# Statistical Inference: Xác suất (Probability)

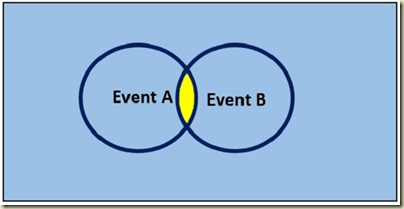
## Các khái niệm cơ bản

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.

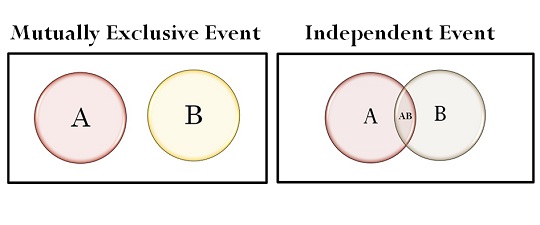
Nếu ta thu được n các biến cố có thể xảy ra gồm e1, e2…en thì tổng xác suất của các biến cố này bằng 1. Nếu các biến cố xảy ra như nhau (likely), như ví dụ tung xúc sắc trên, thì xác suất của mỗi biến cố sẽ bằng 1/n.

**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).



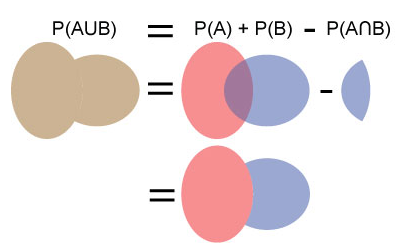
Nếu A và B là hai biến cố độc lập **(Independent events)** nhau thì xác suất xảy ra đồng thời hai biến cố trên sẽ bằng tích xác suất xảy ra các biến cố đó. P(A&B) = P(A) \* P(B).

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

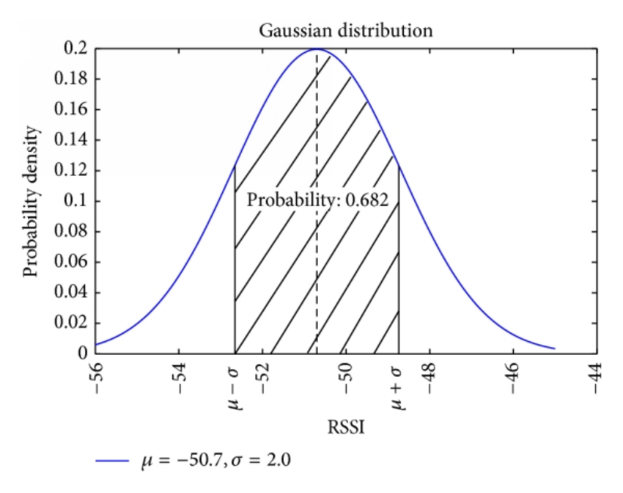
**Hàm độ lớn xác suất (pmf – probability mass function)** của biến ngẫu nhiên rời rạc là hàm gán xác suất cho từng giá trị của X, ký hiệu P(x). Giả sử ta có một đồng xu với x=0 thể hiện biến cố xảy ra mặt phải (head), x=1 thể hiện biến cố xảy ra mặt trái (tail). Nếu p là xác suất biến cố xảy ra mặt phải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



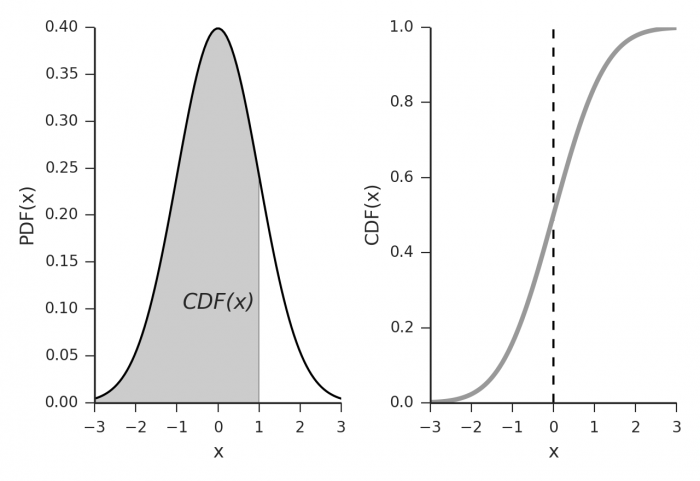
Phân phối được gọi là liên tục nếu biến ngẫu nhiên nhận giá trị trong một miền vô hạn không đếm được. Hàm phân bố tích lũy tạo thành một đường cong liên tục. Nếu X là một biến ngẫu nhiên liên tục, ta không thể sử dụng hàm độ lớn xác suất (pmf) cho X. Ta chỉ có thể tính xác suất cho một khoảng giá trị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



Tóm lại:

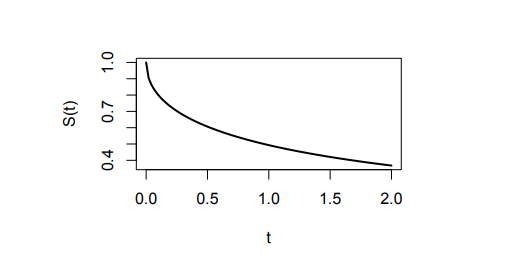
* CDF là tích phân của PDF
* PDF là đạo hàm của CDF

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.

Trong đó:

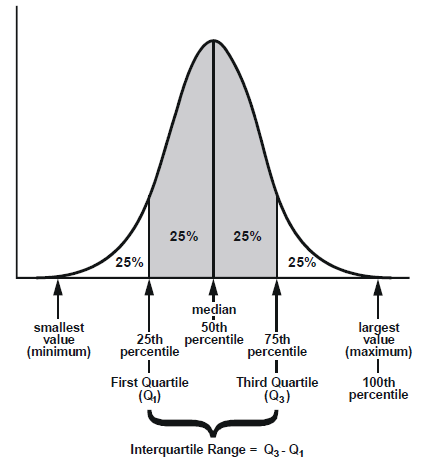
T là response variable. T ≥ 0



Khi t dao động từ 0 đến ∞, hàm sống sót có thuộc tính sau

* Nó không tăng
* Tại thời điểm t = 0, S (t) = 1. Nói cách khác, xác suất tồn tại thời gian 0 là 1
* Tại thời điểm t = ∞, S (t) = S (∞) = 0. Khi thời gian trôi qua vô hạn, đường cong sống còn là 0.

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.

# Statistical Inference: Tiệm cận (Asymptotics)

## Luật số lớn

Luật số lớn (Law of Large Numbers (LLN)) phát biểu rằng giá trị trung bình (mean) có xu hướng tiệm cận về giá trị mà nó đang ước lượng. Nghĩa là kích thước mẫu dữ liệu càng lớn thì ước lượng quần thể càng chính xác. Như chúng ta tung một con xúc sắc nhiều lần liên tiếp, nó có xu hướng hội tụ về giá trị xác suất 0.5. LLN tạo cơ sở cho kiểu suy luận dựa trên tần xuất. Lấy ví dụ minh hoạ với hàm coinPlot(). Hàm này nhận vào đối số n là số lượt tung xúc sắc. Sau mỗi lần tung, hàm này sẽ cộng dồn các giá trị 1 (head) và 0 (tail) sau đó tính trung bình cộng lại. Sau đó, hàm này sẽ xuất ra biểu đồ biến thiên của giá trị trung bình vừa mới tính được này.

coinPlot <- function(n){

means <- cumsum(sample(0 : 1, n , replace = TRUE)) / (1 : n)

g <- ggplot(data.frame(x = 1 : n, y = means), aes(x = x, y = y))

g <- g + geom\_hline(size=1.5 ,yintercept = 0.5,alpha=0.6,

linetype="longdash") + geom\_line(size = 1)

if(n<100){

g <- g + geom\_point(colour="red",size=3,alpha=0.8)

}

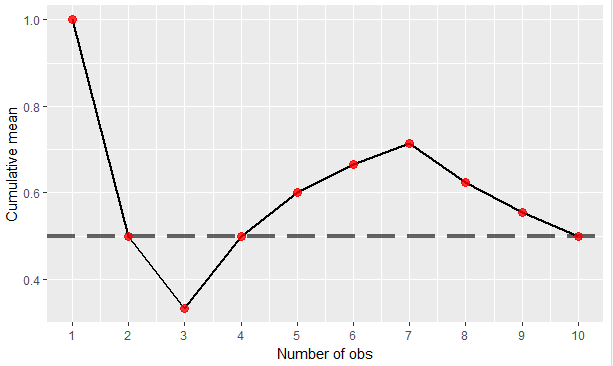
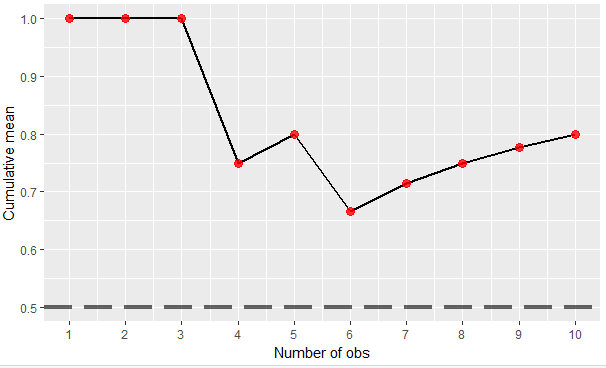
g <- g + labs(x = "Number of obs", y = "Cumulative mean")

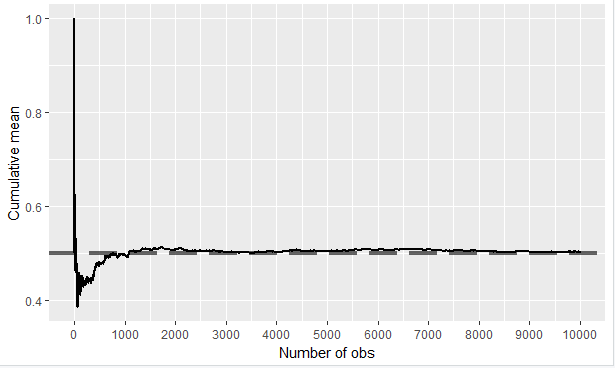
g <- g + scale\_x\_continuous(breaks=seq(0,n+1,ceiling(n/10)))

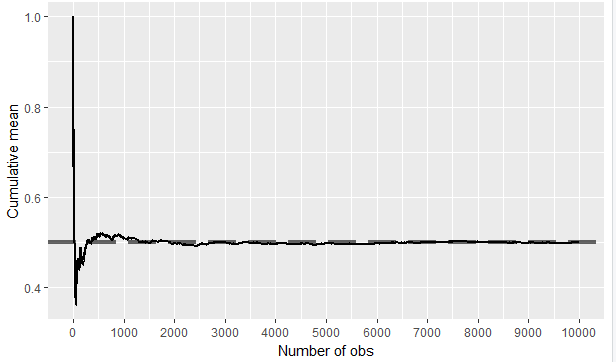
print(g)

invisible()

}

Ví dụ tung xúc sắc 10 lần

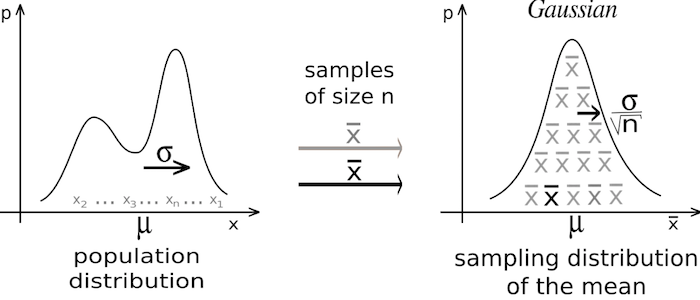
Biểu đồ xuất ra phụ thuộc vào giá trị ngẫu nhiên mà R tạo được. Vì vậy, biểu đồ trên có vẻ khác một chút khi chạy trên máy tính của bạn. Nếu bạn thử chạy hàm trên coinPlot(10) vài lần bạn sẽ thấy mỗi lần như vậy sẽ cho ra kết quả khác một chút. Bây giờ, ta thử tung xúc sắc 10,000 lần



Bạn đã thấy được sự khác biệt phải không nào. Đây là minh chứng cho lý thuyết tiệm cận (asymptotics). Đường biến thiên hội tụ tại giá trị 0.5. Ta bảo rằng một ước lượng thống kê như vậy thì đồng nhất nếu nó hội tụ tại giá trị mà nó đang ước lượng. Định lý số lớn bảo rằng trung bình mẫu của biến ngẫu nhiên iid thì đồng nhất với trung bình của quần thể. Tương tự như vậy, phương sai mẫu (sample variance) cũng có xu hướng hội tụ về phương sai của quần thể (population variance).

## Định lý giới hạn trung tâm

Định lý giới hạn trung tâm (Central Limit Theorem (CLT)) là một trong những định lý quan trọng trong thống kê. Định lý phát biểu rằng phân phối của các giá trị trung bình của các biến ngẫu nhiên iid hội tụ về các giá trị chuẩn khi kích thước mẫu dữ liệu tăng lên.



Để minh hoạ CLT trực quan hơn ta quan sát hình sau.

nosim <- 1000

cfunc <- function(x, n) sqrt(n) \* (mean(x) - 3.5) / 1.71

dat <- data.frame(

  x = c(apply(matrix(sample(1 : 6, nosim \* 10, replace = TRUE),

                     nosim), 1, cfunc, 10),

        apply(matrix(sample(1 : 6, nosim \* 20, replace = TRUE),

                     nosim), 1, cfunc, 20),

        apply(matrix(sample(1 : 6, nosim \* 30, replace = TRUE),

                     nosim), 1, cfunc, 30)

  ),

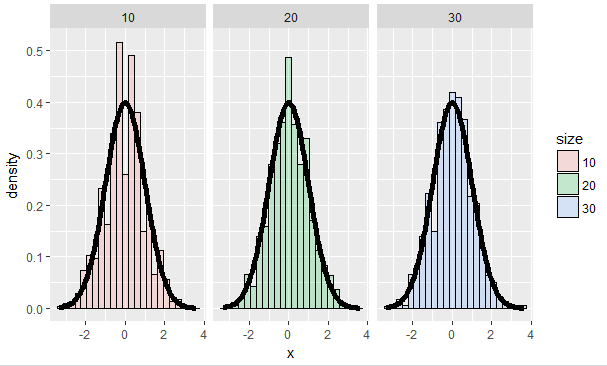
  size = factor(rep(c(10, 20, 30), rep(nosim, 3))))

g <- ggplot(dat, aes(x = x, fill = size)) + geom\_histogram(alpha = .20, binwidth=.3, colour = "black", aes(y = ..density..))

g <- g + stat\_function(fun = dnorm, size = 2)

g <- g + facet\_grid(. ~ size)

print(g)

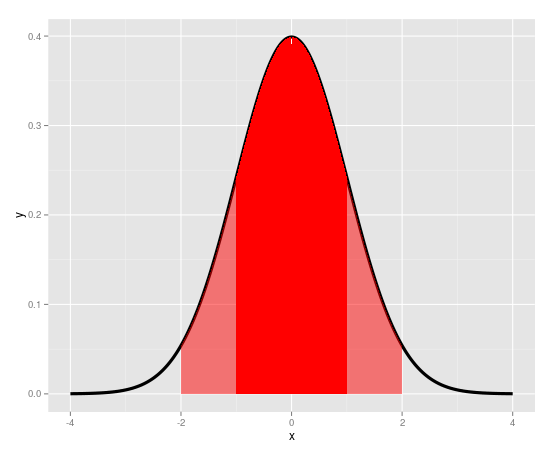


Giải thích: Hình trên biểu diễn 3 histogram của 1000 giá trị trung bình khi thực hiện 1000 lần lấy mẫu trên kích thước mẫu n (10, 20, 30). Mỗi giá trị trung bình của n mẫu (n=10, 20, 30) được chuẩn hoá bằng cách trừ cho mean (3.5) sau đó chia cho độ lỗi chuẩn (standard error), sqrt(2.92/n). Việc chuẩn hoá này khiến cho các histogram có dạng như phân phối chuẩn (với mean bằng 0 và độ lệch chuẩn bằng 1).

Chú ý rằng CLT không nói gì về phân bố ban đầu của quần thể là phân phối chuẩn. Ta có thể giả định tính phân phối chuẩn của trung bình quần thể bất chấp phân bố của quần thể ta đang có. Chỉ cần kích thước mẫu dữ liệu đủ lớn và các mẫu dữ liệu độc lập với nhau.

## Khoảng tin cậy

Bây giờ chúng ta sẽ nói về khoảng tin cậy (confidence intervals). Từ CLT với n đủ lớn, trung bình mẫu là mean µ và độ lệch chuẩn . Ta biết rằng 95% diện tích phía dưới đường cong nằm trong khoảng độ lệch chuẩn này.



*Standard normal (là các giá trị của hàm phân bố tích lũy của phân phối chuẩn)*

Biểu đồ trên là phân phối chuẩn với µ = 0 và = 1, để minh họa điều trên, toàn bộ bóng đỏ mô tả diện tích nằm trong độ lệch chuẩn 2, bóng đậm hơn mô tả 68% diện tích nằm trong độ lệch chuẩn 1. Theo đó, 5% diện tích còn lại không được tô bóng đỏ. Nghĩa là xác suất trung bình mẫu là 5%.

Tương tự, xác suất nằm trong giới hạn này là 95%. Tất nhiên ta có thể có nhiều khoảng tin cậy khác nhau. Nếu ta muốn giá trị khác 95%, ta có thể sử dụng phân vị (quantile) thay cho 2. Nếu ta muốn tìm khoảng tin cậy 90% thì phân vị chúng ta được tính như sau (100-90)/2 + 90 = 95. Ta dùng hàm qnorm của R để tìm phân vị 95 cho phân phối chuẩn. Hàm này lấy đối số là xác suất. Các đối số còn lại ta có thể lấy mặc định.

*1 qnorm(.95)*

*2 [1] 1.644854*

|  |  |  |
| --- | --- | --- |
| **-2-2** | **-2** | **0** |
| **-2** | **2-2** | **0** |
| **0** | **0** | **-1-2** |