|  |
| --- |
| HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG |
|  |
| BÁO CÁO MÔN HỌC |
| **HỌC PHẦN: Nhập môn Khoa học dữ liệu**  Nhóm học phần: 07  Giảng viên hướng dẫn: Vũ Hoài Thư  Nhóm bài tập: 13  Sinh viên thực hiện:  Trương Quốc Bình – B22DCKH008  Đào Văn Công – B22DCKH010  Đinh Văn Khải – B22DCKH064  Nguyễn Chính Nghĩa – B19DCCN465 |
| HÀ NỘI, 2025 |

Mục lục

[1. Giới thiệu bài toán và thư viện sử dụng 2](#_Toc209801326)

[1.1. Giới thiệu bài toán 2](#_Toc209801327)

[1.2. Giới thiệu thư viện sử dụng 3](#_Toc209801328)

[1.2.1. Mục tiêu bài toán 3](#_Toc209801329)

[1.2.2. Bộ thư viện sử dụng 4](#_Toc209801330)

[2. Cơ sở lý thuyết 5](#_Toc209801331)

[2.1. Giả thuyết không, giả thuyết đối 5](#_Toc209801332)

[2.2. Khoảng tin cậy 5](#_Toc209801333)

[2.2.1. Định nghĩa 5](#_Toc209801334)

[2.2.2. Ước lượng khoảng tin cậy 5](#_Toc209801335)

[2.2.3. Bài toán ước lượng khoảng tin cậy 5](#_Toc209801336)

[2.3. Chuẩn hóa dữ liệu thu được bằng z-score 6](#_Toc209801337)

[2.3.1. Định nghĩa 6](#_Toc209801338)

[2.3.2. Ý nghĩa 6](#_Toc209801339)

[2.3.3. Mục đích chuẩn hóa bằng z\_score 7](#_Toc209801340)

[2.4. Bài toán kiểm định thống kê 7](#_Toc209801344)

[3. Thực nghiệm và kết quả 9](#_Toc209801345)

[3.1. Khảo sát để thu thập dữ liệu mẫu 9](#_Toc209801346)

[3.2. Code kiểm định giả thuyết thống kê vừa nêu cho kỳ vọng với mức ý nghĩa 5%......................1](#_Toc209801347)1

[3.3. Giải thích chương trình hoạt động 12](#_Toc209801351)

[3.3.1. Chuẩn bị và làm sạch dữ liệu: 1](#_Toc209801352)2

[3.3.2. Thống kê Mô tả Cơ bản 1](#_Toc209801359)3

[3.3.3. Chuẩn hóa Dữ liệu (Z-score) 1](#_Toc209801368)3

[3.3.4. Suy luận Thống kê (Ước lượng Khoảng Tin cậy cho Tỷ lệ) 1](#_Toc209801369)4

[3.4. Kết quả chương trình 1](#_Toc209801370)4

Đề tài: Theo khảo về dinh dưỡng và sức khỏe năm 2024-2025 xác định mức khuyến nghị tiêu chuẩn về lượng nước uống hàng ngày là 2 lít/ngày. Tuy nhiên nhóm em nghi ngờ con số này đã cải thiện so với trước đây. Bởi vậy nhóm em tiến hành khảo sát 56 sinh viên và tính được trung bình mẫu là X điểm và phương sai hiệu chỉnh là Y. Nhóm em tiến hành các thí nghiệm để kết luận nghi ngờ trên với mức ý nghĩa là 5%

# Giới thiệu bài toán và thư viện sử dụng

## 1.1. Giới thiệu bài toán

Trong các nghiên cứu về dinh dưỡng và sức khỏe, một cuộc khảo sát trước đây (hoặc một khuyến nghị y tế phổ biến) thường xác định **mức khuyến nghị tiêu chuẩn** về lượng nước uống hàng ngày là **1.8 lít/ngày**. Đây được xem như **mốc tham chiếu (μ0​=1.8 lít)** để đánh giá thói quen uống nước.

Tuy nhiên, nhóm nghiên cứu đặt ra nghi vấn rằng **lượng nước uống trung bình thực tế hiện nay của người trong khoa đã có sự thay đổi**, có thể **thấp hơn** hoặc **khác biệt** so với mốc 1.8 lít do đặc thù công việc và môi trường. Để kiểm chứng điều này, nhóm đã tiến hành một **khảo sát mới** trên một mẫu gồm 100 người trong khoa, thu thập dữ liệu về lượng nước uống hàng ngày (đơn vị: lít) và tính toán các thông số mẫu: kích thước mẫu, trung bình mẫu và phương sai mẫu.

Mục tiêu của nghiên cứu là sử dụng các phương pháp thống kê để:

* + - * **Thu thập và chuẩn hóa dữ liệu** (ví dụ: bằng Z-score) nhằm phân tích phân phối mẫu.
      * **Ước lượng kỳ vọng** (μ) lượng nước uống trung bình thông qua **khoảng tin cậy 95%**.
      * **Ước lượng khoảng tin cậy 95% cho tỷ lệ** người uống ≥1.8 lít nước/ngày.
      * Thực hiện **kiểm định giả thuyết thống kê** với mức ý nghĩa (α) là **5%** để xác định liệu lượng nước uống trung bình hiện tại **có khác biệt** so với mức 1.8 lít/ngày hay không.

Với cách tiếp cận này, nghiên cứu sẽ trả lời câu hỏi:

*“Lượng nước uống trung bình hàng ngày của người trong khoa có* ***khác biệt*** *so với mức khuyến nghị 1.8 lít hay không?”*

## 1.2. Giới thiệu thư viện sử dụng

### 

### **1.2.1. Mục tiêu bài toán**

* + - * Dữ liệu: Cột điểm tổng từ file Excel
      * Cần:

+ Chuẩn hóa dữ liệu

+ Ước lượng khoảng tin cậy

+ Kiểm định giả thuyết 1 mẫu cho trung bình

### **1.2.2. Bộ thư viện sử dụng**

1. Pandas – Đọc và xử lý dữ liệu bảng

* Lý do: Thao tác trực tiếp với file dữ liệu Excel (Book1.xlsx), làm sạch dữ liệu thô (chuẩn hóa định dạng số, trích xuất số), và tính toán các thống kê mô tả cơ bản của mẫu.
* Hàm và thuộc tính chính:

pd.read\_excel(), Series.str.replace(), Series.str.extract(), Series.astype(float), Series.mean(), Series.median(), Series.var(), Series.std(), Series.dropna().

1. Numpy – Tính toán mảng số học

* Lý do: Cung cấp các hàm toán học hiệu suất cao, đặc biệt là căn bậc hai, cần thiết cho việc tính toán thống kê kiểm định và sai số chuẩn.
* Hàm và thuộc tính chính:

np.sqrt(), np.isnan(), np.zeros().

1. Scikit-learn – chuẩn hoá dữ liệu

* Lý do: Cung cấp công cụ chuẩn hóa chuyên nghiệp, mặc dù bạn đã định nghĩa lại hàm Z-score thủ công, thư viện này là chuẩn mực để áp dụng các phép biến đổi dữ liệu.
* Hàm và thuộc tính chính:

StandardScaler().

# Cơ sở lý thuyết

## 2.1. Giả thuyết không, giả thuyết đối

- Giả sử cần nghiên cứu tham số *θ* của biến ngẫu nhiên X và có cơ sở nào đó để nêu lên giả thuyết *θ* = *θ*0 . Giả thuyết này ký hiệu là *H0* , còn gọi là giả thuyết cần kiểm định.

- Mệnh đề đối lập với giả thuyết *H*0 , ký hiệu là *H*1​, còn gọi là đối thuyết.   
Dạng tổng quát nhất của *H1* là *θ* ≠ *θ*0  .  
Trong nhiều trường hợp giả thuyết đối được phát biểu cụ thể là

H1 : *θ* > *θ*0 hoặc H1 : *θ* < *θ*0

## 2.2. Khoảng tin cậy

**2.2.1. Định nghĩa**

Giả sử chưa biết đặc trưng θ nào đó của biến ngẫu nhiên X. Ước lượng khoảng của θ là chỉ ra một khoảng số (g1, g2) nào đó chứa θ, tức là có thể ước lượng g1

< θ < g2.

**2.2.2. Ước lượng khoảng tin cậy**

Với α > 0 khá bé, ta tìm được P(g1 < θ < g2) = 1 − α = β thì ta kết luận: với độ tin cậy 1 − α = β, tham số θ nằm trong khoảng (g1, g2). Khi đó:

1. (g1, g2) được gọi là khoảng tin cậy của θ với độ tin cậy β = 1 − α.
2. 1 − α = β được gọi là độ tin cậy của ước lượng.
3. l = g2 − g1 được gọi là độ dài khoảng tin cậy.

### 

### **2.2.3. Bài toán ước lượng khoảng tin cậy**

Giả sử biến ngẫu nhiên X tuân theo luật phân phối chuẩn *N(μ,o²)* với kỳ vọng *E(X)=μ* chưa biết.

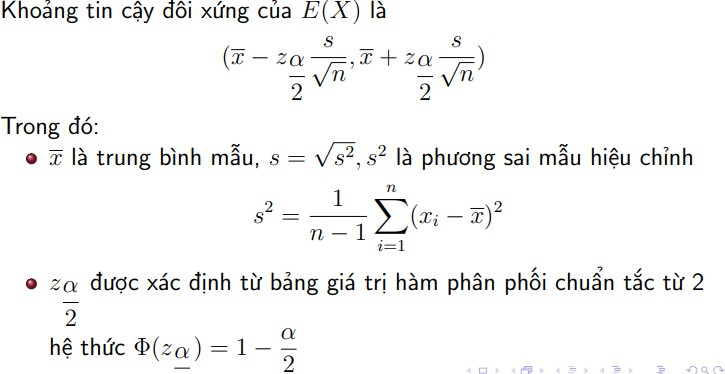
Hãy ước lượng *E(X)*. Bài toán được xét trong 3 trường hợp:

- Trường hợp 1:đã biết phương sai V(X)=σ²

- Trường hợp 2:chưa biết phương sai V(X)=σ²và kích thướcmẫu n>30

- Trường hợp 3:chưa biết phương sai V(X)=σ²và kích thướcmẫu n<30

số lượng sinh viên mà nhóm đã khảo sát là 100 người và chưa biết phương sai V(X) thì bài toán này sẽ rơi vào **trường hợp 2: Chưa biết phương sai V(X) và khước thước mẫu n > 30**

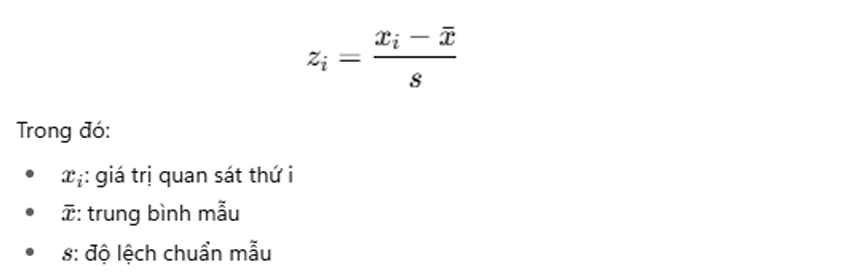
****

## 2.3. Chuẩn hóa dữ liệu thu được bằng z-score

### 

### **2.3.1. Định nghĩa**

**Chuẩn hoá Z-score (hay chuẩn hoá theo độ lệch chuẩn)** là kỹ thuật biến đổi dữ liệu gốc xi thành dạng chuẩn hoá zi. Chuẩn hoá bằng Z-score là bước trung gian quan trọng giúp chuẩn bị và làm sạch dữ liệu trước khi tiến hành phân tích thống kê.



### **2.3.2. Ý nghĩa**

**Z-score = 0**: giá trị đúng bằng trung bình.

**Z-score > 0**: giá trị lớn hơn trung bình, càng lớn → càng xa trung bình.

**Z-score < 0**: giá trị nhỏ hơn trung bình.

**|Z-score| ≥ 2 hoặc 3**: thường được coi là ngoại lệ (*outlier*).

## **2.3.3. Mục đích chuẩn hóa bằng z\_score**

## Đưa dữ liệu về cùng thang đo

* Nếu các biến có thang đo khác nhau (ví dụ điểm Toán 0–100, điểm Văn 0–10) → chuẩn hoá để so sánh công bằng.

## Phát hiện ngoại lệ (Outlier Detection)

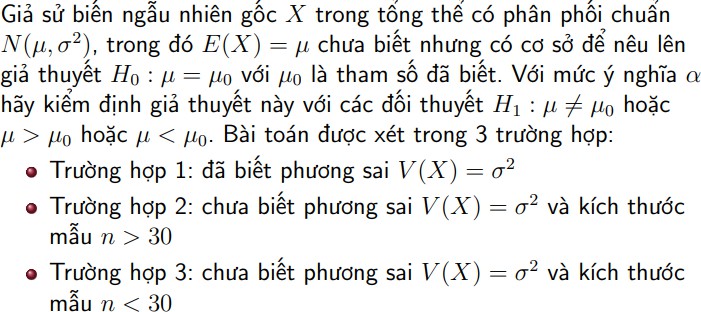
* Các giá trị cực trị sẽ có |Z| rất lớn, dễ dàng nhận ra.

## Giúp dữ liệu gần với phân phối chuẩn

* Sau chuẩn hoá, dữ liệu thường gần với N(0,1)N(0,1)N(0,1), thuận lợi cho các kiểm định thống kê (z-qs, t-qs…).

## 

## 2.4. Bài toán kiểm định thống kê

****

P-value là mức ý nghĩa nhỏ nhất có thể dẫn đến việc bác bỏ giả thuyết H0 với dữ liệu đã cho.

số lượng thí sinh mà nhóm đã khảo sát là 96 người thì bài toán này sẽ rơi vào

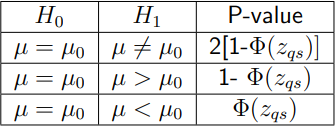
**trường hợp 2: Chưa biết phương sai V(X) và khước thước mẫu n > 30**

**Bước 1:** Đặt giả thuyết H0 và đối thuyết H1

**Bước 2:** Xác định giá trị quan sát



**Bước 3:** Xác định P-value

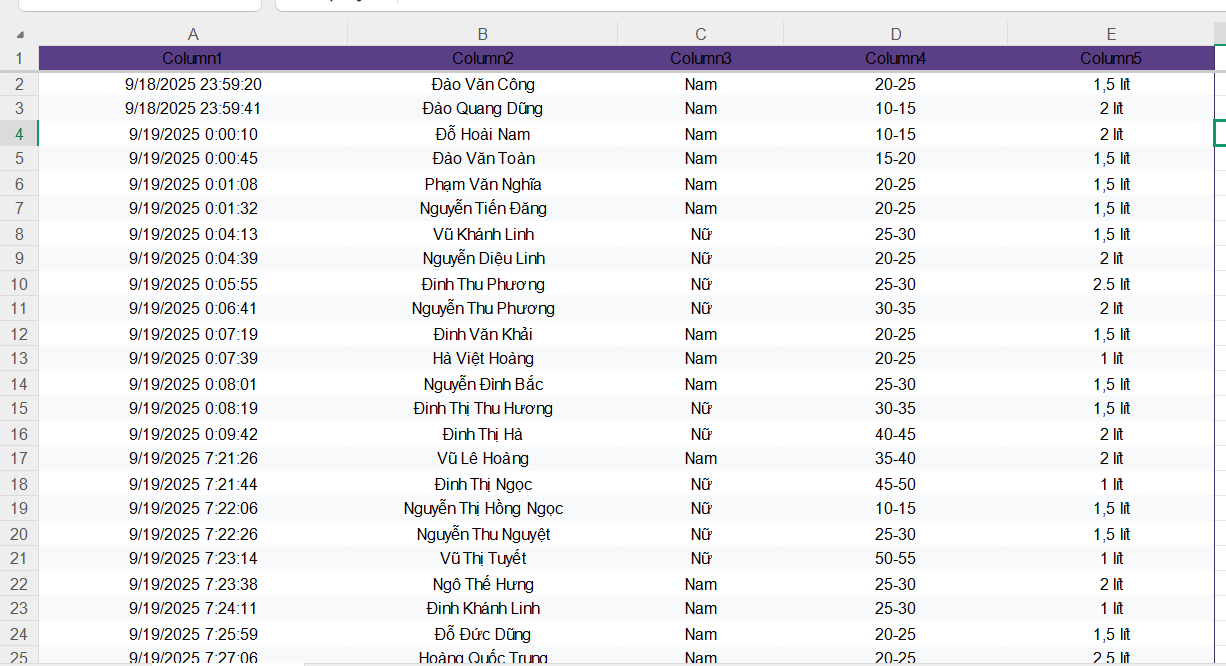
****

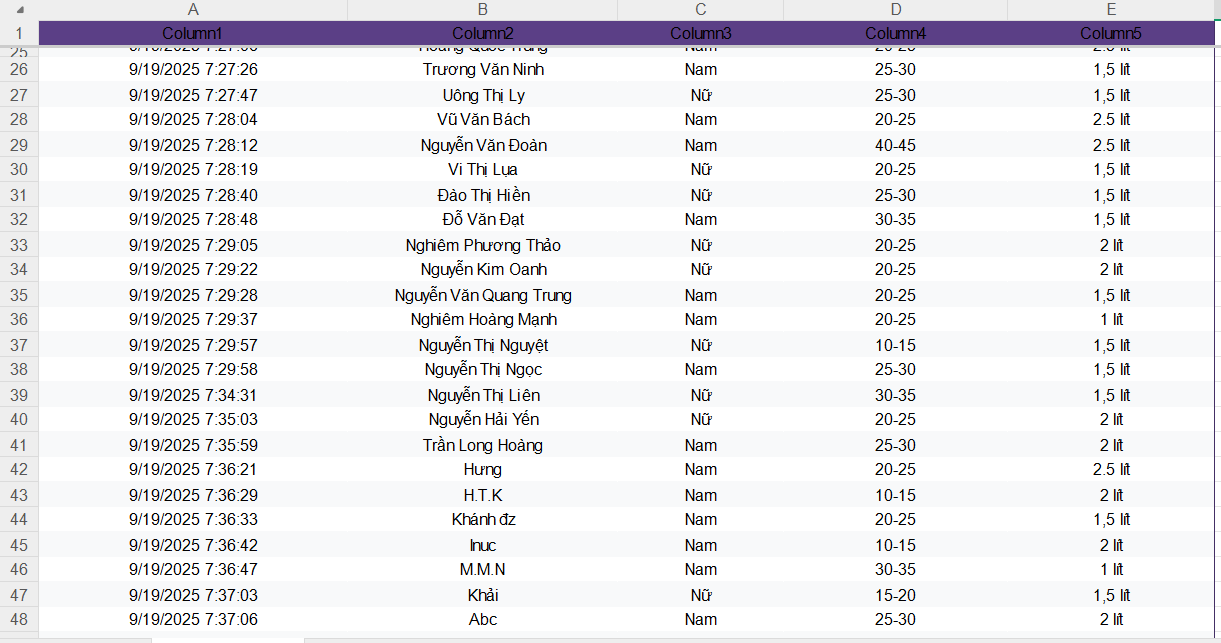
**Bước 4:** Kết luận

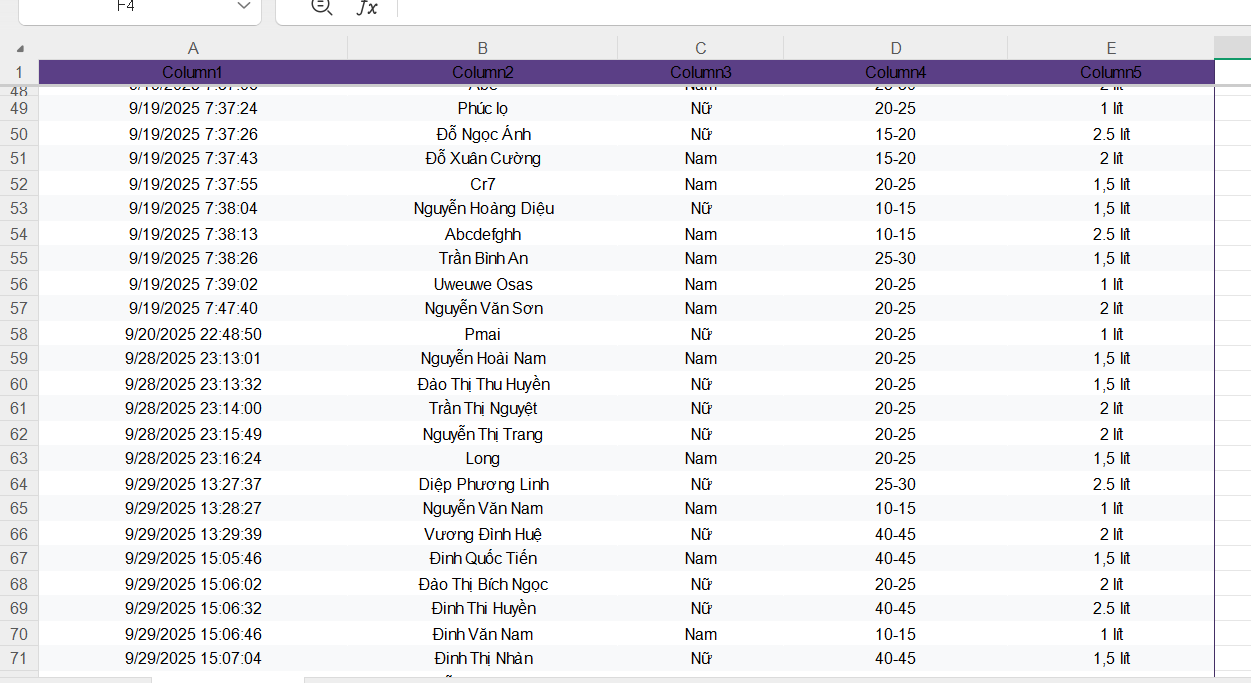
****

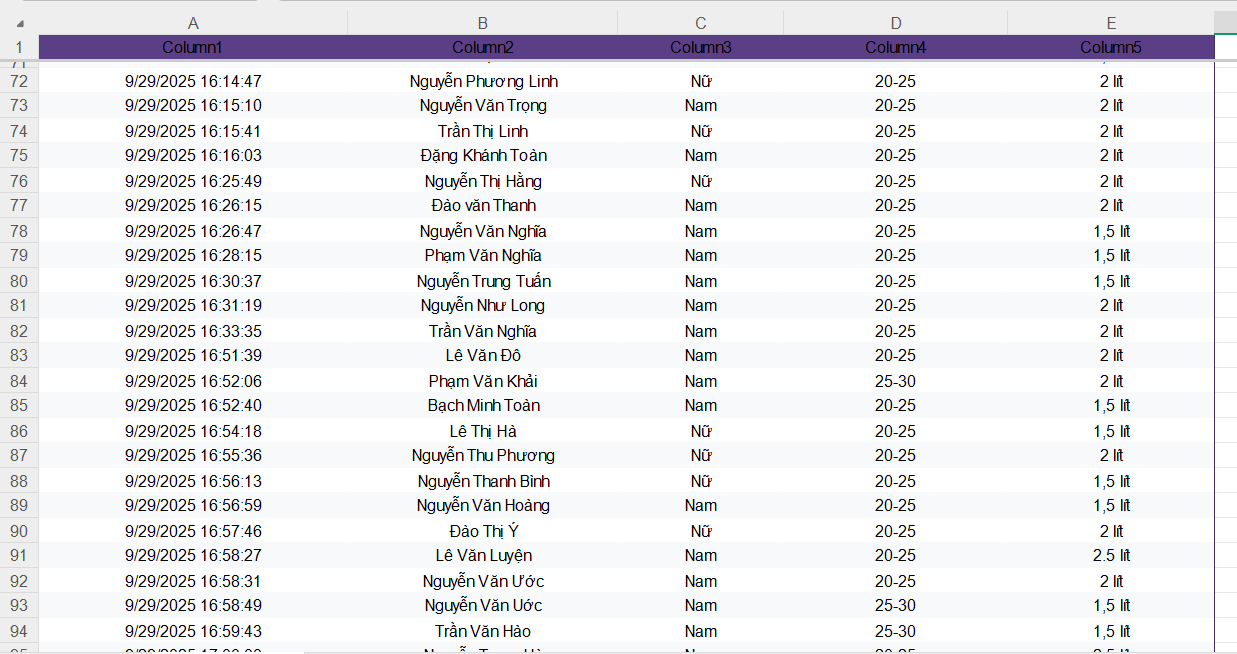
# Thực nghiệm và kết quả

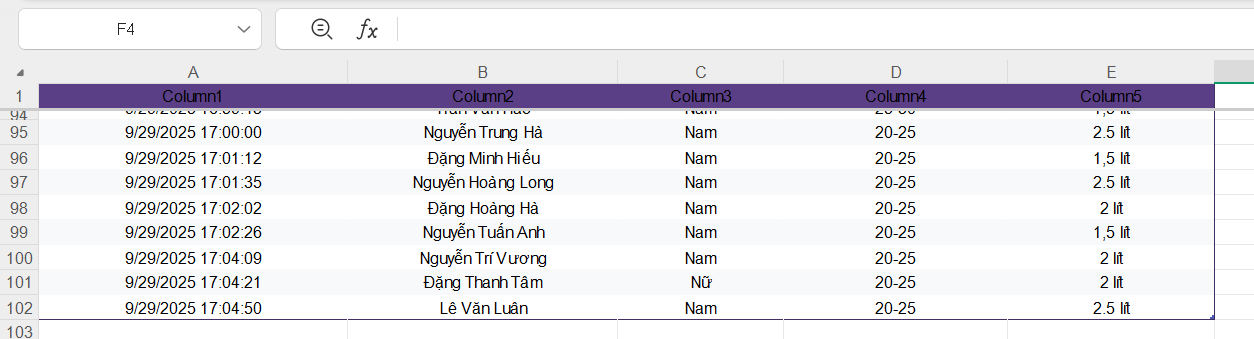
## 3.1. Khảo sát để thu thập dữ liệu mẫu



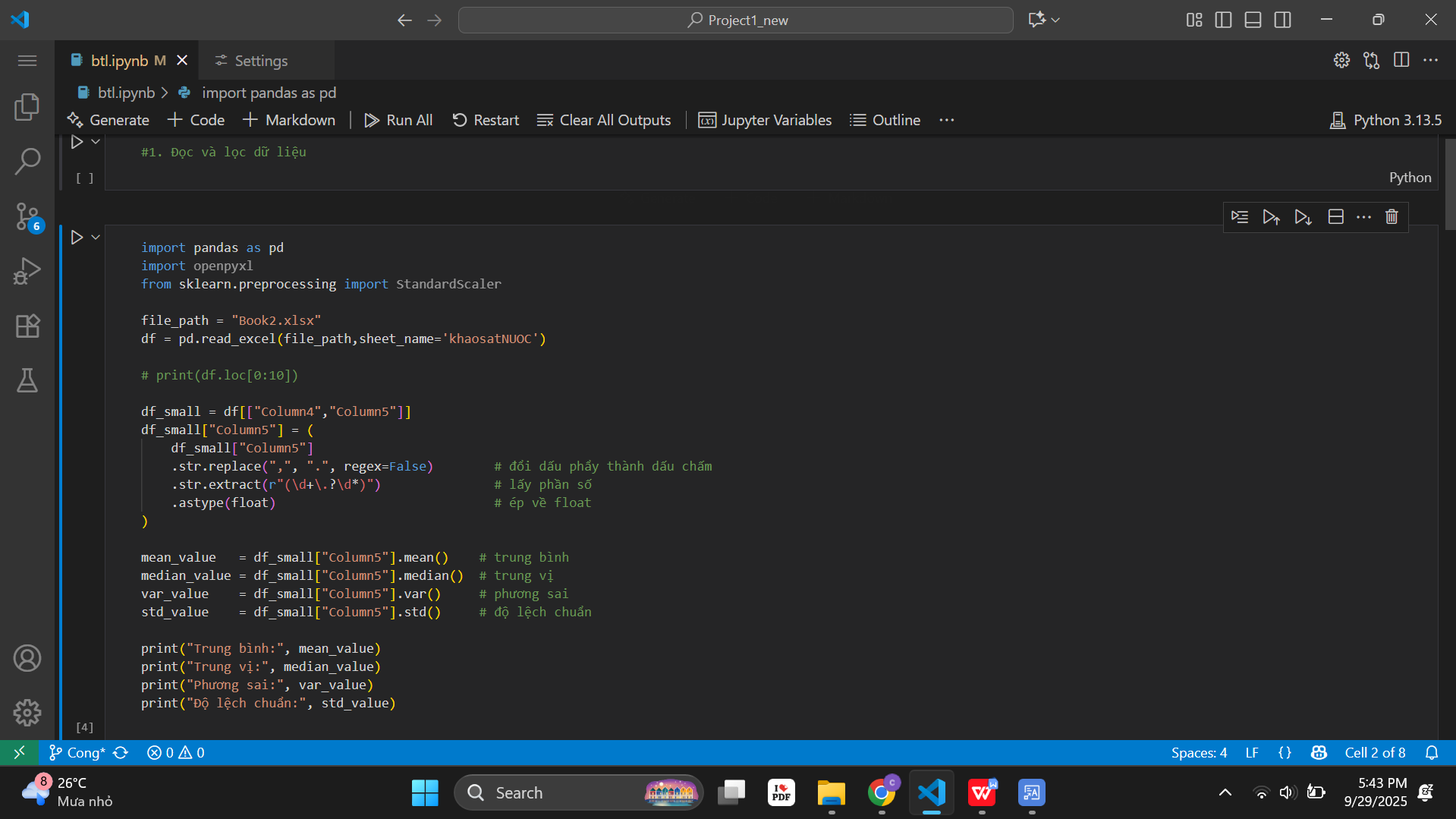


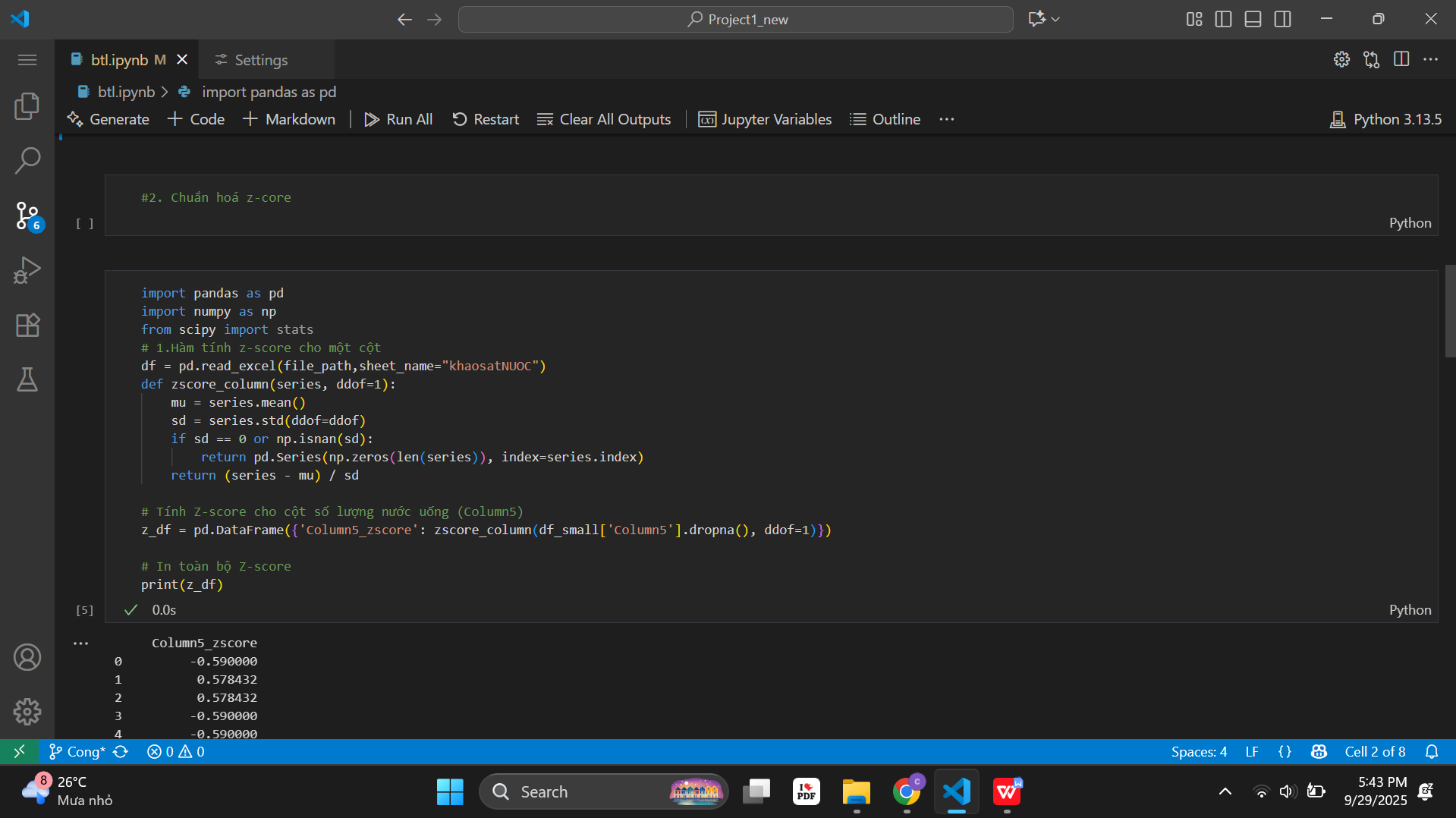


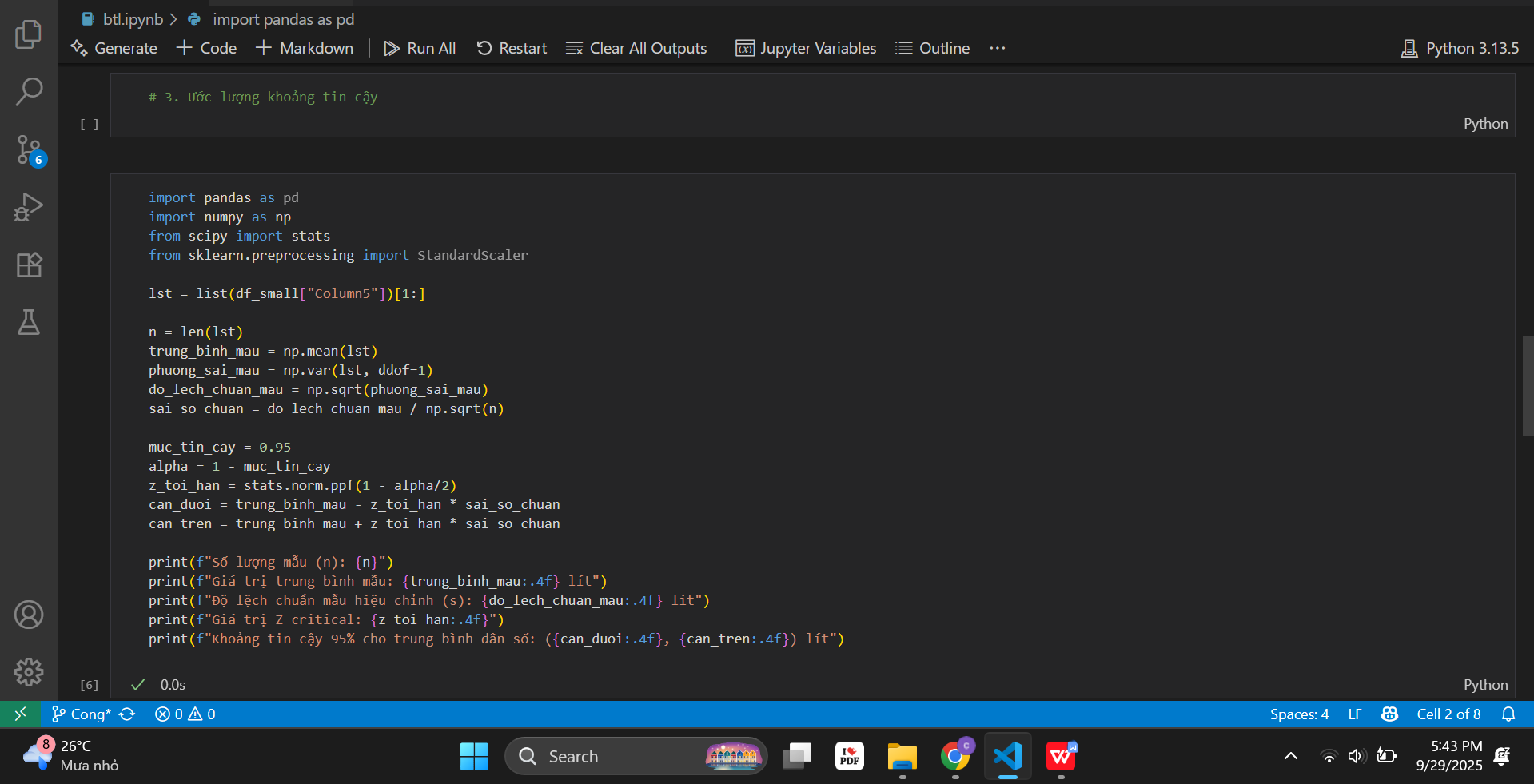




## 3.2. Chương trình code kiểm định giả thuyết thống kê vừa nêu cho kỳ vọng với mức ý nghĩa 5%







## 3.3. Giải thích chương trình hoạt động

### 

### **3.3.1. Chuẩn bị và làm sạch dữ liệu:**

## pd.read\_excel(...)

Đọc dữ liệu từ file "Book1.xlsx" tại sheet "khaosatNUOC" vào DataFrame (df)

## df\_small = df[["Column4","Column5"]]

Tạo DataFrame mới (df\_small) chỉ chứa hai cột quan tâm: **Độ tuổi** (Column4) và **Số lượng nước uống** (Column5).

## .str.replace(",", ".", regex=False)

Xử lý dữ liệu thô: Chuyển đổi dấu phẩy (,) thành dấu chấm (.) trong cột Column5 để chuẩn bị cho việc ép kiểu số.

## .str.extract(r"(\d+\.?\d\*)")

Lọc bỏ các ký tự không phải số, chỉ giữ lại phần số (bao gồm cả số thập phân) từ chuỗi, nhằm loại bỏ các câu trả lời không chuẩn (ví dụ: "1.8 lít/ngày")

## astype(float)

Ép kiểu dữ liệu của cột Column5 về kiểu số thực (float), cho phép thực hiện các phép tính thống kê.

## return Summary(...)

Trả về một đối tượng Summary (dataclass bạn định nghĩa ở trên) chứa:

* + - n: số quan sát
    - mean: trung bình mẫu
    - var: phương sai mẫu
    - std: độ lệch chuẩn mẫu

### 

### **3.3.2. Thống kê Mô tả Cơ bản**

## df\_small["Column5"].mean()

## Giá trị trung tâm, cho biết lượng nước uống trung bình của mẫu khảo sát.

## df\_small["Column5"].median()

## Giá trị ở giữa của tập dữ liệu đã sắp xếp, giúp nhận biết mức độ lệch của phân phối (so sánh với Trung bình).

## df\_small["Column5"].var()

## Đo lường mức độ phân tán của dữ liệu so với giá trị trung bình (bình phương của độ lệch chuẩn).

## df\_small["Column5"].std()

## Đo lường mức độ phân tán theo đơn vị gốc (lít), là thước đo phổ biến về sự biến động của lượng nước uống giữa các sinh viên.

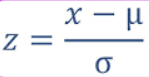
### 

### **3.3.3. Chuẩn hóa Dữ liệu (Z-score)**

1. **def zscore\_column(...)**

Định nghĩa hàm tùy chỉnh để tính **Z-score** (điểm chuẩn) cho một cột.

1. **(series - mu) / sd**

Công thức Z-score: 

trong đó μ là trung bình mẫu và σ là độ lệch chuẩn mẫu.

1. **z\_df = ...**

Z-score cho biết mỗi sinh viên uống lượng nước ít hơn hay nhiều hơn trung bình bao nhiêu độ lệch chuẩn

### **3.3.4. Suy luận Thống kê (Ước lượng Khoảng Tin cậy cho Tỷ lệ)**

1. **p\_hat = (df\_small["Column5"] >= 2).mean()**

Tỷ lệ sinh viên trong mẫu uống ≥2 lít nước mỗi ngày.

1. **n = 100**

Kích thước mẫu n

1. **z\_score = 1.96**

Z=1.96 được sử dụng cho mức tin cậy **95%** (α=0.05).

1. **margin\_of\_error = z\_score \* np.sqrt(p\_hat \* (1 - p\_hat) / n)**

Sai số biên (ME): 

Khoảng cách tối đa mà tỷ lệ mẫu (p) có thể sai lệch so với tỷ lệ tổng thể (p) với mức tin cậy 95%.

1. **ci\_lower = p\_hat - margin\_of\_error; ci\_upper = p\_hat + margin\_of\_error**

Khoảng tin cậy (CI): 

Khoảng giá trị mà tỷ lệ sinh viên uống ≥2 lít nước **thực tế của toàn bộ tổng thể** có 95% khả năng nằm trong đó.

## 3.4. Kết quả chương trình

Có **bằng chứng thống kê** cho thấy **trung bình điểm của năm khảo sát hiện tại cao hơn 20.11** (so với mốc trung bình năm 2019–2020).

→ Như vậy, nghi ngờ "điểm trung bình đã cải thiện" là **có cơ sở**.

