Contents

Combination – Tổ Hợp:

1. Hoán Vị (Permutation)?

$$P(n,k) = A_n^k = \frac{n!}{(n-k)!}$$

2. Hoán Vị Với Phần Tử Lặp Nhưng Không Được Liền Nhau?

⇨ Cho tập S gồm n kí tự khác nhau, ta muốn chọn từ S k kí tự để làm 1 chuỗi, các kí tự trong chuỗi có thể giống nhau, nhưng không được liền nhau, ví dụ "aabc" thì không được, "abac" thì được

⇨ Số các chuỗi thỏa mãn là

$$n(n-1)^{k-1}$$

⇨ Nghĩa là ta có n cách chọn kí tự đầu tiên, ở kí tự thứ 2, ta không thể chọn lại kí tự thứ nhất, do đó còn n – 1 cách chọn, ở kí tự thứ 3, ta không thể chọn lại kí tự thứ 2, do đó còn n – 1 cách chọn, …

⇨ Ví dụ, cho tập S = {a, b, c, d, e}, số chuỗi 3 kí tự được tạo từ S sao cho không có 2 kí tự liền kề giống nhau là

$$5(5-1)^{3-1} = 80$$

3. Hoán Vị Với Phần Tử Lặp Nhưng 1 Phần Tử Không Được Xuất Hiện Liên Tiếp?

⇨ Cho tập S gồm các kí tự khác nhau, ta muốn chọn từ S k kí tự để làm 1 chuỗi, các kí tự trong chuỗi có thể giống nhau, tuy nhiên không cho phép n kí tự A được đứng liền kề nhau, ví dụ nếu n = 3 và A = "g", thì "gggd" không hợp lệ, "ggdg" hợp lệ

⇨ Ý tưởng là dùng công thức truy hồi

⇨ Ví dụ, cho tập S = {a, b, c, d}, ta muốn tạo ra các chuỗi gồm k = 10 kí tự, sao cho không có 3 kí tự "a" nào đứng liền nhau

⇨ Bước 1, ta có các chuỗi kí tự khởi đầu sau

"b", "c", "d", "ab", "ac", "ad", "aab", "aac", "aad"

⇨ Như vậy tổng cộng có 3 chuỗi có 1 kí tự, 3 chuỗi có 2 kí tự và 3 chuỗi có 3 kí tự

⇨ Từ các chuỗi khởi đầu, ta sẽ bù vào các kí tự đằng sau còn thiếu, số cách điền chính = lời giải cho bài toán trên nhưng với k = số kí tự bù

⇨ Bước 2, lập công thức truy hồi, R(n) là lời giải của bài toán trên với k = n

$$R(0) = 1, R(1) = 4, R(2) = 16$$
$$R(n) = 3R(n-1) + 3R(n-2) + 3R(n-3)$$

⇨ Bước 3, giải công thức truy hồi

⇨ Phương trình đặc trưng

$$x^3 - 3x^2 - 3x - 3 = 0 \Leftrightarrow \begin{cases} x_1 \approx 3.951 \\ x_2 \approx -0.476 + 0.73i \\ x_3 \approx -0.476 - 0.73i \end{cases}$$

⇨ Ta có nghiệm tổng quát

$$R(n) = C_1 x_1^n + C_2 x_2^n + C_3 x_3^n$$

⇨ Giải phương trình với điều kiện ban đầu để tìm hệ số tự do

$$\begin{cases} R(0) = 1 \\ R(1) = 4 \\ R(2) = 16 \end{cases} \Leftrightarrow \begin{cases} C_1 + C_2 + C_3 = 1 \\ x_1 C_1 + x_2 C_2 + x_3 C_3 = 4 \\ x_1^2 C_1 + x_2^2 C_2 + x_3^2 C_3 = 16 \end{cases} \Leftrightarrow \begin{cases} C_1 \approx 1.022 \\ C_2 \approx -0.011 + 0.032i \\ C_3 \approx -0.011 - 0.032i \end{cases}$$
$$\Rightarrow R(10) = C_1 x_1^{10} + C_2 x_2^{10} + C_3 x_3^{10} = 947808$$

4. Tổ Hợp (Combination)?

$$C(n,k) = C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$
$$C_n^k = \frac{A_n^a}{A_k^a} C_{n-a}^{k-a}, a \in N \cap [0,k]$$
$$C_n^k = \frac{n-k+1}{k} C_n^{k-1}$$
$$C_n^k = \frac{n}{n-k} C_{n-1}^k$$
$$C_n^k = \frac{A_n^k}{k!}$$

5. Tổ Hợp Với Phần Tử Lặp?

⇨ Cho 1 tập hợp gồm n phần tử, ta muốn chọn ra bộ k phần tử không quan tâm thứ tự, các phần tử có thể giống nhau, số cách chọn là

$$C_{n+k-1}^k$$

⇨ Ví dụ 1

⇨ Cho tập S = {0, 1} có 2 phần tử, chọn ra các bộ 3 phần tử không quan tâm thứ tự, các phần tử có thể giống nhau, dễ thấy có 4 bộ 3 là 000, 001, 011, 111

$$C_{2+3-1}^3 = 4$$

⇨ Ví dụ 2, tìm số nghiệm nguyên không âm của phương trình sau

$$x_1 + x_2 + x_3 + x_4 + x_5 = 17$$

⇨ Bài toán tương đương số cách đánh 5 mốc vào 18 ô Index 0, 1, 2, …, 17, sao cho mốc thứ 1 ở ô Index x_1, mốc thứ 2 ở ô Index $x_1 + x_2$, mốc thứ 3 ở ô Index $x_1 + x_2 + x_3$, mốc thứ 4 ở ô Index $x_1 + x_2 + x_3 + x_4$, mốc thứ 5 ở ô Index 17

⇨ Như vậy bản chất bài toán là chọn 4 phần tử không quan tâm thứ tự trong 18 phần tử, cho phép lặp, do đó số cách chọn hay số nghiệm không âm của phương trình là

$$C_{18+4-1}^4 = C_{17+5-1}^{17} = 5985$$

⇨ Tổng quát, với phương trình

$$x_1 + x_2 + x_3 + \cdots + x_n = a$$

⇨ Ta có số nghiệm nguyên không âm là

C_{a+n-1}^{a}

⇨ Ví dụ 3, tìm số nghiệm nguyên của phương trình sau, sao cho x_1 và $x_2 \geq 3$, x_3, x_4, và $x_5 \geq 1$

$x_1 + x_2 + x_3 + x_4 + x_5 = 17$

⇨ Đặt

$x_1 = y_1 + 3, x_2 = y_2 + 3, x_3 = y_3 + 1, x_4 = y_4 + 1, x_5 = y_5 + 1$

⇨ Bài toán trở thành tìm số nghiệm nguyên không âm của phương trình sau

$y_1 + 3 + y_2 + 3 + y_3 + 1 + y_4 + 1 + y_5 + 1 = 17 \Leftrightarrow y_1 + y_2 + y_3 + y_4 + y_5 = 8$

⇨ Áp dụng công thức, ta có số nghiệm là

$C_{8+5-1}^{8} = 495$

⇨ Ví dụ 4, tìm số nghiệm nguyên của phương trình sau, sao cho $1 \leq x_1 \leq 3$, $1 \leq x_2 \leq 4$, x_3, x_4, và $x_5 \geq 1$

$x_1 + x_2 + x_3 + x_4 + x_5 = 17$

⇨ Bước 1, tìm số nghiệm nguyên với điều kiện x_1, x_2, x_3, x_4, và $x_5 \geq 1$

⇨ Đặt

$x_1 = y_1 + 1, x_2 = y_2 + 1, x_3 = y_3 + 1, x_4 = y_4 + 1, x_5 = y_5 + 1$

⇨ Bài toán trở thành tìm số nghiệm nguyên không âm của phương trình sau

$y_1 + y_2 + y_3 + y_4 + y_5 = 12$

⇨ Áp dụng công thức, ta có số nghiệm là

$C_{12+5-1}^{12} = 1820$

⇨ Bước 2, tìm số nghiệm nguyên với điều kiện $x_1 > 3$, x_2, x_3, x_4, và $x_5 \geq 1$

⇨ Đặt

$x_1 = y_1 + 4, x_2 = y_2 + 1, x_3 = y_3 + 1, x_4 = y_4 + 1, x_5 = y_5 + 1$

⇨ Bài toán trở thành tìm số nghiệm nguyên không âm của phương trình sau

$y_1 + y_2 + y_3 + y_4 + y_5 = 9$

⇨ Áp dụng công thức, ta có số nghiệm là

$C_{9+5-1}^{9} = 715$

⇨ Bước 3, tìm số nghiệm nguyên với điều kiện $x_2 > 4$, x_1, x_3, x_4, và $x_5 \geq 1$

⇨ Đặt

$x_1 = y_1 + 1, x_2 = y_2 + 5, x_3 = y_3 + 1, x_4 = y_4 + 1, x_5 = y_5 + 1$

⇨ Bài toán trở thành tìm số nghiệm nguyên không âm của phương trình sau

$y_1 + y_2 + y_3 + y_4 + y_5 = 8$

⇨ Áp dụng công thức, ta có số nghiệm là

$C_{8+5-1}^{8} = 495$

⇨ Bước 4, tìm số nghiệm nguyên với điều kiện $x_1 > 3$, $x_2 > 4$, x_3, x_4, và $x_5 \geq 1$

⇨ Đặt

$x_1 = y_1 + 4, x_2 = y_2 + 5, x_3 = y_3 + 1, x_4 = y_4 + 1, x_5 = y_5 + 1$

⇨ Bài toán trở thành tìm số nghiệm nguyên không âm của phương trình sau

$y_1 + y_2 + y_3 + y_4 + y_5 = 5$

⇨ Áp dụng công thức, ta có số nghiệm là

$C_{5+5-1}^{5} = 126$

⇨ Vậy tóm lại số nghiệm của phương trình đầu với điều kiện đầu là

$1820 - 715 - 495 + 126 = 736$

6. Hoán Vị Của 1 Dãy Có Các Phần Tử Giống Nhau?

⇨ Cho 1 dãy gồm n phần tử, trong đó có n_1 phần tử có giá trị 1, n_2 phần tử có giá trị 2, …, n_k phần tử có giá trị k, khi này số hoán vị của dãy này là

$$\frac{n!}{n_1!n_2!...n_k!}$$

7. Transposition?

⇨ Là 1 phép hoán vị mà trong đó, chỉ có hoán đổi vị trí của 2 phần tử, các phần tử khác giữ nguyên

8. Hoán Vị Chẵn?

⇨ Là hoán vị mà chỉ có thể phân tách thành số chẵn lần các Transposition

⇨ Ví dụ

⇨ Phép hoán vị (1, 4, 2, 3) là hoán vị chẵn vì nó = tráo vị trí 2, 4 + rồi tráo vị trí 3, 4

9. Hoán Vị Lẻ?

⇨ Là hoán vị mà chỉ có thể phân tách thành số lẻ lần các Transposition

Central Tendency – Khuynh Hướng Tập Trung:

1. Trung Bình Điều Hòa (Harmonic Mean)?

⇨ Cho dãy X là x_1, x_2, …, x_n, khi đó Harmonic Mean của X là

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

2. Trung Vị (Median)?

⇨ Cho X là biến ngẫu nhiên, khi đó điểm mà tại đó CDF của X có giá trị = 0.5 gọi là Median của X

⇨ X có thể có nhiều Median

⇨ Cho X là 1 dãy số bất kì được sắp xếp từ nhỏ đến lớn, nếu số phần tử của X lẻ thì Median là số nằm chính giữa dãy, nếu số phần tử là chẵn thì lấy trung bình 2 số ở trung tâm

⇨ Nếu số liệu thống kê theo kiểu phân khoảng, ví dụ thay vì nói thẳng ra số người 25 tuổi là bao nhiêu thì người ta lại nói số người có tuổi từ 24 đến 26, thì khoảng đầu tiên làm tần số tích lũy ≥ 1 nửa tần số tổng thể F sẽ là khoảng chứa Median, cho khoảng này là [a, b], f_1 là tần số tích lũy của khoảng đằng trước, f_0 là tần số của khoảng này, khi đó Median được tính theo công thức sau

$$M_e(X) = a + (b - a)\frac{0.5F - f_1}{f_0}$$

⇨ Công thức này có nghĩa là Median sẽ nằm trong khoảng [a, b] sao cho khi nội suy tuyến tính thì Median sẽ là điểm làm tần số tích lũy = 0.5F

3. Tứ Phân Vị (Quartiles)?

⇨ Cho X là dãy số có n phần tử được sắp xếp từ nhỏ đến lớn, khi đó

⇨ Phân vị thứ 1 là phần tử có số thứ tự là (n + 1) / 4

⇨ Phân vị thứ 2 chính là Median

⇨ Phân vị thứ 3 là phần tử có số thứ tự là 3(n + 1) / 4

⇨ Nếu số thứ tự không phải số nguyên thì nội suy tuyến tính phần thập phân

⇨ Trường hợp số liệu thống kê theo kiểu phân khoảng thì nội suy tuyến tính như Median

4. Mốt (Mode)?

⇨ Cho X là biến ngẫu nhiên, nếu rời rạc, thì Mode của X là giá trị x_0 sao cho

$P(X = x_0)$ đạt giá trị lớn nhất, nếu liên tục, thì Mode của X là điểm mà tại đó PDF của X có giá trị lớn nhất

⇨ X có thể có nhiều Mode

⇨ Trường hợp số liệu thống kê theo kiểu phân khoảng, cho khoảng có Mode là [a, b], f_0 là tần số của khoảng này, f_1 là tần số của khoảng đằng trước, f_2 là tần số của khoảng đằng sau, khi đó, Mode sẽ được tính theo công thức sau

$$M_0(X) = a + (b - a)\frac{f_0 - f_1}{2f_0 - f_1 - f_2}$$

⇨ Công thức này có nghĩa là Mode sẽ nằm trong khoảng [a, b], nếu tần số của khoảng đằng trước gần với tần số của khoảng chứa Mode hơn tần số của khoảng đằng sau, thì Mode sẽ lệch về trước, và ngược lại

Parameter Estimation – Ước Lượng Tham Số::

1. Tổng Thể (Population)?

⇨ Là 1 tập hợp gồm tất cả các đối tượng cùng có 1 thuộc tính X nào đó

2. Mẫu (Sample)?

⇨ Là 1 tập con của Population, giá trị thuộc tính X của mỗi phần tử trong Sample được kí hiệu là $X_1, X_2, \ldots, X_n$, n là kích thước Sample

3. Mô Hình (Model)?

⇨ Là phương trình nêu lên mối quan hệ gần đúng giữa 1 nhóm thuộc tính nào đó với 1 thuộc tính khác cho tất cả phần tử trong Population

⇨ Ví dụ

⇨ Model A cho ta biết chiều cao gần đúng của 1 người từ cân nặng của người đó

4. Thống Kê (Statistic)?

⇨ Là 1 giá trị nào đó được tính toán từ số liệu Sample

5. Trung Bình Mẫu (Sample Mean)?

⇨ 1 thống kê được tính như sau

$$\bar{X} = \sum_{i=1}^{n} X_i$$

⇨ Kì vọng và Variance của Sample Mean khi lấy nhiều Sample

$$E[\bar{X}] = E[X]$$
$$Var[\bar{X}] = \frac{1}{n}Var[X]$$

6. Phương Sai Mẫu Có Hiệu Chỉnh (Sample Variance)?

⇨ Là 1 thống kê được tính như sau

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

⇨ Chứng minh

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 \Leftrightarrow \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

⇨ Kì vọng và Variance của Sample Variance khi lấy nhiều Sample

$$E[S_X^2] = Var[X]$$
$$Var[S_X^2] = pass$$

⇨ Chứng minh

$$E[S_X^2] = E\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] =$$

$$\frac{1}{n-1}(nE[X^2] - nE[\bar{X}^2]) = \frac{1}{n-1}\left(nE[X^2] - n\left(\frac{1}{n^2}(nE[X^2] + 2C_n^2 E[X]^2)\right)\right) =$$
$$\frac{1}{n-1}(nE[X^2] - E[X^2] - (n-1)E[X]^2) = E[X^2] - E[X]^2 = Var[X]$$

⇨ Nếu X ~ N(μ, σ^2), thì Y ~ χ^2(n – 1), với

$$Y = \frac{(n-1)S_X^2}{\sigma^2}$$

7. Khoảng Biến Thiên (Sample Range)?

⇨ Là 1 thống kê có giá trị = giá trị lớn nhất – giá trị nhỏ nhất trong Sample

8. Phương Pháp Hàm Ước Lượng (Method Of Moments)?

⇨ Là phương pháp dùng để ước lượng tham số của Population = cách tính tham số của Sample

⇨ Ví dụ

⇨ Công thức tính Sample Variance chính là 1 hàm ước lượng cho Population Variance, đây là hàm ước lượng không chệch (Unbiased Estimator) vì kì vọng của nó đúng = Population Variance

9. Ước Lượng Hợp Lí Tối Đa (Maximum Likelihood Estimation)?

⇨ Giả sử ta có 1 dãy số liệu thống kê được cho từ 1 Sample, bài toán đặt ra là làm cách nào để tìm được 1 phân phối hợp nhất với dãy số liệu này, bằng cách dựa vào đặc điểm phân bố, ta có thể chọn kiểu phân phối phù hợp

⇨ Ví dụ

⇨ Nếu số liệu phân bố đối xứng và tập trung nhiều ở giữa, thì nghĩ ngay đến phân phối chuẩn

⇨ Sau khi đã chọn được kiểu phân phối, bước tiếp theo là xác định các tham số của phân phối này như Mean, Variance, … bằng phương pháp ước lượng hợp lí tối đa, nghĩa là chọn các tham số để tạo ra 1 phân phối sao cho khả năng để ta chọn được số liệu đã cho từ phân phối này là cao nhất, nghĩa là tích Likelihood của tất cả số liệu phải cao nhất

⇨ Cho dãy số liệu $x_1, x_2, \ldots, x_n$, giả sử f(x, θ) là PDF hoặc PMF của phân phối ứng với dãy số liêu này, khi đó θ chính là tham số mà chúng ta cần tìm, ta có tích Likelihood của tất cả số liệu là

$$L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

⇨ Như vậy bây giờ ta phải tìm θ sao cho L(θ) đạt cực đại, cách dễ nhất là lấy Logarithm, vì khi đó điểm cực đại sẽ không thay đổi

$$ln\big(L(\theta)\big) = ln(\prod_{i=1}^{n} f(x_i, \theta)) = \sum_{i=1}^{n} ln\big(f(x_i, \theta)\big) = g(\theta)$$

⇨ Bây giờ tiến hành tìm điểm cực đại của g(θ) = cách tìm điểm để đạo hàm = 0

$$\frac{\partial}{\partial\theta} g(\theta) = 0 \Leftrightarrow \frac{\partial}{\partial\theta} \sum_{i=1}^{n} ln\big(f(x_i, \theta)\big) = 0 \Leftrightarrow \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i, \theta)}{f(x_i, \theta)} = 0$$

⇨ Như vậy, tóm lại, để tìm tham số θ, ta chỉ cần giải phương trình sau

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(x_i, \theta)}{f(x_i, \theta)} = 0$$

⇨ Thông thường 1 phân phối chỉ có 1 cực đại nên không cần thiết phải sử dụng đạo hàm cấp 2 để kiểm tra cực đại hay cực tiểu

10. Ước Lượng = Khoảng Tin Cậy (Confidence Interval)?

⇨ Giả sử ta có 1 dãy số liệu thống kê $x_1, x_2, \ldots, x_n$ được cho từ 1 Sample, và ta đã biết X thuộc kiểu phân phối có PDF là f(x, θ), bài toán đặt ra là bây giờ phải xác

định xem tham số θ của phân phối này nằm trong khoảng nào thay vì dự đoán nó là 1 giá trị cụ thể, khoảng này gọi là khoảng tin cậy [a, b]

⇨ Độ tin cậy (Confidence Level) = 1 – α chính là xác suất để θ rơi vào khoảng tin cậy, với α là mức ý nghĩa do chúng ta đặt

⇨ Để xác định khoảng tin cậy, đầu tiên ta cần tạo biến ngẫu nhiên $Y = g(X_1, X_2, \ldots, X_n, \theta)$ sao cho mặc dù ta không biết θ nhưng vẫn biết được phân phối của Y

⇨ Ví dụ

⇨ Giả sử n = 9, $X \sim N(\mu, \sigma^2)$, σ^2 đã biết và = 1, $\bar{x}$ đã tính được và = 2.4, μ chưa biết, thì μ chính là tham số mà ta cần tìm khoảng tin cậy, dễ thấy

$$Y = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} = 3(\bar{X}-\mu) \sim N(0,1)$$

⇨ Như vậy mặc dù ta không biết μ nhưng ta vẫn biết phân phối của Y

⇨ Tiếp theo, với độ tin cậy = 1 – α đã đặt ra, ta tìm 2 cận c và d sao cho $P(Y \in [c, d]) = 1 - \alpha$, nghĩa là tìm khoảng mà xác suất Y rơi vào = 1 – α, dễ thấy sẽ có vô số c và d thỏa mãn điều kiện này nhưng ta sẽ chỉ chọn khoảng [c, d] ngắn nhất, vì càng ngắn thì càng chắc chắn

⇨ Ví dụ

⇨ Tiếp tục với ví dụ trên, giả sử ta đặt mức ý nghĩa α = 5%, khi đó độ tin cậy sẽ là 95%, như vậy ta cần phải tìm khoảng [c, d] ngắn nhất sao cho xác suất Y rơi vào khoảng này là 95%, dễ thấy khoảng này là [–2, 2], đây còn gọi là khoảng đối xứng, 1 cách tổng quát, nếu Y là phân phối chuẩn thì muốn khoảng [c, d] ngắn nhất thì nó phải đối xứng qua Mean

⇨ Sau khi đã xác định được khoảng [c, d] mà Y khả năng cao sẽ rơi vào, từ công thức của Y, ta sẽ suy ra được khoảng [a, b] mà θ khả năng cao sẽ rơi vào, khi đó giá trị (b – a) / 2 gọi là dung sai (Margin Of Error)

⇨ Ví dụ

⇨ Tiếp tục với ví dụ trên, ta có

$$P(Y \in [-2,2]) = P(3(\bar{X}-\mu) \in [-2,2]) = P\left((\bar{X}-\mu) \in \left[-\frac{2}{3}, \frac{2}{3}\right]\right) =$$
$$P\left((-\mu) \in \left[-\frac{2}{3} - \bar{X}, \frac{2}{3} - \bar{X}\right]\right) = P\left(\mu \in \left[\bar{X} - \frac{2}{3}, \bar{X} + \frac{2}{3}\right]\right) =$$
$$P\left(\mu \in \left[2.4 - \frac{2}{3}, 2.4 + \frac{2}{3}\right]\right) = P\left(\mu \in \left[\frac{26}{15}, \frac{46}{15}\right]\right) = 95\%$$

⇨ Như vậy, ta đã tìm được khoảng tin cậy của μ là [26 / 15, 46 / 15] với độ tin cậy = 95%

⇨ Một cách tổng quát, với $X \sim N(\mu, \sigma^2)$, σ^2 đã biết, $\bar{x}$ đã tính, c = d = C, thì khoảng tin cậy đối xứng của μ là

$$\mu \in \left[\bar{x} - \frac{C\sigma}{\sqrt{n}}, \bar{x} + \frac{C\sigma}{\sqrt{n}}\right]$$

⇨ Nếu ta chọn c và d sao cho b = ∞, thì [a, b] = [a, ∞] gọi là khoảng tin cậy tối thiểu, còn nếu chọn c và d sao cho a = –∞, thì [a, b] = [–∞, b] gọi là khoảng tin cậy tối đa

⇨ Trường hợp $X \sim N(\mu, \sigma^2)$ mà σ^2 chưa biết, ta tạo biến ngẫu nhiên Z như sau

$$Z = \sqrt{n}\frac{\bar{X}-\mu}{S_X}$$

⇨ Dễ dàng chứng minh được $Z \sim St(n-1)$, ta có

$$Z = \sqrt{n}\frac{\bar{X}-\mu}{S_X} = \sqrt{n}\frac{\bar{X}-\mu}{\sigma} / \frac{S_X}{\sigma} = Y\sqrt{n-1} / \sqrt{\frac{(n-1)S_X^2}{\sigma^2}} = \sqrt{n-1}\frac{Y}{\sqrt{W}}$$
$$W = \frac{(n-1)S_X^2}{\sigma^2}$$

⇨ Do Y ~ N(0, 1) và W ~ χ^2(n – 1) nên Z ~ St(n – 1)

⇨ Các bước để tìm khoảng tin cậy cũng tương tự, thay vì dùng Y có phân phối chuẩn thì dùng Z có phân phối Student, 1 cách tổng quát, với X ~ N(μ, σ^2), σ^2 chưa biết, $\bar{x}$ và s_x đã tính, c = d = C, thì khoảng tin cậy đối xứng của μ là

$$\mu \in \left[\bar{x} - \frac{Cs_x}{\sqrt{n}}, \bar{x} + \frac{Cs_x}{\sqrt{n}}\right]$$

⇨ Trường hợp X ~ N(μ, σ^2) mà μ đã biết, σ^2 là tham số cần tìm khoảng tin cậy, ta tạo biến ngẫu nhiên U như sau

$$U = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$$

⇨ Dễ thấy U ~ χ^2(n)

⇨ Các bước để tìm khoảng tin cậy cũng tương tự, thay vì dùng Y có phân phối chuẩn thì dùng U có phân phối chi bình phương, 1 cách tổng quát, với trường hợp trên, μ đã biết, chọn c và d sao cho P(U ≤ c) = P(U ≥ d), thì khoảng tin cậy 2 phía của σ^2 là

$$\sigma^2 \in \left[\frac{1}{d}\sum_{i=1}^{n}(x_i-\mu)^2, \frac{1}{c}\sum_{i=1}^{n}(x_i-\mu)^2\right]$$

⇨ Trường hợp X ~ N(μ, σ^2) mà μ chưa biết, σ^2 là tham số cần tìm khoảng tin cậy, ta dùng lại biến ngẫu nhiên W đã tạo ở trên

⇨ Các bước để tìm khoảng tin cậy cũng tương tự, thay vì dùng Y có phân phối chuẩn thì dùng W có phân phối chi bình phương, với trường hợp trên, μ chưa biết, s_x đã tính, chọn c và d sao cho P(W ≤ c) = P(W ≥ d), thì khoảng tin cậy 2 phía của σ^2 là

$$\sigma^2 \in \left[\frac{(n-1)s_x^2}{d}, \frac{(n-1)s_x^2}{c}\right]$$

⇨ Trường hợp X ~ Bernoulli(p), p là tham số cần tìm khoảng tin cậy, ta tạo biến ngẫu nhiên V như sau

$$V = \sqrt{n}\frac{\bar{X}-p}{\sqrt{p(1-p)}}$$

⇨ Dễ thấy V ~ N(0, 1), về cơ bản, V chính là Y nhưng cho phân phối Bernoulli

⇨ Các bước để tìm khoảng tin cậy cũng tương tự, 1 cách tổng quát, với trường hợp trên, $\bar{x}$ đã tính, c = d = C, thì khoảng tin cậy đối xứng của p là

$$p \in \left[\bar{x} - C\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}, \bar{x} + C\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\right]$$

⇨ Trường hợp X ~ Bernoulli(p) nhưng ta lại muốn ước lượng n để sai số ≤ ε_0, ta có

$$C\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \le \varepsilon_0 \Leftrightarrow n \ge \frac{C^2\bar{x}(1-\bar{x})}{\varepsilon_0^2} \le \frac{C^2}{4\varepsilon_0^2}$$

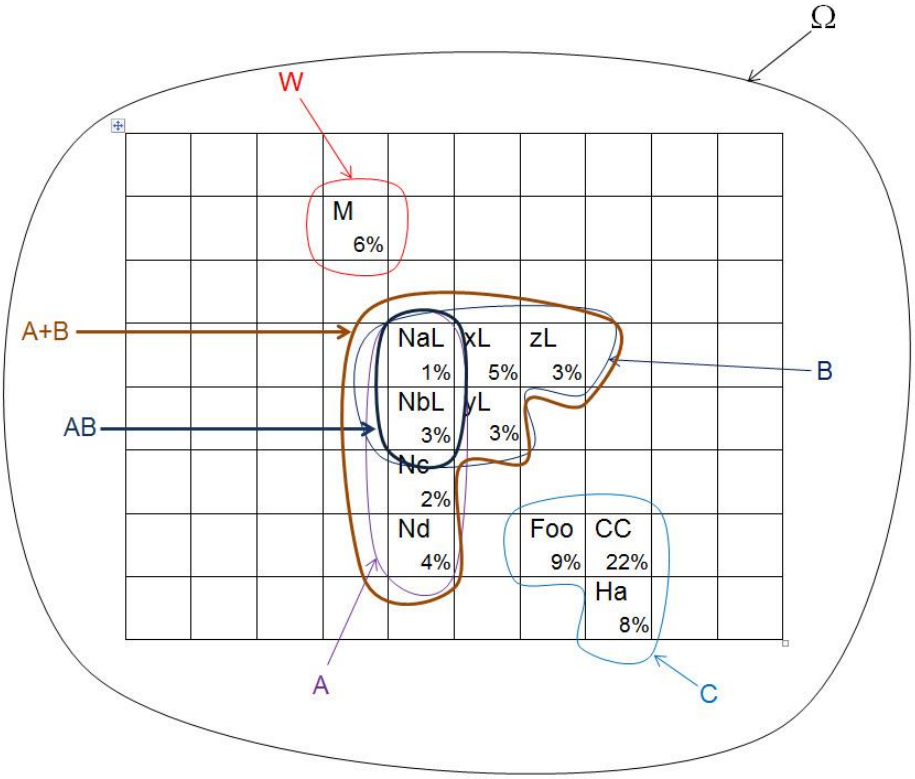
⇨ Vậy để cho chắc kèo sai số sẽ ≤ ε_0 cho trước, thì chọn n sao cho

$$n \ge \frac{C^2}{4\varepsilon_0^2}$$

Probability – Xác Suất:

1. Mô Hình Hóa Xác Suất?

⇨ Cho cái bảng sau



⇨ Không gian mẫu (Sample Space) Ω, chính là nguyên cái bảng này

⇨ Phép thử (Experiment) T, chính là cái hành động chọn ngẫu nhiên 1 trong các ô này

⇨ Biến cố sơ cấp (Elementary Event) là việc bạn chọn được 1 ô nào đó, ví dụ, biến cố sơ cấp W = “chọn được ô có tên M”

⇨ Biến cố (Event) là việc bạn chọn được ô trong 1 nhóm ô nào đó, ví dụ, biến cố A = “chọn được ô có chữ cái đầu tiên là N”

⇨ Giả sử bạn chọn n lần, thì số lần biến cố nào đó xảy ra gọi là tần số của biến cố đó, lấy tần số chia cho n, được tần suất của biến cố

⇨ Giả sử n = ∞, thì tần suất sẽ trở thành xác suất, ví dụ xác suất của biến cố B = “chọn được ô có chữ cái cuối cùng là L” là P(B) = 15%

⇨ Mức ý nghĩa (Significance Level) là 1 số mà nếu xác suất của biến cố nào đó < số này thì ta coi như biến cố đó sẽ không thể nào xảy ra, ví dụ ta đặt mức ý nghĩa = 1.5%, thì biến cố “chọn được ô có tên NaL” theo ta sẽ không thể xảy ra, vì xác suất của nó = 1%

⇨ Tổng biến cố A + B là biến cố “chọn được ô trong nhóm A hoặc nhóm B”

⇨ Tích biến cố AB là biến cố “chọn được ô nằm trong cả nhóm A và nhóm B”

⇨ $\bar{A}$ là biến cố “chọn được ô không nằm trong nhóm A”, A và $\bar{A}$ gọi là 2 biến cố đối lập (Complementary Events)

⇨ A và C gọi là 2 biến cố xung khắc (Mutually Exclusive/Disjoint Events) nếu biến cố “chọn được ô nằm trong cả nhóm A và nhóm C” không thể xảy ra

⇨ A và B gọi là 2 biến cố độc lập (Independent Events) chỉ khi P(AB) = P(A)P(B), còn nếu P(AB) ≠ P(A)P(B), thì A và B gọi là 2 biến cố phụ thuộc (Dependent

Events), ví dụ P(AB) = 4% ≠ P(A)P(B) = 10% * 15% = 1.5%, nên A và B là 2 biến cố phụ thuộc

2. Ví Dụ Cụ Thể Về Mô Hình Hóa Xác Suất?

⇨ Giả sử ta tiến hành khảo sát 1 nhóm sinh viên xem họ thích những môn nào trong các môn A, B

⇨ Ta dễ dàng lập được bảng

Ω

Thích A đéo thích B	Thích cả A và B
Đéo thích môn nào	Thích B đéo thích A

⇨ Dễ thấy biến cố W = "khảo sát trúng sinh viên thích môn A nhưng không thích môn B" là 1 biến cố sơ cấp

⇨ Biến cố C = "khảo sát trúng sinh viên thích môn A" là tổng biến cố W + E, E là biến cố sơ cấp "khảo sát trúng sinh viên thích cả môn A và môn B"

⇨ Biến cố E là tích biến cố CF, F là biến cố "khảo sát trúng sinh viên thích môn B"

3. Biểu Đồ Tần Suất (Histogram) Là Gì?

⇨ Là biểu đồ mà trục hoành chia thành các đoạn mà thường là = nhau, độ cao mỗi đoạn là tần suất của biến cố "giá trị trong đoạn này xuất hiện"

4. Xác Suất Của 1 Sự Kiện Khi Biết 1 Sự Kiện Khác Đã Xảy Ra?

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

⇨ A là sự kiện muốn tính xác suất xảy ra khi B đã xảy ra

⇨ Ví dụ

⇨ Tính xác suất tôi bị bệnh biết tôi dương tính với Virus, A là sự kiện tôi bị bệnh, còn B là sự kiện tôi dương tính với Virus, dễ thấy P(A | B) = 100% nếu xét nghiệm lúc nào cũng đúng

⇨ Nếu A và B là 2 sự kiện độc lập thì dễ thấy

$$P(A|B) = P(A)$$

5. Dương Tính Giả (False Positive) Là Gì?

⇨ Là 1 kết quả dự đoán, khi ta dự đoán 1 bệnh nhân dương tính, nhưng sự thật thì anh ta âm tính, kết quả này chính là dương tính giả

6. Định Lý Bayes?

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

7. Cách Dùng Định Lý Bayes Để Cập Nhật Lại Sự Tin Tưởng?

⇨ Giả sử đặt trước mặt bạn 1 người lạ hoắc, và bạn phải dự đoán xem thằng đó là nông dân hay lập trình viên, lúc đầu khi chưa có gợi ý gì, thì bạn 10% tin rằng người đó là lập trình viên, vì cứ 100 người trong xã hội thì có 90 người là nông dân, 10 người là lập trình viên, vậy P(A) = 10%, A là sự kiện người này là lập trình viên

⇨ Bây giờ cho thêm gợi ý rằng người này sở hữu máy tính lượng tử, thì bạn tính toán rằng cứ 90 người nông dân thì có 30 người sở hữu máy tính lượng tử, và cứ 10 lập trình viên thì có 8 người sở hữu máy tính lượng tử, vậy
P(B) = (30 + 8) / (90 + 10) = 38%, B là sự kiện người này có máy tính lượng tử, dễ thấy P(B | A) = 8 / 10 = 80%, thay số vào công thức Bayes, được
P(A | B) ≈ 21.05%

⇨ Như vậy, bây giờ ta 21.05% tin rằng người này là lập trình viên, > 10% trước đó, niềm tin của ta được củng cố

8. Phép Thử Bernoulli?

⇨ Là phép thử mà không gian mẫu chỉ có 2 phần tử

9. Công Thức Bernoulli?

⇨ Xét 1 phép thử Bernoulli, gọi xác suất trả về phần tử thứ 1 là p, phép thử được thực hiện n lần, khi đó, xác suất để có đúng k lần trả về phần tử thứ 1 là

$$P_n(k) = C_n^k p^k (1-p)^{n-k}$$

10. Xác Suất Để Bạn Được N Câu Trở Lên Trong 1 Bài Kiểm Tra?

⇨ Cho bài kiểm tra gồm n câu, mỗi câu gồm m đáp án, khi này xác xuất để bạn làm đúng k câu trở lên khi khoanh lụi hoàn toàn là

$$\frac{1}{m^n}\sum_{i=0}^{i=n-k}(m-1)^i C_n^i$$

⇨ Ví dụ

⇨ Bài kiểm tra gồm 15 câu, mỗi câu 4 đáp án, xác suất để bạn làm đúng 9 câu trở lên khi khoanh lụi là

$$\frac{1}{4^{15}}\sum_{i=0}^{i=15-9}(4-1)^i C_{15}^i = \frac{1}{4^{15}}\left(3^0 C_{15}^0 + 3^1 C_{15}^1 + 3^2 C_{15}^2 + \cdots + 3^6 C_{15}^6\right) \approx 0.42\%$$

⇨

Discrete Distribution – Phân Phối Rời Rạc:

1. Bảng Phân Phối Xác Suất?

⇨ Là bảng có dạng

X	x_1	x_2	x_3	…	x_4
P	p_1	p_2	p_3	…	p_4

⇨ X là biến ngẫu nhiên rời rạc

$$p_n = P(X = x_n)$$

2. Hàm Xác Suất (PMF – Probability Mass Function)?

⇨ Là hàm có dạng

$$f(x) = P(X = x)$$

3. Hàm Phân Phối Tích Lũy (CDF – Cummulative Distribution Function)?

⇨ Là hàm có dạng

$$f(x) = P(X \le x)$$

4. Phân Phối Bernoulli (Bernoulli Distribution)?

⇨ Cho X ~ Bernoulli(p), khi đó PMF của X là

$$f(x) = \begin{cases} p, x = 1 \\ 1-p, x = 0 \\ 0, x \notin \{0,1\} \end{cases}, p \in [0,1]$$

⇨ Dễ thấy

$$E[X] = p$$
$$Var[X] = p(1-p)$$

5. Mô Hình Xác Xuất Hình Học (Geometric Probability Model)?

⇨ Cho X ~ Geom(p), khi đó PMF của X là

$$f(x) = \begin{cases} p(1-p)^{x-1}, x \in \{1,2,\dots,n\} \\ 0, x \notin \{0,1,2,\dots,n\} \end{cases}$$

⇨ Kì vọng và Variance của X là

$$E[X] = \frac{1}{p}$$
$$Var[X] = \frac{1-p}{p^2}$$

⇨ Cho 1 phép thử với tỉ lệ thành công là p, dễ thấy f(x) chính là xác xuất để sau đúng x – 1 phép thử thất bại, phép thử thứ x thành công

6. Phân Phối Nhị Thức (Binomial Distribution)?

⇨ Cho X ~ B(n, p), khi đó PMF của X là

$$f(x) = \begin{cases} C_n^x p^x (1-p)^{n-x}, x \in \{0,1,2,\dots,n\} \\ 0, x \notin \{0,1,2,\dots,n\} \end{cases}$$

⇨ f(x) chính là xác suất để trong n phép thử Bernoulli với Ω = {0, 1} có đúng x lần trả về 1

⇨ Kì vọng và Variance của X là

$$E[X] = np$$
$$Var[X] = np(1-p)$$

⇨ Chứng minh

⇨ Dễ thấy X chính là tổng của n biến ngẫu nhiên độc lập với cùng 1 phân phối Bernoulli

$$X = \sum_{i=1}^{n} X_i, X_i \sim Bernoulli(p), \forall i$$
$$\Rightarrow E[X] = \sum_{i=1}^{n} E[X_i] = np$$
$$\Rightarrow Var[X] = \sum_{i=1}^{n} Var[X_i] = np(1-p)$$

⇨ Mode của X là

$$M_0(X) \in N \cap [(n+1)p - 1, (n+1)p]$$

⇨ Chứng minh

$$\frac{f(x)}{f(x-1)} = \frac{C_n^x p^x (1-p)^{n-x}}{C_n^{x-1} p^{x-1} (1-p)^{n-x+1}} = \frac{\frac{n-x+1}{x} C_n^{x-1}}{C_n^{x-1}} \frac{p}{1-p} =$$
$$\frac{(n-x+1)p}{x(1-p)} = 1 + \frac{(n-x+1)p - x(1-p)}{x(1-p)} = 1 + \frac{(n+1)p - x}{x(1-p)}$$
$$\Rightarrow x_0 = M_0(X) \Leftrightarrow \begin{cases} (n+1)p - x_0 \geq 0 \\ (n+1)p - (x_0+1) \leq 0 \\ x_0 \in N \end{cases} \Leftrightarrow \begin{cases} x_0 \in [(n+1)p - 1, (n+1)p] \\ x_0 \in N \end{cases}$$

7. Phân Phối Đa Thức (Multinomial Distribution)?

⇨ Đây là dạng mở rộng của phân phối nhị thức ra nhiều biến, PMF của nó là

$$f(x_1, x_2, \dots, x_k) = \begin{cases} n! \sum_{i=1}^{k} \frac{p_i^{x_i}}{x_i!}, \sum_{i=1}^{k} x_i = n, x_i \in \{0,1,2,\dots,n\}, \forall i \\ 0, otherwise \end{cases}$$

⇨ Cho bài toán thực hiện n phép thử, n phép thử này y chang phép thử Bernoulli, chỉ có điều có > 2 khả năng xảy ra, Ω = {1, 2, 3, …, k}, với xác suất tương ứng là {p_1, p_2, p_3, …, p_k}, câu hỏi đặt ra là tính xác suất để trong n phép thử này, số lần

xuất hiện của 1 đúng = x_1, số lần xuất hiện của 2 đúng = x_2, … là bao nhiêu, khi đó PMF trên sẽ cho ta câu trả lời

⇨ Dễ thấy, $X_i \sim B(n, p_i)$

8. Phân Phối Siêu Bội (Hypergeometric Distribution)?

⇨ Cho $X \sim H(N, K, n)$, khi đó PMF của X là

$$f(x) = \begin{cases} \frac{C_K^x C_{N-K}^{n-x}}{C_N^n}, x \in [max(0, n-N+K), min(n, K)] \\ 0, x \notin [max(0, n-N+K), min(n, K)] \end{cases}$$

⇨ Công thức này có nghĩa là bạn có N quả bóng, K quả màu đỏ, còn lại màu xanh, bạn bốc ngẫu nhiên n quả, khi đó f(x) chính là xác suất để trong n quả đó, có đúng x quả màu đỏ

⇨ Kì vọng và Variance của X là

$$E[X] = \frac{nK}{N}$$
$$Var[X] = \frac{nK}{N}\left(1 - \frac{K}{N}\right)\left(\frac{N-n}{N-1}\right)$$

⇨ Chứng minh

$$E[A_X^a] = \sum_{x=0}^{n} A_x^a P(X = x) = \sum_{x=1}^{n} A_x^a \frac{\frac{A_K^a}{A_x^a} C_{K-a}^{x-a} C_{(N-a)-(K-a)}^{(n-a)-(x-a)}}{\frac{A_N^a}{A_n^a} C_{N-a}^{n-a}} =$$

$$\frac{A_n^a A_K^a}{A_N^a} \sum_{x=1}^{n} \frac{C_{K-a}^{x-a} C_{(N-a)-(K-a)}^{(n-a)-(x-a)}}{C_{N-a}^{n-a}} = \frac{A_n^a A_K^a}{A_N^a}$$

$$\Rightarrow E[X] = E[A_X^1] = \frac{A_n^1 A_K^1}{A_N^1} = \frac{nK}{N}$$

$$\Rightarrow Var[X] = E[X^2] - (E[X])^2 = E[X(X-1) + X] - (E[X])^2 =$$

$$E[X(X-1)] + E[X] - (E[X])^2 = E[A_X^2] + E[X] - (E[X])^2 = \frac{A_n^2 A_K^2}{A_N^2} + \frac{nK}{N} - \left(\frac{nK}{N}\right)^2 =$$

$$\frac{n(n-1)K(K-1)}{N(N-1)} + \frac{nK}{N} - \left(\frac{nK}{N}\right)^2 = \frac{nK}{N}\left(1 - \frac{K}{N}\right)\left(\frac{N-n}{N-1}\right)$$

⇨ Khi n << N, thì

$$Var[X] \approx \frac{nK}{N}\left(1 - \frac{K}{N}\right)$$

⇨ Dễ thấy, khi n << N, đồ thị của PMF của X sẽ gần giống với đồ thị của PMF của biến ngẫu nhiên $Y \sim B(n, K / N)$

9. Phân Phối Poisson (Poisson Distribution)?

⇨ Cho $X \sim Poisson(\lambda)$, khi đó PMF của X là

$$f(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!}, x \in N \\ 0, x \notin N \end{cases}$$

⇨ f(x) chính là xác suất để trong n phép thử Bernoulli với $\Omega = \{0, 1\}$ có đúng x lần trả về 1, với n cực lớn và p là xác suất trả về 1 cực thấp, sao cho $np = \lambda$

⇨ Chứng minh

⇨ Cho $Y \sim B(n, p)$, $x \in N \cap [0, n]$, khi đó

$$P(Y = x) = C_n^x p^x (1-p)^{n-x} = \frac{A_n^x}{n^x} \frac{(np)^x}{x!} (1-p)^{n-x} = \frac{A_n^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} \frac{\lambda^x}{x!}$$

⇨ Ta có

$$\lim_{n \to \infty} \frac{A_n^x}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} = \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^{n-x} = \lim_{n \to \infty} \left(\left(1 - \frac{1}{\frac{n}{\lambda}}\right)^{-\frac{n}{\lambda}}\right)^{-\lambda + \frac{\lambda}{n}x} = e^{-\lambda}$$

⇨ Vậy với n cực lớn và p cực thấp, thì

$$P(Y = x) \approx e^{-\lambda}\frac{\lambda^x}{x!}$$

⇨ Như vậy, thay vì dùng phân phối nhị phân thì ta có thể dùng phân phối Poisson để biểu diễn Y

⇨ Kì vọng và Variance của X là

$$E[X] = Var[X] = \lambda$$

⇨ Chứng minh

⇨ Dễ thấy bản chất của phân phối Poisson là phân phối nhị phân với n rất lớn và p rất nhỏ nên kì vọng của phân phối Poisson là E[X] = np = λ

⇨ Tương tự Var[X] = np(1 – p) = np = λ

Continuous Distribution – Phân Phối Liên Tục:

1. Hàm Mật Độ Xác Suất (PDF – Probability Density Function)?

⇨ Là hàm f(x) sao cho

$$\int_a^b f(x)dx = P(X \in [a, b])$$

⇨ X là biến ngẫu nhiên liên tục

⇨ Dễ thấy PDF là đạo hàm của CDF

2. Khả Năng Xảy Ra (Likelihood)?

⇨ Cho X là biến ngẫu nhiên liên tục, khi đó Likelihood của giá trị x_0 là $f(x_0)$, f(x) là PDF của X

3. Giá Trị Tới Hạn (Critical Value)?

⇨ Cho X là biến ngẫu nhiên liên tục có f(x) là PDF, x_a sẽ được gọi giá trị tới hạn mức a của X khi

$$P(X \geq x_a) = a$$

4. Hệ Số Đối Xứng (Skewness)?

⇨ Cho X là biến ngẫu nhiên liên tục có f(x) là PDF, độ lệch chuẩn σ_X, khi đó Skewness của X là

$$\gamma_1(X) = \frac{E[(X-E[X])^3]}{\sigma_X^3}$$

⇨ Dễ thấy Skewness chỉ = 0 khi đồ thị f(x) đối xứng

⇨ Nếu Skewness > 0, đồ thị f(x) lệch phải, < 0 thì lệch trái

⇨ Nếu X là biến ngẫu nhiên rời rạc thì Skewness của X vẫn được tính = công thức trên

5. Hệ Số Nhọn (Kurtosis)?

⇨ Cho X là biến ngẫu nhiên liên tục có f(x) là PDF, độ lệch chuẩn σ_X, khi đó Kurtosis của X là

$$\gamma_2(X) = \frac{E[(X-E[X])^4]}{\sigma_X^4}$$

⇨ Dễ thấy nếu đồ thị f(x) càng nhọn, thì phần đuôi càng rộng, do đó Kurtosis càng lớn

⇨ Nếu X là biến ngẫu nhiên rời rạc thì Kurtosis của X vẫn được tính = công thức trên

6. Phân Phối Đều Liên Tục (Continuous Uniform Distribution)?

⇨ Cho X ~ U[a, b], khi đó PDF của X là

$$f(x) = \begin{cases} \frac{1}{b-a}, x \in [a, b] \\ 0, x \notin [a, b] \end{cases}$$

⇨ Kì vọng và Variance của X là

$$E[X] = \frac{a+b}{2}$$
$$Var[X] = \frac{(b-a)^2}{12}$$

7. Phân Phối Chuẩn (Normal Distribution)?

⇨ Cho X ~ N(μ, σ^2), khi đó PDF của X là

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

⇨ Kì vọng và Variance của X là

$$E[X] = \mu$$
$$Var[X] = \sigma^2$$

⇨ CDF của X là

$$\Phi(x) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)$$

⇨ erf(x) gọi là hàm lỗi, cho biến ngẫu nhiên Y ~ N(0, 0.5) có PDF là g(x), khi đó

$$erf(x) = 2\int_0^x g(t)dt$$

⇨ MGF của X là

$$M_X(t) = e^{\mu t + \frac{1}{2}(\sigma t)^2}$$

⇨ Ước lượng xác suất phân phối chuẩn

$$P(X \in [\mu - \sigma, \mu + \sigma]) \approx 68\%$$
$$P(X \in [\mu - 2\sigma, \mu + 2\sigma]) \approx 95\%$$
$$P(X \in [\mu - 3\sigma, \mu + 3\sigma]) \approx 99.7\%$$

⇨ Có thể dùng phân phối chuẩn để ước lượng phân phối nhị thức, giả sử ta có biến ngẫu nhiên Y ~ B(n, p), vì bản chất phân phối nhị thức là tổng của nhiều biến ngẫu nhiên với cùng 1 phân phối Bernoulli, do đó, theo định luật giới hạn trung tâm, với n lớn, ta có thể coi phân phối nhị thức là phân phối chuẩn, hay Y ~ N(np, np(1 – p))

8. Phân Phối Gamma (Gamma Distribution)?

⇨ Cho X ~ Γ(α, β), khi đó PDF của X là

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, x \geq 0 \\ 0, x < 0 \end{cases}$$

⇨ Kì vọng và Variance của X là

$$E[X] = \alpha\beta$$
$$Var[X] = \alpha\beta^2$$

9. Phân Phối Chi Bình Phương (Chi Squared Distribution)?

⇨ Cho X ~ χ^2(r), khi đó PDF của X là

$$f(x) = \begin{cases} \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, x \geq 0 \\ 0, x < 0 \end{cases}$$

⇨ Dễ thấy X ~ Γ(r / 2, 2), do đó phân phối chi bình phương chỉ là 1 dạng đặc biệt của phân phối Gamma

⇨ Kì vọng và Variance của X là

$$E[X] = r$$
$$Var[X] = 2r$$

⇨ Bản chất của X chính là tổng bình phương của r biến ngẫu nhiên độc lập với phân phối chuẩn tắc

10. Phân Phối Mũ (Exponential Distribution)?

⇨ Cho X ~ Exp(λ), khi đó PDF của X là

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases}$$

⇨ Dễ thấy X ~ Γ(1, 1 / λ), do đó phân phối mũ chỉ là 1 dạng đặc biệt của phân phối Gamma

⇨ Kì vọng và Variance của X là

$$E[X] = \frac{1}{\lambda}$$
$$Var[X] = \frac{1}{\lambda^2}$$

11. Phân Phối Student (Student's T Distribution)?

⇨ Cho X ~ St(n), khi đó PDF của X là

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

⇨ Cho Y ~ N(0, 1), Z ~ χ^2(n), Y và Z độc lập, khi đó

$$X = \sqrt{n}\frac{Y}{\sqrt{Z}}$$

⇨ Kì vọng và Variance của X là

$$E[X] = 0$$
$$Var[X] = \frac{n}{n-2}$$

⇨ Dễ thấy nếu n cực lớn thì phân phối Student tương đương phân phối chuẩn tắc, thông thường với n ≥ 30, ta coi phân phối Student là phân phối chuẩn tắc luôn

12. Phân Phối Fisher (F Distribution)?

⇨ Cho X ~ F(n, m), khi đó PDF của X là

$$f(x) = \begin{cases} \frac{\sqrt{\frac{m^m (nx)^n}{(nx+m)^{n+m}}}}{x\mathrm{B}\left(\frac{n}{2}, \frac{m}{2}\right)}, x \geq 0 \\ 0, x < 0 \end{cases}$$

⇨ Cho Y ~ χ^2(n), Z ~ χ^2(m), Y và Z độc lập, khi đó

$$X = \frac{m}{n}\frac{Y}{Z}$$

⇨ Kì vọng và Variance của X là

$$E[X] = \frac{n}{n-2}$$
$$Var[X] = \frac{2m^2(n+m^2-2)}{n(m-2)^2(m-4)}$$

13. Có Phải Giá Trị Của Mọi PDF Tại Vô Cực Đều = 0?

⇨ Đúng, cho f(x) và g(x) là 2 PDF của 2 biến ngẫu nhiên nào đó, ta có

$$\int_{-\infty}^{\infty} f(x)g'(x)\,dx = -\int_{-\infty}^{\infty} f'(x)g(x)\,dx$$

Statistical Hypothesis Testing – Kiểm Định Giả Thuyết Thống Kê:

1. Bản Chất Của Kiểm Định Giả Thuyết Thống Kê?

⇨ Là bạn đặt ra 1 giả thuyết gốc H_0 (Null Hypothesis) từ 1 cơ sở nào đó, sau đó dựa vào dữ liệu mẫu để quyết định bác bỏ hoặc chấp nhận giả thuyết này

⇨ Thông thường đi đôi với H_0 còn có đối thuyết H_1 (Alternative Hypothesis), H_0 và H_1 là cặp mệnh đề xung khắc

⇨ Có 3 loại giả thuyết

⇨ Giả thuyết về quy luật phân phối của biến ngẫu nhiên

⇨ Ví dụ

H_0 = "X có phân phối chuẩn" H_1 = "X không có phân phối chuẩn"

⇨ Giả thuyết về tham số của biến ngẫu nhiên

⇨ Ví dụ

H_0 = "$\mu = 15$" H_1 = "$\mu > 15$"

⇨ Giả thuyết về tính độc lập của các biến ngẫu nhiên

⇨ Ví dụ

H_0 = "X và Y độc lập" H_1 = "X và Y không độc lập"

2. Kiểm Định Giả Thuyết Về Tham Số Population?

⇨ Để giải bài toán kiểm định này thì yêu cầu $X \sim N(\mu, \sigma^2)$, nếu không thì kích thước Sample n phải > 30 để định luật giới hạn trung tâm xảy ra

⇨ Trường hợp σ^2 đã biết và ta có cơ sở để cho rằng $\mu = \mu_0$ nào đó, ta có cặp giả thuyết thống kê

H_0 = "$\mu = \mu_0$"

⇨ Tùy thuộc vào yêu cầu bài toán mà ta sẽ chọn 1 trong các đối thuyết sau

H_1 = "$\mu \neq \mu_0$" H_1 = "$\mu > \mu_0$" H_1 = "$\mu < \mu_0$"

⇨ Tạo biến ngẫu nhiên Y như sau

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

⇨ Giả sử ta có độ tin cậy = 1 – α, với α là mức ý nghĩa, tìm khoảng [c, d] sao cho xác suất Y rơi vào khoảng này = độ tin cậy, nếu H_1 = "$\mu \neq \mu_0$" thì đây là khoảng đối xứng, nếu H_1 = "$\mu > \mu_0$" thì đây là khoảng tối đa, nếu H_1 = "$\mu < \mu_0$" thì đây là khoảng tối thiểu, sau đó tính y theo Y

$y = \sqrt{n}\frac{\bar{x}-\mu_0}{\sigma}$

⇨ Nếu y nằm ngoài khoảng [c, d] thì bác bỏ H_0 và thừa nhận H_1, nếu nằm trong thì chưa có cơ sở để bác bỏ H_0

⇨ Các trường hợp khác cũng làm tương tự, vì bản chất nó giống với việc bạn ước lượng = khoảng tin cậy

⇨ Trường hợp kiểm định về so sánh Population Mean của 2 Population khác nhau U và V, ta tạo biến ngẫu nhiên W như sau

$W = \frac{\bar{U}-\bar{V}}{\sqrt{\frac{\sigma_U^2}{n_U}+\frac{\sigma_V^2}{n_V}}} \sim N(0,1)$

⇨ Giả thuyết H_0 và H_1 tương tự, chỉ cần thay "μ" thành "μ_U" và "μ_0" thành "μ_V"

⇨ Cũng trường hợp trên, nhưng Variance của U và V chưa biết, thì chỉ cần thay σ_U thành S_U, σ_V thành S_V

⇨ Trường hợp U ~ Bernoulli(p_U) và V ~ Bernoulli(p_V), thì cũng làm tương tự như trên, nhưng thay 2 cái σ^2 thành p(1 – p), với

$$p = \frac{n_U\bar{U}+n_V\bar{V}}{n_U+n_V}$$

⇨ Trường hợp kiểm định về so sánh Population Variance của 2 Population khác nhau U và V, ta tạo biến ngẫu nhiên G như sau

$$G = \frac{S_U^2}{S_V^2} \sim F(n_U - 1, n_V - 1)$$

⇨ Giả thuyết H_0 và H_1 tương tự như trên, chỉ cần thay Mean thành Variance

3. So Sánh 2 Bộ Số Liệu?

⇨ Cho bộ số liệu quan sát được $x_1, x_2, \ldots, x_n$, và bộ số liệu theo lí thuyết $y_1, y_2, \ldots, y_n$

⇨ Ví dụ

⇨ Ta có 1 lô hàng gồm 200 sản phẩm, theo người bán, trong này sẽ có 10% lỗi, 20% kém chất lượng, 30% nát, 40% tốt, thì đáng lẽ sẽ có 20 lỗi, 40 kém chất lượng, 60 nát và 80 tốt, nhưng thế đéo nào khi kiểm tra thì thấy có 35 lỗi, 30 kém chất lượng, 65 nát và 70 tốt, vậy câu hỏi đặt ra ở đây là người bán có xạo lồn hay không

⇨ Ta có

⇨ H_0 = "2 bộ số liệu giống nhau"

⇨ H_1 = "2 bộ số liệu khác nhau"

⇨ Tạo biến ngẫu nhiên Q

$$Q = \sum_{i=1}^{n} \frac{(X_i - Y_i)^2}{Y_i} \sim \chi^2(n - 1)$$

⇨ Có phân phối cụ thể rồi thì kiểm định như bình thường

⇨ Trường hợp so sánh k bộ số liệu quan sát được với nhau mà đéo có bộ số liệu lí thuyết thì đầu tiên, cộng hợp toàn bộ các bộ lại với nhau, được bộ tổng, rồi chuẩn hóa bộ tổng, được tỉ lệ phần trăm theo lí thuyết, coi các bộ ban đầu là 1 bảng, rồi từ tỉ lệ phần trăm theo lí thuyết tạo ra bảng lí thuyết và áp dụng công thức trên, lưu ý khi này Q ~ χ^2((k – 1)(n – 1))

4. Kiểm Định Về Luật Phân Phối?

⇨ Giả sử cho bộ số liệu quan sát được nào đó, ta đặt ra giả thuyết là liệu nó có kiểu phân phối A nào đó hay không, ví dụ như phân phối chuẩn chẳng hạn

⇨ Để kiểm tra giả thuyết trên, đầu tiên là ước lượng các tham số của A, sau đó dùng A để tính bộ số liệu lí thuyết, rồi so sánh bộ số liệu quan sát với lí thuyết

Error – Sai Số:

1. Số Chữ Số Có Nghĩa?

⇨ Là số chữ số bắt đầu từ chữ số khác 0 đầu tiên từ bên trái sang

⇨ Ví dụ

0.000123 có 3 chữ số có nghĩa
0.0001230 có 4 chữ số có nghĩa

0.0001203 có 4 chữ số có nghĩa

2. Làm Tròn Số?

4.778 làm tròn lên 4.8
4.728 làm tròn xuống 4.7
4.75 làm tròn lên 4.8
4.85 làm tròn xuống 4.8

⇨ Gặp đúng 1 số 5 thì làm tròn lên nếu lẻ, xuống nếu chẵn

3. Dụng Cụ Đo Chia Vạch?

⇨ Cấp chính xác của dụng cụ này = nửa độ chia nhỏ nhất, vị dụ độ chia nhỏ nhất = 1m thì cấp chính xác = 0.5m

⇨ Cấp chính xác là độ lệch tối đa của giá trị đo với giá trị thực, ví dụ đầu cu ở giữa vạch 12 và 13, giả sử nó là 12.4, thì ta chọn 12, độ lệch = 0.4

⇨ Sai số hệ thống của dụng cụ này = độ chia nhỏ nhất, khi đo chiều dài 1 vật, ta đo 2 đầu, tức là mỗi đầu sẽ sai lệch 0.5 so với vạch, tổng lại thành 1

4. Sai Số Tuyệt Đối (Absolute Error)?

⇨ Giả sử ta thực hiện phép đo chiều dài cu n lần

⇨ Giá trị trung bình là Ā, được làm tròn như sau, giả sử giá trị mỗi lần đo có số chữ số sau dấu phẩy là n, thì làm tròn kết quả trung bình tới n + 1 chữ số, ví dụ đo được 7.4, 7.5, 7.5, thì trung bình = 7.47, giá trị lần đo thứ t là A_t, ta có sai số tuyệt đối của lần đo đó là

$$\Delta A_t = |A_t - \bar{A}|$$

⇨ Sai số tuyệt đối trung bình là

$$\overline{\Delta A} = \frac{1}{n}\sum_{t=1}^{t=n} \Delta A_t$$

⇨ Giá trị trên = sai số ngẫu nhiên, khi n từ 5 trở lên, còn không thì sai số ngẫu nhiên = sai số tuyệt đối lớn nhất trong n lần đo

⇨ Xét 1 số vô tỉ, ta làm tròn nó, khi đó sai số tuyệt đối ứng với hằng số này = $0.5*10^{-n}$, n là số chữ số sau phần thập phân

⇨ Ví dụ

π = 3.14 nên Δπ = 0.005
g = 9.872 nên Δg = 0.0005

5. Sai Số Tuyệt Đối Của Tổng Các Biến Sai Số?

⇨ Bằng tổng sai số tuyệt đối của các biến

6. Tính Sai Số Tuyệt Đối Của Biến Là Một Hàm Phức Tạp Của Nhiều Biến Sai Số Khác?

⇨ Lưu ý giá trị đạo hàm được tính với giá trị trung bình

$$x = f(a,b,c) \Rightarrow \Delta x = \left|\frac{\partial x}{\partial a}\right| \Delta a + \left|\frac{\partial x}{\partial b}\right| \Delta b + \left|\frac{\partial x}{\partial c}\right| \Delta c$$

7. Tính Sai Số Tương Đối?

$$\delta x = \frac{\Delta x}{\bar{x}} \times 100\%$$

⇨ Δx là sai số tuyệt đối

⇨ x̄ là giá trị trung bình

8. Sai Số Tương Đối Của Tích Các Biến Sai Số?

⇨ Bằng tổng sai số tương đối của các biến

⇨ Trong các công thức, nếu thế các hằng số như g, π, … thành giá trị cụ thể thì phải chọn giá trị sao cho đóng góp vào sai số tương đối của hằng số này phải < 0.1 lần tổng đóng góp của các đại lượng khác

9. Sai Số Toàn Phương Trung Bình?

⇨ = độ lệch chuẩn có hiệu chỉnh = σ

⇨ Giả sử đo nhiều hơn 10 lần, thì có thể biểu diễn kết quả của 1 lần đo như sau, giả sử lần đó đo được giá trị x

$$x \pm \sigma$$

10. Sai Số Toàn Phương Trung Bình Của Trung Bình?

⇨ Giả sử ta muốn thực hiện k đợt, mỗi đợt đo con cu m lần, thì giá trị trung bình mỗi đợt sẽ có độ lệch chuẩn đã hiệu chỉnh = sai số toàn phương trung bình của trung bình = công thức sau, σ là sai số toàn phương trung bình của đợt đầu tiên, những đợt sau ta không làm

$$\bar{\sigma}_x = \frac{\sigma}{\sqrt{m}}$$

⇨ Giả sử giá trị trung bình đo được trong đợt đầu = x̄, sai số hệ thống = Δx_{ht}, thì ta có thể biểu diễn giá trị chiều dài con cu như sau

$$\bar{x} \pm (\bar{\sigma}_x + \Delta x_{ht})$$

11. Sai Số Toàn Phần?

⇨ = tổng sai số hệ thống ΔA' và sai số ngẫu nhiên

$$\Delta A = \overline{\Delta A} + \Delta A'$$

⇨ Giả sử phép đo con cu thu được chiều dài trung bình Ā, sai số toàn phần ΔA, thì chiều dài con cu có thể biểu diễn dưới dạng

$$A = \bar{A} \pm \Delta A$$

⇨ Hoặc

$$\bar{A} - \Delta A < A < \bar{A} + \Delta A$$

⇨ Trước khi viết, ta phải làm tròn số, nếu chữ số có nghĩa đầu tiên của ΔA từ 3 trở lên, thì chỉ làm tròn sao cho còn sót lại đúng 1 chữ số có nghĩa

⇨ Nếu từ 2 trở xuống thì làm tròn 2 chữ số có nghĩa

⇨ Ví dụ

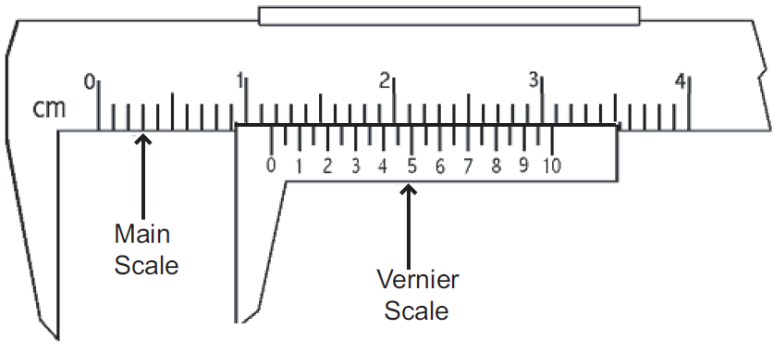
0.00498 làm tròn lên 0.005
0.00284 làm tròn xuống 0.0028
0.0031 làm tròn xuống 0.003

⇨ Làm tròn Ā sao cho số chữ số sau dấu phẩy của nó = số chữ số sau dấu phẩy của ΔA

⇨ Ví dụ

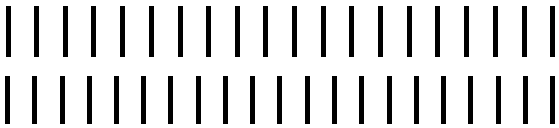
Ā = 1.23456, ΔA = 0.1, thì viết A = 1.2 ± 0.1
Ā = 123456, ΔA = 400, thì viết A = 123500 ± 400 = (123.5 ± 0.4) x 10^3

12. Thước Kẹp (Vernier Caliper)?



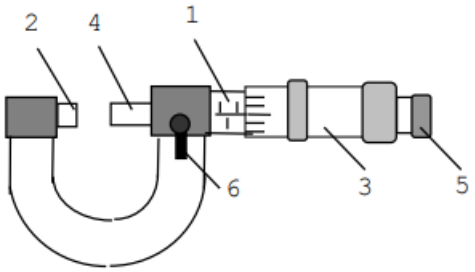
⇨ Giới hạn đo kí hiệu là <Cận Dưới> ÷ <Cận Dưới> <Đơn Vị>, ví dụ 0 ÷ 150mm

- ⇨ Giả sử độ chia nhỏ nhất trên thanh to là 1mm, chiều dài từ 0 đến 10 trên thanh nhỏ, hay du xích, là 19mm
- ⇨ Đặt vậy vào 2 thanh, kẹp chặt, xác định vạch có trị số lớn nhất trên thanh mà vạch 0 trên thanh nhỏ nằm bên phải nó, ví dụ 123mm, xác định vạch trên thanh nhỏ thẳng hàng nhất với vạch trên thanh to, giả sử vạch này ghi 4.5, thì kích thước vật = 123.45mm
- ⇨ Minh họa



- ⇨ Ở trên là 19 khoảng 1mm
- ⇨ Ở dưới là 20 khoảng 0.95mm
- ⇨ Giả sử chèn 1 vật đằng trước thanh dưới, kích thước 0.05mm, thì rõ ràng vạch thứ 2 từ trái sang của 2 thanh sẽ trùng nhau, mà vạch này lại ghi 0.5, ví dụ khác, vật có kích thước 0.1 mm, thì rõ ràng vạch thứ 3 của 2 thanh sẽ trùng nhau, mà vạch này lại ghi 1
- ⇨ Ngoài ra, trên thanh lớn còn 2 cái kẹp khác để đo kích thước bên trong vật, thay vì kẹp chặt vào thì đẩy chặt ra, sau đít thanh to cũng có 1 cái thanh nhọn chĩa ra đo độ sâu
- ⇨ Sai số hệ thống của thước kẹp = độ chia nhỏ nhất của nó, ví dụ 123.45mm thì 0.05mm = 1 / 20 vạch trên thanh nhỏ

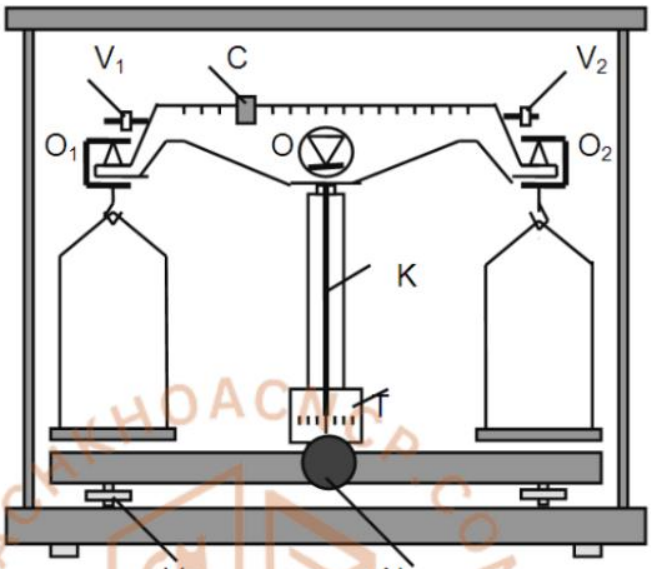
13. Thước Panme (Micrometer)?



Hình 4

- ⇨ Vỏ trụ 3 khắc 50 vạch ứng với 0 đến 0.5mm, các vạch được đánh số
- ⇨ Lõi 1 chứa dãy vạch trên và dưới ngăn cách bởi đường d ở giữa, trên ứng với 0, 1, 2, …, dưới ứng với 0.5, 1.5, 2.5, …, đơn vị mm
- ⇨ Khoảng cách giữa đầu 2 và 4 = vạch trên lõi gần nhất với mép vỏ 3, ví dụ vạch 6.5mm, + giá trị vạch thẳng hàng nhất với d trên vỏ 3, ví dụ 0.2mm, được 6.7mm, công thức tổng quát là 0.5k + 0.01m, k là tổng số vạch trên và dưới ở lõi 1 ta nhìn thấy được, không tính vạch 0, m là vạch trên vỏ 3 trùng đường d
- ⇨ Độ chính xác = 0.01mm = độ chia nhỏ nhất của vỏ 3
- ⇨ Cần gạt 6 có tác dụng hàm đầu 4, gạt sang phải sẽ mở hãm
- ⇨ Để đẩy đầu 4 tiến tới đầu 2, vặn lõi 5, nếu đầu 4 chạm vật thì sẽ nghe tách tách, khi đó dừng lại
- ⇨ Để thả đầu 4 ra, vặn vỏ 3
- ⇨ Trước khi đo, cần đẩy đầu 2 và 4 chạm nhau để kiểm xem giá trị đo có = 0, nghĩa là đường d phải thẳng hàng với vạch 0 của vỏ 3, nếu đường d thấp hơn vạch 0 của vỏ 3 n vạch thì kết quả đo phải trừ đi 0.01n

14. Cân Kỹ Thuật?



Hình 4

- ⇨ Cái thanh trên hình quần sịp là đòn cân, nó có kẻ 50 vạch, vạch đầu 0 gam, vạch cuối 1 gam, như vậy độ chia nhỏ nhất = 0.02 gam, trên đó có con mã C, cạnh trái nó chạm vào vạch nào thì khối lượng tương ứng sẽ được cộng vào cân bên phải
- ⇨ Ban đầu khi núm xoay N vặn hết về bên trái, đòn cân sẽ không quay được do bị đỡ bởi khung đỡ ở dưới, khi văn núm xoay N vào bên phải, đòn cân từ từ nâng lên, rời khung đỡ và có thể quay tự do, khi này kim K cũng sẽ quay theo đòn cân
- ⇨ Kim K lệch về bên nào thì bên đấy nhẹ hơn
- ⇨ Trước khi bỏ vật lên cân, phải vặn núm xoay N về hết phải, trượt C về 0, sau đó vặn ốc V_1 và V_2 để kim K chỉ vào vạch 0 là vạch ở giữa trên thước T ở dưới, tức là cân đang cân bằng, sau đó vặn N về hết trái rồi mới bỏ vật lên cân trái, các quả cân lên cân phải, rồi vặn N về hết phải, rồi chỉnh con mã C để cân bằng
- ⇨ Ta chỉ quan tâm tổng khối lượng các quả cân bên phải và giá trị con mã C, không quan tâm thước T
- ⇨ Các quả cân có trong hộp cân từ nhỏ đến lớn bao gồm 1, 2, 2, ,5, 10, 20, 20, 50, 100 gam
- ⇨ Độ nhạy S của cân kỹ thuật = kim K lệch bao nhiêu vạch trên T khi đặt 1 vật 1mg vào 1 bên của cân, bên còn lại trống, đơn vị là độ chia / mg
- ⇨ Độ chính xác α của cân kỹ thuật = 1 / S
- ⇨ Sai số hệ thống = độ chia nhỏ nhất của con mã C = 0.02 gam

Expected Value – Kì Vọng:

1. Kì Vọng Của 1 Biến Ngẫu Nhiên?

- ⇨ Cho X là 1 biến ngẫu nhiên, khi đó u(X) cũng là 1 biến ngẫu nhiên
- ⇨ Nếu X là biến ngẫu nhiên rời rạc, khi đó kì vọng của u(X) là

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)P(X = x)$$

- ⇨ Nếu X là biến ngẫu nhiên liên tục, khi đó kì vọng của u(X) là

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f(x)\,dx$$

- ⇨ f(x) là PDF của X

2. Kì Vọng Của Tổng 2 Biến Ngẫu Nhiên?

$$E[X+Y] = E[X] + E[Y]$$

3. Kì Vọng Của Tích 2 Biến Ngẫu Nhiên Độc Lập?

$$E[XY] = E[X]E[Y]$$

⇨ Chứng minh

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyP(X = x \,\&\, Y = y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyP(X = x)P(Y = y) = \int_{-\infty}^{\infty} xP(X = x)\int_{-\infty}^{\infty} yP(Y = y) = E[X]E[Y]$$

Variance – Phương Sai:

1. Variance Của 1 Biến Ngẫu Nhiên?

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

⇨ X là biến ngẫu nhiên

2. Độ Lệch Chuẩn (Standard Deviation) Của 1 Biến Ngẫu Nhiên?

$$\sigma_X = \sqrt{Var[X]}$$

3. Tại Sao Cần Độ Lệch Chuẩn Trong Khi Đã Có Variance?

⇨ Vì Variance có đơn vị đo không giống X trong khi độ lệch chuẩn thì giống

4. Variance Của Tổng 2 Biến Ngẫu Nhiên?

$$Var[X+Y] = Var[X] + 2Cov(X,Y) + Var[Y]$$

⇨ Nếu X và Y độc lập, ta có

$$Var[X+Y] = Var[X] + Var[Y]$$

⇨ Chứng Minh

$$Var[X+Y] = E[(X+Y-E[X+Y])^2] = E[(X-E[X]+Y-E[Y])^2] = E[(X-E[X])^2 + 2(X-E[X])(Y-E[Y]) + (Y-E[Y])^2] = E[(X-E[X])^2] + 2E[(X-E[X])(Y-E[Y])] + E[(Y-E[Y])^2] = Var[X] + 2Cov(X,Y) + Var[Y]$$

5. Variance Của 1 Số Lần Biến Ngẫu Nhiên?

$$Var[kX] = k^2 Var[X], \forall k \in R$$

6. Tại Sao Thay Gốc Tính Moment = Giá Trị Khác Thì Variance Thu Được Sẽ Luôn Lớn Hơn Khi So Với Gốc Là Mean?

⇨ Giả sử giá trị khác là v, lấy đạo hàm của Variance theo v

$$Var[X] = \frac{1}{n}\sum(x-v)^2 \Rightarrow \frac{d}{dv}Var[X] = -\frac{2}{n}\sum x - v \Rightarrow \frac{d}{dv}Var[X] = 0 \Leftrightarrow \sum x - v = 0 \Rightarrow n\mu - nv = 0 \Rightarrow v = \mu$$

⇨ Dễ thấy v = μ thì đạo hàm = 0 nên giá trị Variance nhỏ nhất khi v = μ

7. Variance Của 1 Biến Ngẫu Nhiên Dạng Vector Mà Các Phần Tử Của Nó Độc Lập?

⇨ Bằng tổng Variance của mỗi phần tử

8. Hiệp Phương Sai (Covariance) Của 2 Biến Ngẫu Nhiên Là Gì?

⇨ Ở đây ta xét 1 biến ngẫu nhiên là Vector 2D, có 2 phần tử là biến ngẫu nhiên X và biến ngẫu nhiên Y

⇨ Nếu 2 biến ngẫu nhiên độc lập, không có liên hệ gì với nhau, thì Covariance = 0, nếu chúng có phần nào phụ thuộc tuyến tính thì Covariance sẽ khác 0, ví dụ nếu

X tăng mà Y cũng tăng, thì Covariance > 0, nếu X tăng mà Y giảm, Covariance < 0

9. Công Thức Tính Covariance Cho Population?

$$Cov(X,Y) = \frac{1}{n}\sum(x-\bar{x})(y-\bar{y})$$

⇨ n là kích thước Population

10. Hệ Số Tương Quan (Correlation Coefficient) Là Gì?

⇨ Là Variance được chuẩn hóa, có giá trị trong đoạn [–1, 1], có giá trị - 1 hoặc 1 khi 2 biến ngẫu nhiên phụ thuộc tuyến tính 100%

11. Công Thức Hệ Số Tương Quan?

$$r = \frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}$$

12. Chứng Minh Hệ Số Tương Quan Có Độ Lớn Không Vượt Quá 1?

⇨ Do độ lớn tích vô hướng của 2 Vector luôn không vượt quá tích Module 2 Vector, nên

$$(\sum(x-\bar{x})(y-\bar{y}))^2 \le (\sum(x-\bar{x}))^2(\sum(y-y))^2 \Rightarrow$$
$$\frac{\frac{1}{n^2}(\sum(x-\bar{x})(y-\bar{y}))^2}{\frac{1}{n^2}(\sum(x-\bar{x}))^2(\sum(y-y))^2} \le 1 \Rightarrow \left|\frac{\frac{1}{n}\sum(x-\bar{x})(y-\bar{y})}{\frac{1}{n}\sum(x-\bar{x})\sum(y-y)}\right| \le 1 \Rightarrow$$
$$\left|\frac{Cov(X,Y)}{\sqrt{Var[X]Var[Y]}}\right| \le 1 \Rightarrow |r| \le 1$$

Moment:

1. Moment?

⇨ Cho điểm A và B, AB^k * 1 đại lượng nào đó tại điểm A, đây chính là Moment bậc k của A so với B

⇨ Trong thống kê, đại lượng nào đó ở đây chính là Likelihood

2. Moment Thô?

⇨ Là Moment so với gốc tọa độ

3. Standardized Moment?

⇨ Là Moment nhân với 1 hệ số

4. Các Moment Trong Thống Kê?

⇨ Mean là Moment thô bậc 1

⇨ Variance là Moment bậc 2 so với Mean

⇨ Skewness là Standardized Moment bậc 3 so với Mean

⇨ Kurtosis là Standardized Moment bậc 4 so với Mean

5. MGF (Moment Generating Function)?

⇨ Là cái hàm mà đạo hàm bậc k tại 0 của nó có giá trị = Moment thô bậc k

6. MGF Của 1 Biến Ngẫu Nhiên?

$$M_X(t) = E[e^{tX}]$$

⇨ X là biến ngẫu nhiên

⇨ Ví dụ

⇨ Giả sử X là biến ngẫu nhiên thuộc phân phối mũ có $\lambda = 1$, khi đó MGF của X là

$$M_X(t) = E[e^{tX}] = \int_0^\infty e^{tx}e^{-x}dx = \int_0^\infty e^{(t-1)x}dx = \frac{1}{1-t}$$
$$M_X'(0) = E[X] = 1$$
$$M_X''(0) = E[X^2] = 2 \Rightarrow Var[X] = E[X^2] - (E[X])^2 = 1$$

$M_X'''(0) = E[X^3] = 6$
...

⇨ Chứng Minh

⇨ Dùng chuỗi Taylor

$$M_X(t) = E[e^{tX}] = E\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots\right] =$$
$$1 + tE[X] + \frac{t^2}{2!}E[X^2] + \frac{t^3}{3!}E[X^3] + \cdots$$

⇨ Dễ thấy đạo hàm bậc k của MGF tại 0 có giá trị = Moment thô bậc k

7. Công Thức MGF Của Biến Ngẫu Nhiên Của Sample Mean?

$$M_{\bar{X}}(t) = M_X^n\left(\frac{t}{n}\right)$$

⇨ n là kích thước Sample

8. Chứng Minh Công Thức MGF Của Biến Ngẫu Nhiên Của Sample Mean?

$$M_{\bar{X}}(t) = E\left[e^{t\bar{X}}\right] = E\left[e^{\frac{t}{n}X_1}\right]E\left[e^{\frac{t}{n}X_2}\right]E\left[e^{\frac{t}{n}X_3}\right]\dots E\left[e^{\frac{t}{n}X_n}\right] = \left(E\left[e^{\frac{t}{n}X}\right]\right)^n = M_X^n\left(\frac{t}{n}\right)$$

Central Limit Theorem – Định Luật Giới Hạn Trung Tâm:

1. Định Luật Giới Hạn Trung Tâm Nói Về Điều Gì?

⇨ Tổng của vô hạn biến ngẫu nhiên với cùng 1 PDF là 1 biến ngẫu nhiên khác có phân phối chuẩn

2. Chứng Minh Định Luật Giới Hạn Trung Tâm?

⇨ Gọi X_1, X_2, X_3, …, X_n là các biến ngẫu nhiên với cùng 1 PDF có Mean = μ và Variance = σ^2, n = ∞

⇨ Cho biến nhiên sau

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

⇨ Ta có

$$M_Z(t) = E[e^{tZ}] = E\left[e^{t\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}\right] = e^{-\frac{t\mu\sqrt{n}}{\sigma}}E\left[e^{\frac{t\bar{X}\sqrt{n}}{\sigma}}\right] = e^{-\frac{t\mu\sqrt{n}}{\sigma}}M_{\bar{X}}\left(\frac{t\sqrt{n}}{\sigma}\right) =$$
$$e^{-\frac{t\mu\sqrt{n}}{\sigma}}M_X^n\left(\frac{t}{\sigma\sqrt{n}}\right) \Rightarrow$$
$$ln\big(M_Z(t)\big) = -\frac{t\mu\sqrt{n}}{\sigma} + nln\left(M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right) =$$
$$-\frac{t\mu\sqrt{n}}{\sigma} + nln\left(1 + \frac{t}{\sigma\sqrt{n}}E[X] + \frac{t^2}{\sigma^2 n}\frac{E[X^2]}{2!} + \cdots\right) =$$
$$-\frac{t\mu\sqrt{n}}{\sigma} +$$
$$n\left(\left(\frac{t}{\sigma\sqrt{n}}E[X] + \frac{t^2}{\sigma^2 n}\frac{E[X^2]}{2!} + \cdots\right) - \frac{1}{2}\left(\frac{t}{\sigma\sqrt{n}}E[X] + \frac{t^2}{\sigma^2 n}\frac{E[X^2]}{2!} + \cdots\right)^2 + \cdots\right) =$$
$$-\frac{t\mu\sqrt{n}}{\sigma} + n\frac{t}{\sigma\sqrt{n}}E[X] + n\frac{t^2}{\sigma^2 n}\frac{E[X^2]}{2!} - \frac{1}{2}n\left(\frac{t}{\sigma\sqrt{n}}E[X]\right)^2 = \frac{t^2}{2\sigma^2}(E[X^2] - (E[X])^2) =$$
$$\frac{t^2}{2\sigma^2}Var[X] = \frac{t^2}{2} \Rightarrow$$
$$M_Z(t) = e^{\frac{t^2}{2}} \Rightarrow Z \sim N(0,1)$$

⇨ Dễ thấy Z có phân phối chuẩn tắc nên $\bar{X}$ cũng có phân phối chuẩn nên tổng

$X_1 + X_2 + X_3 + \ldots$ có phân phối chuẩn