**ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC BÁCH KHOA**

**---------------o0o---------------**

****

**BÁO CÁO BÀI TẬP LỚN**

**MÔN HỌC: XÁC SUẤT THỐNG KÊ**

**GVHD: NGUYỄN KIỀU DUNG**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **Họ và tên** | **MSSV** | **Lớp/Tổ** | **Ngành học** | **Ký tên tham dự** |
| 1 | Phí Thành Danh | 1910914 | L05 | Điện-Điện tử |  |
| 2 | Trần Quang Duy | 1910096 | L05 | Điện-Điện tử |  |
| 3 | Phạm Vũ Hoàng | 1911204 | L05 | Điện-Điện tử |  |
| 4 | Lê Đình Hưng | 1911301 | L05 | Điện-Điện tử |  |
| 5 | Huỳnh Nhật Minh | 1911595 | L05 | Điện-Điện tử |  |
| 6 | **Nguyễn Ngọc Thanh Tuấn** | 1912365 | L05 | Điện-Điện tử |  |

**TP. HỒ CHÍ MINH, THÁNG 12 NĂM 2020**

**MỤC LỤC**

[**ĐỀ BÀI 2**](#_Toc31519)

[**1. CÂU 1: 3**](#_Toc23237)

[**2. CÂU 2: 6**](#_Toc21610)

[**3. CÂU 3: 10**](#_Toc28710)

[**4. CÂU 4: 19**](#_Toc8774)

[**5. CÂU 5: 23**](#_Toc15171)

[**6. CÂU 6: 29**](#_Toc2952)

**ĐỀ BÀI**

Xây dựng một ( hay nhiều) tập dữ liệu có số quan sát n > 100; có ít nhất 1 biến định

tính và 3 biến định lượng về cùng 1 chủ đề. Các tập dữ liệu này phải được ghi rõ

nguồn gốc, đường link (nếu là tham khảo) hoặc có minh chứng về bảng câu hỏi,

file dữ liệu excel (nếu do tự thu thập) gửi kèm bài nộp.

Sử dụng toàn bộ hoặc một phần của tập dữ liệu thực hiện các công việc sau:

Câu 1: Chọn 1 biến định lượng nào đó và thực hiện:

- Tìm các đặc trưng từ mẫu dữ liệu.

- Tìm các khoảng tin cậy cho giá trị trung bình và phương sai của tổng thể.

Câu 2: Chọn dữ liệu 2 biến định lượng (hoặc xử lý số liệu theo nhóm cho phù hợp)

để lập bài toán kiểm định so sánh 2 trung bình tổng thể.

Câu 3: Chọn dữ liệu cho k biến (k >= 3) (hoặc xử lý số liệu theo nhóm cho phù

hợp) để lập bài toán so sánh về trung bình k tổng thể.

Câu 4: Chọn dữ liệu cho 2 biến (hoặc xử lý số liệu theo nhóm cho phù hợp) để lập

bài toán kiểm định so sánh về sự phân tán. Trình bày các bước thực hiện và

nhận xét kết quả.

Câu 5: Chọn dữ liệu phù hợp để lập bài toán so sánh các tỉ lệ ( hay là bài toán kiểm

định tính độc lập). Trình bày các bước thực hiện và nhận xét kết quả.

Câu 6: Khảo sát hệ số tương quan giữa 2 biến định lượng cụ thể, dự đoán phương

trình đường hồi quy tuyến tính giữa chúng ( có hình vẽ minh họa) và nhận xét về mối quan hệ tuyến tính giữa 2 biến.

1. **CÂU 1:**

# **LẬP BÀI TOÁN**

Cho bảng số liệu thống kê thời gian ngủ thường ngày (giờ) của một mẫu sinh viên bách khoa K19

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Thời gian ngủ | 3.5 | 4.5 | 5.5 | 6.5 | 7.5 | 9.5 |
| Số sinh viên | 6 | 22 | 24 | 32 | 14 | 3 |

-Hãy tìm đặc trưng của bảng số liệu

-Hãy tìm khoảng tin cậy cho giá trị trung bình và phương sai của tổng thể

# **B. CƠ SỞ LÝ THUYẾT**

Đặc trưng mẫu

* Trung bình cộng được tính bằng cách cộng tất cả các giá trị quan sát trong tập dữ liệu lại rồi đem kết quả đó chia cho kích thước . Công thức tính

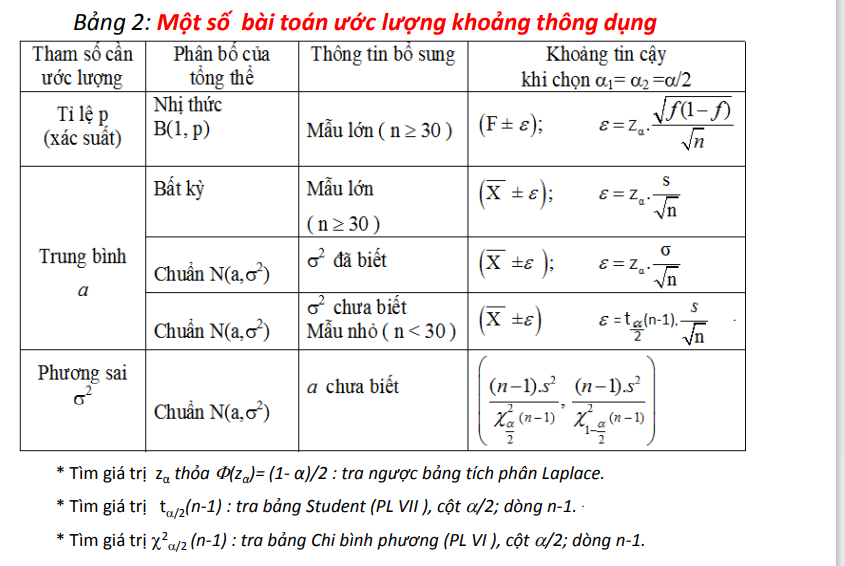
=

* Phương sai mẫu được định nghĩa bằng trung bình của bình phương sai lệch giữa các giá trị với trung bình cộng của nó. Công thức tính

2

* Độ lệch mẫu là căn bậc hai của phương sai : =
* Phương sai mẫu hiệu chỉnh : s2 = 2 =
* Độ lệch mẫu hiệu chỉnh là s =

Khoảng ước lượng cho giá trị trung bình và phương sai tổng thể :

[](file:///E:\~$Câu%201.xlsx)

# **C. LỜI GIẢI**

Gọi: x là thời gian ngủ thường ngày của sinh viên K19

m là số sinh viên ngủ 7.5 tiếng

* Các đặc trưng:
* Kích thước của mẫu: n =101
* Trung bình mẫu:
* Phương sai mẫu:
* Độ lệch mẫu:

=1.2886

* Phương sai mẫu hiệu chỉnh:
* Độ lệch mẫu hiệu chỉnh: s = = 1.2950
* Tỉ lệ mẫu:

Ước lượng khoảng tin cậy:

- Khoảng ước lượng giá trị trung bình tổng thể có dạng: (

+ Chọn độ lệch chuẩn tổng thể: σ = 1.31

+ Chọn mức ý nghĩa α = 5%

+ Ta có:

🡺

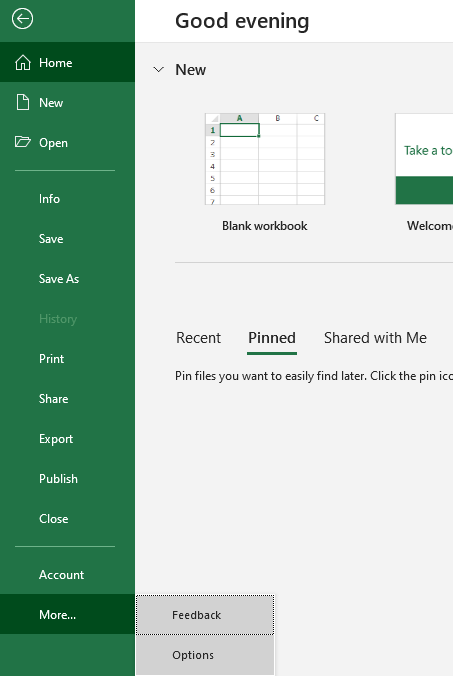
Vì n>30, đã biết σ nên ta có

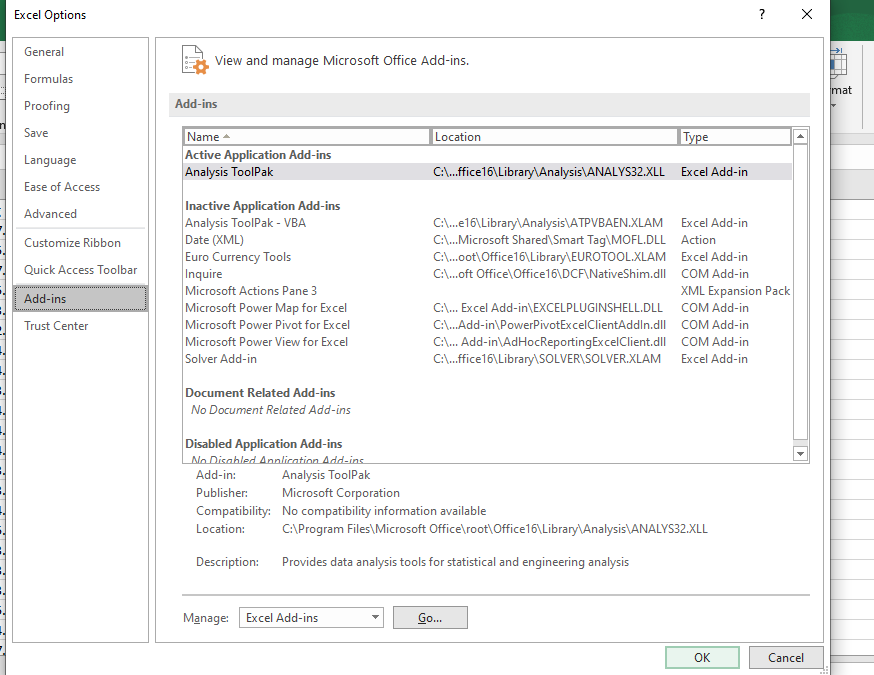
+ Vậy khoảng tin cậy trung bình tổng thể: (

- Khoảng tin cậy phương sai là:  
 = ( 1,294 ; 2,259)

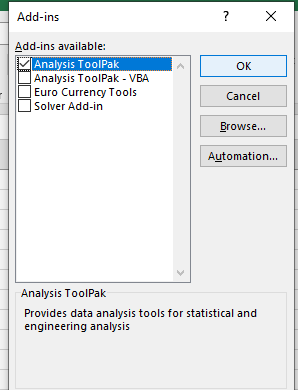
= 129.56 , = 74.22 tra từ bảng phân phối Chi-square trên mạng

# **LỜI GIẢI SỬ DỤNG EXCEL**

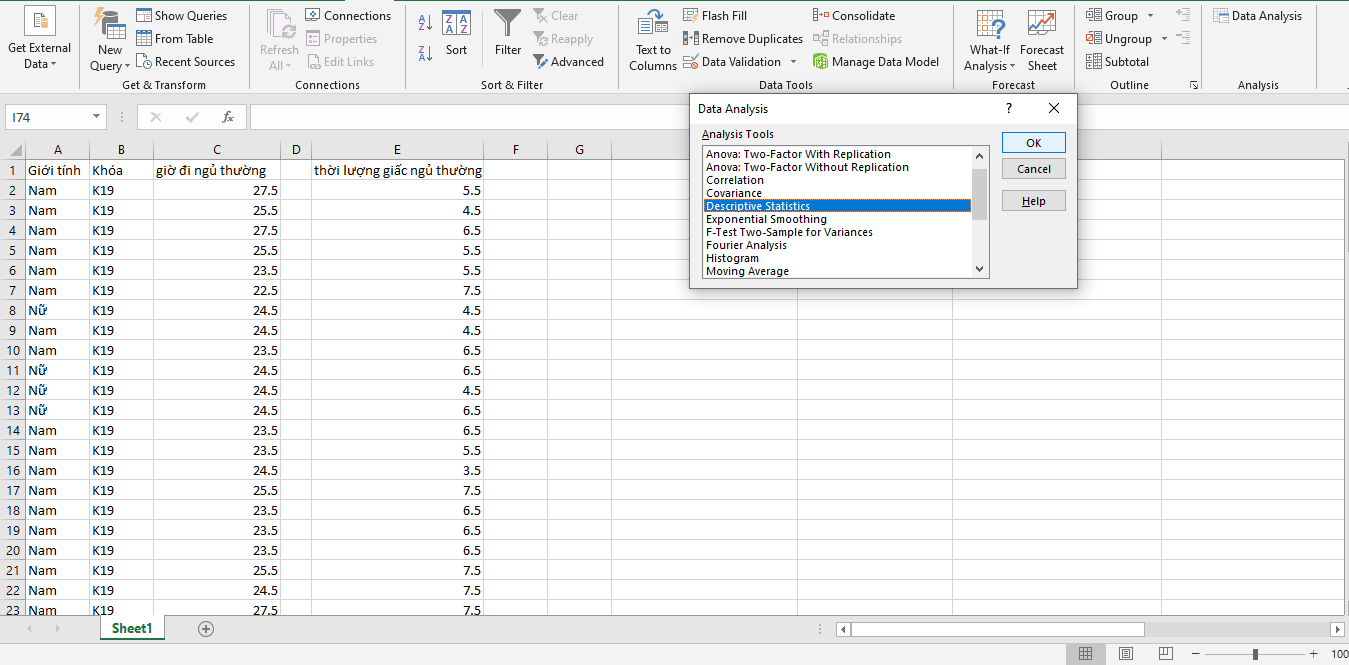
**Bước1:** Options Add-ins Analysis Toolpalk Go



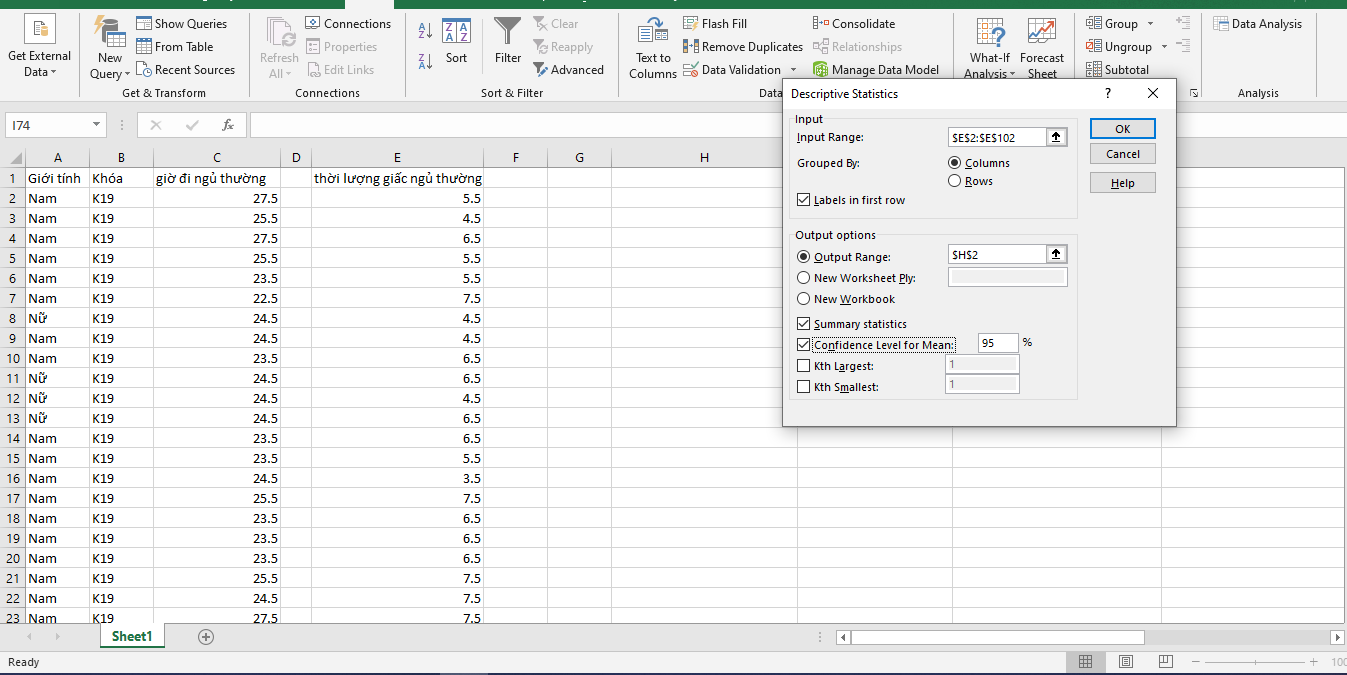
**Bước 2:** Tích chọn Analysis Toolpalk



**Bước 3:** Chọn Data Data Analysis Descriptive Statistics OK



**Bước 4:** Nhập dữ liệu cần tính vào Input Range, tích chọn Summary statistics, vị trí xuất kết quả ở Output Range, chọn độ tin cậy (Confidence Level for Mean) = (1-α)%



1. **CÂU 2:**
2. **LẬP BÀI TOÁN**

Cho bảng số liệu tình trạng tinh thần sau khi dậy của sinh viên HCMUT và mức độ căng thẳng trong học tập với alpha=0,05:

|  |  |  |
| --- | --- | --- |
| Mức căng thẳng | Nam | Nữ |
| 0 | 12 | 3 |
| 1 | 22 | 13 |
| 2 | 44 | 27 |
| 3 | 25 | 17 |
| 4 | 11 | 2 |
| Trung bình | 2,00877193 | 2,032258065 |
| Phương sai đã hiệu chỉnh | 1,230010868 | 0,81861449 |

- Lập bai toán so sánh trung bình 2 tổng thể kể trên.

**B. CƠ SỞ LÝ THUYẾT**

Khái niệm về kiểm định: Việc dùng kết quả của mẫu để khẳng định hay bác bỏ một giả thiết H nào đó được gọi là kiểm định giả thiết H. Khi kiểm định ta có thể mắc 1 trong 2 loại sai lầm sau:

- Sai lầm loại 1: Là sai lầm khi ta bác bỏ H nếu H đúng. Ta ký hiệu xác suất để mắc sai lầm loại này là  và gọi là mức ý nghĩa.

- Sai lầm loại 2: Là sai lầm mắc phải khi ta công nhận H khi H sai. Ta ký hiệu xác suất mắc sai lầm loại này là và gọi 1- là lực kiểm định.

Quy trình kiểm định:

+ Tiêu chuẩn kiểm định: xét mẫu W = . Chọn thống kê: G = f.

Điều kiện đặt ra đối với thống kê G là nếu H đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định .

+Miền bác bỏ giả thuyết H: Cho trước , miền bác bỏ  là miền thỏa:

. Có vô số miền  như vậy, trong số đó ta sẽ chọn miền bác bỏ  sao cho xác suất mắc sai lầm loại 2 là nhỏ nhất nghĩa là sao cho:

.

Giả thuyết: H: P = P0.

Giả thuyết đối lập  - (thiếu miền bác bỏ bên trái).

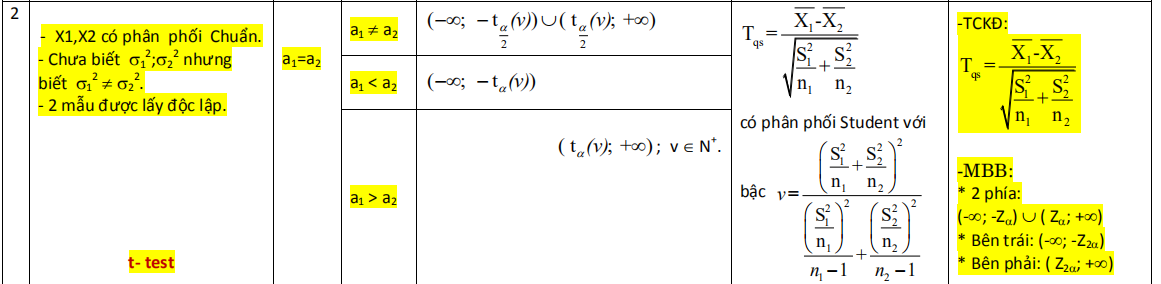
- (thừa miền bác bỏ bên phải).

- (đối xứng).

\* Kiểm định giả thiết về giá trị trung bình: (cho nhiều mẫu phát triển từ 2 mẫu trở lên).

Kí hiệu trung bình của tổng thể 1,2 là ( cả hai chưa biết).Từ các tổng thể lấy các mẫu kích thước  có trung bình mẫu  và phương sai hiệu chỉnh mẫu . Với mức ý nghĩa , hãy kiểm định giả thiết:

H: 



# **C. LỜI GIẢI**

H0: tình trạng mệt sau khi ngủ của sinh viên không phụ thuộc vào giới tính (X1=X2=Xtb)

Đặt giải thuyết đối lập

Ta có những thông số:

* Kích thước mẫu: n1=114; n2=62
* Trung bình mẫu ( ):
* Phương sai hiệu chỉnh ():

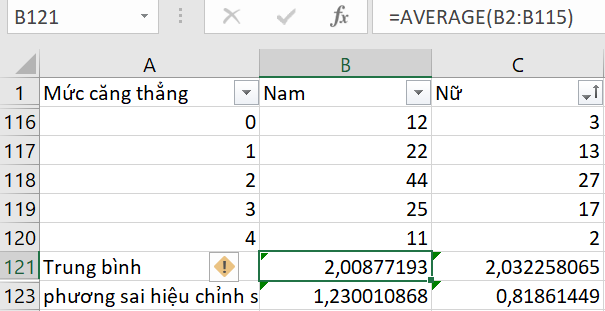
= 1,230010868; = 0,81861449

Tra bảng phân phối chuẩn Laplace (n1,n2>30), ta có hàng 20 cột 6

Vậy miền bác bỏ là (; ) (;)

* X1; X2 phân phối chuẩn (giả thuyết đề); chưa biết phương sai tổng thể (mình chỉ có của mẫu); và 2 mẫu độc lập nên ta dùng công thức

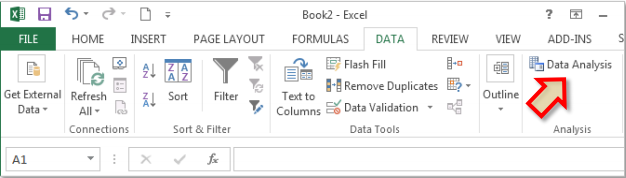
Zqs =0,151688932

**Kết luận:** Vì **Tqs không thuộc miền bác bỏ** nên chưa thể bác bỏ được giả thiết thể trạng tinh thần trung bình sau khi ngủ của nam và nữ khác nhau

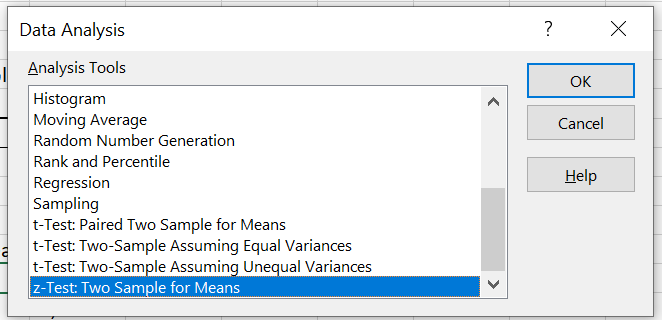
# **D. LỜI GIẢI SỬ DỤNG EXCEL**

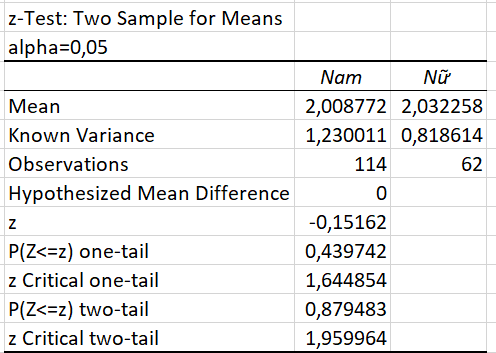
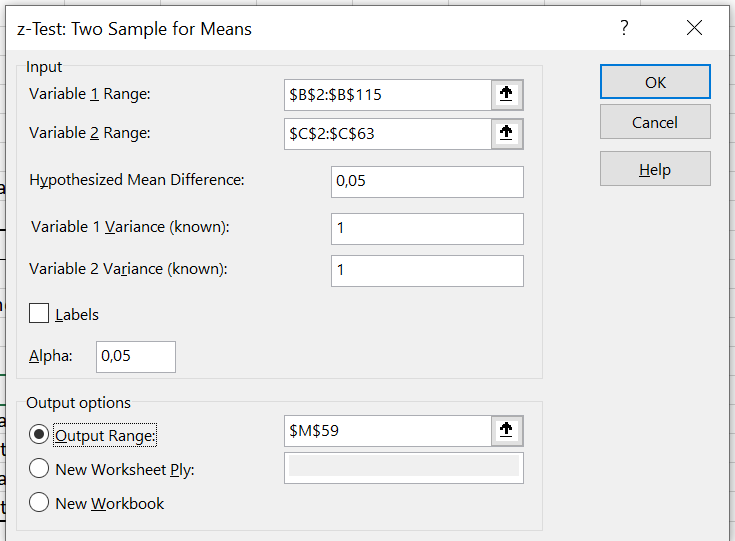
**Bước 1:** Tạo 3 cột là các biến độc lập ứng với 2 giới nam, nữ và hàng là mức độ mệt của các sinh viên sau khi dậy.

**Bước 2:** Dùng công cụ Data Analysis ở Data. Nếu chưa có phải trở về File-> Option->Excell Option-> Add -Ins -> Analysis Toolpack -> Go -> Analysis Toolpack -> Go



**Bước 3:** Sử dụng công cụ Data Analysis và chọn z-Test: Two-sample Assuming Unqual Variance ->Ok.



**Bước 4:** Chọn vùng dữ liệu nam vừa thêm ở bước 1 vào Variable 1 Range và nữ ở V2R, chọn vị trí xuất bảng giá trị sau khi tính toán ở Output Range và nhấn OK. Ta sẽ được các giá trị cần tính. Sau đó ta đọc giá trị để kết luận việc có bác bỏ giá thiết hay không.

1. **CÂU 3:**
2. **LẬP BÀI TOÁN**

Cho bảng số liệu thời gian ngủ trưa (phút) của sinh viên ĐHBK các khóa 2017,2018 và 2020:

|  |  |  |
| --- | --- | --- |
| K17 | K18 | K20 |
| 135 | 135 | 105 |
| 75 | 15 | 45 |
| 15 | 105 | 45 |
| 45 | 75 | 75 |
| 15 | 15 | 45 |
| 45 | 135 | 75 |
| 75 | 15 | 45 |
| 45 | 15 | 45 |
| 165 | 45 | 75 |
| 105 | 75 | 75 |
| 135 | 135 | 105 |
| 15 | 45 | 15 |
| 45 | 165 | 75 |
| 15 | 75 | 45 |
| 15 | 75 | 45 |
| 15 | 75 | 75 |
| 75 |  | 75 |
| 105 |  | 45 |
|  |  | 45 |
|  |  | 75 |
|  |  | 75 |
|  |  | 45 |
|  |  | 135 |
|  |  | 15 |
|  |  | 15 |
|  |  | 45 |
|  |  | 75 |
|  |  | 45 |

- Lập bài toán so sánh trung bình 3 tổng thể kể trên với mức ý nghĩa  = 5%.

**B. CƠ SỞ LÝ THUYẾT**

Khái niệm về kiểm định: Việc dùng kết quả của mẫu để khẳng định hay bác bỏ một giả thiết H nào đó được gọi là kiểm định giả thiết H.Khi kiểm định ta có thể mắc 1 trong 2 loại sai lầm sau:

- Sai lầm loại 1: Là sai lầm khi ta bác bỏ H nếu H đúng. Ta ký hiệu xác suất để mắc sai lầm loại này là  và gọi là mức ý nghĩa.

- Sai lầm loại 2: Là sai lầm mắc phải khi ta công nhận H khi H sai. Ta ký hiệu xác suất mắc sai lầm loại này là và gọi 1- là lực kiểm định.

Quy trình kiểm định:

+ Tiêu chuẩn kiểm định: xét mẫu W = . Chọn thống kê: G = f.

Điều kiện đặt ra đối với thống kê G là nếu H đúng thì quy luật phân phối xác suất của G phải hoàn toàn xác định .

+Miền bác bỏ giả thuyết H: Cho trước , miền bác bỏ  là miền thỏa:

. Có vô số miền  như vậy, trong số đó ta sẽ chọn miền bác bỏ  sao cho xác suất mắc sai lầm loại 2 là nhỏ nhất nghĩa là sao cho:

.

Giả thuyết: H: P = P0.

Giả thuyết đối lập  - (thiếu miền bác bỏ bên trái).

- (thừa miền bác bỏ bên phải).

- (đối xứng).

\* Phân tích phương sai: mục tiêu của phân tích phương sai là so sánh trung bình của nhiều nhóm (tổng thể), dựa trên các giá trị trung bình của các mẫu quan sát từ các nhóm này, và thông qua sự kiểm định giả thuyết để kết luận về sự bằng nhau của các trung bình tổng thể này. Trong nghiên cứu, phân tích phương sai được dùng như một công cụ để xem xét ảnh hưởng của 1 yếu tố nguyên nhân (định tính) đến 1 yếu tố kết quả (định lượng).

Chúng ta đề cập đến 2 mô hình phân tích phương sai: phân tích phương sai 1 yếu tố và 2 yếu tố. Cụm từ yếu tố ở đây ám chỉ số lượng yếu tố nguyên nhân ảnh hưởng đến số lượng yếu tố kết quả đang nghiên cứu.

- Phân tích phương sai một yếu tố: là phân tích ảnh hưởng của 1 yếu tố nguyên nhân (dạng biến định tính) đến 1 yếu tố kết quả (dạng biến định lượng) đang nghiên cứu.

- Trường hợp k tổng thể có phân phối bình thường và phương sai bằng nhau:

\*Cần có 3 giả định sau về các nhóm tổng thể để phân tích:

+ Các tổng thể này có phân phối bình thường.

+ Các phương sai tổng thể bằng nhau.

+ Các quan sát được lấy mẫu là độc lập nhau.

Nếu trung bình của các tổng thể lần lượt làthì khi các giả định trên được đáp ứng, mô hình phân tích phương sai một yếu tố ảnh hưởng được mô tả dưới dạng kiểm định lý thuyết như sau:

H0 : 

Giả thuyết H0 cho rằng trung bình k tổng thể đều bằng nhau (về mặt nghiên cứu liên hệ thì giả thuyết này cho rằng yếu tố nguyên nhân không có tác động gì đến vấn đề mà ta nghiên cứu).

Giả thuyết đối H1 : Tồn tại ít nhất 1 cặp trung bình tổng thể khác nhau.

- Để kiểm định giả thuyết này ta thực hiện các bước như sau:

**Bước 1: Tính trung bình mẫu của các nhóm.**



**Bước 2: Tính tổng các chênh lệch bình phương.**

Tổng chênh lệch bình phương trong nội bộ nhóm (SSW).

Nhóm 1: 

Nhóm 2: 

Tương tự với. SSW = .

Hay SSW = .

Tổng chênh lêch bình phương giữa các nhóm (SSG):

SSG = .

Tổng chênh lệch bình phương toàn bộ (SST):

SST = . Hoặc ta có thể dễ dàng chứng minh SST = SSG + SSW.

**Bước 3: Tìm các phương sai:**

Phương sai trong nội bộ nhóm (MSW) :

MSW = .

Phương sai giữa các nhóm (MSG):

MSG = .

**Bước 4: Kiểm định giả thuyết:**

.

Ta bác bỏ giả thuyết H0 cho rằng trị trung bình của k tổng thể bằng nhau khi:

.

 là giá trị tra bảng Fisher có = 0.05 ở cột k - 1 và hàng n - k.

# **C. LỜI GIẢI**

H0: Thời gian ngủ trưa của sinh viên trường đại học bách khoa không phụ thuộc vào năm học. (X1 = X2 =X3 = Xtb).

H1: Tồn tại ít nhất 2 năm học mà thời gian ngủ trưa của sinh viên 2 năm học đó là khác nhau.().

Với n = 62 và k = 3. Tra bảng Fisher cột 2 hàng 59 để thu được giá trị cận dưới của vùng bác bỏ  3.15 => .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | K17 | K18 | K20 | Toàn mẫu |
|  | 18 | 16 | 28 | 62 |
|  | 63,33 | 75 | 60 | 64.84 |
|  | 112050 | 126000 | 121500 | 359550 |
|  | 39849.84 | 36000 | 20700.12 | SSW = 96549.96 |
|  | 41.04 | 1651.61 | 655.92 | SSG = 2348.57 |
|  |  |  |  | SST = SSG + SSW = 98898.53 |

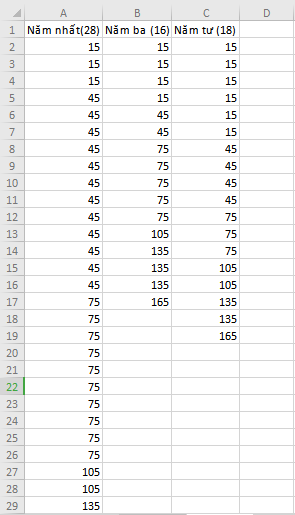
MSW = == 1636.44

MSG = = 1174.29

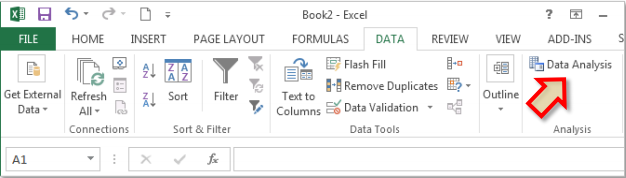
= 0.72 =>  => Chưa thể bác bỏ H0 => Thời gian ngủ trưa của sinh viên trường ĐHBK không phụ thuộc vào năm học.

**D. LỜI GIẢI SỬ DỤNG EXCEL**

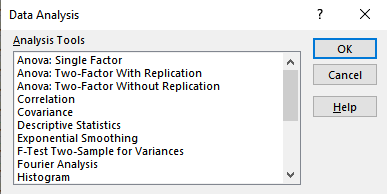
**Bước 1:** Tạo 3 cột là các biến độc lập ứng với 3 khóa K17, K18, K20 và hàng là thời gian ngủ của các sinh viên.

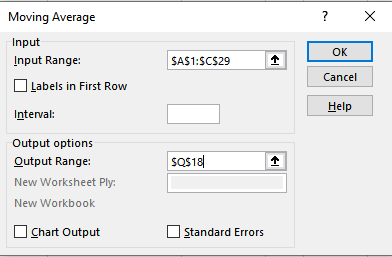
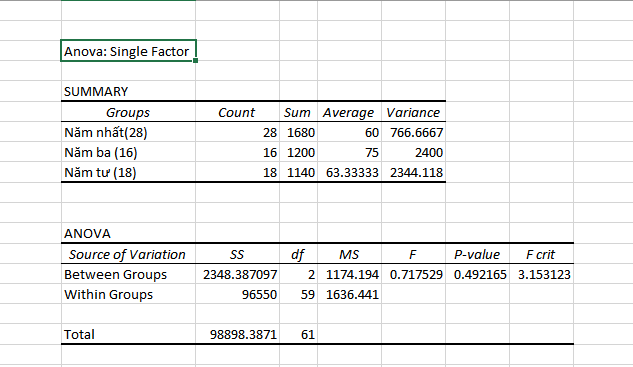


**Bước 2:** Dùng công cụ Data Analysis ở Data. Nếu chưa có phải trở về File-> Option->Excell Option-> Add -Ins để thêm công cụ Data Analysis vào



**Bước 3:** Sử dụng công cụ Data Analysis và chọn Analysis Tools là Anova : Single Factor.



**Bước 4:** Chọn vùng dữ liệu vừa thêm ở bước 1 vào Input Range, chọn vị trí xuất bảng giá trị sau khi tính toán ở Output Range và nhấn OK. Ta sẽ được các giá trị cần tính. Sau đó ta đọc giá trị F và Fcrit để kết luận việc có bác bỏ giá thiết hay không.

1. **CÂU 4:**

**A. LẬP BÀI TOÁN**

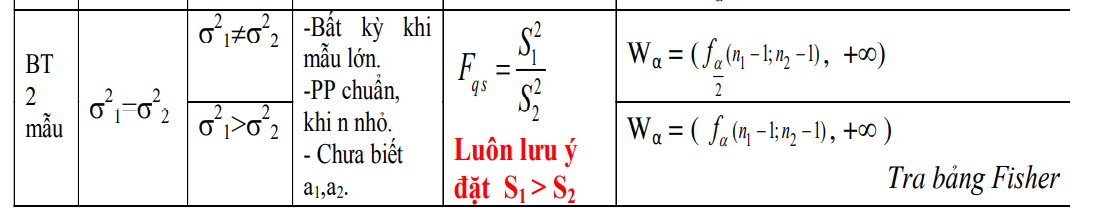
Bảng số liệu dưới đây là thời điểm đi ngủ vào những ngày nghỉ của sinh viên K19 và sinh viên K20 của Trường Đại học Bách khoa Thành phố Hồ Chí Minh. Hãy kiểm định sự phân tán trong thời điểm đi ngủ ngày nghỉ của hai nhóm sinh viên K19 và K20 (Mức ý nghĩa 5%) ?

|  |  |
| --- | --- |
| K19 | K20 |
| 21.5 | 21.5 |
| 21.5 | 22.5 |
| 21.5 | 22.5 |
| 21.5 | 22.5 |
| 21.5 | 22.5 |
| 21.5 | 23.5 |
| 21.5 | 23.5 |
| 21.5 | 23.5 |
| 21.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 23.5 |
| 22.5 | 24.5 |
| 22.5 | 24.5 |
| 22.5 | 24.5 |
| 22.5 | 24.5 |
| 23.5 | 24.5 |
| 23.5 | 24.5 |
| 23.5 | 24.5 |
| 23.5 | 24.5 |
| 23.5 | 24.5 |
| 23.5 | 25.5 |
| 23.5 | 26.5 |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 23.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 24.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 25.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 26.5 |  |
| 27.5 |  |
| 27.5 |  |
| 27.5 |  |

**B. CƠ SỞ LÝ THUYẾT**

Độ phân tán dữ liệu thường được diễn tả bởi phương sai (Variance) hay độ lệch chuẩn.

Với phương sai là trung bình cộng của bình phương khoảng cách các điểm giá trị Xi so với giá trị trung bình. Công thức biểu diễn như sau:



# **C. LỜI GIẢI**

Áp dụng công thức phương sai ( như ở cơ sở lý thuyết) ta được phương sai mẫu hiệu chỉnh của K19 và K20 lần lượt là: ;

Theo đề, ta có: ( Với n1,n2 lần lượt là kích thước mẫu của lần lượt K19 và K20)

Ta có giả thiết kiểm định

Miền bác bỏ: ( ( Tra bảng Fisher với cột n1-1 và hang n2-1)

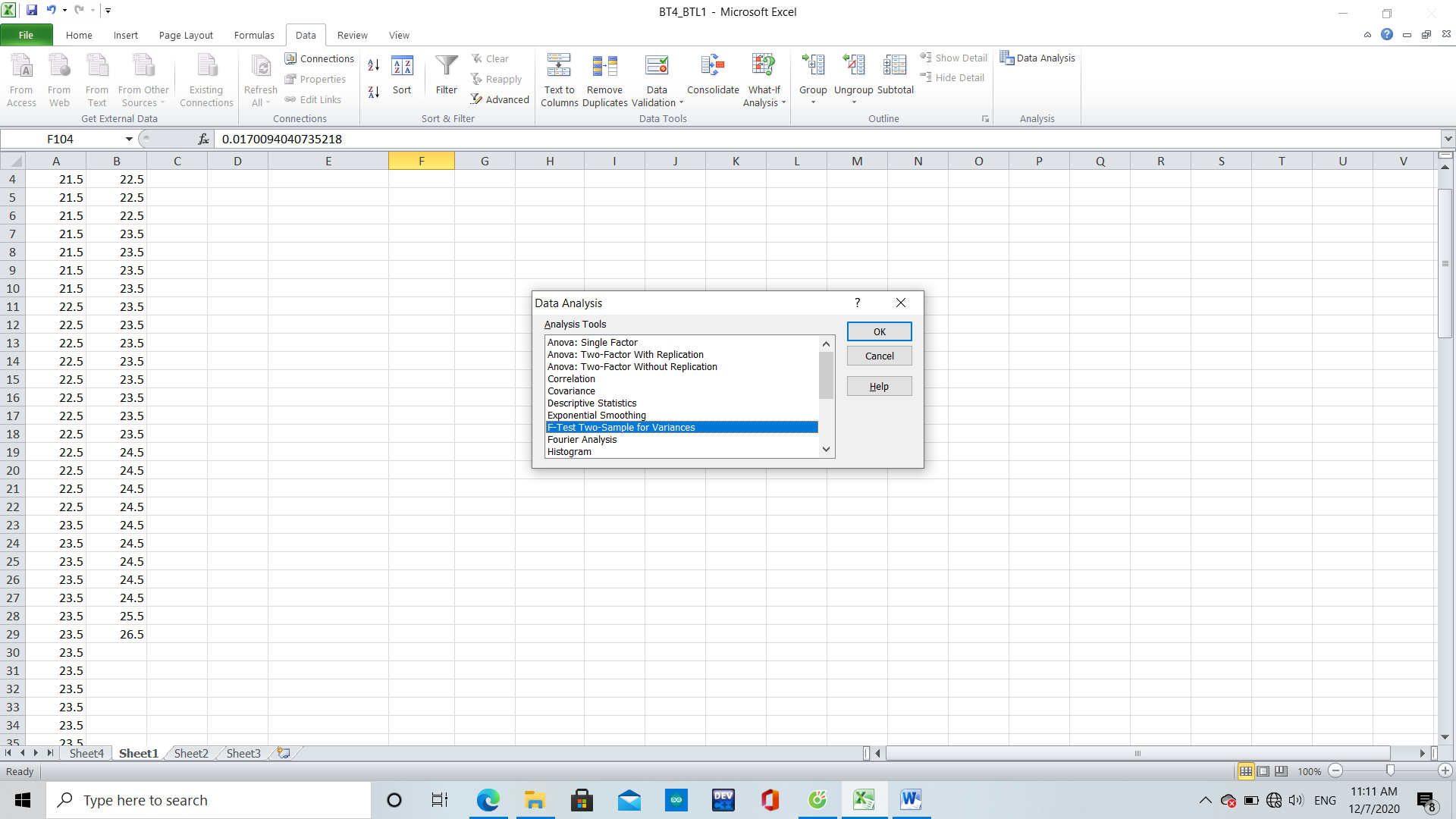
Tiêu chuẩn kiểm định là:

Kết luận: Do đó, bác bỏ H0 và chấp nhận H1. Tức độ phân tán của K19 lớn hơn K20

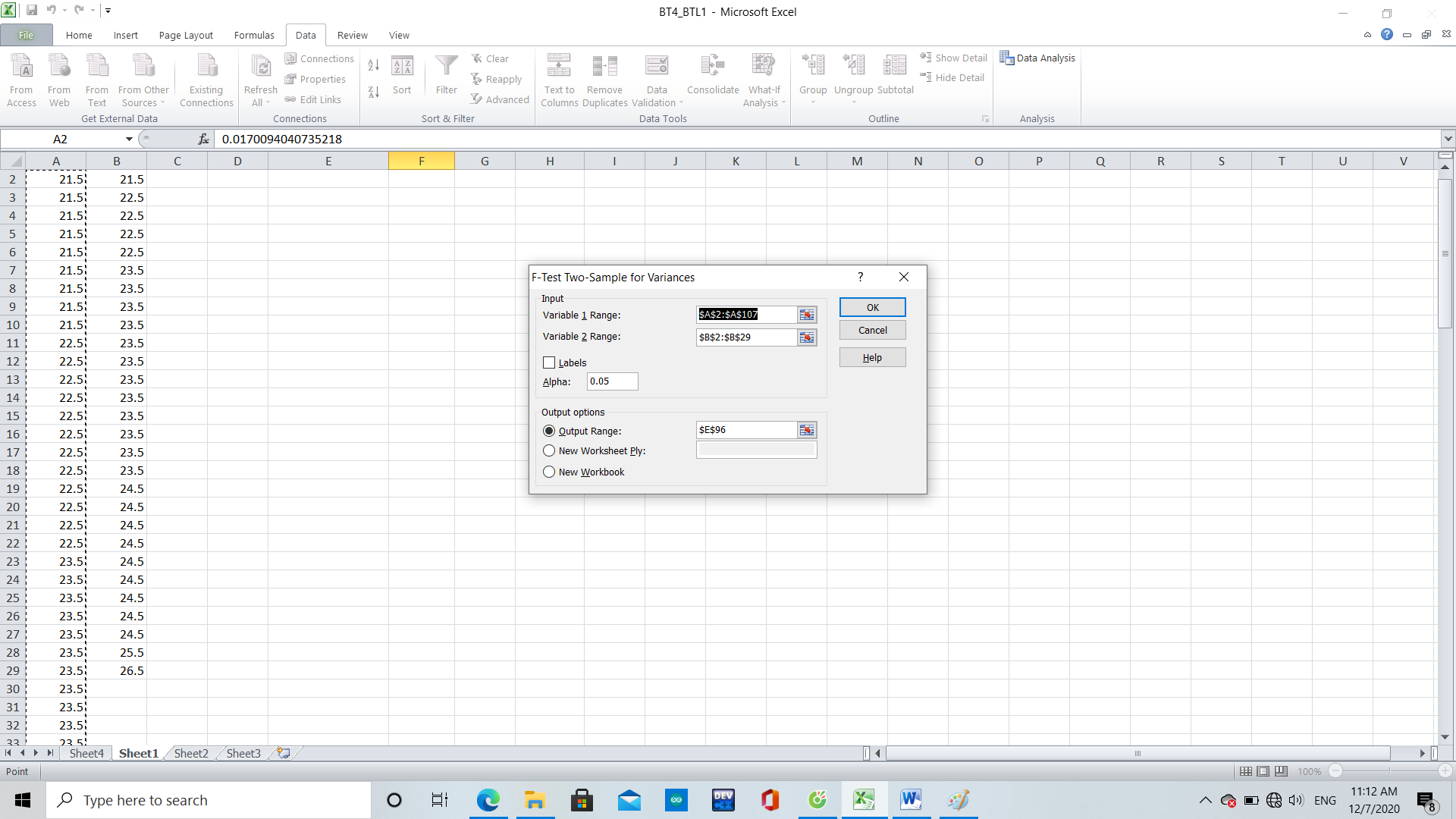
**D. LỜI GIẢI SỬ DỤNG EXCEL**

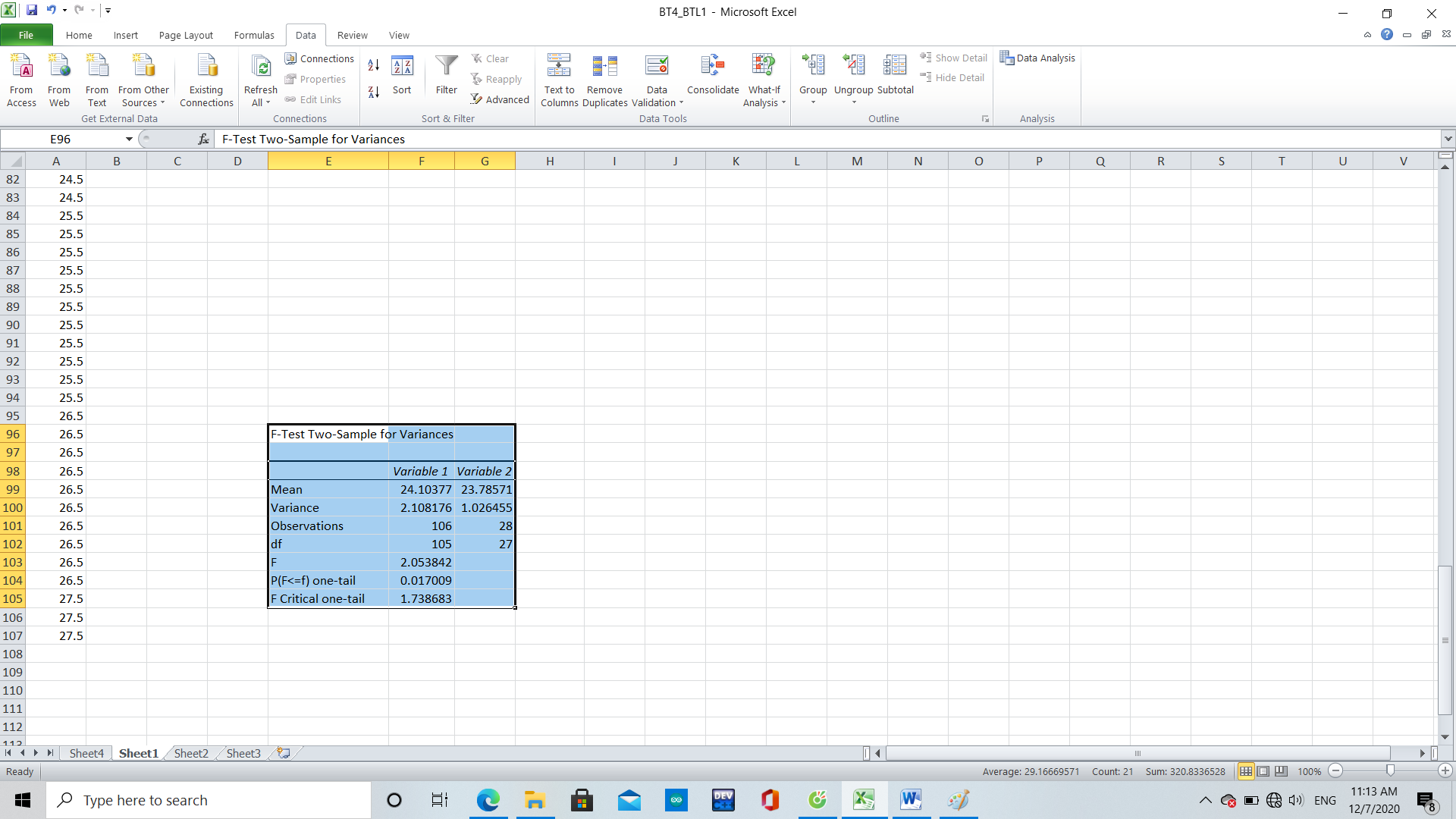
**Bước 1:** Nhập dữ liệu thời điểm ngủ vào các ngày ngủ của sinh viên K19, K20 khảo sát được (trích từ bảng số liệu đã khảo sát)

**Bước 2:** Chọn chức năng Data/Data Analysis/F-Test Two-Sample for Variances. Nhấp OK.



**Bước 3:** Thực hiện chọn dữ liệu cột của K19 vào Variable 1 Range và cột dữ liệu của K20 vào Variable 2 Range. Chọn vị trí kết quả ở output range.Sau đó,nhấp OK.



**-Kết quả:**

1. **CÂU 5:**

# **LẬP BÀI TOÁN**

Bảng số liệu dưới đây cho ta biết thời điểm đi ngủ mỗi đêm của sinh viên Bách Khoa theo khoá. Hãy cho biết thời gian đi ngủ của sinh viên có khác nhau theo từng khoá hay không? (Mức ý nghĩa 𝛼 = 5%)

| Giờ ngủ | <12h | >=12h |
| --- | --- | --- |
| Khoá |
| K20 | 22 | 6 |
| K19 | 15 | 21 |
| >=K18 | 19 | 17 |

Bảng A.1

# **CƠ SỞ LÝ THUYẾT**

## Khái niệm

Bài toán trên là bài toán kiểm định so sánh nhiều tỷ lệ để nhận xét tính độc lập của thời điểm đi ngủ của sinh viên với việc sinh viên đó học khoá mấy. Ta dùng hàm CHITEST để thực hiện việc kiểm định vì trắc nghiệm “khi” bình phương (X2) cho phép so sánh không chỉ hai mà nhiều tỷ số (hay tỷ lệ hoặc xác suất). X2 là phân phối về xác suất, không có tính đối xứng và chỉ có giá trị ≥ 0.

Ví dụ: với N thử nghiệm độc lập, mỗi thử nghiệm có k kết quả và mỗi kết quả mang một xác suất thực nghiệm là Pi (i = 1, 2,...,k). Nếu gọi Pi,0 là các giá trị lý thuyết tưng ứng với Pi thì các tần số lý thuyết sẽ là Ei = NPi,0. Nói một cách dễ hiểu, Pi là xác suất của mỗi trường hợp (mỗi ô) trong bảng BNN định tính hai chiều (X,Y) với X nhận các giá trị A1, A2,…,Ar và Y nhận B1, B2,…,Bc.

Điều kiện để áp dụng trắc nghiệm X2 một cách thành công là các tần số lý thuyết Ei phải ≥ 5.

## Hướng dẫn giải

* Giả thiết kiểm định H0: “Các cặp Pi và Pi,0 giống nhau”

Giả thuyết đối H1: “Ít nhất một cặp Pi và Pi,0 khác nhau”

* Miền bác bỏ : tra bảng VI, phụ lục 2, Giáo trình Xác suất và thống kê với bậc tự do và mức ý nghĩa .

Trong đó r: số hàng

c: số cột

* Giả thiết các cặp Pi và Pi,0 là giống nhau nên xác suất tính theo lý thuyết là:

Với ni: tổng hàng

mi: tổng cột

⇒ Tần số lý thuyết là

* Tiêu chuẩn kiểm định:

Oi,j: Tần số thực nghiệm của ô thuộc hàng i và cột j

Ei, j: tần số lý thuyết của ô thuộc hàng i và cột j.

* Kết luận: nếu thì bác bỏ 𝐻0 và ngược lại.

# **LỜI GIẢI**

Giả thiết H0: “Thời gian đi ngủ vào ban đêm của sinh viên Bách Khoa mỗi khoá là như nhau”.

Giả thiết đối H1: “Thời gian đi ngủ ban đêm của sinh viên Bách Khoa phụ thuộc vào việc sinh viên đó đang học năm mấy”.

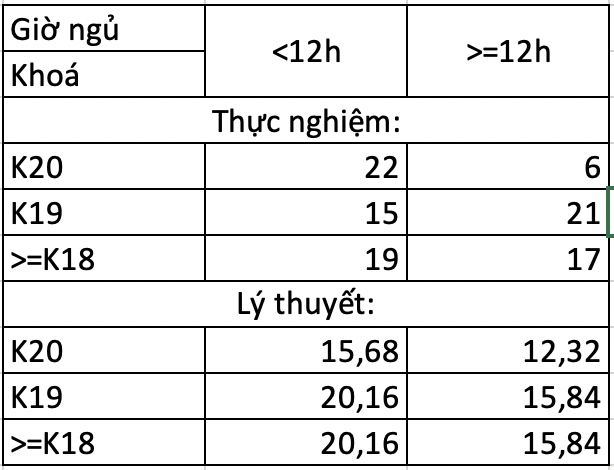
Bậc tự do

Miền bác bỏ

Ta có bảng tần số thực nghiệm như trên Bảng A.1

Tính tần số lý thuyết với với lần lượt là tổng số phần tử theo cột và theo hàng.

Ta có bảng tần số sau:



Bảng C.1

Bác bỏ giả thiết H0.

**KẾT LUẬN:** Vậy thời gian đi ngủ của sinh viên Bách Khoa phụ thuộc vào việc sinh viên đó là sinh viên năm thứ mấy.

# **LỜI GIẢI SỬ DỤNG EXCEL**

**Bước 1:** Đặt giả thiết:

Gọi H0: “Thời gian đi ngủ vào ban đêm của sinh viên Bách Khoa mỗi khoá là như nhau”.

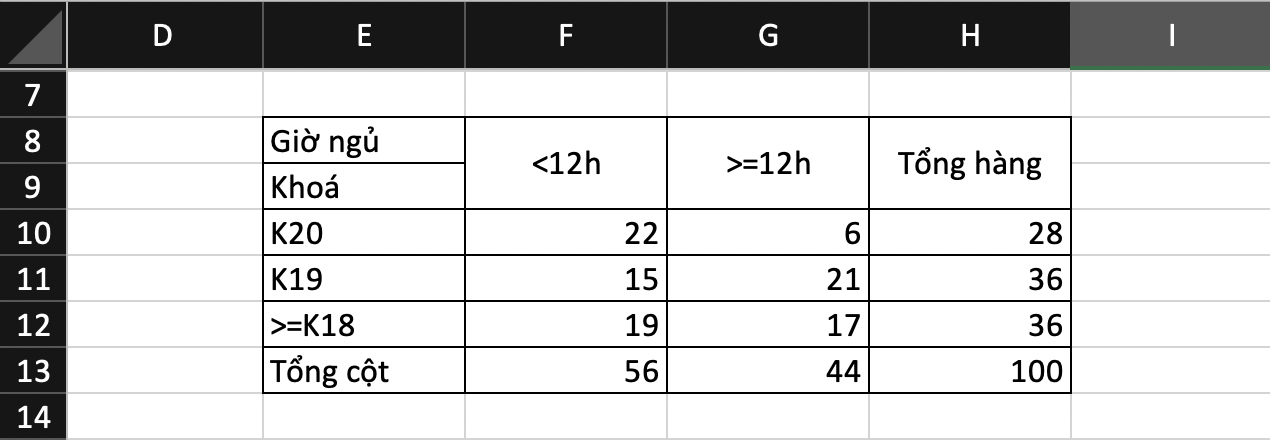
Giả thiết đối H1: “Thời gian đi ngủ ban đêm của sinh viên Bách Khoa phụ thuộc vào việc sinh viên đó đang học năm mấy”

**Bước 2:** Nhập bảng số liệu; tính tổng hàng và tổng cột

Tổng hàng (Row totals): Tại H10 nhập =SUM(F10:G10) và nhấn Enter. Chọn ô H10 và kéo xuống đến ô H12 để tính tổng 2 hàng dưới

Tổng cột (Column totals): Tại F13 nhập =SUM(F10:F12) và nhấn Enter. Chọn ô F13 và kéo sang phải đến ô H13 để tính tổng hai cột còn lại.

Tổng cộng (Grand totals): Tại H13 nhập =SUM(H10:H12) và nhấn Enter.



Hình D.1

**Bước 3:** Tính các tần số lý thuyết:

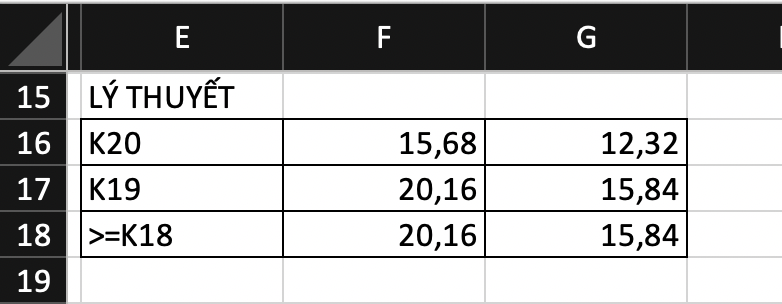
Tần số lý thuyết = (tổng hàng × tổng cột) / tổng cộng

K20 ngủ trước 12h đêm: tại F16 nhập =F13\*H10/H13

K20 ngủ sau 12h đêm: tại G16 nhập =G13\*H10/H13

K19 ngủ trước 12h đêm: tại F17 nhập =F13\*H11/H13

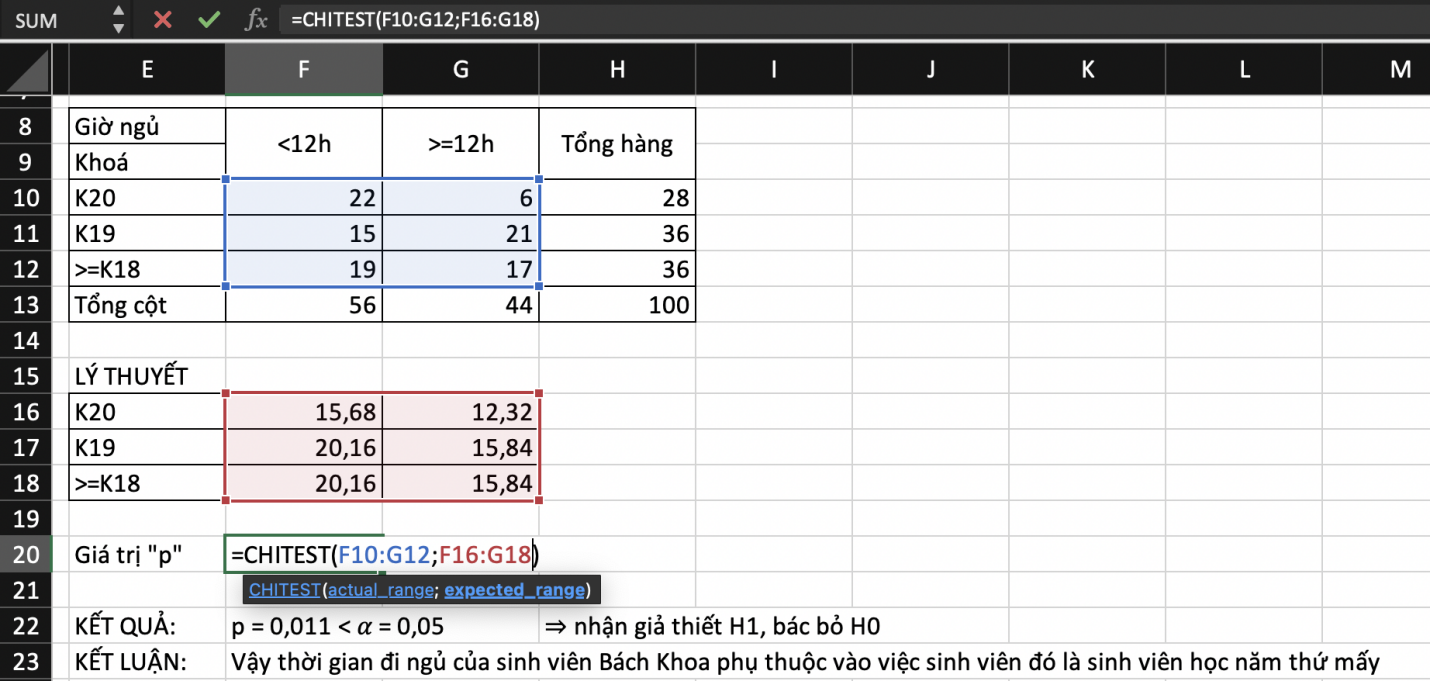
Tương tự đối với K19 ngủ sau 12h đêm; >=K18 ngủ trước, sau 12h đêm.



Hình D.2

**Bước 4:** Dùng hàm CHITEST để tính xác suất P(X > 𝜒2).

Tại F20 nhập =CHITEST(F10:G12;F16:G18) là phạm vi của 2 bảng tần số thực nghiệm và lý thuyết. Hàm CHITEST cho ta kết quả là xác suất P(X > 𝜒2).

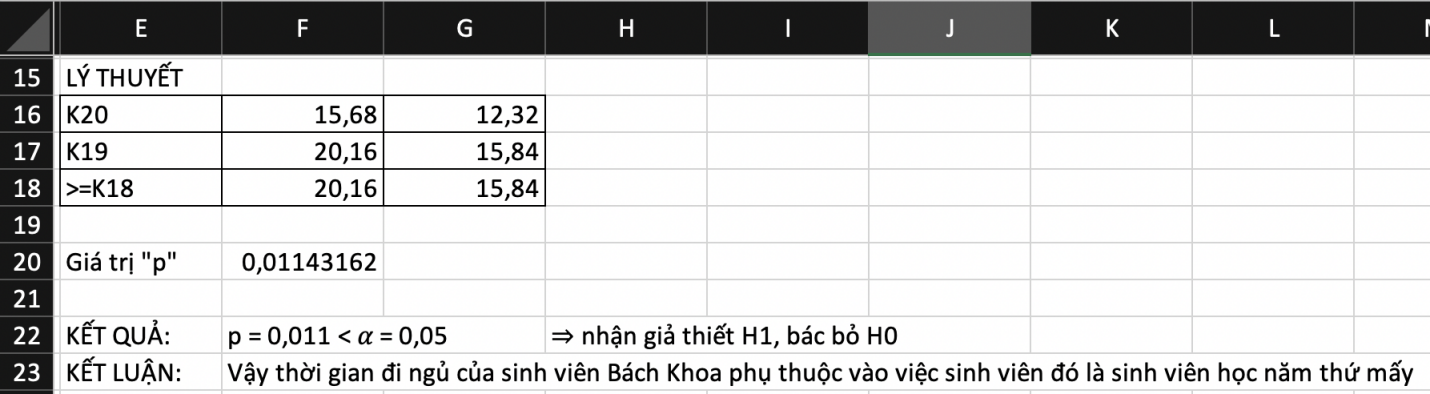


Hình D.3

**Bước 5:** Kết luận

Vì P(X > 𝜒2) < 𝛼 (0,011 < 0,05) nên ta nhận giả thiết H1, bác bỏ H0.

Vậy thời gian đi ngủ của sinh viên Bách Khoa phụ thuộc vào việc sinh viên đó là sinh viên năm thứ mấy.



Hình D.4

1. **CÂU 6:**

# **LẬP BÀI TOÁN**

Khảo sát cho thấy thời gian đi ngủ (X) và thời lượng giấc ngủ (Y) của các bạn sinh viên bách khoa cho bảng số liệu như sau.

1. Hãy khảo sát hệ số tương quan giữa X và Y
2. Dự đoán phương trình hồi quy tuyến tính giữa chúng
3. Nhận xét về mối tương quan tuyến tính giữa 2 biến.

|  |  |
| --- | --- |
| Thời gian đi ngủ (X) | Thời lượng ngủ (Y) |
| 26.5 | 5.5 |
| 25.5 | 4.5 |
| 26.5 | 6.5 |
| 25.5 | 5.5 |
| 23.5 | 5.5 |
| 22.5 | 6.5 |
| 22.5 | 7.5 |
| 21.5 | 7.5 |
| 24.5 | 4.5 |
| 22.5 | 6.5 |
| 24.5 | 6.5 |
| 23.5 | 4.5 |
| 24.5 | 4.5 |
| 23.5 | 6.5 |
| 24.5 | 6.5 |
| 21.5 | 6.5 |
| 24.5 | 5.5 |
| 24.5 | 4.5 |
| 23.5 | 6.5 |
| 24.5 | 6.5 |

1. **Hệ số tương quan giữa X và Y:**

# **CƠ SỞ LÝ THUYẾT**

Khi cần kiểm tra 2 biến định lượng có mối quan hệ với nhau hay không, chúng ta xét tới hệ số tương quan giữa chứng

Hệ số tương quan là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số.

Với : thì X, Y tương quan thuận.

Với : thì X, Y tương quan nghịch.

Với : thì X, Y không tương quan với nhau.

Với : thì X, Y có quan hệ tuyến tính với nhau.

Với : thì X, Y có tương quan mạnh

Với : thì X, Y có tương quan yếu

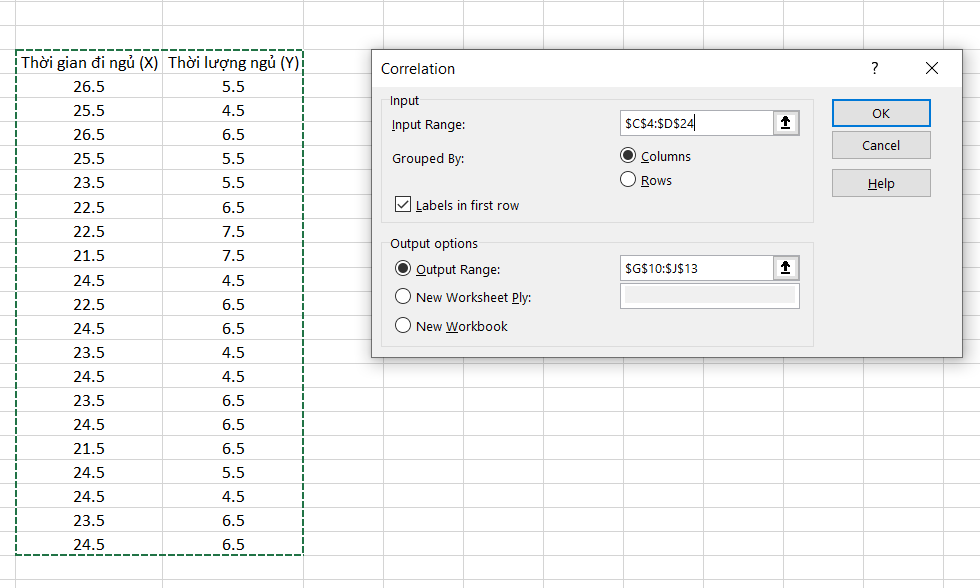
# **LỜI GIẢI**

Ta có:

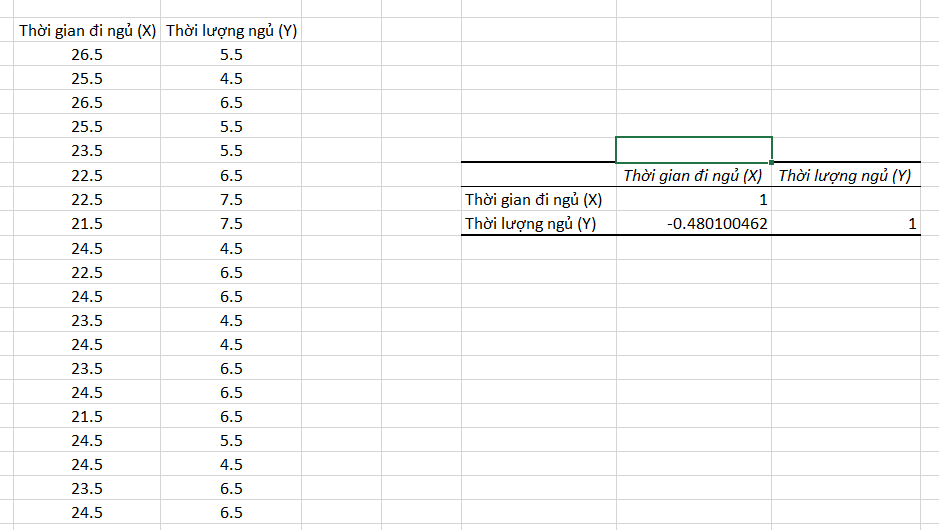
Nhận xét : chứng tỏ giữa thời gian đi ngủ và thời lượng giấc ngủ có tương quan ở mức trung bình, và là tương quan nghịch.

**Thực hiện trên excel:**

* Nhập dữ liệu cần tính (copy từ bảng số liệu đã khảo sát)
* Chọn chức năng Data/Data Analysis/Correlation.
* Chọn đoạn dữ liệu cần tính ở input range và chọn vị trí kết quả ở output range, nhấp OK



* Kết quả



**b. Dự đoán phương trình hồi quy tuyến tính giữa X và Y**

# **B. CƠ SỞ LÝ THUYẾT**

Lý thuyết hồi quy (đơn biến) nghiên cứu bài toán dự báo biến ngẫu nhiên Y trên cơ sở đã biết về biến ngẫu nhiên X.

ĐN hồi quy: Hàm hồi quy của Y theo X chính là kỳ vọng có điều kiện của Y đối với X, tức là E(Y|X). Nếu đồ thị hàm hồi quy là đường thẳng thì ta nói hồi quy tuyến tính.

Hệ số tương quan tuyến tính

Đường hồi quy tuyến tính mẫu: Đường hồi quy tuyến tính mẫu của Y theo X là đường thẳng có dạng ,“ xấp xỉ gần nhất” với đường hồi quy mẫu.

Người ta dùng phương pháp bình phương bé nhất để tìm các hệ số A, B này, tức là (A,B) làm cho hàm Q đạt giá trị nhỏ nhất.

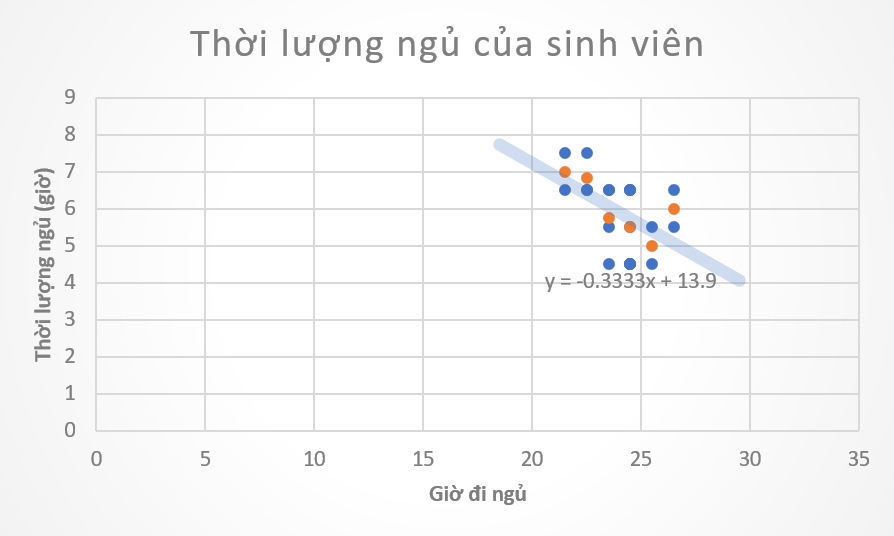
Với Giải bài toán cực trị tự do ta có : là cặp giá trị để hàm Q(A,B) nhỏ nhất

Khi đó giá trị được tìm qua đường hồi quy tuyến tính mẫu thường được viết dưới dạng , nó là dự đoán giá trị trung bình ()của biến Y khi .

**C. LỜI GIẢI:**

Ta có:

Vậy đường hồi quy tuyến tính mẫu của Y theo X là

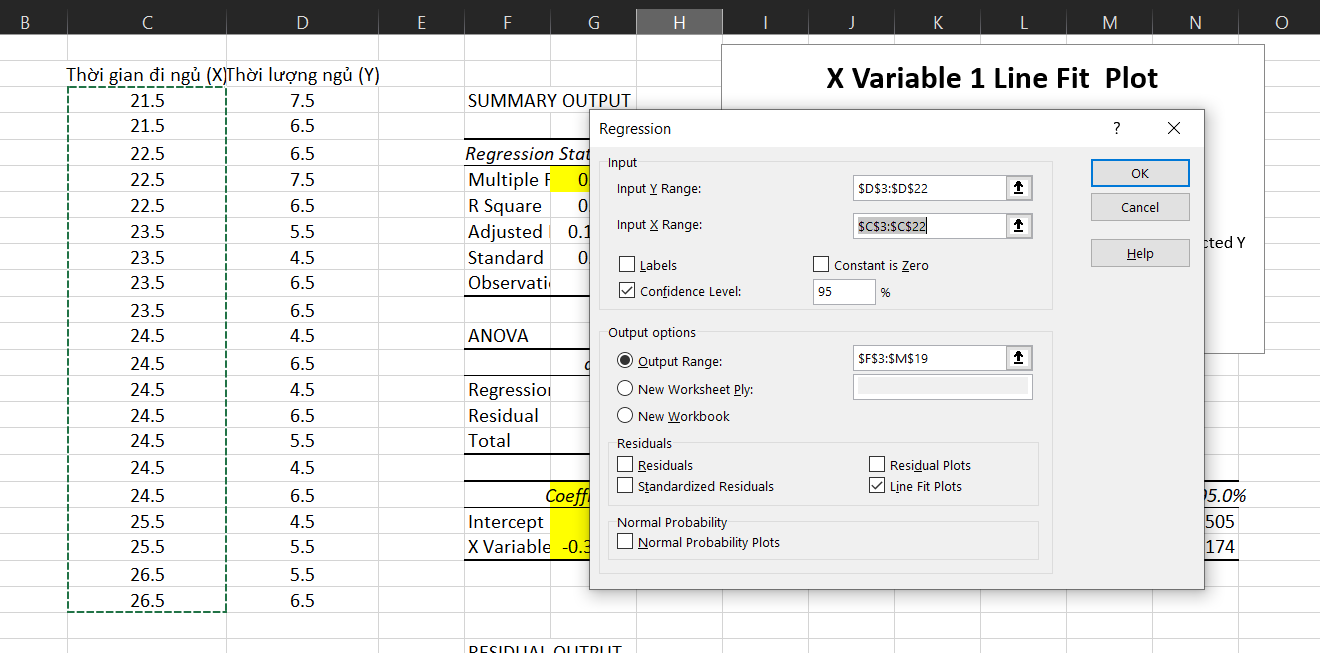


# **LỜI GIẢI SỬ DỤNG EXCEL**

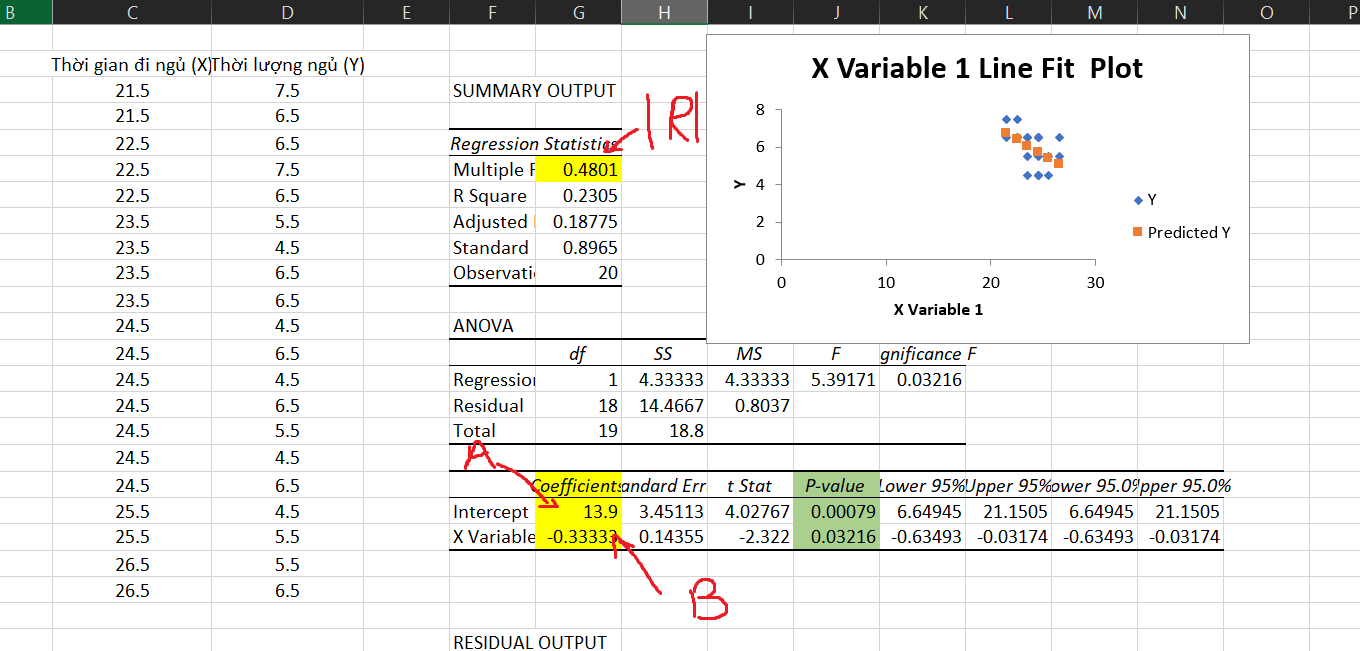
**Bước 1:** Nhập dữ liệu cần tính (copy từ bảng số liệu đã khảo sát)

**Bước 2:** Chọn chức năng Data/Data Analysis/ Regression.

**Bước 3:** Chọn đoạn dữ liệu cần tính: Y ở Input Y Range và X ở Input X Range, chọn độ tin cậy ở Confidence Level (95%), nhấp OK



* **Kết quả**



1. **Nhận xét về mối tương quan tuyến tính của 2 biến**
2. **CƠ SỞ LÝ THUYẾT**

* Kiểm định phương trình hồi quy:

Giả thiết : Phương trình hồi quy tuyến tính không thích hợp ()

Giả thiết : Phương trình hồi quy tuyến tính thích hợp ()

Tiêu chuẩn kiếm định:

Tra bảng Student(myn/2,n-2)

* Kiểm định hệ số A,B

Giả thiết : Hệ số hồi quy không có ý nghĩa ()

Giả thiết : Hệ số hồi quy không có ý nghĩa

Tiêu chuẩn kiểm định:

Tra bảng Student(myn/2, n-2)

Miền bác bỏ .

1. **LỜI GIẢI**

* Kiểm định phương trình hồi quy:

+ Giả thiết : Thời gian đi ngủ và thời lượng giấc ngủ **không** có tương quan tuyến tính ( )

+ Giả thiết : Thời gian đi ngủ và thời lượng giấc ngủ **có** tương quan tuyến tính ()

Ta có

Tiêu chuẩn kiểm định:

Tra bảng Student (0.05/2, 20-2) có t= 2.101 ( hoặc tra bảng Fisher(1,18) = 4.41 nếu so sánh với )

Miền bác bỏ

* Vì nên bác bỏ , chấp nhận

Kết luận: Thời gian đi ngủ và thời lượng giấc ngủ **có** tương quan tuyến tính

* Kiểm định hệ số A, B có nghĩa thống kê (a là hệ số tự do, b là hệ số góc)

+ Giả thiết H0: A, B không có ý nghĩa thống kê

+ Giả thiết H1: A, B có ý nghĩa thống kê

Tiêu chuẩn kiếm định:

Tra bảng Student(0.025, 18) có t= 2.101

Miền bác bỏ

* Vì nên bác bỏ , chấp nhận do đó B có ý nghĩa thống kê
* Vì nên bác bỏ , chấp nhận do đó A có ý nghĩa thống kê

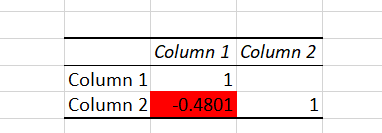
Kết luận: cả A,B đều có ý nghĩa thống kê

Nhận xét: Thời gian đi ngủ và thời lượng giấc ngủ **có** tương quan tuyến tính theo phương trình

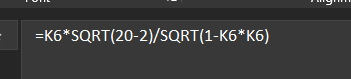
# **LỜI GIẢI SỬ DỤNG EXCEL**

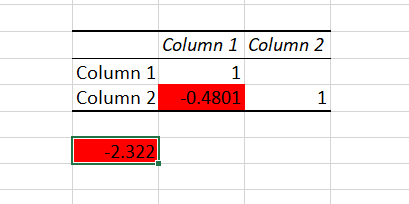
**Bước 1:** Nhập bảng số liệu

**Bước 2:** Lấy bảng Correlation từ câu a



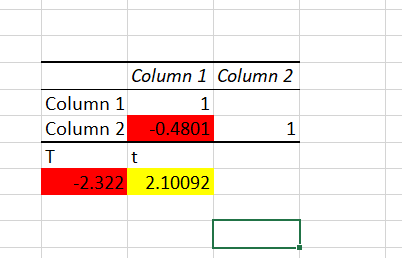
**Bước 3:** Tính giá trị T theo các hàm đã cho





**Bước 4:** Tính giá trị t (bảng student), bằng lệnh =TINV (myn, n-2)



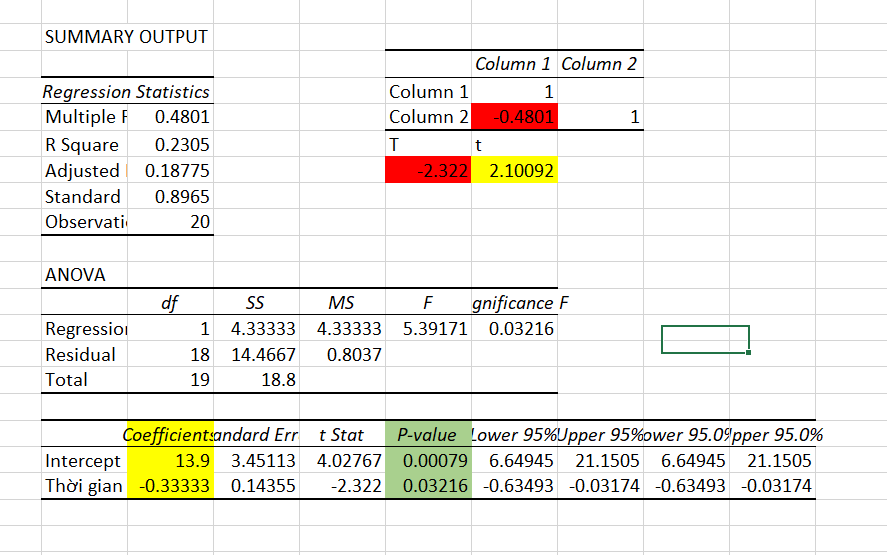


**Bước 5:** So sánh giá trị tuyệt đối của 2 kết quả

* Vì nên bác bỏ , chấp nhận

**Bước 6:** Sử dụng bảng Regression ở câu b

Đọc giá trị P value



**Bước 8:** So sánh với mức ý nghĩa

* P-value(A)= 0.00079 <0.05 nên bác bỏ , chấp nhận do đó A có ý nghĩa thống kê
* P-value(B)= 0.03216 <0.05 nên bác bỏ , chấp nhận do đó B có ý nghĩa thống kê