THỐNG KÊ MÔ TẢ

Nguyễn Thị Hiên

Ngày 10 tháng 11 năm 2023

Biến và dữ liệu

- **Biến:** một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang vật, hiện tượng khác. Biến gồm hai loại: Biến định tính (qualitative variable) và biến định lượng (quantitative variable).

Biến và dữ liệu

- **Biến:** một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang vật, hiện tượng khác. Biến gồm hai loại: Biến định tính (qualitative variable) và biến định lượng (quantitative variable).
- **Biến định tính:** biểu diễn tính chất của đặc trưng mà nó thể hiện, có tác dụng phân loại; ví dụ: nhóm máu (A, B, AB, O), giới tính (nam,nữ),....

Biến và dữ liệu

- **Biến:** một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang vật, hiện tượng khác. Biến gồm hai loại: Biến định tính (qualitative variable) và biến định lượng (quantitative variable).
- **Biến định tính:** biểu diễn tính chất của đặc trưng mà nó thể hiện, có tác dụng phân loại; ví dụ: nhóm máu (A, B, AB, O), giới tính (nam,nữ),....
- **Biến định lượng:** biểu diễn độ lớn của đặc trưng mà nó thể hiện; ví dụ: chiều cao, cân nặng, thời gian,...

Biến và dữ liệu

- **Biến:** một đặc trưng mà thay đổi từ người hay vật, hiện tượng này sang vật, hiện tượng khác. Biến gồm hai loại: Biến định tính (qualitative variable) và biến định lượng (quantitative variable).
- **Biến định tính:** biểu diễn tính chất của đặc trưng mà nó thể hiện, có tác dụng phân loại; ví dụ: nhóm máu (A, B, AB, O), giới tính (nam,nữ),....
- **Biến định lượng:** biểu diễn độ lớn của đặc trưng mà nó thể hiện; ví dụ: chiều cao, cân nặng, thời gian,...
- Biến định lượng bao gồm biến rời rạc (discrete variable) và biến liên tục (continuous variable).

Biến và dữ liệu

- Thông thường, biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể; ví dụ: số sản phẩm hỏng trong 1 lô hàng, số con trong 1 gia đình, số cuộc điện thoại đến tổng đài trong 1 giờ,... trong khi biến liên tục liên quan đến sự đo đạc; ví dụ: cân nặng của 1 sản phẩm, chiều cao của 1 cái cây, cường độ dòng điện, nhiệt độ,...

Biến và dữ liệu

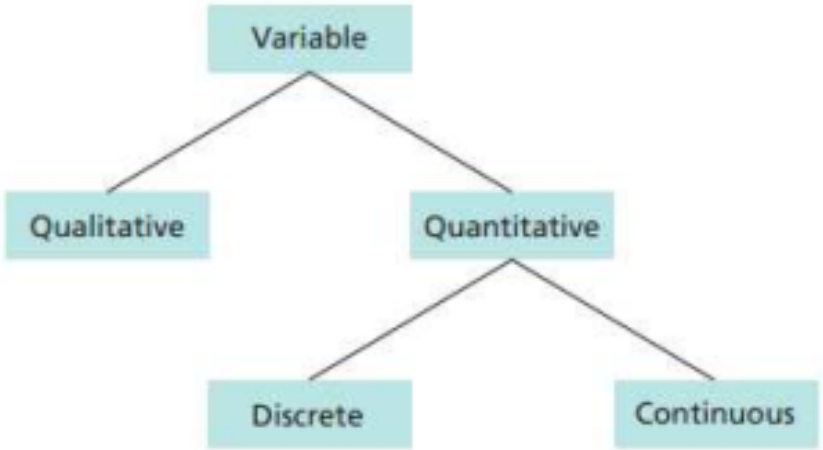
- Thông thường, biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể; ví dụ: số sản phẩm hỏng trong 1 lô hàng, số con trong 1 gia đình, số cuộc điện thoại đến tổng đài trong 1 giờ,... trong khi biến liên tục liên quan đến sự đo đạc; ví dụ: cân nặng của 1 sản phẩm, chiều cao của 1 cái cây, cường độ dòng điện, nhiệt độ,...
- **Dữ liệu (data):** các giá trị của một biến. Tập hợp tất cả những quan trắc cho một biến cụ thể được gọi là một tập dữ liệu (Data set).

Biến và dữ liệu

- Thông thường, biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể; ví dụ: số sản phẩm hỏng trong 1 lô hàng, số con trong 1 gia đình, số cuộc điện thoại đến tổng đài trong 1 giờ,... trong khi biến liên tục liên quan đến sự đo đạc; ví dụ: cân nặng của 1 sản phẩm, chiều cao của 1 cái cây, cường độ dòng điện, nhiệt độ,...
- **Dữ liệu (data):** các giá trị của một biến. Tập hợp tất cả những quan trắc cho một biến cụ thể được gọi là một tập dữ liệu (Data set).

Biến và dữ liệu

- Thông thường, biến rời rạc liên quan đến bài toán đếm số các phần tử của một tổng thể; ví dụ: số sản phẩm hỏng trong 1 lô hàng, số con trong 1 gia đình, số cuộc điện thoại đến tổng đài trong 1 giờ,... trong khi biến liên tục liên quan đến sự đo đạc; ví dụ: cân nặng của 1 sản phẩm, chiều cao của 1 cái cây, cường độ dòng điện, nhiệt độ,...
- **Dữ liệu (data):** các giá trị của một biến. Tập hợp tất cả những quan trắc cho một biến cụ thể được gọi là một tập dữ liệu (Data set).



Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.
  - Điểm trung bình của tất cả các sinh viên trong một trường đại học.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.
  - Điểm trung bình của tất cả các sinh viên trong một trường đại học.
- Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.
  - Điểm trung bình của tất cả các sinh viên trong một trường đại học.
- Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:
  - Số phần tử của tổng thể rất lớn.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.
  - Điểm trung bình của tất cả các sinh viên trong một trường đại học.
- Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:
  - Số phần tử của tổng thể rất lớn.
  - Thời gian và kinh phí không cho phép.

Tổng thể và mẫu

- **Tổng thể (population):** Tập hợp tất cả những phần tử mang đặc trưng quan tâm hay cần nghiên cứu.
- Ví dụ tổng thể:
  - Số cử tri đăng ký đi bầu cử.
  - Thu nhập của các hộ gia đình trong thành phố.
  - Điểm trung bình của tất cả các sinh viên trong một trường đại học.
- Thông thường, ta không thể chọn hết được tất cả các phần tử của tổng thể để nghiên cứu bởi vì:
  - Số phần tử của tổng thể rất lớn.
  - Thời gian và kinh phí không cho phép.
  - Có thể làm hư hại các phần tử của tổng thể.

Tổng thể và mẫu

Do đó, ta chỉ có thể thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

Tổng thể và mẫu

Do đó, ta chỉ có thể thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

- **Mẫu (sample):** là một tập con được chọn ra từ tổng thể. Ta thường ký hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.

Tổng thể và mẫu

Do đó, ta chỉ có thể thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

- **Mẫu (sample):** là một tập con được chọn ra từ tổng thể. Ta thường ký hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.
- **Tham số (parameter):** là một đặc trưng cụ thể của một tổng thể.

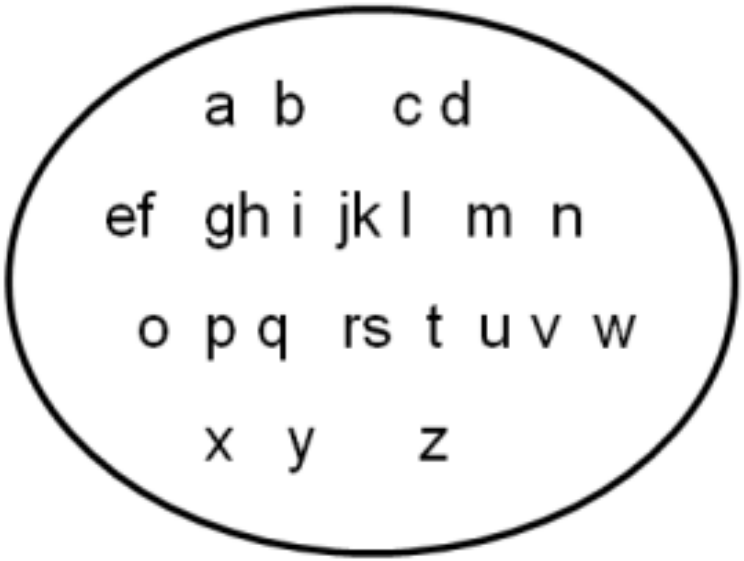
Tổng thể và mẫu

Do đó, ta chỉ có thể thực hiện nghiên cứu trên các mẫu được chọn ra từ tổng thể.

- **Mẫu (sample):** là một tập con được chọn ra từ tổng thể. Ta thường ký hiệu N để chỉ số phần tử của tổng thể và n để chỉ cỡ mẫu.
- **Tham số (parameter):** là một đặc trưng cụ thể của một tổng thể.
- **Thống kê (statistic):** là một đặc trưng cụ thể của một mẫu.

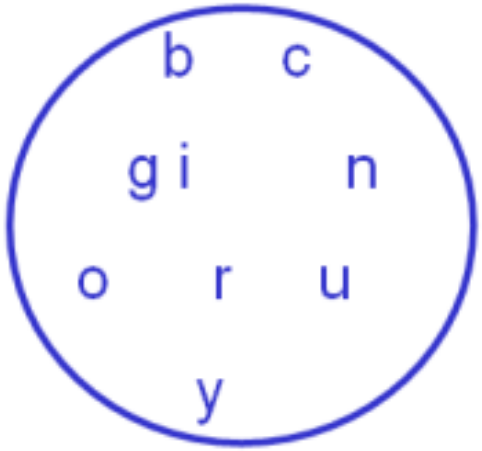
Tổng thể và mẫu

Population



Những giá trị tính từ dữ liệu của tổng thể gọi là các tham số

Sample



Những giá trị được tính từ dữ liệu của mẫu gọi là các thống kê

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- ◊ Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- ◊ Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- ◊ Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- Mọi mẫu cỡ n cũng có cùng khả năng được chọn từ tổng thể.

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- Mọi mẫu cỡ n cũng có cùng khả năng được chọn từ tổng thể.

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- ◊ Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- ◊ Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- ◊ Mọi mẫu cỡ n cũng có cùng khả năng được chọn từ tổng thể.

Phương pháp chọn mẫu ngẫu đơn giản (simple random sampling):

- ♡ Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.

Chọn mẫu ngẫu nhiên

Một mẫu ngẫu nhiên (random sample) gồm n phần tử được chọn ra từ tổng thể phải thỏa các điều kiện sau:

- ◊ Mỗi phần tử trong tổng thể phải được chọn ngẫu nhiên và độc lập.
- ◊ Mỗi phần tử trong tổng thể có khả năng được chọn như nhau (xác suất được chọn bằng nhau).
- ◊ Mọi mẫu cỡ n cũng có cùng khả năng được chọn từ tổng thể.

Phương pháp chọn mẫu ngẫu đơn giản (simple random sampling):

- ♡ Đánh số các phần tử của tổng thể từ 1 đến N. Lập các phiếu cũng đánh số như vậy.
- ♡ Trộn đều các phiếu, sau đó chọn có hoàn lại n phiếu. Các phần tử của tổng thể có số thứ tự trong phiếu lấy ra sẽ được chọn làm mẫu.

Thống kê mô tả

Thống kê mô tả (Descriptive statistics): là quá trình thu thập, tổng hợp và xử lý dữ liệu để biến đổi dữ liệu thành thông tin.

- Thu thập dữ liệu: khảo sát, đo đạc,...

Thống kê mô tả

Thống kê mô tả (Descriptive statistics): là quá trình thu thập, tổng hợp và xử lý dữ liệu để biến đổi dữ liệu thành thông tin.

- Thu thập dữ liệu: khảo sát, đo đạc,...
- Biểu diễn dữ liệu: dùng bảng tần số và đồ thị.

Thống kê mô tả

Thống kê mô tả (Descriptive statistics): là quá trình thu thập, tổng hợp và xử lý dữ liệu để biến đổi dữ liệu thành thông tin.

- Thu thập dữ liệu: khảo sát, đo đạc,...
- Biểu diễn dữ liệu: dùng bảng tần số và đồ thị.
- Tổng hợp dữ liệu: tính các tham số như trung bình mẫu, phương sai mẫu, trung vị,...

Thống kê suy luận

- Suy luận là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.

Thống kê suy luận

- Suy luận là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- Thống kê suy luận (Inferential statistics): xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở để dự đoán (predictions), dự báo (forecasts) và ước lượng (estimations).

Thống kê suy luận

- Suy luận là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- Thống kê suy luận (Inferential statistics): xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở để dự đoán (predictions), dự báo (forecasts) và ước lượng (estimations).
  - **Ước lượng:** ví dụ ước lượng tỷ lệ sản phẩm kém chất lượng trong 1 nhà máy; ước lượng trong lượng trung bình.

Thống kê suy luận

- Suy luận là một quá trình rút ra các kết luận hoặc đưa ra các quyết định về một tổng thể dựa vào các kết quả nghiên cứu từ mẫu.
- Thống kê suy luận (Inferential statistics): xử lý các thông tin có được từ thống kê mô tả, từ đó đưa ra các cơ sở để dự đoán (predictions), dự báo (forecasts) và ước lượng (estimations).
  - **Ước lượng:** ví dụ ước lượng tỷ lệ sản phẩm kém chất lượng trong 1 nhà máy; ước lượng trong lượng trung bình.
  - **Kiểm định giả thuyết:** ví dụ cần kiểm định khẳng định trọng lượng trung bình của 1 sản phẩm là 20 kg.

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:
  - Bảng

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:
  - Bảng
  - Đồ thị

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:
  - Bảng
  - Đồ thị
- Các loại đồ thị được sử dụng sẽ phụ thuộc vào biến được tổng hợp.

Giới thiệu

- Để đưa ra quyết đinh từ dữ liệu dạng thô thường khó. Do vậy, cần phải tổ chức lại dữ liệu.
- Các dạng tổ chức dữ liệu:
  - Bảng
  - Đồ thị
- Các loại đồ thị được sử dụng sẽ phụ thuộc vào biến được tổng hợp.
- Các dạng đồ thị quan trọng thường dùng: biểu đồ đường (line chart), đồ thị tổ chức tần số (histogram), đồ thị stem-and-leaf, đồ thị phân tán (scatter plot).

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,
- và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,
- và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,
- và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

Tại sao sử dụng phân phối tần số?

- phân phối tần số dùng để tổng hợp dữ liệu,

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,
- và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

Tại sao sử dụng phân phối tần số?

- phân phối tần số dùng để tổng hợp dữ liệu,
- biến đổi dữ liệu tho thành các thông tin hữu ích hơn,

Phân phối tần số

Phân phối tần số (Frequency distribution) là gì?

- là một danh sách hoặc bảng,
- chứa các khoảng được phân nhóm theo dữ liệu quan trắc,
- và các tần số tương ứng của dữ liệu nằm bên trong từng khoảng.

Tại sao sử dụng phân phối tần số?

- phân phối tần số dùng để tổng hợp dữ liệu,
- biến đổi dữ liệu tho thành các thông tin hữu ích hơn,
- sử dụng một cái nhìn trực quan để giải thích dữ liệu.

Xây dựng một phân phối tần số

Trong bảng phân phối tần số:

- Mỗi nhóm có bề rộng bằng nhau.

Xây dựng một phân phối tần số

Trong bảng phân phối tần số:

- Mỗi nhóm có bề rộng bằng nhau.
- Bề rộng của mỗi nhóm được xá định bởi

$$\frac{\text{Giá trị lớn nhất} - \text{Giá trị bé nhất}}{\text{Số khoảng cần chia}}$$

Xây dựng một phân phối tần số

Trong bảng phân phối tần số:

- Mỗi nhóm có bề rộng bằng nhau.
- Bề rộng của mỗi nhóm được xá định bởi

$$\frac{\text{Giá trị lớn nhất} - \text{Giá trị bé nhất}}{\text{Số khoảng cần chia}}$$

- Tối thiểu là 5 khoảng, không nhiều hơn 20 khoảng.

Xây dựng một phân phối tần số

Trong bảng phân phối tần số:

- Mỗi nhóm có bề rộng bằng nhau.
- Bề rộng của mỗi nhóm được xá định bởi

$$\frac{\text{Giá trị lớn nhất} - \text{Giá trị bé nhất}}{\text{Số khoảng cần chia}}$$

- Tối thiểu là 5 khoảng, không nhiều hơn 20 khoảng.
- Các khoảng không trùng nhau.

Xây dựng một phân phối tần số -ví dụ

Ví dụ 1

Chọn ngẫu nhiên 20 ngày mùa đông và đo nhiệt độ (Đv: Độ F) được số liệu như sau

24	35	17	21	24	37	26	46	58	30
32	13	12	38	41	43	44	27	53	27

Hãy lập bảng phân phối tần số cho số liệu này.

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$
- Xác định miền dữ liệu (range): $58 - 12 = 46$

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$
- Xác định miền dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5 (thông thường từ 5 đến 15)

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$
- Xác định miền dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5 (thông thường từ 5 đến 15)
- Xác định độ rộng của khoảng: 10 (làm tròn 46/5)

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$
- Xác định miền dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5 (thông thường từ 5 đến 15)
- Xác định độ rộng của khoảng: 10 (làm tròn $46/5$)
- Xác định biên của các khoảng: từ 10 đến dưới 20, từ 20 đến dưới 30,..., từ 50 đến dưới 60.

Xây dựng một phân phối tần số-ví dụ

Các bước thực hiện:

- Sắp xếp dữ liệu theo thứ tự tăng dần
  $12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58$
- Xác định miền dữ liệu (range): $58 - 12 = 46$
- Chọn số khoảng cần chia: 5 (thông thường từ 5 đến 15)
- Xác định độ rộng của khoảng: 10 (làm tròn 46/5)
- Xác định biên của các khoảng: từ 10 đến dưới 20, từ 20 đến dưới 30,. . . , từ 50 đến dưới 60.
- Đếm số giá trị dữ liệu nằm trong mỗi khoảng

Xây dựng một phân phối tần số-ví dụ

Dữ liệu được sắp tăng:

$$12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 423, 44, 46, 53, 58$$

Xây dựng một phân phối tần số-ví dụ

Dữ liệu được sắp tăng:

$$12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 423, 44, 46, 53, 58$$

Khoảng	Tần số	Tần suất
$[10, 20)$	3	0.15
$[20, 30)$	6	0.30
$[30, 40)$	5	0.25
$[40, 50)$	4	0.20
$[50, 60)$	2	0.10
Tổng	20	1.00

Xây dựng một phân phối tần số-ví dụ

Dữ liệu được sắp tăng:

$$12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 423, 44, 46, 53, 58$$

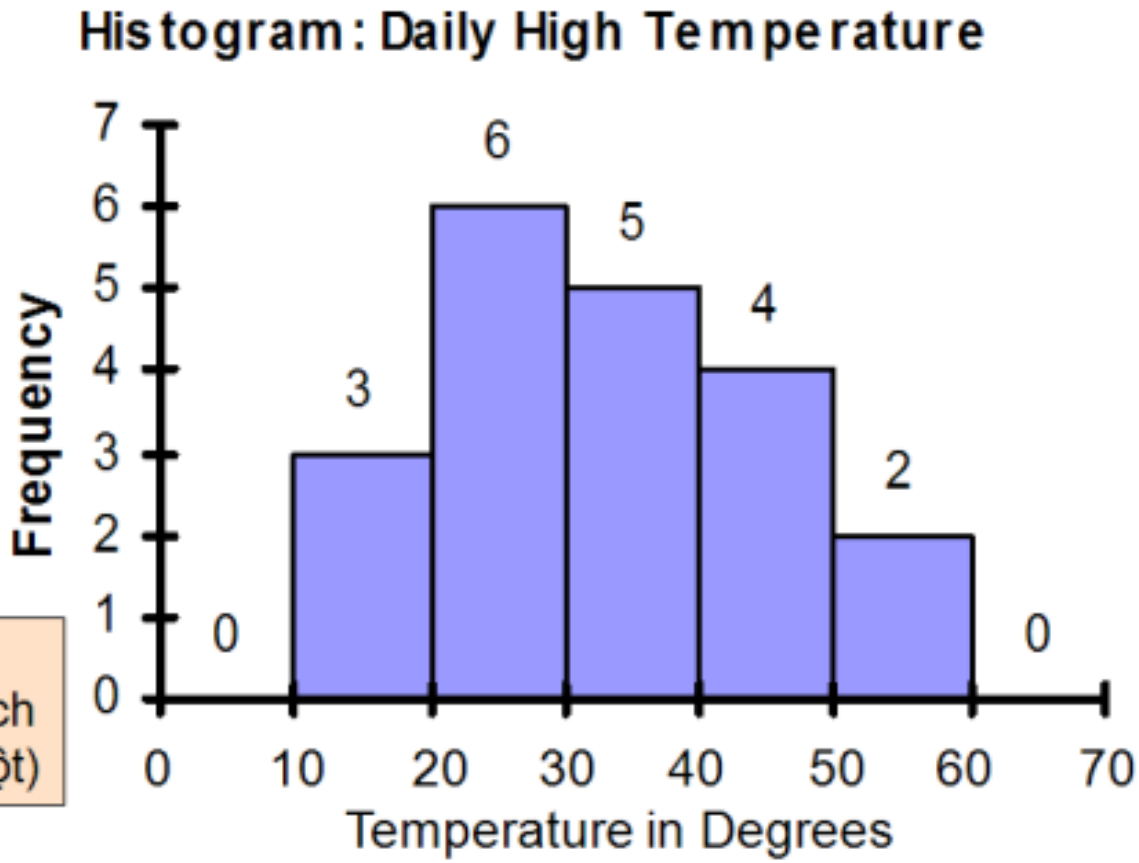
Khoảng	Tần số	Tần suất
$[10, 20)$	3	0.15
$[20, 30)$	6	0.30
$[30, 40)$	5	0.25
$[40, 50)$	4	0.20
$[50, 60)$	2	0.10
Tổng	20	1.00

Đồ thị biểu diễn bảng phân phối tần số gọi là đồ thị tổ chức tần số (histogram).

Xây dựng một phân phối tần số-ví dụ

Khoảng	Tần số
[10, 20)	3
[20, 30)	6
[30, 40)	5
[40, 50)	4
[50, 60)	2

(Không có khoảng cách giữa các cột)



Xây dựng một phân phối tần số

Câu hỏi: chia dữ liệu thành bao nhiêu khoảng là tốt? Chọn các điểm cắt thế nào cho phù hợp?

- Là quá trình "thử " và "sai".

Xây dựng một phân phối tần số

Câu hỏi: chia dữ liệu thành bao nhiêu khoảng là tốt? Chọn các điểm cắt thế nào cho phù hợp?

- Là quá trình "thử " và "sai".
- Mục tiêu là tạo được 1 phân phối không quá "lởm chởm",không có nhiều đỉnh và không có dạng "khối".

Xây dựng một phân phối tần số

Câu hỏi: chia dữ liệu thành bao nhiêu khoảng là tốt? Chọn các điểm cắt thế nào cho phù hợp?

- Là quá trình "thử " và "sai".
- Mục tiêu là tạo được 1 phân phối không quá "lởm chởm",không có nhiều đỉnh và không có dạng "khối".
- Mục tiêu là chỉ ra được sự biến thiên trong dữ liệu.

Xây dựng một phân phối tần số

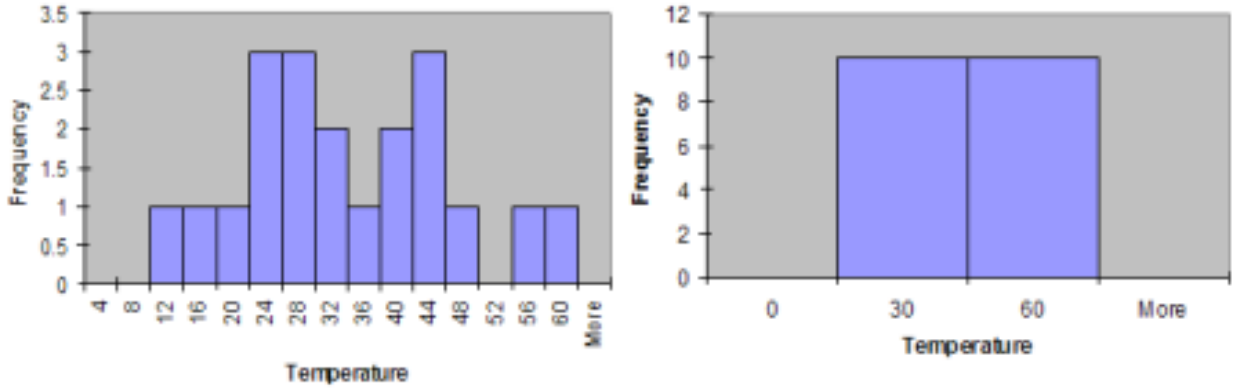
Câu hỏi: chia dữ liệu thành bao nhiêu khoảng là tốt? Chọn các điểm cắt thế nào cho phù hợp?

- Là quá trình "thử " và "sai".
- Mục tiêu là tạo được 1 phân phối không quá "lởm chởm",không có nhiều đỉnh và không có dạng "khối".
- Mục tiêu là chỉ ra được sự biến thiên trong dữ liệu.

Xây dựng một phân phối tần số

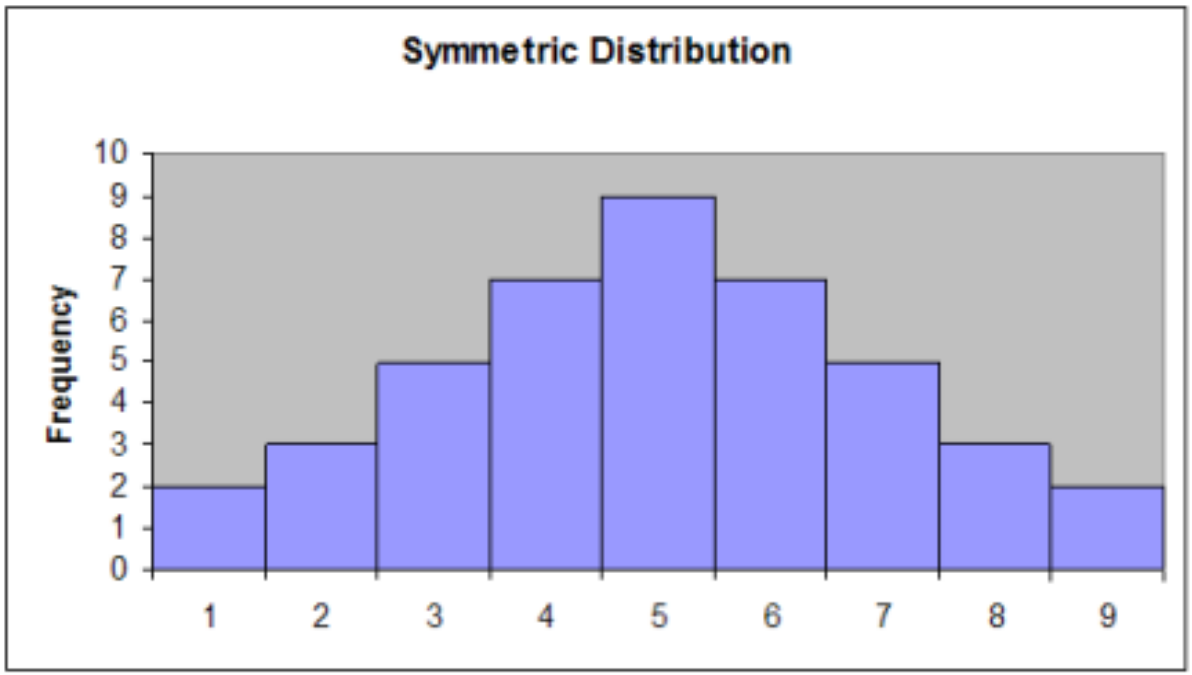
Câu hỏi: chia dữ liệu thành bao nhiêu khoảng là tốt? Chọn các điểm cắt thế nào cho phù hợp?

- Là quá trình "thử " và "sai".
- Mục tiêu là tạo được 1 phân phối không quá "lởm chởm",không có nhiều đỉnh và không có dạng "khối".
- Mục tiêu là chỉ ra được sự biến thiên trong dữ liệu.



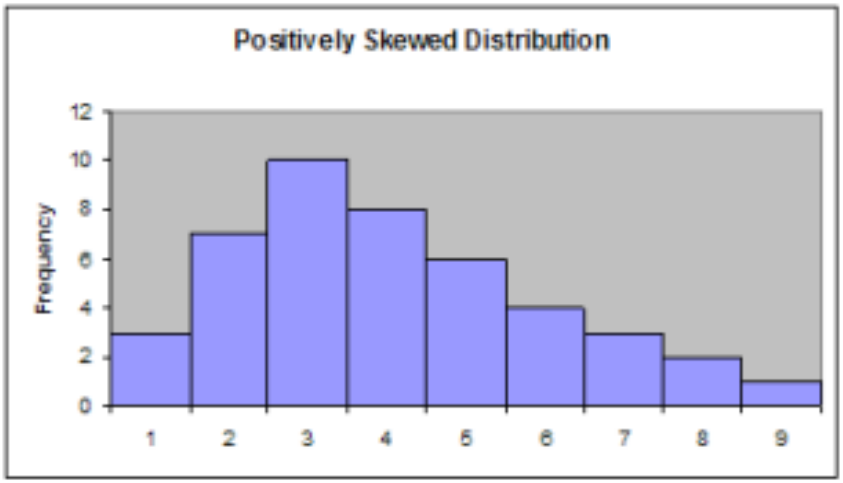
Hình dạng của phân phối

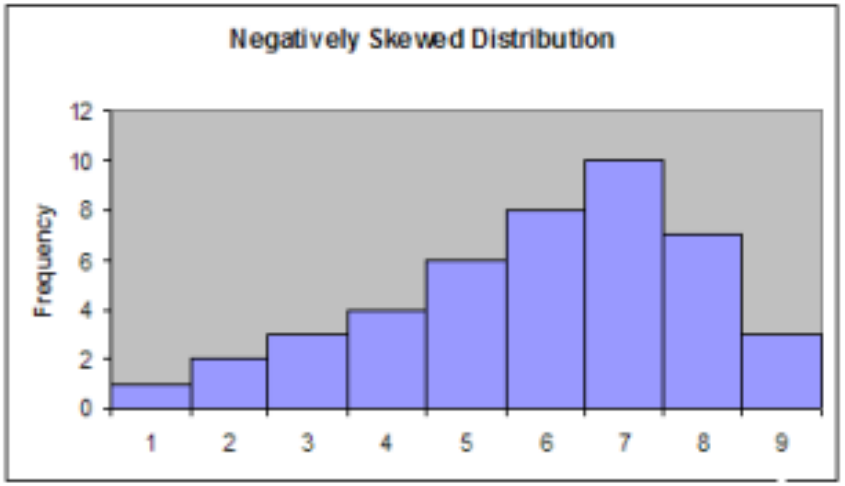
Hình dạng của phân phối gọi là đối xứng (symmetric) nếu các giá trị quan trắc cân bằng xung quanh trung tâm.



Hình dạng của phân phối

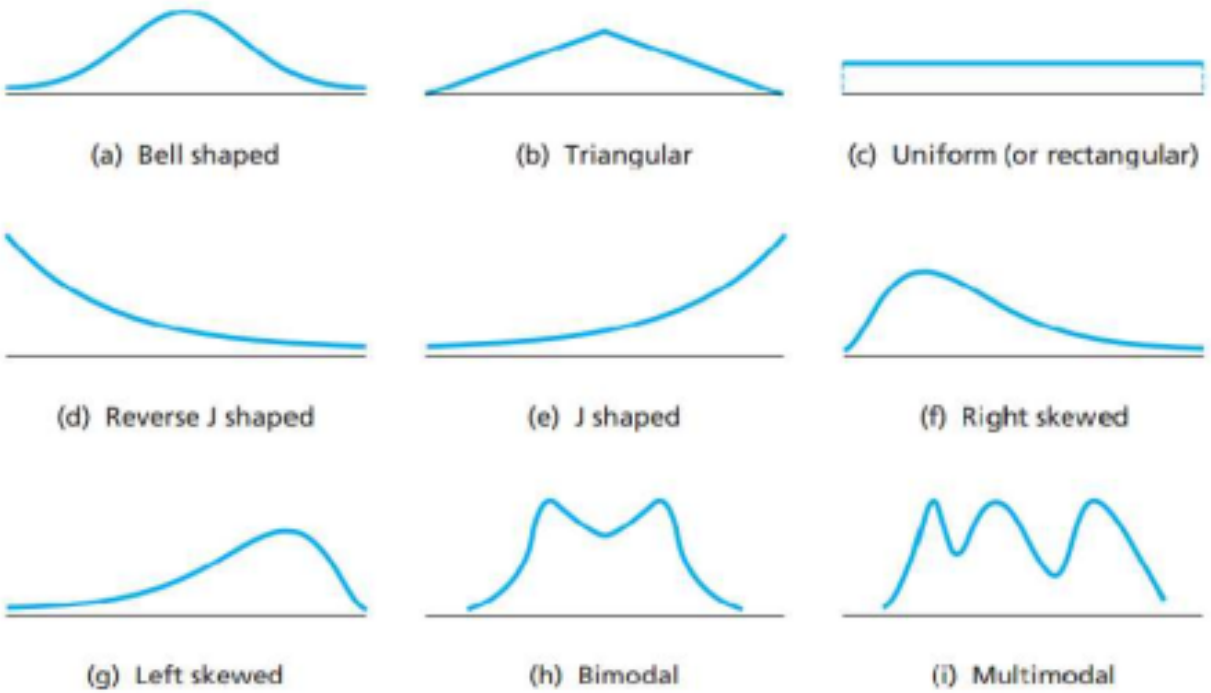
Hình dạng của phân phối gọi là bất đối xứng (skewed) nếu dữ liệu quan trắc không phân bố đối xứng xung quanh trung tâm.





Hình dạng của phân phối

Sử dụng đồ thị histogram để nhận biết phân phối xác suất của một đại lượng ngẫu nhiên.



Đồ thị Stem-and-Leaf

- Một phương pháp đơn giản để nhận biết các chi tiết của phân phối trong một tập dữ liệu.

Đồ thị Stem-and-Leaf

- Một phương pháp đơn giản để nhận biết các chi tiết của phân phối trong một tập dữ liệu.
- Phương pháp: sắp xếp dữ liệu theo thứ tự tăng dần, chia các giá trị đã sắp xếp thành hai phần: phần thứ nhất gồm các chữ số dẫn đầu (stem) và phần thứu hai là chữ số đuôi (leaf).

Đồ thị Stem-and-Leaf - ví dụ

Sắp xếp dữ liệu:

$$21, 24, 24, 26, 27, 27, 30, 32, 38, 41$$

Đồ thị Stem-and-Leaf - ví dụ

Sắp xếp dữ liệu:

$$21, 24, 24, 26, 27, 27, 30, 32, 38, 41$$

Sử dụng đơn vị hàng chục cho đơn vị của stem:

- 21 được ghi là
- 38 được ghi là

Stem	Leaf
2	1
3	8

Đồ thị Stem-and-Leaf - ví dụ

Sắp xếp dữ liệu:

$$21, 24, 24, 26, 27, 27, 30, 32, 38, 41$$

Hoàn thành đồ thị stem-and-leaf:

Stem	Leaves
2	1 4 4 6 7 7
3	0 2 8
4	1

Đồ thị Stem-and-Leaf - ví dụ

Sử dụng đơn vị hàng trăm cho stem: Đồ thị stem-and-leaf:

- Làm tròn chữ số hàng chục để làm leaf
  - 613 được ghi là → 6 | 1
  - 776 được ghi là → 7 | 8
  - ...
  - 1224 được ghi là → 12 | 2

Stem	Leaf
6	1
7	8
12	2

Đồ thị Stem-and-Leaf - ví dụ

Sử dụng đơn vị hàng trăm cho stem: Đồ thị stem-and-leaf:

Data:

613, 632, 658, 717,
722, 750, 776, 827,
841, 859, 863, 891,
894, 906, 928, 933,
955, 982, 1034,
1047,1056, 1140,
1169, 1224

Stem	Leaves
6	1 3 6
7	2 2 5 8
8	3 4 6 6 9 9
9	1 3 3 6 8
10	3 5 6
11	4 7
12	2

Đồ thị Stem-and-Leaf - ví dụ

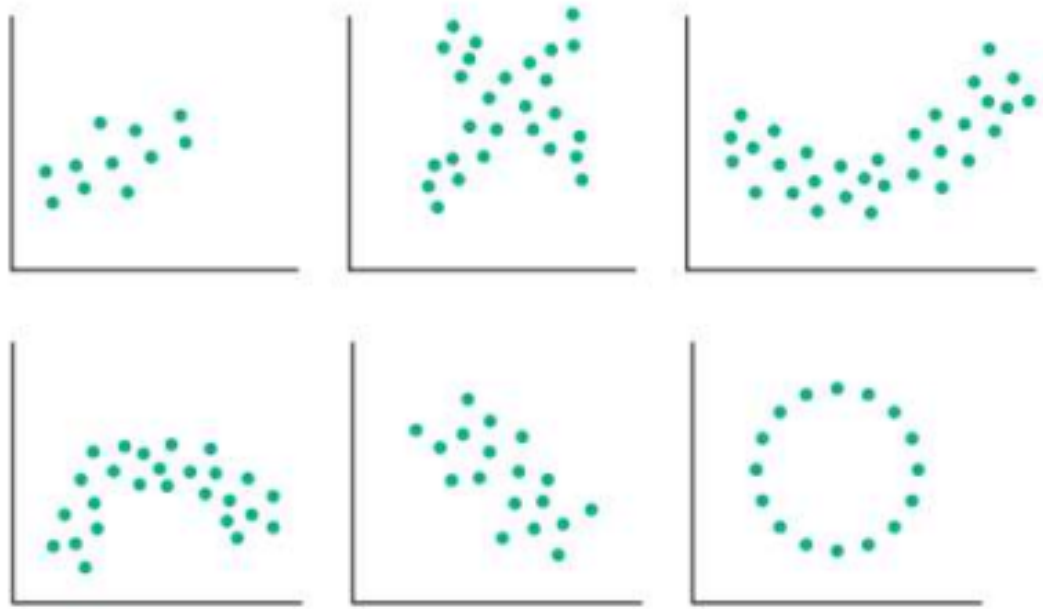
Ví dụ 2

Vẽ đồ thị stem-and-leaf cho tập dữ liệu sau:

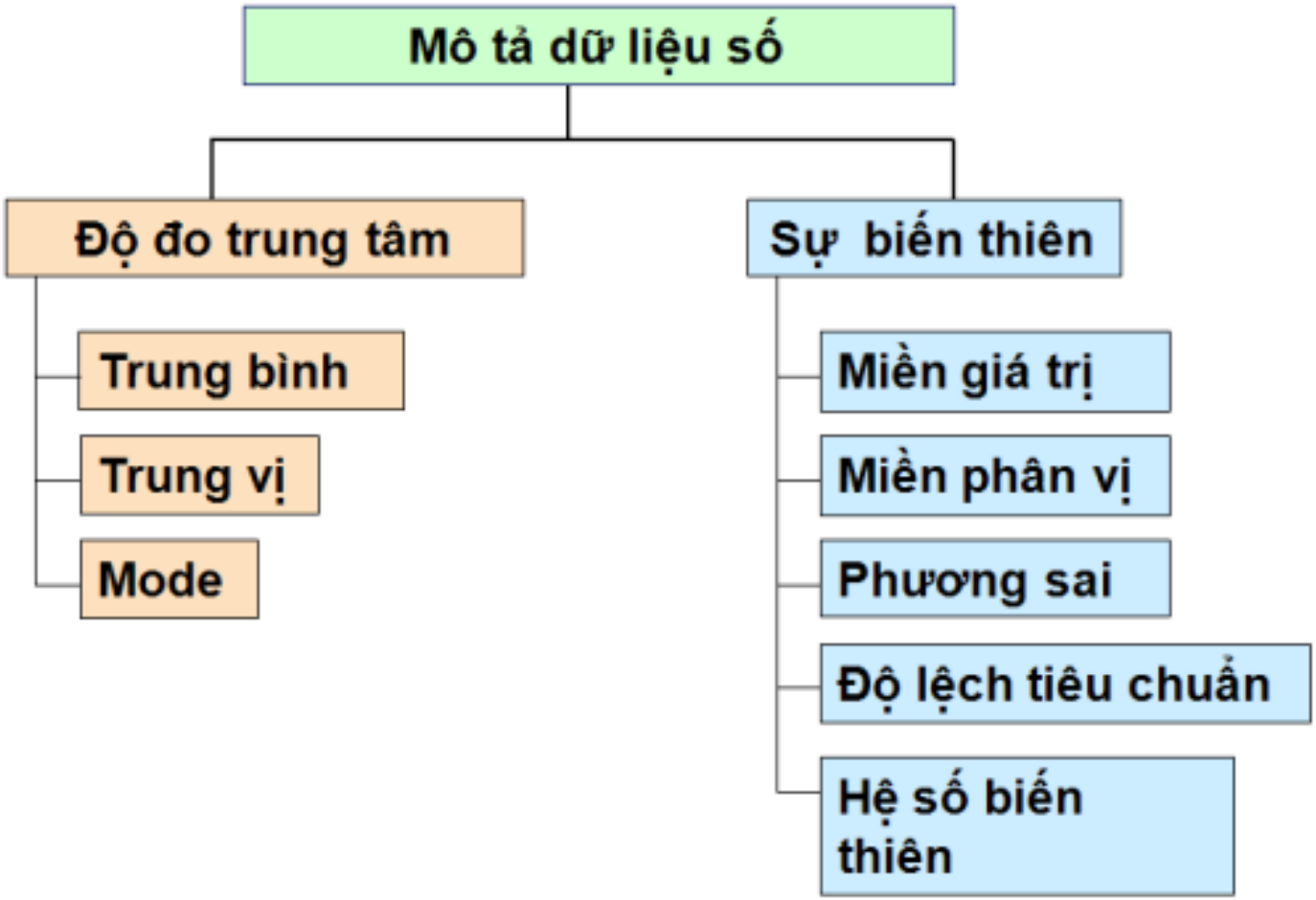
61	63	70	71	71	81	83	84	64	65
65	66	84	87	73	75	92	93	77	78
78	88	88	95	79					

Đồ thị phân tán

Đồ thị phân tán (Scatter plot) được sử dụng để xác định mối liên hệ giữa hai biên X và Y.



Giới thiệu


Mô tả dữ liệu số
Độ đo trung tâm
Trung bình
Trung vị
Mode
Sự biến thiên
Miền giá trị
Miền phân vị
Phương sai
Độ lệch tiêu chuẩn
Hệ số biến thiên

Các độ đo xu hướng trung tâm

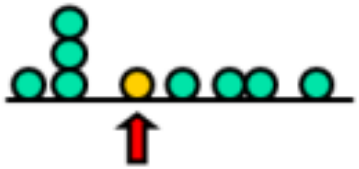
Độ đo trung tâm

Trung bình

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

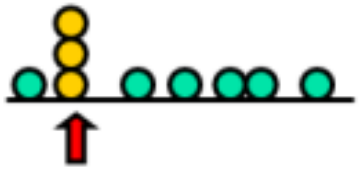
Trung bình số học

Trung vị



Điểm chính giữa của dữ liệu đã sắp xếp

Mode



Giá trị thường gặp nhất

Trung bình

- Trung bình (mean) là đại lượng thường được sử dụng để đo gía trị trung tâm của dữ liệu.

Trung bình

- Trung bình (mean) là đại lượng thường được sử dụng để đo gía trị trung tâm của dữ liệu.
- Với một tổng thể có N phần tử, trung bình của tổng thể được tính như sau:

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{1}$$

Trung bình

- Trung bình (mean) là đại lượng thường được sử dụng để đo gía trị trung tâm của dữ liệu.
- Với một tổng thể có N phần tử, trung bình của tổng thể được tính như sau:

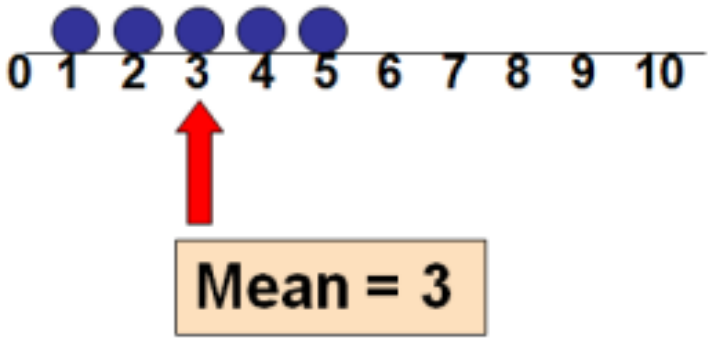
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \tag{1}$$

- Với một mẫu cỡ n, trung bình mẫu được tính:

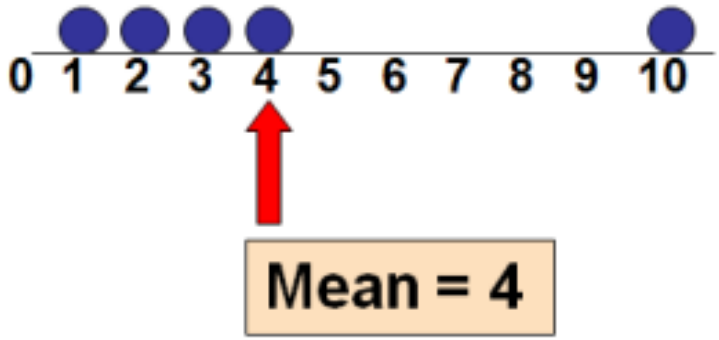
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{2}$$

Trung bình

Trung bình bị ảnh hưởng bởi điểm ngoại lai (outliers)



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Trung vị

- Trong một tập dữ liệu được sắp xếp (tăng), trung vị (median) là giá trị "chính giữa" của dữ liệu (50% bên trên, 50% bên dưới).

Trung vị

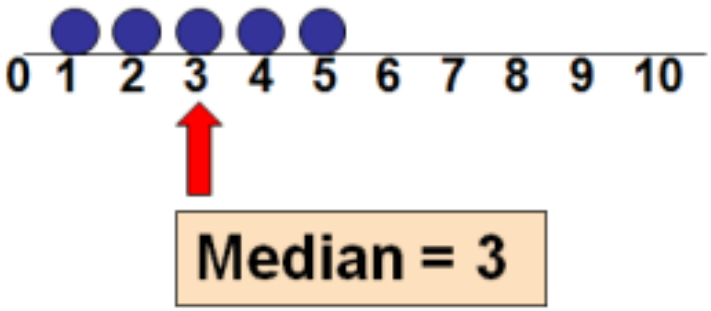
- Trong một tập dữ liệu được sắp xếp (tăng), trung vị (median) là giá trị "chính giữa" của dữ liệu (50% bên trên, 50% bên dưới).
- Trung vị không bị ảnh hưởng bởi các điểm outliers.

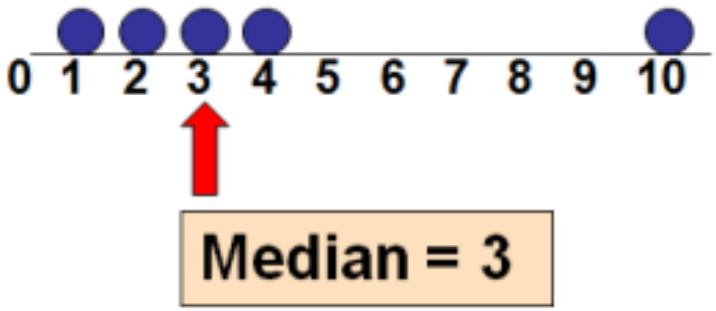
Trung vị

- Trong một tập dữ liệu được sắp xếp (tăng), trung vị (median) là giá trị "chính giữa" của dữ liệu (50% bên trên, 50% bên dưới).
- Trung vị không bị ảnh hưởng bởi các điểm outliers.

Trung vị

- Trong một tập dữ liệu được sắp xếp (tăng), trung vị (median) là giá trị "chính giữa" của dữ liệu (50% bên trên, 50% bên dưới).
- Trung vị không bị ảnh hưởng bởi các điểm outliers.





Trung vị

Vị trí của trung vị: sắp xếp theo thứ tự tăng dần, gọi i là vị trí của trung vị:

$$i = \frac{n+1}{2}$$

- Nếu i chẵn, trung vị $= x_i$,
- Nếu i lẻ, trung vị $= \frac{x_{[i]}+x_{[i]+1}}{2}$, với $[i]$ là phần nguyên của i.

Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.

Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.

Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.
- Không bị ảnh hưởng bởi các điểm ngoại lai.

Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.
- Không bị ảnh hưởng bởi các điểm ngoại lai.
- Có thể sử dụng cho cả dữ liệu dạng số và dữ liệu phân loại.

Mode

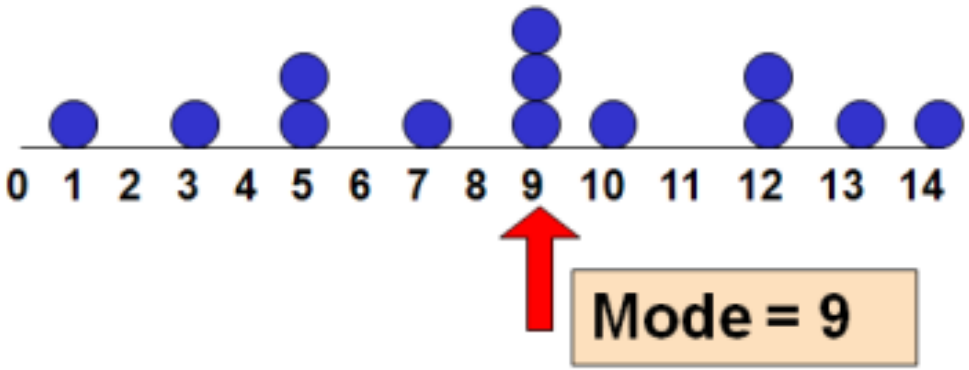
- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.
- Không bị ảnh hưởng bởi các điểm ngoại lai.
- Có thể sử dụng cho cả dữ liệu dạng số và dữ liệu phân loại.
- Có thể có nhiều mode hoặc không tồn tại mode trong dữ liệu

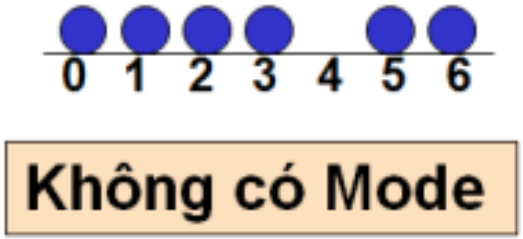
Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.
- Không bị ảnh hưởng bởi các điểm ngoại lai.
- Có thể sử dụng cho cả dữ liệu dạng số và dữ liệu phân loại.
- Có thể có nhiều mode hoặc không tồn tại mode trong dữ liệu

Mode

- Mode là một địa lượng để đo xu hướng trung tâm của dữ liệu.
- là giá trị thường xảy ra nhất.
- Không bị ảnh hưởng bởi các điểm ngoại lai.
- Có thể sử dụng cho cả dữ liệu dạng số và dữ liệu phân loại.
- Có thể có nhiều mode hoặc không tồn tại mode trong dữ liệu





Độ đo nào là tốt nhất ?

- Trung bình luôn luôn được sử dung, nếu các điểm outliers không tồn tại.

Độ đo nào là tốt nhất ?

- Trung bình luôn luôn được sử dung, nếu các điểm outliers không tồn tại.
- Trung vị thường được dùng vì trung vị không bị ảnh hưởng bởi các điểm outliers.

Độ đo nào là tốt nhất ?

- Trung bình luôn luôn được sử dung, nếu các điểm outliers không tồn tại.
- Trung vị thường được dùng vì trung vị không bị ảnh hưởng bởi các điểm outliers.
- Vị trí của trung vị và trung bình ảnh hưởng bởi hình dạng của phân phối:

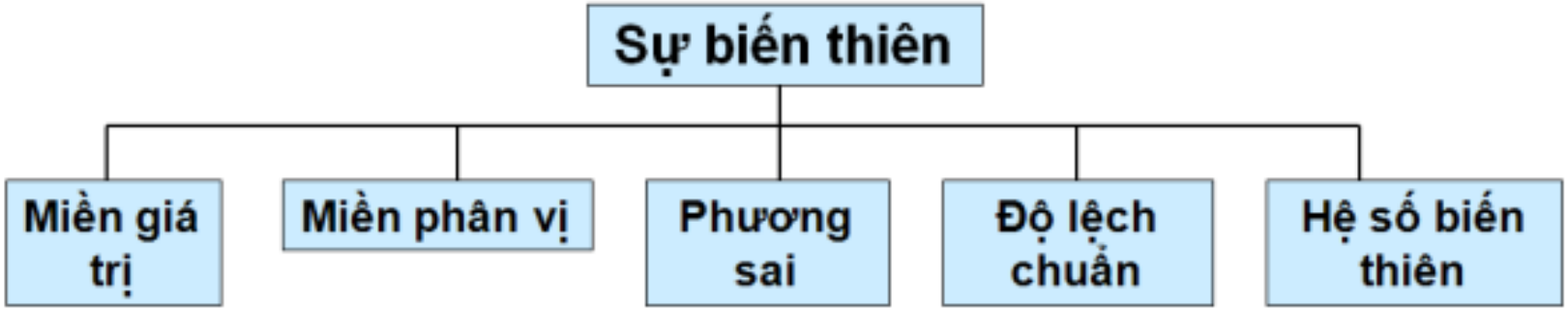
Độ đo nào là tốt nhất ?

- Trung bình luôn luôn được sử dung, nếu các điểm outliers không tồn tại.
- Trung vị thường được dùng vì trung vị không bị ảnh hưởng bởi các điểm outliers.
- Vị trí của trung vị và trung bình ảnh hưởng bởi hình dạng của phân phối:

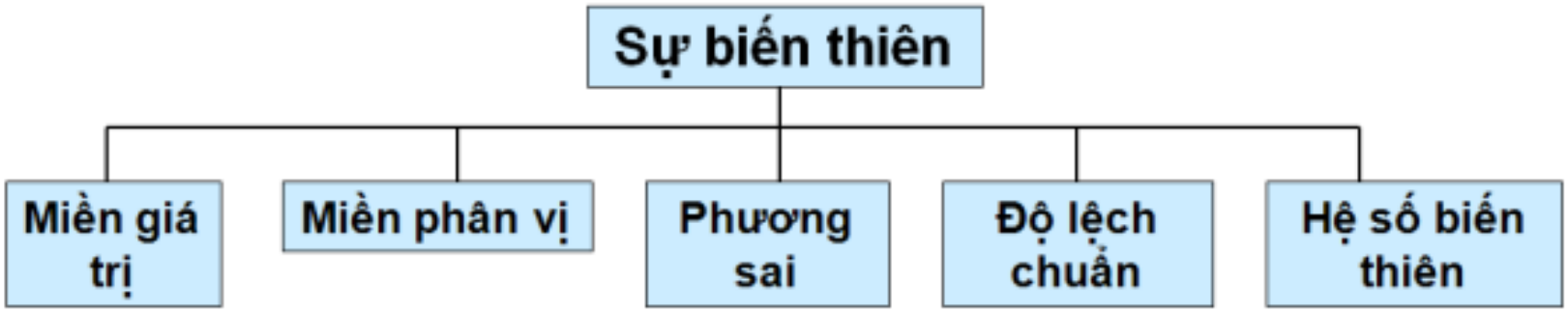
Độ đo nào là tốt nhất ?

- Trung bình luôn luôn được sử dung, nếu các điểm outliers không tồn tại.
- Trung vị thường được dùng vì trung vị không bị ảnh hưởng bởi các điểm outliers.
- Vị trí của trung vị và trung bình ảnh hưởng bởi hình dạng của phân phối:

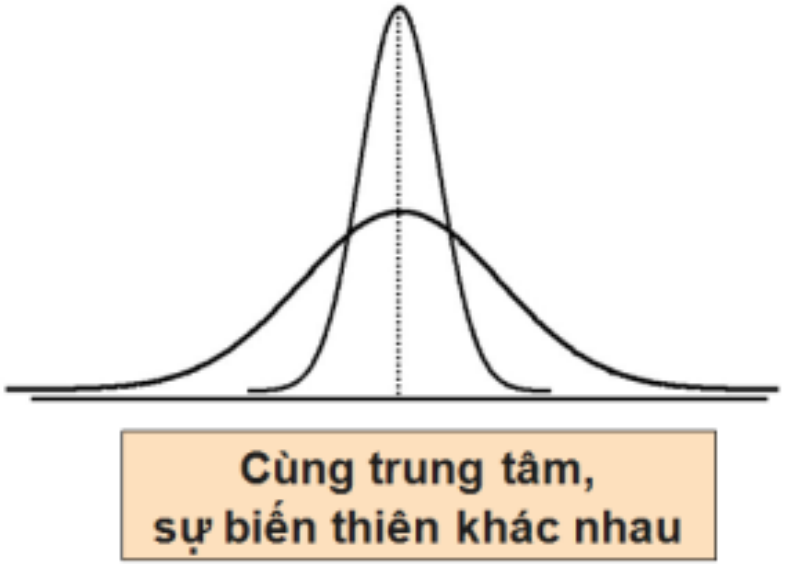
Độ đo sự biến thiên



Độ đo sự biến thiên



Độ đo sự biến thiên cho biết thông tin về độ phân tán hay sự biến thiên của dữ liệu.



Miền giá trị

- Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.

Miền giá trị

- Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.
- Là độ chênh lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc

$$\text{Miền giá trị (Range)} = x_{\max} - x_{\min}$$

Miền giá trị

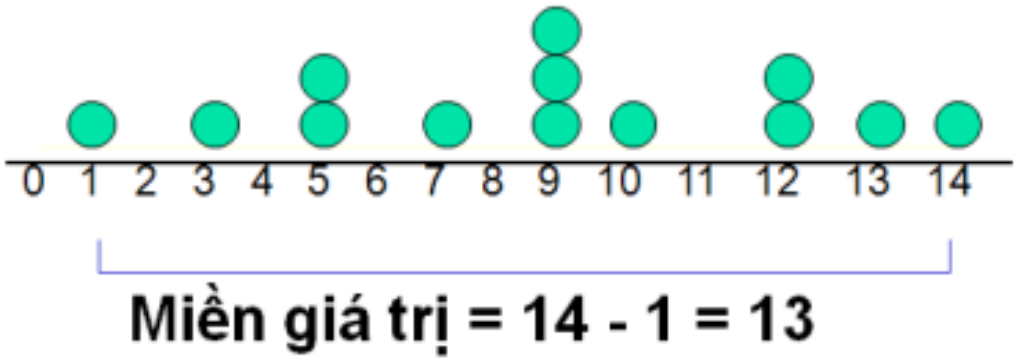
- Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.
- Là độ chênh lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc

$$\text{Miền giá trị (Range)} = x_{\max} - x_{\min}$$

Miền giá trị

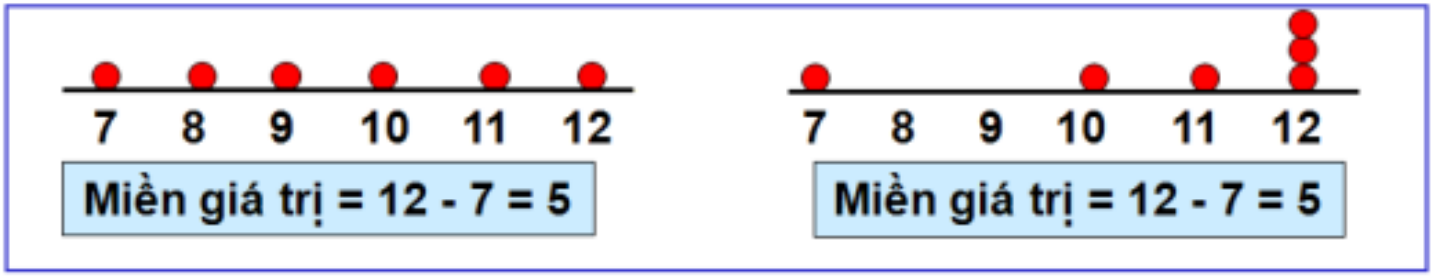
- Miền giá trị (range) là độ đo sự biến thiên đơn giản nhất.
- Là độ chênh lệch giữa giá trị lớn nhất và bé nhất của dữ liệu quan trắc

$$\text{Miền giá trị (Range)} = x_{\max} - x_{\min}$$



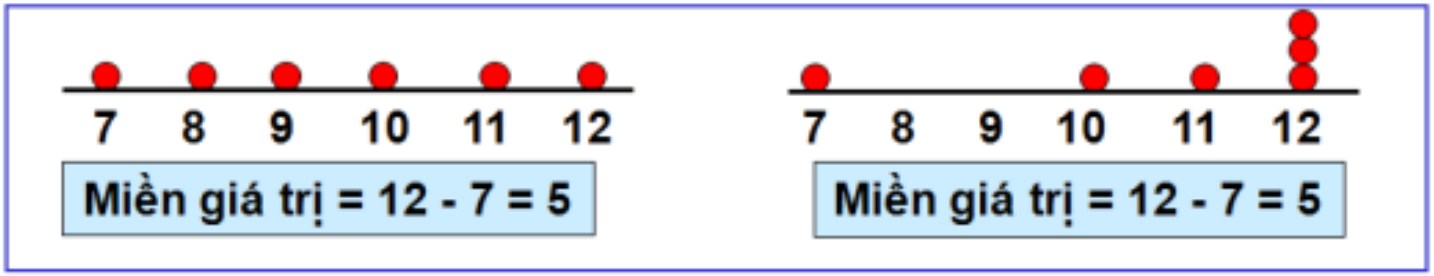
Nhược điểm của miền giá trị

- Bỏ qua phân bố của dữ liệu



Nhược điểm của miền giá trị

- Bỏ qua phân bố của dữ liệu



- Bị ảnh hưởng bởi các điểm ngoại lai

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

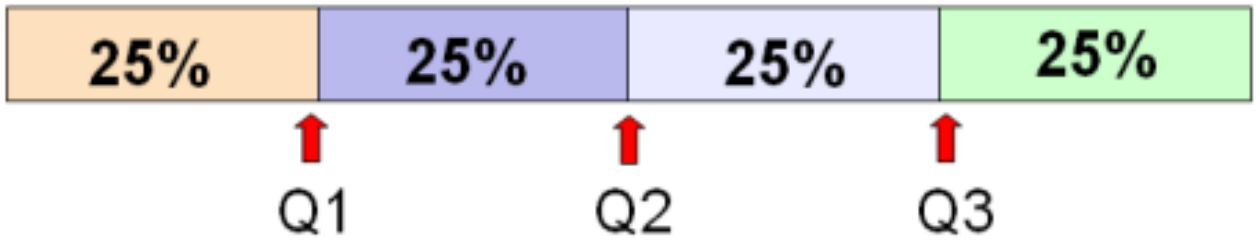
Miền giá trị = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Miền giá trị = 120 - 1 = 119

Miền phân vị

- Ta có thể loại bỏ các điểm outliers bằng cách sử dụng miền phân vị (interquartile range).

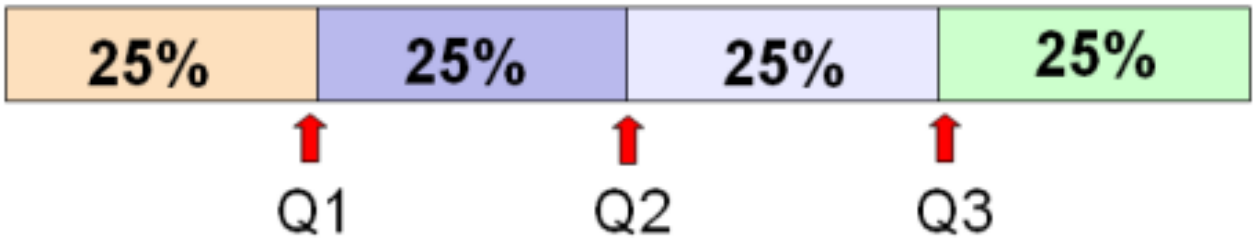


Miền phân vị

- Ta có thể loại bỏ các điểm outliers bằng cách sử dụng miền phân vị (interquartile range).
- Công thức tính miền phân vị:

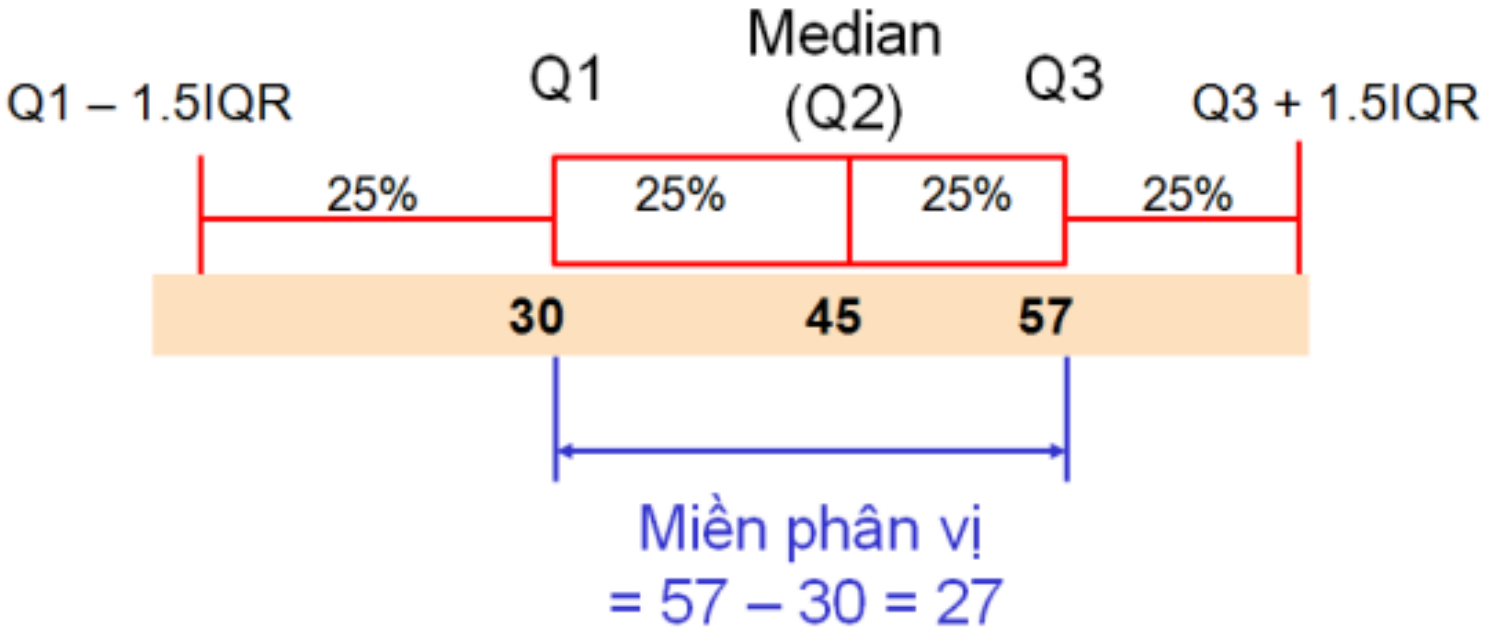
$$IQR = Q_3 - Q_1 \tag{3}$$

với Q_1 là phân vị thứ 1 (mức 25%) và Q_3 là phân vị thứ 3 (mức 75%) của dữ liệu.



Đồ thị Boxplot

Để biểu diễn miền phân vị và các điểm outliers ta dùng đồ thị boxplot.



Công thức tìm phân vị

Sắp xếp dữ liệu theo thứ tự tăng dần, gọi Q_1, Q_2 (trung vị), Q_3 lần lượt là phân vị thứ 1,2 và 3 của dữ liệu; vị trí của Q_1, Q_2 và Q_3 được xác định như sau:

$$\begin{aligned} \text{Vị trí} Q_1 &= 0.25(n+1) \\ \text{Vị trí } Q_2 &= 0.5(n+1) \\ \text{Vị trí } Q_3 &= 0.75(n+1) \end{aligned}$$

với n là số giá trị quan trắc.

Phương sai

- Phương sai (Variance) là trung bình của bình phương độ lệch các giá trị so với trung bình.

Phương sai

- Phương sai (Variance) là trung bình của bình phương độ lệch các giá trị so với trung bình.
- Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.

Phương sai

- Phương sai (Variance) là trung bình của bình phương độ lệch các giá trị so với trung bình.
- Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.
- Phương sai tổng thể

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{4}$$

với N là số phần tử của tổng thể, μ là trung bình tổng thể, x_i là giá trị thứ i của biến x.

Phương sai

- Phương sai (Variance) là trung bình của bình phương độ lệch các giá trị so với trung bình.
- Phương sai phản ánh độ phân tán hay sự biến thiên của dữ liệu.
- Phương sai tổng thể

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{4}$$

với N là số phần tử của tổng thể, μ là trung bình tổng thể, x_i là giá trị thứ i của biến x.

- Phương sai mẫu

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \tag{5}$$

với $\bar{X}$ là trung bình mẫu, n là cỡ mẫu, X_i là giá trị thứ i của biến X.

Độ lệch tiêu chuẩn

- Sử dụng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình.

Độ lệch tiêu chuẩn

- Sử dụng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình.
- Có cùng đơn vị đo với dữ liệu gốc.

Độ lệch tiêu chuẩn

- Sử dụng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình.
- Có cùng đơn vị đo với dữ liệu gốc.
- Độ lệch chuẩn của tổng thể, ký hiệu là σ :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{6}$$

Độ lệch tiêu chuẩn

- Sử dụng để đo sự biến thiên, biểu diễn sự biến thiên xung quanh trung bình.
- Có cùng đơn vị đo với dữ liệu gốc.
- Độ lệch chuẩn của tổng thể, ký hiệu là σ :

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{6}$$

- Độ lệch chuẩn của mẫu:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \tag{7}$$

Ví dụ độ lệch tiêu chuẩn

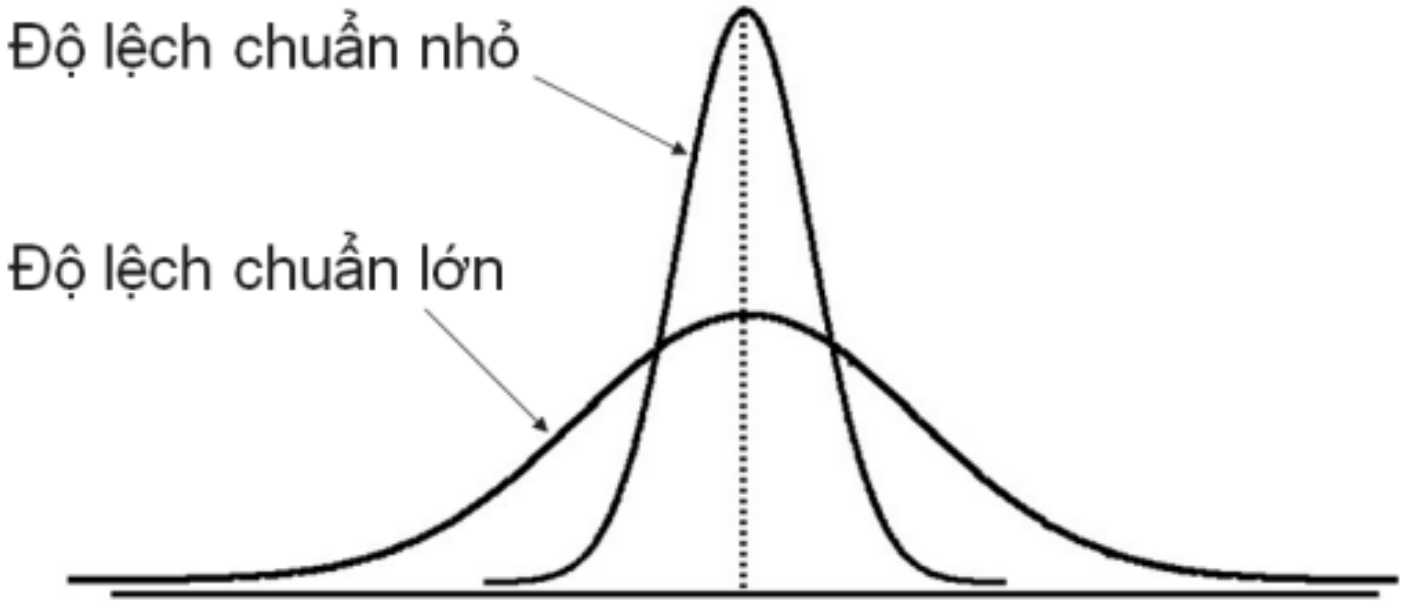
Dữ liệu mẫu x_i :

10 12 14 15 17 18 18 24

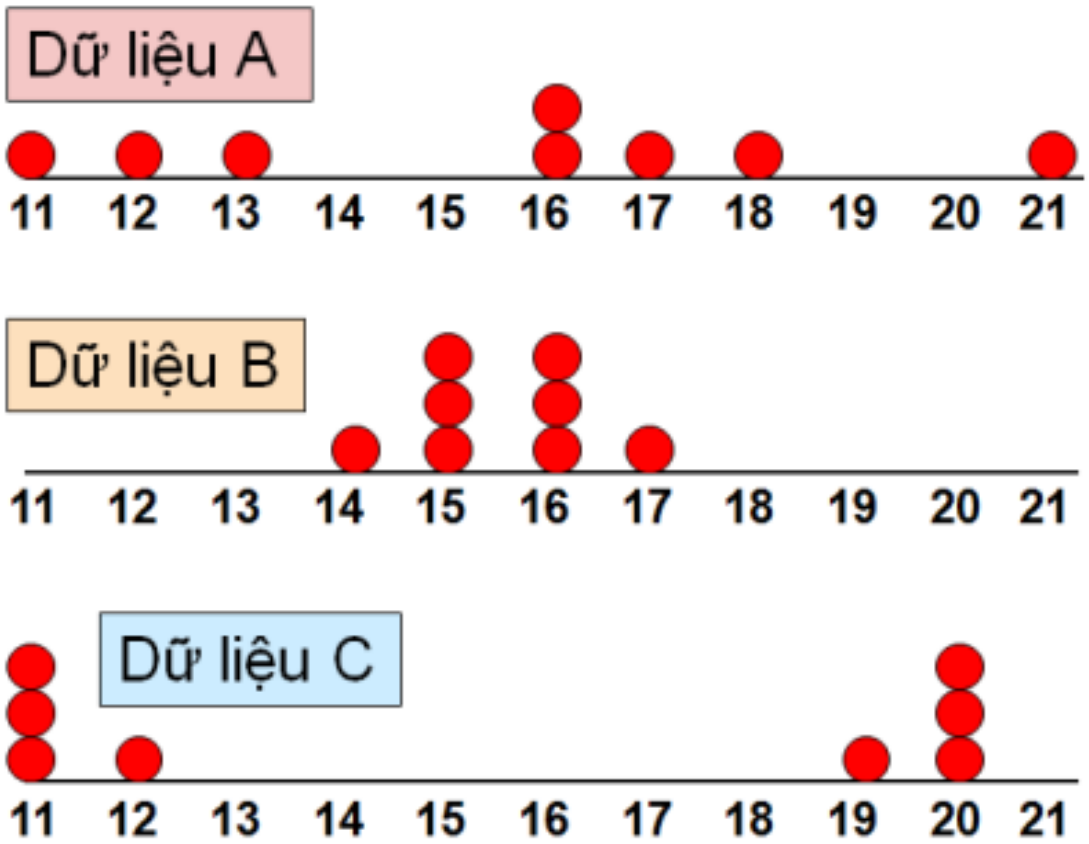
$$n = 8 \qquad \bar{x} = 16$$

$$\begin{aligned} s &= \sqrt{\frac{(10-\bar{x})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \cdots + (24-\bar{x})^2}{n-1}} \\ &= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}} \\ &= \sqrt{\frac{126}{7}} = 4.2426 \end{aligned}$$

Độ đo sự biến thiên



So sánh các độ lệch chuẩn



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.926

Mean = 15.5
s = 4.570

Hệ số biến thiên

- Hệ số biến thiên (Coefficient of Variation) được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.

Hệ số biến thiên

- Hệ số biến thiên (Coefficient of Variation) được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.
- Đo mối liên hệ giữa sự biến thiên và trung bình.

Hệ số biến thiên

- Hệ số biến thiên (Coefficient of Variation) được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.
- Đo mối liên hệ giữa sự biến thiên và trung bình.
- Đơn vị tính bằng %.

Hệ số biến thiên

- Hệ số biến thiên (Coefficient of Variation) được sử dụng để so sánh sự biến thiên của hai hay nhiều tập dữ liệu, có thể đo ở các đơn vị khác nhau.
- Đo mối liên hệ giữa sự biến thiên và trung bình.
- Đơn vị tính bằng %.
- Công thức

$$CV = \frac{S}{\bar{X}} 100\% \tag{8}$$

So sánh hệ số biến thiên

- Cổ phiếu A :

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

$$CV_A = \frac{S_A}{\bar{x}_A} 100\% = \frac{5}{50} 100\% = 10\%$$

- Cổ phiếu B :

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

$$CV_A = \frac{S_A}{\bar{x}_A} 100\% = \frac{5}{50} 100\% = 10\%$$

- Cổ phiếu B :
  - Giá trung bình $\bar{x}_B = \$100$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

$$CV_A = \frac{S_A}{\bar{x}_A} 100\% = \frac{5}{50} 100\% = 10\%$$

- Cổ phiếu B :
  - Giá trung bình $\bar{x}_B = \$100$
  - Độ lệch chuẩn $s_B = \$5$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

$$CV_A = \frac{S_A}{\bar{x}_A} 100\% = \frac{5}{50} 100\% = 10\%$$

- Cổ phiếu B :
  - Giá trung bình $\bar{x}_B = \$100$
  - Độ lệch chuẩn $s_B = \$5$

So sánh hệ số biến thiên

- Cổ phiếu A :
  - Giá trung bình $\bar{x}_A = \$50$
  - Độ lệch chuẩn $s_A = \$5$

$$CV_A = \frac{S_A}{\bar{x}_A} 100\% = \frac{5}{50} 100\% = 10\%$$

- Cổ phiếu B :
  - Giá trung bình $\bar{x}_B = \$100$
  - Độ lệch chuẩn $s_B = \$5$

$$CV_B = \frac{S_B}{\bar{x}_B} 100\% = \frac{5}{100} 100\% = 5\%$$

- Cả hai tập dữ liệu có cùng độ lệch chuẩn, nhưng dữ liệu B biến thiên ít hơn so với giá trị của nó.

Ví dụ 3

Khảo sát chiều cau của 15 sinh viên trong một lớp học:

160 165 155 162 167 145 158 170 165 155
158 160 170 175 169

a) *Vẽ đồ thị Stem-and-Leaf.*

b) *Tính các thống kê mẫu.*

Ví dụ 3

Khảo sát chiều cau của 15 sinh viên trong một lớp học:

160 165 155 162 167 145 158 170 165 155
158 160 170 175 169

a) *Vẽ đồ thị Stem-and-Leaf.*

b) *Tính các thống kê mẫu.*

Ví dụ 4

Thời gian tự học của 90 SV của một trường đại học cho bởi bảng sau:

Thời gian tự học	*1*	*2*	*3*	*4*	*5*	*6*
Số SV	*7*	*8*	*17*	*24*	*20*	*14*

Tính các thống kê mẫu.

Định nghĩa 1 (Chi-squared distribution)

Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(0, +\infty)$ được gọi là có phân phối chi bình phương với n bậc tự do, ký hiệu $X \sim \chi^2(n)$,, nếu hàm mật độ xác suất có dạng

$$f(x) = \begin{cases} 0, & \text{với } x \leq 0, \\ \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, & \text{với } x > 0, \end{cases}$$

trong đó $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ là hàm Gamma.

Định nghĩa 1 (Chi-squared distribution)

Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(0, +\infty)$ được gọi là có phân phối chi bình phương với n bậc tự do, ký hiệu $X \sim \chi^2(n)$,, nếu hàm mật độ xác suất có dạng

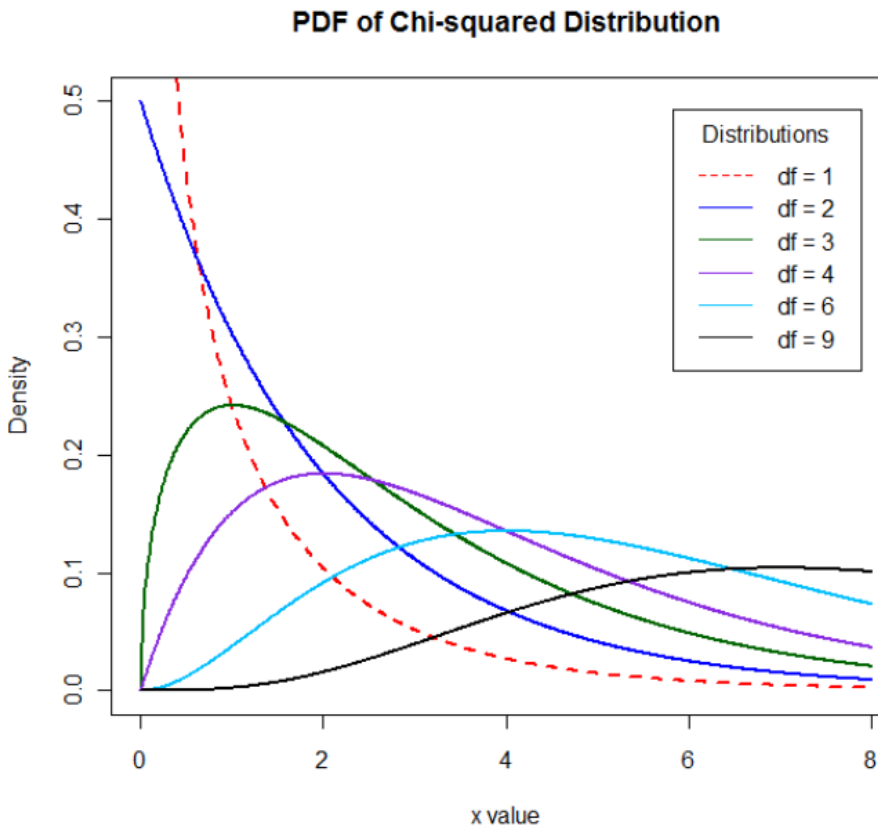
$$f(x) = \begin{cases} 0, & \text{với } x \leq 0, \\ \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}, & \text{với } x > 0, \end{cases}$$

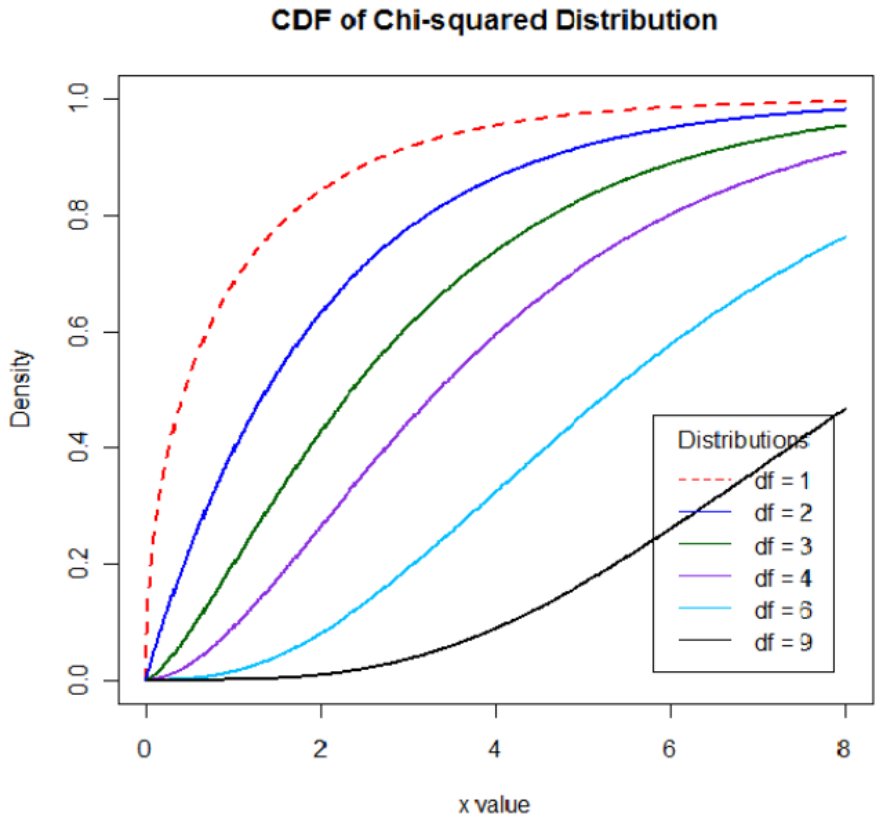
trong đó $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ là hàm Gamma.

Chú ý 1

Để thấy phân phối chi bình phương xuất phát từ phân phối chuẩn người ta con định nghĩa $X \sim \chi^2(n)$ nếu $X = \sum_{i=1}^n X_i^2$ với X_i là các biến ngẫu nhiên độc lập và $X_i \sim N(0, 1)$.

Phân phối chi bình phương





Phân phối chi bình phương

Định lý 1 (Các đặc trưng của biến ngẫu nhiên có phân phối chi bình phương)

Cho X là biến ngẫu nhiên có phân phối chi bình phương với n bậc tự do thì

- *Kỳ vọng* $E(X) = n$,

Phân phối chi bình phương

Định lý 1 (Các đặc trưng của biến ngẫu nhiên có phân phối chi bình phương)

Cho X là biến ngẫu nhiên có phân phối chi bình phương với n bậc tự do thì

- *Kỳ vọng* $E(X) = n$,
- *Phương sai* $Var(X) = 2n$,

Phân phối chi bình phương

Định lý 1 (Các đặc trưng của biến ngẫu nhiên có phân phối chi bình phương)

Cho X là biến ngẫu nhiên có phân phối chi bình phương với n bậc tự do thì

- *Kỳ vọng $E(X) = n$,*
- *Phương sai $Var(X) = 2n$,*
- *Nếu $X \sim \chi^2(n), Y \sim \chi^2(m)$ và X, Y là hai biến ngẫu nhiên độc lập thì $X + Y \sim \chi^2(m + n)$.*

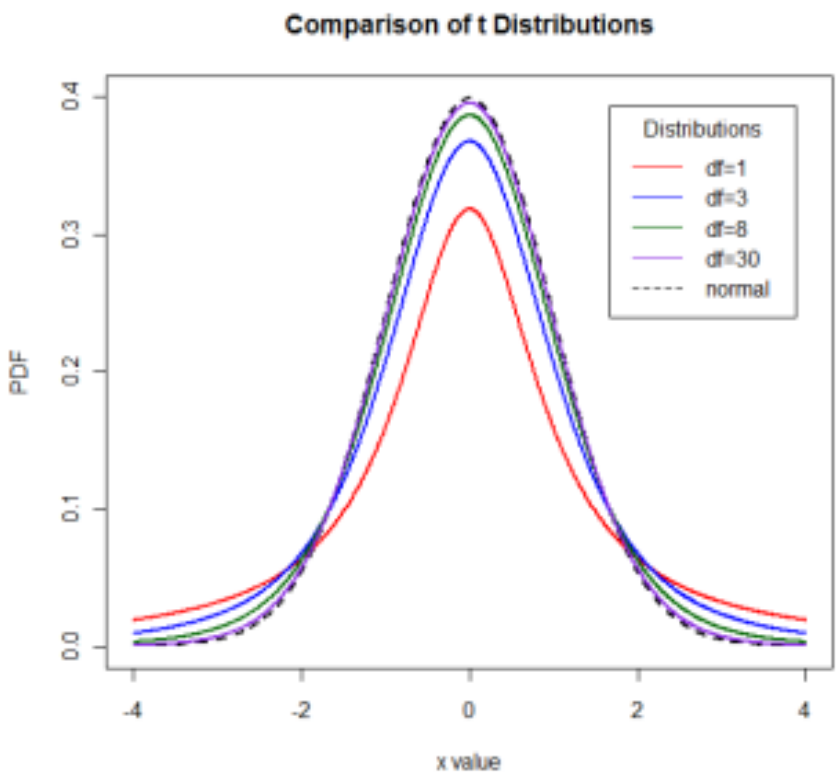
Định nghĩa 2 (Student distribution)

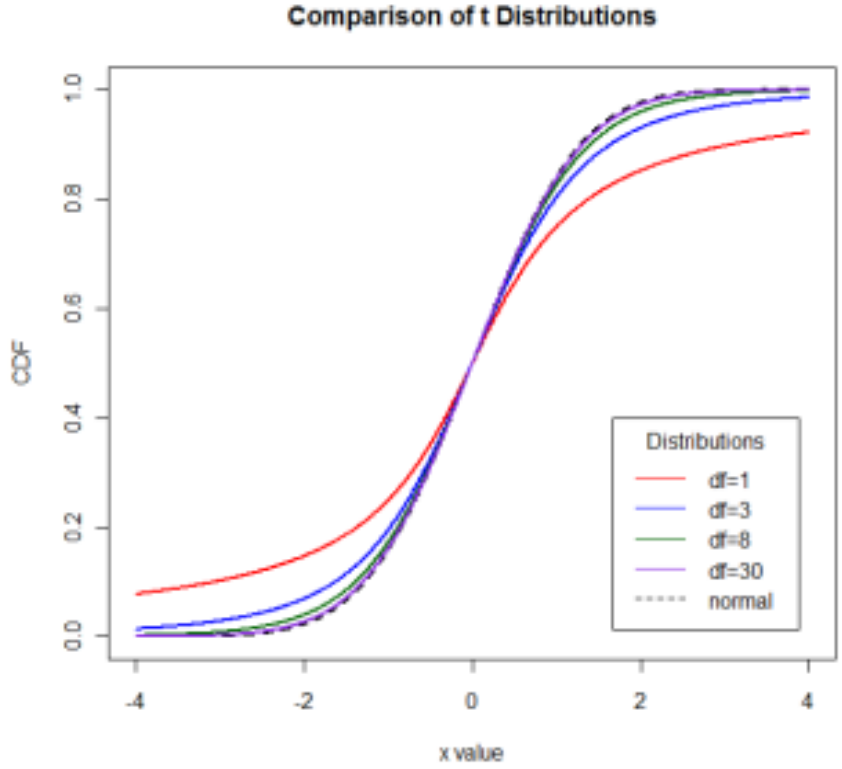
Biến ngẫu nhiên liên tục X nhận giá trị trong khoảng $(-\infty, +\infty)$ được gọi là có phân phối Student với n bậc tự do, ký hiệu: $X \sim t(n)$, nếu hàm mật độ xác suất có dạng

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}},$$

trong đó $\Gamma(x)$ là hàm Gamma.

Phân phối Student





Chú ý 2

- *Đồ thị của hàm mật độ của phân phối Student co hình dạng hình chuông như đồ thị hàm mật độ của phân phối chuẩn, nhưng có phần đỉnh thấp hơn và hai phần đuôi cao hơn so với phân phối chuẩn.*

Chú ý 2

- *Đồ thị của hàm mật độ của phân phối Student co hình dạng hình chuông như đồ thị hàm mật độ của phân phối chuẩn, nhưng có phần đỉnh thấp hơn và hai phần đuôi cao hơn so với phân phối chuẩn.*
- *Để thấy phân phối Student xuất phát từ phân phối chuẩn và phân phối* $\chi^2(n)$ *người ta còn định nghĩa* $X \sim t(n)$ *nếu* $X = \frac{Z}{\sqrt{\frac{Y}{n}}}$ *với* $Z \sim N(0,1), Y \sim \chi^2(n)$ *và* Z, Y *là các biến ngẫu nhiên độc lập.*

Định lý 2 (Các đặc trưng của biến ngẫu nhiên có phân phối Student)

Cho $X \sim t(n)$, *ta có*

- *Kỳ vọng* $E(X) = 0$ *nếu* $n > 1$, *các trường hợp còn lại* $E(X)$ *không được định nghĩa.*

Định lý 2 (Các đặc trưng của biến ngẫu nhiên có phân phối Student)

Cho $X \sim t(n)$, ta có

- *Kỳ vọng $E(X) = 0$ nếu $n > 1$, các trường hợp còn lại $E(X)$ không được định nghĩa.*
- *Phương sai $Var(X) = \frac{n}{n-2}$ nếu $n > 2$, $Var(X) = \infty$ nếu $1 < n \leq 2$, các trường hợp còn lại $Var(X)$ không được định nghĩa.*

Định nghĩa 3

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể và hàm giá trị thực (hay vector) $T(x_1, x_2, \ldots, x_n)$. Thì biến ngẫu nhiên hay vector ngẫu nhiên $Y = T(X_1, X_2, \ldots, X_n)$ được coi là một thống kê. Phân phối xác suất của thống kê Y được gọi là phân phối mẫu của Y.

Định nghĩa 3

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể và hàm giá trị thực (hay vector) $T(x_1, x_2, \ldots, x_n)$. Thì biến ngẫu nhiên hay vector ngẫu nhiên $Y = T(X_1, X_2, \ldots, X_n)$ được coi là một thống kê. Phân phối xác suất của thống kê Y được gọi là phân phối mẫu của Y.

Những phân phối mẫu được khảo sát:

- ★ Phân phối mẫu của trung bình.

Định nghĩa 3

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể và hàm giá trị thực (hay vector) $T(x_1, x_2, \ldots, x_n)$. Thì biến ngẫu nhiên hay vector ngẫu nhiên $Y = T(X_1, X_2, \ldots, X_n)$ được coi là một thống kê. Phân phối xác suất của thống kê Y được gọi là phân phối mẫu của Y.

Những phân phối mẫu được khảo sát:

- ★ Phân phối mẫu của trung bình.
- ★ Phân phối mẫu của phương sai.

Định nghĩa 3

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ tổng thể và hàm giá trị thực (hay vector) $T(x_1, x_2, \ldots, x_n)$. Thì biến ngẫu nhiên hay vector ngẫu nhiên $Y = T(X_1, X_2, \ldots, X_n)$ được coi là một thống kê. Phân phối xác suất của thống kê Y được gọi là phân phối mẫu của Y.

Những phân phối mẫu được khảo sát:

- ★ Phân phối mẫu của trung bình.
- ★ Phân phối mẫu của phương sai.
- ★ Phân phối mẫu của tỷ lệ.

Định nghĩa 4

Trung bình mẫu là trung bình số học của các giá trị trong một mẫu ngẫu nhiên, cho bởi công thức

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{9}$$

Định nghĩa 4

Trung bình mẫu là trung bình số học của các giá trị trong một mẫu ngẫu nhiên, cho bởi công thức

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{9}$$

Phương sai mẫu là thống kê cho bởi công thức:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{10}$$

Định lý 3

Nếu tổng thể X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ và $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên từ tổng thể trên thì

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Định lý 3

Nếu tổng thể X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ và $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên từ tổng thể trên thì

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- $\frac{(n-1)}{\sigma^2} S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.

Định lý 3

Nếu tổng thể X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ và $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên từ tổng thể trên thì

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- $\frac{(n-1)}{\sigma^2} S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.
- $\frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$.

Định lý 3

Nếu tổng thể X có phân phối chuẩn $X \sim N(\mu, \sigma^2)$ và $(X_1, \ldots, X_n)$ là một mẫu ngẫu nhiên từ tổng thể trên thì

- $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
- $\frac{(n-1)}{\sigma^2} S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1)$.
- $\frac{(\bar{X} - \mu)\sqrt{n}}{S} \sim t(n-1)$.
- $\bar{X}$ *và* S^2 *là hai biến ngẫu nhiên độc lập.*

Phân phối mẫu của trung bình và phương sai

Trong trường hợp tổng thể không có phân phối chuẩn, từ định lý giới hạn trung tâm ta suy ra rằng

$$\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \xrightarrow{\mathrm{D}} N(0,1)$$

$$\frac{(\bar{X}-\mu)\sqrt{n}}{S} \xrightarrow{\mathrm{D}} N(0,1)$$

Phân phối mẫu của trung bình và phương sai

Trong trường hợp tổng thể không có phân phối chuẩn, từ định lý giới hạn trung tâm ta suy ra rằng

$$\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \xrightarrow{\mathrm{D}} N(0,1)$$

$$\frac{(\bar{X}-\mu)\sqrt{n}}{S} \xrightarrow{\mathrm{D}} N(0,1)$$

Từ kết quả này, trong thực hành, khi mẫu có kích thước, n, đủ lớn ta có các phân phối xấp xỉ chuẩn sau

$$\frac{(\bar{X}-\mu)\sqrt{n}}{\sigma} \approx N(0,1)$$

$$\frac{(\bar{X}-\mu)\sqrt{n}}{S} \approx N(0,1)$$

Sai số chuẩn của trung bình

Định nghĩa 5

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ một tổng thể có trung bình μ và phương sai $\sigma^2 < \infty$. Sai số chuẩn (Standard Error) của trung bình, ký hiệu $\sigma_{\bar{X}}$ được định nghĩa như sau

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{11}$$

Sai số chuẩn của trung bình

Định nghĩa 5

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ một tổng thể có trung bình μ và phương sai $\sigma^2 < \infty$. Sai số chuẩn (Standard Error) của trung bình, ký hiệu $\sigma_{\bar{X}}$ được định nghĩa như sau

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{11}$$

Ý nghĩa:

- $\sigma_{\bar{X}}$ đo độ biến thiên của $\bar{X}$ xung quanh μ.

Sai số chuẩn của trung bình

Định nghĩa 5

Xét $X_1, X_2, \ldots, X_n$ là một mẫu ngẫu nhiên chọn từ một tổng thể có trung bình μ và phương sai $\sigma^2 < \infty$. Sai số chuẩn (Standard Error) của trung bình, ký hiệu $\sigma_{\bar{X}}$ được định nghĩa như sau

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{11}$$

Ý nghĩa:

- $\sigma_{\bar{X}}$ đo độ biến thiên của $\bar{X}$ xung quanh μ.
- Sai số chuẩn càng nhỏ, ước lượng tham số từ tổng thể càng tốt và độ tin cậy cao.

Sai số chuẩn của trung bình

$\sigma_{\bar{X}}$ bị ảnh hưởng bởi hai yếu tố:

- Cỡ mẫu n : Cỡ mẫu càng lớn $\Rightarrow$ sai số chuẩn càng nhỏ, chú ý rừng khi $n = 1$ thì $\sigma_{\bar{X}} = \sigma$.

Sai số chuẩn của trung bình

$\sigma_{\bar{X}}$ bị ảnh hưởng bởi hai yếu tố:

- Cỡ mẫu n : Cỡ mẫu càng lớn $\Rightarrow$ sai số chuẩn càng nhỏ, chú ý rừng khi $n = 1$ thì $\sigma_{\bar{X}} = \sigma$.
- Độ biến thiên của tổng thể σ : σ càng lớn $\Rightarrow$ sai số chuẩn càng lớn.

Phân phối mẫu của tỷ lệ

- Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của một tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1, & \text{nếu thoả } \mathcal{A} \\ 0, & \text{nếu không thoả } \mathcal{A} \end{cases}$$

thu được mẫu ngẫu nhiên $X_1, \ldots, X_n$ với $X_i \sim B(p)$, p là tỷ lệ phần tử thoả đặc trưng $\mathcal{A}$.

Phân phối mẫu của tỷ lệ

- Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của một tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1, & \text{nếu thoả } \mathcal{A} \\ 0, & \text{nếu không thoả } \mathcal{A} \end{cases}$$

  thu được mẫu ngẫu nhiên $X_1, \ldots, X_n$ với $X_i \sim B(p)$, p là tỷ lệ phần tử thoả đặc trưng $\mathcal{A}$.
- Đặt $X = \sum_{i=1}^{n} X_i$ là số phần tử thoả đặc trưng $\mathcal{A}$ trong mẫu khảo sát, thì $X \sim B(n, p)$.

Phân phối mẫu của tỷ lệ

- Giả sử cần khảo sát đặc trưng $\mathcal{A}$ của một tổng thể, khảo sát n phần tử và đặt

$$X_i = \begin{cases} 1, & \text{nếu thoả } \mathcal{A} \\ 0, & \text{nếu không thoả } \mathcal{A} \end{cases}$$

  thu được mẫu ngẫu nhiên $X_1, \ldots, X_n$ với $X_i \sim B(p)$, p là tỷ lệ phần tử thoả đặc trưng $\mathcal{A}$.
- Đặt $X = \sum_{i=1}^{n} X_i$ là số phần tử thoả đặc trưng $\mathcal{A}$ trong mẫu khảo sát, thì $X \sim B(n, p)$.
- Tỷ lệ mẫu $\hat{P}$ là một ước lượng của tỷ lệ p xác định bởi

$$\hat{P} = \frac{X}{n} \tag{12}$$

Phân phối mẫu của tỷ lệ

- Kỳ vọng và phương sai của $\hat{P}$ bằng

$$E(\hat{P}) = p;\ Var(\hat{P}) = \frac{p(1-p)}{n}$$

Phân phối mẫu của tỷ lệ

- Kỳ vọng và phương sai của $\hat{P}$ bằng

$$E(\hat{P}) = p;\ Var(\hat{P}) = \frac{p(1-p)}{n}$$

- Theo định lý giới hạn trung tâm ta có

$$\frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{p}}} \rightsquigarrow N(0,1)$$

Vì vậy trong thực hành, khi $np \geq 5, n(1-p) \geq 5$, ta có $\hat{P} \approx N\left(p, \frac{p(1-p)}{n}\right)$.