Tuần 1

# **Bài 1**

1. **Phân biệt xác suất và thống kê**

Xác suất là thước đo khả năng xảy ra một sự kiện. Vì xác suất là một thước đo được lượng hóa nên nó phải được phát triển với nền tảng toán học. Cụ thể, cấu trúc toán học về xác suất này được gọi là lý thuyết xác suất. Thống kê là bộ môn thu thập, tổ chức, phân tích, giải thích và trình bày dữ liệu

**Xác suất**: thông tin,quy luật của quần thể suy ra các thông tin cơ bản từng trường hợp. Populations -> Sample

**Thống kê**: Từ thông tin cơ bản từng trường hợp suy ra các quy luật cho quần thể. Sample -> Populations

1. **Phân biệt thống kê với học máy**

Khác biệt chính giữa học máy và thống kê là mục đích của chúng:

* Các mô hình học máy được thiết kế để đưa ra những dự đoán chính xác nhất có thể.
* Các mô hình thống kê được thiết kế để suy luận về mối quan hệ giữa các biến.

1. **Liệt kê yếu tố thống kê trong dữ liệu thị giác**

**-**Hàm phân bố xác suất trong histogram equalization

1. **Cho ví dụ về thống kê**

Dịch bệnh Covid - 19:

* Thống kê tình hình số ca nhiễm Covid - 19 theo vùng:
  + Số người mất, khỏi bệnh, tái bệnh, … => Phân vùng số người mắc Covid ( vùng xanh, đỏ, vàng,...)
* Thống kê về Vaccine:
  + Số người tiêm (từng loại Vaccine),số tuổi, giới tính , triệu chứng từng người,... => Triệu chứng sau tiêm tùy thuộc tính chất (từng loại Vaccine) => Phân bổ Vaccines

1. **Ví dụ ứng dụng thống kê: Sample -> Population**

Ví dụ về nhà ở sinh viên

Một cuộc khảo sát ước tính tỷ lệ tất cả sinh viên đại học sống ở nhà trong học kỳ hiện tại. Trong số 3.838 sinh viên đại học theo học tại trường, một mẫu ngẫu nhiên gồm 100 người đã được khảo sát.

Quần thể: Tất cả 3.838 sinh viên đại học

Mẫu: 100 sinh viên đại học được khảo sát

=>Chúng ta có thể sử dụng dữ liệu thu thập được từ mẫu 100 sinh viên để suy luận về quần thể 3,838 sinh viên.

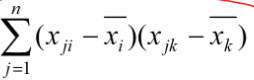
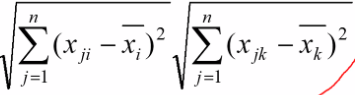
# **Bài 2**

1. **Sample mean có ý nghĩa thống kê không, đã đủ để thể hiện bản chất dữ liệu chưa**

Ý nghĩa: Chưa đủ để thể hiện được bản chất của dữ liệu. Thứ nhất vì là trung bình mẫu chứ không phải trung bình từng cá thể trong quần thể. Thứ hai vì cách lấy, chọn data tác động rất nhiều lên Sample mean, Do đó nếu chỉ dựa trên mean, bản chất dữ liệu sẽ bị tác động dựa trên cách chọn data, khiến cho bản chất không còn chính xác và dễ bị thay đổi.

1. **Tại sao dùng phép nhân để biết quan hệ giữa 2 biến ( Tại sao xài phép nhân trong Sample covariance)**
2. **Chứng minh Rik nằm trong khoảng nào?(Sample correlation coefficient)**

Áp dụng bất đẳng thức bunhiacopxki ta có

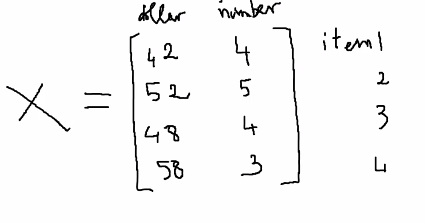
** <= **

Nên ta có |Rik|<=1 → -1<=Rik<=1

1. **Bài tập thống kê:**

**V1 (dollar sales) - Giá tiền**

**V2 (numbers of books) - Số sách**

****

**a)Tính Sample mean**

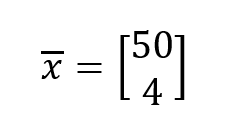
**Với n = 4, p = 2**

**=>**

**=> Số tiền bán trung bình (doanh thu trung bình)**

**=>**

**=> Số lượng sách bán trung bình**

****

**b)Tính Sample Variance, Standard Variance**

**Với n = 4, p = 2**

**=> Sự chênh lệch số lượng giữa các loại sách nhiều.**

**=> Sự chênh lệch số lượng giữa giá các loại sách ít.**

**c)Tính Sample Covariance**

**=> Số tiền và số sách có quan hệ ngược nhau**

**d)Tính Correlation Coefficient**

**=> Cường độ tương quan giữa hai biến ( giá tiền - sách )**

1. **Diễn giải ý nghĩa của các biến(Sample) ở bài trên**

-Sample mean: Trung bình mỗi loại sách bán được

Trung bình giá tiền mỗi loại sách

-Sample Variance, Standard Variance:

Độ lệch phương sai

Lấy số sách trung bình trừ độ lệch phương sai sách = số sách nên nhập mỗi ngày

Lấy số tiền trung bình trừ độ lệch phương sai tiền = số tiền nên sài từ số tiền bán sách

-Sample Covariance:

Bán ít sách nhưng doanh thu nhiều và ngược lại

-Correlation Coefficient : chỉ rõ hơn độ tương quan giữa số tiền thu được và số sách bán được.