

IS403

BDA

NGUYEN MINH NHUT

INFORMATION SYSTEM ENGINEERING


1

statistics

plural noun

US  /stəˈtɪs-tɪks/

(infml stats, US /stæts/)

Add to word list 

a collection of numerical facts or measurements, as about people, business conditions, or weather:



Statistics in daily life



2

Utilizing Statistics to Measure Results



Trong chương này chúng ta sẽ học những nội dung:

- Khái niệm dự báo (Forecasting), dự đoán (Predicting)
- Hiệp phương sai (Covariance)
- Độ tương quan (Correlation)
- Mục tiêu và ý nghĩa của phân tích độ tương quan trong các mô hình dự báo, dự đoán.

LAB01 – PART 2

CORRELATION VÀ COVARIANCE



PHẦN 1

KHÁI NIỆM DỰ BÁO & DỰ ĐOÁN ³

Dự báo và dự đoán đều liên quan đến việc ước lượng tương lai, với dự báo thường sử dụng mô hình toán học và dữ liệu thống kê, trong khi dự đoán có thể dựa trên sự hiểu biết và kinh nghiệm cá nhân.



Phần 1. Khái niệm về dự báo và dự đoán

- **Khái niệm dự báo (Forecasting):**

- Dự báo là quá trình **ước lượng giá trị** của một biến trong tương lai dựa trên thông tin **lịch sử và xu hướng hiện tại**.
- **Ví dụ 1:** Dự báo thời tiết, dự báo giá vàng, dự báo nguồn lực lao động Việt Nam,...
- **Chủ yếu dùng mô hình:** Thống kê, máy học, học sâu,....
- **Dự báo có cơ chế lấy quá khứ dự báo cho tương lai.**

Phần 1. Khái niệm về dự báo và dự đoán

• Khái niệm dự đoán (Predicting):

- Dự đoán là quá trình ước lượng giá trị của một biến dựa trên **thông tin hiện tại** mà chưa có thông tin lịch sử.
- **Ví dụ 2:** Dự đoán người đó có bị Covid 19 không thông qua các triệu chứng và danh sách các ca nhiễm hiện tại.
- **Ví dụ 3:** Dự đoán đó là bất thường giao dịch hay không thông qua các dấu hiệu giao dịch trong 1 giờ....
- Cả hai đều sử dụng các kỹ thuật thống kê và mô hình học máy, nhưng mục tiêu của chúng có thể khác nhau tùy thuộc vào ngữ cảnh và mục đích sử dụng.

Phần 1. Khái niệm về dự báo và dự đoán

• Bài tập 1

Điền vào chỗ trống **dự báo** hay **dự đoán** các trường hợp sau:

- (1) giá nhà năm 2024.
- (2) ... số lượng ca nhiễm Covid 19 tháng tới.
- (3) ... dự báo tỉ lệ ung thư.
- (4) ... có khả năng tốt nghiệp trễ hạn.
- (5) ... nguồn lực lao động tích cực của công ty.
- (6) ... tình trạng tỉnh táo của tài xế khi lái xe.



PHẦN 2

ĐỘ TƯƠNG QUAN & HIỆP PHƯƠNG SAI

Độ Tương Quan: Đo mức độ tuyến tính giữa hai biến. Được đo bằng hệ số tương quan Pearson (-1 đến 1). 1 hoặc -1 biểu thị mối quan hệ hoàn toàn tuyến tính, 0 là không có mối quan hệ.

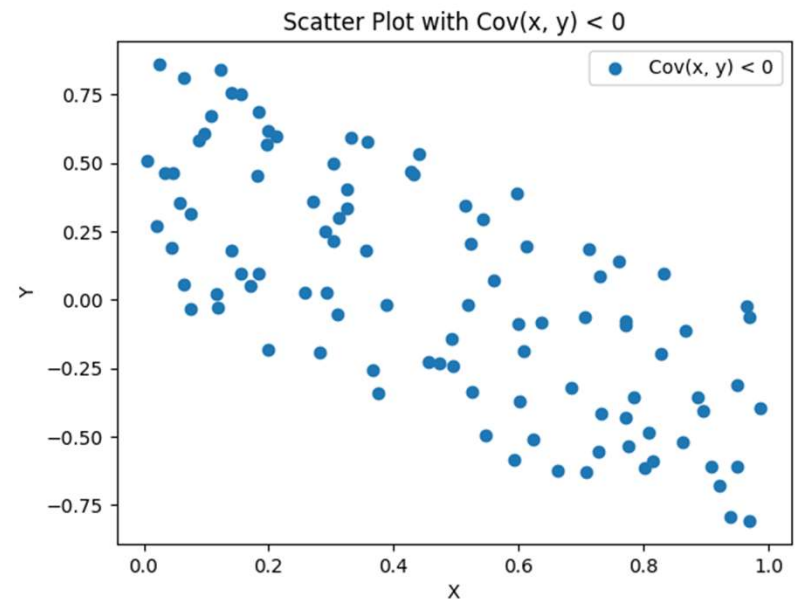
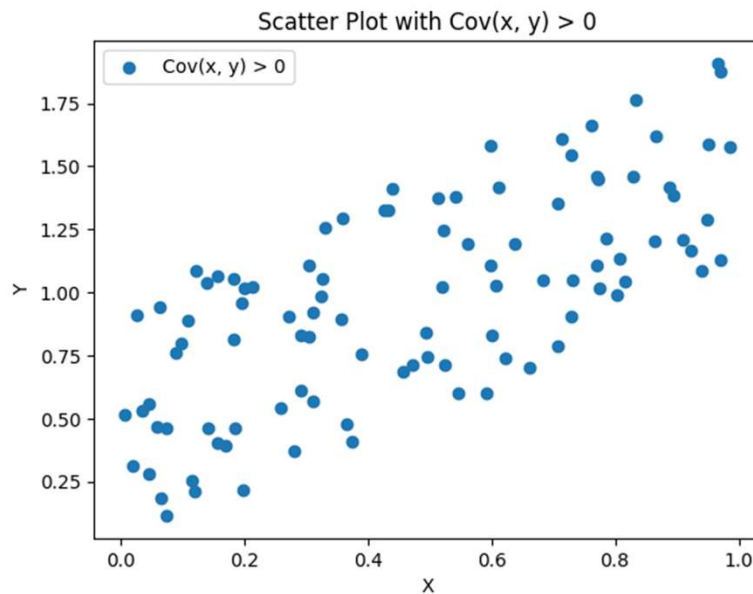
Hiệp Phương Sai: Đo mức độ biến động chung của hai biến. Tính bằng cách kết hợp độ lệch chuẩn và độ tương quan. Giá trị cao thể hiện biến động mạnh.



Phần 2. Hiệp phương sai và Độ tương quan

- **Hiệp phương sai**

- Giả sử biến X và Y muốn đo xem X và Y biến động giống nhau hay không, độ đo để thực hiện được điều này gọi là **Hiệp phương sai**.



Phần 2. Hiệp phương sai và Độ tương quan

- **Hiệp phương sai**

- Hiệp phương sai mẫu của X và Y ký hiệu là **Cov(X,Y)** được xác định bởi công thức sau đây:

Hiệp Phương Sai X và Y:
$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

X	5	6	4	2	4
Y	3	4	2	1	3

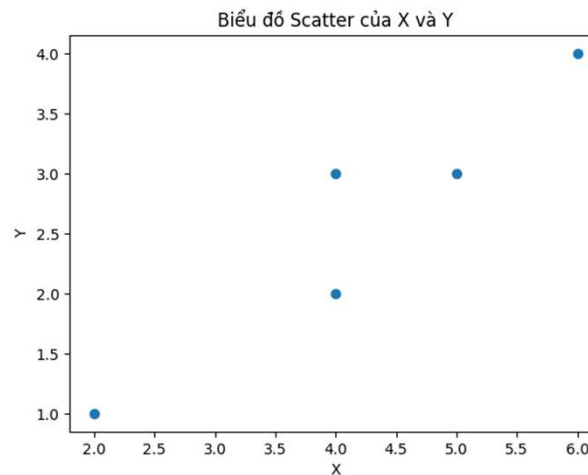
- Thực hiện tính **Cov(X,Y)**?

Phần 2. Hiệp phương sai và Độ tương quan

- **Hiệp phương sai**

- Hiệp phương sai mẫu của X và Y ký hiệu là **Cov(X,Y)** được xác định bởi công thức sau đây:

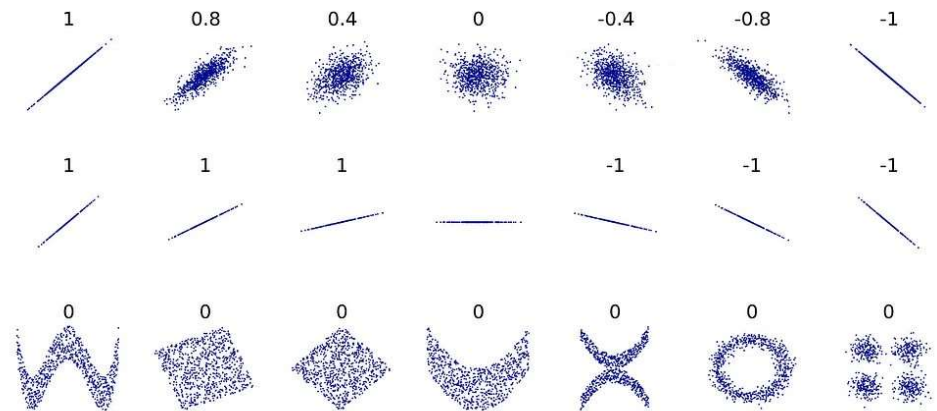
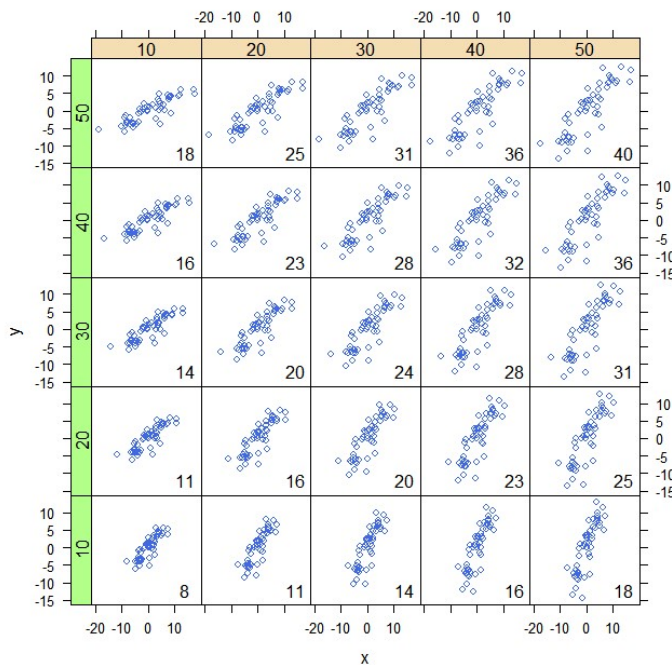
X	5	6	4	2	4
Y	3	4	2	1	3



Phần 2. Hiệp phương sai và Độ tương quan

- **Hiệp phương sai**

- Hiệp phương sai mẫu của X và Y ký hiệu là **Cov(X,Y)** được xác định bởi công thức sau đây:



Phần 2. Hiệp phương sai và Độ tương quan

- **Hiệp phương sai**

- Hiệp phương sai mẫu của X và Y ký hiệu là **Cov(X,Y)** được xác định bởi công thức sau đây:

Hiệp Phương Sai X và Y:
$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- $Cov(X, Y) > 0$ thì ta nói X và Y tương quan thuận với nhau
- **$Cov(X, Y) < 0$ thì ta nói X và Y tương quan nghịch với nhau**
- $Cov(X, Y) \approx 0$ thì X và Y không tương quan với nhau
- Mục tiêu của **hiệp phương sai** dùng để làm gì?
 - **Đo lường hướng của mối quan hệ:** Nếu hiệp phương sai dương, có thể nói rằng có mối quan hệ tích cực giữa hai biến. Nếu hiệp phương sai âm, có mối quan hệ tiêu cực.
 - **Đo lường độ mạnh của mối quan hệ:** Giá trị của hiệp phương sai càng lớn, mối quan hệ giữa hai biến càng mạnh.

Phần 2. Hiệp phương sai và Độ tương quan

- Ma trận hiệp phương sai:

$$S = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) & \text{Cov}(X, Z) \\ \text{Cov}(X, Y) & \text{Var}(Y) & \text{Cov}(Y, Z) \\ \text{Cov}(X, Z) & \text{Cov}(Y, Z) & \text{Var}(Z) \end{bmatrix}$$

- $\text{Var}(X)$: là phương sai của X
- $\text{Cov}(X, Y)$: là hiệp phương sai của X, Y

Phần 2. Hiệp phương sai và Độ tương quan

- Bài tập 2:

X	5	6	4	2	4
Y	3	4	2	1	3
Z	3	5	6	8	9

- Tính hiệp phương sai giữa các biến đưa về dạng **Covariance Matrix**

Phần 2. Hiệp phương sai và Độ tương quan

- Điểm yếu của Hiệp Phương Sai:

Hiệp Phương Sai X và Y:
$$Cov(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Hiệp phương sai là một đại lượng thống kê đo lường mối quan hệ tuyến tính giữa hai biến ngẫu nhiên. Nó biểu thị mức độ thay đổi chung của hai biến đối với nhau. Một điểm yếu của hiệp phương sai là nó **không chuẩn hóa**, nên giá trị của nó có thể bị **ảnh hưởng bởi đơn vị đo lường của các biến**. Ví dụ, nếu hai biến đo lường chiều cao và cân nặng của một nhóm người, thì hiệp phương sai sẽ khác nhau tùy theo đơn vị đo là mét và kilôgam hay inch và pound. Do đó, **hiệp phương sai** không thể so sánh được giữa các cặp biến khác nhau.
- Ngoài ra hiệp phương sai chỉ thể hiện tính tương quan không thể hiện tính mạnh yếu.

Phần 2. Hiệp phương sai và Độ tương quan

- **Độ tương quan (Correlation Pearson):**

- Độ tương quan Pearson giải quyết các vấn đề của Hiệp phương sai. Ký hiệu của độ tương quan giữa hai biến X và Y là: r

Độ Tương Quan giữa hai biến X và Y:
$$r_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

- Khi nhận xét về độ tương quan:
 - $r > 0$ mối quan hệ giữa X và Y tương quan thuận.
 - $r < 0$ mối quan hệ giữa X và Y tương quan yếu.
 - $r = 0$ mối quan hệ giữa X và Y là không tương quan.
 - r càng gần 1 hoặc -1 thì độ tương quan càng mạnh. Ngược lại càng gần 0 thì càng tương quan yếu.

Reference

**Python
Statistics**

Nguyen Minh Nhut
UIT – Information System

- [1] Peter Bruce (Author), Andrew Bruce (Author), Peter Gedeck (Author), Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python 2nd Edition
- [2] James R Evans, Business Analytics-Pearson (2017)
- [3] <http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke/861-thong-ke-mo-ta-trong-nghien-cuu-dai-luong-tuong-quan#:~:text=C%E1%BA%A3%20hai%20thu%E1%BA%ADt%20ng%E1%BB%AF%20%C4%91%E1%BB%81u,tuy%E1%BA%BFn%20t%C3%ADnh%20gi%E1%BB%AFa%20hai%20bi%E1%BA%BFn.> ¹⁷



CẢM ƠN ĐÃ THEO DÕI



ftisu.vn



minhnhut.ftisu@gmail.com



0939013911

