

PHÂN TÍCH THỐNG KÊ SỬ DỤNG EXCEL®

Nguyễn Ngọc Anh
Nguyễn Đình Chúc
Đoàn Quang Hưng

PHÂN TÍCH THỐNG KÊ SỬ DỤNG EXCEL®

Tác giả

Nguyễn Ngọc Anh
Nguyễn Đình Chúc
Đoàn Quang Hưng

Copyright notice

This material is copyrighted by DEPOCEN® . Authorized users may be allowed to use this material for their personal educational and research purposes. Other use, storage, reproduction, and distribution is strictly prohibited.

MỤC LỤC

| | | |
|-----|--|----|
| 1 | GIỚI THIỆU | 4 |
| 2 | NHẬP DỮ LIỆU | 5 |
| 3 | BỘ CÔNG CỤ DATA ANALYSIS TOOLPACT | 5 |
| 4 | THỐNG KÊ MÔ TẢ | 6 |
| 5 | PHÂN PHỐI CHUẨN* | 8 |
| 6 | XÂY DỰNG KHOẢNG TIN CẬY CHO TRUNG BÌNH TỔNG THỂ | 12 |
| 6.1 | Khi qui mô của mẫu thống kê lớn (n lớn hơn 30) | 12 |
| 6.2 | Mẫu nhỏ (ít hơn 30 quan sát) | 15 |
| 7 | KIỂM ĐỊNH GIẢ THUYẾT VỀ TRUNG BÌNH TỔNG THỂ | 16 |
| 8 | KIỂM ĐỊNH SỰ KHÁC BIỆT GIỮA HAI TRUNG BÌNH TỔNG THỂ..... | 18 |
| 8.1 | Mẫu lớn: | 18 |
| 8.2 | Mẫu nhỏ: Một trong hai mẫu có số lượng các quan sát nhỏ hơn 30. | 22 |
| 9 | TƯƠNG QUAN TUYẾN TÍNH VÀ PHÂN TÍCH HỒI QUI* | 26 |
| 9.1 | Phân tích tương quan tuyến tính..... | 27 |
| 9.2 | Phân tích hồi qui..... | 29 |

Lưu ý: Những mục đánh dấu * sẽ được học viên đọc thêm

1 GIỚI THIỆU

EXCEL là một chương trình bảng tính do Microsoft® phát triển. Đây là một chương trình bảng tính được sử dụng rộng rãi nhất. Trong EXCEL có bộ công cụ cho phép người sử dụng tiến hành phân tích dữ liệu thống kê. **EXCEL** có thể được sử dụng để tổ chức sắp xếp dữ liệu, trình bày dữ liệu, lập bảng, vẽ đồ thị và phân tích thống kê (thống kê mô tả, kiểm định giả thuyết và phân tích hồi qui).¹

The screenshot shows the Microsoft Excel interface with a data table. The table has 8 columns: 'Số nhận dạng (ID)', 'Tên', 'Tuổi', 'Giới tính', 'Văn đi học', 'Số năm đi học', 'Sức khoẻ', and 'Hỗ trợ của Tỉnh'. The data rows are numbered 1 to 22. Two callout boxes are present: one labeled 'Tên biến' pointing to the 'Tên' column, and another labeled 'Số liệu' pointing to the 'Số năm đi học' column.

| | Số nhận dạng (ID) | Tên | Tuổi | Giới tính | Văn đi học | Số năm đi học | Sức khoẻ | Hỗ trợ của Tỉnh |
|----|-------------------|----------------------|------|-----------|------------|---------------|----------|-----------------|
| 1 | 1 | dinh nghĩa l?i | 83 | 1 | 2 | 2 | 2 | 2 |
| 2 | 2 | dinh th? yên | 33 | 2 | 2 | 5 | 2 | 1 |
| 3 | 3 | dinh hoài phong | 19 | 1 | 2 | 4 | 2 | 2 |
| 4 | 4 | dinh th? thu nhân | 17 | 2 | 1 | 4 | 2 | 2 |
| 5 | 5 | nguy?n van ti?n | 39 | 1 | 2 | 3 | 2 | 1 |
| 6 | 6 | nguy?n th? lu?n | 36 | 2 | 2 | 3 | 2 | 1 |
| 7 | 7 | nguy?n th? bích trâm | 14 | 2 | 1 | 3 | 2 | 2 |
| 8 | 8 | nguy?n th? sang | 10 | 2 | 1 | 2 | 2 | 2 |
| 9 | 9 | nguy?n th? thanh nhấ | 3 | 2 | | | 2 | |
| 10 | 10 | nguy?n van phi | 39 | 1 | 2 | 2 | 2 | 1 |
| 11 | 11 | hoàng th? m? liên | 32 | 2 | 2 | 2 | 2 | 1 |
| 12 | 12 | nguy?n th? m? h?o | 4 | 2 | | | 2 | |
| 13 | 13 | nguy?n phú thiên | 2 | 1 | | | 2 | |
| 14 | 14 | vuong van th?o | 36 | 1 | 2 | 2 | 2 | 1 |
| 15 | 15 | thái th? mo | 36 | 2 | 2 | 3 | 2 | 1 |
| 16 | 16 | vuong th? m? phu?ng | 15 | 2 | 2 | 3 | 2 | 2 |
| 17 | 17 | vuong van tin | 13 | 1 | 1 | 3 | 2 | 2 |
| 18 | 18 | vuong th? th?t | 2 | 2 | | | 2 | |
| 19 | 19 | nguy?n vaen tu?n | 38 | 1 | 2 | 9 | 2 | 1 |
| 20 | 20 | luong th? siêm | 42 | 2 | 2 | 1 | 2 | 1 |
| 21 | 21 | nguy?n th? kim thoa | 18 | 2 | 2 | 3 | 2 | 2 |

Hình 1: Ví dụ về số liệu trong EXCEL

¹ Để thực hiện các phân tích thống kê phức tạp hơn, chúng ta phải sử dụng các phần mềm thống kê chuyên dụng khác như SPSS, SAS, Splus, R, STATA, GAUSS. Trong số các phần mềm nêu trên, phần mềm R là phần mềm miễn phí nhưng lại có ưu điểm vượt trội hơn khá nhiều phần mềm thương mại khác.

Một số lưu ý: Dòng trên cùng cho người sử dụng biết tên các biến số. Mỗi dòng trong bảng số liệu gọi là một quan sát. Đơn vị quan sát có thể ở cấp cá nhân (số liệu về các cá nhân), hộ gia đình (số liệu về gia đình), công ty, quận, tỉnh, quốc gia. Số liệu không nhất thiết phải ở dạng con số (numerics), mà có thể ở dạng chữ (string). Trong Hình 1, cột thứ 2, thể hiện biến số **Tên** cho ta thấy số liệu là tên người ở dạng chữ.

2 NHẬP DỮ LIỆU

Để có số liệu như trong Hình 1, thông thường người sử dụng/nhà nghiên cứu phải tiến hành nhập số liệu vào trong EXCEL. Việc nhập dữ liệu trong Excel rất đơn giản. Một bảng EXCEL (worksheet) được chia thành các dòng và các cột. Dòng được đánh dấu bằng số và cột được đánh số bằng chữ. Dòng và cột tạo ra các ô trong worksheet. Mỗi ô đều có địa chỉ theo số của dòng và chữ của cột. Để có thể nhập dữ liệu vào một ô, chúng ta cần phải ô cần nhập dữ liệu là ô đang hoạt động. Để làm điều này, chúng ta nhấn chuột vào ô đó.

Mỗi ô có thể chứa các dãy ký tự, các giá trị bằng số, giá trị logic hoặc chứa công thức. Dãy ký tự có thể bao gồm chữ, số hoặc ký hiệu. Giá trị bằng số là những con số tự nhiên mà chúng ta biết và chỉ có con số mới có thể được dùng trong tính toán. Giá trị logic là giá trị cho ta biết một điều gì đó “đúng” hoặc “sai”. Công thức cho phép chúng ta thực hiện việc tính toán một cách tự động đối với giá trị của các ô khác.

3 BỘ CÔNG CỤ DATA ANALYSIS TOOLPACK

Microsoft Excel có một bộ công cụ có thể dùng để phân tích dữ liệu được gọi là **Analysis Toolpack** mà chúng ta có thể sử dụng để phân tích dữ liệu. Nếu như lệnh **Data Analysis** đã hiển thị trên thanh công cụ **Tool menu**, thì bộ công cụ **Analysis Toolpack** đã được cài trên hệ thống. Nếu không chúng ta có thể tiến hành cài bộ công cụ này như sau. Trước hết bạn chọn thanh công cụ **Tool**, sau đó chọn **Add-ins**, sau đó nhấn nút **OK**.

Nếu như, mục Analysis Toolpack không được liệt kê trong cửa sổ **Add-ins** thì bạn bấm nút Browse để tìm tệp **Analys32.xll** thường ở tại program **files\microsoft office\office\library\analysis**. Sau khi đã tìm và chọn được tệp **analyse32.xll**, bạn nhấn nút OK. Sau khi làm các thao tác này, bộ công cụ **Analysis Toolpack** sẽ được cài đặt và bạn có thể sử dụng.

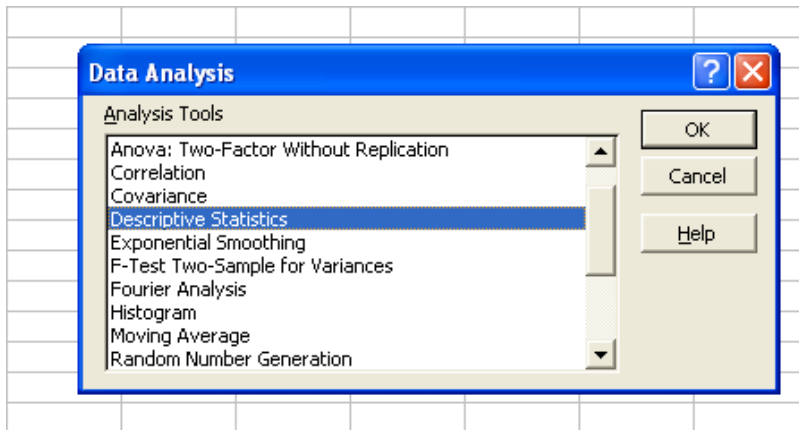
Microsoft Excel là một phần mềm bảng tính rất mạnh được sử dụng để duy trì thông tin và dữ liệu theo cột và hàng. Phần mềm Excel thực hiện các công việc theo **workbooks**, và mỗi **workbook** lại có các **worksheet**, và **worksheet** là nơi mà chúng ta sẽ liệt kê và phân tích dữ liệu với Excel. Khi chúng ta bắt đầu kích hoạt phần mềm Excel, một **worksheet** trắng sẽ được hiển thị, bao gồm nhiều ô trên bảng tính. Mỗi ô trên bảng tính được dẫn chiếu thông qua tọa độ của chúng.

4 THỐNG KÊ MÔ TẢ

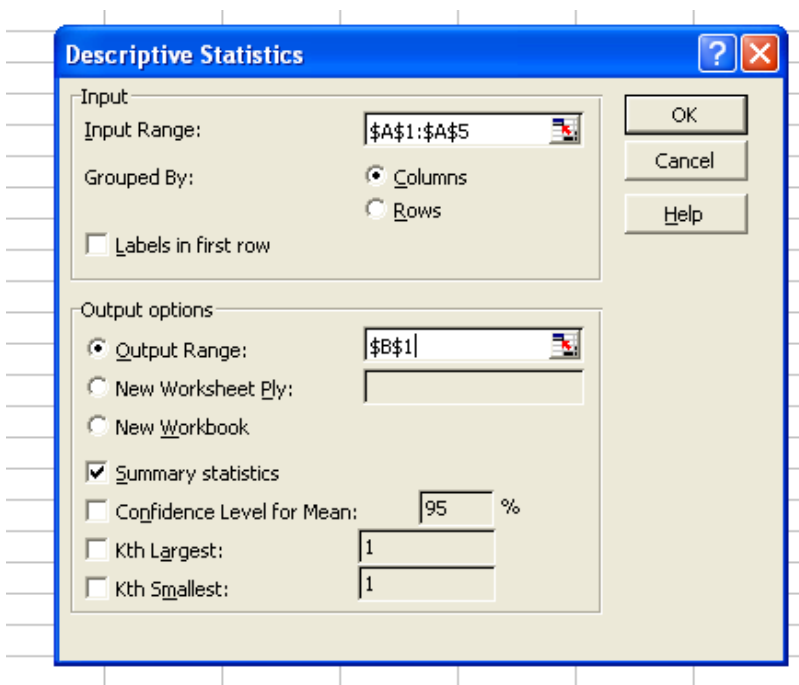
Bộ công cụ Data Analysis Toolpack có một bộ công cụ con để chúng ta có thể tiến hành thực hiện các phương pháp thống kê mô tả. Để tiến hành tìm các đại lượng trong thống kê mô tả, ta thực hiện các bước như sau

Bước 1. Từ menu chúng ta chọn **Tool**, nếu như chúng ta thấy lệnh **data analysis** có hiển thị, chúng ta chọn lệnh này, nếu không chúng ta chọn **add-ins** để cài đặt **Analysis Toolpack** như đã nêu ở trên.

Bước 2. Sau khi đã chọn data analysis, chúng ta chọn **descriptive statistics**.



Bước 3. Khi xuất hiện cửa sổ Descriptive statistics, chúng ta sẽ nhập khoảng dữ liệu, sau đó chúng ta sẽ chọn ô để Excel xuất kết quả.



Sau đó bấm OK và xem xét kết quả thu được

| A | B | C |
|---|--------------------|-------------|
| 1 | Column1 | |
| 2 | | |
| 3 | Mean | 3 |
| 4 | Standard Error | 0.707106781 |
| 5 | Median | 3 |
| | Mode | #N/A |
| | Standard Deviation | 1.58113883 |
| | Sample Variance | 2.5 |
| | Kurtosis | -1.2 |
| | Skewness | 0 |
| | Range | 4 |
| | Minimum | 1 |
| | Maximum | 5 |
| | Sum | 15 |
| | Count | 5 |

Ta thấy Excel cho ta các đại lượng thống kê mô tả cơ bản như trung bình (mean), độ lệch chuẩn (standard deviation), phương sai (variance), dải biến thiên (range), số quan sát (count), giá trị tối đa và giá trị tối thiểu, trung vị (median), sai số chuẩn của trung bình mẫu (standard error).

5 PHÂN PHỐI CHUẨN

Giả sử chúng ta muốn tìm xác suất của một biến X nhận giá trị nhỏ hơn một giá trị nhất định nào đó. Chúng ta giả sử là điểm số của các cá nhân trong lớp là phân bố theo phân phối chuẩn có trị trung bình là 500 và độ lệch chuẩn là 100. Các câu hỏi mà chúng ta phải trả lời là

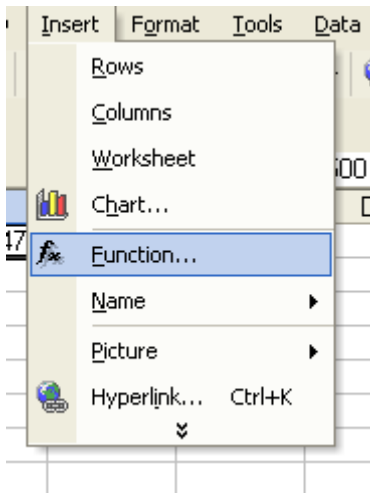
- Xác suất để một sinh viên được chọn ngẫu nhiên có điểm số thấp hơn 600 là bao nhiêu?
- Xác suất để một sinh viên được chọn ngẫu nhiên có điểm số cao hơn 600 là bao nhiêu?
- Xác suất để một sinh viên được chọn ngẫu nhiên có điểm số nằm trong khoảng 400-600 là bao nhiêu?

Gợi ý: Khi sử dụng Excel chúng ta có thể tìm được xác suất của một biến X nhận giá trị nhỏ hơn hoặc bằng một giá trị cho trước nào đó. Và khi chúng ta đã biết trị trung bình và độ lệch chuẩn, chúng ta phải suy nghĩ một cách “thông minh” để tính toán vì chúng ta biết rằng tổng xác suất nằm dưới đường cong chuẩn là bằng 1.

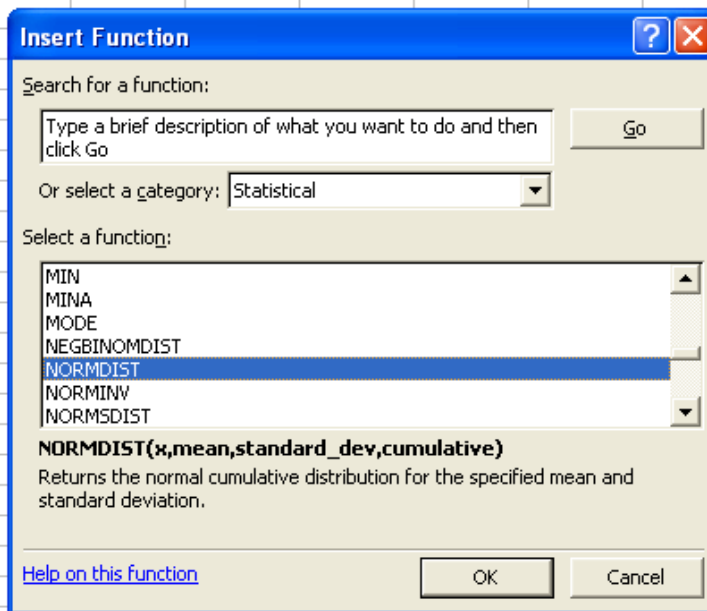
Giải đáp

Bước 1. Chọn ô mà ta muốn Excel xuất kết quả, sau đó chọn **Insert**

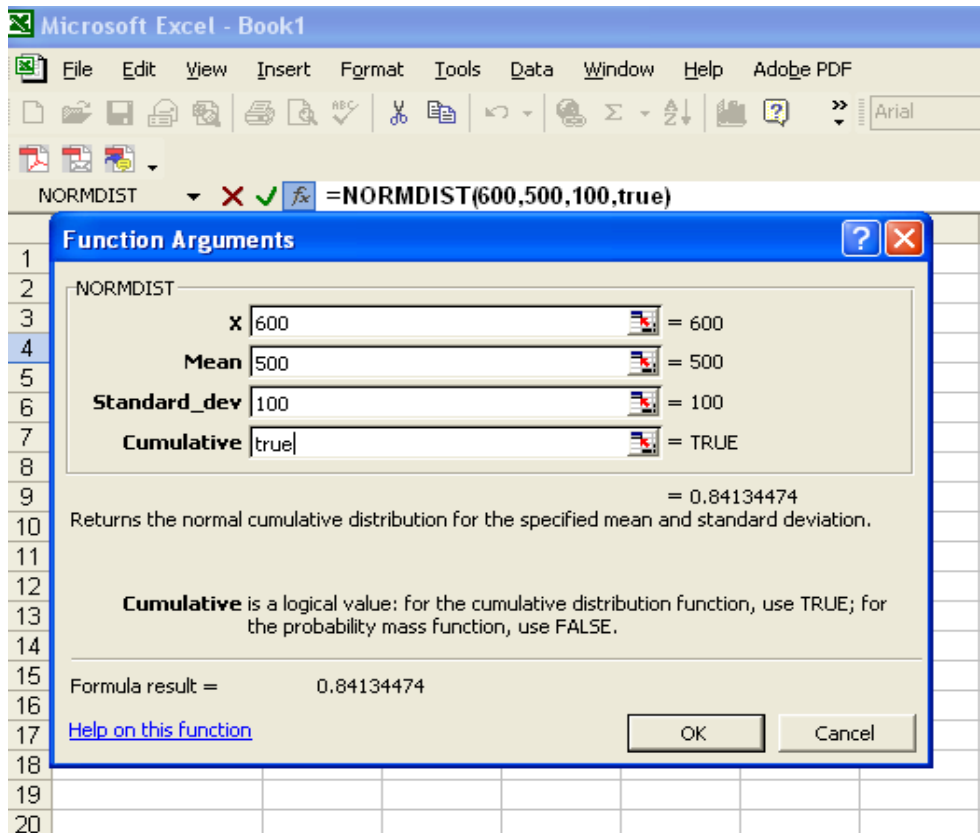
Bước 2. Sau khi bấm vào **insert** chúng ta chọn **Function**



Bước 3. Sau khi chúng ta bấm vào **Function**, cửa sổ **insert function** sẽ xuất hiện. Chúng ta sẽ chọn **statistical**, và sau đó chọn **Normdist** trong số các hàm có sẵn trong Excel



Bước 4. Sau khi nhấn OK, cửa sổ Normdist sẽ xuất hiện, và chúng ta cung cấp các thông số cần thiết. Chúng ta điền 600 vào X, 500 vào ô mean, 100 vào ô standard deviation, và điền true vào ô cumulative box, và sau đó nhấn OK.



Chúng ta sẽ có kết quả sau

| | A1 | | | | |
|---|------------|---|---|---|---|
| | | | | | |
| | A | B | C | D | E |
| 1 | 0.84134474 | | | | |
| 2 | | | | | |

Như chúng ta thấy, xác suất để một học sinh được chọn ngẫu nhiên có số điểm thấp hơn 600 là 0.84134474. Để trả lời được câu b, chúng ta lấy 1 trừ đi con số này và kết quả là 0.158653. Đây là xác suất để một học sinh được chọn ngẫu nhiên có số điểm lớn hơn 600. Thực hiện các bước như trên và suy nghĩ một cách hợp lý chúng ta có thể tính được xác suất một học sinh được chọn ngẫu nhiên sẽ có số điểm nằm trong khoảng 400-600. Người đọc nên lấy đây làm bài tập cho chính mình để thực hiện thành thạo các bước ở trên.

6 XÂY DỰNG KHOẢNG TIN CẬY CHO TRUNG BÌNH TỔNG THỂ

Giả sử chúng ta muốn xây dựng khoảng tin cậy cho trung bình của một tổng thể. Tùy theo qui mô của mẫu thống kê mà chúng ta có thể sử dụng một trong số các trường hợp sau

6.1 Khi qui mô của mẫu thống kê lớn (n lớn hơn 30)

Công thức chung để xây dựng khoảng tin cậy cho trung bình tổng thể là

$$\bar{x} \pm Z^* (S / \sqrt{n})$$

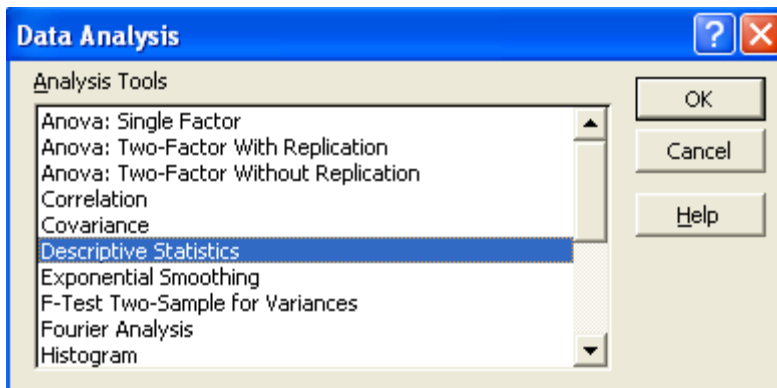
trong đó \bar{x} là trung bình mẫu; Z là hệ số khoảng tin cậy chúng ta có thể tìm thấy trong bảng phân phối chuẩn (ví dụ, hệ số khoảng tin cậy cho khoảng tin cậy 95% là 1.96). S là độ lệch chuẩn của mẫu và n là kích thước của mẫu (số lượng các quan sát của mẫu).

Chúng ta muốn sử dụng Excel để xây dựng khoảng tin cậy cho trung bình tổng thể dựa trên các thông tin của mẫu thống kê. Như chúng ta sẽ thấy, để sử dụng được công thức trên, chúng ta cần có trung bình mẫu \bar{x} , và biên độ sai số $Z^* (S / \sqrt{n})$. Excel sẽ tính toán các đại lượng này cho chúng ta. Điều duy nhất mà chúng ta phải làm là cộng biên độ sai số $Z^* (S / \sqrt{n})$ vào trung bình mẫu \bar{x} để tìm chặn trên của khoảng tin cậy và lấy trung bình mẫu \bar{x} trừ đi biên độ sai số $Z^* (S / \sqrt{n})$ để tìm chặn dưới của khoảng tin cậy.

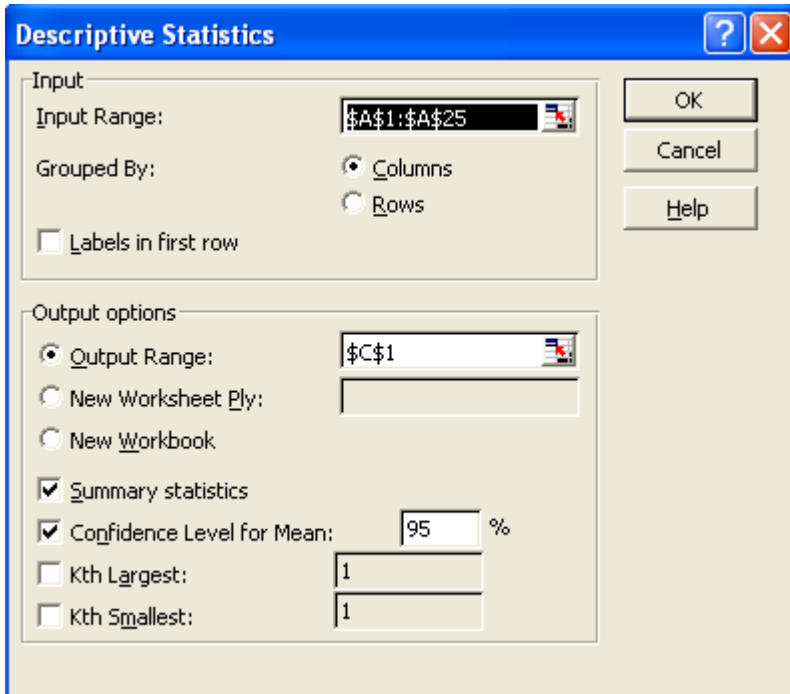
Sau khi nhập dữ liệu vào Excel, chúng ta lại thực hiện các bước như chúng ta đã thực hiện đối với việc tính toán các đại lượng thống kê mô tả. Công việc duy nhất khác với việc tính toán các đại lượng thống kê mô tả là lần này chúng ta sẽ chọn ô confidence interval (khoảng tin cậy) trong cửa sổ descriptive statistics (thống kê mô tả) và chọn mức tin cậy (confidence level), và trong trường hợp này chúng ta chọn 95%. Các bước cụ thể như sau

Bước 1. Nhập dữ liệu: 6, 8, 6.5, 7, 7, 6.5, 8, 6.5, 7, 7, 7.5, 6, 6, 6, 7.5, 8, 7, 6.5, 7, 8, 6, 6, 6.5, 7, 8, 7.5.

Bước 2. Chọn **Tool** và sau đó chọn **Data Analysis**, rồi chọn **Descriptive Statistics**



Bước 3. Trên cửa sổ **Descriptive statistics**, chúng ta chọn Summary Statistics. Sau khi chúng ta đã thực hiện các bước này, chúng ta chọn confidence interval và chọn mức tin cậy là 95%. Ở mục chọn ô để Excel xuất kết quả, chúng ta có thể chọn ô bất kỳ không trùng đè lên các dữ liệu.



Sau khi bấm OK, chúng ta sẽ nhận được kết quả như sau.

| C | D |
|-------------------------|--------------|
| <i>Column1</i> | |
| Mean | 6.94 |
| Standard Error | 0.14525839 |
| Median | 7 |
| Mode | 7 |
| Standard Deviation | 0.726291952 |
| Sample Variance | 0.5275 |
| Kurtosis | -1.214091534 |
| Skewness | 0.137033139 |
| Range | 2 |
| Minimum | 6 |
| Maximum | 8 |
| Sum | 173.5 |
| Count | 25 |
| Confidence Level(95.0%) | 0.299798521 |

Như chúng ta thấy, trung bình của mẫu là $\bar{x} = 6.94$ và giá trị tuyệt đối của sai số

$\left| \pm Z * (S / \sqrt{n}) \right| = 0.2997$. Khoảng tin cậy 95% có chặn trên là $6.94+0.2997$ và chặn dưới

là 6.94-0.2997. Lưu ý rằng chúng ta có thể nói rằng các khoảng tin cậy được xây dựng theo cách này 95% chúng sẽ chứa trung bình tổng thể.

6.2 Mẫu nhỏ (ít hơn 30 quan sát)

Nếu như qui mô của mẫu ít hơn 30 quan sát, chúng ta phải sử dụng một qui trình đối với mẫu nhỏ để xây dựng độ tin cậy cho trung bình của tổng thể. Công thức chung để xây dựng khoảng tin cậy cho trung bình tổng thể dựa trên mẫu qui mô nhỏ là

$$\bar{x} \pm t_{\alpha/2} * (S / \sqrt{n})$$

Trong công thức này \bar{x} là trung bình mẫu, $t_{\alpha/2}$ là hệ số khoảng tin cậy có thể tìm được trong bảng phân phối t với $n-1$ độ tự do (ví dụ hệ số khoảng tin cậy 90% là 1.833 nếu như mẫu có 10 quan sát). S là độ lệch chuẩn của mẫu và n là số quan sát hay kích thước mẫu.

Bây giờ chúng ta sẽ xem Excel được sử dụng để xây dựng khoảng tin cậy của trung bình tổng thể dựa trên một mẫu thống kê kích thước nhỏ. Như chúng ta đã thấy, để sử dụng công thức này, chúng ta phải tính được trung bình mẫu \bar{x} và biên độ sai số $t_{\alpha/2} * (S / \sqrt{n})$ (margin of error). Tương tự như trên điều duy nhất mà chúng ta phải làm là cộng biên độ sai số vào trung bình mẫu để tính chặn trên và lấy trung bình mẫu trừ đi biên độ sai số để tính chặn dưới của khoảng tin cậy.

7 KIỂM ĐỊNH GIẢ THUYẾT VỀ TRUNG BÌNH TỔNG THỂ

Tương tự như trên, chúng ta cần phải phân biệt hai trường hợp là mẫu lớn và mẫu nhỏ.

Mẫu lớn (khi $n > 30$): Ở phần này chúng ta sẽ trình bày cách sử dụng Excel để tiến hành kiểm định giả thuyết về trung bình tổng thể. Chúng ta sẽ sử dụng dữ liệu

Mục tiêu của chúng ta là tiến hành kiểm định giả thuyết trống H_0 nào đó, ví dụ trong trường hợp này chúng ta muốn kiểm định giả thuyết là trị trung bình của một biến ngẫu nhiên nào đó có giá trị là 7 như sau:

$$H_0: \mu = 7$$

với giả thuyết thay thế

$$H_1: \mu \neq 7$$

Ở đây ra sẽ lặp lại các bước để tính các đại lượng thống kê mô tả như ở trên. Điều khác biệt là ta sau đó tính toán giá trị các đại lượng kiểm định.

Bước 1: Chọn **Tool**, sau đó chọn **Data Analysis**, rồi chọn **Descriptive statistics**.

Bước 2: Để tính toán được giá trị đại lượng kiểm định, chúng ta cần biết trị trung bình (mean) và sai số chuẩn (standard error). Ta có thể tìm được các giá trị trên trong bảng kết quả trong **Excel**. Ví dụ, trong bảng kết quả mô tả thống kê ở trên, chúng ta thấy trị trung bình nằm tại ô D3 và sai số chuẩn nằm tại ô D4.

Bước 3: Để tính được giá trị đại lượng kiểm định ta làm như sau: chọn một ô trên bảng tính để hiển thị kết quả, sau đó nhập công thức cho ô đó là $=(C3-7)/C4$. Ở đây ta thấy C3 là giá trị trung bình của mẫu, 7 là giá trị của giả thuyết trống, và C4 là sai số chuẩn, và công thức này là công thức cho phép ta tính giá trị kiểm định Z.

Bước 4: Nếu như giá trị Z lớn nằm ngoài khoảng -1.96 tới +1.96 chúng ta sẽ bác bỏ giả thuyết trống với mức ý nghĩa là 95%, nếu như giá trị Z nằm trong khoảng -1.96 tới +1.96, chúng ta sẽ không bác bỏ giả thuyết trống.

| E1 | | fx =(D3-7)/D4 | | | | |
|----|-----|-------------------------|---------|--------------|----------|---|
| | A | B | C | D | E | F |
| 1 | 6 | | Column1 | | -0.41306 | |
| 2 | 8 | | | | | |
| 3 | 6.5 | Mean | | 6.94 | | |
| 4 | 7 | Standard Error | | 0.14525839 | | |
| 5 | 7 | Median | | 7 | | |
| 6 | 8 | Mode | | 7 | | |
| 7 | 6.5 | Standard Deviation | | 0.726291952 | | |
| 8 | 7 | Sample Variance | | 0.5275 | | |
| 9 | 7 | Kurtosis | | -1.214091534 | | |
| 10 | 7.5 | Skewness | | 0.137033139 | | |
| 11 | 6 | Range | | 2 | | |
| 12 | 6 | Minimum | | 6 | | |
| 13 | 6 | Maximum | | 8 | | |
| 14 | 7.5 | Sum | | 173.5 | | |
| 15 | 8 | Count | | 25 | | |
| 16 | 7 | Confidence Level(95.0%) | | 0.299798521 | | |
| 17 | 6.5 | | | | | |

Mẫu nhỏ (n<30): Lập lại các bước đã sử dụng khi ta có mẫu lớn, Excel có thể được sử dụng để tiến hành kiểm định trong trường hợp chúng ta có mẫu nhỏ. Giả sử chúng ta cũng muốn kiểm định với giả thuyết trống và giả thuyết thay thế như trên

$$H_0: \mu=7$$

với giả thuyết thay thế

$$H_1: \mu \neq 7$$

Lập lại các bước như trên với mẫu nhỏ, nhưng lần này miền giá trị chấp nhận của đại lượng kiểm định t sẽ khác với miền chấp nhận của kiểm định Z. Nếu giá trị đại lượng kiểm định t nằm trong khoảng -2.064 đến +2.064 đối với mức ý nghĩa $\alpha/2=0.025$ và 24 độ tự do, thì chúng ta sẽ không bác bỏ giả thuyết trống, nếu giá trị đại lượng kiểm định t nằm ngoài khoảng này ta sẽ bác bỏ giả thuyết trống. (Với mức ý nghĩa $\alpha/2=0.025$ và 10 độ tự do thì miền giá trị sẽ là -2.228 đến +2.228).

8 KIỂM ĐỊNH SỰ KHÁC BIỆT GIỮA HAI TRUNG BÌNH TỔNG THỂ

8.1 Mẫu lớn:

Tại phần này chúng ta sẽ trình bày cách sử dụng Excel để tiến hành kiểm định về sự chênh lệch hay khác biệt giữa trung bình của hai tổng thể. Giả thiết cơ bản ở đây là hai tổng thể này có phương sai bằng nhau. Giả sử trước khi tiến hành đưa một sản phẩm mới ra thị trường, chúng ta muốn tìm hiểu xem sức mua của người dân thuộc hai thành phố Hà Nội và Hồ Chí Minh có tương đương như nhau hay không và chúng ta tiến hành điều tra về mức thu nhập của người dân tại hai thành phố này. Giả sử mẫu ngẫu nhiên của chúng ta gồm có 35 quan sát thể hiện ở bảng dưới đây. Thu nhập của người dân ở từng thành phố có thể được ký hiệu là X1 và X2 để dễ khái quát hoá.

| | Thu nhập tại Hà Nội X1 | Thu nhập tại Hồ Chí Minh X2 |
|----|---------------------------|--------------------------------|
| 1 | 6 | 6 |
| 2 | 6 | 6 |
| 3 | 6 | 6 |
| 4 | 6 | 6 |
| 5 | 6 | 6 |
| 6 | 6 | 6.5 |
| 7 | 6 | 6.5 |
| 8 | 6.5 | 6.5 |
| 9 | 6.5 | 6.5 |
| 10 | 6.5 | 6.5 |
| 11 | 6.5 | 7 |
| 12 | 6.5 | 7 |
| 13 | 7 | 7 |
| 14 | 7 | 7 |
| 15 | 7 | 7 |
| 16 | 7 | 7.5 |
| 17 | 7 | 7.5 |
| 18 | 7 | 7.5 |
| 19 | 7 | 7.5 |

| | | |
|----|-----|-----|
| 20 | 7.5 | 8 |
| 21 | 7.5 | 8 |
| 22 | 7.5 | 8 |
| 23 | 7.5 | 8 |
| 24 | 7.5 | 8 |
| 25 | 7.5 | 8 |
| 26 | 8 | 8 |
| 27 | 8 | 8.5 |
| 28 | 8 | 8.5 |
| 29 | 8 | 8.5 |
| 30 | 8 | 8.5 |
| 31 | 8 | 8.5 |
| 32 | 8 | 9 |
| 33 | 8 | 9 |
| 34 | 8 | 9 |
| 35 | 8 | 9 |

Miền giá trị của X1 nằm trong khoảng 6-8 còn X2 biến động trong khoảng 6-9. Mục đích chính của chúng ta ở đây là muốn tiến hành kiểm định xem có sự khác biệt đáng kể về giá trị trung bình của hai tổng thể hay không. Giả thuyết trống là hai tổng thể có trị trung bình là như nhau, $H_0: \mu_1 = \mu_2$, và giả thuyết thay thế trung bình hay tổng thể là khác nhau $H_1: \mu_1 \neq \mu_2$, với μ_1 và μ_2 là trung bình của tổng thể của X1 và X2.

Sử dụng công cụ thống kê mô tả trình bày ở trên chúng ta có thể tính toán được trung bình và phương sai của hai mẫu. Excel khi tiến hành kiểm định sự chênh lệch giữa hai tổng thể cần thông tin về phương sai của hai tổng thể. Do chúng ta không biết phương sai của tổng thể (trong hầu hết các trường hợp thì các tham số của tổng thể như kỳ vọng toán hay phương sai là những đại lượng chưa biết), nên chúng ta sử dụng phương sai của mẫu để thay thế cho phương sai tổng thể. Thống kê mô tả cho chúng ta thấy phương sai của mẫu X1 là 0.57, và phương sai của mẫu X2 là 0.98.

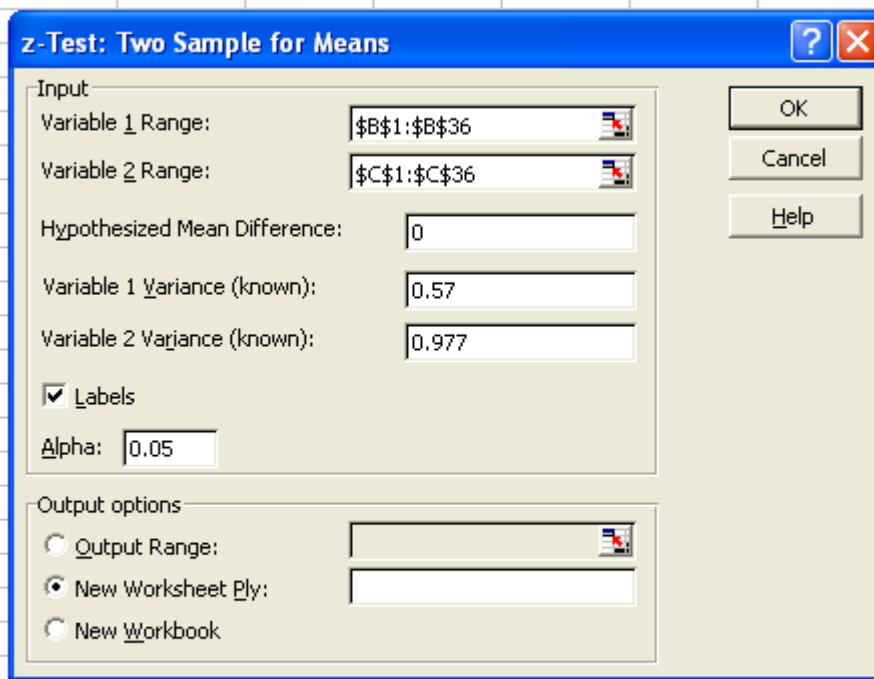
| | A | B | C | D | E |
|----|--------------------|----------|--------------------|----------|---|
| 1 | X1 | | X2 | | |
| 2 | | | | | |
| 3 | Mean | 7.1 | Mean | 7.485714 | |
| 4 | Standard Error | 0.127681 | Standard Error | 0.167138 | |
| 5 | Median | 7 | Median | 7.5 | |
| 6 | Mode | 8 | Mode | 8 | |
| 7 | Standard Deviation | 0.755373 | Standard Deviation | 0.988803 | |
| 8 | Sample Variance | 0.570588 | Sample Variance | 0.977731 | |
| 9 | Kurtosis | -1.38568 | Kurtosis | -1.2219 | |
| 10 | Skewness | -0.19758 | Skewness | -0.05486 | |
| 11 | Range | 2 | Range | 3 | |
| 12 | Minimum | 6 | Minimum | 6 | |
| 13 | Maximum | 8 | Maximum | 9 | |
| 14 | Sum | 248.5 | Sum | 262 | |
| 15 | Count | 35 | Count | 35 | |
| 16 | | | | | |
| 17 | | | | | |
| 18 | | | | | |

Để tiến hành kiểm định giả thuyết về chênh lệch trung bình giữa hai tổng thể với Excel ta thực hiện các bước sau:

Bước 1. Chọn **Tools**, sau đó chọn **Data Analysis** như chúng ta vẫn làm.

Bước 2. Khi cửa sổ Data analysis xuất hiện, chúng ta sẽ chọn **Z-test: two sample for means**, và chọn OK.

Bước 3. Khi cửa sổ **z-test: Two samples for means** xuất hiện, chúng ta sẽ điền khoảng dữ liệu vào khoảng **variable 1 range** và **variable 2 range** tương ứng với X1 và X2. Tiếp sau đó chúng ta sẽ điền 0 (zero) vào ô **Hypothesis mean difference** (về mặt nguyên tắc chúng ta có thể điền bất kỳ giá trị nào mà ta muốn), sau đó ta điền giá trị của phương sai mẫu của biến X1 và X2 vào ô **variable 1 variance** và **variable 2 variance** một cách tương ứng. Tiếp đó chúng ta phải chọn mức ý nghĩa α , chúng ta có thể chọn 0.05 như ví dụ ở trên hoặc chọn bất kỳ giá trị nào mà ta muốn. Cuối cùng chúng ta chọn nơi để Excel xuất kết quả, và chọn OK.



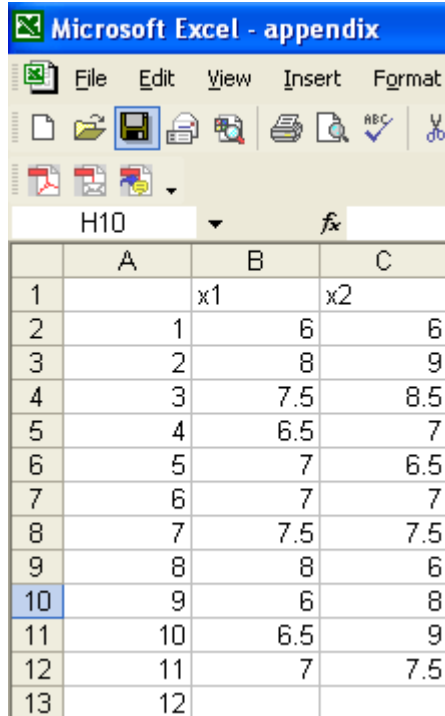
Sau khi bấm OK ta sẽ có cửa sổ kết quả như sau

| | A | B | C |
|----|------------------------------|----------|----------|
| 1 | z-Test: Two Sample for Means | | |
| 2 | | | |
| 3 | | X | Y |
| 4 | Mean | 7.1 | 7.485714 |
| 5 | Known Variance | 0.57 | 0.977 |
| 6 | Observations | 35 | 35 |
| 7 | Hypothesized Mean Difference | 0 | |
| 8 | z | -1.83466 | |
| 9 | P(Z<=z) one-tail | 0.033278 | |
| 10 | z Critical one-tail | 1.644853 | |
| 11 | P(Z<=z) two-tail | 0.066557 | |
| 12 | z Critical two-tail | 1.959963 | |
| 13 | | | |

Ta để ý sẽ thấy một số giá trị tới hạn của đại lượng z với kiểm định 1 bên và kiểm định 2 bên. Tuy theo yêu cầu của đầu bài ta sẽ chọn giá trị tới hạn là một bên hay hai bên cho phù hợp. Nếu như giá trị của đại lượng kiểm định z nằm trong khoảng -1.96 tới $+1.96$ chúng ta sẽ không bác bỏ giả thuyết trống. Ngược lại nếu z nằm ngoài khoảng này ta sẽ bác bỏ giả thuyết trống.

8.2 Mẫu nhỏ: Một trong hai mẫu có số lượng các quan sát nhỏ hơn 30.

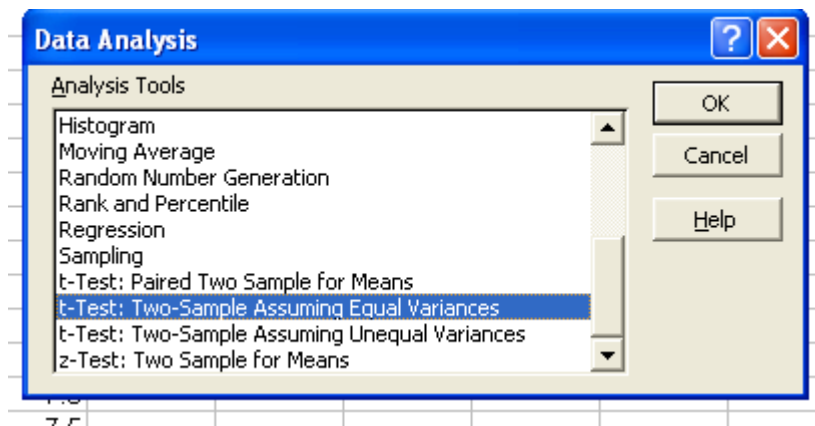
Tại phần này chúng ta sẽ trình bày các sử dụng Excel để kiểm định giả thuyết về sự khác biệt giữa hai trung bình tổng thể khi hai tổng thể có phương sai bằng nhau và số lượng các quan sát trong mẫu nhỏ. Tương tự như trên, mục tiêu chính của việc kiểm định là để đánh giá xem hai trung bình tổng thể có khác nhau hay không. Giả thuyết trống là hai tổng thể có trị trung bình là như nhau, $H_0: \mu_1 = \mu_2$, và giả thuyết thay thế trung bình hay tổng thể là khác nhau $H_1: \mu_1 \neq \mu_2$, với μ_1 và μ_2 là trung bình của tổng thể của X1 và X2. Giả sử lần này chúng ta chỉ có dữ liệu về X1 và X2 như sau



| | A | B | C |
|----|----|-----|-----|
| 1 | | x1 | x2 |
| 2 | 1 | 6 | 6 |
| 3 | 2 | 8 | 9 |
| 4 | 3 | 7.5 | 8.5 |
| 5 | 4 | 6.5 | 7 |
| 6 | 5 | 7 | 6.5 |
| 7 | 6 | 7 | 7 |
| 8 | 7 | 7.5 | 7.5 |
| 9 | 8 | 8 | 6 |
| 10 | 9 | 6 | 8 |
| 11 | 10 | 6.5 | 9 |
| 12 | 11 | 7 | 7.5 |
| 13 | 12 | | |

Ở phần trên chúng ta phải tính phương sai của hai mẫu, ở phần này Excel sẽ tự động tính toán các đại lượng này và sử dụng các đại lượng này trong việc tính toán giá trị của đại lượng kiểm định. Để kiểm định giả thuyết trong trường hợp này chúng ta thực hiện các bước sau

Bước 1. Trong Excel chọn **Tool**, sau đó chọn **Data Analysis**, khi cửa sổ **Data Analysis** xuất hiện, chúng ta chọn: **t-test: Two sample assuming equal variance**, sau đó nhấn OK.



Bước 2. Khi cửa sổ t-test: Two samples assuming equal variance xuất hiện, chúng ta nhập khoảng dữ liệu của biến X1 vào ô **variable 1 range**, và khoảng dữ liệu của biến X2 vào ô **variable 2 range**, sau đó điền số 0 (zero) vào ô **Hypothesis mean difference** (nếu như chúng ta muốn kiểm định với một giá trị khác, ví dụ như $\mu_1 - \mu_2 = 2$, thì ta có thể điền giá trị đó vào ô này), ta chọn **labels** để Excel nhận biết rằng dòng trên cùng của hai cột dữ liệu X1 và X2 không phải là số liệu dùng để tính toán, tiếp đó ta chọn mức ý nghĩa α , ở trong ví dụ này ta chọn 0.05 là mức vẫn thường được sử dụng, mặc dù chúng ta có thể chọn bất kỳ mức α nào mà ta muốn, sau đó ta chọn một nơi để Excel xuất kết quả, và bấm OK.

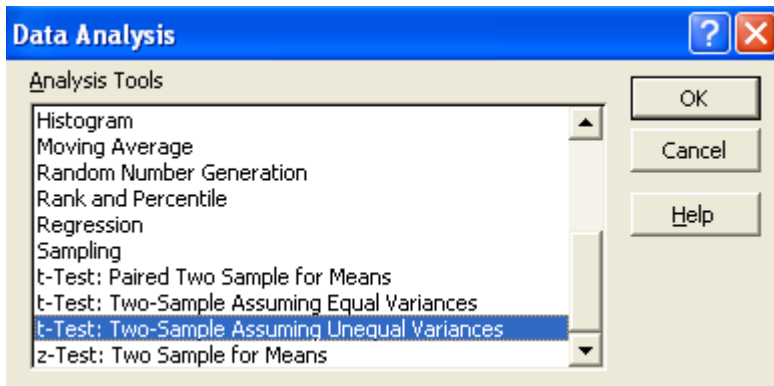
The screenshot shows an Excel spreadsheet with the following data:

| | A | B | C | D | E |
|----|---|----------|----------|---|---|
| 1 | t-Test: Two-Sample Assuming Equal Variances | | | | |
| 2 | | | | | |
| 3 | | x_1 | x_2 | | |
| 4 | Mean | 7 | 7.454545 | | |
| 5 | Variance | 0.5 | 1.172727 | | |
| 6 | Observations | 11 | 11 | | |
| 7 | Pooled Variance | 0.836364 | | | |
| 8 | Hypothesized Mean Difference | 0 | | | |
| 9 | df | 20 | | | |
| 10 | t Stat | -1.16563 | | | |
| 11 | P(T<=t) one-tail | 0.128739 | | | |
| 12 | t Critical one-tail | 1.724718 | | | |
| 13 | P(T<=t) two-tail | 0.257477 | | | |
| 14 | t Critical two-tail | 2.085962 | | | |
| 15 | | | | | |

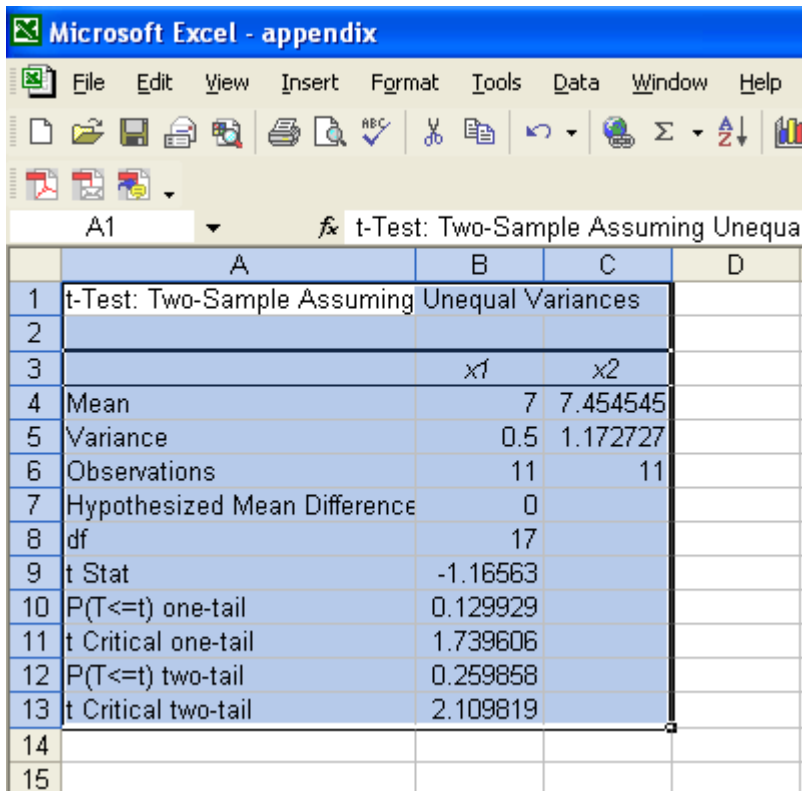
Giá trị của đại lượng kiểm định t là -1.16563. Chúng ta sẽ bác bỏ giả thuyết về hai tổng thể có kỳ vọng toán (trung bình tổng thể) bằng nhau nếu đại lượng kiểm định $t < -2.086$ hoặc $t > +2.086$. Các giá trị này có thể được tìm bằng cách tra bảng phân phối t , trong đó t là giá trị của biến ngẫu nhiên tuân thủ phân phối t có $n_1 + n_2 - 2$ bậc tự do với mức ý nghĩa $\alpha = 0.05$. Trong bảng kết quả Excel, ta thấy giá trị của đại lượng kiểm định không nằm ngoài khoảng từ -2.086 đến +2.086, nên ta không bác bỏ giả thuyết trống. Ta cũng có thể sử dụng đại lượng thống kê p , nếu ta so sánh đại lượng này với mức ý nghĩa α , ta cũng sẽ không bác bỏ giả thuyết trống.

Lưu ý: Ở trên chúng ta vừa tiến hành kiểm định dựa trên giả thiết là hai tổng thể có phương sai bằng nhau. Nếu chúng ta không muốn sử dụng giả thiết này, chúng ta có thể chọn **t-test: Two samples assuming unequal variances**. Trên thực tế ứng dụng, hai kiểm định này trong hầu hết các trường hợp là cho kết quả như nhau. Tuy nhiên, chúng ta sẽ “an toàn” hơn khi sử dụng kiểm định t và không giả thiết là hai tổng thể có phương sai

bằng nhau. Để thực hiện kiểm định này, chúng ta chọn **t-test: Two samples assuming unequal variance** như sau:



Sau đó lặp lại các bước như trong trường hợp hai tổng thể có phương sai bằng nhau ta có kết quả



| | A | B | C | D |
|----|---|----------|----------|---|
| 1 | t-Test: Two-Sample Assuming Unequal Variances | | | |
| 2 | | | | |
| 3 | | x1 | x2 | |
| 4 | Mean | 7 | 7.454545 | |
| 5 | Variance | 0.5 | 1.172727 | |
| 6 | Observations | 11 | 11 | |
| 7 | Hypothesized Mean Difference | 0 | | |
| 8 | df | 17 | | |
| 9 | t Stat | -1.16563 | | |
| 10 | P(T<=t) one-tail | 0.129929 | | |
| 11 | t Critical one-tail | 1.739606 | | |
| 12 | P(T<=t) two-tail | 0.259858 | | |
| 13 | t Critical two-tail | 2.109819 | | |
| 14 | | | | |
| 15 | | | | |

9 TƯƠNG QUAN TUYẾN TÍNH VÀ PHÂN TÍCH HỒI QUI

Tại phần này chúng ta tìm hiểu xem liệu giữa hai biến ngẫu nhiên x và y có tương quan với nhau hay không. Sau đó chúng ta sẽ xây dựng một mô hình để có thể dự đoán một biến này thông qua một biến khác. Có rất nhiều ví dụ mà chúng ta có thể sử dụng, nhưng chúng ta sẽ đề cập tới một ví dụ hay được sử dụng trong kinh doanh. Thông thường biến độc lập (biến giải thích) được ký hiệu bằng chữ X và biến phụ thuộc được ký hiệu bằng chữ Y . Một nhà kinh doanh muốn xem xét xem liệu có mối quan hệ giữa số lượng hộp soda bán được và nhiệt độ trong những ngày hè nóng dựa trên những thông tin trong quá khứ. Đồng thời nhà kinh doanh này cũng muốn ước lượng số lượng hộp soda mà anh ta có thể bán trong một ngày hè nóng. Để làm được điều này, nhà kinh doanh ghi chép cẩn thận nhiệt độ và số lượng hộp soda bán được trong những ngày này. Bảng dữ liệu sau đây cho ta biết số liệu từ ngày 1/6 đến ngày 13/6. Người dự báo thời tiết trên truyền hình dự

báo là nhiệt độ sẽ lên tới 94 độ F vào ngày 14/6, và nhà kinh doanh muốn đáp ứng tất cả nhu cầu cho khách hàng đổi vào ngày 14/6.

| Ngày | 1/6 | 2/6 | 3/6 | 4/6 | 5/6 | 6/6 | 7/6 | 8/6 | 9/6 | 10/6 | 11/6 | 12/6 | 13/6 |
|--------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|------|------|------|
| Hộp soda bán | 57 | 59 | 65 | 67 | 75 | 81 | 86 | 88 | 88 | 84 | 82 | 80 | 83 |
| Nhiệt độ | 56 | 58 | 63 | 66 | 73 | 78 | 85 | 85 | 87 | 84 | 88 | 84 | 89 |

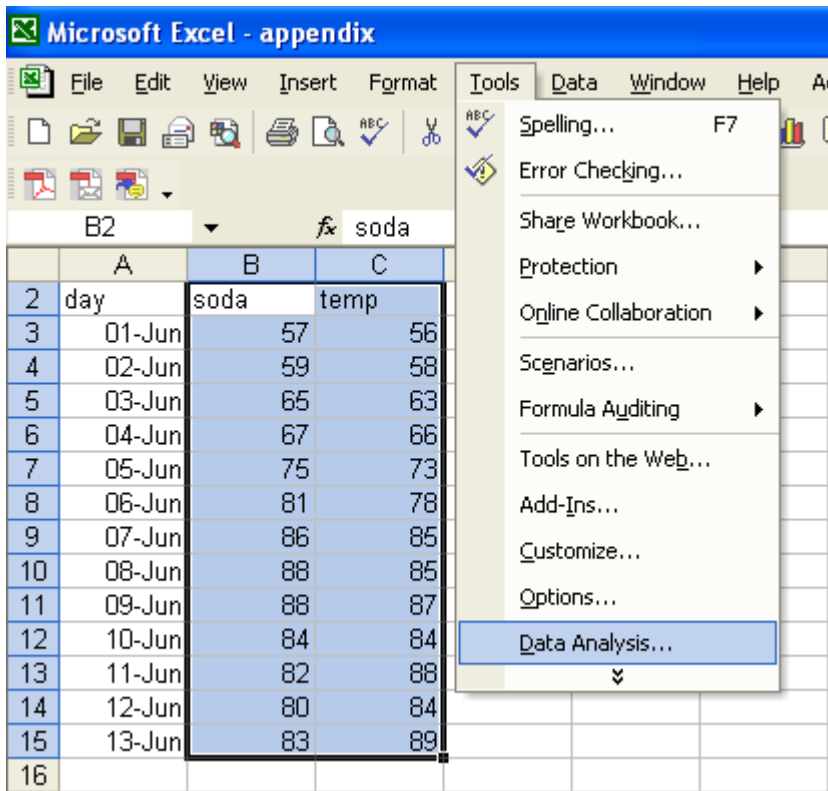
Trước hết, chúng ta hãy sử dụng Excel để tìm hệ số tương quan tuyến tính giữa lượng hộp soda đã bán và nhiệt độ trong ngày. Sau đó ta sẽ sử dụng Excel để tìm đường hồi qui.

9.1 Phân tích tương quan tuyến tính

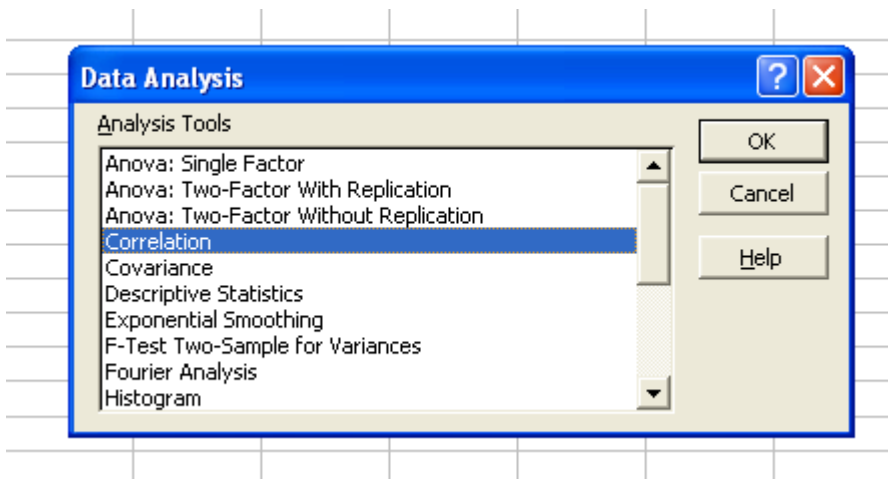
Hệ số tương quan tuyến tính là một đại lượng nằm trong khoảng -1 và +1. Đại lượng này được ký hiệu bằng r. Để tìm r ta thực hiện các bước sau:

Bước 1. Mở Excel và nhập dữ liệu sau đó tô đậm dữ liệu ta cần phân tích, tiếp theo đó từ thanh menu ta chọn **Tool** và chọn **Data analysis**

Khi ta tô đậm dữ liệu cần phân tích như ở trên, thì ở bước sau Excel sẽ rất thông minh để nhận biết dữ liệu ta cần phân tích là khoảng dữ liệu nào, và ta sẽ không phải điền khoảng dữ liệu ở bước 3.

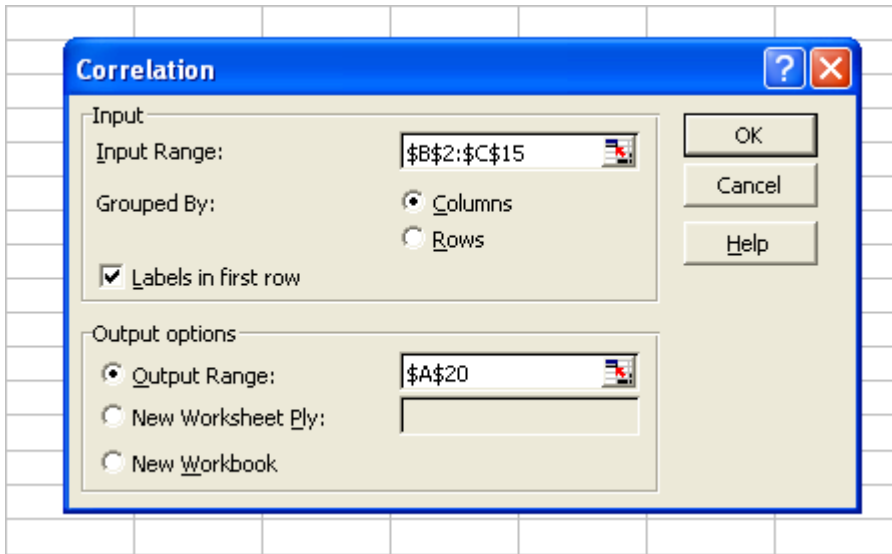


Bước 2. Khi cửa sổ **Data Analysis** xuất hiện, ta chọn **correlation**



Bước 3. Khi cửa sổ tương quan xuất hiện ta điền khoảng dữ liệu vào mục **input range** của cửa sổ này, sau đó nhấn OK. Nếu như ở bước 1 ta đã tô đậm khoảng dữ liệu rồi thì Excel sẽ nhận biết điều này và ta sẽ không phải điền vào khoảng dữ liệu vào mục **input**

range nữa. Đồng thời ta chọn Labels in first row để Excel nhận biết và ta cũng chọn khoảng dữ liệu đầu ra **output range** là nơi để Excel xuất kết quả phân tích.



Sau đó ta sẽ thu được kết quả như sau

| | | | |
|----|------|----------|------|
| 19 | | | |
| 20 | | soda | temp |
| 21 | soda | | 1 |
| 22 | temp | 0.966599 | 1 |
| 23 | | | |

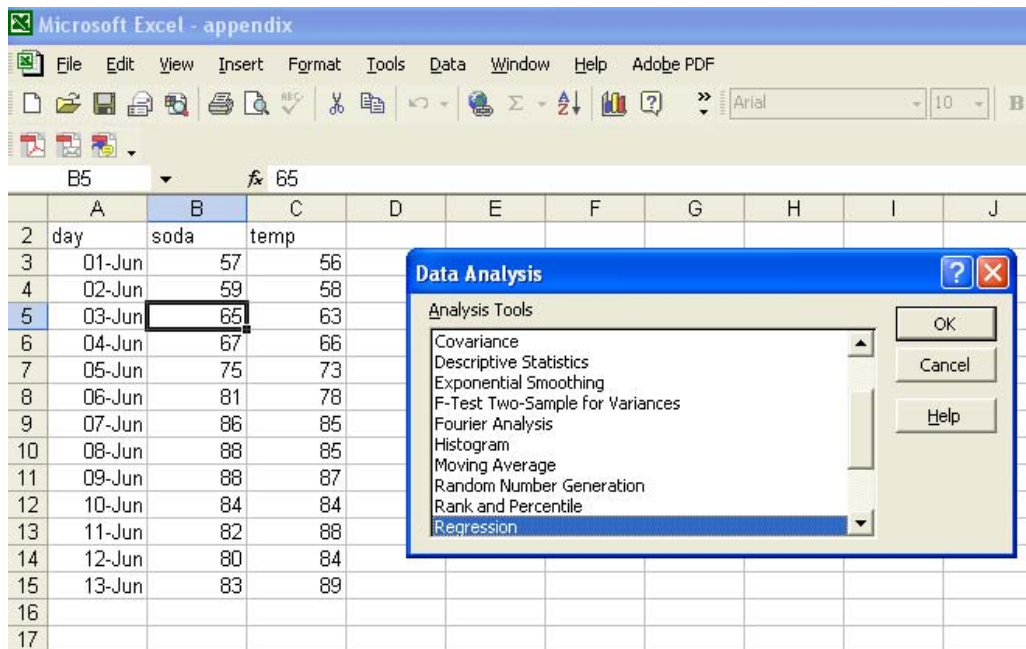
Như ta thấy hệ số tương quan là rất gần +1, như vậy quan hệ tương quan giữa hai biến là rất mạnh. Điều này có nghĩa là khi nhiệt độ tăng lên thì nhu cầu đối với nước uống soda hộp cũng tăng lên.

9.2 Phân tích hồi qui

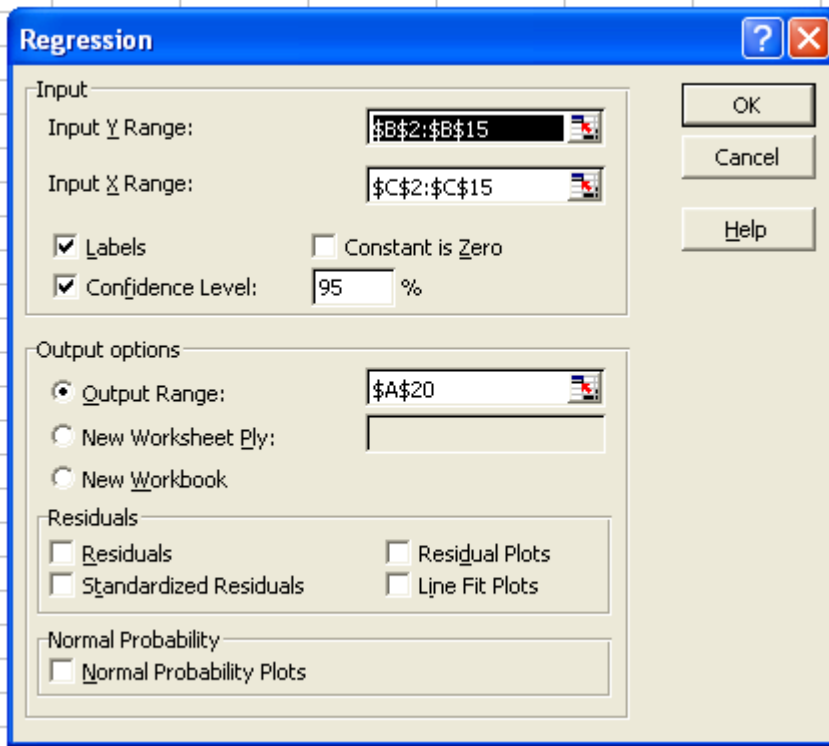
Để tìm đường hồi qui, ta cũng thực hiện các bước tương tự như vậy.

Bước 1: Sau khi đã nhập dữ liệu, ta chọn **Tool** và chọn **data analysis**

Bước 2: Khi cửa sổ **data analysis** xuất hiện, ta chọn **regression**



Bước 3: Khi cửa sổ regression xuất hiện, ta điền khoảng dữ liệu vào cho biến phụ thuộc Y và biến độc lập X, đồng thời chọn **Labels**. Ở đây biến X và Y hoàn toàn do ta lựa chọn. Người nghiên cứu phải thận trọng trong việc tiến hành phân tích hồi qui. Excel chỉ là một công cụ và nó chỉ thực hiện các lệnh mà ra yêu cầu nó thực hiện.



Bước 4: Sau đó tiến hành chọn nơi để Excel xuất kết quả ra. Ta thực hiện điều này bằng cách cung cấp thông tin cho Excel bằng cách điền vào **output range**, sau đó ấn OK.

| SUMMARY OUTPUT | | | | | | |
|------------------------------|---------------------|-----------------------|---------------|----------------|-----------------------|------------------|
| <i>Regression Statistics</i> | | | | | | |
| Multiple R | | 0.966598577 | | | | |
| R Square | | 0.934312809 | | | | |
| Adjusted R Square | | 0.928341246 | | | | |
| Standard Error | | 2.919383191 | | | | |
| Observations | | 13 | | | | |
| <i>ANOVA</i> | | | | | | |
| | <i>df</i> | <i>SS</i> | <i>MS</i> | <i>F</i> | <i>Significance F</i> | |
| Regression | 1 | 1333.479989 | 1333.48 | 156.4603 | 7.58511E-08 | |
| Residual | 11 | 93.75078034 | 8.522798 | | | |
| Total | 12 | 1427.230769 | | | | |
| | <i>Coefficients</i> | <i>Standard Error</i> | <i>t Stat</i> | <i>P-value</i> | <i>Lower 95%</i> | <i>Upper 95%</i> |
| Intercept | 9.17800767 | 5.445742836 | 1.685355 | 0.120045 | -2.80799756 | 21.1640129 |
| temp | 0.879202711 | 0.07028892 | 12.50841 | 7.59E-08 | 0.724497763 | 1.033907659 |

Quan hệ giữa số hộp soda bán được và nhiệt độ là: $Y=0.879*X+9.178$. Sử dụng công thức này ta có thể dự đoán một cách xấp xỉ số lượng hộp soda có thể bán được vào ngày

14/6. Nhiệt độ được dự đoán là có thể lên tới 94F, và như vậy số hộp soda có thể bán được là:

$$Y=0.879*94+9.178 = \text{khoảng } 92 \text{ hộp.}$$

Ở trên ta mới chỉ xem xét hàm hồi qui tuyến tính đơn giản, trong đó biến phụ thuộc chỉ chịu ảnh hưởng của một biến độc lập. Chúng ta có thể mở rộng mô hình hồi qui này bằng cách đưa thêm các biến khác vào mô hình. Điều này có thể được thực hiện trong Excel vô cùng đơn giản. Ở bước 3 vừa nêu trên, khi chúng ta điền khoảng dữ liệu cho biến X ta sẽ chọn nhiều hơn một cột trong bảng tính Excel.