

Bài giảng 1: Xác suất và Thống kê cho Máy học và KHDL

TS. Tô Đức Khanh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Trường hè Toán học sinh viên năm 2025

Nội dung buổi học

1 *Giới thiệu về Khoa học thống kê*

2 *Xác suất và Biến ngẫu nhiên*

3 *Vector ngẫu nhiên*

1 *Giới thiệu về Khoa học thống kê*

2 *Xác suất và Biến ngẫu nhiên*

3 *Vector ngẫu nhiên*

Khoa học thống kê: Mô tả và Suy luận

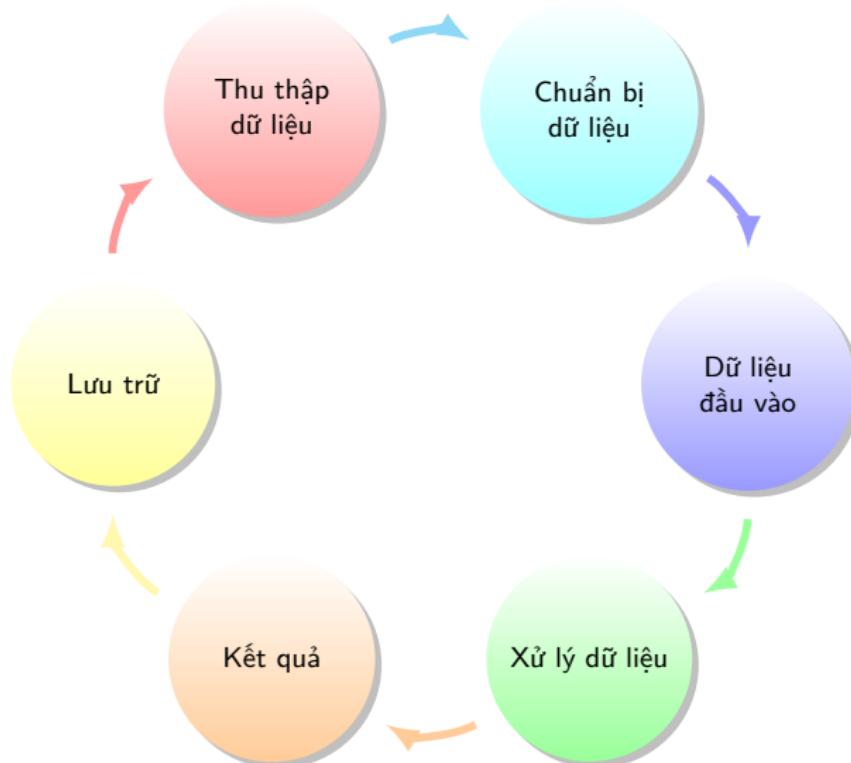
Khoa học thống kê - Statistical Science

Khoa học thống kê là khoa học nghiên cứu phát triển và áp dụng các phương pháp **thu thập, phân tích** và **diễn giải** dữ liệu. ([Agresti and Kateri, 2021](#))

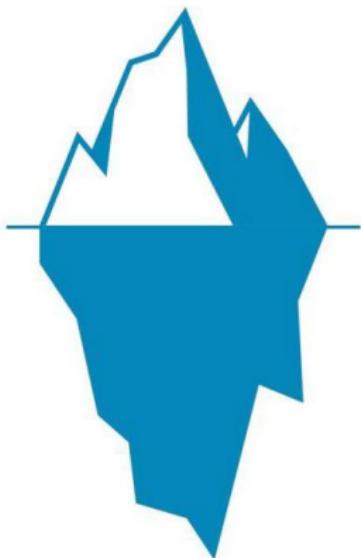
Cụ thể, có 3 khía cạnh chính:

- 1 Thiết kế:** lên kế hoạch thu thập dữ liệu có liên quan đến chủ đề quan tâm.
 - 2 Mô tả:** tổng hợp và tóm tắt dữ liệu.
 - 3 Suy luận:** đưa ra các đánh giá, chẳng hạn như ước tính và dự đoán, dựa trên dữ liệu.

Khoa học thống kê: Mô tả và Suy luận



Khoa học thống kê: Mô tả và Suy luận



Phân tích dữ liệu

Trực quan hóa dữ liệu

Thu thập dữ liệu

Statistical learning

Phương pháp tính

Lý thuyết Thống kê

Lý thuyết Xác suất

Toán học

Hồi quy Phân loại Phân cụm

Phương pháp ước lượng điểm Ước lượng khoảng Lý thuyết kiểm định

Biến ngẫu nhiên Phân phối xác suất Định lý giới hạn trung tâm

Đại số tuyến tính

Giải tích

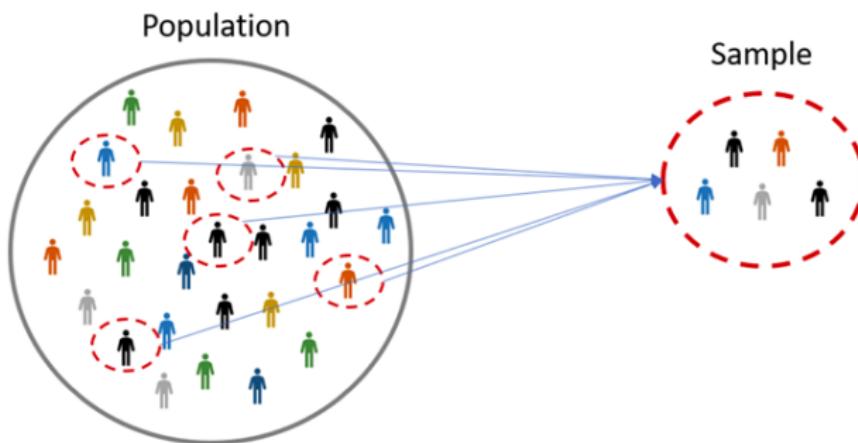
Ví tích phân

Dữ liệu thống kê là gì?

- Một dữ liệu thống kê là một bộ *mẫu ngẫu nhiên* (**sample**) của các *đối tượng nghiên cứu*, nó bao gồm các thông tin mô tả *tính chất* của các *đối tượng nghiên cứu*.
- Một bộ dữ liệu thống kê có thể được thu được thông qua việc quan sát và thu mẫu thực tế (*observations*), hoặc thu mẫu các đối tượng trong phòng thí nghiệm *experiments*.
- Một quần thể (**population**) là bộ dữ liệu lớn bao gồm tất cả đối tượng nghiên cứu có thể được bao hàm trong câu hỏi nghiên cứu.
- Một biến (**variable**) là 1 tính chất (đặc trưng) có thể thay đổi giá trị giữa các đối tượng trong một mẫu hoặc quần thể.

Chú ý: một đối tượng nghiên cứu được coi là một đơn vị thống kê (*statistical unit*).

Dữ liệu thống kê là gì?



Dữ liệu thống kê là gì?

Dữ liệu từ nghiên cứu của Nurit Tal-Or, Jonanathan Cohen, Yariv Tasfati và Albert Gunther (2010) về đánh giá tác động giả định của truyền thông đối với người khác và sự thay đổi trong thái độ.

	cond	pmi	import	reaction	gender	age
1	1	7.0	6	5.25	1	51.0
2	0	6.0	1	1.25	1	40.0
3	1	5.5	6	5.00	1	26.0
4	0	6.5	6	2.75	2	21.0
5	0	6.0	5	2.50	1	27.0
6	0	5.5	1	1.25	1	25.0
7	0	3.5	1	1.50	2	23.0
8	1	6.0	6	4.75	1	25.0
9	0	4.5	6	4.25	1	22.0
10	0	7.0	6	6.25	1	24.0
11	1	1.0	3	1.25	2	22.0
12	0	6.0	3	2.75	2	21.0
13	1	5.0	4	3.75	2	23.0

- cond: mức độ quan trọng của truyền thông (thấp: 0, cao: 1).
- pmi: mức độ ảnh hưởng giả định
- import: mức độ quan trọng của vấn đề.
- reaction: đối tượng đánh giá mức độ đồng ý về các phản ứng có thể có đối với câu chuyện.
- gender: giới tính (1: male, 2: female)
- age: độ tuổi

Tham số và ước lượng

Tham số - Parameters

Tham số (**Parameter**) là một số tổng hợp của một quần thể.

Trong các bài toán thống kê, ta thường quan tâm tới:

- μ : trung bình của quần thể,
- σ^2 : phương sai của quần thể,
- p : tỷ lệ của 1 đặc tính nào đó của quần thể,

và nhiều tham số khác.

Trong thực tế, các tham số là:

- không biết giá trị chính xác,
- và thường được giả định là 1 hằng số.

Tham số và ước lượng

Ước lượng - Estimator

Ước lượng (**Estimator**) là một giá trị đại diện cho tham số, và được xác định dựa trên 1 mẫu ngẫu nhiên (hay 1 bộ dữ liệu).

Trong các dữ liệu, ta thường quan tâm tới:

- \bar{x} : trung bình mẫu của 1 cột dữ liệu x ,
- $\hat{\sigma}_x^2$: phương sai mẫu của 1 cột dữ liệu x ,
- \hat{p} : tỷ lệ mẫu của 1 đặc tính nào đó của đối tượng quan sát,

và nhiều tham số khác.

Dữ liệu?

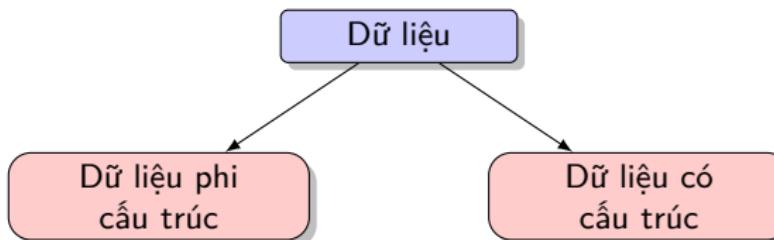
Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- âm thanh;
- video.

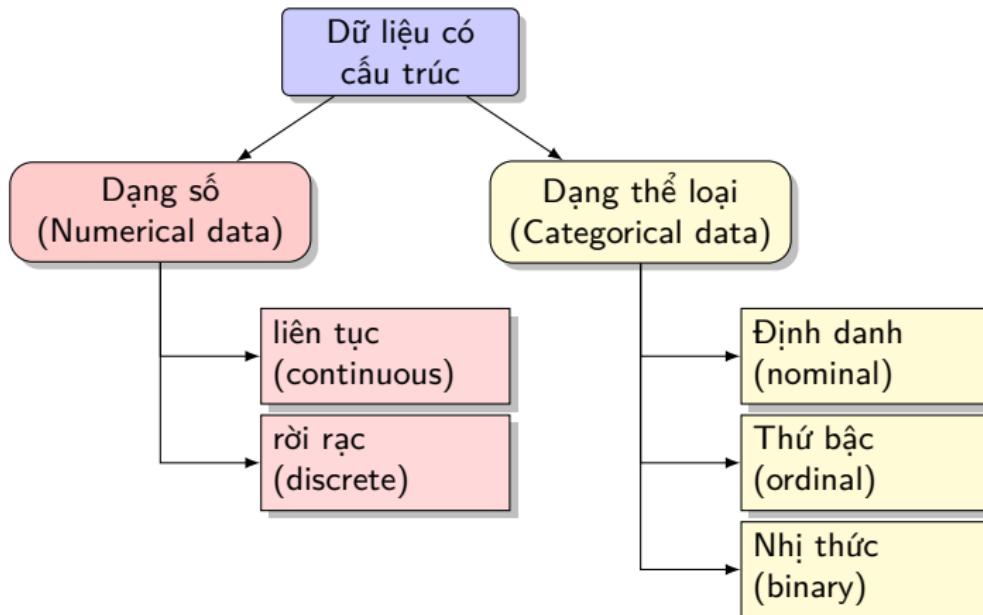
Dữ liệu?

Dữ liệu có thể tới từ nhiều nguồn:

- cảm biến (sensor);
- sự kiện;
- văn bản (text);
- hình ảnh (images);
- âm thanh;
- video.



Dữ liệu có cấu trúc - Structured data



Dữ liệu có cấu trúc - Structured data

Dữ liệu dạng số (*numerical data* hay *quantitative data*) là dữ liệu được cấu trúc với giá trị là số:

- liên tục (*continuous*);
- rời rạc hay số đếm (*discrete*),

ghi lại/biểu diễn số lượng của một đối tượng nào đó.

Dữ liệu dạng phân loại (*categorical data* hay *qualitative data*) là dữ liệu được cấu trúc với giá trị là các thể loại khác nhau của người hoặc vật. Thông thường ta có hai dạng phổ biến sau:

- biến định danh (*nominal*)
- biến thứ bậc (*ordinal*)

Đặc biệt, đối với dữ liệu định danh, ta có:

- dữ liệu nhị định danh hay dữ liệu nhị phân (*binary data*), khi biến có 2 giá trị;
- dữ liệu đa định danh hay dữ liệu đa thức (*nominal multinomial data* hoặc *multinomial data*).

Dữ liệu có cấu trúc - Structured data

Ví dụ:

- số liên tục: tốc độ gió, doanh thu, khoảng thời gian, chi phí, ...
- số rời rạc: số lượng lượt truy cập web, số lượng khách hàng trong một ngày, ...
- đa định danh: thể loại màn hình TV: plasma, LCD, LED, ...
- thứ bậc: điểm đánh giá sản phẩm, thăng hạng lòng về chất lượng dịch vụ, ...
- nhị phân: có/không, đúng/sai, ...

Ví dụ dữ liệu cấu trúc

Chi phí quảng cáo trên các nền tảng:

- truyền hình (TV)
- đài phát thanh (radio)
- nhật báo in (newspaper)

và doanh thu bán sản phẩm của công ty đều là các dữ liệu dạng số liên tục.

	A	B	C	D	E
1		TV	radio	newspaper	sales
2	1	230.1	37.8	69.2	22.1
3	2	44.5	39.3	45.1	10.4
4	3	17.2	45.9	69.3	9.3
5	4	151.5	41.3	58.5	18.5
6	5	180.8	10.8	58.4	12.9
7	6	8.7	48.9	75	7.2
8	7	57.5	32.8	23.5	11.8
9	8	120.2	19.6	11.6	13.2
10	9	8.6	2.1	1	4.8
11	10	199.8	2.6	21.2	10.6
12	11	66.1	5.8	24.2	8.6

Dữ liệu phi cấu trúc - Unstructured data

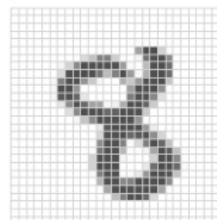
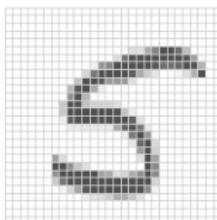
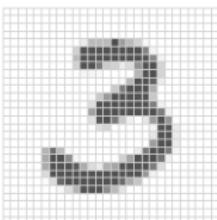
Dữ liệu phi cấu trúc là những dữ liệu được thu thập không có dạng số hoặc dạng thể loại:

- hình ảnh - images: là tập hợp các pixel với mỗi pixel chứa thông tin màu RGB (đỏ - red, xanh lá - green, xanh dương - blue);
- văn bản - text: là chuỗi các từ và ký tự không phải từ, thường được sắp xếp theo các phần, phần phụ, v.v;
- dòng nhấp chuột - clickstreams: là chuỗi hành động do người dùng tương tác với một ứng dụng hoặc trang web.

Để áp dụng các mô hình của statistical learning/machine learning, dữ liệu không có cấu trúc phải được xử lý và thao tác thành dạng có cấu trúc.

Dữ liệu phi cấu trúc - Unstructured data

0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9
0 1 2 3 4 5 6 7 8 9



Dữ liệu phi cấu trúc - Unstructured data

Dữ liệu về hình chụp phiết máu mỏng của hai người:

- nhiễm ký sinh trùng (a);
- không bị nhiễm ký sinh trùng (b).



(a) Parasitized



(b) Uninfected

Thu thập dữ liệu và sự ngẫu nhiên

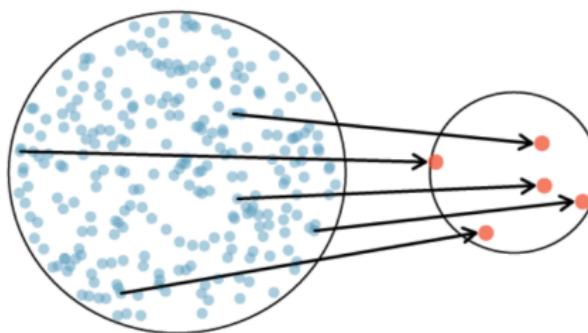
Sự ngẫu nhiên hóa - Randomization

Ngẫu nhiên hóa (**Randomization**) là một cơ chế để đạt được một mẫu ngẫu nhiên (hay dữ liệu) đủ “tốt” để đại diện cho một quần thể trong một cuộc khảo sát hoặc một thí nghiệm.

Sự ngẫu nhiên hóa được thể hiện qua việc lựa chọn không thiên vị, không bị chi phối bởi các yếu tố chủ quan.

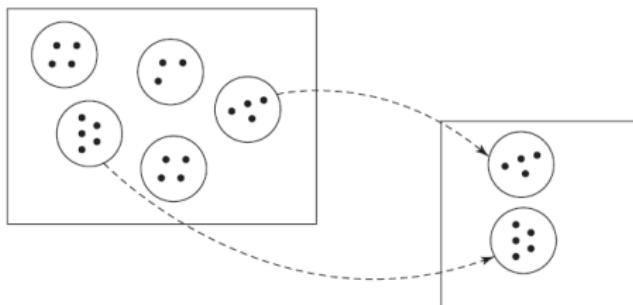
Thu thập dữ liệu và sự ngẫu nhiên

Lấy mẫu ngẫu nhiên đơn giản (*simple random sampling*) là một cách lấy mẫu ngẫu nhiên mà trong đó, ta chọn ngẫu nhiên n đối tượng từ quần thể sao cho mỗi đối tượng một lần với xác suất được chọn là như nhau, các đối tượng là độc lập nhau.



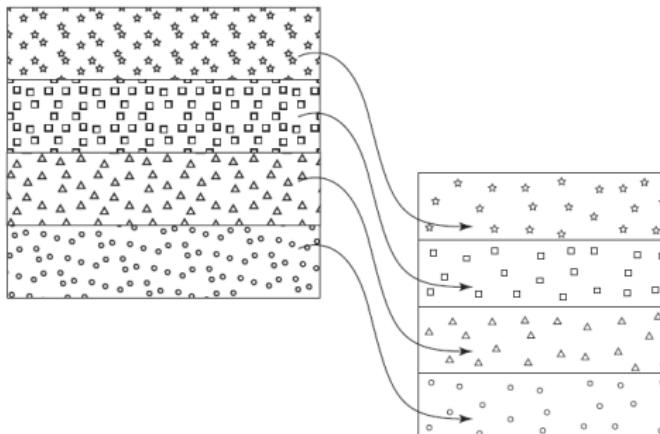
Thu thập dữ liệu và sự ngẫu nhiên

Lấy mẫu ngẫu nhiên theo cụm (*random cluster sample*) là một cách lấy mẫu ngẫu nhiên mà trong đó, ta sẽ chọn ngẫu nhiên **toàn bộ** thành viên trong một nhóm thay vì chỉ định một thành viên duy nhất.



Thu thập dữ liệu và sự ngẫu nhiên

Lấy mẫu ngẫu nhiên phân tầng (*stratified random sample*) một mẫu ngẫu nhiên phân tầng được chọn bằng cách trước tiên chia quần thể thành các tầng, tức là tập hợp các cá thể đồng nhất. Sau đó, nhiều mẫu ngẫu nhiên đơn giản được lấy - một mẫu trong mỗi tầng - và kết hợp lại để tạo thành mẫu.



Thu thập dữ liệu và sự ngẫu nhiên

Ví dụ:

Một nhà nghiên cứu dược phẩm muốn so sánh hiệu quả của hai loại thuốc trong một vài điều kiện bất lợi.

- Cô ấy có 4 bệnh nhân thỏa các điều kiện này.
- Cô ấy mong muốn chọn 2 bệnh nhân để thử 1 loại thuốc.

Để lựa chọn ngẫu nhiên, cô ấy đánh số các bệnh nhân lần lượt là P_1, P_2, P_3 và P_4 .

Đối với việc sử dụng thuốc 1, số cặp ngẫu nhiên có thể là:

$$(P_1, P_2), (P_1, P_3), (P_1, P_4), (P_2, P_3), (P_2, P_4), (P_3, P_4).$$

Tổng quát hóa, cho trường hợp N đối tượng trong 1 quần thể. Số mẫu với n đối tượng có thể có là C_n^N , tức là:

$$C_n^N = \frac{N!}{n!(N-n)!}.$$

Một ví dụ khác: lựa chọn bi màu trong hộp kín.

Thống kê mô tả: tổng hợp dữ liệu

Dữ liệu của một biến được tổng hợp (mô tả) bởi ba yếu tố chính:

- giá trị trung tâm - central tendency,
- độ biến động - variability,
- phân phối của dữ liệu.

Cụ thể

giá trị trung tâm cung cấp ước tính về nơi chứa hầu hết dữ liệu

giá trị biến động đo lường xem các giá trị dữ liệu được phân cụm chặt chẽ
hay dàn trải.

phân phối của dữ liệu cung cấp cái nhìn tổng quát về sự phân bố của dữ liệu
(tập trung dày ở một vùng cụ thể, dàn trải trên vùng nào đó).

Thống kê mô tả: tổng hợp dữ liệu

Có nhiều định nghĩa giá trị trung tâm khác nhau:

- trung bình cộng - mean (average)
- trung bình bị cắt bớt - trimmed mean (trimmed average)
- trung bình có trọng số - weighted mean (weighted average)
- trung vị - median (phân vị thứ 2 - second quantile)
- trung vị có trọng số - weighted median

Thống kê mô tả: tổng hợp dữ liệu

Trung bình cộng - mean

Trung bình cộng - mean (average) là ước lượng cơ bản cho giá trị trung tâm của dữ liệu:

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

với n là số lượng dữ liệu được quan sát.

Trung bình bị cắt bớt - trimmed mean

Trung bình bị cắt bớt - trimmed mean là một biến thể của trung bình, nó tính dựa trên bộ dữ liệu đã được cắt bỏ đi $p\%$ dữ liệu nhỏ nhất và lớn nhất (hai đầu của dữ liệu được sắp xếp thứ tự):

$$\text{trimmed mean} = \bar{x}_p = \frac{1}{n - 2\lfloor np \rfloor} \sum_{i=\lfloor np \rfloor + 1}^{n - \lfloor np \rfloor} x_{(i)},$$

với $x_{(i)}$ là giá trị của dữ liệu sau khi được sắp xếp tăng dần.

Thống kê mô tả: tổng hợp dữ liệu

Trung bình có trọng số - weighted mean

Trung bình có trọng số - weighted mean là một biến thể khác của trung bình, ở đó, các giá trị x_i được nhân thêm với một trọng số w_i :

$$\text{weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^n x_i w_i}{\sum_{i=1}^n w_i}.$$

Trung bình có trọng số thường được áp dụng khi:

- khi dữ liệu là kết quả từ các nhóm có kích thước khác nhau: tỷ lệ tội phạm của các thành phố;
- khi dữ liệu được thu thập từ các nguồn biến động không đồng nhất: dữ liệu thu được từ các sensor có chất lượng khác nhau.

Thống kê mô tả: tổng hợp dữ liệu

Trung vị - median

Trung vị - median là một giá trị của dữ liệu quan sát, mà chia dữ liệu thành hai nửa đều nhau, khi dữ liệu được sắp xếp theo thứ tự tăng dần.



Trung vị của một bộ dữ liệu là nghiệm của

$$\text{median} = \arg \min_x \sum_{i=1}^n |x - x_i|$$

Thống kê mô tả: tổng hợp dữ liệu

Trung vị có trọng số - weighted median

Trung vị có trọng số - weighted median là một biến thể của trung vị, được áp dụng cho trường hợp dữ liệu có trọng số.

Nó là điểm x_k trong dữ liệu sao cho

$$\sum_{x_i < x_k} w_i \leq \frac{1}{2} \sum_{i=1}^n w_i, \quad \sum_{x_i > x_k} w_i \leq \frac{1}{2} \sum_{i=1}^n w_i$$

Định nghĩa này tương ứng với trung vị có trọng số là nghiệm của

$$\text{weighted median} = \arg \min_x \sum_{i=1}^n w_i |x - x_i|$$

Thống kê mô tả: tổng hợp dữ liệu

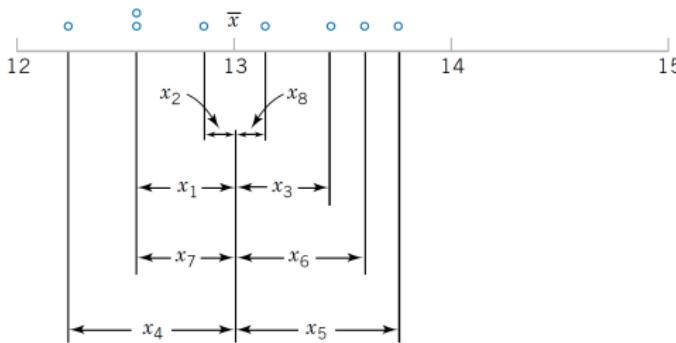
Độ biến động của dữ liệu được miêu tả thông qua các số đo:

- phương sai - variance
- độ lệch chuẩn - standard deviation
- trung bình độ lệch tuyệt đối - mean absolute deviation
- trung vị độ lệch tuyệt đối - median absolute deviation (MAD)
- khoảng biến động - range
- phân vị
- khoảng tứ phân vị - interquartile range (IQR)

Thống kê mô tả: tổng hợp dữ liệu

Phương sai, độ lệch chuẩn

Phương sai và độ lệch chuẩn (variance, standard deviation) là hai đại lượng đo độ biến động của dữ liệu xung quanh giá trị trung bình.



Phương sai được ký hiệu là s^2 và được xác định bởi:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2,$$

Độ lệch chuẩn, được ký hiệu là s và được xác định bởi: $s = \sqrt{s^2}$.

Thống kê mô tả: tổng hợp dữ liệu

Trung bình độ lệch tuyệt đối

Trung bình độ lệch tuyệt đối (mean absolute deviation) là đại lượng biểu thị độ biến động của dữ liệu xung quanh giá trị trung bình thông qua trung bình khoảng cách Euclidean:

$$\text{mean absolute deviation} = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|,$$

- Độ lệch chuẩn dễ diễn giải hơn nhiều so với phương sai vì nó có cùng tỷ lệ với dữ liệu gốc.
- Giá trị của phương sai, độ lệch chuẩn, trung bình độ lệch tuyệt đối càng nhỏ thì dữ liệu càng ít biến động so với trung bình.

Thống kê mô tả: tổng hợp dữ liệu

Trung vị độ lệch tuyệt đối - median absolute deviation (MAD)

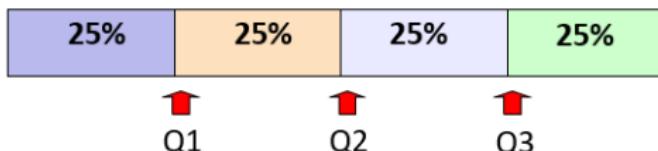
Trung vị độ lệch tuyệt đối - MAD là thước đo độ biến thiên dữ liệu xung quanh điểm trung vị:

$$\text{MAD} = \text{median}(|x_1 - m|, |x_2 - m|, \dots, |x_n - m|),$$

với m là trung vị.

Khoảng tứ phân vị - interquartile range

Khoảng tứ phân vị - interquartile range là khoảng cách giữa phân vị thứ nhất (25% dữ liệu) và phân vị thứ ba (75% dữ liệu).



Thống kê mô tả: tổng hợp dữ liệu

Độ lệch chuẩn có trọng số - weighted standard deviation

Độ lệch chuẩn có trọng số - weighted standard deviation là một biến thể của độ lệch chuẩn, để tương thích với các trường hợp dữ liệu có trọng số.

$$s_w = \sqrt{\frac{\sum_{i=1}^n w_i(x_i - \bar{x}_w)^2}{\frac{n'}{n'-1} \sum_{i=1}^n w_i}},$$

trong đó, n' là số lượng trọng số khác 0.

Tương tự, ta có:

- phương sai có trọng số - weighted variance
- trung vị độ lệch tuyệt đối có trọng số - weighted median absolute deviation (weighted MAD).

Thống kê mô tả: tổng hợp dữ liệu

Bảng tần số - frequency table

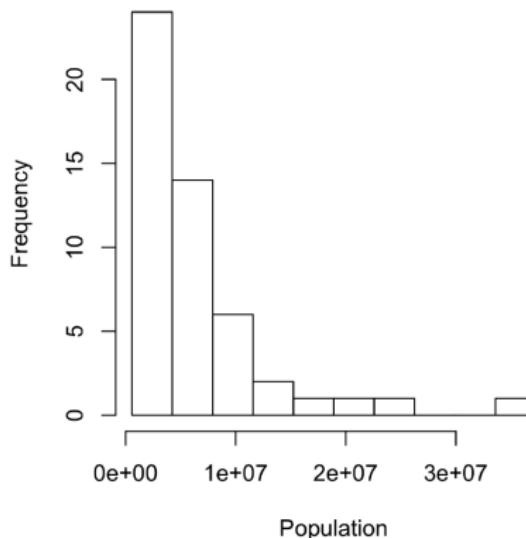
Bảng tần số của một biến chia phạm vi biến thành các phân đoạn cách đều nhau và cho chúng ta biết có bao nhiêu giá trị rơi vào mỗi phân đoạn.

BinNumber	BinRange	Count	States
1	563,626–4,232,658	24	WY, VT, ND, AK, SD, DE, MT, RI, NH, ME, HI, ID, NE, WV, NM, NV, UT, KS, AR, MS, IA, CT, OK, OR
2	4,232,659– 7,901,691	14	KY, LA, SC, AL, CO, MN, WI, MD, MO, TN, AZ, IN, MA, WA
3	7,901,692– 11,570,724	6	VA, NJ, NC, GA, MI, OH
4	11,570,725– 15,239,757	2	PA, IL
5	15,239,758– 18,908,790	1	FL
6	18,908,791– 22,577,823	1	NY
7	22,577,824– 26,246,856	1	TX
8	26,246,857– 29,915,889	0	
9	29,915,890– 33,584,922	0	
10	33,584,923– 37,253,956	1	CA

Thống kê mô tả: tổng hợp dữ liệu

Biểu đồ tần số - histogram

Biểu đồ tần số là một cách để trực quan hóa bảng tần số, với các cột trên trục x và số lượng dữ liệu trên trục y .



Thống kê mô tả: tổng hợp dữ liệu

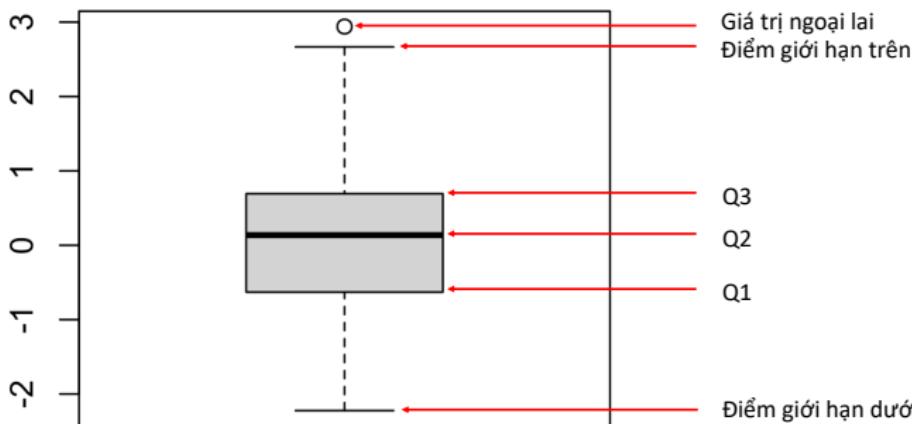
Về mặt kỹ thuật, một biểu đồ tần số được vẽ sao cho:

- Cột rỗng được bao gồm trong biểu đồ (khoảng không có dữ liệu).
- Các cột có chiều rộng bằng nhau.
- Số cột (hoặc tương đương, kích thước của cột) tùy thuộc vào người dùng.
- Các cột xếp kề nhau - không có khoảng trống nào hiển thị giữa các cột, trừ khi có cột trống.

Thống kê mô tả: tổng hợp dữ liệu

Biểu đồ hộp

Biểu đồ hộp hay **boxplot** là một biểu đồ dạng hộp dùng để miêu tả sự phân phối của dữ liệu, bằng việc biểu diễn các điểm tứ phân vị.



- $\text{Điểm giới hạn trên} = \min \{\max(x), Q_3 + 1.5 \times IQR\}$.
- $\text{Điểm giới hạn dưới} = \max \{\min(x), Q_1 - 1.5 \times IQR\}$.

Thống kê mô tả: tổng hợp dữ liệu

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Thống kê mô tả: tổng hợp dữ liệu

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Giá trị ngoại lai có thể là:

- một quan sát hiếm khi xảy ra: lượng mưa lớn đột ngột, số lượng lớn khác hàng trong một giờ;
- kết quả của lỗi dữ liệu: đơn vị không đồng nhất giữa các giá trị, sai sót trong đo lường hoặc ghi chép.

Thống kê mô tả: tổng hợp dữ liệu

Outliers - Extreme values

Giá trị ngoại lai (outliers/extreme values) là bất kỳ giá trị nào khác rất xa so với các giá trị còn lại trong tập dữ liệu.

129, (1.65), 132, 133, 137, 138, (1308), 140, 140, 141, 143, 143, (1405)

Giá trị ngoại lai có thể là:

- một quan sát hiếm khi xảy ra: lượng mưa lớn đột ngột, số lượng lớn khác hàng trong một giờ;
- kết quả của lỗi dữ liệu: đơn vị không đồng nhất giữa các giá trị, sai sót trong đo lường hoặc ghi chép.

Nhận xét:

- các giá trị ngoại lai có ảnh hưởng xấu tới ước lượng của trung bình và phương sai cũng như độ lệch chuẩn;
- các ước lượng: median, MAD là các ước lượng “robust” với outliers (tức là ít bị ảnh hưởng bởi outliers).

Thống kê mô tả: tổng hợp dữ liệu

Để nhận dạng được sự hiện diện của outliers trong dữ liệu, ta có thể dùng:

- khoảng tứ phân vị - IQR, cụ thể, các điểm dữ liệu nằm ngoài khoảng

$$[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$$

thì được coi là outliers trong dữ liệu;

- median và MAD, cụ thể, ta xét khoảng

$$[\text{median} - 3 \times \text{MAD}, \text{median} + 3 \times \text{MAD}]$$

nếu các điểm dữ liệu không nằm trong khoảng này, thì được coi là outliers.

Cách xét này được gọi là bộ lọc Hampel - Hampel filter.

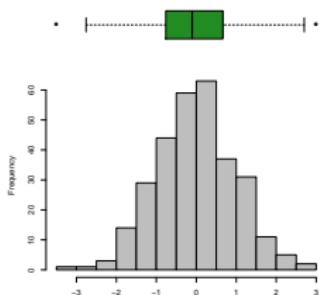
Thống kê mô tả: tổng hợp dữ liệu

Các hình dạng phân bố thường gặp

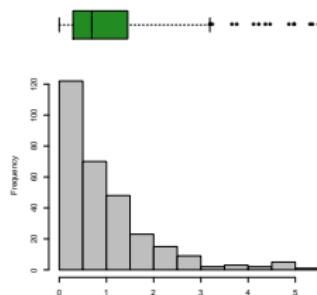
- Unimodal distribution: một đỉnh, hầu hết dữ liệu được phân bố xung quanh một giá trị;
- Bimodal (multimodal) distribution: hai (hoặc nhiều) đỉnh, dữ liệu được nhóm xung quanh hai (hoặc nhiều) đỉnh;
- Symmetric distribution: một đỉnh, phía bên trái của phân bố phản ánh phía bên phải.
- Positive or right-skewed distribution: đỉnh bên trái, đuôi dài bên phải, kéo dài (nghiêng) sang bên phải, một vài giá trị lớn, trung vị nhỏ hơn trung bình, nếu đơn hình thì mode nhỏ hơn trung vị nhỏ hơn trung bình;
- Negative or left-skewed distribution: đỉnh bên phải, đuôi dài bên trái, kéo dài (nghiêng) sang trái, một vài giá trị nhỏ, trung vị lớn hơn trung bình, nếu unimodal thì mode lớn hơn trung vị lớn hơn trung bình.

Thống kê mô tả: tổng hợp dữ liệu

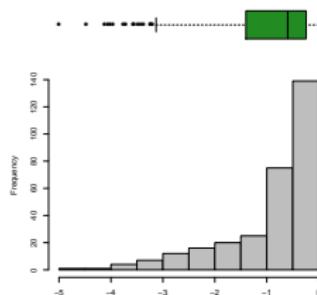
symmetric



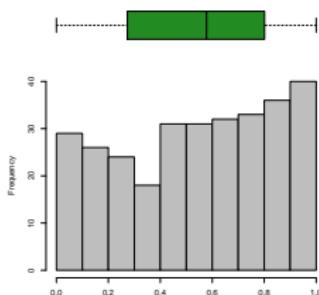
right-skewed



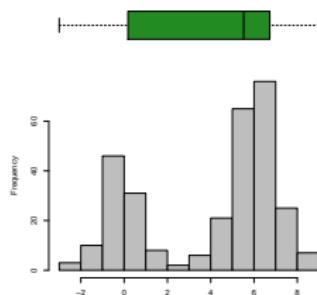
left-skewed



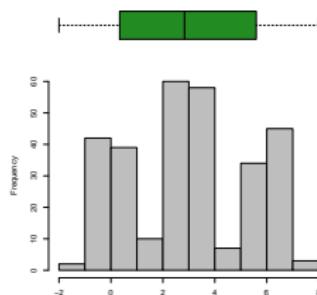
uniform



bimodal



multimodal



Thống kê mô tả: tổng hợp dữ liệu

Modality

unimodal



bimodal



multimodal



uniform

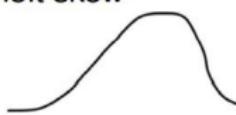


Skewness

right skew



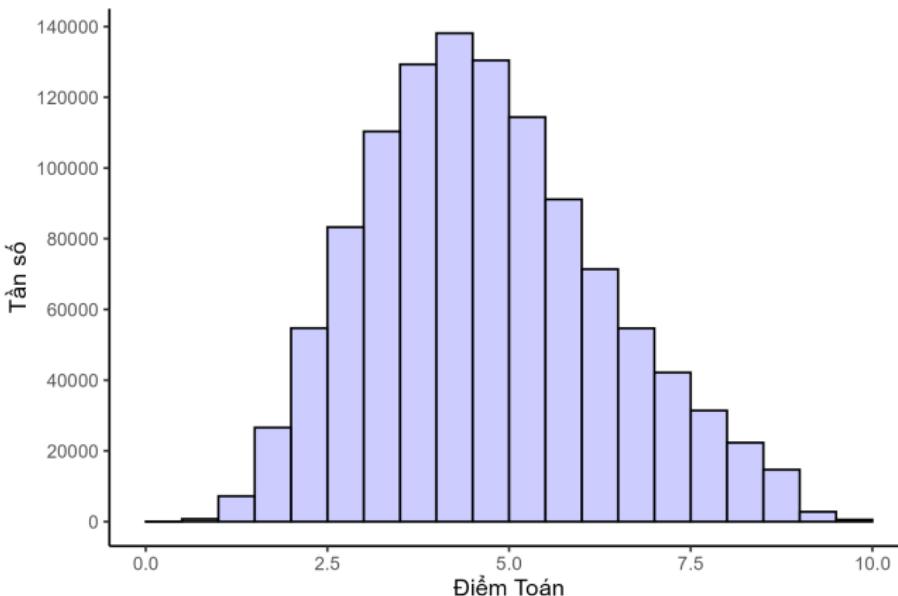
left skew



symmetric



Ví dụ: điểm thi Toán tốt nghiệp THPT 2025

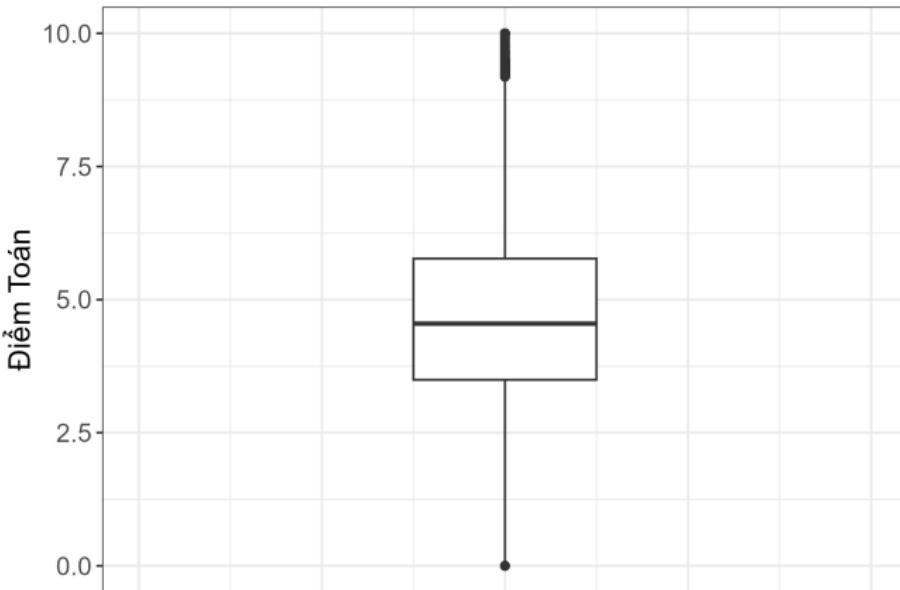


Ví dụ: điểm thi Toán tốt nghiệp THPT 2025

Một số thống kê cơ bản:

	Giá trị
Tổng số thí sinh	1,126,172
Trung bình điểm	4.78
Trung vị	4.6
Độ lệch chuẩn	1.68
MAD	1.35
IQR	2.28

Ví dụ: điểm thi Toán tốt nghiệp THPT 2025



Thống kê mô tả: tổng hợp dữ liệu đa chiều

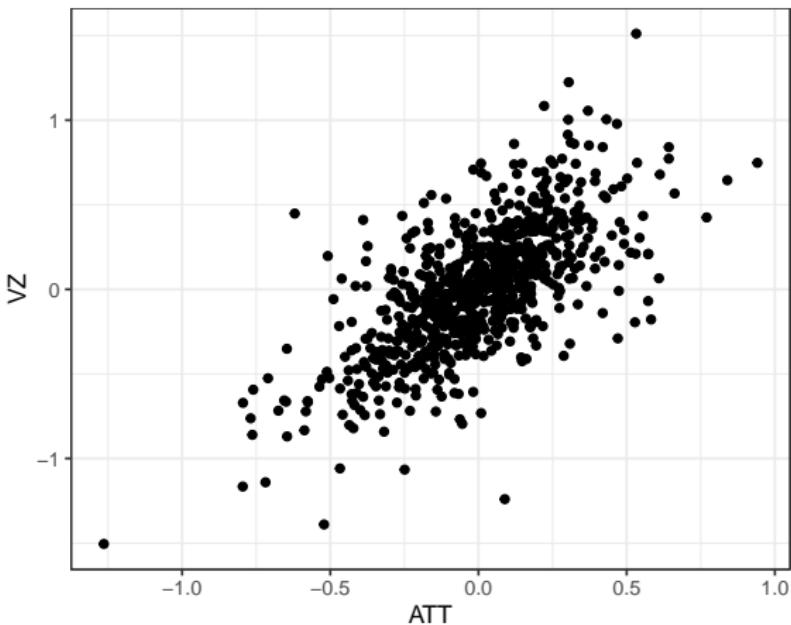
Trong thống kê, ta thường quan tâm tới bài việc mô tả thống hợp cho nhiều biến cùng 1 lúc, nhằm ta ra các xu hướng giữa:

- hai biến định lượng
- biến định lượng vs. biến định tính
- hai biến định tính

Ta có thể dùng các công cụ:

- biểu đồ phân tán (scatter plot);
- biểu đồ hộp;
- contingency table.

Thống kê mô tả: tổng hợp dữ liệu đa chiều

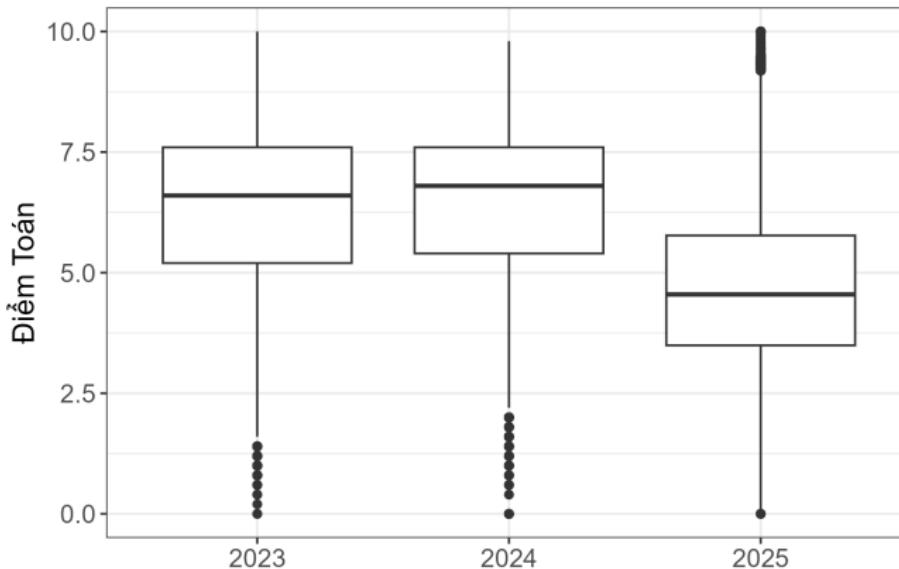


Thống kê mô tả: tổng hợp dữ liệu đa chiều

Bảng tổng hợp dưới đây là một ví dụ về contingency table của màu tóc (X) và màu mắt (Y) của 6,800 nam giới Đức.

		Màu tóc			
		Nâu	Đen	Vàng nhạt	Đỏ
Màu mắt	Nâu	438	288	115	16
	Xám hoặc Xanh lá	1387	746	946	53
	Xanh dương	807	189	1768	47

Thống kê mô tả: tổng hợp dữ liệu đa chiều



- Biến định tính: năm thi tốt nghiệp THPT (2023, 2024, 2025)
- Biến định lượng: điểm thi môn Toán

1 *Giới thiệu về Khoa học thống kê*

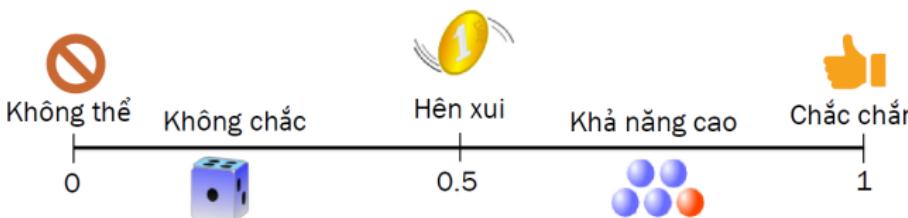
2 *Xác suất và Biến ngẫu nhiên*

3 *Vector ngẫu nhiên*

Không gian xác suất

Xác suất là gì?

Xác suất là một chuyên ngành Toán, nghiên cứu về khả năng xảy ra của một sự kiện trong một phép thử ngẫu nhiên hay quan sát ngẫu nhiên.



Xác suất xảy ra của một sự kiện A , thường được ký hiệu là $P(A)$, luôn có giá trị từ 0 tới 1, trong đó,

- $P(A) = 0$ có nghĩa là sự kiện A chắc chắn không thể xảy ra;
- $P(A) = 1$ có nghĩa là sự kiện A chắc chắn xảy ra;
- các giá trị $P(A)$ càng gần 0 thì khả năng xảy ra sự kiện A càng nhỏ, và ngược lại khi giá trị $P(A)$ càng gần 1;
- đặc biệt, khi $P(A) = 0.5$, ta thường nói là 50/50 hay là "hên xui".

Không gian xác suất

Ví dụ: ta thường có những câu hỏi liên quan tới xác suất chẳng hạn như:

- khả năng trúng 1 giải khi mua vé số?
- khả năng xảy ra trời mưa trong một ngày theo dự báo thời tiết?
- xác suất một kênh dẫn thông tin truyền sai tín hiệu?
- tỷ lệ tử vong khi bị nhiễm covid-19?

Không gian xác suất

Phép thử ngẫu nhiên

Một **phép thử ngẫu nhiên** là một phép thử mà khi ta lặp lại sẽ có thể thu được kết quả khác so với lần thử trước đó.

Một số phép thử ngẫu nhiên:

- tung một đồng xu (cân đối, đồng chất) một lần;
- gieo một con xúc xắc một lần;
- dự đoán thời tiết trong một ngày.

Thử một số ví dụ tại trang web: <https://www.random.org/>

Không gian xác suất

Một phép thử ngẫu nhiên gồm có các thành phần sau:

đối tượng là một đối tượng nhất định được quan tâm trong nghiên cứu, ví dụ: đồng xu, con xúc xắc, thời tiết, con người, mạch điện, ...

không gian mẫu là tập hợp gồm tất cả các kết quả của một phép thử ngẫu nhiên, được ký hiệu là Ω (Omega-hoa);

biến cỗ sơ cấp là một kết quả đơn lẻ được bao hàm trong không gian mẫu, được ký hiệu là ω (omega);

biến cỗ/sự kiện là một tập hợp gồm 1 hoặc nhiều biến cỗ sơ cấp, được ký hiệu bởi các chữ in hoa, ví dụ, A, B, C.

Không gian xác suất

Ví dụ 1: Xét phép thử tung một đồng xu (cân đối, đồng chất) một lần.

Ta có cây sự kiện như sau:



Trong đó:

- phép thử sơ cấp ω : mặt sấp (S), mặt ngửa (N);
- không gian mẫu $\Omega = \{S, N\}$.

Giả sử ta có

- biến cố A : thu được mặt sấp, $A = \{S\}$;
- biến cố B : thu được mặt sấp hoặc ngửa, $B = \{S, N\}$;
- biến cố C : không thu được mặt sấp hoặc không thu được mặt ngửa, $C = \emptyset$.

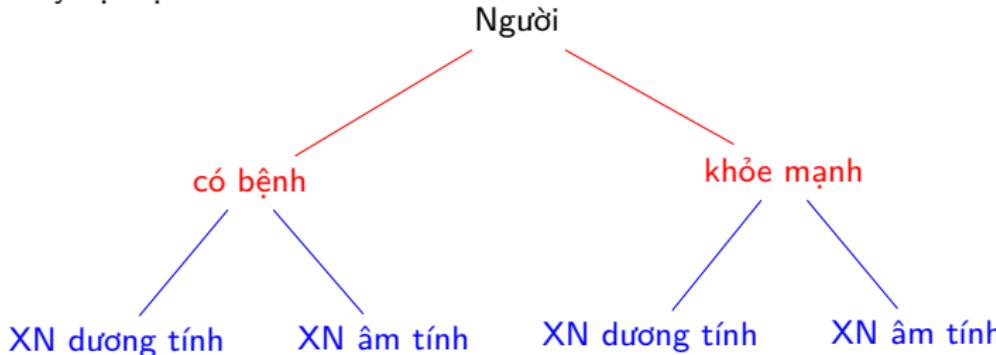
Biến cố B là **biến cố chắc chắn**, tức là chắc chắn xảy ra.

Biến cố C là **biến có không**, tức là không bao giờ xảy ra,

Không gian xác suất

Ví dụ 2: Xét nghiệm y tế.

Ta có cây sự kiện như sau:



Trong đó

- phép thử sơ cấp ω : BD, BA, KD, KA;
- không gian mẫu $\Omega = \{BD, BA, KD, KA\}$.

Khi đó, ta có

- biến cố A: người bệnh có kết quả dương tính $A = \{BD\}$;
- biến cố B: kết quả âm tính, $B = \{BA, KA\}$;
- biến cố C: người khỏe mạnh, $C = \{KD, KA\}$.

Không gian xác suất

Ví dụ 3: Xét phép thử chọn 1 số ngẫu nhiên trong $[0, 1]$ tức là từ 0 tới 1.

Khi đó,

- phép thử sơ cấp ω : 0, 0.001, 0.0011, ..., 1 ;
- không gian mẫu $\Omega = \{0, 0.001, 0.0011, \dots, 1\} = [0, 1]$.

Ta có

- biến cố A: thu được số 0.2, $A = \{0.2\}$;
- biến cố B: thu được 1 số trong khoảng $[0, 0.15]$

$$B = \{\omega \in \Omega : 0 \leq \omega \leq 0.15\} \equiv [0, 0.15];$$

- biến cố C: thu được 1 số lớn hơn 1, $C \equiv (1, +\infty)$.

Không gian xác suất

Trong các ví dụ 1 và 2, ta nhận thấy rằng không gian mẫu Ω :

- ví dụ 1: $\Omega = \{S, N\}$
- ví dụ 2: $\Omega = \{BD, BA, KD, KA\}$

đều là các tập hợp với số phần tử đếm được, lần lượt 2 và 4.

↪ những không gian mẫu có số phần tử đếm được hoặc vô hạn đếm được, thì được gọi là **không gian mẫu rời rạc**.

Trong ví dụ 3, không gian mẫu $\Omega = \{0, 0.001, 0.0011, \dots, 1\}$, chứa vô hạn phần tử, và là một khoảng số thực, $[0, 1]$.

↪ những không gian mẫu là khoảng số thực, thì được gọi là **không gian mẫu liên tục**.

Hay tương đương $\Omega \subseteq \mathbb{R}$.

Các phép toán trong không gian xác suất

Nhắc lại: một biến cố hay sự kiện là một tập hợp gồm 1 hoặc nhiều biến cố sơ cấp, thường được ký hiệu bởi các chữ cái in hoa, ví dụ, A, B, C

Giả sử rằng ta thực hiện phép thử ngẫu nhiên, với không gian mẫu Ω . Ta có một số tính chất cơ bản sau:

- Gọi A là biến cố được quan tâm, khi đó, $A \subset \Omega$ (tức là, A là tập con của Ω);
- tập rỗng \emptyset là biến cố không (tức là chắc chắn không xảy ra);
- không gian mẫu Ω là biến cố chắc chắn (tức là chắc chắn xảy ra).

Ví dụ 1 (tiếp theo): không gian mẫu $\Omega = \{S, N\}$.

Xét biến cố A: thu được mặt sấp. Khi đó, ta viết: $A = \{S\} \Rightarrow A \subset \Omega$.

Ví dụ 2 (tiếp theo): không gian mẫu $\Omega = \{BD, BA, KD, KA\}$.

Xét biến cố B: xét nghiệm âm tính. Khi đó, ta viết: $B = \{BA, KA\} \Rightarrow B \subset \Omega$.

Ví dụ 3 (tiếp theo): không gian mẫu $\Omega = [0, 1]$.

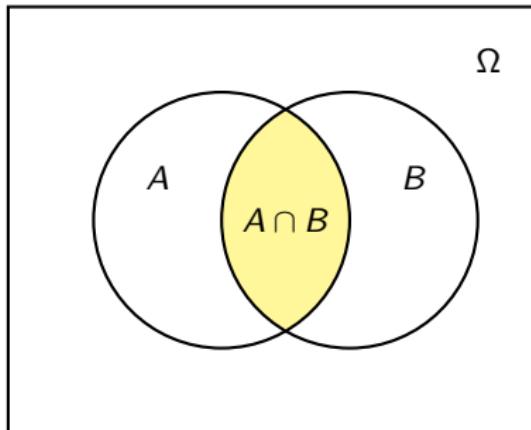
Xét biến cố C: thu được số x bất kỳ trong khoảng $[0.1, 0.3]$. Khi đó, ta viết:
 $C = \{x \in [0, 1] : 0.1 \leq x \leq 0.3\} \Rightarrow C \subset \Omega$.

Các phép toán trong không gian xác suất

Xét hai biến cố A và B , **phép toán giao** giữa A và B ký hiệu là $A \cap B$ (hay viết tắt AB), kết quả là một biến cố mới có các biến cố sơ cấp vừa thuộc A mà cũng vừa thuộc B , ta viết

$$C = A \cap B.$$

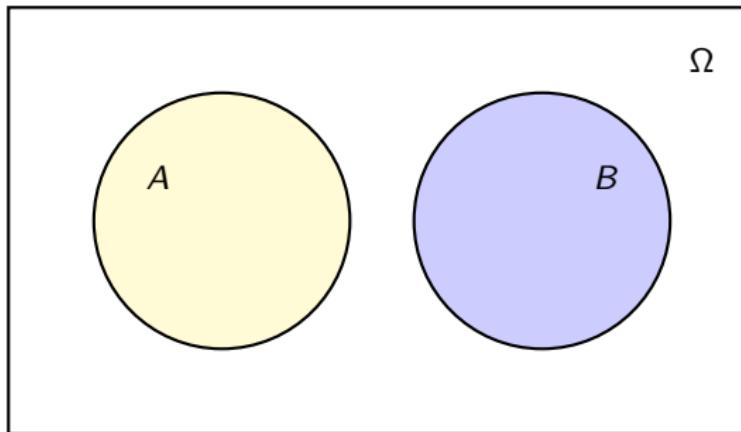
Phép toán được miêu tả bởi biểu đồ Venn dưới đây.



Các phép toán trong không gian xác suất

Nếu $A \cap B = \emptyset$, thì ta nói A và B là hai biến cố **xung khắc**.

Tương ứng đồ Venn như sau.



Các phép toán trong không gian xác suất

Ví dụ 2 (tiếp theo): Nhắc lại không gian mẫu cho xét nghiệm y tế,
 $\Omega = \{BD, BA, KD, KA\}$. Xét các biến cỗ:

- $A = \{BD, BA\}$;
- $B = \{BD, KD, KA\}$;
- $C = \{KD, KA\}$

Khi đó,

- $A \cap B = \{BD\}$;
- $A \cap C = \emptyset \Rightarrow A$ và C là biến cỗ ???;
- $B \cap C = ???$;
- $A \cap B \cap C = ???$.

Các phép toán trong không gian xác suất

Ví dụ 3 (tiếp theo): Nhắc lại không gian mẫu cho phép thử chọn ngẫu nhiên một số x trong khoảng $[0, 1]$ là $\Omega = [0, 1]$. Xét các biến cố:

- $A = \{x \in [0, 1] : 0 \leq x \leq 0.15\}$ - tức là thu được số x bất kỳ trong khoảng $[0, 0.15]$;
- $B = \{x \in [0, 1] : 0.1 < x \leq 0.3\}$ - tức là thu được số x bất kỳ trong khoảng $(0.1, 0.3]$;
- $C = \{x \in [0, 1] : 0 \leq x \leq 0.1\}$.

Khi đó,

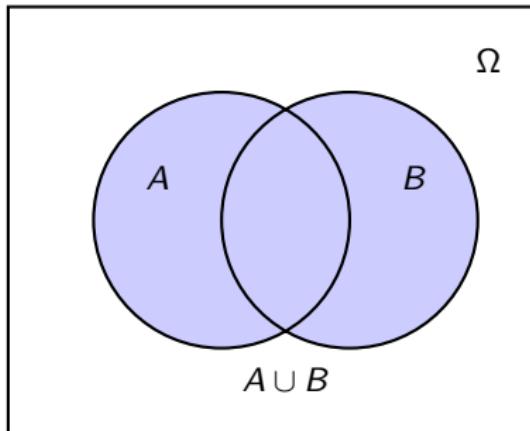
- $A \cap B = \{x \in [0, 1] : 0.1 < x \leq 0.15\};$
- $A \cap C = ???;$
- $B \cap C = ???;$
- $A \cap B \cap C = ???.$

Các phép toán trong không gian xác suất

Xét hai biến cố A và B , **phép toán hợp** giữa A và B ký hiệu là $A \cup B$ (hay viết tắt $A + B$), kết quả là một biến cố mới có tất cả các biến cố sơ cấp của cả A và B , ta viết

$$C = A \cup B$$

Phép toán được miêu tả bởi biểu đồ Venn dưới đây.



Các phép toán trong không gian xác suất

Ví dụ 2 (tiếp theo): Nhắc lại không gian mẫu cho xét nghiệm y tế,

$\Omega = \{BD, BA, KD, KA\}$. Xét các biến cố:

- $A = \{BD, BA\}$;
- $B = \{BD, KD, KA\}$;
- $C = \{KD, KA\}$

Khi đó,

- $A \cup B = \{BD, BA, KD, KA\} = \Omega$;
- $B \cup C = ???$;
- $A \cup B \cup C = ???$;
- $A \cup B \cap C = ???$;
- $A \cup (B \cap C) = ???$.

Các phép toán trong không gian xác suất

Ví dụ 3 (tiếp theo): Nhắc lại không gian mẫu cho phép thử chọn ngẫu nhiên một số x trong khoảng $[0, 1]$ là $\Omega = [0, 1]$. Xét các biến cố:

- $A = \{x \in [0, 1] : 0 \leq x \leq 0.15\};$
- $B = \{x \in [0, 1] : 0.1 < x \leq 0.3\};$
- $C = \{x \in [0, 1] : 0 \leq x \leq 0.1\}.$

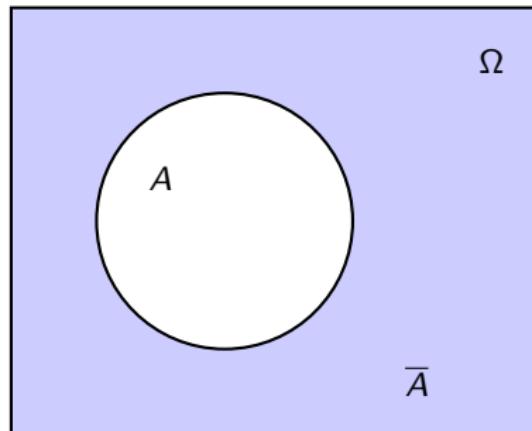
Khi đó,

- $A \cup B = \{x \in [0, 1] : 0 \leq x \leq 0.3\};$
- $A \cup C = ???;$
- $A \cup B \cup C = ???;$
- $B \cap C \cup A = ???;$
- $(A \cup B) \cap C = ???.$

Các phép toán trong không gian xác suất

Xét biến cố A , **phép toán phần bù** của A được ký hiệu là \bar{A} (hoặc A^c), kết quả là một biến cố mới có các biến cố sơ cấp thuộc Ω nhưng không thuộc A ; ta gọi \bar{A} là **biến cố đối lập** của A .

Phép toán được miêu tả bởi biểu đồ Venn dưới đây.

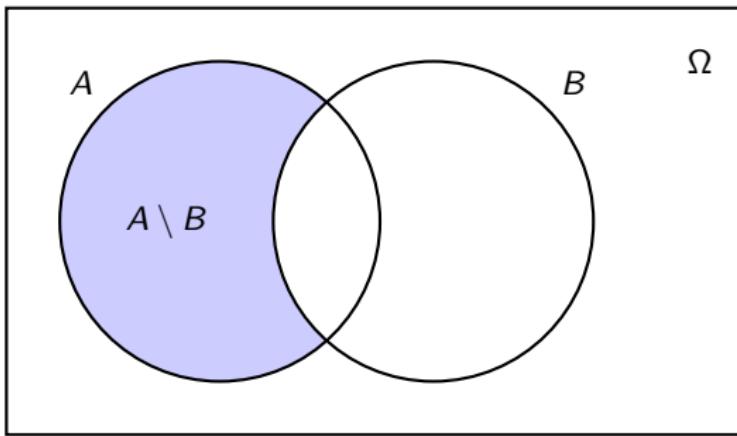


Nhận xét: $A \cup \bar{A} = \Omega$.

Các phép toán trong không gian xác suất

Cho hai biến cố A và B sao cho $A \cap B \neq \emptyset$, thì khi đó ta có phép toán $A \setminus B$ (A bù B , hay A hiệu B), được xác định là biến cố có các biến cố sơ cấp thuộc A nhưng không thuộc B :

$$A \setminus B = \{\omega \in \Omega : \omega \in A \text{ và } \omega \notin B\}$$



Đặc biệt, nếu $A \equiv \Omega$, ta có

$$\Omega \setminus B = \overline{B} = \{\omega \in \Omega : \omega \notin B\}$$

Các phép toán trong không gian xác suất

Ví dụ 2 (tiếp theo): Nhắc lại không gian mẫu cho xét nghiệm y tế,
 $\Omega = \{BD, BA, KD, KA\}$. Xét các biến cố:

- $A = \{BD, BA\}$;
- $B = \{BD, KD, KA\}$;
- $C = \{KD, KA\}$

Khi đó,

- $A \setminus B = \{BA\}$;
- $\bar{C} = \{BD, BA\}$;
- $B \setminus C = ???$;

Các phép toán trong không gian xác suất

Ví dụ 3 (tiếp theo): Nhắc lại không gian mẫu cho phép thử chọn ngẫu nhiên một số x trong khoảng $[0, 1]$ là $\Omega = [0, 1]$. Xét các biến cố:

- $A = \{x \in [0, 1] : 0 \leq x \leq 0.15\}$;
- $B = \{x \in [0, 1] : 0.1 < x \leq 0.3\}$;
- $C = \{x \in [0, 1] : 0 \leq x \leq 0.1\}$.

Khi đó,

- $A \setminus B = \{x \in [0, 1] : 0 \leq x \leq 0.1\}$;
- $B \setminus A = ???$;
- $A \setminus C = ???$;
- $\Omega \setminus (A \cup B) = ???$.

Các phép toán trong không gian xác suất

Ta có một số kết quả cho phép toán của biến cố, theo luật De Morgan:

- $\overline{A \cup B} = \overline{A} \cap \overline{B}$
- $\overline{A \cap B} = \overline{A} \cup \overline{B}$
- $\overline{(A \cup B) \cap C} = (\overline{A} \cap \overline{B}) \cup \overline{C}$
- $A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$
- $A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$
- $\overline{A_1 \cup A_2 \cup \dots \cup A_k} = \overline{A_1} \cap \overline{A_2} \dots \cap \overline{A_k}$
- $\overline{A_1 \cap A_2 \cap \dots \cap A_k} = \overline{A_1} \cup \overline{A_2} \dots \cup \overline{A_k}$.

Các công thức xác suất

Định nghĩa xác suất cổ điển

Xét không gian mẫu rời rạc Ω (đếm được), gọi A là một biến cố được quan tâm. Khi đó, xác suất để biến cố A xảy ra được tính bởi:

$$P(A) = \frac{\text{số biến cố sơ cấp trong } A}{\text{số biến cố sơ cấp trong } \Omega} = \frac{|A|}{|\Omega|}.$$

Ví dụ: Xét phép thử tung đồng xu cân đối đồng chất 1 lần.

- không gian mẫu $\Omega = \{S, N\}$;
- biến cố $A = \{S\}$.

Khi đó, xác suất biến cố A xảy ra là

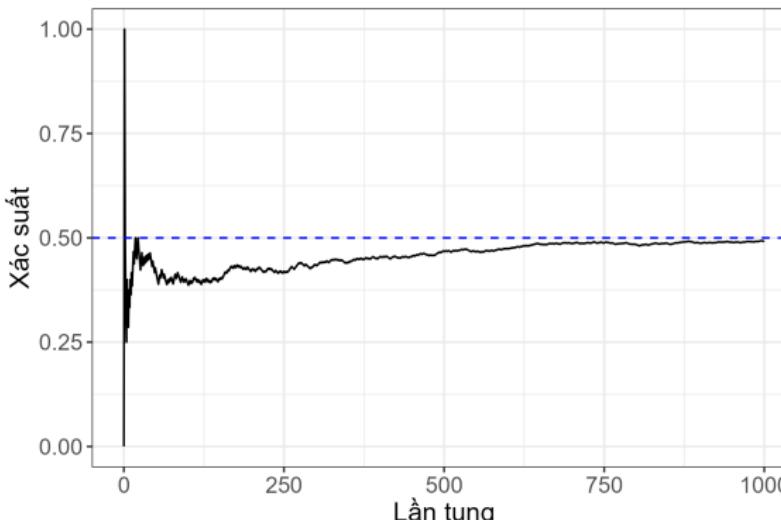
$$P(A) = \frac{1}{2}.$$

Các công thức xác suất

Định nghĩa xác suất hiện đại

Đối với việc quan sát một hiện tượng ngẫu nhiên, xác suất của một kết quả cụ thể là tỷ lệ số lần kết quả đó xảy ra trong một chuỗi dài vô hạn các quan sát giống nhau, trong cùng điều kiện.

Ví dụ: thực hiện tung đồng xu trong nhiều lần và ghi chép lại kết quả. Khi đó, xác suất thu được biểu diễn như trong hình:



Các công thức xác suất

Tiên đề xác suất

Xác suất là một số được gán cho mỗi thành viên của tập hợp các biến cố từ một phép thử ngẫu nhiên thỏa mãn các tính chất sau:

Nếu Ω là không gian mẫu và A là biến cố ngẫu nhiên bất kỳ,

$$(1) \ P(\Omega) = 1;$$

$$(2) \ 0 \leq P(A) \leq 1;$$

$$(3) \ Vói hai biến cố A và B sao cho A \cap B = \emptyset, \text{ khi đó}$$

$$P(A \cup B) = P(A) + P(B).$$

Từ 3 tính chất trên của tiên đề xác suất, ta có 2 kết quả sau:

$$\blacksquare \ P(\emptyset) = 0;$$

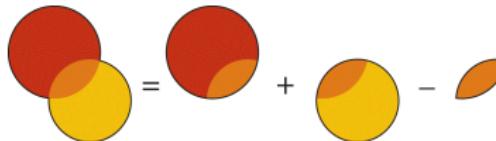
$$\blacksquare \ P(\overline{A}) = 1 - P(A).$$

Các công thức xác suất

Xác suất của hội hai biến cố

Cho không gian mẫu Ω , xét hai biến cố bất kỳ A và B . Xác suất của $A \cup B$ được viết là:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



Đặc biệt, nếu $A \cap B = \emptyset$ (tức là A và B là xung khắc), thì

$$P(A \cup B) = P(A) + P(B).$$

Các công thức xác suất

Xác suất của hội ba biến cố

Cho không gian mẫu Ω , xét hai biến cố bất kỳ A , B và C . Xác suất của $A \cup B \cup C$ được viết là:

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) \\ &\quad - P(B \cap C) + P(A \cap B \cap C). \end{aligned}$$

Đặc biệt, nếu $A \cap B = \emptyset$, $A \cap C = \emptyset$ và $B \cap C = \emptyset$ (tức là A , B và C là xung khắc nhau từng đôi một), thì

$$P(A \cup B \cup C) = P(A) + P(B) + P(C).$$

Xác suất của hội nhiều biến cố xung khắc

Cho không gian mẫu Ω , xét bộ biến cố A_1, A_2, \dots, A_k , sao cho, chúng xung khắc với nhau từng đôi một, tức là $A_i \cap A_j = \emptyset$, với mọi $i \neq j$ từ 1 tới k . Khi đó:

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

Các công thức xác suất

Xác suất điều kiện (cơ bản)

Cho không gian mẫu Ω , xét hai biến cố bất kỳ A và B . Xác suất xảy ra A khi biết B xảy ra, được viết là:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Ví dụ: Xét kết quả của 1 bài toán phân loại thư rác (spam) của 1 hệ thống lọc thư tự động.

		Nhận thật		Tổng
Kết quả	Thư thường	Thư rác		
	Thư thường	600	120	720
	Thư rác	50	400	450
Tổng		650	520	1170

Các công thức xác suất

- Đặt A là biến cỗ 1 bức thư được phân loại là thư rác.
- Đặt B là biến cỗ 1 bức thư có nhãn là thư thường.
 $\hookrightarrow A \cap B$: 1 bức thư là thư thường và bị phân thành thư rác.

Khi đó, từ bảng tổng hợp ta có

$$\begin{aligned} ■ P(A \cap B) &= \frac{50}{1170}; & ■ P(B) &= \frac{650}{1170} \end{aligned}$$

Xác suất mà ta quan tâm là $P(A|B)$, được tính như sau:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{50/1170}{650/1170} = \frac{50}{650} \approx 0.0769$$

Các công thức xác suất

Luật nhân xác suất

Cho không gian mẫu Ω , xét hai biến cố bất kỳ A và B . Xác suất của $A \cap B$ được viết là:

$$P(A \cap B) = P(A|B)P(B),$$

hay tương đương với

$$P(A \cap B) = P(B|A)P(A).$$

Công thức này được suy ra từ công thức xác suất điều kiện.

- Hai biến cố A và B là **độc lập** khi và chỉ khi $P(A \cap B) = P(A)P(B)$;
- Nếu A_1, A_2, \dots, A_k là độc lập, thì

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1)P(A_2) \dots P(A_k)$$

Các công thức xác suất

Ví dụ: Xét kết quả của 1 bài toán phân loại thư rác (spam) của 1 hệ thống lọc thư tự động.

		Nhận thật		Tổng
		Thư thường	Thư rác	
Kết quả	Thư thường	400	320	720
	Thư rác	250	200	450
Tổng		650	520	1170

- Đặt A là biến cố 1 bức thư được phân loại là thư rác.
- Đặt B là biến cố 1 bức thư có nhãn là thư thường.
 $\rightarrow A \cap B$: 1 bức thư là thư thường và bị phân thành thư rác.

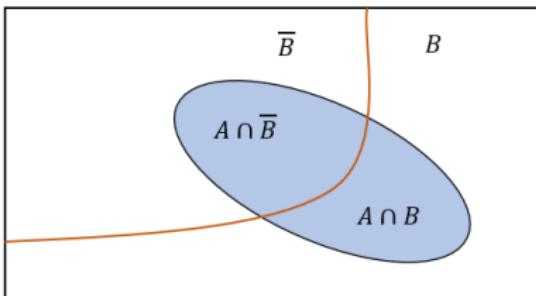
Từ bảng tổng hợp, ta chứng minh được A và B là độc lập.

Tức là việc phân loại thư không phụ thuộc vào đặc tính của bức thư.

Các công thức xác suất

Đôi khi xác suất của một biến cố A được cung cấp theo từng điều kiện của một biến cố B .

↪ Câu hỏi là làm sao tính xác suất của biến cố A , $P(A)$.



Thông qua biểu diễn bằng biểu đồ ven, ta nhận thấy rằng:

$$A = (A \cap \bar{B}) \cup (A \cap B)$$

Do đó,

$$\begin{aligned} P(A) &= P((A \cap \bar{B}) \cup (A \cap B)) \\ &= P(A \cap \bar{B}) + P(A \cap B) \\ &= P(A|\bar{B})P(\bar{B}) + P(A|B)P(B). \end{aligned}$$

Các công thức xác suất

Công thức tính xác suất $P(A)$ như trên, tức là:

$$P(A) = P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)$$

được gọi là **công thức xác suất toàn phần**.

Xác suất toàn phần (tổng quát)

Cho không gian mẫu Ω , xét dãy các biến cố B_1, B_2, \dots, B_k xung khắc nhau từng đôi một, và $B_1 \cup B_2 \cup \dots \cup B_k = \Omega$ (tức là hệ đầy đủ). Khi đó, xác suất của một biến cố A được xác định bởi:

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k).$$

Dịnh lý Bayes

Trong các phần trước, ta đã biết công thức tính xác suất điều kiện $P(A|B)$, trong đó:

- A là biến cố được quan tâm;
- B là biến cố điều kiện.

Tuy nhiên, trong một số trường hợp, ta lại mong muốn tính xác suất $P(B|A)$.

Về mặt bản chất,

$$P(B|A) = \frac{P(B \cap A)}{P(A)},$$

\Rightarrow ta cần $P(B \cap A)$ và $P(A)$.

Tuy nhiên, trong 1 số trường hợp, ta chỉ có

- $P(B)$;
- $P(A|B)$;
- $P(A|\bar{B})$.

Dịnh lý Bayes

Để giải quyết vấn đề này thiếu hụt thông tin, ta cần sử dụng một số biến đổi để đưa công thức về dạng có thông tin mà ta có.

- $P(B \cap A) \Rightarrow$ sử dụng luật nhân xác suất: $P(B \cap A) = P(A|B)P(B)$;
- $P(A) \Rightarrow$ sử dụng công thức xác suất toàn phần:

$$P(A) = P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)$$

Suy ra,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)}$$

Công thức này, là hoàn toàn tính được dựa trên các xác suất đã được cung cấp.

Dịnh lý Bayes

Công thức tính xác suất $P(B|A)$ như trên, tức là:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)}$$

được gọi là công thức xác suất Bayes - được tên theo nhà thống kê học người Anh, Thomas Bayes (1702 - 1761).

Dịnh lý Bayes

Cho không gian mẫu Ω , xét dãy các biến cố B_1, B_2, \dots, B_k xung khắc nhau từng đôi một, và $B_1 \cup B_2 \cup \dots \cup B_k = \Omega$ (tức là hệ đầy đủ). Khi đó, xác suất điều kiện của biến cố B_i khi biết một biến cố A bất kỳ, được xác định bởi:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \dots + P(A|B_k)P(B_k)}$$

với i là một số bất kỳ trong $1, 2, \dots, k$.

Dịnh lý Bayes

Ví dụ: Xét kết quả của 1 bài toán phân loại thư rác (spam) của 1 hệ thống lọc thư tự động.

- Đặt A là biến cố 1 bức thư được phân loại là thư rác.
- Đặt B là biến cố 1 bức thư có nhãn là thư rác.

Quá trình kiểm tra thực nghiệm với 1170 thư (trong đó, 650 thư thường và 520 thư rác), cho ra các báo cáo như sau:

- Khả năng phân loại đúng thư thường là: $P(\bar{A}|\bar{B}) = 0.923$
- Khả năng phân loại đúng thư rác là: $P(A|B) = 0.769$

Ta quan tâm tới xác suất:

Nếu một bức thư bị phân vào hộp thư rác thì khả năng nó là thực sự là thư rác là bao nhiêu?

Tức là, ta cần tính xác suất $P(B|A)$.

Áp dụng công thức Bayes ta tính được:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|\bar{B})P(\bar{B}) + P(A|B)P(B)} \approx 0.889$$

Như vậy, cứ 100 thư bị phân thành thư rác, thì có khoảng 11 thư là thông thường!

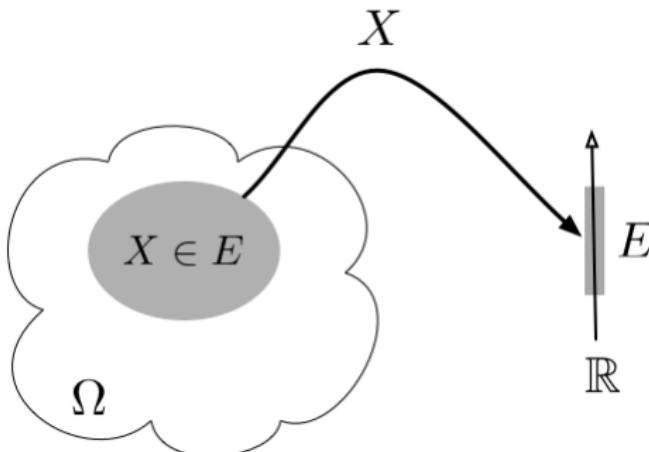
Biến ngẫu nhiên

Định nghĩa biến ngẫu nhiên - công thức toán học

Xét phép thử ngẫu nhiên, với không gian mẫu Ω . Gọi ω là một biến cỗ sơ cấp (hay kết quả) trong không gian mẫu Ω . Khi đó, hàm số X được xác định bởi:

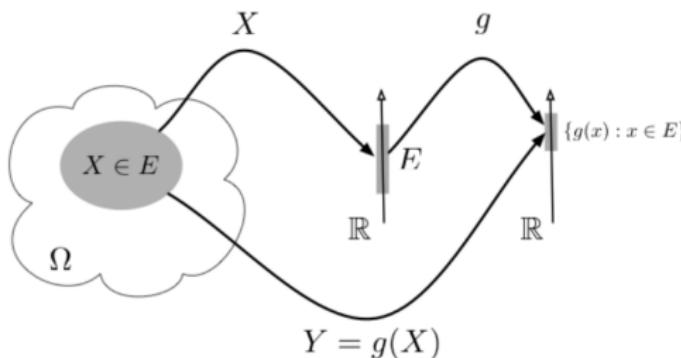
$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X(\omega) \end{aligned}$$

được gọi là một biến ngẫu nhiên.



Biến ngẫu nhiên

Một phép biến đổi được áp dụng trên 1 biến ngẫu nhiên được gọi là hàm ngẫu nhiên (random function)



Biến ngẫu nhiên

Xét về mặt giá trị số, ta cơ bản chia biến ngẫu nhiên thành hai loại:

- biến ngẫu nhiên rời rạc;
- biến ngẫu nhiên liên tục.

Biến ngẫu nhiên rời rạc

Biến ngẫu nhiên rời rạc là biến ngẫu nhiên có dạng số rời rạc nhau (số tự nhiên, số nguyên) có phạm vi hữu hạn (hoặc vô hạn) đếm được.

Ví dụ: số vết xước trên bề mặt, số các bộ phận bị lỗi trong 1000 bộ phận được kiểm tra, số lượng bit truyền nhận bị lỗi.

Ví dụ: 0, 1, 2, 3, 4, ...

Biến ngẫu nhiên liên tục

Biến ngẫu nhiên liên tục là một biến ngẫu nhiên có vô số kết quả có thể xảy ra, được biểu thị bằng một khoảng trên trực số thực hoặc toàn bộ trực số thực.

Ví dụ: chiều dài, áp suất, nhiệt độ, thời gian, điện áp, trọng lượng.

Ví dụ cho giá trị: 0.1, 0.123, 2.36, 3.85, 4.235, ...

Các đặc trưng của biến ngẫu nhiên rời rạc

Xét biến ngẫu nhiên rời rạc X , với không gian giá trị $S = \{x_1, x_2, \dots, x_k\}$, ta quan tâm tới

$$P(X = x),$$

với x là 1 giá trị bất kỳ trong S .

Ví dụ 1: (tiếp theo). Quan sát lại phép thử tung 1 đồng xu cân đối đồng chất 2 lần liên tiếp. Không gian mẫu

$$\Omega = \{SS, SN, NS, NN\}$$

Gọi X là biến ngẫu nhiên miêu tả số mặt sấp xuất hiện sau 2 lần tung liên tiếp. Không gian giá trị của X là

$$S = \{0, 1, 2\}.$$

Theo định nghĩa của X , ta có thể viết $X(NN) = 0$, $X(SN) = X(NS) = 1$ và $X(SS) = 2$. Do đó,

$$P(X = 0) = P(X(NN) = 0) = P(\{NN\}).$$

Theo công thức xác suất cổ điển,

$$P(X = 0) = P(X(NN) = 0) = P(\{NN\}) = \frac{1}{4}.$$

Tương tự, $P(X = 1) = P(\{SN\} \cup \{NS\}) = \frac{1}{2}$ và $P(X = 2) = P(\{SS\}) = \frac{1}{4}$.

Các đặc trưng của biến ngẫu nhiên rời rạc

Từ ví dụ trên, để tính xác suất $P(X = x)$ của một biến ngẫu nhiên rời rạc, ta cần quay về không gian mẫu với biến cố sơ cấp ban đầu Ω .

Từ kết quả trên, ta có thể lập một bảng tổng xác suất cho biến X như sau:

X		0	1	2
$P(X = x)$		0.25	0.5	0.25

Bảng này được gọi là bảng **phân phối xác suất** của biến ngẫu nhiên rời rạc.

Bảng phân phối xác suất

Xét biến ngẫu nhiên rời rạc X , giả sử các giá trị của X là x_1, x_2, \dots, x_k , và ký hiệu $P(X = x_i) = p_i$, với $i = 1, 2, \dots, k$. Khi đó, bảng phân phối xác suất của biến ngẫu nhiên rời rạc X là:

X		x_1	x_2	\dots	x_k
$P(X = x)$		p_1	p_2	\dots	p_k

Các đặc trưng của biến ngẫu nhiên rời rạc

Từ bảng phân phối xác suất, ta nhận thấy:

- $p_i \geq 0$ và $p_i \leq 1$, với mọi $i = 1, 2, \dots, k$;
- $p_1 + p_2 + \dots + p_k = 1$ (xác suất của toàn không gian mẫu);
- mỗi giá trị của x_i tương ứng một giá trị p_i .
 - ↪ $P(X = x_i) = p_i$ là một hàm số của x_i .
 - ↪ ta có thể viết $f(x_i) = p_i$, hàm f như vậy được gọi là hàm trọng lượng xác suất (*probability mass function*) - hay hàm xác suất.

Hàm trọng lượng xác suất

Xét biến ngẫu nhiên rời rạc X , với các giá trị có thể $x_1, x_2, x_3, \dots, x_k$, một hàm trọng lượng xác suất f của X là một hàm thỏa các tính chất sau:

- $f(x_i) = P(X = x_i)$;
- $0 \leq f(x_i) \leq 1$;
- $\sum_{i=1}^k f(x_i) = f(x_1) + f(x_2) + \dots + f(x_k) = 1$.

Trong một số trường hợp, hàm f có thể có công thức tường minh, nhưng cũng có thể chỉ được biểu thị qua bảng phân phối xác suất.

Các đặc trưng của biến ngẫu nhiên rời rạc

Ví dụ 1: (tiếp theo). Ta có bảng phân phối xác suất cho biến X như sau:

X		0	1	2
$P(X = x)$		0.25	0.5	0.25

Khi đó, hàm trọng lượng xác suất của X được xác định là: $f(0) = 0.25$, $f(1) = 0.5$ và $f(2) = 0.25$.

Các đặc trưng của biến ngẫu nhiên rời rạc

Trong phần trước, ta đã làm quen với

- biến ngẫu nhiên rời rạc X
- hàm trọng lượng xác suất $f(x)$ của X .

Ta cũng đã biết cách tính xác suất $P(X \leq x)$ bởi:

$$P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} f(x_i),$$

tức là tổng tất cả các xác suất $P(X = x_i)$ sao cho $x_i \leq x$, với $i = 1, 2, \dots, n$.

Nhận thấy rằng, $P(X \leq x)$ là:

- một hàm số của X thay đổi theo x ;
- tổng tích lũy của các xác suất của biến cố $\{X = x_i\}$.

Trong lý thuyết xác suất cho biến ngẫu nhiên, ta gọi $P(X \leq x)$ là *hàm phân phối tích lũy* của X .

Các đặc trưng của biến ngẫu nhiên rời rạc

Hàm phân phối tích lũy

Cho biến ngẫu nhiên rời rạc X , với các giá trị có thể x_1, x_2, \dots, x_n và hàm trọng lượng xác suất $f(x)$. Hàm phân phối xác suất tích lũy (*cumulative distribution function*) $F(x)$ của X là

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

Thỏa những điều kiện sau:

- $0 \leq F(x) \leq 1$;
- nếu $x \leq y$ thì $F(x) \leq F(y)$.

Ghi nhớ

Ta có một số công thức tính xác suất dựa theo $F(x)$ như sau:

- $P(X \leq x) = F(x)$;
- $P(X > x) = 1 - P(X \leq x) = 1 - F(x)$;
- $P(x \leq X \leq y) = P(X \leq y) - P(X \leq x - 1) = F(y) - F(x - 1)$.

Các đặc trưng của biến ngẫu nhiên rời rạc

Ví dụ 1: (tiếp theo). Xét lại biến ngẫu nhiên X biểu thị số mặt sấp xuất hiện sau 2 lần tung liên tiếp 1 đồng xu cân đối đồng chất:

X	0	1	2
$f(x)$	0.25	0.5	0.25

Ta dễ dàng tính được:

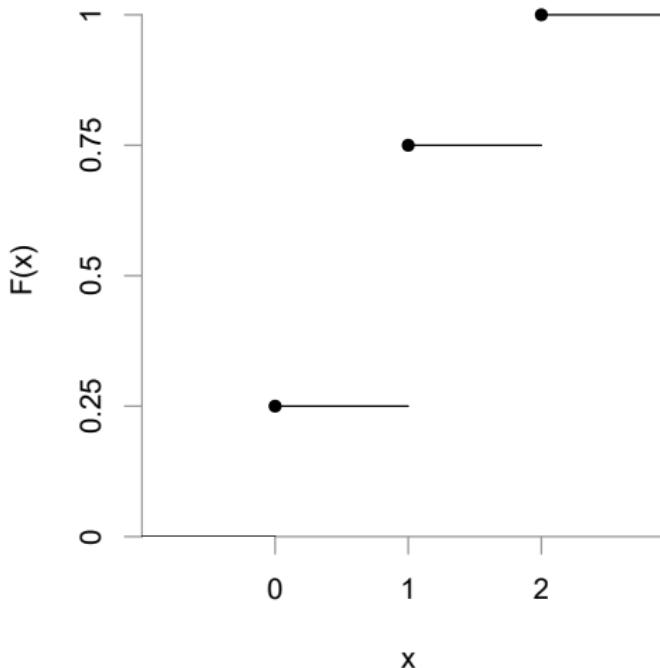
- $P(X < 0) = 0,$
- $F(0) = 0.25,$
- $F(1) = 0.25 + 0.5 = 0.75,$
- $F(2) = 0.75 + 0.25 = 1,$
- $F(3) = 1.$

Khi đó, ta biểu diễn hàm phân phối tích lũy như sau:

$$F(x) = \begin{cases} 0 & \text{nếu } x < 0 \\ 0.25 & \text{nếu } 0 \leq x < 1 \\ 0.75 & \text{nếu } 1 \leq x < 2 \\ 1 & \text{nếu } x \geq 2 \end{cases}$$

Các đặc trưng của biến ngẫu nhiên rời rạc

Ta vẽ minh họa $F(x)$ trên biểu đồ như sau



Các đặc trưng của biến ngẫu nhiên rời rạc

Kỳ vọng của biến ngẫu nhiên rời rạc

Xét biến ngẫu nhiên rời rạc X với $S = \{x_1, x_2, \dots, x_k\}$, và hàm trọng lượng xác suất $f(x)$. **Kỳ vọng** (hay **trung bình**) của X được ký hiệu là $\mathbb{E}(X)$ hay μ_X , xác định bởi:

$$\mathbb{E}(X) = \sum_{i=1}^k x_i f(x_i).$$

Ví dụ 1: (tiếp theo). Xét lại biến ngẫu nhiên X biểu thị số mặt sấp xuất hiện sau 2 lần tung liên tiếp 1 đồng xu cân đối đồng chất:

X	0	1	2
$f(x)$	0.25	0.5	0.25

Khi đó, trung bình của X được tính bởi:

$$\mathbb{E}(X) = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1$$

Các đặc trưng của biến ngẫu nhiên rời rạc

Phương sai của biến ngẫu nhiên rời rạc

Phương sai của X được ký hiệu là $\text{Var}(X)$ hay σ_X^2 , xác định bởi:

$$\text{Var}(X) = \sum_{i=1}^k (x_i - \mathbb{E}(X))^2 f(x_i)$$

hay tương đương với

$$\text{Var}(X) = \sum_{i=1}^k x_i^2 f(x_i) - \mathbb{E}(X)^2.$$

Căn bậc hai của phương sai, $\sqrt{\text{Var}(X)}$, được gọi là **độ lệch chuẩn** của X , ký hiệu là σ_X .

Các đặc trưng của biến ngẫu nhiên rời rạc

Ví dụ 1: (tiếp theo). Xét lại biến ngẫu nhiên X biểu thị số mặt sấp xuất hiện sau 2 lần tung liên tiếp 1 đồng xu cân đối đồng chất:

X		0	1	2
$f(x)$		0.25	0.5	0.25

Khi đó, trung bình của X được tính bởi:

$$\mathbb{V}\text{ar}(X) = (0^2 \times 0.25 + 1^2 \times 0.5 + 2^2 \times 0.25) - 1^2 = 0.5$$

Các đặc trưng của biến ngẫu nhiên rời rạc

Ta có một số tính chất cần ghi nhớ sau của trung bình và phương sai:

- $\mathbb{E}(c) = c$, với c là một hằng số thực bất kỳ.
- $\mathbb{E}(cX + b) = c\mathbb{E}(X) + b$, với X là một biến ngẫu nhiên bất kỳ, c và b là hai hằng số.
- Cho hai biến ngẫu nhiên X và Y , ta có

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

- Cho hai biến ngẫu nhiên X và Y , nếu X và Y là độc lập thì:

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

- $\text{Var}(X) \geq 0$ và $\sigma = \sqrt{\text{Var}(X)} \geq 0$.
- $\text{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$.

Các đặc trưng của biến ngẫu nhiên rời rạc

- $\text{Var}(c) = 0$, với c là một hằng số thực bất kỳ.
- $\text{Var}(cX + b) = c^2\text{Var}(X)$, với X là một biến ngẫu nhiên bất kỳ, c và b là hai hằng số.
- Cho $h(X)$ là một hàm số của X , khi đó

$$\mathbb{E}(h(X)) = h(x_1)f(x_1) + h(x_2)f(x_2) + \dots + h(x_n)f(x_n),$$

và

$$\text{Var}(h(X)) = \mathbb{E}(h(X)^2) - \mathbb{E}(h(X))^2.$$

- Nếu X có đơn vị (ví dụ, m, s, VND, kg, ...) thì trung bình và độ lệch chuẩn của X sẽ có cùng đơn vị với X , trong khi đó, đơn vị của phương sai sẽ là bình phương của đơn vị ban đầu (ví dụ, m^2 , s^2 , VND^2 , ...).
- Cho hai biến ngẫu nhiên X và Y , nếu X và Y là độc lập thì:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y).$$

Phân phối Bernoulli

Như đã giới thiệu, biến ngẫu nhiên là hàm gán một số thực cho mỗi kết quả trong không gian mẫu Ω của một phép thử ngẫu nhiên.

Một phép thử ngẫu nhiên cụ thể sẽ cho ta một biến ngẫu nhiên cụ thể.

Ví dụ 4: Xét các phép thử ngẫu nhiên sau:

- Tung 1 đồng xu cân đối đồng chất 1 lần. Gọi X là số lần mặt sấp xuất hiện.
- Chọn ngẫu nhiên 1 sinh viên trong lớp. Gọi X là số sinh viên nam được chọn.
- Chọn ngẫu nhiên 1 người ở tp. Hồ Chí Minh. Gọi X là số người bị bệnh tiểu đường.
- Truyền đi một bit thông tin. Gọi X là số bít được truyền đúng.

Trong các phép thử trên, $X = 0$ hoặc $X = 1$, luôn luôn.

→ những phép thử như trên được gọi là phép thử Bernoulli¹.

¹được đặt theo họ của nhà Toán học Jacob Bernoulli (1655 - 1705), người Thụy Sĩ

Phân phối Bernoulli

Phép thử Bernoulli

Một phép thử ngẫu nhiên gồm nhiều phép thử, trong đó, mỗi phép thử là độc lập nhau, và chỉ có thể có hai giá trị, thì khi đó mỗi phép thử được gọi là một phép thử Bernoulli.

Phân phối Bernoulli

Xét một phép thử Bernoulli, gọi X là biến ngẫu nhiên biểu thị số lần thành công (ví dụ, số lần xuất hiện mặt sấp, số sinh viên nam, ...):

- $X \in \{0, 1\}$, với $X = 1$ nếu phép thử thành công;
- $P(X = 1) = p$ và $P(X = 0) = 1 - p$.

↪ X được gọi là một biến ngẫu nhiên Bernoulli, với hàm xác suất

$$f(x) \equiv P(X = x) = p^x(1 - p)^{1-x},$$

$x = 0, 1$. Khi đó, ta viết $X \sim B(p)$, tức là, X tuân theo phân phối Bernoulli với xác suất thành công p . Ta có:

- $\mathbb{E}(X) = p$;
- $\text{Var}(X) = p(1 - p)$.

Phân phối nhị thức

Ví dụ 5: Xét phép thử ngẫu nhiên, truyền 4 bít thông tin. Gọi X là tổng số bit thông tin bị truyền lỗi trong tổng số 4 bit:

- mỗi bit thông tin là độc lập nhau;
- chỉ có hai kết quả có thể xảy ra: lỗi (E) hoặc đúng (C);
- trong 1 bit được truyền, nếu xuất hiện lỗi được gọi là thành công;
 ↳ ta đang quan sát chuỗi 4 phép thử Bernoulli.

Bảng giá trị của X :

Kết quả	x	Kết quả	x
CCCC	0	ECEC	2
CCCE	1	EECC	2
CCEC	1	CEEC	2
CECC	1	CEEE	3
ECCC	1	ECEE	3
CCEE	2	EECE	3
CECE	2	EEECC	3
ECCE	2	EEEE	4

Tính $P(X = 2)$, $P(X = 3)$. Biết xác suất xảy ra lỗi $P(E) = 0.1$.

Phân phối nhị thức

Từ bảng tổng hợp ta có, tập hợp các kết quả đê $X = 2$ là:

$$A = \{CCEE, CECE, ECCE, ECEC, EECC, CEEC\}.$$

Khi đó, vì các biến cố này là đôi một xung khắc nhau, nên

$$\begin{aligned} P(X = 2) \equiv P(A) &= P(\{CCEE\}) + P(\{CECE\}) + P(\{ECCE\}) \\ &\quad + P(\{ECEC\}) + P(\{EECC\}) + P(\{CEEC\}). \end{aligned}$$

Vì các bit thông tin là độc lập nên ta có:

$$P(\{CCEE\}) = P(C)P(C)P(E)P(E) = 0.9^2 \times 0.1^2 = 0.0081$$

Do đó, ta tính được

$$P(X = 2) = 6 \times 0.0081 = 0.0486$$

Phân phối nhị thức

Từ bảng tổng hợp ta có, tập hợp các kết quả để $X = 3$ là:

$$B = \{CEEE, ECEE, EECE, EEEC\}.$$

Khi đó, vì các biến cố này là đôi một xung khắc nhau, nên

$$\begin{aligned} P(X = 3) \equiv P(B) &= P(\{CEEE\}) + P(\{ECEE\}) + P(\{EECE\}) \\ &\quad + P(\{EEEC\}). \end{aligned}$$

Vì các bit thông tin là độc lập nên ta có:

$$P(\{CEEE\}) = P(C)P(C)P(E)P(E) = 0.9^1 \times 0.1^3 = 0.0009$$

Do đó, ta tính được

$$P(X = 3) = 4 \times 0.0009 = 0.0036$$

Phân phối nhị thức

Từ các kết quả trên,

- $P(X = 2) = 6 \times 0.1^2 \times 0.9^2$
- $P(X = 3) = 4 \times 0.1^3 \times 0.9^1$

ta thấy rằng 6 hoặc 4 là số lần kết quả cho ra giá trị $x = 2$ hoặc $x = 3$.

Sử dụng phép tính tổ hợp:

- $x = 2$ chọn được 2 bit lỗi trong tổng số 4 bit
 \Rightarrow có $C_4^2 = 6$ cách chọn;
- $x = 3$ chọn được 3 bit lỗi trong tổng số 4 bit
 \Rightarrow có $C_4^3 = 4$ cách chọn;

Ta có thể tổng quát hóa công thức trên:

$$P(X = x) = C_4^x \times 0.1^x \times 0.9^{4-x}$$

Áp dụng công thức này, ta dễ dàng tính được $P(X = 0)$, $P(X = 1)$, $P(X = 4)$.

Biến X trong ví dụ này được gọi là biến ngẫu nhiên nhị thức.

Phân phối nhị thức

Phân phối nhị thức

Một phép thử ngẫu nhiên gồm n phép thử Bernoulli, sao cho

- các phép thử là độc lập nhau;
- mỗi phép thử chỉ có hai khả năng có thể xảy ra, được đánh nhãn là thành công hoặc thất bại;
- xác suất thành công của một lần thử được ký hiệu là p .

Đặt biến ngẫu nhiên X biểu thị số lần thành công trong n lần thử, khi đó, X là một biến ngẫu nhiên nhị thức (*binomial random variable*) với 2 tham số $0 < p < 1$ và $n = 1, 2, \dots$, được ký hiệu $X \sim B(n, p)$. Hàm xác suất của X :

$$f(x) = P(X = x) = C_n^x p^x (1 - p)^{n-x},$$

với $x = 0, 1, 2, \dots, n$.

Chú ý: tên phân phối nhị thức được đặt dựa theo khai triển nhị thức Newton bởi vì tổng của hàm xác suất tại tất cả các điểm giá trị x , là tương đương với khai triển nhị thức Newton của p và $1 - p$

$$C_n^0 p^0 (1 - p)^n + C_n^1 p^1 (1 - p)^{n-1} + \dots + C_n^n p^n (1 - p)^0 = (p + (1 - p))^n$$

và bằng 1.

Phân phối nhị thức

Kỳ vọng và phương sai của phân phối nhị thức

Xét biến ngẫu nhiên $X \sim B(n, p)$, với $n \in N$ và $0 < p < 1$. Khi đó, ta có:

$$\mathbb{E}(X) = np, \quad \text{và} \quad \text{Var}(X) = np(1 - p)$$

Chứng minh:

- Theo định nghĩa kỳ vọng của biến ngẫu nhiên rời rạc:

$$E(X) = \sum_{x=0}^n xf(x) = \sum_{x=0}^n xC_n^x p^x (1-p)^{n-x} = \sum_{x=1}^n xC_n^x p^x (1-p)^{n-x}.$$

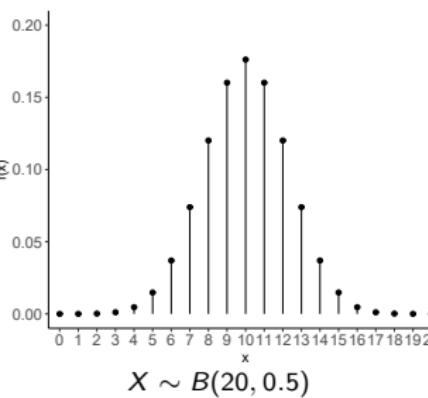
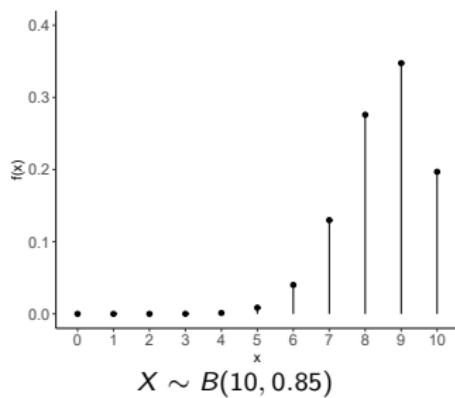
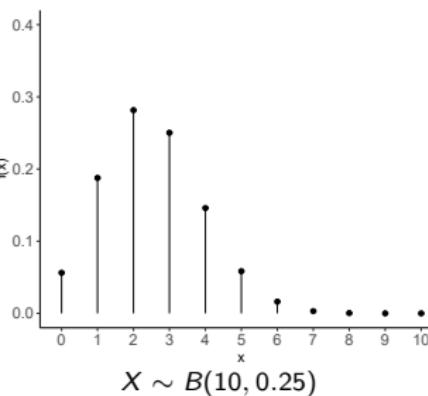
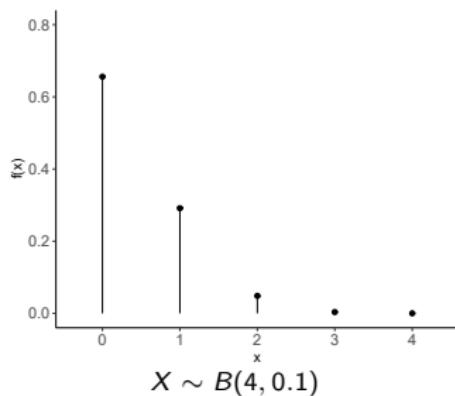
Nhận xét rằng, $xC_n^x = nC_{n-1}^{x-1}$, biến đổi về dạng nhị thức Newton cho p và $1 - p$, sau đó, ta thu được kết quả.

- Áp dụng công thức phương sai cho biến ngẫu nhiên rời rạc:

$$\text{Var}(X) = E(X^2) - E(X)^2 = \sum_{x=1}^n x^2 C_n^x p^x (1-p)^{n-x} - n^2 p^2.$$

Nhận xét rằng, $x^2 C_n^x = xnC_{n-1}^{x-1}$, biến đổi về dạng nhị thức Newton cho p và $1 - p$, sau đó, ta thu được kết quả.

Phân phối nhị thức



Phân phối nhị thức

Định lý

Xét $X \sim B(n, p)$, trong đó $\lim_{n \rightarrow +\infty} np = \lambda$, với $\lambda > 0$. Khi đó, ta có:

$$\lim_{n \rightarrow +\infty} P(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!},$$

với $x = 0, 1, 2, \dots$.

Chứng minh:

Ta có

$$P(X = x) = C_n^x p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}.$$

Do đó,

$$\lim_{n \rightarrow +\infty} P(X = x) = \lim_{n \rightarrow +\infty} \frac{n!}{k!(n-k)!} p^x (1 - p)^{n-x}.$$

Áp dụng tính chất $\lim_{n \rightarrow +\infty} np = \lambda$, ta thu được

$$\lim_{n \rightarrow +\infty} P(X = x) = \lim_{n \rightarrow +\infty} \frac{n!}{x!(n-x)!} \frac{\lambda^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} = \frac{\lambda^x \exp(-\lambda)}{x!}.$$

Phân phối Poisson

Phân phối Poisson được xây dựng để mô hình quá trình đếm trong 1 khoảng thời gian vô hạn.

Giả sử ta đếm số lượng khách hàng tới 1 cửa hàng trong 1 khoảng thời gian.

- Để thực hiện đếm, ta chia khung thời gian thành n đoạn đều nhau, với n lớn.
- Trong mỗi khoảng thời gian i , ta thực hiện 1 thí nghiệm Bernoulli, với xác suất 1 người xuất hiện là p .
- Giả sử rằng việc khách hàng xuất hiện tại các khoảng thời gian i là độc lập nhau.
- Trong mỗi khoảng i , ta thu được $X_i \sim Ber(p)$.

Khi đó, $X_n = \sum_{i=1}^n X_i \sim B(n, p)$, với $E(X_n) = np$.

Khi $n \rightarrow +\infty$, ta định rằng $\lim_{n \rightarrow +\infty} np = \lambda$ tức là tỷ số khách hàng xuất hiện trong khung thời gian T .

Phân phối Poisson

Cho qua giới hạn $n \rightarrow +\infty$, ta có

$$\lim_{n \rightarrow +\infty} P(X_n = x) = \frac{\lambda^x \exp(-\lambda)}{x!}.$$

Hàm xác suất này được gọi là phân phối Poisson (đặt theo tên của nhà Toán học người pháp Siméon Denis Poisson, 1781 - 1840).

Phân phối Poisson

Gọi X là biến ngẫu nhiên miêu tả số lần một đối tượng xuất hiện trong một khoảng thời gian t của một khoảng thời gian quan sát T . Khi đó, giá trị của X là $0, 1, 2, 3, \dots$. Nếu hàm xác suất của X được miêu tả bởi:

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

với tham số $\lambda > 0$ và $x = 0, 1, 2, 3, \dots$, thì ta nói X là biến ngẫu nhiên Poisson, ký hiệu là $X \sim P(\lambda)$. Ta tính được:

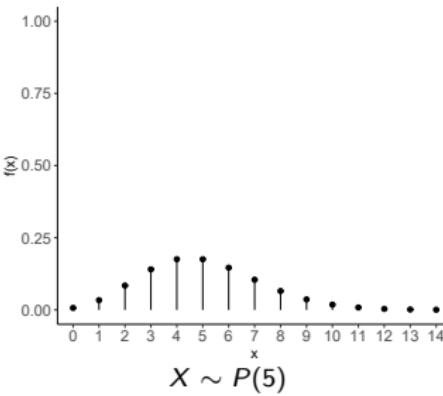
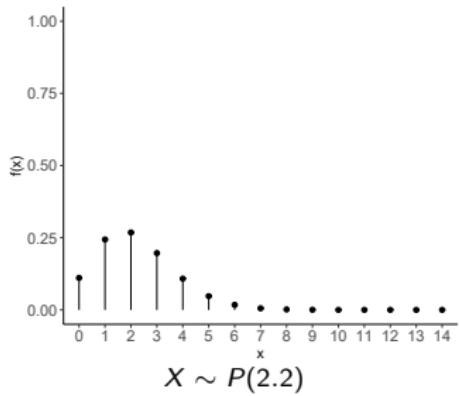
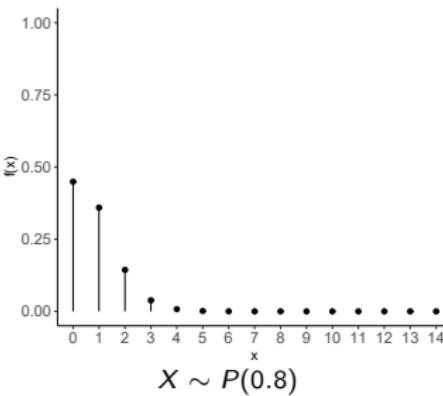
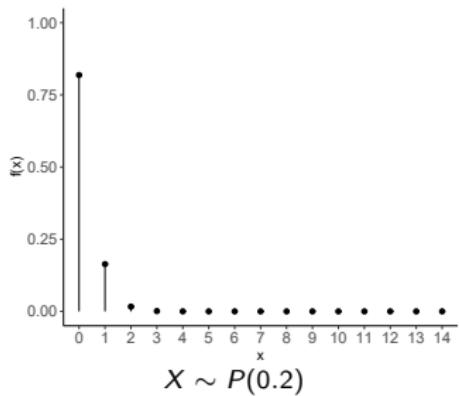
$$\mathbb{E}(X) = \text{Var}(X) = \lambda.$$

Phân phối Poisson

Thông thường, phân phối Poisson thường được dùng để mô hình hóa các biến:

- Số lỗi in trong một (hoặc một số) trang sách.
- Số người sống lâu trên 100 tuổi trong một cộng đồng dân cư.
- Số người đến một bưu điện nào đó trong một ngày.
- Số tai nạn hoặc sự cố giao thông xảy ra tại một điểm giao thông trong một ngày.
- Số đột biến gen mỗi chiều dài DNA.
- Số cuộc gọi đến trung tâm mỗi giờ.

Phân phối Poisson



Các đặc trưng của biến ngẫu nhiên liên tục

Biến ngẫu nhiên liên tục

Biến ngẫu nhiên X là **liên tục** khi không gian giá trị S của X là một khoảng số thực, có dạng (a, b) hoặc là toàn không gian số thực \mathbb{R} .

Do số lượng giá trị có thể của X là vô hạn, nên:

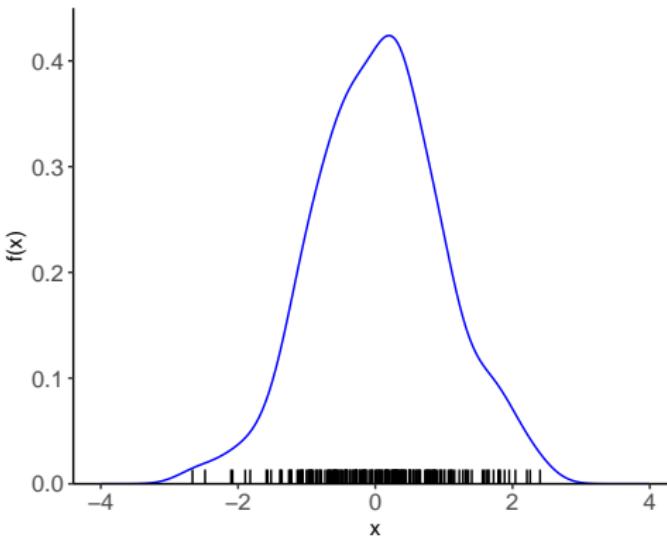
- $P(X = x) = 0$;
- không thể sử dụng bảng phân phối xác suất để mô tả X .

Thay vào đó, ta biểu diễn X :

- trên trục số thực \mathbb{R} , kèm theo
- hàm $f(x)$ miêu tả mật độ tập trung của các giá trị của X trong một khoảng nhất định.

Các đặc trưng của biến ngẫu nhiên liên tục

Ví dụ 6: Xét là một biến ngẫu nhiên liên tục X có miền giá trị khả kiến $(-4, 4)$. Ta biểu diễn mật độ tập trung của các giá trị x bởi hàm $f(x)$ như trong hình.



Giá trị $f(x)$ càng lớn thì mật độ tập trung các giá trị xung quanh x càng dày đặc.

Các đặc trưng của biến ngẫu nhiên liên tục

Hàm mật độ xác suất

Hàm miêu tả mật độ tập trung của một biến ngẫu nhiên liên X trong không gian giá trị S , được gọi là *hàm mật độ xác suất*, ký hiệu là $f(x)$, với các tính chất sau:

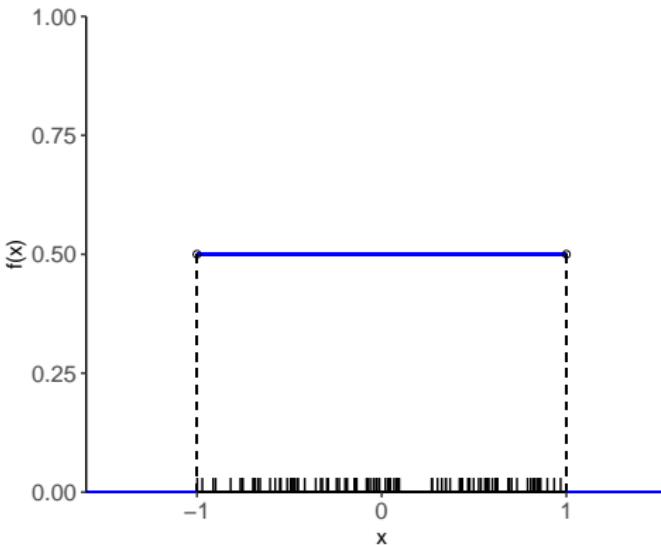
- $f(x)$ là khả tích;
- $f(x) \geq 0$, với mọi giá trị của x ;
- $\int_S f(x) dx = 1$;
- $P(a \leq X \leq b) = \int_a^b f(x) dx$, tức là diện tích phía dưới đường cong xác định bởi hàm $f(x)$ trong khoảng $[a, b]$ của X .

Hàm mật độ xác suất $f(x)$ được mô tả theo công thức tường minh.

Các đặc trưng của biến ngẫu nhiên liên tục

Ví dụ 7: Xét biến ngẫu nhiên liên tục X có không gian giá trị S trên đoạn $[-1, 1]$, với hàm mật độ xác suất:

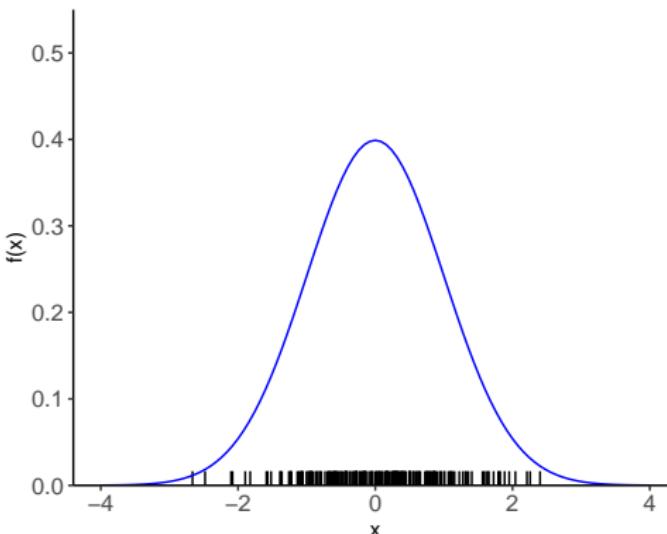
$$f(x) = \begin{cases} 0.5 & \text{nếu } -1 \leq x \leq 1; \\ 0 & \text{nếu } x < -1 \text{ hoặc } x > 1. \end{cases}$$



Các đặc trưng của biến ngẫu nhiên liên tục

Ví dụ 8: Xét biến ngẫu nhiên liên tục X có không gian giá trị $S \equiv \mathbb{R}$, với hàm mật độ xác suất:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$



Các đặc trưng của biến ngẫu nhiên liên tục

Hàm phân phối tích lũy liên tục

Xét biến ngẫu nhiên liên tục X có không gian giá trị \mathbb{R} và hàm mật độ $f(x)$.
Hàm phân phối tích lũy của X được định nghĩa:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(s) ds.$$

Ta có một số tính chất của $F(x)$ như sau:

- $F(x)$ là hàm liên tục, khả vi;
- đạo hàm của $F(x)$ là $f(x)$, tức là $\frac{dF(x)}{dx} = f(x)$.

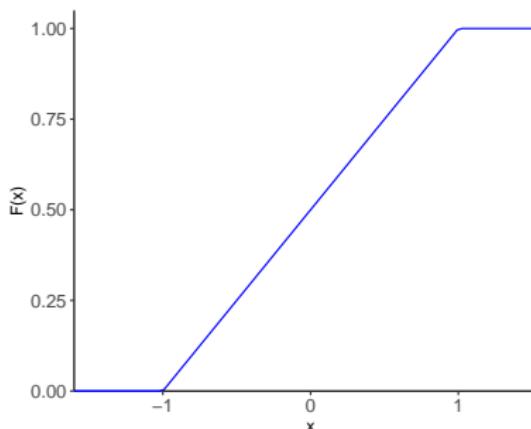
Các đặc trưng của biến ngẫu nhiên liên tục

Ví dụ 7: (tiếp theo). Xét biến ngẫu nhiên liên tục X có không gian giá trị S trên đoạn $[-1, 1]$, với hàm mật độ xác suất:

$$f(x) = \begin{cases} 0.5 & \text{nếu } -1 \leq x \leq 1; \\ 0 & \text{nếu } x < -1 \text{ hoặc } x > 1. \end{cases}$$

Ta tính được:

$$F(x) = \begin{cases} 0 & \text{nếu } x < -1 \\ \frac{x+1}{2} & \text{nếu } -1 \leq x \leq 1; \\ 1 & \text{nếu } x > 1. \end{cases}$$



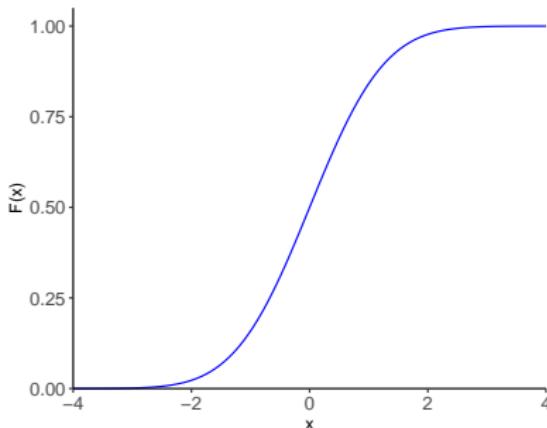
Các đặc trưng của biến ngẫu nhiên liên tục

Ví dụ 8: Xét biến ngẫu nhiên liên tục X có không gian giá trị $S \equiv \mathbb{R}$, với hàm mật độ xác suất:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Ta tính được:

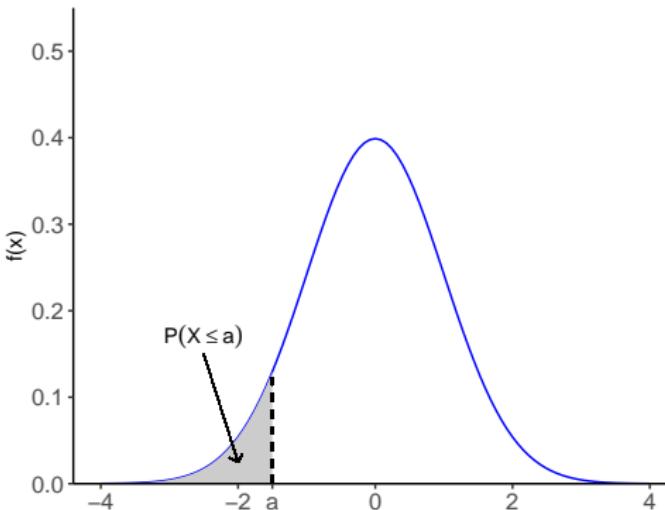
$$F(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds$$



Các đặc trưng của biến ngẫu nhiên liên tục

Xác suất $P(X \leq a)$ là diện tích bên dưới hàm mật độ $f(x)$

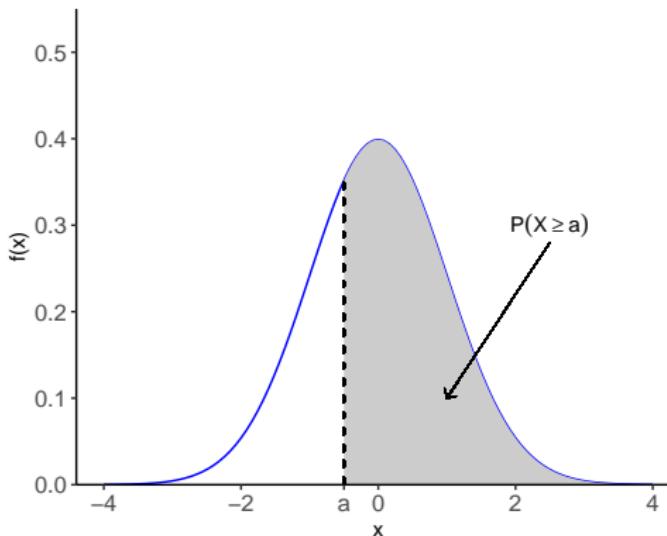
$$P(X \leq a) = \int_{-\infty}^a f(x) dx = F(a)$$



Các đặc trưng của biến ngẫu nhiên liên tục

Xác suất $P(X \geq a)$ là diện tích bên dưới hàm mật độ $f(x)$

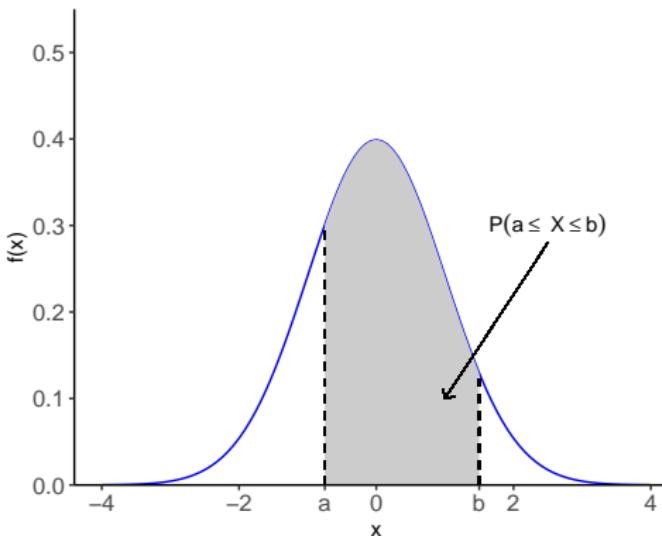
$$P(X \geq a) = \int_a^{\infty} f(x) dx = 1 - P(X \leq a) = 1 - F(a)$$



Các đặc trưng của biến ngẫu nhiên liên tục

Xác suất $P(a \leq X \leq b)$ là diện tích bên dưới hàm mật độ $f(x)$

$$P(a \leq X \leq b) = \int_a^b f(x) dx = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$



Các đặc trưng của biến ngẫu nhiên liên tục

Chú ý: do tính liên tục của hàm $F(x)$, ta có:

$$P(X < a) = P(X \leq a),$$

tương tự

$$P(a < X < b) = P(a \leq X \leq b).$$

Các đặc trưng của biến ngẫu nhiên liên tục

Cũng giống như biến ngẫu nhiên rời rạc, trong trường hợp biến ngẫu nhiên liên tục, ta có xét 2 đặc trưng quan trọng:

- kỳ vọng
- phương sai (hoặc độ lệch chuẩn).

Kỳ vọng và phương sai của biến ngẫu nhiên liên tục

Xét X là một biến ngẫu nhiên liên tục trên \mathbb{R} , có hàm mật độ xác suất $f(x)$. Khi đó, kỳ vọng (hay trung bình) của X , được ký hiệu là $\mathbb{E}(X)$ hay μ_X :

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

Phương sai của X được ký hiệu là $\text{Var}(X)$ hay σ_X^2 :

$$\text{Var}(X) = \int_{-\infty}^{+\infty} (x - \mathbb{E}(X))^2 f(x) dx = \int_{-\infty}^{+\infty} x^2 f(x) dx - \mathbb{E}(X)^2.$$

Căn bậc 2 của phương sai, $\sqrt{\text{Var}(X)}$ được gọi là **độ lệch chuẩn** của X , ký hiệu là σ_X .

Phân phối Chuẩn

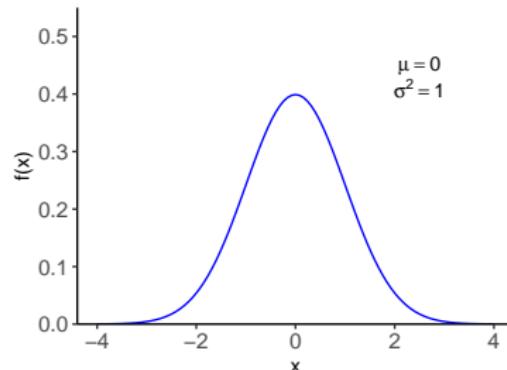
Phân phối chuẩn

Một biến ngẫu nhiên liên tục X được gọi là tuân theo *phân phối chuẩn - normal distribution* nếu hàm mật độ:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

với $-\infty < x < +\infty$, và 2 tham số $-\infty < \mu < +\infty$, $\sigma > 0$. Hàm phân phối tích lũy của X là:

$$F(x) = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(s-\mu)^2}{2\sigma^2}\right) ds.$$



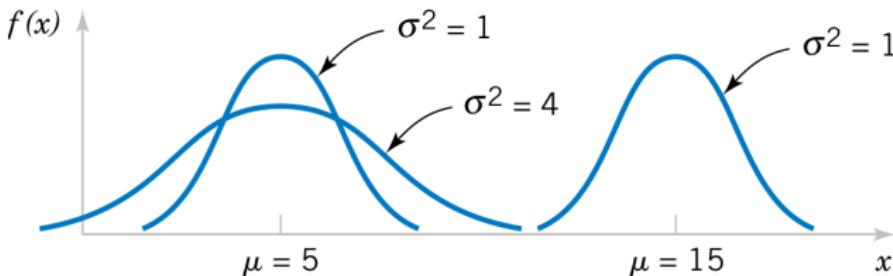
Trung bình và phương sai lần lượt là $\mathbb{E}(X) = \mu$ và $\text{Var}(X) = \sigma^2$.

Ta ký hiệu $X \sim N(\mu, \sigma^2)$, tức là X tuân theo phân phối chuẩn với trung bình μ và phương sai σ^2 .

Phân phối Chuẩn

Khi ta thay đổi giá trị của μ và σ^2 ta sẽ có các dạng khác nhau của $f(x)$:

- cố định σ , thay đổi μ vị trí của $f(x)$ sẽ di chuyển theo vị trí của μ ;
- cố định μ , tăng hoặc giảm σ , độ rộng của $f(x)$ sẽ tăng hoặc giảm,
- hình dạng hàm mật độ xác suất $f(x)$ luôn là hình chuông (*bell shape*), đối xứng qua μ ,

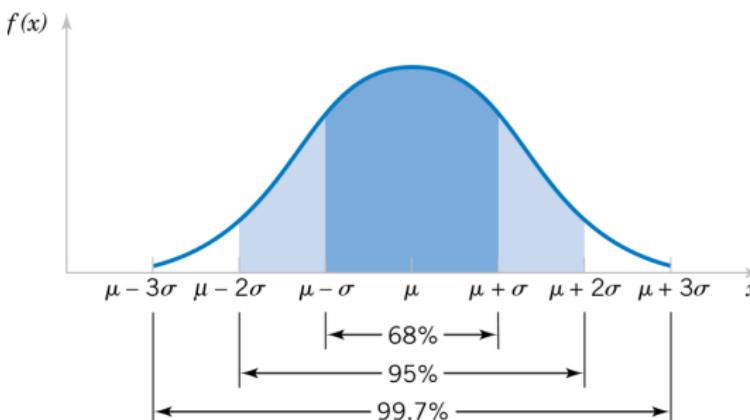


Phân phối Chuẩn

Cho $X \sim N(\mu, \sigma^2)$, ta có các kết quả sau:

- $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.6827,$
- $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545,$
- $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973.$

Kết quả này được gọi là “quy tắc $k\sigma$ ”.



Phân phối Chuẩn

Phân phối chuẩn tắc - Gauss

Biến ngẫu nhiên $Z \sim N(\mu, \sigma^2)$, nếu $\mu = 0$ và $\sigma = 1$, ta nói Z tuân theo phân phối chuẩn tắc (Gauss), $Z \sim N(0, 1)$. Hàm phân phối tích lũy của Z được ký hiệu là $\Phi(z)$:

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

Ta có các tính chất:

- $\Phi(a) = 1 - \Phi(-a)$, với a bất kỳ,
- $P(Z > a) = 1 - \Phi(a)$,
- $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$,

Phân phối Chuẩn

Phép đổi biến chuẩn tắc

Xét $X \sim N(\mu, \sigma^2)$, ta có một phép đổi biến chuẩn tắc:

$$Z = \frac{X - \mu}{\sigma}$$

khi đó $Z \sim N(0, 1)$.

Phép đổi biến này cho phép ta tính xác suất của X thông qua Z , cụ thể:

- $P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P\left(Z \leq \frac{x - \mu}{\sigma}\right) = \Phi\left(\frac{x - \mu}{\sigma}\right).$
- $P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{b - \mu}{\sigma}\right) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right),$
- $P(X \geq x) = 1 - P(X \leq x) = 1 - \Phi\left(\frac{x - \mu}{\sigma}\right),$
- $P(X \leq a) = \Phi\left(\frac{a - \mu}{\sigma}\right) = 1 - \Phi\left(-\frac{a - \mu}{\sigma}\right).$

Phân vị của phân phối xác suất

Xét biến ngẫu nhiên X , có hàm phân phối xác suất $F_X(x)$.

Phân vị thứ p (p -th quantile) hay $100p$ percentile là điểm q sao cho $F_X(q) = p$.

Ví dụ: phân vị 0.5 (trung vị) là điểm q sao cho $F_X(q) = 0.5$.

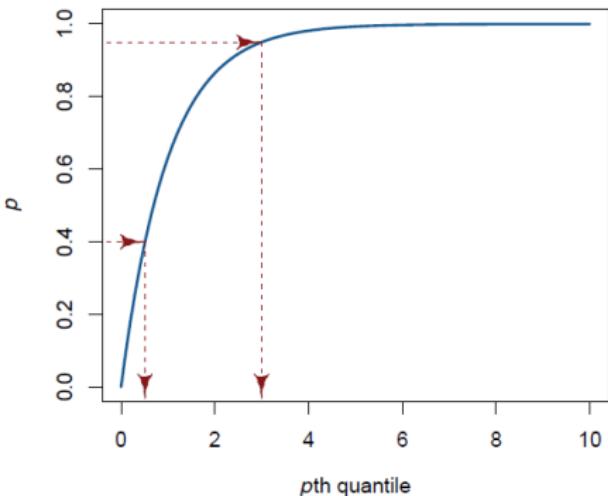
- Nếu hàm $F_X(x)$ là liên tục thì

$$q = F_X^{-1}(p)$$

- Nếu hàm $F_X(x)$ là rời rạc thì

$$q = \inf \{t : F_X(t) \geq p\}.$$

Phân vị của phân phối xác suất



Phép đổi biến và áp dụng

Cho biến ngẫu nhiên liên tục X có hàm c.d.f $F_X(x)$.

Xét biến đổi ngẫu nhiên $Y = F_X(X)$, khi đó Y chỉ có giá trị trong $[0, 1]$.

Hàm c.d.f của Y được xác định bởi:

$$F_Y(y) = P(Y \leq y) = P(F_X(Y) \leq y) = P(Y \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

Do đó, $F_Y(y)$ là hàm cdf của phân phối đều trên $[0, 1]$.

Hơn nữa theo cách đặt, ta có

$$X = F_X^{-1}(Y).$$

Do đó, nếu ta tạo ngẫu nhiên các giá trị y của Y theo $U(0, 1)$ (phân phối đều trên $[0, 1]$), thì ta hoàn toàn xác định được $x = F_X^{-1}(y)$.

Phép đổi biến và áp dụng

Ví dụ: Xét biến $X \sim \text{Exp}(\lambda)$ (Phân phối mũ). Hàm mật độ xác suất là

$$f(x) = \lambda e^{-\lambda x},$$

với $x \geq 0$, và tham số $\lambda > 0$. Hàm xác suất tích lũy được cho bởi

$$F_X(x) = \begin{cases} 1 - \exp(-\lambda x), & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Từ đây ta có hàm ngược $F_X^{-1}(u)$

$$F_X^{-1}(u) = -\log(1-u)/\lambda.$$

Như vậy, cho trước $\lambda > 0$,

- ta tạo ngẫu nhiên $U \sim \mathcal{U}(0, 1)$;
- xác định X thông qua $F_X^{-1}(U)$.

Phép đổi biến và áp dụng

Khi **không** có công thức tường minh của $F_X^{-1}(\cdot) \Rightarrow$ xấp xỉ hàm này.

Ta sử dụng phương pháp nội suy tuyến tính (linear interpolation)

- dựa trên vùng xác định của hàm mật độ f ;
- tạo một lưới x_1, \dots, x_m ;
- tính xấp xỉ $u_i = F_X(x_i)$;
- tạo ngẫu nhiên $U \sim \mathcal{U}(0, 1)$
- xác định X theo đường tuyến tính nội suy giữa hai điểm liền kề $u_i \leq U \leq u_j$

$$X = \frac{u_j - U}{u_j - u_i}x_i + \frac{U - u_i}{u_j - u_i}x_j.$$

Ví dụ: Áp dụng phương pháp này để tạo biến ngẫu nhiên theo phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ với hàm cdf:

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) dx.$$

Phép đổi biến và áp dụng

Xét biến ngẫu nhiên liên tục X có pdf là $f_X(x)$ và cdf $F_X(x)$.

Xét g là một hàm one-to-one, liên tục và khả vi. Ta có biến đổi $Y = g(X)$.

Khi đó, ta có cdf của Y được xác định bởi:

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Thêm vào đó, thực hiện đạo hàm cho hàm $F_Y(y)$ ta thu được pdf của Y :

$$f_Y(y) = \frac{1}{|g'(g^{-1}(y))|} f_X(g^{-1}(y)),$$

kết quả này có được nhờ áp dụng quy tắc Chain rule và đạo hàm cho hàm ngược.

1 *Giới thiệu về Khoa học thống kê*

2 *Xác suất và Biến ngẫu nhiên*

3 *Vector ngẫu nhiên*

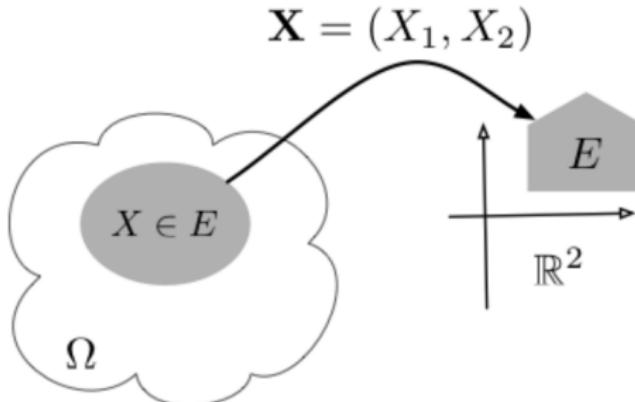
Vector ngẫu nhiên

Định nghĩa vector ngẫu nhiên - công thức toán học

Xét phép thử ngẫu nhiên, với không gian mẫu Ω . Gọi ω là một biến cố sơ cấp (hay kết quả) trong không gian mẫu Ω . Khi đó, vector $\mathbf{X} = (X_1, \dots, X_p)$ gồm p biến ngẫu nhiên được xác định bởi:

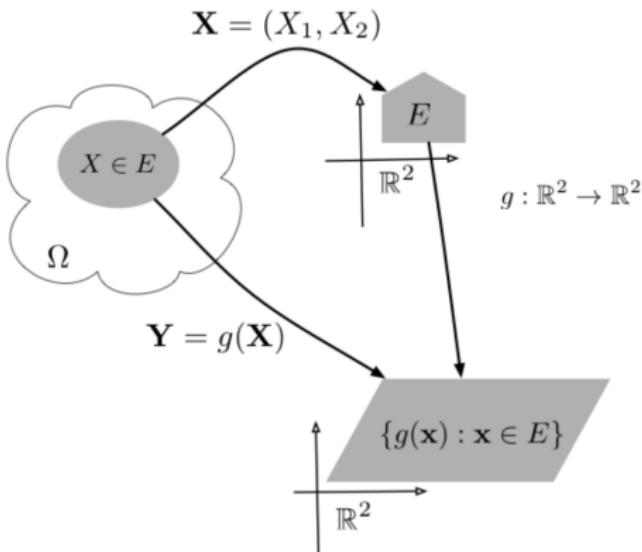
$$\begin{aligned} X_i : \Omega &\rightarrow \mathbb{R} \\ \omega &\mapsto X_i(\omega) \end{aligned}$$

với $i = 1, \dots, p$, và do đó, $\mathbf{X} : \Omega \rightarrow \mathbb{R}^p$.



Vector ngẫu nhiên

Một phép biến đổi được áp dụng trên 1 vector ngẫu nhiên được gọi là hàm vector ngẫu nhiên (function of random vector)



Vector ngẫu nhiên

Xét về mặt giá trị số, ta cơ bản chia vector ngẫu nhiên thành ba loại:

- vector ngẫu nhiên **rời rạc** - tất cả các biến thành phần là rời rạc;
- vector ngẫu nhiên **liên tục** - tất cả các biến thành phần là liên tục;
- vector ngẫu nhiên **hỗn hợp** - biến thành phần là liên tục hoặc rời rạc;

Ví dụ: xét nghiên cứu về tập dữ liệu của khách hàng với các biến:

- X_1 : tuổi \Rightarrow liên tục
- X_2 : chiều cao \Rightarrow liên tục
- X_3 : giới tính \Rightarrow rời rạc
- X_4 : thu nhập \Rightarrow liên tục
- X_5 : tình trạng sức khỏe \Rightarrow rời rạc

Khi đó,

- $\mathbf{X} = (X_1, X_2, X_4)$ là vector ngẫu nhiên liên tục.
- $\mathbf{Y} = (X_3, X_5)$ là vector ngẫu nhiên rời rạc.
- $\mathbf{Z} = (X_1, X_2, X_3)$ là vector ngẫu nhiên hỗn hợp.

Vector ngẫu nhiên

Ta định nghĩa

- xác suất của vector ngẫu nhiên liên tục là

$$P(\mathbf{X} \in \mathbf{S}) = P(X_1 \in S_1, X_2 \in S_2, \dots, X_p \in S_p),$$

với $\mathbf{S} = S_1 \times S_2 \times \dots \times S_p$, S_i là các miền xác định giá trị cho biến X_i .

- xác suất của vector ngẫu nhiên rời rạc là

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1, X_2 = x_2, \dots, X_p = x_p),$$

với $\mathbf{x} = (x_1, x_2, \dots, x_p)$.

Vector ngẫu nhiên

Ví dụ: Xét vec tơ $\mathbf{X} = (X_1, X_2) \in [-1, 1] \times [-1, 1]$, với $X_i \sim \mathcal{U}(-1, 1)$, $i = 1, 2$.

Đặt C là đường tròn bán kính 1, tâm tại $(0, 0)$, tức là

$$C = \{(x_1, x_2) \in [-1, 1] \times [-1, 1] : x_1^2 + x_2^2 \leq 1\}$$

Xác suất quan sát các điểm (x_1, x_2) thuộc vào hình tròn đơn vị C là:

$$P((X_1, X_2) \in C) = P\left(X_1 \in [-1, 1], X_2 \in \left[-\sqrt{1 - x_1^2}, \sqrt{1 - x_1^2}\right]\right).$$

Phân phối đồng thời

Xét vector ngẫu nhiên $\mathbf{X} = (X_1, \dots, X_p)$.

Hàm xác suất tích lũy đồng thời (joint cdf) được định nghĩa trong 2 trường hợp:

- vector liên tục

$$F_{\mathbf{X}}(\mathbf{x}) = P(X_1 \leq x_1, \dots, X_p \leq x_p);$$

- vector rời rạc

$$F_{\mathbf{X}}(\mathbf{x}) = \sum_{y_1 \leq x_1} \dots \sum_{y_p \leq x_p} P(X_1 = y_1, \dots, X_p = y_p).$$

Trong trường hợp vector liên tục, ta xác định hàm mật độ xác suất đồng thời (joint pdf) là

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^p}{\partial x_1 \dots \partial x_p} F_{\mathbf{X}}(x_1, \dots, x_p),$$

Trong trường hợp vector rời rạc, ta xác định hàm trọng lượng xác suất đồng thời (joint pmf) là

$$p_{\mathbf{X}}(\mathbf{x}) = P(X_1 = x_1, \dots, X_p = x_p).$$

Phân phối đồng thời

Một số tính chất

1. Tích phân trên toàn không gian của joint pdf bằng 1:

$$\int_{\mathbb{R}^p} \dots \int_{\mathbb{R}^p} f_{\mathbf{x}}(x_1, \dots, x_p) dx_1 \dots dx_p = 1.$$

2. Tổng trên toàn không gian của joint pmf bằng 1:

$$\sum_{x_1} \dots \sum_{x_p} p_{\mathbf{x}}(\mathbf{x}) = 1.$$

3. Joint cdf được xác định bởi

$$F_{\mathbf{x}}(\mathbf{x}) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_p} f_{\mathbf{x}}(t_1, \dots, t_p) dt_1 \dots dt_p$$

4. Joint cdf là hàm đơn điệu tăng trên \mathbb{R} :

$$F_{\mathbf{x}}(a_1, \dots, a_p) \leq F_{\mathbf{x}}(b_1, \dots, b_p),$$

với $a_i \leq b_i$, với mọi $i = 1, \dots, p$.

Phân phối đồng thời

5. Giới hạn đồng thời:

$$\lim_{x_1 \rightarrow -\infty} \dots \lim_{x_p \rightarrow -\infty} F_{\mathbf{X}}(x_1, \dots, x_p) = 0,$$

và

$$\lim_{x_1 \rightarrow +\infty} \dots \lim_{x_p \rightarrow +\infty} F_{\mathbf{X}}(x_1, \dots, x_p) = 1,$$

6. Xác suất của vector liên tục

$$P(\mathbf{X} \in A) = \int \dots \int_A f_{\mathbf{X}}(x_1, \dots, x_p) dx_1 \dots dx_p.$$

7. Xác suất của vector liên tục

$$P(\mathbf{X} \in A) = \sum_{\mathbf{x} \in A} p_{\mathbf{X}}(x_1, \dots, x_p).$$

Phân phối lề

Hàm mật độ xác suất lề (marginal pdf) là hàm mật độ xác suất của một biến X_i trong vector ngẫu nhiên

- vector liên tục

$$f_{X_i}(x_i) = \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} f_{\mathbf{X}}(x_1, \dots, x_p) dx_1 \dots dx_p,$$

không lấy tích phân đối với biến x_i .

- vector rời rạc

$$p_{X_i}(x_i) = \sum_{x_1} \dots \sum_{x_p} p_{\mathbf{X}}(x_1, \dots, x_p),$$

không lấy tổng đối với biến x_i .

Phân phối lề

Ví dụ: Xét vector ngẫu nhiên (X_1, X_2) có joint pdf như sau

$$f_{X_1, X_2}(x_1, x_2) = \begin{cases} \frac{1}{\pi}, & \text{nếu } x^2 + y^2 \leq 1, \\ 0, & \text{chỗ khác.} \end{cases}$$

Khi đó, ta tính được hàm mật độ lề cho X_1 là

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_{X_1, X_2}(x_1, x_2) dx_2 = \frac{2}{\pi} \sqrt{1 - x_1^2},$$

nếu $-1 < x_1 < 1$ và $f_{X_1}(x_1) = 0$ nếu $|x_1| \geq 1$.

Tương tự, ta tính được hàm mật độ lề cho X_2 là

$$f_{X_2}(x_2) = \begin{cases} \frac{2}{\pi} \sqrt{1 - x_2^2}, & \text{nếu } -1 < x_2 < 1, \\ 0, & \text{chỗ khác.} \end{cases}$$

Phân phối điều kiện

Nhắc lại công thức xác suất điều kiện

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Xét vector ngẫu nhiên rời rạc (X_1, X_2) . Đặt

- biến cố A là " $X_1 = x_1$ ",
- biến cố B là " $X_2 = x_2$ ".

Áp dụng công thức xác suất điều kiện, ta có

$$P(X_1 = x_1 | X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{p_{X_1, X_2}(x_1, x_2)}{p_{X_2}(x_2)},$$

với $p_{X_1, X_2}(x_1, x_2)$ là joint pmf, $p_{X_2}(x_2)$ là hàm trọng lượng xác suất lè của X_2 .

Phân phối điều kiện

Định nghĩa 1

Xét vector ngẫu nhiên rời rạc, $\mathbf{X} = (X_1, X_2)$, khi đó, hàm trọng lượng xác suất điều kiện của X_1 khi biết $X_2 = x_2$ được ký hiệu là $p_{X_1|X_2=x_2}(x_1)$ và được xác định bởi

$$p_{X_1|X_2=x_2}(x_1) = P(X_1 = x_1 | X_2 = x_2) = \frac{p_{X_1, X_2}(x_1, x_2)}{p_{X_2}(x_2)}.$$

Hàm phân bố điều kiện của X_1 khi biết $X_2 = x_2$ là

$$F_{X_1|X_2=x_2}(x_1) = \sum_{y \leq x_1} p_{X_1|X_2=y}(x_1).$$

Thêm vào đó, áp dụng định nghĩa của xác suất lề và công thức Bayes ta có:

$$p_{X_1|X_2=x_2}(x_1) = \frac{p_{X_1, X_2}(x_1, x_2)}{\sum_y p_{X_2|X_1=y}(x_2)p_{X_1}(y)}.$$

Phân phối điều kiện

Định nghĩa 2

Xét vector ngẫu nhiên liên tục, $\mathbf{X} = (X_1, X_2)$, khi đó, hàm mật độ xác suất điều kiện của X_1 khi biết $X_2 = x_2$ được ký hiệu là $f_{X_1|X_2=x_2}(x_1)$ và được xác định bởi

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1, X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

Hàm phân bố điều kiện của X_1 khi biết $X_2 = x_2$ là

$$F_{X_1|X_2=x_2}(x_1) = \int_{-\infty}^{x_1} f_{X_1|X_2=x_2}(s)ds.$$

Thêm vào đó, áp dụng định nghĩa của xác suất lề và công thức Bayes ta có:

$$f_{X_1|X_2=x_2}(x_1) = \frac{f_{X_1, X_2}(x_1, x_2)}{\int_{-\infty}^{x_1} f_{X_2|X_1=s}(x_2) f_{X_1}(s) ds}.$$

Phân phối điều kiện

Trường hợp vector ngẫu nhiên hỗn hợp.

Xét vector ngẫu nhiên (X_1, X_2) xác định trên không gian $(0, +\infty) \times \{1, 2\}$, sao cho với một tập bất kỳ $A \subset (0, +\infty)$ và $x_2 = 1, 2$, ta có hàm xác suất đồng thời:

$$P(X_1 \in A, X_2 = x_2) = \frac{1}{2} \int_A x_2 \exp(-x_1 x_2) dx_1.$$

Phân phối đồng thời của (X_1, X_2) không phải liên tục, cũng không phải rời rạc.

Ta xác định được xác suất lề cho X_1 như sau:

$$\begin{aligned} F_{X_1}(x_1) &= P(X_1 \leq x_1) = P(X_1 \leq x_1, X_2 = 1) + P(X_1 \leq x_1, X_2 = 2) \\ &= 1 - \frac{1}{2} [\exp(-x_1) + \exp(-2x_1)], \end{aligned}$$

với $x_1 \in (0, +\infty)$.

Xác suất lề cho X_2

$$P(X_2 = x_2) = P(X_1 \in (0, +\infty), X_2 = 1) = \frac{1}{2} \int_0^{+\infty} x_2 \exp(-x_1 x_2) dx_1 = \frac{1}{2}.$$

Phân phối điều kiện

Xác suất có điều kiện của X_1 khi biết $X_2 = x_2$ là

$$F_{X_1|X_2=x_2}(x) = P(X_1 \leq x_1 | X_2 = x_2) = 1 - \exp(-x_1 x_2),$$

với $x_1 \in (0, +\infty)$.

Xác suất có điều kiện của X_2 khi biết $X_1 = x_1$ là

$$p_{X_2|X_1=x_1}(x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{f_{X_1}(x_1)} = \frac{f_{X_1|X_2=x_2}(x_1) p_{X_2}(x_2)}{f_{X_1}(x_1)}.$$

Nhận xét rằng,

$$f_{X_1|X_2=x_2}(x_1) = \frac{d}{dx_1} F_{X_1|X_2=x_2}(x_1) = x_2 \exp(-x_1 x_2),$$

với $x_1 \in (0, +\infty)$ và

$$f_{X_1}(x_1) = \frac{d}{dx_1} F_{X_1}(x_1) = \frac{1}{2} [\exp(-x_1) + \exp(-2x_1)]$$

Khi đó, ta thu được

$$p_{X_2|X_1=x_1}(x_2) = \frac{x_2 \exp(-x_1 x_2)}{\exp(-x_1) + \exp(-2x_1)},$$

với $x_2 = 1, 2$.

Kỳ vọng, ma trận hiệp phương sai

Xét vector ngẫu nhiên $\mathbf{X} = (X_1, \dots, X_p)$, ta định nghĩa kỳ vọng đồng thời của (X_1, \dots, X_p) như sau:

- vector ngẫu nhiên rời rạc

$$\mathbb{E}(X_1 \dots X_p) = \sum_{x_1} \dots \sum_{x_p} x_1 \dots x_p p_{\mathbf{X}}(x_1, \dots, x_p);$$

- vector ngẫu nhiên liên tục

$$\mathbb{E}(X_1 \dots X_p) = \int \dots \int_{\mathbb{R}^p} x_1 \dots x_p f_{\mathbf{X}}(x_1, \dots, x_p) dx_1 \dots dx_p.$$

Kỳ vọng đồng thời, $\mathbb{E}(X_1 \dots X_p)$ là một số.

Kỳ vọng, ma trận hiệp phương sai

Hơn nữa, dựa vào thông tin của xác suất lề cho X_j , ta định nghĩa được kỳ vọng cho từng biến X_j

$$\mathbb{E}(X_j) = \int_{\mathbb{R}} x_j f_{X_j}(x_j) u dx_j,$$

hoặc

$$\mathbb{E}(X_j) = \sum_{x_j} x_j p_{X_j}(x_j)$$

trong trường hợp rời rạc.

Khi đó ta có vector kỳ vọng $\mathbb{E}(\mathbf{X}) = \mu_{\mathbf{X}} = (\mathbb{E}(X_1), \dots, \mathbb{E}(X_p))$.

Kỳ vọng, ma trận hiệp phương sai

Ma trận hiệp phương sai (covariance matrix) của \mathbf{X} được định nghĩa bởi

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \dots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_1, X_2) & \text{Var}(X_2) & \dots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_p) & \text{Cov}(X_2, X_p) & \dots & \text{Var}(X_p) \end{pmatrix},$$

là ma trận đối xứng, xác định dương và khả nghịch, trong đó

- $\text{Var}(X_j)$ là phương sai của biến ngẫu nhiên thành phần X_j ;
- $\text{Cov}(X_j, X_k)$ là hiệp phương sai (covariance) giữa thành phần X_j và X_k :

$$\text{Cov}(X_j, X_k) = \mathbb{E}[(X_j - \mathbb{E}(X_j))(X_k - \mathbb{E}(X_k))] = \mathbb{E}(X_j X_k) - \mathbb{E}(X_j)\mathbb{E}(X_k).$$

Kỳ vọng, ma trận hiệp phương sai

Xét vector ngẫu nhiên $\mathbf{X} = (X_1, X_2)$, ta có định nghĩa kỳ vọng điều kiện như sau:

- vector ngẫu nhiên rời rạc

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_{X_1 | X_2 = x_2} = \sum_{x_1} x_1 p_{X_1 | X_2 = x_2}(x_1);$$

- vector ngẫu nhiên liên tục

$$\mathbb{E}(X_1 | X_2 = x_2) = \mu_{X_1 | X_2 = x_2} = \int_{\mathbb{R}} x_1 f_{X_1 | X_2 = x_2}(x_1) dx_1.$$

Ta có tính chất:

$$\mathbb{E}_{X_2}(\mathbb{E}(X_1 | X_2)) = \mathbb{E}(X_2).$$

Kỳ vọng, ma trận hiệp phương sai

Xét vector ngẫu nhiên $\mathbf{X} = (X_1, X_2)$, ta có định nghĩa phương sai điều kiện như sau:

- vector ngẫu nhiên rời rạc

$$\text{Var}(X_1 | X_2 = x_2) = \sigma_{X_1 | X_2 = x_2}^2 = \sum_{x_1} x_1^2 p_{X_1 | X_2 = x_2}(x_1) - \mu_{X_1 | X_2 = x_2}^2;$$

- vector ngẫu nhiên liên tục

$$\text{Var}(X_1 | X_2 = x_2) = \sigma_{X_1 | X_2 = x_2}^2 = \int_{\mathbb{R}} x_1^2 f_{X_1 | X_2 = x_2}(x_1) dx_1 - \mu_{X_1 | X_2 = x_2}^2.$$

Xem thêm trong sách của [Severini \(2005\)](#) và của [Gut \(2009\)](#).

Sự độc lập của hai biến ngẫu nhiên

Xét vector ngẫu nhiên rời rạc $\mathbf{X} = (X_1, X_2)$. Ta nói biến ngẫu nhiên X_1 và X_2 là độc lập nếu và chỉ nếu:

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2),$$

với mọi cặp giá trị (x_1, x_2) trên không gian xác định.

Trong trường hợp liên tục, X_1 và X_2 là độc lập nếu và chỉ nếu, 1 trong 3 kết quả sau là đúng:

- với mọi giá trị của x_1 và x_2 trên \mathbb{R}

$$F_{X_1, X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2),$$

- với mọi giá trị của x_1 và x_2 trên \mathbb{R}

$$f_{X_1, X_2}(x_1, x_2) = f_{X_1}(x_1)f_{X_2}(x_2),$$

- với mọi hàm thực bị chặn g_1 và g_2 :

$$\mathbb{E}[g_1(X_1)g_2(X_2)] = \mathbb{E}[g_1(X_1)]\mathbb{E}[g_2(X_2)].$$

Sự độc lập của hai biến ngẫu nhiên

Hết quả

Khi hai biến ngẫu nhiên X_1 và X_2 là độc lập thì

$$\mathbb{E}(X_1 X_2) = \mathbb{E}(X_1) \mathbb{E}(X_2),$$

và do đó $\text{Cov}(X_1, X_2) = 0$. Tuy nhiên, điều ngược lại là không đúng.

Ví dụ: Đặt X là một biến ngẫu nhiên theo phân phối Laplace với hàm mật độ

$$f(x) = \frac{1}{2} \exp(-|x|),$$

với $-\infty < x < \infty$. Khi đó, ta tính được $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = 2$ và $\mathbb{E}(X^3) = 0$.

Đặt $Y = X^2$, khi đó, $\mathbb{E}(Y) = \mathbb{E}(X^2) = 2$ và

$$\text{Cov}(X, Y) = \mathbb{E}[X(Y - 2)] = \mathbb{E}[X^3 - 2X] = 0.$$

Tuy nhiên, thực tế thì Y là phụ thuộc vào X .

Sự tương quan

Sự không độc lập

Nếu X và Y là không độc lập, khi đó, với mọi x, y , ta có

$$F(x, y) = F_{Y|X}(y|x)F_X(x) = F_{X|Y}(x|y)F_Y(y).$$

Tức là, với một giá trị của X ta có thể suy ra luật phân phối cho Y (hoặc ngược lại). $\Rightarrow X$ và Y có quan hệ với nhau.

Ta có một số loại phụ thuộc như sau:

- phụ thuộc tuyến tính;
- phụ thuộc phi tuyến tính;
- phụ thuộc nhân quả (causality);
- sự liên hệ.

Trong đó, phụ thuộc tuyến tính và phụ thuộc phi tuyến xuất hiện khi xét mối quan hệ giữa hai biến định lượng.

Sự tương quan

Đặc biệt, mối quan hệ giữa hai biến định lượng được gọi là **sự tương quan - correlation**.

Như vậy, sự tương quan là một dạng của sự phụ thuộc (hay mối quan hệ):

- tương quan tuyến tính;

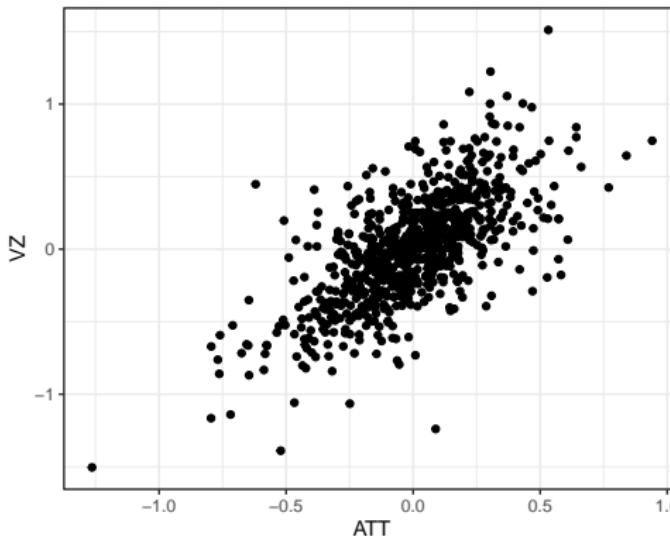
- tương quan phi tuyến tính.

→ ta có thể mô hình Y bởi $g(X)$, với g là một hàm tuyến tính hoặc phi tuyến.

Sự tương quan

Tương quan tuyến tính

Hai biến định lượng được gọi là có tương quan tuyến tính nếu các cặp giá trị quan sát của hai biến thể hiện một xu hướng tăng (hoặc giảm) tuyến tính, tức là theo một đường thẳng.

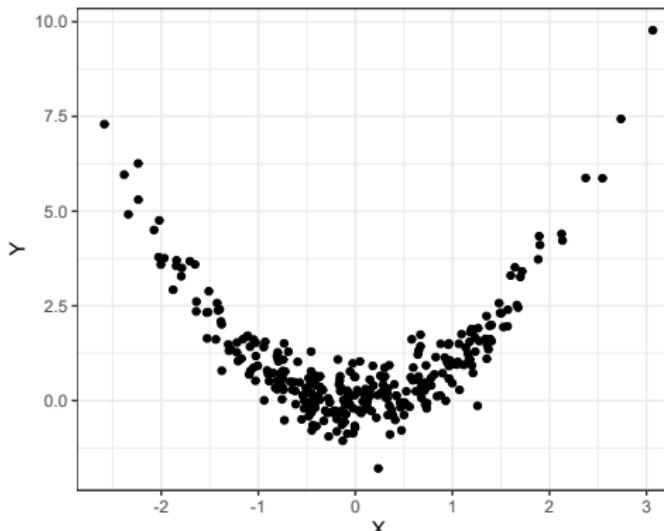


Sự tương quan

Tương quan phi tuyến tính

Hai biến định lượng được gọi là có tương quan phi tuyến tính nếu các cặp giá trị quan sát của hai biến thể hiện một xu hướng tăng (hoặc giảm) theo một hàm phi tuyến:

- đơn điệu (monotonic function),
- phi đơn điệu.



Sự tương quan

Sự tương quan giữa hai biến ngẫu nhiên, có thể được đo độ mạnh trong một số trường hợp cụ thể:

- tương quan tuyến tính,
- tương quan phi tuyến đơn điệu.

Thăng đo độ mạnh của sự tương quan, được gọi là **hệ số tương quan - correlation coefficient**.

Có nhiều hệ số tương quan khác nhau đã được phát triển, mỗi loại tương ứng cho một trường hợp cụ thể:

- hệ số tương quan tuyến tính Pearson - **Pearson correlation coefficient**, dành cho tương quan tuyến tính,
- hệ số tương quan hạng Spearman - **Spearman's correlation coefficient**, dành cho tương quan phi tuyến đơn điệu.

Sự tương quan

Hệ số tương quan tuyến tính Pearson

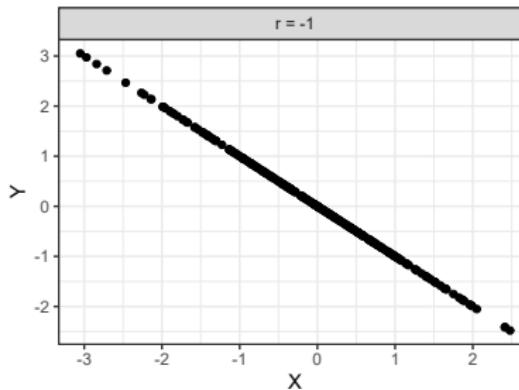
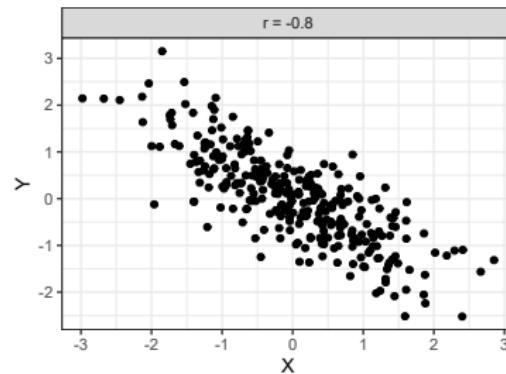
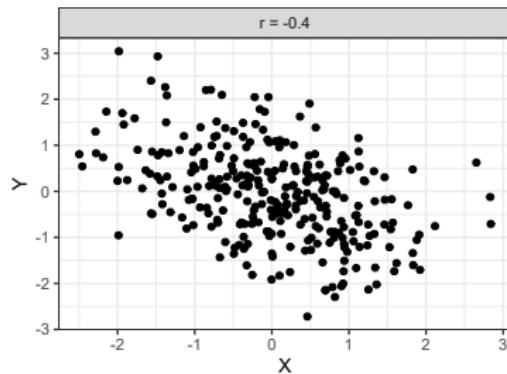
Hệ số tương quan tuyến tính Pearson - Pearson correlation coefficient là một hệ số dùng để đo độ mạnh của một tương quan tuyến tính giữa hai biến ngẫu nhiên liên tục X và Y :

$$r = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}.$$

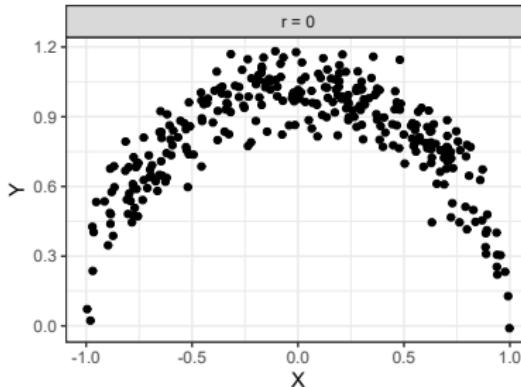
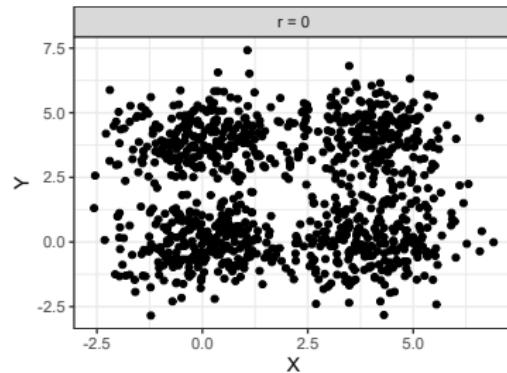
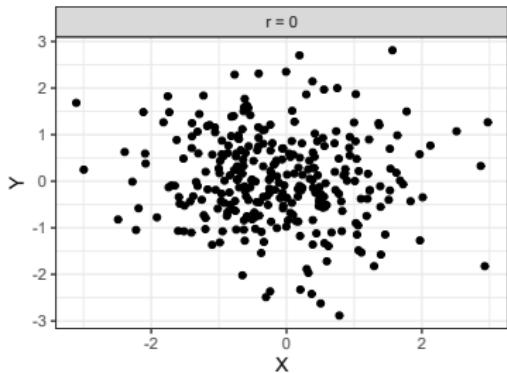
Hệ số r có giá trị nằm trong $[-1, 1]$:

- nếu $-1 < r < 0$: tương quan tuyến tính giữa X và Y là tương quan nghịch, tức là X tăng thì Y giảm;
- nếu $0 < r < 1$: tương quan tuyến tính giữa X và Y là tương quan thuận, tức là X tăng thì Y tăng;
- nếu $r = 0$: tương quan tuyến tính giữa X và Y là không tồn tại;
- nếu $r = -1$: tương quan tuyến tính giữa X và Y là tương quan nghịch ngặt;
- nếu $r = 1$: tương quan tuyến tính giữa X và Y là tương quan thuận ngặt;

Sự tương quan



Sự tương quan



Cách hệ số tương quan tuyến tính hoạt động

Nhắc lại rằng, với hai biến ngẫu nhiên X và Y , ta có:

- $\text{Cov}(X, Y) = \mathbb{E}((X - \mu_X)(Y - \mu_Y))$;
- $Z_X = \frac{X - \mu_X}{\sqrt{\text{Var}(X)}}$ với $\mathbb{E}(Z_X) = 0$ và $\text{Var}(Z_X) = 1$;
- $Z_Y = \frac{Y - \mu_Y}{\sqrt{\text{Var}(Y)}}$ với $\mathbb{E}(Z_Y) = 0$ và $\text{Var}(Z_Y) = 1$.

Từ đây, ta suy ra

$$r = \mathbb{E}\left(\frac{(X - \mu_X)(Y - \mu_Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}\right) = \mathbb{E}(Z_X Z_Y).$$

Nếu X và Y là không tương quan, tức là chúng độc lập, thì ta có

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$

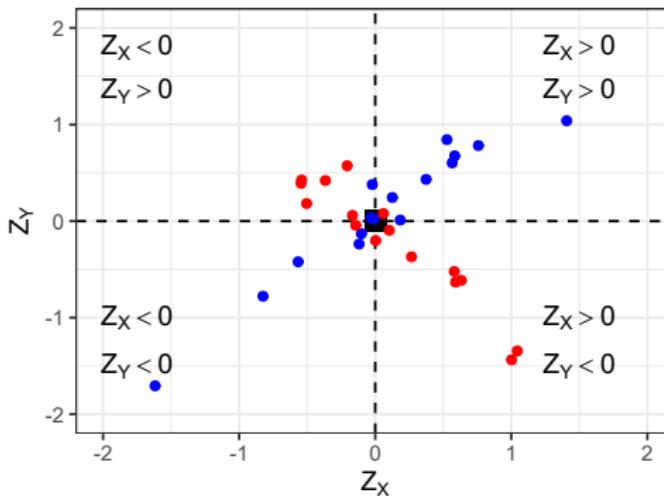
do đó,

$$\mathbb{E}(Z_X Z_Y) = \mathbb{E}(Z_X)\mathbb{E}(Z_Y) = 0$$

hay $r = 0$. Nhưng, điều ngược lại thì không đúng!

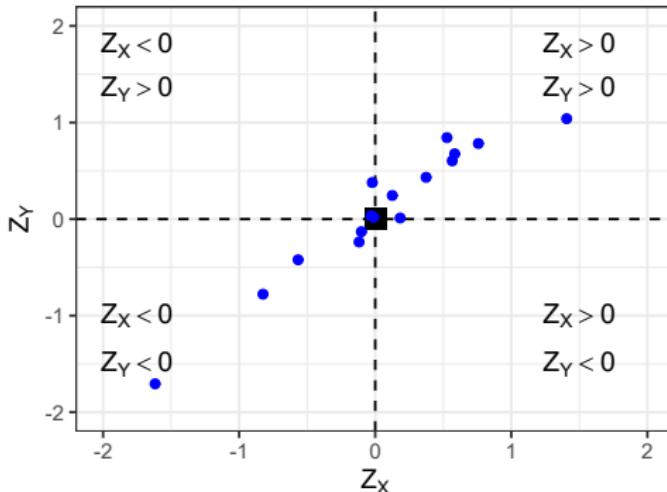
Cách hệ số tương quan tuyến tính hoạt động

Chú ý, phép đổi biến Z_X và Z_Y là tuyến tính, nên không thay đổi tính tương quan của X và Y .



Nếu X và Y có tương quan tuyến tính, các điểm dữ liệu z_x và z_y sẽ có xu hướng nằm thành 1 đường chéo đi qua gốc tọa độ (trung bình Z_X và Z_Y).

Cách hệ số tương quan tuyến tính hoạt động

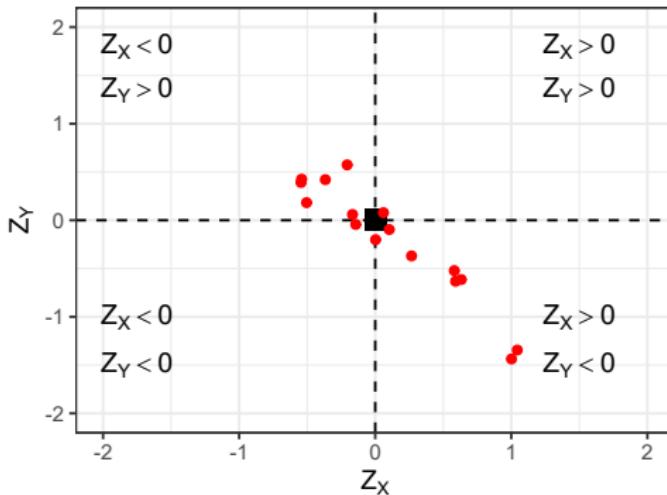


Nếu đa phần các điểm nằm vào góc phần tư thứ I và thứ III, thì

$$Z_X Z_Y > 0,$$

với hầu hết dữ liệu quan sát, dẫn tới kết quả là $r > 0$. (Về mặt lý thuyết, ta có thể chứng minh bằng bổ đề Fatou - Fatou's Lemma)

Cách hệ số tương quan tuyến tính hoạt động



Nếu đa phần các điểm nằm vào góc phần tư thứ II và thứ IV, thì

$$Z_X Z_Y < 0,$$

với hầu hết dữ liệu quan sát, dẫn tới kết quả là $r < 0$.

Một số kết quả cho hệ số tương quan tuyến tính

Định lý sau cung cấp các kết quả quan trọng của hệ số tương quan tuyến tính.
(Xem thêm chứng minh trong Severini, 2005, Chương 4, trang 97).

Định lý

Gọi X và Y là hai biến ngẫu nhiên, thỏa $\mathbb{E}(X^2) < \infty$ và $\mathbb{E}(Y^2) < \infty$. Giả sử rằng $\text{Var}(X) > 0$ và $\text{Var}(Y) > 0$. Khi đó,

- (i) r không thay đổi dưới phép biến đổi tỷ lệ của X và Y ;
- (ii) $r^2 \leq 1$;
- (iii) $r^2 = 1$ khi và chỉ khi tồn tại các hằng số thực a, b sao cho

$$\Pr(Y = aX + b) = 1.$$

- (iv) $r^2 = 0$ khi và chỉ khi, với bất kỳ hằng số thực a, b sao cho

$$\mathbb{E}\{[Y - (aX + b)]^2\} \geq \text{Var}(Y).$$

Từ (ii), ta có kết quả rằng, nếu $r^2 = 1$ thì chắc chắn rằng Y là một hàm tuyến tính của X .

Từ (iv), $r^2 = 0$ nếu và chỉ nếu hàm tuyến tính của X là tiên đoán tốt nhất cho Y theo nghĩa $\mathbb{E}\{[Y - (aX + b)]^2\}$ là hàm số với $a = 0$ và $b = \mathbb{E}(Y)$.

Một số kết quả cho hệ số tương quan tuyến tính

Tức là X sẽ không hữu ích trong việc tiên đoán Y , ít nhất khi ta cố áp đặt một hàm tuyến tính của X .

Ví dụ: Đặt X là một biến ngẫu nhiên theo phân phối Laplace với hàm mật độ

$$f(x) = \frac{1}{2} \exp(-|x|),$$

với $-\infty < x < \infty$. Khi đó, ta tính được $\mathbb{E}(X) = 0$, $\mathbb{E}(X^2) = 2$ và $\mathbb{E}(X^3) = 0$.

Đặt $Y = X^2$, khi đó, $\mathbb{E}(Y) = \mathbb{E}(X^2) = 2$ và

$$\text{Cov}(X, Y) = \mathbb{E}[X(Y - 2)] = \mathbb{E}[X^3 - 2X] = 0.$$

Suy ra, $r = 0$. Do đó, tất cả các hàm tuyến tính của X là không hữu ích cho việc tiên đoán Y (không có tương quan tuyến tính).

Tuy nhiên, thực tế thì X^2 tạo ra Y .

Ví dụ này cũng minh họa cho việc $r = 0$ chỉ suy ra **không có tương quan tuyến tính**, nhưng **không chắc chắn là không có tương quan** giữa X và Y , trong trường hợp tổng quát.

Phân phối chuẩn nhiều chiều

Xét vector ngẫu nhiên liên tục $\mathbf{X} = (X_1, X_2, \dots, X_p)$, có vector kỳ vọng $\mu_{\mathbf{X}} = (\mu_1, \mu_2, \dots, \mu_p)$ và ma trận hiệp phương sai $\Sigma_{\mathbf{X}}$ xác định bởi:

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p} & \sigma_{2p} & \dots & \sigma_p^2 \end{pmatrix}.$$

Nếu joint pdf của (X, Y) có dạng:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2} |\Sigma_{\mathbf{X}}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu_{\mathbf{X}})^\top \Sigma_{\mathbf{X}}^{-1} (\mathbf{x} - \mu_{\mathbf{X}}) \right\},$$

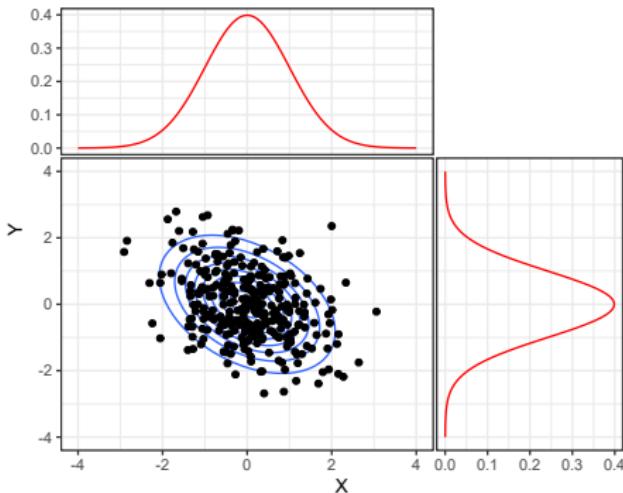
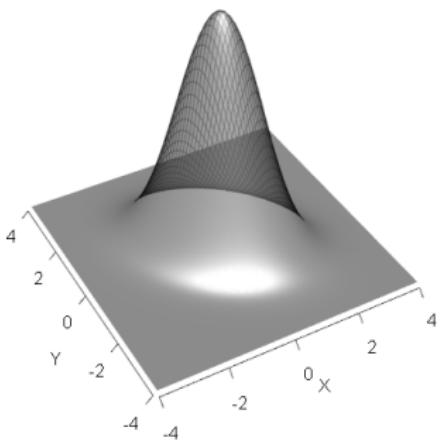
trong đó, $|\Sigma_{\mathbf{X}}|$ là định thức của $\Sigma_{\mathbf{X}}$, khi đó, ta nói \mathbf{X} tuân theo phân phối chuẩn p chiều, $\mathbf{X} \sim \mathcal{N}_p(\mu_{\mathbf{X}}, \Sigma_{\mathbf{X}})$.

Phân phối lề của X_j là phân phối chuẩn $\mathcal{N}(\mu_j, \sigma_j^2)$, với $j = 1, \dots, p$.

Phân phối chuẩn nhiều chiều

Ví dụ: phân phối chuẩn 2 chiều:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & -0.4 \\ -0.4 & 1 \end{pmatrix} \right)$$



Phân phối chuẩn nhiều chiều

Phân phối chuẩn hai chiều

Ta xét trường hợp đặc biệt, khi (X, Y) tuân theo phân phối chuẩn hai chiều:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & r\sigma_X\sigma_Y \\ r\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right)$$

Ta tính được phân phối điều kiện của Y khi biết X là một phân phối chuẩn với trung bình điều kiện

$$\mu_{Y|X} = \mu_Y + r \frac{\sigma_Y}{\sigma_X} (X - \mu_X)$$

và phương sai điều kiện $\sigma_{Y|X}^2 = (1 - r^2)\sigma_Y^2$.

Từ kết quả trung bình điều kiện $\mu_{Y|X}$, ta thấy rằng nếu $r = 0$, thì $\mu_{Y|X} = \mu_Y$, tức là không phụ thuộc vào X , và khi đó, không có tương quan tuyến tính giữa X và Y .

Phân phối chuẩn nhiều chiều

Hơn nữa, khi $r = 0$, ta cũng có $\sigma_{Y|X}^2 = \sigma_Y^2$.

↪ Phân phối điều kiện của Y khi biết X chính là phân phối lè của Y .

Như vậy, ta có

$$F_{X,Y}(x,y) = F_X(x)F_{Y|X}(y|x) = F_X(x)F_Y(y)$$

tức là, X và Y là độc lập, hay không có tương quan giữa X và Y .

Trường hợp này cho ta hai kết luận:

- Nếu (X, Y) là vectơ ngẫu nhiên tuân theo phân phối chuẩn 2 chiều, thì tương quan giữa X và Y nếu tồn tại thì chỉ là tương quan tuyến tính.
- Nếu $r = 0$ thì nhận xét “không tồn tại tương quan giữa X và Y ” chỉ đúng trong trường hợp (X, Y) là vectơ ngẫu nhiên tuân theo phân phối chuẩn 2 chiều.

Kết luận thứ hai là bản lề cho kiểm định giả thuyết $r = 0$ mà ta sẽ xét trong phần sau.

Phân phối đa thức - Multinomial distribution

Xét vectơ k chiều $Y = (Y_1, Y_2, \dots, Y_k)^\top$ sao cho:

- $Y_j \sim B(n, p_j)$;
- $\sum p_j = 1$;
- $\sum Y_j = n$.

Khi đó, ta nói $Y \sim \text{Multi}(n, (p_1, p_2, \dots, p_k))$, hàm xác suất

$$f(y_1, y_2, \dots, y_k; n, p_1, p_2, \dots, p_k) = \frac{n!}{y_1! y_2! \dots y_k!} p_1^{y_1} p_2^{y_2} \dots p_k^{y_k}.$$

Tính chất:

- vectơ trung bình $\mathbb{E}(Y) = (np_1, np_2, \dots, np_k)^\top$;
- $\text{Var}(Y_j) = np_j(1 - p_j)$;
- $\text{Cov}(Y_j, Y_r) = -np_j p_r$;
- ma trận hiệp phương sai:

$$\Omega = \begin{pmatrix} np_1(1 - p_1) & -np_1 p_2 & \dots & -np_1 p_k \\ -np_1 p_2 & np_2(1 - p_2) & \dots & -np_2 p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_1 p_k & -np_2 p_k & \dots & np_k(1 - p_k) \end{pmatrix}$$

Phân phối đa thức - Multinomial distribution

Ví dụ: Xét kết quả xuất hiện khi gieo một con xúc xắc 6 mặt, 30 lần. Trong mỗi lần tung:

- số chấm xuất hiện ở mặt trên là kết quả được ghi nhận: 1, 2, 3, 4, 5, 6;
- chỉ 1 trong 6 kết quả (tính chất) có thể xuất hiện;
- xác suất 1 kết quả xuất hiện là $1/6$.

Kết quả trong 30 lần có thể là (D_j là kết quả chấm j xuất hiện)

Lần tung	Số chấm	D_1	D_2	D_3	D_4	D_5	D_6
1	1	1	0	0	0	0	0
2	3	0	0	1	0	0	0
:	:	:	:	:	:	:	:
29	4	0	0	0	1	0	0
30	6	0	0	0	0	0	1

Đặt $Y_j = \sum_{i=1}^{30} D_{ij}$ là tổng số lần mặt j xuất hiện trong 30 lần, khi đó

$$(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)^\top \sim \text{Multi}(30, (1/6, 1/6, 1/6, 1/6, 1/6, 1/6))$$

Biến định tính và chuyển đổi số

Biến định tính X là một biến miêu tả các đặc tính hoặc tính chất không thể cân-đong-đo-đếm của một đối tượng:

- giới tính,
- màu sắc,
- nhóm thể loại,
- trạng thái bệnh,
- tình trạng hút thuốc.

Thông thường, dữ liệu quan sát của biến định tính được ghi nhận dưới dạng chữ, thay vì số

	Pain	Treatment	Age	Gender	Duration
1	eased	Treated	76	M	36
2	notEased	Treated	52	M	22
3	notEased	NotTreated	80	F	33
4	notEased	Treated	77	M	33
5	notEased	Treated	73	F	17
6	notEased	NotTreated	82	F	84

Biến định tính và chuyển đổi số

Về mặt kỹ thuật, ta có thể đưa biến X về dạng biến số rời rạc, bằng cách sử dụng giả biến (dummy variable).

Xét X là biến định tính có hai tính chất. Đặt

- tính chất 1 = 0
- tính chất 2 = 1 (được quan tâm)

khi này, X sẽ là một biến ngẫu nhiên Bernoulli với 2 kết quả:

- thành công $\Leftrightarrow 1$, với xác suất p ;
- thất bại $\Leftrightarrow 0$, với xác suất $1 - p$.

Hơn nữa, nếu ta có một mẫu ngẫu nhiên X_1, X_2, \dots, X_n , độc lập cùng phân phối Bernoulli, $B(p)$, thì

- $\mathbb{E}(X_i) = p$,
- $\text{Var}(X_i) = p(1 - p)$.

Đặt $Y = \sum X_i$, thì

- Y là tổng số lượng đối tượng có tính chất 2, và
- $Y \sim B(n, p)$ - phân phối nhị thức.

Biến định tính và chuyển đổi số

Tổng quát, khi biến định tính X có k tính chất ($k > 2$), ta sẽ đặt một tính chất tương ứng là một biến Bernoulli:

X	D_1	D_2	\dots	D_{k-1}	D_k
tính chất 1	1	0	\dots	0	0
tính chất 2	0	1	\dots	0	0
\vdots	\vdots	\vdots	\dots	\vdots	\vdots
tính chất $k - 1$	0	0	\dots	1	0
tính chất k	0	0	\dots	0	1

$\Rightarrow X$ tương đương một vectơ ngẫu nhiên Bernoulli k chiều, $(D_1, D_2, \dots, D_k)^\top$:

- vectơ xác suất thành công (p_1, p_2, \dots, p_k) ,
- $p_1 + p_2 + \dots + p_k = 1$;
- vectơ $(D_1, D_2, \dots, D_k)^\top \sim \text{Multi}(1, (p_1, p_2, \dots, p_k))$, phân phối đa thức.

Biến định tính và chuyển đổi số

Giả sử rằng, ta có một mẫu ngẫu nhiên của X với n quan sát, X_1, X_2, \dots, X_n . Khi đó, ta có ma trận kết quả cỡ $n \times k$ tương ứng biểu diễn vectơ ngẫu nhiên Bernoulli k chiều:

$$\begin{pmatrix} D_{11} & D_{12} & \dots & D_{1k} \\ D_{21} & D_{22} & \dots & D_{2k} \\ \vdots & \vdots & \dots & \vdots \\ D_{n1} & D_{n2} & \dots & D_{nk} \end{pmatrix}$$

trong đó,

$$D_{ij} = \begin{cases} 1, & \text{tính chất } j, \\ 0, & \text{tính chất khác,} \end{cases}$$

với $i = 1, \dots, n$ và $j = 1, \dots, k$. Đặt $Y_j = \sum_{i=1}^n D_{ij}$, khi đó,

- Y_j là tổng số lượng đối tượng có tính chất j , và
- $Y_j \sim B(n, p_j)$ - phân phối nhị thức,
- $\sum Y_j = n$,
- vector rời rạc $(Y_1, Y_2, \dots, Y_k)^\top \sim \text{Multi}(n, (p_1, p_2, \dots, p_k))$, phân phối đa thức.

Main references I

Agresti, A. and Kateri, M. (2021). *Foundations of statistics for data scientists: with R and Python*. Chapman and Hall/CRC.

Gut, A. (2009). *An Intermediate Course in Probability*. Springer.

Severini, T. A. (2005). *Elements of distribution theory*, volume 17. Cambridge University Press.