Thống kê dữ liệu

1. Các đặc tính của dữ liệu

1.1. Kiểu dữ liệu trong thực tế

Dữ liệu trong thực tế được chia thành nhiều kiểu dữ liệu khác nhau. Các kiểu dữ liệu này được nhóm thành hai nhóm chính: dữ liệu định tính (qualitative data) và dữ liệu định lượng (Quantitative data) hoặc dữ liệu số học (numerical data).

- Dữ liệu định tính là những kiểu dữ liệu không thể được biểu thị bằng giá trị số hoặc các toán tử, mà biểu thị bằng ngôn ngữ và các loại (category). Ví dụ: màu tóc, các loại trái cây, giới tính, quốc gia, etc. Ta có thể mã hoá các kiểu dữ liệu này dưới dạng các con số, tuy nhiên, về bản chất, đây vẫn là kiểu dữ liệu định tính.
- Dữ liệu định lượng là dữ liệu được biểu thị bằng giá trị số và các ký tự toán học. Ta có thể thực hiện các phép toán trên những kiểu dữ liệu này. Dữ liệu định lượng được chia làm hai nhóm con:
  - Dữ liệu rời rạc: Dữ liệu được biểu diễn dưới các giá trị là các số tự nhiên hoặc số nguyên.
  - Dữ liệu liên tục: Dữ liệu được biểu diễn dưới các giá trị là các số thực.

1.2. Hình dạng dữ liệu

Ta có ba thành phần cấu thành nên mỗi bảng dữ liệu (data table) trong cơ sở dữ liệu (database): cột, hàng và giá trị.

- Mỗi cột trong bảng dữ liệu chứa thông tin về các biến (variable) hoặc đặc trưng (feature) của dữ liệu.
- Mỗi hàng trong bảng dữ liệu chứa thông tin về các bản ghi dữ liệu (data record) hoặc mẫu dữ liệu (data sample).
- Mỗi giá trị trong bảng dữ liệu đại diện cho từng đặc trưng tương ứng của mỗi bản ghi. Các giá trị này được gọi là quan sát (observation).

Index	Name	Age	Team
1	Customer 1	20	team A
2	Customer 2	21	team A
3	Customer 3	22	team B
4	Customer 4	31	team B
5	Customer 5	30	team B
6	Customer 6	29	team C
7	Customer 7	28	team C
8	Customer 8	27	team D

2. Mẫu thống kê và thống kê mô tả

Dữ liệu là kết quả của việc đếm khi quan sát, của đo đạc nhờ các thiết bị đo ... và dữ liệu cần được thu thập, lưu trữ và phản tích, những công việc này được gọi là thống kê mô tả. Phần dữ liệu được thu thập và thống kê được gọi là tập mẫu (sample set) và nó có nguồn gốc từ một tập dữ liệu lớn hơn goi là tập thế giới (population). Tập mẫu sẽ mang thông tin nào đó về tập population, và các tập mẫu khác nhau có thể sẽ phản ánh những thông tin khác nhau của tập population.

Để phân tích dữ liệu được chính xác nhất, ta phải làm việc với tập population, nhưng trong thực tế, tập population thường quá lớn và đòi hỏi chi phí xử lý cao. Do đó, ta chỉ có thể làm việc với tập mẫu và kỳ vọng rằng tập mẫu có thể phản ảnh được hầu như toàn bộ bản chất của tập population. Điều này dẫn đến bài toán về việc chọn mẫu thống kê.

2.1. Vấn đề chọn mẫu thống kê

2.1.1. Phương pháp chọn mẫu ngẫu nhiên

Vì mỗi phần tử của tập population đã có xác suất được chọn được xác định từ trước cả khi chọn mẫu, và nó cũng chính là một khía cạnh của bản chất của tập population, nên chọn mẫu ngẫu nhiên cho phép đánh giá khách quan hơn các đặc trưng của tập population.

Một số cách để chọn mẫu ngẫu nhiên như:

- Chọn mẫu ngẫu nhiên đơn giản: Lựa chọn một cách hoàn toàn ngẫu nhiên, mọi phần tử của tập population có đồng khả năng lọt vào mẫu. Do tính ngẫu nhiên nên mẫu có tính đại diện cao và tin cậy. Tuy nhiên, phương pháp đòi hỏi phải biết toàn bộ tập population và vì thế chi phí chọn mẫu khá lớn.
- Chọn mẫu phân nhóm: Đầu tiên ta chia tập population thành các nhóm tương đối thuần nhất, sau đó từ mỗi nhóm trích ra một mẫu ngẫu nhiên; tập hợp tất cả các mẫu đó cho ta một mẫu ngẫu nhiên phân nhóm. Để phương pháp này hiệu quả, ta phải có hiểu biết nhất định về cấu trúc tập population để phân chia nhóm hợp lý. Sau này mỗi nhóm sẽ có vai trò khác nhau phụ thuộc vào độ quan trọng của chúng trong tập population. Hạn chế của phương pháp là tính chủ quan khi phân chia nhóm.

2.1.2. Phương pháp chọn mẫu có suy luận

Phương pháp chọn mẫu này dựa trên ý kiến các chuyên gia về đối tượng nghiên cứu và điều này kéo theo hạn chế về tính chủ quan của mẫu và chất lượng của mẫu phụ thuộc nhiều vào trình độ và kinh nghiệm của chuyên gia.

2.2. Các đặc trưng của mẫu thống kê

Với một mẫu dữ liệu gồm có k giá trị khác nhau $x_1, x_2, \ldots, x_k$, mỗi giá trị có tương ứng $n_1, n_2, \ldots, n_k$ phần tử và tổng số phần từ $n = n_1 + n_2 + \cdots + n_k$

2.2.1. Trung bình / kỳ vọng (Mean / expectation)

Trung bình của mẫu hay được gọi cách khác là kỳ vọng mẫu được ký hiệu là $\bar{X}$ và được tính bằng công thức sau:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} x_i n_i$$

Giá trị trung bình là một giá trị đặc trưng quan trọng của một mẫu dữ liệu. Giá trị trung bình thường được dùng làm đại diện cho tất cả các phần từ trong mẫu dữ liệu.

Tuy nhiên, giá trị trung bình trong một số trường hợp không thể đại diện được cho một mẫu dữ liệu, ví dụ như:

- Mẫu dữ liệu có một số ít các phần tử ngoại lai (outlier) có giá trị lớn.
- Mẫu dữ liệu có phân bố không đều.
- ...

2.2.2. Phương sai - độ lệch chuẩn (Variance - standard deviation)

Phương sai của mẫu được ký hiệu là S^2 và được tính bằng công thức sau:

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

Và ta cũng có độ lệch chuẩn (standard deviation - std) được tính bằng căn bậc hai của phương sai của mẫu và được ký hiệu là S.

Phương sai hoặc độ lệch chuẩn là giá trị thể hiện độ phân tán của dữ liệu. Đối với những bộ dữ liệu có phương sai nhỏ (bộ dữ liệu phân tán ít), ta có thể chỉ cần sử dụng một mẫu dữ liệu nhỏ để phân tích và đánh giá. Ngược lại, với những bộ dữ liệu có phương sai lớn (bộ dữ liệu phân tán nhiều), ta có thể cần phải sử dụng một mẫu dữ liệu lớn.

2.2.3. Giá trị lớn nhất, giá trị nhỏ nhất (Max, min)

Giá trị lớn nhất và giá trị nhỏ nhất là hai giá trị đặc trưng cơ bản khi thống kê dữ liệu. Hai giá trị này cung cấp thông tin về khoảng phân bố dữ liệu từ đó:

- Giúp ta có được cái nhìn sơ bộ về mẫu dữ liệu và đánh giá tổng quan về chất lượng của mẫu dữ liệu.
- Giúp lựa chọn tối ưu kiểu dữ liệu để lưu trữ mẫu dữ liệu.
- Kết hợp cùng các giá trị như trung bình, trung vị giúp đánh giá sâu hơn về phân bố của dữ liệu.
- ...

2.2.4. Trung vị mẫu (Median)

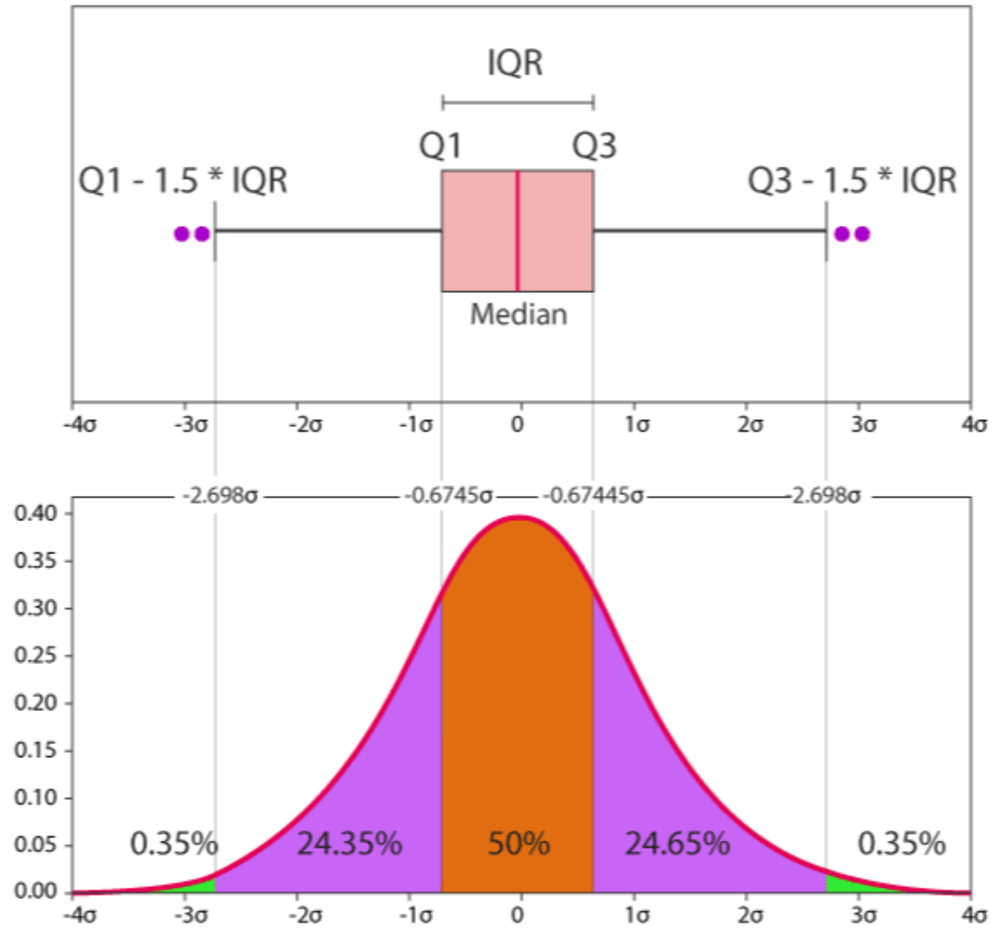
Trung vị của mẫu được ký hiệu là M_e và được tính bằng cách sắp xếp mẫu thành dãy theo thứ tự tăng dần hoặc giảm dần:

- Đối với mẫu có số lượng phần tử là số lẻ, trung vị của mẫu là số nằm ở đúng vị trí chính giữa của dãy.
- Đối với mẫu có số lượng phần tử là số chẵn, trung vị của mẫu đươc tính bằng trung bình cộng của hai số ở giữa của dãy.

Trung vị trong một số trường hợp cũng được sử dụng để làm đại diện cho tất cả các phần từ trong mẫu dữ liệu. Đặc biệt, trong trường hợp giá trị trung bình không thể đại diện tốt được cho một mẫu dữ liệu như mẫu dữ liệu có một số ít các phần tử ngoại lai có giá trị lớn.

Tuy nhiên, trong trường hợp mẫu dữ liệu có nhiều các phần tử ngoại lai có giá trị nhỏ, trung vị không thể đại diện được tốt cho mẫu dữ liệu.

Đi kèm với trung vị, ta có thêm giá trị đặc trưng tứ phân vị (quantiles): tứ phân vị thứ nhất (giá trị ở vị trí 25%), tứ phân vị thứ hai (giá trị ở vị trí 50% hay là trung vị), tứ phân vị thứ ba (giá trị ở vị trí 75%). Các giá trị này được sử dụng để xây dựng biểu đồ Box plot và Violin plot, được sử dụng nhiều trong quá trình trực quan hoá dữ liệu.



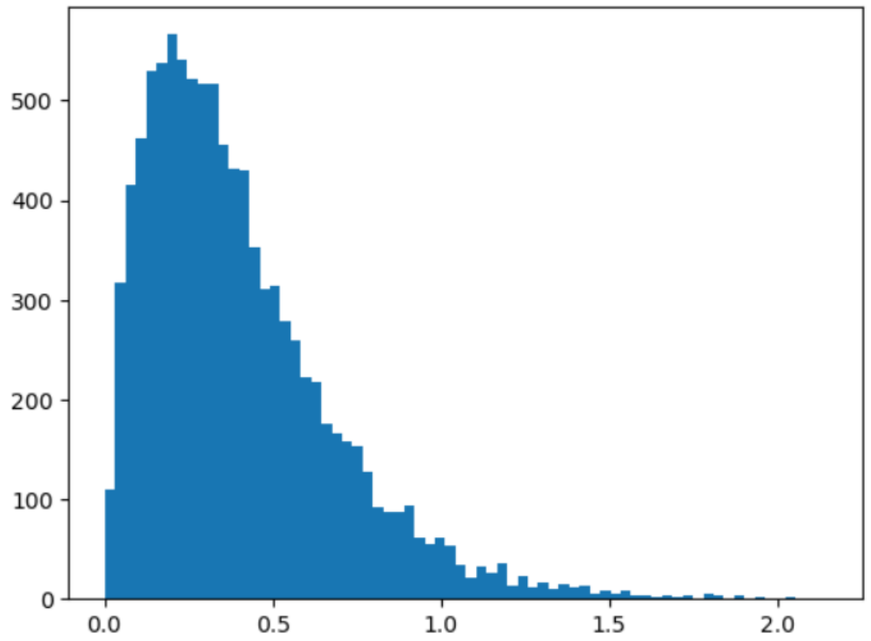
2.2.5. Mốt (Mode)

Mốt là giá trị có tần suất xuất hiện lớn nhất trong một lân cận nào đó của nó. Như vậy mốt có thể chỉ là cực đại địa phương hoặc cực đại toàn cục và một phân phối dữ liệu có thể có một mốt hoặc nhiều mốt.

3. Một số hình dạng của phân phối khi thống kê dữ liệu

3.1. Phân phối long tail - phân phối lệch (Long tail distribution - Skewed distribution)

Khi thống kê dữ liệu, ta có thể bắt gặp những trường hợp bộ dữ liệu có những giá trị lớn hoặc giá trị nhỏ xuất hiện rất ít lần. Với những bộ dữ liệu như vậy, việc vẽ phân phối tần suất xuất hiện sẽ xuất hiện trạng thái được gọi là long tail. Phân phối khi đó được gọi là phân phối long tail (long tail distribution).



Nếu như đối với phân phối long tail, ta tập trung góc nhìn vào phần "đuôi" của phân phối, thì phân phối lệch là một góc nhìn khác khi ta tập trung vào phần đỉnh của phân phối.

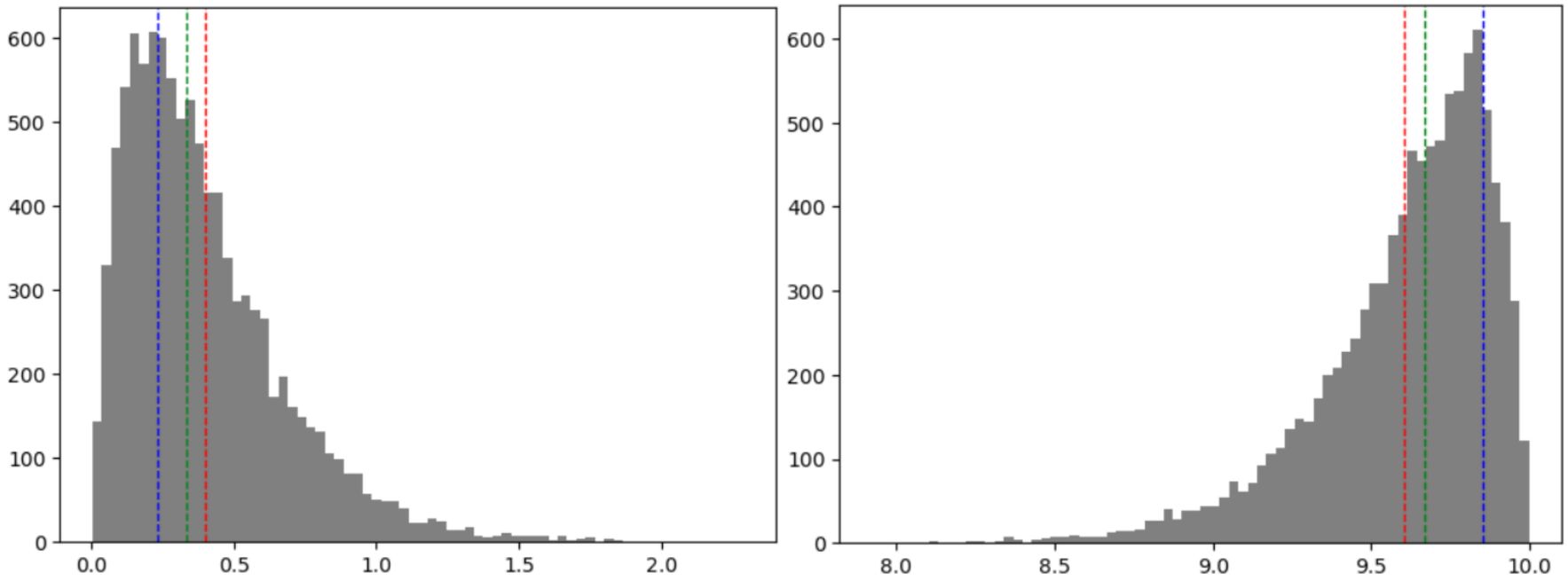
Độ lệch của phân phối được sử dụng để so sánh với phân phối chuẩn và được tính bằng công thức:

$$\text{skewness} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

Có ba loại độ lệch:

- skewness = 0, phân phối đối xứng và được xem như là phân phối chuẩn
- skewness < 0, phân phối lệch trái có độ lệch âm (negative skewness), khi đó, giá trị kỳ vọng < giá trị trung vị < giá trị mốt
- skewness > 0, phân phối lệch phải có độ lệch dương (positive skewness), khi đó, giá trị kỳ vọng > giá trị trung vị > giá trị mốt

Độ lệch được coi là đáng kể nếu độ lớn của giá trị tuyệt đối của nó lớn hơn 0.5.



3.2. Phân phối nhọn (Kurtosis distribution)

Độ nhọn của phân phối được sử dụng để so sánh với phân phối chuẩn và được tính bằng công thức:

$$\text{kurtosis} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

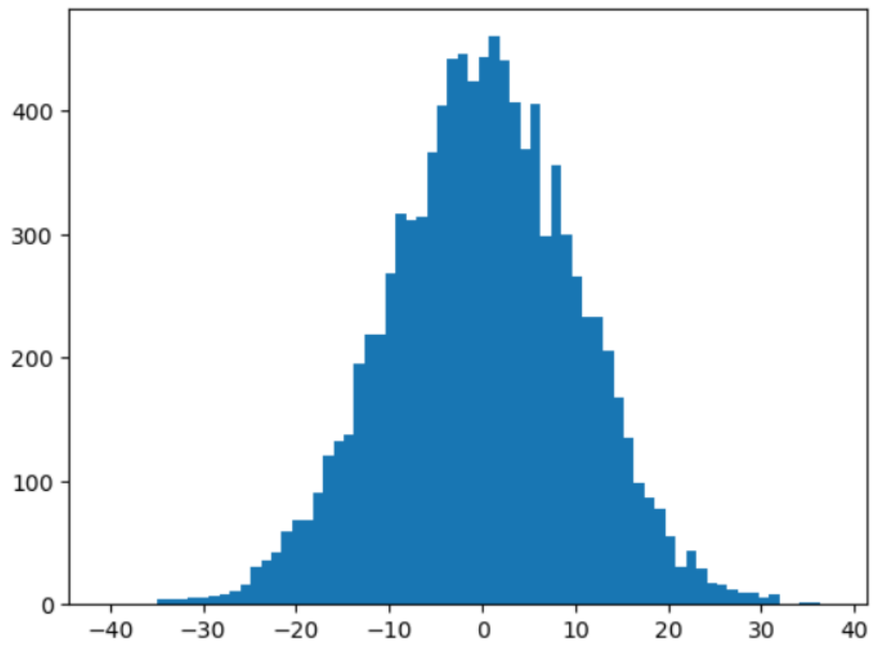
Đối với phân phối chuẩn có độ nhọn = 3, ta tính toán chỉ số

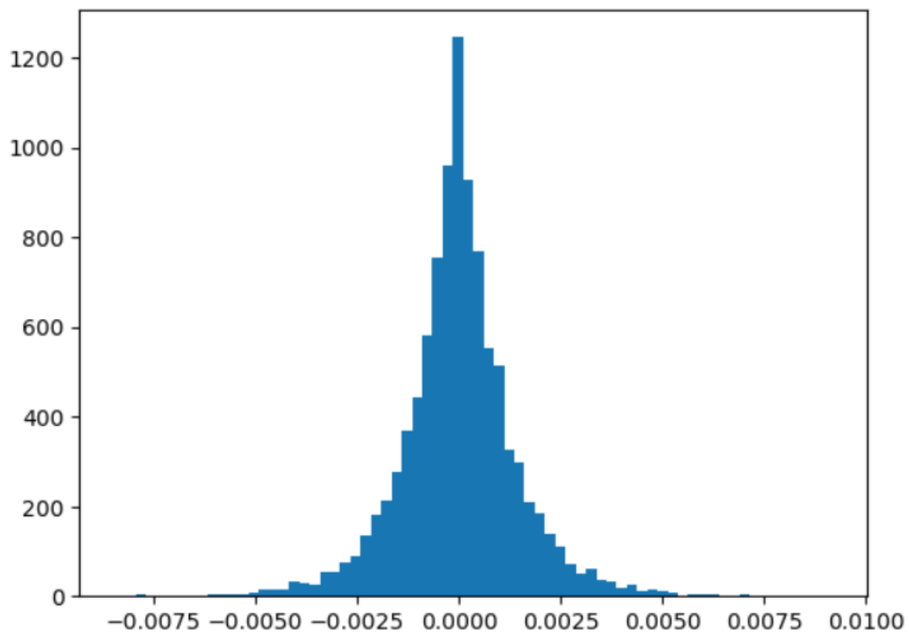
$$\text{excess kurtosis} = \text{kurtosis} - 3$$

Có ba loại độ nhọn:

- excess kurtosis = 0 được gọi là mesokurtic và được xem như là phân phối chuẩn
- excess kurtosis > 0 được gọi là leptokurtic, khi đó phân phối có dạng nhọn
- excess kurtosis < 0 được gọi là platykurtic, khi đó phân phối có dạng rộng

Độ nhọn được coi là đáng kể nếu độ lớn của giá trị tuyệt đối của nó lớn hơn 1.





3.3. Phân phối đa thức (Multimodal distribution)

Phân phối đa thức (multimodal distribution) là phân phối dữ liệu có nhiều hơn một đỉnh (một mốt). Phân phối có một đỉnh (một mốt) được gọi là unimodal. Phân phối có hai đỉnh (hai mốt) được gọi là bimodal Phân phối có nhiều đỉnh (nhiều mốt) được gọi là multimodal.