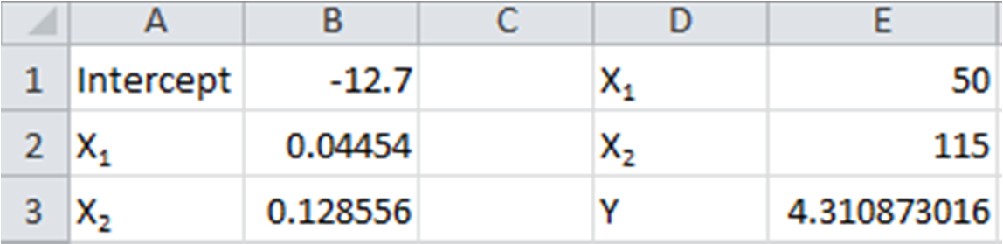
Hàm lượng thực nghiệm (Y)

Hàm lượng dự đoán (Y’)

#### Dự đoán hiệu suất của phản ứng bằng phương trình hồi quy tại nhiệt thời gian (X1) 50 phút, nhiệt độ (X2) 115oC:

* + Thực hiện:

Công thức ô *E3*: *=B1+B2\*E1+B3\*E2*



 **Kết quả: 4.3109**

#### Vậy hiệu suất của phản ứng bằng phương trình hồi quy tại nhiệt thời gian (X1) 50 phút, nhiệt độ (X2) 115oC là 4.3109