**LAB 1**

***Bài 1a.*** Dùng công cụtrực quan hóa dữ liệu (Data Visualization) và Thống kê mô tả (Descriptive Statistics) thực hiện yêu cầu sau:

Theo số liệu về kinh tế Việt Nam tại website:

<http://finance.vietstock.vn/du-lieu-vi-mo/Default.htm>

Với tập tin dữ liệu:

*GDP binh quan cua VN*

1. Sử dụng ***MS Excel, ngôn ngữ R*** và Ngôn ngữ Python với số liệu về GDP bình quân của VN, hãy tính và giải thích ý nghĩa của các giá trị: Count, Min, Max, Mean, Median, Mode, Quantile, Range, Mode, Variance, Standard Deviation, Coefficient of Deviation, Skewness, Kurtosis.

2. Dùng chức năng Visualization: Histogram, Box Plot hãy hiển thị một số giá trị trên

Bài làm:

**1. Ý nghĩa của các giá trị:**

***Số lượng (Count):*** là số lượng giá trị đang quan sát

***Giá trị nhỏ nhất (Min):*** là giá trị nhỏ nhất của giá trị đang quan sát

***Giá trị nhỏ lớn nhất (Max):*** là giá trị lớn nhất của giá trị đang quan sát

***Giá trị trung bình (Mean)***: Là giá trị trung bình số học của một biến, được tính bằng tổng các giá trị quan sát chia cho số quan sát. Đây là dạng công cụ thường được dùng cho dạng đo khoảng cách và tỷ lệ. Giá trị trung bình có đặc điểm là chịu sự tác động của các giá trị ở mỗi quan sát, do đó đây là thang đo nhạy cảm nhất đối với sự thay đổi của các giá trị quan sát.

***Mode:*** Là giá trị có tần suất xuất hiện lớn nhất của một tập hợp các số đo, dạng này thường được dùng đối với dạng dữ liệu thang biểu danh. Giống như trung vị, mode không bị ảnh hưởng bởi giá trị đầu mút của dãy phân phối.

***Trung vị (Median):*** Là số nằm giữa (nếu lượng quan sát là số lẻ) hoặc là giá trị trung bình của hai quan sát nằm giữa (nếu số lượng quan sát là số chẳn) của một dãy quan sát được xắp xếp theo thứ tự từ nhỏ đến lớn. Đây là dạng công cụ thống kê thường được dùng để đo lường mức độ tập trung của dạng dữ liệu thang đo thứ tự, nó có đặc điểm là không bị ảnh hưởng của các giá trị đầu mút của dãy phân phối, do đó rất thích hợp để phân tích đối với dữ liệu có sự chênh lệch lớn về giá trị ở hay đầu mút của dãy phân phối.

***Phương sai (Variance)****:* Dùng để đo lường mức độ phân tán của một tập các giá trị quan sát xung quanh giá trị trung bình của tập quan sát đó. Phương sai bằng trung bình các bình phương sai lệch giữa các giá trị quan sát đối với giá trị trung bình của các quan sát đó. Công thức tính phương sai:



***Độ lệch chuẩn (SD: Standard deviation):*** Một công cụ khác dùng để đo lường độ phân tán của dữ liệu xung quanh giá trị trung bình của nó. Độ lệch chuẩn chính bằng căn bậc hai của phương sai. Vì phương sai là trung bình của các bình phương sai lệch của các giá trị quan sát từ giá trị trung bình, việc khảo sát phương sai thường cho các giá trị rất lớn, do đó sử dụng phương sai sẽ gặp khó khăn trong việc diễn giải kết quả. Sử dụng độ lệch chuẩn sẽ giúp dễ dàng cho việc diễn giải do các kết quả sai biệt đưa ra sát với dữ liệu gốc hơn. Công thức tính độ lệch chuẩn:



***Khoảng biến thiên (Range):***Là khoảng cách giữa giá trị quan sát nhỏ nhất đến giá trị quan sát lớn nhất.  
 R = Xmax – Xmin

***Hệ số đối xứng Skewness:*** cho biết dạng phân phối của các giá trị quan sát có thể được sử dụng để kiểm nghiệm tính phân phối chuẩn. Một giá trị Skewness dương cho biết nhánh của phân phối này lệch qua bên phải và ngược lại một trị âm chỉ ra nhánh của phân phối này dài qua bên trái. Công thức tính hệ số đối xứng:

A picture containing text, clock

Description automatically generated

***Hệ số tập trung Kurtosis:*** Dùng để đánh giá đỉnh của đường cong quan sát với dạng đường cong phân phối chuẩn. Công thức tính hệ số tập trung:

Diagram

Description automatically generated

**2. Với tập tin “*GDP binh quan cua VN”* tính các giá trị trên:**

***Ngôn ngữ R:***

Text

Description automatically generated

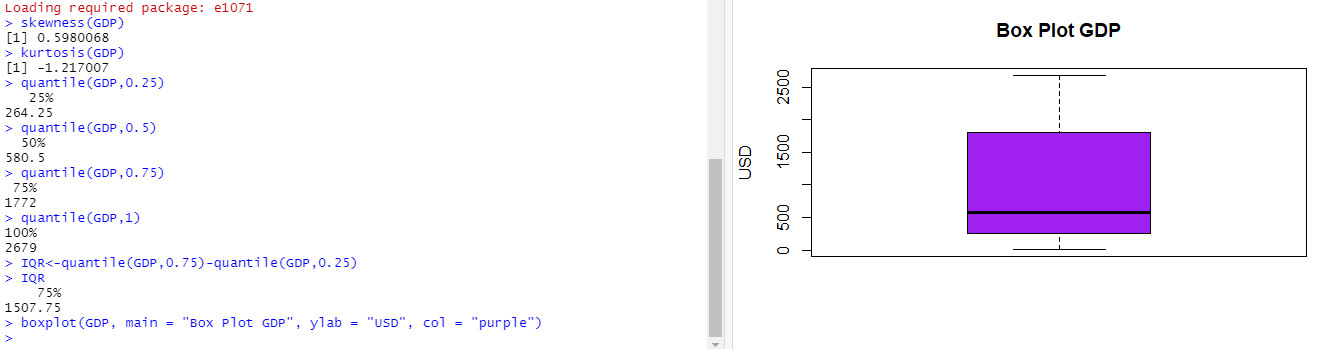
*Hình 1: Tiến hành đọc file GDP.csv*

***Tính các giá trị theo yêu cầu: hình 2,3,4***

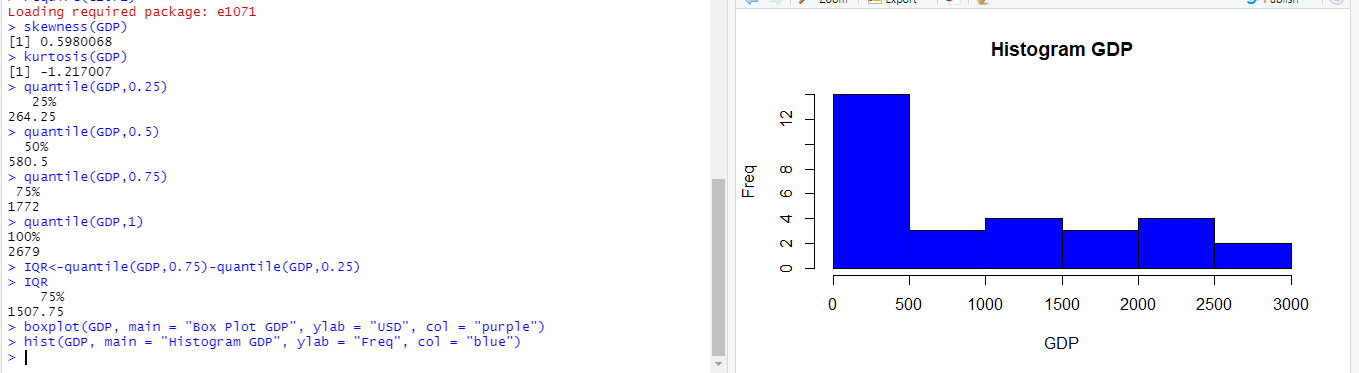
Ảnh có chứa văn bản

Mô tả được tạo tự động

*Hình 2:*



*Hình 3:*



*Hình 4:*

***Ngôn ngữ Python***

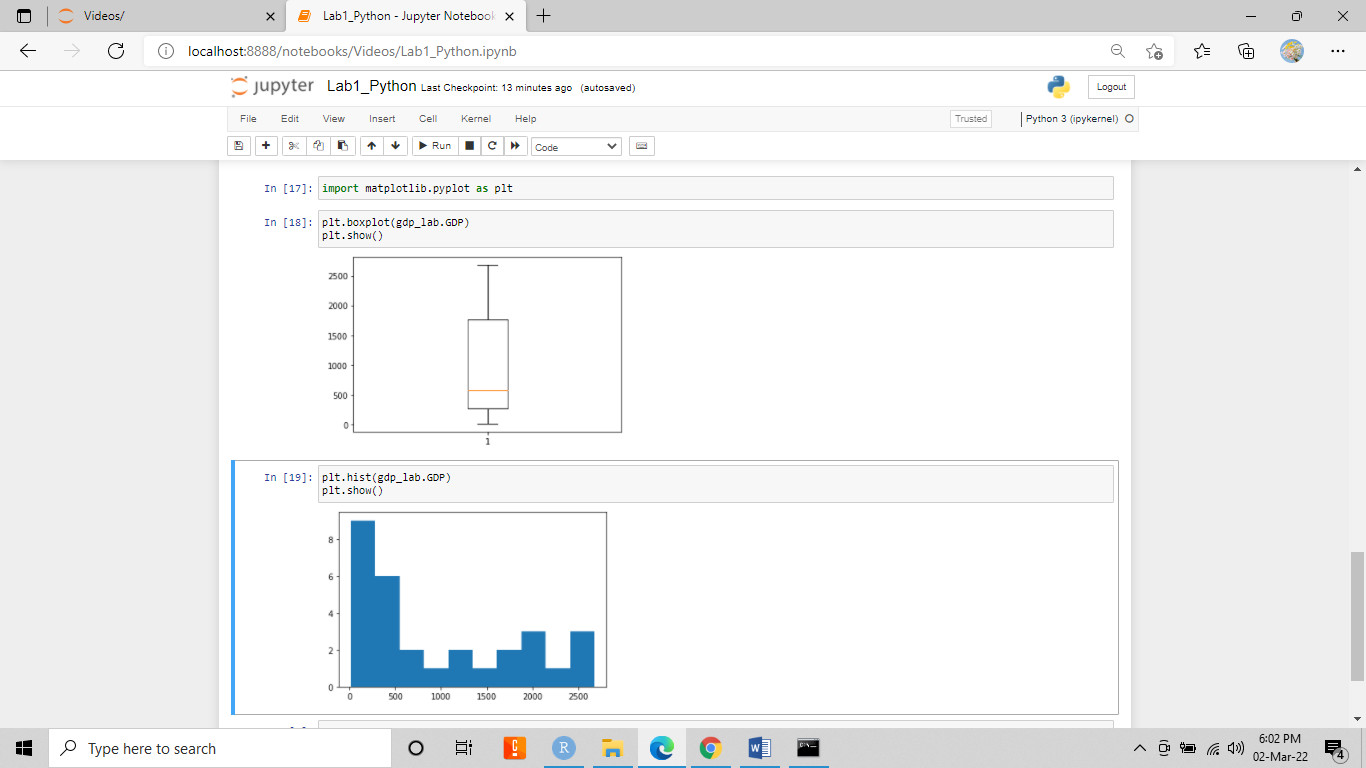
Graphical user interface, text, application, email

Description automatically generated

Ảnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự động

***MS Excel***Ảnh có chứa bàn

Mô tả được tạo tự động

***Bài 1b.***

Thực hiện các yêu cầu tương tự bài 1a với các tập tin dữ liệu:

*Purchase Orders*

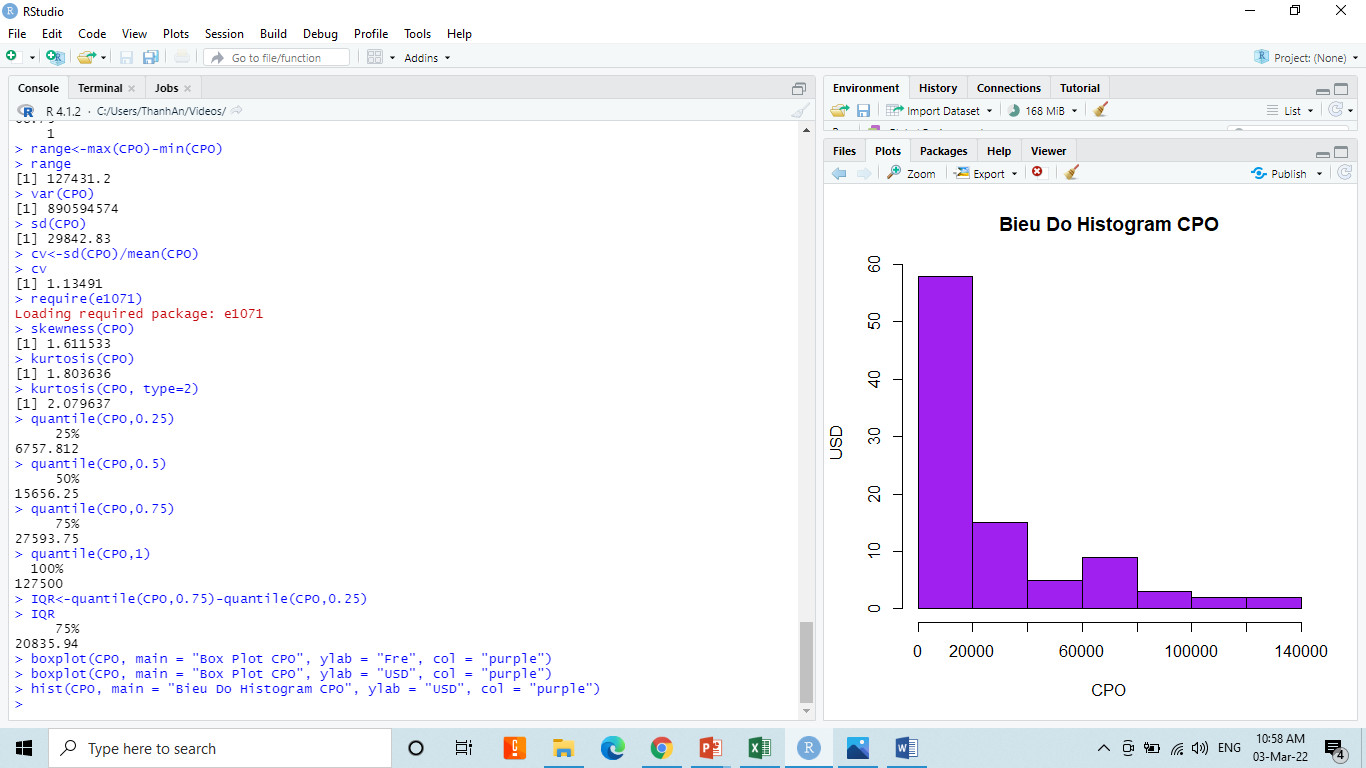
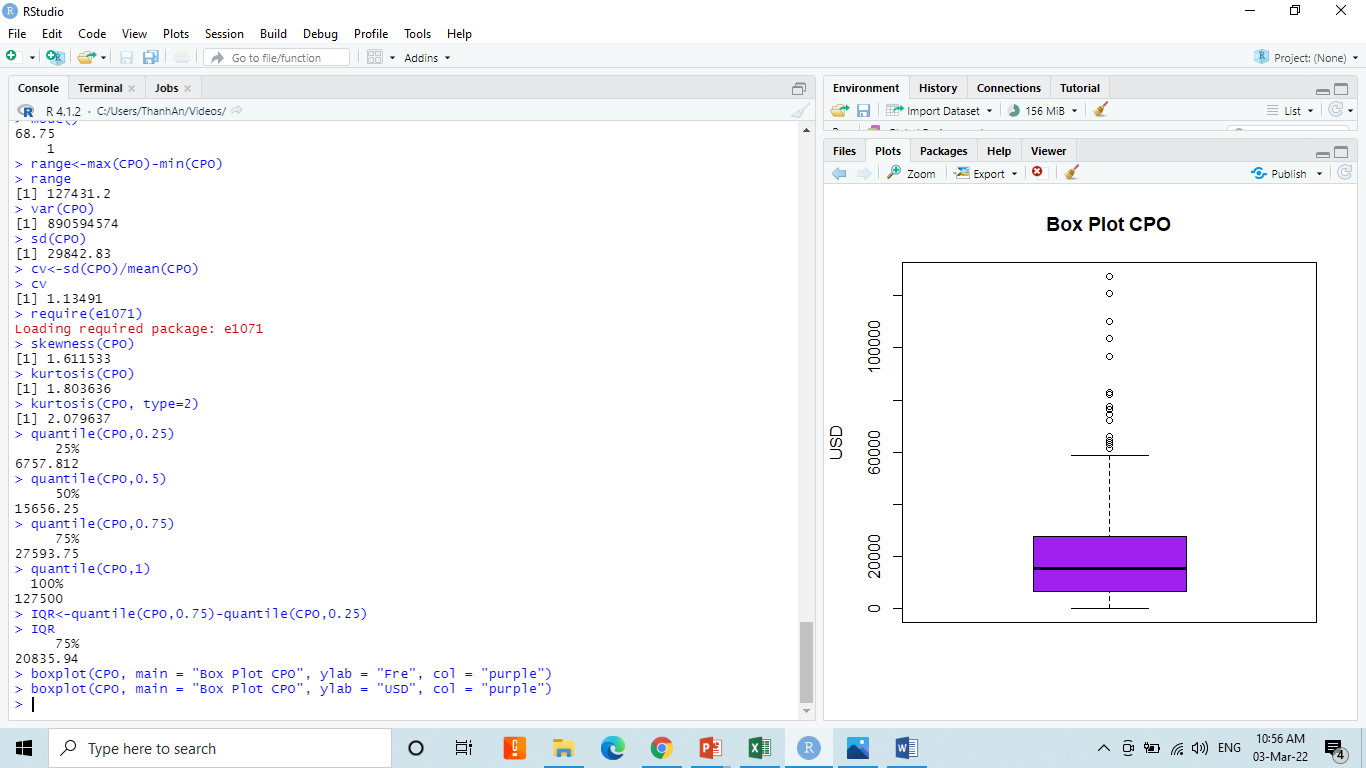
*Computer Repair Times*

***Bài làm:***

**Purchase oders:**

***Ngôn ngữ R:Ảnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự động***

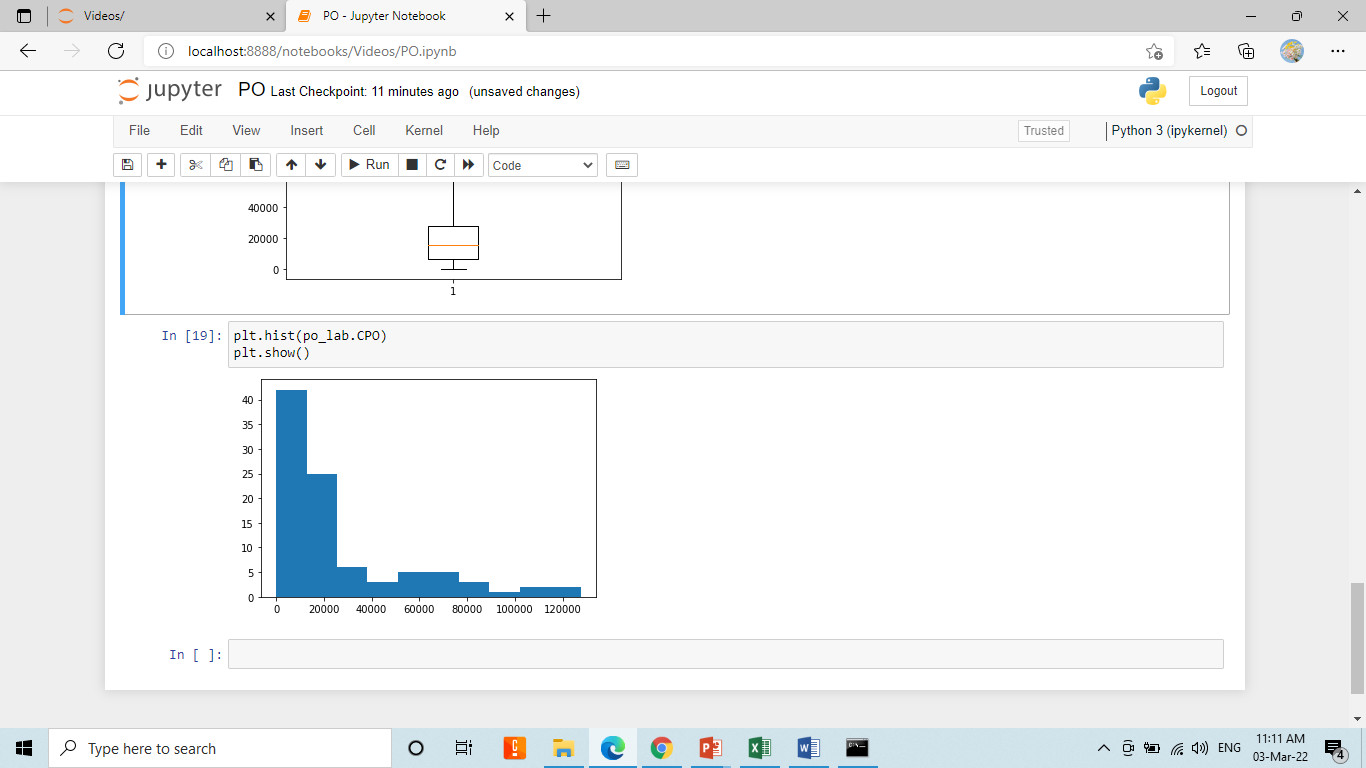
***Ngôn ngữ Python:***

Ảnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự độngẢnh có chứa văn bản

Mô tả được tạo tự động

***MS Excel:***

Ảnh có chứa bàn

Mô tả được tạo tự động

**Computer Repair Time**

***Ngôn ngữ R***

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Chart, histogram

Description automatically generated

***Ngôn ngữ Python***

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

***MS Excel***

Graphical user interface, application, table, Excel

Description automatically generated

**Bài 1C**

1. **Phân biệt**
2. **Null hypothesis và Alternative hypothesis**

* Null hypothesis là phần giả thuyết gốc.
* Alternative hypothesis là phần giả thuyết bù với Null hypothesis

|  |  |  |
| --- | --- | --- |
| Cơ sở | Null hypothesis | Alternative hypothesis |
| Ý nghĩa | Null hypothesis là phần giả thuyết gốc. | Alternative hypothesis là phần giả thuyết mà đơn giản là phần nghịch đảo với Null hypothesis. |
| Ký hiệu | H0 | H1 |
| Nó là gì? | Là những gì mà nhà nghiên cứu cố gắng bác bỏ | Là những gì mà nhà nghiên cứu cố gắng chứng minh. |
| Chấp nhận | Không thay đổi ý kiến ​​hay hành động | Thay đổi ý kiến ​​hoặc hành động |

1. **Type I error và Type II error**

|  |  |  |
| --- | --- | --- |
| Cơ sở | Type I error | Type II error |
| Ý nghĩa | Type I error đề cập đến việc không chấp nhận giả thuyết mà lẽ ra phải được chấp nhận. | Type II error là sự chấp nhận giả thuyết mà lẽ ra phải bị bác bỏ. |
| Nó là gì? | Đó là sự bác bỏ giả thuyết không đúng sự thật. | Đó là sự chấp nhận sai giả thuyết vô hiệu. |
| Kí hiệu | Chữ cái Hy Lạp 'α' | Chữ cái Hy Lạp 'β' |

1. **Giá trị: α và p-value**

* Điểm khác nhau là, α được chọn trước, còn p-value thì do tính toán mà ra. Khi kiểm nghiệm một giả thuyết, chúng ta chọn một mức ý nghĩa (α), và sau đó tính toán số liệu thống kê dựa trên các dữ liệu để lấy được một mức ý nghĩa quan sát được (p-value). Cuối cùng, bạn so sánh 2 mức này (α và p-value) để đưa ra quyết định bác bỏ giả thuyết hoặc không bác bỏ giả thuyết.
* α tương ứng với critical value (được chọn trước).
* p-value tương ứng to the calculated value (được tính toán ra).
* Nếu p-value < α thì bác bỏ giả thuyết, ngược lại nếu p > α thì không bác bỏ giả thuyết.

1. **Áp dụng**
2. **Example 7.5**

* H0: thời gian trả lời trung bình lớn hơn hoặc bằng 25.
* H1: thời gian trả lời trung bình nhỏ hơn 25.

**Tóm tắt**

n = 44

df = n – 1 = 43

α = 0.05

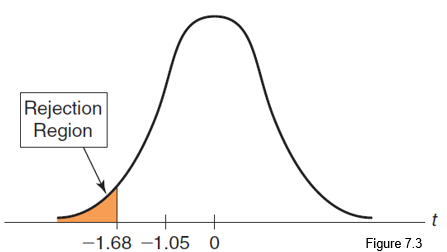
t = -1.05

Tính giá trị quan trọng

Tra bảng phân phối STUDENT

Với df = 43 và α = 0.05

Ta thấy: Critical Value = 1.684 là giao điểm của cột α = 0.05 và df = 40 (≈43)



Ta thấy: t = -1.05 không nằm trong khu vực từ chối

* Không thể bác bỏ H0
* Không thể kết luận rằng thời gian trả lời trung bình là nhỏ hơn 25 phút

Ảnh có chứa bàn

Mô tả được tạo tự động

1. **Example 7.15, 7.16, 7.17**

* H0: giới tính và thương hiệu là độc lập
* H1: giới tính và thương hiệu là phụ thuộc.

Ảnh có chứa bàn

Mô tả được tạo tự động

**Bước 1: Tính giá trị kỳ vọng**

Ảnh có chứa bàn

Mô tả được tạo tự động

Công thức: fe = (Tổng cột i \* Tổng hàng j) / Tổng quan sát được

* Ô B2: fe = (34 \* 37) /100 = 12.58
* Ô C2: fe = (23 \* 37) / 100 = 8.51
* Ô D2: fe = (43 \* 37) / 100 = 15.91
* Ô B3: fe = (34 \* 63) / 100 = 21.42
* Ô C3: fe = (23 \* 63) / 100 = 14.49
* Ô D3: fe = (43 \* 63) / 100 = 27.09

**Bước 2: Tính giá trị X2**

Ảnh có chứa bàn

Mô tả được tạo tự động

Công thức: X2 = ∑

Ô B7: X2 = (9 – 12.58)2 / 12.58 = 1.0188

Ô C7: X2 = (6 – 8.51)2 / 8.51 = 0.7403

Ô D7: X2 = (22 – 15.91)2 / 15.91 = 2.3311

Ô B8: X2 = (25 – 21.42)2 / 21.42 = 0.5983

Ô C8: X2 = (17 – 14.49)2 / 14.49 = 0.4348

Ô D8: X2 = (21 – 27.09)2 / 27.09 = 1.3691

Tổng cộng X2 = 6.4924

**Bước 3: Tính giá trị X2 tới hạn**

df = (r – 1) \* (c – 1) = (2 – 1) \* (3 – 1)= 2

Tra bảng phân phối các giá trị X2 tới hạn

Với α = 0.05 và df = 2

Ảnh có chứa văn bản, biên lai

Mô tả được tạo tự động

* Tra bảng ta thấy giá trị X2 tới hạn = 5.991 là giao điểm của hàng df và cột α = 0.05

Ta thấy: X2 = 6.4942 > 5.991 => Ta bác bỏ H0

**Bước 4: Tính giá trị p-value**

p-value = CHISQ.TEST(B2:D3, B7:D8) = 0.0389

Ảnh có chứa bàn

Mô tả được tạo tự động

Ta thấy p-value = 0.039 < α = 0.05 => Ta bác bỏ H0

* Ta kết luận giới tính và thương hiệu là phụ thuộc

**THÔNG TIN CÁC THÀNH VIÊN**

1. Hoàng Trí Tâm - 19522159
2. Phạm Minh Thắng - 19522215
3. Nguyễn Hữu Thắng - 19522208
4. Đỗ Thị Thanh An - 19520365
5. Trương Thị Kim Thành – 0278

**BẢNG PHÂN CÔNG CÔNG VIỆC**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Họ tên** | **Bài 1a** | **Bài 1b** | **Bài 1c** | | | |
| **Phân biệt** | | | **Áp dụng** |
| **a** | **b** | **c** |
| Hoàng Trí Tâm |  |  | **x** |  |  | **x** |
| Phạm Minh Thắng |  |  |  |  | **x** | **x** |
| Nguyễn Hữu Thắng |  |  |  | **x** |  | **x** |
| Trương Thị Kim Thành | **x** | **x** |  |  |  |  |
| Đỗ Thị Thanh An | **x** | **x** |  |  |  |  |