Bài giảng 1

THỐNG KÊ ỨNG DỤNG

NỘI DUNG CHÍNH

- Thống kê và các ứng dụng trong kinh tế
- Dữ liệu
- Nguồn dữ liệu
- Thống kê mô tả
- Thống kê suy luận

THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ

Thống kê là một *Nghệ thuật* và *Khoa học về:*

- Thu thập
- Phân tích
- Trình bày
- Và giải thích DỮ LIỆU

THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ

Ứng dụng trong kinh tế:

- Các ứng dụng của thống kê rất hiển nhiên trong nhiều lãnh vực kinh tế
- Thống kê được sử dụng để:
  - Thông báo cho công chúng
  - Dự báo cho việc lập kế hoạch và ra quyết định

THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ

Các phần mềm thống kê so với Excel

- Các phần mềm thống kê thường là "Hộp đen"
  - **EVIEWS:** Economic Views

THỐNG KÊ ỨNG DỤNG TRONG KINH TẾ

Các phần mềm thống kê so với Excel

Sử dụng Excel để phân tích thống kê bởi vì:

- Excel sẵn có ở các văn phòng
- Excel đủ mạnh để giải quyết các vấn đề thống kê thường gặp
- Người sử dụng có thể hiểu được ý nghĩa của các vấn đề thống kê

Các nhà quản lý và ra quyết định thành công là những người có thể hiểu và sử dụng các thông tin một cách hiệu quả nhất

DỮ LIỆU

Dữ liệu

- **Dữ liệu** là các sự kiện và con số được thu thập, phân tích và tổng kết để trình bày và giải thích
- **Tập dữ liệu** là tất cả các dữ liệu được thu thập cho một nghiên cứu cụ thể
- **Thang đo**
- **Dữ liệu định tính so với định lượng**
- **Dữ liệu chéo so với chuỗi thời gian**

DỮ LIỆU

Thang đo

Xác định lượng thông tin có trong dữ liệu và chỉ ra sự tổng kết dữ liệu và phân tích thống kê nào là thích hợp nhất

- **Thang đo chỉ danh**
- **Thang đo thứ tự**
- **Thang đo khoảng**
- **Thang đo tỉ lệ**

DỮ LIỆU

- **Thang đo chỉ danh**

  Sử dụng nhãn hiệu hoặc tên để nhận dạng một thuộc tính của phần tử → bằng số hoặc không bằng số

- **Thang đo thứ tự**

  Có đặc tính của thang đo chỉ danh và có thể dùng để sắp hạng hoặc thứ tự dữ liệu → bằng số hoặc không bằng số

- **Thang đo khoảng**

  Có đặc tính của thang đo thứ tự và khoảng cách giữa các quan sát được diễn tả dưới dạng các đơn vị đo lường cố định → luôn luôn bằng số

- **Thang đo tỉ lệ**

  Có đặc tính của thang đo khoảng và tỉ lệ của 2 giá trị là có ý nghĩa → luôn luôn bằng số

  (Chứa giá trị **Zero** → Có nghĩa là không có gì)

DỮ LIỆU

Dữ liệu định tính so với định lượng

- **Dữ liệu định tính**
  - Dữ liệu định tính là các nhãn hiệu hay tên được dùng để nhận dạng và đặc trưng cho mỗi phần tử
  - Blến định tính là biến với dữ liệu định tính
  - Dữ liệu định tính sử dụng thang đo chỉ danh hoặc thang đo thứ tự; có thể đo bằng số hoặc không bằng số

DỮ LIỆU

Dữ liệu định tính so với định lượng

- **Dữ liệu định lượng**
  - Dữ liệu định lượng là dữ liệu cho biết số lượng bao nhiêu của một đại lượng nào đó
  - Biến định lượng là biến với dữ liệu định lượng
  - Dữ liệu định lượng sử dụng thang đo khoảng hoặc thang đo tỷ lệ; luôn đo bằng số

DỮ LIỆU

Dữ liệu định tính so với định lượng

- **Sự khác nhau** giữa dữ liệu định lượng và định tính
  - Các phép tính số học thông thường chỉ có ý nghĩa đối với dữ liệu định lượng
  - Tuy nhiên, khi dữ liệu định tính được ghi nhận như các giá trị bằng số thì các phép tính số học sẽ cho ra các kết quả không có ý nghĩa

DỮ LIỆU

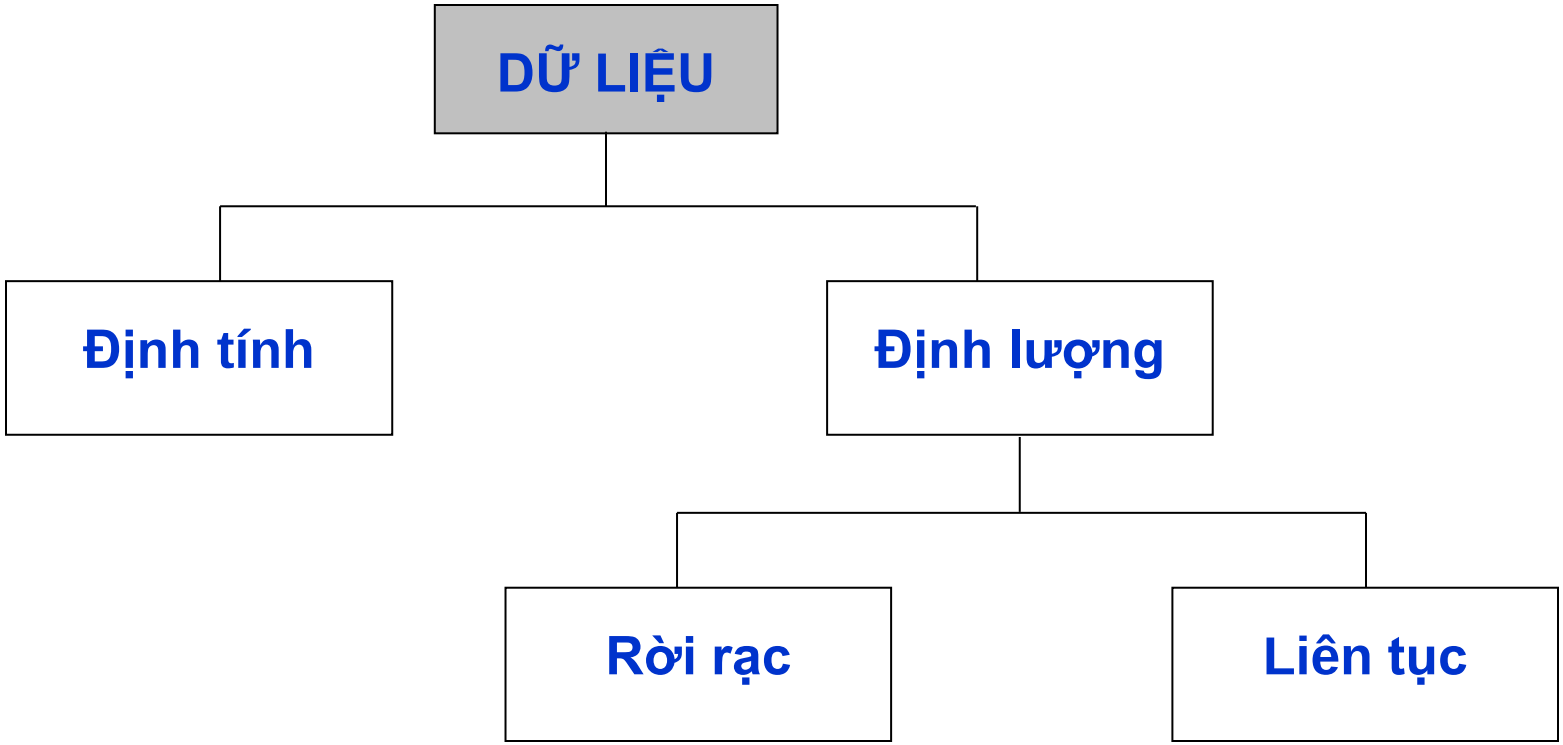
Dữ liệu định tính so với định lượng

- **Sự khác nhau** giữa dữ liệu định lượng và định tính
  - Các phép tính số học thông thường chỉ có ý nghĩa đối với dữ liệu định lượng
  - Tuy nhiên, khi dữ liệu định tính được ghi nhận như các giá trị bằng số thì các phép tính số học sẽ cho ra các kết quả không có ý nghĩa

DỮ LIỆU

- **Biến liên tục** là một biến có thể nhận tất cả giá trị nhiều vô hạn tương ứng với một khoảng vạch.
- **Biến rời rạc** chỉ có thể nhận một số có thể đếm được các giá trị

DỮ LIỆU



Câu hỏi ?

Hãy phát biểu xem các biến sau đây biến nào là biến định tính, biến nào là biến định lượng và hãy chỉ ra thang đo thích hợp cho mỗi biến.

- Tuổi
- Giới tính
- Thứ hạng trong lớp
- Nhiệt độ
- Thu nhập

DỮ LIỆU

Dữ liệu chéo và dữ liệu chuỗi thời gian

- Dữ liệu chéo là các dữ liệu được thu thập trong cùng hay gần cùng một thời điểm
- Dữ liệu chuỗi thời gian là các dữ liệu được thu thập trong các thời điểm liên tiếp nhau

NGUỒN DỮ LIỆU

Nguồn dữ liệu có thể thu thập từ:

- **Các nguồn hiện có:**

  Internet đã trở thành một nguồn dữ liệu quan trọng

- **Các nghiên cứu thống kê:**
  - Nghiên cứu thí nghiệm
  - Nghiên cứu quan sát

NGUỒN DỮ LIỆU

Các sai số của thu thập dữ liệu

- Một sai số trong thu thập dữ liệu xảy ra khi giá trị của dữ liệu thu thập được không bằng với giá đúng/thực có được từ một qui trình thu thập đúng
- Sử dụng dữ liệu sai có thể xấu hơn không sử dụng bất kỳ dữ liệu nào
- **GIGO** "Garbage In Garbage Out – Rác vào Rác Ra"

THỐNG KÊ MÔ TẢ

- **Thống kê mô tả**: Thu thập, Tổng kết và Mô tả dữ liệu
- **Các phương pháp** được sử dụng để tổng kết dữ liệu:
  - Lập Bảng
  - Đồ Thị
  - Bằng số

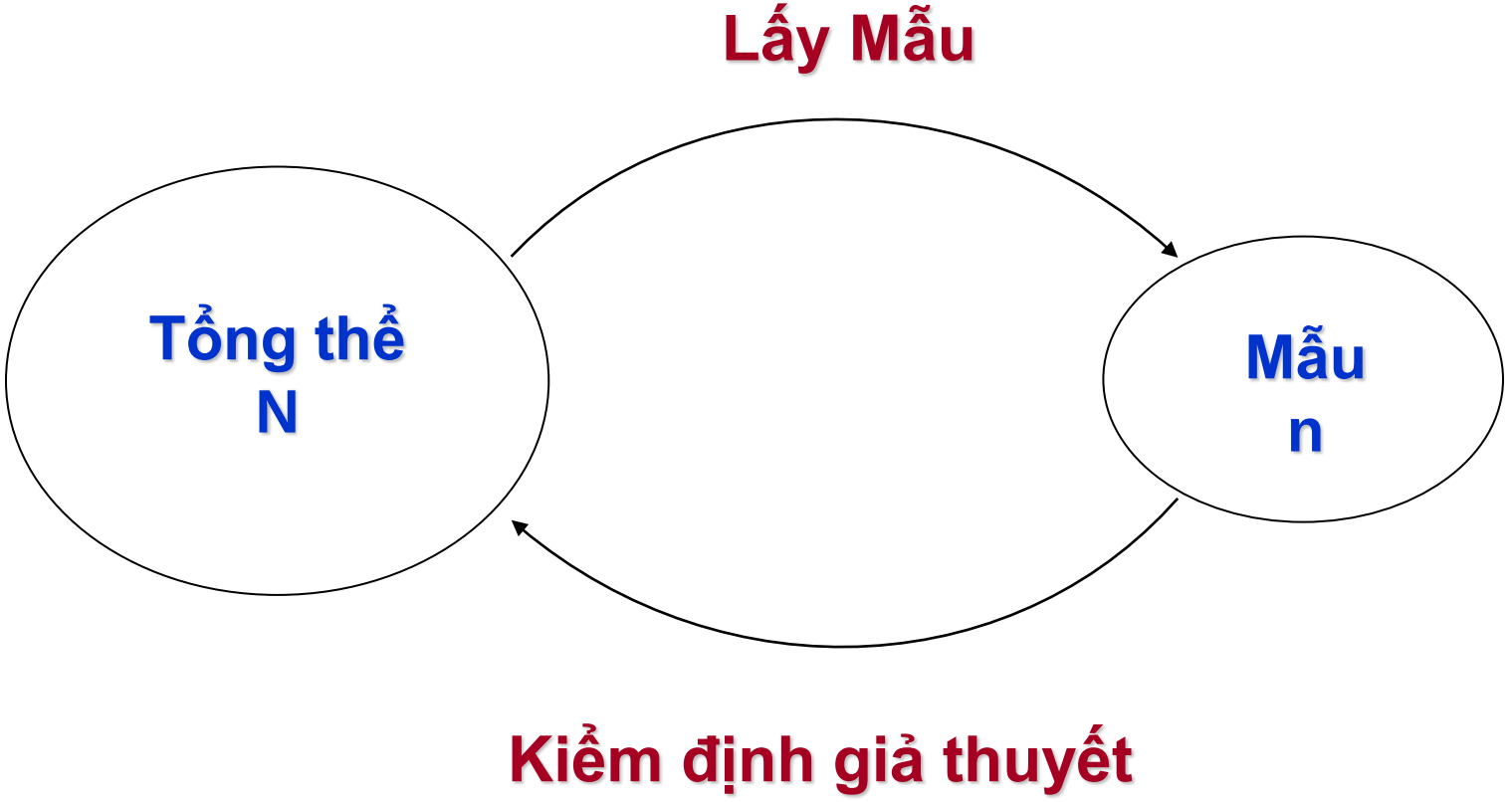
THỐNG KÊ MÔ TẢ

- **Thống kê mô tả**:
  - Các tham số thống kê
  - Tần số
  - Phân phối xác suất

THỐNG KÊ SUY LUẬN

- **Tổng thể** là tập tất cả các phần tử cần quan tâm trong một nghiên cứu cụ thể
- **Mẫu** là một tập con của tổng thể
- **Thống kê suy luận**: là quá trình sử dụng dữ liệu thu thập được từ mẫu để ước lượng hoặc kiểm định các giả thuyết thống kê về các đặc trưng của tổng thể

THỐNG KÊ SUY LUẬN



THỐNG KÊ MÔ TẢ

- Đại lượng về vị trí / số định tâm
- Đại lượng về sự biến thiên
- Đại lượng về dạng phân phối, vị trí tương đối và nhận dạng các điểm cá biệt
- Đại lượng về sự liên hệ giữa 2 biến

GIỚI THIỆU

- **Một đại lượng mô tả** là một con số đơn giản được tính toán từ dữ liệu mẫu để cung cấp thông tin về dữ liệu tổng thể
- Có hai loại đại lượng mô tả:
  - Đại lượng về vị trí
  - Đại lượng về sự biến thiên

GIỚI THIỆU

- **Tham số của tổng thể** (population parameter) là một giá trị bằng số được dùng như một đại lượng tổng kết đối với một dữ liệu của tổng thể

- **Các trị thống kê của mẫu** (sample statistics) được dùng như một đại lượng tổng kết đối với một mẫu

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ
(measure of location)

Một số **các đại lượng về vị trí** là:

- Số trung bình (Mean)
- Số trung vị (Median)
- Số yếu vị (Mode)
- Số phân vị (Percentiles)
- Số tứ phân (Quartiles)

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ

Số trung bình

- Số trung bình được sử dụng phổ biến nhất để đo lường vị trí
- Trung bình của tổng thể: $$\mu = \frac{\sum x}{N}$$
- Trung bình của mẫu: $$\bar{x} = \frac{\sum x}{n}$$

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ

Số trung vị (Md)

Số trung vị là giá trị ở giữa tập dữ liệu đã được sắp xếp theo thứ tự

- n là số lẻ, Md là giá trị ở giữa tập dữ liệu
- n là số chẵn, Md là trung bình của hai giá trị ở giữa tập dữ liệu

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ

Số yếu vị (Mo)

Số yếu vị là giá trị dữ liệu xuất hiện với tần số lớn nhất

- Bimodal → có hai số yếu vị
- Multimodal → > two hai số yếu vị

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ

Số phân vị

- Số phân vị p^{th} là giá trị có ít nhất p % số hạng của tập dữ liệu có giá trị nhỏ hơn hoặc bằng giá trị này, và có ít nhất (100-p) % số hạng của tập dữ liệu có giá trị lớn hơn hoặc bằng giá trị này
- Phân vị 50^{th} là số trung vị

CÁC ĐẠI LƯỢNG VỀ VỊ TRÍ

Số tứ phân

Số tứ phân chỉ đơn thuần là các số phân vị cụ thể, sẽ chia tập dữ liệu ra làm 4 phần, được gọi tên là:

- Q_1 = số tứ phân thứ nhất = $P_{25\%}$
- Q_2 = số tứ phân thứ hai = $P_{50\%}$ = Median
- Q_3 = số tứ phân thứ ba = $P_{75\%}$

CÁC ĐẠI LƯỢNG VỀ SỰ BIẾN THIÊN

- **Đại lượng về sự biến thiên** được sử dụng để mô tả xu hướng của các giá trị dữ liệu phân tán xung quanh giá trị trung bình.
- Một số đại lượng về sự biến thiên:
  - Khoảng biến thiên (Range)
  - Khoảng biến thiên nội tứ phân (Interquartile Range)
  - Phương sai (Variance)
  - Độ lệch chuẩn (Standard Deviation)
  - Hệ số biến thiên (Coefficient of variation)

CÁC ĐẠI LƯỢNG VỀ SỰ BIẾN THIÊN

- **Khoảng biến thiên**
  - Range = Giá trị lớn nhất – Giá trị nhỏ nhất
    hay
  - Range = Max – Min

- **Khoảng biến thiên nội tứ phân** (IQR)
  - IQR = Q3 – Q1

CÁC ĐẠI LƯỢNG VỀ SỰ BIẾN THIÊN

- **Phương sai**
  - Phương sai của tổng thể: $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$
  - Phương sai của mẫu: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

CÁC ĐẠI LƯỢNG VỀ SỰ BIẾN THIÊN

- **Độ lệch chuẩn**

  Độ lệch chuẩn là căn bậc hai của phương sai. Độ lệch chuẩn và phương sai được sử dụng phổ biến để đo lường sự biến thiên

$$\sigma = \sqrt{\sigma^2} \qquad s = \sqrt{s^2}$$

- **Hệ số biến thiên**

$$CV = \frac{\text{Standard Deviation}}{\text{Mean}} * 100 = \frac{S}{\overline{X}} * 100$$

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Dạng phân phối**
  - Độ lệch (Skewness) là đại lượng về dạng của phân phối của tập dữ liệu
    - Đối với dữ liệu lệch về bên trái, độ lệch sẽ âm
    - Đối với dữ liệu lệch về bên phải, độ lệch sẽ dương
    - Nếu dữ liệu đối xứng, độ lệch sẽ bằng 0
  - Đối với phân phối đối xứng, số trung bình và số trung vị sẽ bằng nhau

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Trị thống kê Z** (Z-Scores)

  Trị thống kê Z thường được gọi là giá trị chuẩn hóa

$$Z_i = \frac{x_i - \bar{x}}{s}$$

  Z_i: là số độ lệch chuẩn mà X_i cách xa giá trị trung bình

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Định lý Chebyshev**

  Định lý Chebyshev được sử dụng để phát biểu về phần trăm của các số hạng sẽ nằm trong một con số cụ thể của độ lệch chuẩn tính từ giá trung bình

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Định lý Chebyshev**
  - Tối thiểu (1-1/Z^2) của các số hạng có trong mọi tập dữ liệu sẽ phải nằm trong Z độ lệch chuẩn tính từ số trung bình, khi Z > 1.

    hay
  - Prob $\left(\bar{x} - zs < x < \bar{x} + zs\right) \geq 1 - \frac{1}{z^2}$

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Định lý Chebyshev**

  Đối với mọi tập dữ liệu

  - Prob $\left(\bar{x} - 2s < x < \bar{x} + 2s\right) \geq 75\%$
  - Prob $\left(\bar{x} - 3s < x < \bar{x} + 3s\right) \geq 89\%$
  - Prob $\left(\bar{x} - 4s < x < \bar{x} + 4s\right) \geq 94\%$

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

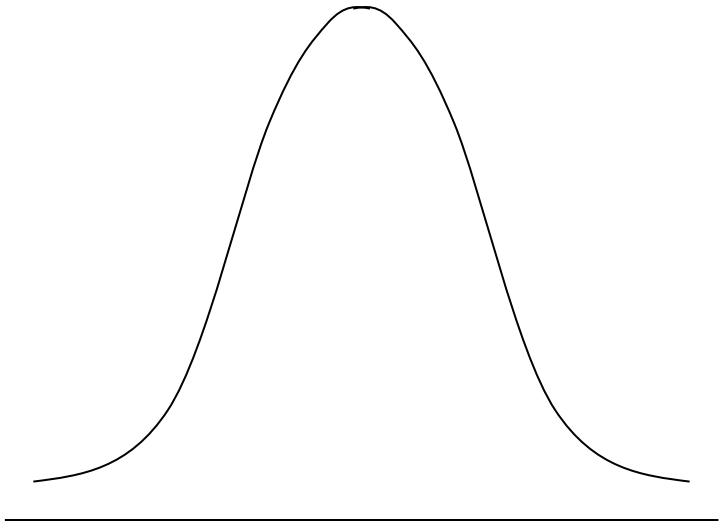
- **Qui tắc kinh nghiệm**

  Đối với mọi tập dữ liệu có phân phối dạng hình chuông:

  - Prob $\left(\bar{x} - 1s < x < \bar{x} + 1s\right) \geq 68\%$
  - Prob $\left(\bar{x} - 2s < x < \bar{x} + 2s\right) \geq 95\%$
  - Prob $\left(\bar{x} - 3s < x < \bar{x} + 3s\right) \geq 99.7\%$

CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

MỘT PHÂN PHỐI DẠNG HÌNH CHUÔNG ĐỐI XỨNG



CÁC ĐẠI LƯỢNG VỀ DẠNG PHÂN PHỐI, VỊ TRÍ TƯƠNG ĐỐI VÀ NHẬN DẠNG CÁC ĐIỂM CÁ BIỆT

- **Nhận dạng các điểm cá biệt** (outliers)
  - Các điểm cá biệt là các giá trị thái cực (lớn khác thường hoặc nhỏ khác thường)
  - Sử dụng Z để nhận dạng điểm cá biệt: mọi giá trị dữ liệu với Z nhỏ hơn –3 hoặc lớn hơn +3 là điểm cá biệt

ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

- **Đồng phương sai** (Covariance)
  - Đồng phương sai đo lường sự liên hệ tuyến tính giữa 2 biến.
  - Đồng phương sai của tổng thể: $$\sigma_{xy} = \frac{\sum (x_i - \mu_x)(y_i - \mu_y)}{N}$$
  - Đồng phương sai của mẫu: $$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

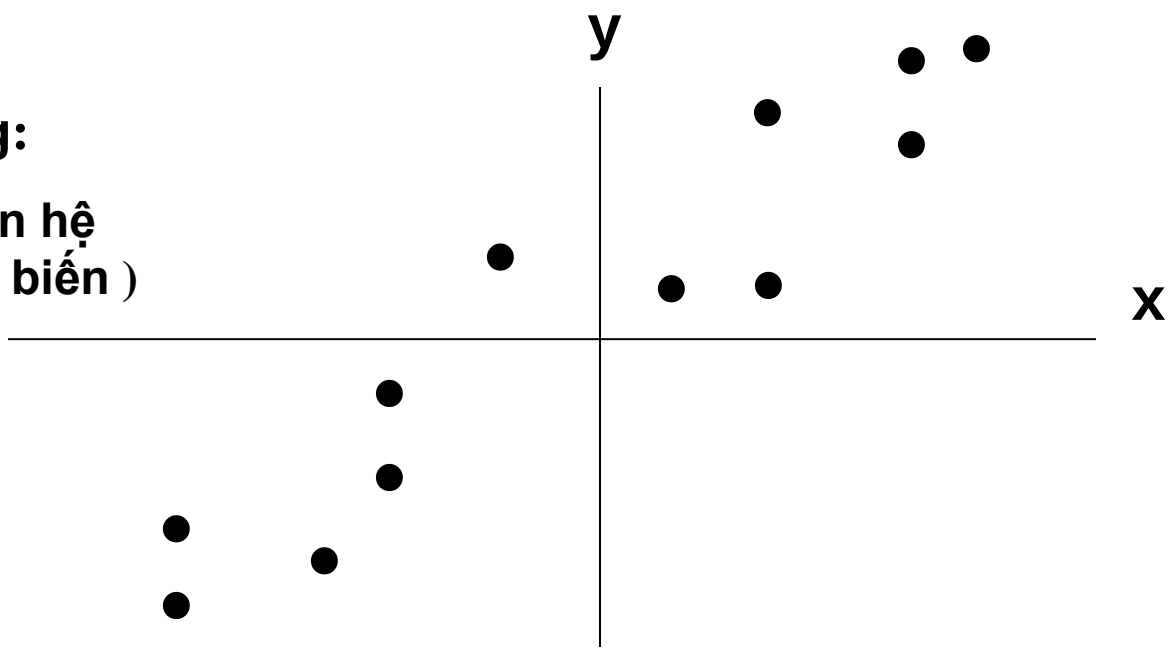
- **Đồng phương sai**
  - $s_{xy} > 0$ ⟶ Quan hệ đồng biến
  - $s_{xy} < 0$ ⟶ Quan hệ nghịch biến
  - Giá trị của đồng phương sai phụ thuộc đơn vị đo lường của x và y

ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

GIẢI THÍCH VỀ ĐỒNG PHƯƠNG SAI CỦA MẪU

S_{xy} dương:

(x và y có quan hệ tuyến tính đồng biến)

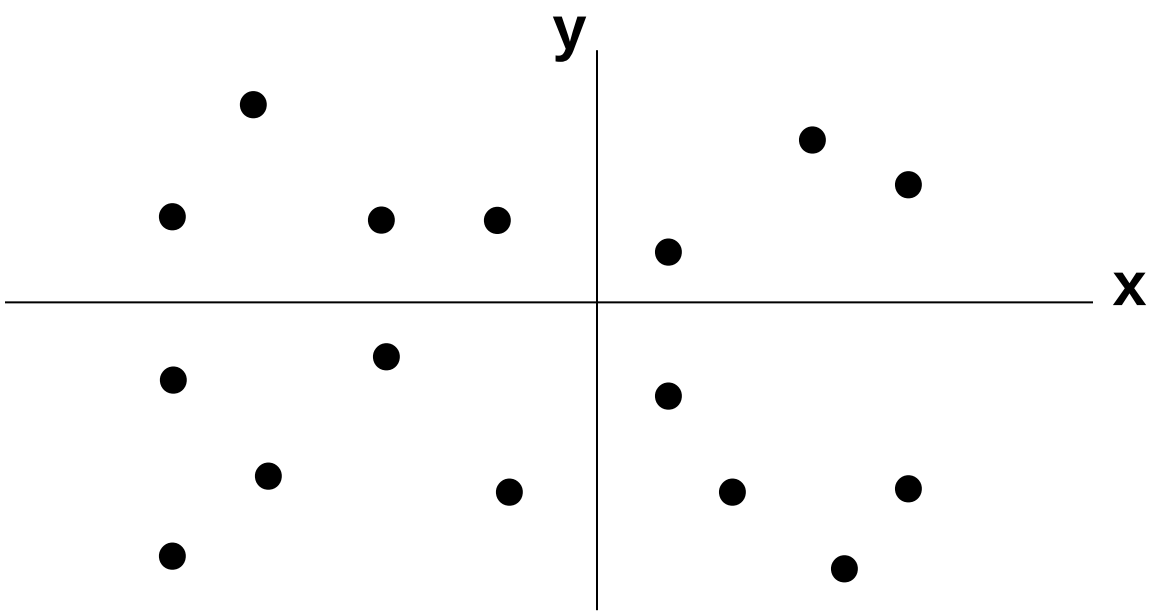


ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

GIẢI THÍCH VỀ ĐỒNG PHƯƠNG SAI CỦA MẪU

S_{xy} gần bằng 0:

(x và y không có quan hệ tuyến tính)

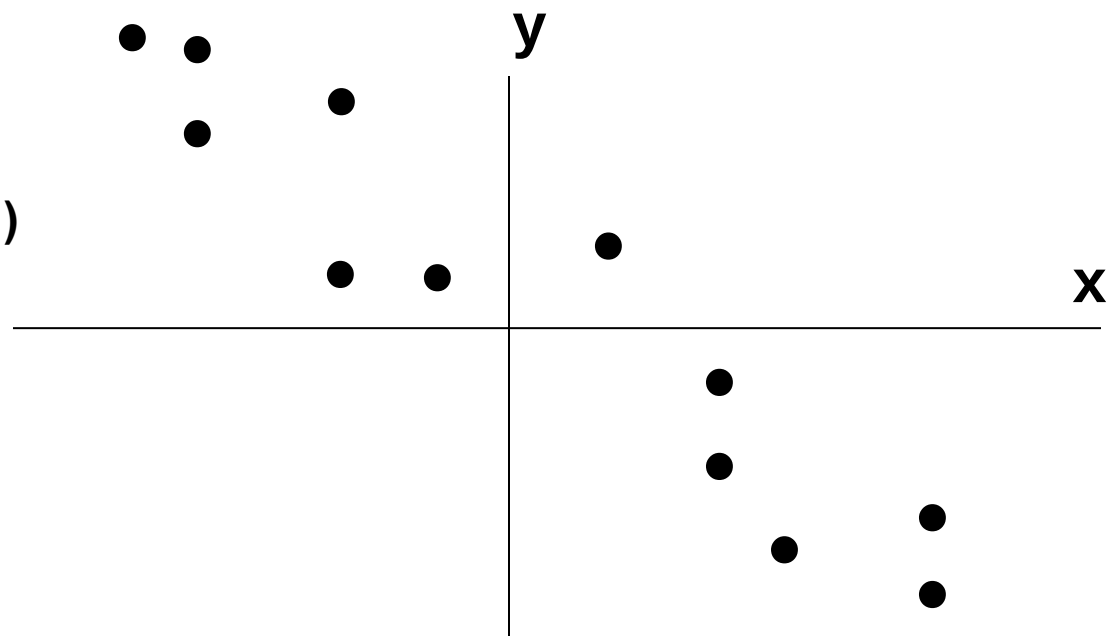


ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

GIẢI THÍCH VỀ ĐỒNG PHƯƠNG SAI CỦA MẪU

S_{xy} âm:

(x và y có quan hệ tuyến tính nghịch biến)



ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

- **Hệ số tương quan** (Correlation Coefficient)
  - Một đại lượng bằng số đo lường mối quan hệ tuyến tính giữa 2 biến
  - Hệ số tương quan Pearson
    - Tổng thể: $$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$
    - Mẫu: $$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

$$r = r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

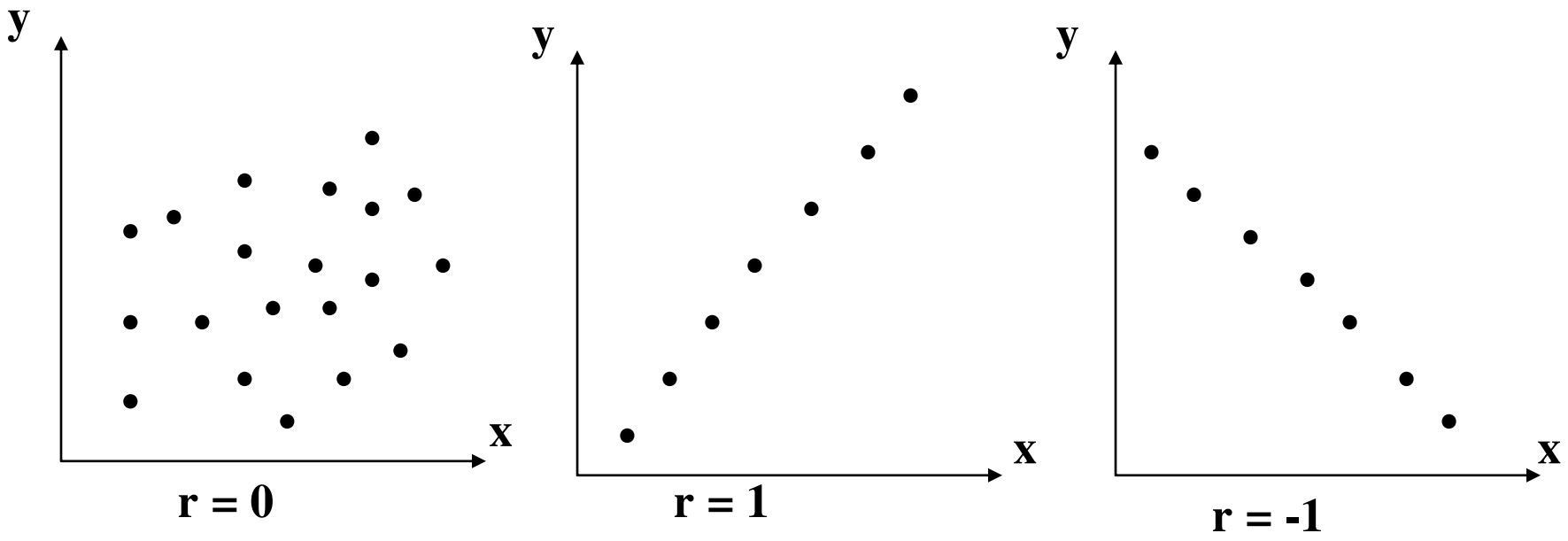
- **Hệ số tương quan**

  Các tính chất quan trọng của r:

  - $-1 \le r \le 1$
  - $|r|$ càng lớn thì mối quan hệ tuyến tính càng mạnh.
  - r = 0 -> không có quan hệ tuyến tính giữa X vàY
  - r = 1 hoặc r = -1 → X và Y tương quan tuyến tính hoàn toàn
  - Dấu của r cho thấy mối quan hệ giữa X và Y là đồng biến hay nghịch biến

ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

Đồ thị phân tán điểm đối với các giá trị r khác nhau



ĐẠI LƯỢNG VỀ SỰ LIÊN HỆ GIỮA 2 BIẾN

Đồ thị phân tán điểm đối với các giá trị r khác nhau