# Statistical Inference: Xác suất (Probability)

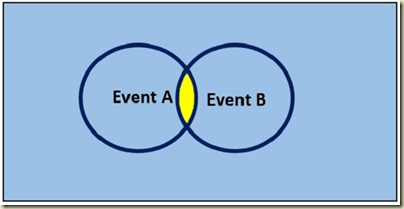
## Các khái niệm cơ bản

Xác suất xảy ra một biến cố cụ thể là tỉ số giữa kích thước của biến cố (event, outcome) với không gian mẫu (sample space).

Xác suất của biến cố E được ký hiệu là P(E), số này nằm trong khoảng 0 và 1. Những biến cố bất khả thi (Impossible events) có xác suất là 0 vì không thể nào xảy ra. Những biến cố chắc chắn (certain) xảy ra có xác suất là 1.

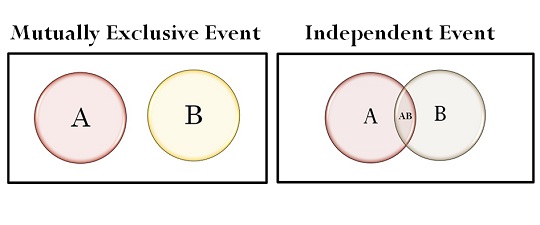
Nếu ta thu được n các biến cố có thể xảy ra gồm e1, e2…en thì tổng xác suất của các biến cố này bằng 1. Nếu các biến cố xảy ra như nhau (likely), như ví dụ tung xúc sắc trên, thì xác suất của mỗi biến cố sẽ bằng 1/n.

**Biến cố giao (Intersection):** Giao của hai biến cố A và B là biến cố chứa các thành phần vừa thuộc A vừa thuộc B. Ký hiệu: P(A U B), P(AB). Công thức tính: P(A U B) = P(A) + P(B).



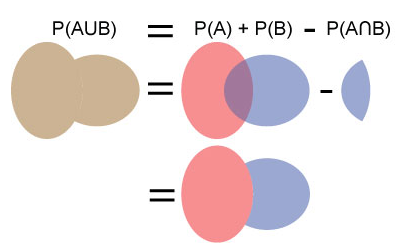
Nếu A và B là hai biến cố độc lập **(Independent events)** nhau thì xác suất xảy ra đồng thời hai biến cố trên sẽ bằng tích xác suất xảy ra các biến cố đó. P(A&B) = P(A) \* P(B).

Nếu xác suất xảy ra một biến cố E có nhiều hơn một cách và các cách này tách rời nhau **(disjoint, mutually exclusive)** thì xác xuất P(E) bằng tổng của từng xác suất xảy ra các biến cố này. P(A or B) = P(A) + P(B) và P(A & B) =0



**Biến cố hợp (union):** Hợp của hai biến cố A và B là biến cố chứa tất cả các thành phần của A và B. Ký hiệu: P(A U B), P(A + B). Công thức tính: P(A U B) = P(A) + P(B) – P(A&B).

Điều này cũng dễ hiểu, khi ta cộng hai xác suất P(A) và P(B), ta đã lỡ đếm xác suất xảy ra biến cố A và B hai lần. Khi A và B là hai biến cố giao nhau (intersection), ta cần loại bớt xác suất P(A&B) này.



# Biến ngẫu nhiên

Biến ngẫu nhiên(Random variables) là ánh xạ từ một tập hợp, xây dựng trên nền không gian mẫu S, vào tập các xác suất có thể xảy ra. Biến ngẫu nhiên rời rạc: nếu nó chỉ có hữu hạn, hoặc vô hạn đếm được các giá trị (ví dụ: X1 = Tổng điểm thi đại học khối A). Biến ngẫu nhiên liên tục (ví dụ: X2 = Chiều cao của 1 người Việt Nam).

Biến ngẫu nhiên liên tục thường được kết hợp với các phép đo thời gian, khoảng cách, hoặc một số quy trình sinh học vì ta có thể xét bất kỳ giá trị nào. Những hạn chế về độ chính xác trong việc đo lường có thể ngụ ý rằng các giá trị là rời rạc.

## Probability mass function

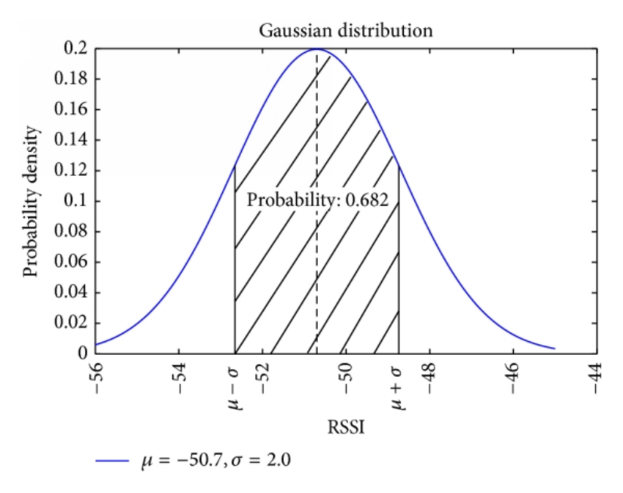
**Hàm độ lớn xác suất (pmf – probability mass function)** của biến ngẫu nhiên rời rạc là hàm gán xác suất cho từng giá trị của X, ký hiệu P(x). Giả sử ta có một đồng xu với x=0 thể hiện biến cố xảy ra mặt phải (head), x=1 thể hiện biến cố xảy ra mặt trái (tail). Nếu p là xác suất biến cố xảy ra mặt phải ta có biểu thức của PMF như sau:

## Probability density function

**Hàm mật độ xác suất (pdf – probability density function)** được đặc trưng cho biến ngẫu nhiên liên tục. Để tìm xác suất của một biến ngẫu nhiên liên tục, thường ta tính diện tích phần dưới đường cong nằm giữa 2 điểm cần tính xác suất.

μ: giá trị trung bình (mean)

σ: độ lệch chuẩn (standard deviation)



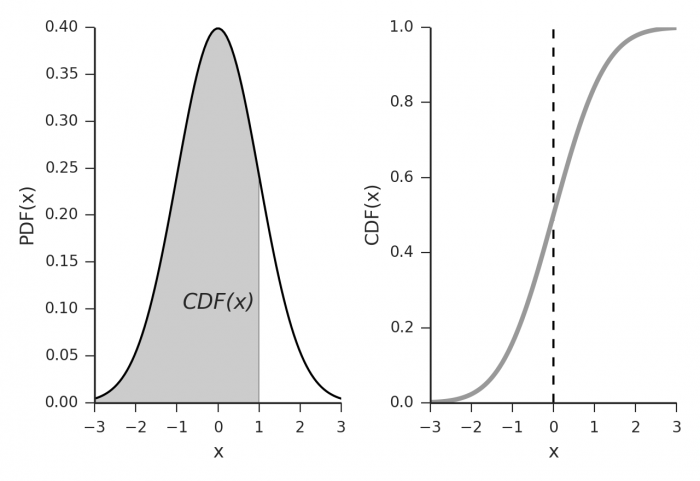
Phân phối được gọi là liên tục nếu biến ngẫu nhiên nhận giá trị trong một miền vô hạn không đếm được. Hàm phân bố tích lũy tạo thành một đường cong liên tục. Nếu X là một biến ngẫu nhiên liên tục, ta không thể sử dụng hàm độ lớn xác suất (pmf) cho X. Ta chỉ có thể tính xác suất cho một khoảng giá trị của X.

## Cumulative distribution function

Hàm phân phối tích lũy (cdf – cumulative distribution function) của biến ngẫu nhiên X được đặc trưng cho biến ngẫu nhiên rời rạc và liên tục. Là hàm F(x) được tính bằng tổng các xác suất của biến ngẫu nhiên X nhỏ hơn hay bằng giá trị a.

Nếu bạn đã từng làm việc với giải tích (calculus), ắt hẳn bạn đã nhận ra khi chúng ta tính diện tích dưới đường cong thật chất chúng ta đi tính tích phân của hàm đó.

Khi biến ngẫu nhiên là liên tục, như ví dụ trên, PDF là đạo hàm của CDF. Do đó, khi ta tính nguyên phân của PDF ta sẽ được CDF. Khi ta tính CDF trong một khoảng [a, b] ta sẽ được kết quả là diện tích phía dưới đường cong.



## Phân vị (quantile)

Hàm sống sót S(x) (survivor function) của một biến ngẫu nhiên X được định nghĩa là một hàm của giá trị x để tính xác suất của biến ngẫu nhiên X có giá trị lớn hơn x.

Như vậy, trong ví dụ của chúng ta, biểu thức thể hiện hàm sống sót là : 1 – (x\*x/4). Vì tổng diện tích của PDF bằng 1 mà hàm sống sót là phần bù của CDF, nên ta chỉ cần lấy phần bù của hàm CDF để tính hàm S(x).

Phân vị (quantile) v của CDF là điểm x\_v mà tại đó CDF có giá trị là v. Cụ thể, F(x\_v) = v. Phần trăm phân vị (percentile) là một phân vị trong đó v được biểu diễn dưới dạng phần trăm (percentage).

Theo đó, phần trăm phân vị của trung vị (median) là 50-th. Nếu ta có F(x)= x^2/4 như ví dụ trên thì điểm phân vị x\_v = \sqrt {x^2} = \sqrt {4\*.5} = \sqrt 2 = 1.414214.

Chúng ta sẽ kết thúc bài viết này với một vài điểm đáng lưu ý. Một mô hình xác suất kết nối dữ liệu đến một quần thể (population) bằng cách sử dụng những giả định. Cần lưu ý sự khác biệt giữa trung bình quần thể (population medians) và trung bình mẫu (sample medians). Trung bình mẫu là một ước lượng (estimator) của trung bình quần thể. Xin chúc mừng, chúng ta đã tìm hiểu được các khái niệm của xác suất.