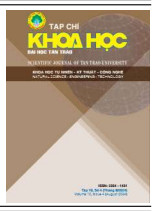
TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/



USING EXCEL SOFTWARE TO PROCESS DATA IN SOME STATISTICAL PROBLEMS TO ENRICH THE CONTENT OF TEACHING PROBABILITY AND STATISTICS AT UNIVERSITY

Bui Linh Phuong, Vu Thi Thu Loan

University of Agriculture and Forestry, Thai Nguyen University

Email address: vuthithuloan@tuaf.edu.vn

https://doi.org/10.51453/2354-1431/2024/1115

Article info

Received:19/10/2024

Revised: 21/11/2024

Accepted: 26/12/2024

Keywords:

Data Analysis

Characteristic Parameters

Confidence intervals

Hypothesis Testing

Correlation and Regression

Abstract:

In this paper, we briefly introduce the theory of some descriptive and inferential statistics problems, along with basic commands for data processing in Excel software. We also present some illustrative examples of calculating common characteristic parameters, confidence intervals for the mean value of a normal distribution, t-test in hypothesis testing problems, and regression correlation problems. The effectiveness of using Excel software is illustrated by comparing the results with theoretical formulas and pocket calculators.

TẠP CHÍ KHOA HỌC ĐẠI HỌC TÂN TRÀO

ISSN: 2354 - 1431

http://tckh.daihoctantrao.edu.vn/

SỬ DỤNG PHẦN MỀM EXCEL ĐỂ XỬ LÝ DỮ LIỆU TRONG MỘT SỐ BÀI TOÁN THỐNG KÊ NHẰM LÀM PHONG PHÚ NỘI DUNG GIẢNG DẠY MÔN XÁC SUẤT THỐNG KÊ Ở TRƯỜNG ĐẠI HỌC

Bùi Linh Phương, Vũ Thị Thu Loan

Trường Đại học Nông Lâm, Đại học Thái Nguyên

Địa chỉ email: vuthithuloan@tuaf.edu.vn

https://doi.org/10.51453/2354-1431/2024/1115

Thông tin bài viết

Ngày nhận bài: 19/10/2024

Ngày sửa bài: 21/11/2024

Ngày duyệt đăng: 26/12/2024

Từ khóa:

Phân tích dữ liệu

Tham số đặc trưng

Khoảng tin cậy

Kiểm định giả thuyết

Tương quan và hồi quy

Tóm tắt

Trong bài báo này, chúng tôi giới thiệu tóm tắt lý thuyết về một số bài toán thống kê mô tả và thống kê suy luận, cùng các câu lệnh cơ bản để xử lý dữ liệu trong phần mềm Excel. Chúng tôi cũng trình bày một số ví dụ minh hoạ về tính toán các tham số đặc trưng thông dụng, khoảng tin cậy cho giá trị trung bình trong phân phối chuẩn, t – Test trong bài toán kiểm định giả thuyết và bài toán tương quan hồi quy. Hiệu quả của việc sử dụng phần mềm Excel được minh hoạ khi so sánh kết quả với việc sử dụng công thức lý thuyết và máy tính bỏ túi.

1. Mở đầu

Xác suất thống kê là một môn học thuộc ngành toán ứng dụng, trang bị các khái niệm cơ bản về lý thuyết xác suất và phân tích thống kê. Môn học này có ứng dụng rất lớn ở hầu hết các lĩnh vực của đời sống. Trong nội dung giảng dạy xác suất thống kê ở hầu hết các trường đại học, sinh viên mới chỉ được học các công thức tính và thực hiện tính toán thủ công bằng máy tính bỏ túi, trong khi thực tế các dữ liệu thường khá lớn nên việc tính thủ công sẽ gặp khó khăn, điều đó dẫn tới nhu cầu sử dụng một phần mềm để giúp việc xử lý dữ liệu tốt hơn.

Có rất nhiều phần mềm có thể giúp xử lý dữ liệu trong thống kê như *SPSS, Stata, R,* …, tuy nhiên trong số đó *Excel* là phần mềm khá đơn giản, thông dụng với sinh viên, có khả năng xử lý dữ liệu gần như chuyên nghiệp.

Thống kê mô tả (*Descriptive Statistics*) là các phương pháp dùng để tóm tắt và mô tả một tập hợp dữ liệu hay một mẫu nghiên cứu. Nó giúp trình bày các thông tin quan trọng về dữ liệu, bao gồm các chỉ số trung bình, phương sai, và phân phối, từ đó giúp người nghiên cứu hiểu rõ hơn về những đặc điểm của dữ liệu đó.

Thống kê suy luận (*Inferential Statistics*) là một nhánh của thống kê sử dụng dữ liệu từ mẫu nghiên cứu để rút ra các kết luận về tổng thể lớn hơn. Nó giúp các nhà nghiên cứu đưa ra những giả thuyết và suy luận dựa trên dữ liệu thu thập được từ mẫu thay vì phải khảo sát toàn bộ quần thể.

Bài viết sẽ giới thiệu một số tiện ích được sử dụng trong *Excel* để giải quyết các bài toán về thống kê mô tả và thống kê suy luận. Đặc biệt, chúng tôi đưa ra các ví dụ thực tế, vừa giải quyết bằng phương pháp sử dụng công thức và máy tính bỏ túi, vừa xử lý bằng phần mềm để bạn đọc thấy rõ hơn tính "ưu việt" của việc sử dụng phần mềm. Hy vọng rằng, sẽ giúp cho nội dung giảng dạy môn xác suất thống kê ở trường đại học thêm phong phú, cũng như giúp sinh viên giải quyết được những bài toán gần với thực tế hơn.

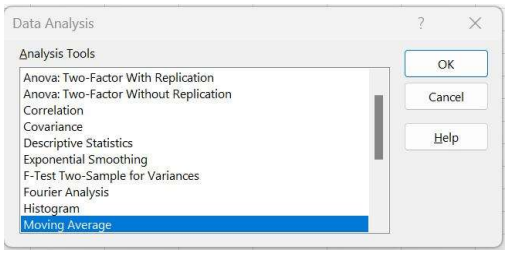
2. Nội dung nghiên cứu

2.1. Hướng dẫn cài đặt tiện ích *Data Analysis* trong Excel [1]

- Với *Excel 2003*, công cụ phân tích dữ liệu *Data Analysis* thuộc đơn lệnh *Tools*. Nếu trong *Tools* không thấy công cụ *Data Analysis* thì ta phải gọi công cụ này bằng cách: *Nhấp vào Tools → Add - ins →Analysis ToolPak → OK.*

Nếu trong đơn lệnh *Tools* cũng không thấy *Add - ins*, ta phải chạy chương trình *Setup*, chọn lệnh *Add/Remove* , … , rồi thực hiện các tùy chọn trong hộp thoại.

- Với *Excel* các phiên bản sau đó không còn tích hợp sẵn trên *menu* nữa. Để kích hoạt tiện ích này ta thực hiện như sau: Nhấp vào *File → Option → Add - ins → Analysis ToolPak → Go → Analysis ToolPak → OK.* Khi đó, ta vào đơn lệnh *Data* sẽ thấy xuất hiện công cụ *Data Analysis*.



Hình 1

2.2. Bài toán tính các đặc trưng mẫu trong thống kê mô tả

2.2.1. Một số tham số đặc trưng thông dụng [2]

Giả sử ta có bộ dữ liệu thực nghiệm gồm *n* phần tử $(x_1, x_2, ..., x_n)$, khi đó:

a/ *Trung bình mẫu*: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ hoặc $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i$ trong đó $\sum_{i=1}^{k} n_i = n$.

b/ *Phương sai điều chỉnh mẫu:* $s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2\right)$

hoặc $s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \left(\bar{x}\right)^2\right)$

c/ *Độ lệch tiêu chuẩn điều chỉnh mẫu:* $s = \sqrt{s^2}$

d / Mode là giá trị xuất hiện nhiều nhất trong tập dữ liệu

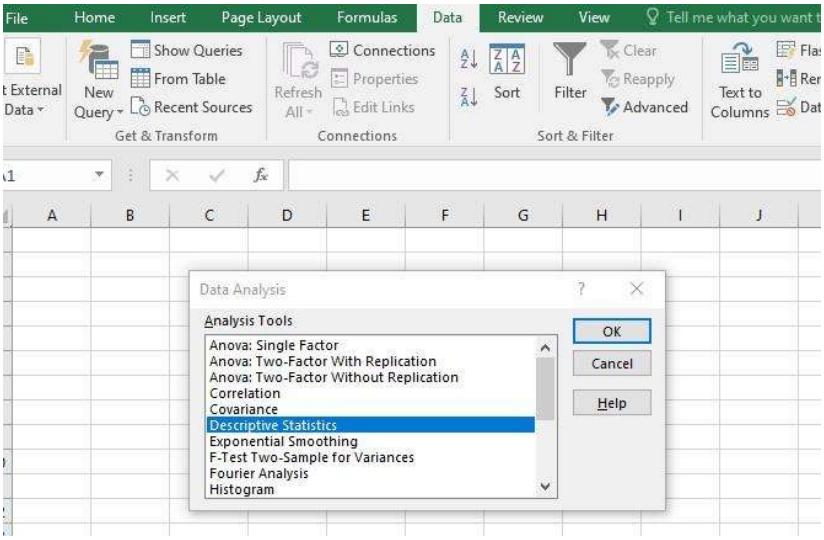
e/ *Trung vị* là giá trị đứng giữa của tập dữ liệu đã được sắp thứ tự. Khi số quan sát *n* là số lẻ thì trung vị là quan sát ở vị trí thứ $\frac{n+1}{2}$. Khi *n* là số chẵn thì trung vị là trung bình cộng của hai quan sát ở vị trí chính giữa của tập dữ liệu, tức là hai quan sát ở vị trí thứ $\frac{n}{2}$ và $\frac{n+2}{2}$.

2.2.2. Tính các đặc trưng mẫu bằng công cụ *Data Analysis* [3]

Bước 1: Nhập dữ liệu vào trang tính *Excel*, sau đó vào *Data* → *Data Analysis*.

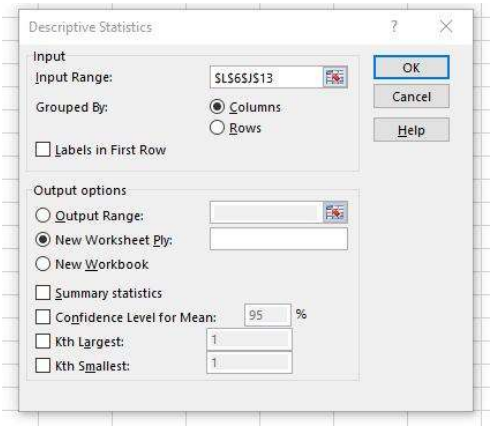
Bước 2: Trong hộp thoại *Data Analysis* vừa mở, chọn *Descriptive Statistics.*

Bước 3: Bấm *OK*.



Hình 2

Bước 4: Trong hộp thoại *Descriptive Statistics*, chúng ta sẽ nhập khoảng dữ liệu, sau đó chúng ta sẽ chọn ô để *Excel* xuất kết quả.



Hình 3

Trong hộp thoại *Descriptive Statistics,* cần lưu ý các tùy chọn:

- *Input Range*: là phạm vi đầu vào, cần chọn biến mà bạn muốn phân tích.

- *Grouped By*: Chọn cách tổ chức biến (theo hàng hoặc theo cột).

- *Output range*: Phạm vi đầu ra.

- *New worksheet ply*: Kết quả xuất hiện trong một trang tính mới.

- *New workbook*: Kết quả xuất hiện trong sổ làm việc mới.

- *Summary statistics:* Tùy chọn này sẽ hiển thị hầu hết các thống kê mô tả.

- *Confidence Level for mean*: Khoảng tin cậy cho giá trị trung bình.

- *Kth Largest*: Hiển thị giá trị cao nhất theo mặc định.

- *Kth Smallest*: Hiển thị giá trị thấp nhất theo mặc định.

Bước 5: Bấm *OK* và xem kết quả thu được.

Cách đọc thông số trên bảng kết quả (chỉ giới thiệu một số đặc trưng thông dụng):

- *Mean*: Giá trị trung bình của tập dữ liệu.

- *Median*: Trung vị.

- *Mode*: hiển thị giá trị xuất hiện nhiều nhất trong dữ liệu.

- *Sample Variance*: Phương sai điều chỉnh mẫu.

- *Standard Deviation*: Độ lệch tiêu chuẩn điều chỉnh mẫu.

- *Minimum*: Hiển thị giá trị nhỏ nhất.

- *Maximum*: Hiển thị giá trị lớn nhất.

- *Sum*: Hiển thị tổng tất cả các giá trị.

- *Count*: Hiển thị kích thước mẫu.

2.2.3. Ví dụ minh họa

Cho dãy số liệu sau: 199, 201, 236, 269, 271, 278, 283, 291, 301, 303, 371. Hãy tính các đặc trưng của mẫu [4].

Cách 1: **Sử dụng công thức và máy tính bỏ túi**

+ Trung bình mẫu: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ = 273

+ Phương sai điều chỉnh mẫu:

$$s^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\bar{x}\right)^2\right)$$

$$= \frac{11}{10}\left(76691{,}36364 - 273^2\right) = 2378{,}6$$

+ Độ lệch tiêu chuẩn điều chỉnh mẫu: $s = \sqrt{2378{,}6}$ = 48,7708

+ Không có *Mode*

+ Trung vị: 278

Cách 2: **Sử dụng đơn lệnh *Data Analysis***

Sau khi nhập dữ liệu vào bảng *Excel*. Trong hộp thoại *Descriptive Statistics*, chúng ta sẽ khai báo dữ liệu đầu vào, vị trí xuất đầu ra, … , cuối cùng ấn *OK*, ta sẽ thu được bảng dưới đây:

199	*Column1*	
201		
236	Mean	273
269	Standard Error	14.70497751
271	Median	278
278	Mode	#N/A
283	Standard Deviation	48.77089296
291	Sample Variance	2378.6
301	Kurtosis	0.77532997
303	Skewness	0.176900515
371	Range	172
	Minimum	199
	Maximum	371
	Sum	3003
	Count	11
	Largest(1)	371
	Smallest(1)	199
	Confidence Level(95.0%)	32.76473171

Hình 4

Nhìn vào bảng trên, ta đọc được các kết quả như sau*:* giá trị trung bình mẫu $\bar{x} = 273;$ độ lệch tiêu chuẩn điều chỉnh mẫu s = 48,7708; trung vị 278; không có *Mode*; phương sai điều chỉnh mẫu $s'^2 = 2378{,}6;$ kích thước mẫu $n = 11$.

2.3. Bài toán ước lượng giá trị trung bình của biến ngẫu nhiên có phân phối chuẩn

2.3.1. Bài toán ước lượng giá trị trung bình của tổng thể bằng khoảng tin cậy [2]

Bài toán: Giả sử X biến ngẫu nhiên có phân phối chuẩn $N(\mu, \delta^2)$ với hai tham số đặc trưng: trung bình tổng thể là μ và phương sai tổng thể σ^2. Hãy ước lượng tham số μ bằng phương pháp ước lượng khoảng tin cậy khi chưa biết

phương sai, với độ tin cậy $1-\alpha$ cho trước.

Bước 1: Dựa vào dữ liệu mẫu, tính các đặc trưng $\bar{x}, s^2, s$.

Bước 2: Với độ tin cậy $1-\alpha$ cho trước:

+ Tìm hai phân vị chuẩn $u_{1-\alpha}$ *hoặc* $u_{1-\frac{\alpha}{2}}$ nếu mẫu có kích thước $n > 30.$

+ Tìm hai phân vị Student $t_{1-\alpha}^{(n-1)}$ *hoặc* $t_{1-\frac{\alpha}{2}}^{(n-1)}$ nếu mẫu có kích thước $n \le 30$.

Bước 3: Tìm khoảng tin cậy:

** Khoảng tin cậy đối xứng:*

$\left(\bar{x}-\frac{s}{\sqrt{n}}.u_{1-\frac{\alpha}{2}};\bar{x}+\frac{s}{\sqrt{n}}.u_{1-\frac{\alpha}{2}}\right)$ với $n>30$

$\left(\bar{x}-\frac{s}{\sqrt{n}}.t_{1-\frac{\alpha}{2}}^{(n-1)};\bar{x}+\frac{s}{\sqrt{n}}.t_{1-\frac{\alpha}{2}}^{(n-1)}\right)$ với $n \le 30$

Đặt $\varepsilon=\frac{s}{\sqrt{n}}.u_{1-\frac{\alpha}{2}}$ hoặc $\varepsilon=\frac{s}{\sqrt{n}}.t_{1-\frac{\alpha}{2}}^{(n-1)}$. Khi đó, ε được gọi là độ chính xác của ước lượng.

Khoảng tin cậy đối xứng có thể viết gọn lại: $(\bar{x}-\varepsilon;\ \bar{x}+\varepsilon)$.

** Khoảng tin cậy bên phải:*

$\left(\bar{x}-\frac{s}{\sqrt{n}}.u_{1-\alpha};+\infty\right);\left(\bar{x}-\frac{s}{\sqrt{n}}.t_{1-\alpha}^{(n-1)};+\infty\right)$

** Khoảng tin cậy bên trái:*

$\left(-\infty;\bar{x}+\frac{s}{\sqrt{n}}.u_{1-\alpha}\right);\left(-\infty;\bar{x}+\frac{s}{\sqrt{n}}.t_{1-\alpha}^{(n-1)}\right)$

2.3.2. Giải bài toán ước lượng giá trị trung bình của tổng thể bằng phần mềm *Excel*[3]

Bước 1: Nhập dữ liệu vào bảng *Excel*, sau đó dùng tiện ích *Data Analysis* để tính các đặc trưng mẫu.

Bước 2: Tìm khoảng ước lượng giá trị trung bình theo công thức: *"= trung bình mẫu ± độ chính xác"*.

2.3.3. Ví dụ minh họa

Phỏng vấn 10 gia đình có ba người ở một huyện vùng cao về chi phí hàng tháng cho nhu yếu phẩm, thu được số liệu như sau: 160, 175, 180, 190, 190, 190, 200, 200, 250, 300 (nghìn đồng). Hãy ước lượng chi phí trung bình hàng tháng cho nhu yếu phẩm của một hộ gia đình có ba người ở huyện này với độ tin cậy 95% [4].

Cách 1: **Sử dụng công thức và máy tính bỏ túi**

Dựa vào số liệu mẫu, ta tính được các đặc trưng:

$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ = 203,5

$s^2=\frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\left(\bar{x}\right)^2\right)$

$=\frac{10}{9}\left(42942{,}5-203{,}5^2\right)$ =1700,2777

$s=\sqrt{1700{,}2777}=41{,}2344$.

Với độ tin cậy 95% ta có $t_{0{,}975}^{(9)}=2{,}262.$

$\varepsilon=\frac{s}{\sqrt{n}}.t_{1-\frac{\alpha}{2}}^{(n-1)}=\frac{41{,}2344}{\sqrt{10}}.2{,}262=29{,}4973$

Vậy khoảng tin cậy đối xứng của chi phí trung bình hàng tháng cho nhu yếu phẩm là:

$(203{,}5-29{,}4973;203{,}5+29{,}4973)$

$=(174{,}0047;232{,}9952)$

Cách 2: **Sử dụng đơn lệnh *Data Analysis***

Sau khi nhập dữ liệu vào trang tính *Excel*. Ta gọi tiện ích *Data Analysis* để xử lý dữ liệu. Lưu ý rằng: 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. Khi đó, ta được kết quả như sau:

		Column1	
160			
175			
180		Mean	203.5
190		Standard Error	13.03947
190		Median	190
190		Mode	190
200		Standard Deviation	41.23442467
200		Sample Variance	1700.277778
250		Kurtosis	2.911219343
300		Skewness	1.729960656
		Range	140
		Minimum	160
		Maximum	300
		Sum	2035
		Count	10
		Confidence Level(95.0%)	29.49733045

Hình 5

Trên bảng kết quả ta có: $\bar{x} = 203{,}5$; $s = 41{,}2344$. Độ chính xác của ước lượng: $\varepsilon = 29{,}4973$. Vậy khoảng tin cậy cần tìm là:

$$(203{,}5 - 29{,}4973; 203{,}5 + 29{,}4973)$$
$$= (174{,}0047; 232{,}9952).$$

2.4. Bài toán kiểm định giả thuyết thống kê giá trị trung bình của tổng thể

2.4.1. Quy trình giải bài toán kiểm định giả thuyết thống kê về giá trị trung bình của tổng thể [2]

Bài toán: Giả sử X biến ngẫu nhiên có phân phối chuẩn $N(\mu, \delta^2)$ với hai tham số đặc trưng: trung bình tổng thể là μ và phương sai tổng thể σ^2. Hãy kiểm định giả thuyết về giá trị trung bình khi chưa biết phương sai.

Bước 1: Đặt giả thuyết $H_0 : \mu = \mu_0$ và đối thuyết $H_1 : \mu \neq \mu_0$ (hoặc $\mu > \mu_0$ hoặc $\mu < \mu_0$)

Bước 2: Chọn tiêu chuẩn kiểm định $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$. Nếu giả thuyết H_0 là đúng thì tiêu chuẩn kiểm định có dạng $T = \dfrac{\overline{X} - \mu_0}{S}\sqrt{n}$.

Bước 3: Dựa vào mẫu cụ thể, ta tính các đặc trưng $\overline{x}, s$ từ đó tính giá trị quan sát T_{qs}.

Bước 4: Với mức ý nghĩa α cho trước, tìm phân vị *chuẩn* $u_{1-\alpha}$ hoặc $u_{1-\frac{\alpha}{2}}$ nếu $n > 30$; tìm hai phân vị *Student* $t_{1-\alpha}^{(n-1)}$ hoặc $t_{1-\frac{\alpha}{2}}^{(n-1)}$ nếu $n \leq 30$.

Bước 5: Tìm miền bác bỏ:

- Nếu $H_1 : \mu \neq \mu_0$ ta dùng miền bác bỏ hai phía:

$$W_\alpha = \left(-\infty; -u_{1-\frac{\alpha}{2}}\right) \cup \left(u_{1-\frac{\alpha}{2}}; +\infty\right).$$

- Nếu $H_1 : \mu > \mu_0$ ta dùng miền bác bỏ bên phải: $W_\alpha = (u_{1-\alpha}; +\infty)$.

- Nếu $H_1 : \mu < \mu_0$ ta dùng miền bác bỏ bên trái: $W_\alpha = (-\infty; -u_{1-\alpha})$.

Lưu ý: trường hợp dùng bảng phân vị *Student* ta cũng có các miền bác bỏ tương tự.

Bước 6: So sánh T_{qs} với W_α rồi rút ra kết luận:

- Nếu $T_{qs} \in W_\alpha$ thì bác bỏ H_0, chấp nhận H_1.

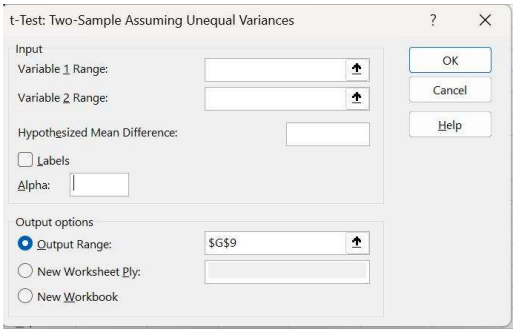
- Nếu $T_{qs} \notin W_\alpha$ thì chưa có cơ sở để bác bỏ H_0, tức là chấp nhận H_0.

2.4.2. Giải bài toán kiểm định giả thuyết về giá trị trung bình của tổng thể bằng phần mềm *Excel* [3]

Sau khi đặt giả thuyết và đối thuyết, nhập dữ liệu vào bảng *Excel*, tiến hành theo các bước:

Bước 1: Chọn *Data*, sau đó chọn *Data Analysis*, chọn *t – Test: two – sample Assuming Unequal Variances*.

Bước 2: Trong hộp thoại *t – Test: two – sample Assuming Unequal Variances*, ta khai báo vị trí biến chính (*Variable 1 range*), biến giả (*Variable 2 range*), giá trị μ_0 (*Hypothesized Mean Difference*), mức ý nghĩa α (*Alpha*), vị trí đầu ra (*Output range*), rồi ấn *OK*:



Hình 6

Bước 3: Màn hình sẽ trả về bảng kết quả với các thông tin về trung bình mẫu, phương sai điều chỉnh mẫu, giá trị quan sát được của tiêu chuẩn kiểm định và giá trị miền bác bỏ.

Bước 4: So sánh T_{qs} với miền bác bỏ để rút ra kết luận.

2.4.3. Ví dụ minh họa

Phòng kỹ thuật của một công ty lắp ráp xe máy theo dõi mức tiêu hao xăng của một loại xe chạy từ địa điểm A đến địa điểm B và ghi lại được số liệu như sau:

Mức xăng tiêu hao (lít)	8,5	9	11	12,5
Số chuyến tương ứng	5	8	10	2

Bảng 1

Với mức ý nghĩa 5%, có thể kết luận mức xăng tiêu hao trung bình thấp hơn 11 lít không? Biết mức xăng tiêu hao là biến ngẫu nhiên tuân theo quy luật phân phối chuẩn [4].

Cách 1: **Sử dụng công thức và máy tính bỏ túi**

Gọi μ là mức xăng tiêu hao trung bình của loại xe này.

Đặt giả thuyết $H_0 : \mu = 11$ và đối thuyết $H_1 : \mu < 11$.

Chọn tiêu chuẩn kiểm định $T = \dfrac{\overline{X} - \mu}{S}\sqrt{n}$, nếu giả thuyết H_0 đúng thì tiêu chuẩn kiểm định có dạng $T = \dfrac{\overline{X} - 11}{S}\sqrt{n}$.

Dựa vào mẫu cụ thể, ta tính các đặc trưng $\overline{x} = 9{,}98;\ s = 1{,}31877$.

$$T_{qs} = \frac{9{,}98 - 11}{1{,}31877}\sqrt{25} = -3{,}86723$$

Với mức ý nghĩa 5% và kích thước mẫu nhỏ, ta có: $t_{0{,}995}^{(24)} = 1{,}711.$

Vì đối thuyết $H_1 : \mu < 11$ ta dùng miền bác bỏ bên trái: $W_\alpha = (-\infty;\ -1{,}711)$.

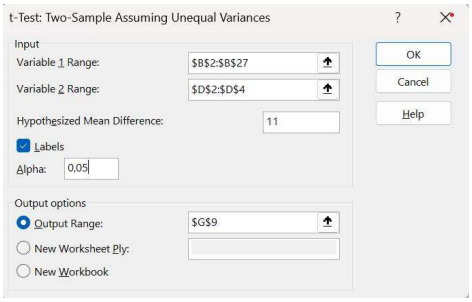
Ta thấy $T_{qs} \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1. Vậy mức xăng hao phí trung bình đúng là thấp hơn 11 lít.

Cách 2: **Sử dụng đơn lệnh *Data Analysis***

Bước đặt giả thuyết và đối thuyết, chọn tiêu chuẩn kiểm định, ta làm như cách 1.

Chọn *Data*, sau đó chọn *Data Analysis*, rồi chọn *t – Test:* two – sample Assuming Unequal Variances.

Trong hộp thoại *t – Test: two – sample Assuming Unequal Variances*, ta khai báo như bảng dưới đây:



Hình 7

Sau khi ấn *OK*, màn hình sẽ trả về bảng kết quả như sau:

t-Test: Two-Sample Assuming Unequal Variances

	Mức xăng hao	*Dummy*	
Mean	9,98	0	Trung bình mẫu
Variance	1,739167	0	Phương sai điều chỉnh
Observati	25	2	kích thước mẫu
Hypothesi	11		a_0
df	24		bậc tự do
t Stat	-3,86723		T_qs
P(T<=t) on	0,000368		
t Critical o	1,710882		Miền bác bỏ t < -1,710882
P(T<=t) tw	0,000736		
t Critical tv	2,063899		

Hình 8

Nhìn vào bảng ta: $T_{qs} = -3{,}86723$. Vì đối thuyết $H_1 : \mu < 11$ nên ta chọn miền bác bỏ bên trái $W_\alpha = (-\infty;\ -1{,}711)$. Ta thấy $T_{qs} \in W_\alpha$ nên bác bỏ H_0, chấp nhận H_1.

2.5. Bài toán tương quan và hồi quy

2.5.1. Hệ số tương quan và phương trình đường hồi quy tuyến tính

a/ Hệ số tương quan lý thuyết [2]

Giả sử *X* và *Y* là hai biến ngẫu nhiên có phương sai dương. *Hệ số tương quan lý thuyết* của hai biến ngẫu nhiên *X* và *Y*, kí hiệu là $\rho(X,Y)$, được xác định như sau:

$$\rho(X,Y) = \frac{E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]}{\sqrt{V(X)}\sqrt{V(Y)}}$$

b/ Hệ số tương quan mẫu [2]

Hệ số tương quan lý thuyết dùng để đo mức độ phụ thuộc tuyến tính giữa hai biến ngẫu nhiên *X* và *Y*. Tuy nhiên, nếu chưa biết phân phối của *(X, Y)* thì việc tính hệ số tương quan lý thuyết gặp khó khăn. Khi đó, hệ số tương quan mẫu được xem như là một ước lượng của hệ số tương quan lý thuyết.

Giả sử tiến hành *n* bước quan sát độc lập đối với cặp biến ngẫu nhiên *(X, Y)* ta có mẫu ngẫu nhiên kích thước *n*: $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$

Hệ số tương quan mẫu của *X* và *Y*, kí hiệu là r được xác định bởi công thức [5]:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

c/ Phương trình đường hồi quy tuyến tính đơn giản [2]

Phương trình đường hồi quy tuyến tính đơn giản của tổng thể là phương trình diễn tả giá trị trung bình của biến phụ thuộc *Y* theo biến độc lập *X* đã biết:

$$Y = A\,X + B$$

d/ Phương trình đường hồi quy tuyến tính mẫu [5]

Phương trình đường hồi quy bình phương trung bình tuyến tính thực nghiệm: $y = a\,x + b$

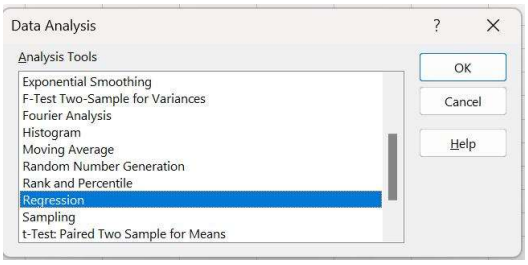
Với $a = \dfrac{\sum_{i=1}^{n} x_i y_i - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$;

$$b = \bar{y} - a\bar{x}$$

2.5.2. Tìm hệ số tương quan và viết phương tình đường hồi quy bằng phần mềm Excel

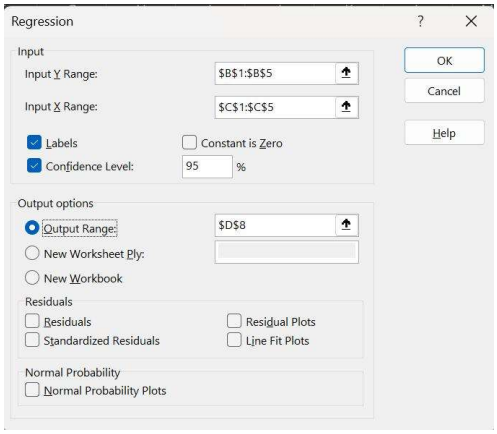
Bước 1: Nhập dữ liệu vào bảng *Excel*.

Bước 2: Mở tiện ích *Data Analysis, sau đó chọn Regression*:



Hình 9

Bước 3: Trong hộp thoại *Regression*, khai báo dữ liệu đầu vào (*Input Y Range, Input X range*), tên biến (*Labels*), độ chính xác (*Confidence Lavel*), đầu ra (*Output range*), rồi ấn *OK*:



Hình 10

Bước 4: Đọc dữ liệu trong bảng kết quả.

2.5.3. Ví dụ minh họa

Để thực hiện một công trình nghiên cứu về mối quan hệ giữa chiều cao *Y* (m) và đường kính *X* (cm) của một loại cây, người ta quan sát trên một mẫu ngẫu nhiên và có kết quả:

x_i	1	3	4	6	8	9	11	14
y_i	1	2	4	4	5	7	8	9

Bảng 2

a/ Hãy tính hệ số tương quan mẫu.

b/ Viết phương trình đường hồi quy bình phương trung bình tuyến tính thực nghiệm của chiều cao của cây theo đường kính của nó [4].

Cách 1: Sử dụng công thức và máy tính bỏ túi

a/ Lập bảng tính toán:

x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	1	1	1	1	$(-6)^2$	$(-4)^2$	24
3	2	9	4	6	$(-4)^2$	$(-3)^2$	12
4	4	16	16	16	$(-3)^2$	$(-1)^2$	3
6	4	36	16	24	$(-1)^2$	$(-1)^2$	1
8	5	64	25	40	1^2	0^2	0
9	7	81	49	63	2^2	2^2	4
11	8	121	64	88	4^2	3^2	12
14	9	196	81	126	7^2	4^2	28
$\sum = 56$	$\sum = 40$	$\sum = 524$	$\sum = 256$	$\sum 364$	$\sum = 132$	$\sum = 56$	$\sum = 84$

Hình 11

Từ bảng trên ta tính được: $\bar{x} = 7; \bar{y} = 5$

a/ Hệ số tương quan mẫu là:

$$r = \frac{\sum_{i=1}^{8}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{8}(x_i - \bar{x})^2 \sum_{i=1}^{8}(y_i - \bar{y})^2}}$$

$$= \frac{84}{\sqrt{132.56}} = 0{,}9770$$

b/ $$a = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{364 - \frac{56.40}{8}}{524 - \frac{56^2}{8}} = 0{,}6364$$

$$b = \bar{y} - a\bar{x} = 5 - 0{,}6364.7 = 0{,}5455$$

Phương trình đường hồi quy bình phương trung bình tuyến tính thực nghiệm của chiều cao cây và đường kính của nó là:

$$y = a\,x + b = 0{,}6364x + 0{,}5455$$

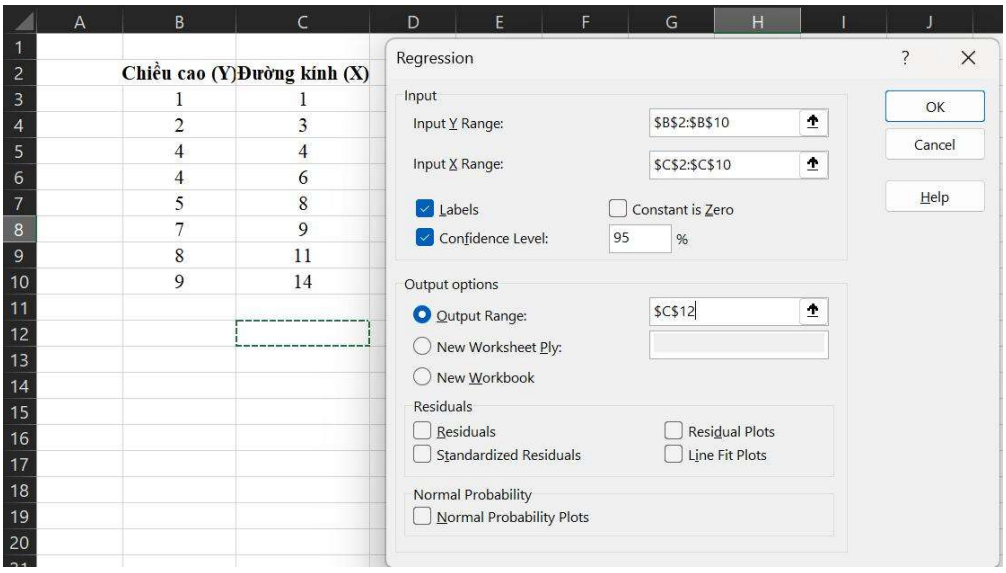
Cách 2: Sử dụng đơn lệnh *Data Analysis*

Nhập dữ liệu vào bảng *Excel* (hình 12):

	A	B	C
1			
2		Chiều cao (Y)	Đường kính (X)
3		1	1
4		3	2
5		4	4
6		6	4
7		8	5
8		9	7
9		11	8
10		14	9

Hình 12

Mở đơn lệnh *Data Analysis*, sau đó chọn *Regression*. Trong hộp thoại *Regression* ta khai báo dữ liệu: vị trí cột đầu vào, đầu ra, độ chính xác, … , sau đó ấn *OK*:



Hình 13

Khi đó bảng *Excel* sẽ xuất hiện bảng kết quả như sau:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,977008
R Square	0,954545
Adjusted R Squa	0,94697
Standard Error	0,651339
Observations	8

ANOVA

	df	*SS*	*MS*	*F*	*gnificance F*
Regression	1	53,45455	53,45455	126	2,99E-05
Residual	6	2,545455	0,424242		
Total	7	56			

	Coefficients	*andard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*	
Intercept	0,545455	0,458818	1,188825	0,279422	-0,57723	1,668142	-0,57723	1,668142	1,349691
Đường kính (X)	0,636364	0,056692	11,22497	2,99E-05	0,497644	0,775083	0,497644	0,775083	1,826982

Hình 14

Nhìn vào bảng kết quả, ta thấy $r = 0{,}9770$ và $y = 0{,}6364x + 0{,}5455$.

3. Kết luận

Xác suất thống kê là một môn học có rất nhiều ứng dụng trong thực tế đặc biệt là cho các ngành kinh tế, tài chính, kế toán, … , tuy nhiên trong chương trình giảng dạy ở trường đại học, môn học Xác suất thống kê mới chỉ dừng lại ở việc cung cấp những kiến thức cơ bản nhất về phân tích thống kê, còn mang nặng tính lý thuyết và việc sử dụng các kiến thức này để xử lý dữ liệu trong thực tế còn khá khó khăn. Bài viết đã giới thiệu tóm tắt kiến thức lý thuyết về một số bài toán thống kê mô tả, thống kê suy diễn và các câu lệnh được sử dụng trong *Excel* để hỗ trợ tính toán. Các ví dụ minh họa cho thấy việc xử lý dữ liệu bằng phần mềm rõ ràng là đơn giản và nhanh gọn hơn rất nhiều so với việc tính toán thủ công. Mặc dù những kiến thức cơ bản về lý xuyết xác suất và thống kê toán là cần thiết khi bắt đầu nhập môn về lĩnh vực thống kê nhưng việc kết hợp bộ công cụ *Data Analysis* trong *Excel* vào môn xác suất thống kê cũng cho thấy tính hữu ích không hề nhỏ trong thời đại 4.0. Hy vọng rằng, việc khai thác chức năng *Data Analysis* trong *Excel* vào giảng dạy xác suất thống kê sẽ giúp người học dễ tiếp cận hơn, học hiệu quả hơn và đặc biệt là bước đầu biết xử lý bộ dữ liệu lớn trong công tác nghiên cứu và làm việc sau này.

TÀI LIỆU THAM KHẢO/ REFERENCES

[1] Nguyen Thanh Tam (2019), *Combine the Data Analysis application in Excel to teach statistics in the subject of statistical probability,* Scientific magazine of Cửu Long University, 13, 86 – 95.

[2] Nguyen Thi Dung, Pham Thanh Hieu, Mai Thi Ngoc Ha (2018), *statistical probability*, Agricultural publisher.

[3] Nguyen Ngoc Anh, Nguyen Dinh Chuc, Doan Quang Hung (2008), *Statistical analysis used in Excel*, Copyright DEPOCEN (Development and policies research center).

[4] Nguyễn Thi Dung, Mai Thi Ngoc Ha, Vi Dieu Minh, Bui Linh Phuong, Duong

Thi Hong (2023), *Statistical probability exercises and practice on R software*, Bach Khoa Ha Noi publisher.

[5] R.E. Walpole, R.H. Myers, S.L. Myers, K. Ye (2011), *Probability & Statistics for Engineers & Scientists*, Prentice Hall, 9th Edition.