B6. Probability

Bổ sung thêm cho bài giảng

2019

Nội dung bổ sung

1. Xác suất
2. Một số phân phối xác suất
3. Quy tắc thực nghiệm
4. Định lý giới hạn trung tâm

1. Xác suất

- **Hoán vị** (*permutation*)**: thay đổi, sắp xếp vị trí**
  - Lấy mẫu <u>không lặp lại</u>, hoán vị n *chọn k* (0 < k ≤ n)

    <u>VD</u>: Chọn 5 trong số 11 cầu thủ đá 11m luân lưu (có thứ tự)

$$P_{n,k} = \frac{n!}{(n-k)!}$$

→ nguyên lý nhân k vị trí đầu tiên sẽ chọn: n(n – 1)...(n – k + 1)

Khi k = n (hoán vị toàn bộ): $P_{n,n} = n!$

<u>VD</u>: Dự đoán kết quả giải thưởng FIFA The Best 2019

  - Lấy mẫu <u>lặp lại</u>, hoán vị n *chọn k* (0 < k ≤ n)

$$P^{*}_{n,k} = n^{k}$$

1. Xác suất (tt.)

- **Hoán vị lặp**

S có n phần tử được phân hoạch thành $S_1, S_2, \ldots, S_k$ (2 ≤ k ≤ n):

S_j gồm các phần tử giống nhau: $|S_j| = n_j$

Tổng số cách hoán vị S:

$$p = \binom{n}{n_1, n_2, \ldots, n_k} = \frac{n!}{n_1! n_2! \ldots n_k!}$$

<u>VD</u>: Số lượng chuỗi ký tự khác nhau được tạo ra từ các chữ cái của từ MISSISSIPPI

n = 11, k = 4 ($S_{'I'}$, $S_{'M'}$, $S_{'P'}$, $S_{'S'}$), $n_{'I'} = 4$, $n_{'M'} = 1$, $n_{'P'} = 2$, $n_{'S'} = 4$

$$p = \frac{11!}{4!1!2!4!} = 34650$$

1. Xác suất (tt.)

- **Tổ hợp** (*combination*)
  - *Tổ hợp* n chập k ($0 < k \le n$): KHÔNG (phân biệt) thứ tự

$$C_n^k = \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

$$C_n^k = C_n^{n-k} \qquad C_n^k + C_n^{k-1} = C_{n+1}^k$$

VD: Kiểm tra chất lượng ngẫu nhiên 2 trong số 10 sản phẩm.

Số khả năng có thể xảy ra:

1. Xác suất (tt.)

- **Tổ hợp** (*combination*)
  - *Chỉnh hợp* n chập k ($0 < k \le n$): CÓ (phân biệt) thứ tự

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

$\rightarrow$ hoán vị không lặp lại

1. Xác suất (tt.)

- **Tập hợp con** (*subset*)
  - Số lượng tập hợp con của A có n phần tử: 2^n
    (kể cả tập A và tập ∅)

1. Xác suất (tt.)

- **VD: Số cách chọn 1 lớp trưởng và sau đó 1 lớp phó của 1 lớp có 25 học viên.**

  Hoán vị không lặp: 25.(25 – 1) = 600

- **VD: Một khoa có 20 giảng viên đạt học vị tiến sĩ. Có bao nhiêu cách thành lập Hội đồng khoa học gồm 7 thành viên ?**

  Tổ hợp: $C_{20}^{7} = \frac{20!}{7!(20-7)!} =$

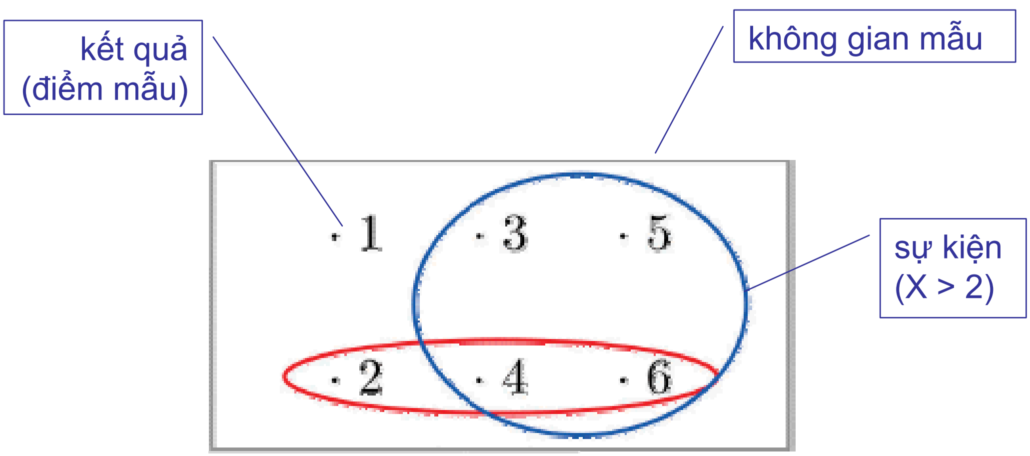
1. Xác suất (tt.)

- **Thí nghiệm (*experiment*): tiến trình (sẽ) diễn ra ngẫu nhiên (không biết trước kết quả) → n <u>lần</u> diễn ra/thực hiện/thử (*trial*)**
  - tung đồng xu 2 lần, số tai nạn máy bay trễ / năm tại sân bay T
- **Kết quả (*outcome*) của <u>1 lần</u> (trial) thí nghiệm được diễn ra**
- **Không gian mẫu (*sample space*): tất cả các kết quả có thể có**
  - tung đồng xu 2 lần: S = { HH (heads), TT (tails), HT, TH }
- **Sự kiện (*event*): tập con của không gian mẫu (một số kết quả)**
  - tung đồng xu 2 lần: kết quả 2 lần không giống nhau (HT, TH)

1. Xác suất (tt.)

- **Một sự kiện E được gọi là "xảy ra" nếu 1 phần tử bất kỳ của E là kết quả của một lần thực hiện thí nghiệm**

"*An event E is said to* **occur** *on <u>a particular trial</u> of the experiment if the outcome observed is an element of the set E.*" [Schmitz]



1. Xác suất (tt.)

- **Không gian mẫu tự nhiên** (*natural sample space*)
  - $|S| = n$, $P(s_i) = 1/n$, $P(E) = |E|/n$
- **Phép đếm trên tập hữu hạn**
  - Nguyên tắc cộng (*addition principle*)

    $S = S_1 \cup S_2 \cup \ldots \cup S_k$, $S_i \cap S_j = \varnothing$: $|S| = |S_1| + |S_2| + \ldots + |S_k|$
  - Nguyên tắc nhân (*multiplication principle*) $\rightarrow$ thí nghiệm k bước

    $S = S_1 \times \ldots \times S_k$: $|S| = |S_1|.|S_2|\ldots|S_k|$
  - Nguyên tắc Dirichlet (*Dirichlet box principle*)

    Nếu có n chim bồ câu ở trong k chuồng thì tồn tại một chuồng có chứa từ $\left\lceil \frac{n}{k} \right\rceil$ bồ câu trở lên.

1. Xác suất (tt.)

- **Các tiên đề**

  Cho không gian mẫu S, các sự kiện rời nhau E, E_1, E_2, …

  $$(i) \quad 0 \le P(E)$$

  $$(ii) \quad P(S) = 1$$

  $$(iii) \quad P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

1. Xác suất (tt.)

- **Một số tính chất cơ bản: E_1, E_2, … rời nhau**

(i) $0 \le P(E) \le 1$

(ii) $P(\varnothing) = 0$

(iii) $A \subseteq B \Rightarrow P(A) \le P(B)$

(iv) $P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i)$

(v) $P(E^C) = 1 - P(E)$

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

(vii) $P(A \cap B) = P(B).P(A \mid B) = P(A).P(B \mid A)$

- **Lưu ý**
  - A, B rời nhau (*disjoint*): A ∩ B = ∅ (không xảy ra đồng thời)
  - A, B độc lập (*independent*): P(A|B) = P(A), P(B|A) = P(B)

1. Xác suất (tt.)

- **Định lý Bayes**

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

$$P(A_i \mid B) = \frac{P(A_i)P(B \mid A_i)}{P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + \ldots + P(A_n)P(B \mid A_n)}$$

1. Xác suất (tt.)

- **<u>VD</u>: Có 5 thanh kim loại có chiều dài lần lượt: 1, 2, 3, 4, 5 (cm). Xác suất bị gẫy tỷ lệ thuận với chiều dài của thanh. Tính xs thanh đầu tiên bị gẫy là thanh có chiều dài không quá 3cm.**

  Gọi s_i là kết quả thanh có chiều dài i bị gẫy đầu tiên ($1 \le i \le 5$).

  Không gian mẫu: $S = \{ s_1, s_2, s_3, s_4, s_5 \}$

  Sự kiện: $E = \{ s_1, s_2, s_3 \}$

  Xác suất thanh i bị gẫy: $p_i = \alpha.i$ (α: hệ số gẫy chưa biết)

  Tổng xác suất: $p_1 + p_2 + p_3 + p_4 + p_5 = 1$

  $15.\alpha = 1 \quad \Rightarrow \alpha = 1/15$

  Xác suất sự kiện E: $P(E) = p_1 + p_2 + p_3 = 6 / 15 = 0.4$

1. Xác suất (tt.)

- **Bài tập: Ex1 – Bài 1**
- **Bài tập: Ex1 – Bài 2**
- **Bài tập: Ex1 – Bài 3**

1. Xác suất (tt.)

- **Biến ngẫu nhiên** (*random variable*)
  - X lấy giá trị số ($\mathbb{R}$) được xác định từ kết quả của 1 thí nghiệm

    $$X : S \rightarrow \mathbb{R}$$

    – tung đồng xu 2 lần, X = số mặt ngửa (heads)
  - phạm vi (*range*) R_X của X: tập hợp miền giá trị của X
    – tung đồng xu 2 lần, X = số mặt ngửa, $R_X = \{ 0, 1, 2 \}$
    – tung đồng xu để có mặt ngửa, X số lần tung, $R_X = \{ 1, 2, \dots \} = N^+$
    – X: thời gian giữa 2 lần nhật thực, $R_X = (0, \infty)$
  - *discrete random variable*: R_X hữu hạn hoặc vô hạn đếm được (*countable*)
  - *continuous random variable*: R_X vô hạn không đếm được

1. Xác suất (tt.)

- **Hàm độ lớn xác suất** (*Probability Mass Function – PMF*), **phân phối xác suất** (*probability distribution*) **của biến X rời rạc**
  - danh sách các xs ứng với từng giá trị $x \in R_X$

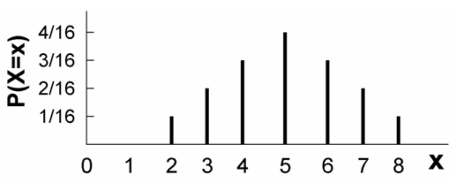
    sự kiện: $E_x = (X = x) = \{ s \in S \mid X(s) = x \}$

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

$$0 \le f(x) \le 1 \qquad \sum_{x \in R_X} f(x) = 1$$

$$A \subseteq R_X : P(X \in A) = \sum f(x)$$

x	f(x)
2	1/16
3	2/16
4	3/16
5	4/16
6	3/16
7	2/16
8	1/16



– tần số

– tần suất

1. Xác suất (tt.)

- **Hàm độ lớn xác suất** (*Probability Mass Function – PMF*), **phân phối xác suất** (*probability distribution*) **của biến X rời rạc**
  - *kỳ vọng* (trên **n lần** thí nghiệm): trung bình có trọng số là các xs

$$E[X] = E(X) = \mu_X = \sum_{x \in R_X} x f(x)$$

E[X] không bắt buộc phải bằng 1 giá trị mà X có thể nhận

  - *phương sai*, *độ lệch chuẩn* (trên **n lần** thí nghiệm)

$$Var(X) = \sigma_X^2 = E[(X-\mu)^2] = \sum_{x \in R_X} (x-\mu)^2 f(x) = \sum_{x \in R_X} x^2 f(x) - \mu^2$$

1. Xác suất (tt.)

- **Hàm độ lớn xác suất** (*Probability Mass Function – PMF*), **phân phối xác suất** (*probability distribution*) **của biến X rời rạc**

$$R_X = \{x_1, x_2, ..., x_m\}$$

$$P(X = x_i) \approx \frac{N_i}{N}$$

$$\text{Average} = \frac{1}{N} \sum_{i=1}^{m} x_i N_i \approx \frac{1}{N} \sum_{i=1}^{m} x_i N P(X = x_i) = E[X]$$

VD: 3 người 170cm, 2 người 165cm

→ TB = (170*3 + 165*2) / 5 = 168cm

1. Xác suất (tt.)

- **Hàm độ lớn xác suất** (*Probability Mass Function – PMF*), **phân phối xác suất** (*probability distribution*) **của biến X <u>rời rạc</u>**

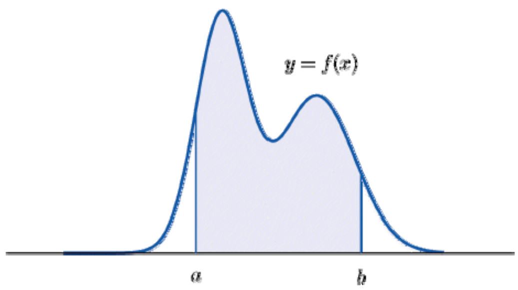
  <u>VD</u>: 10^5 tờ vé số, mỗi tờ giá 10^4.

  Giải thưởng: 1 giải đặc biệt 106, 10^6 giải an ủi 10^5.

  Kỳ vọng của số tiền thu được khi mua 1 tờ vé số ?

1. Xác suất (tt.)

- **Hàm mật độ** (*Probability Density Function – PDF*) **của X <u>liên tục</u>**
  - xs tại 1 giá trị (điểm mẫu) không có ý nghĩa: P(X = x) = 0
  - xs X thuộc 1 khoảng [nửa] đóng/mở
    P(a ≤ X ≤ b): diện tích dưới đường cong giới hạn bởi 2 cận a, b



$$(i) \quad f(x) \geq 0$$

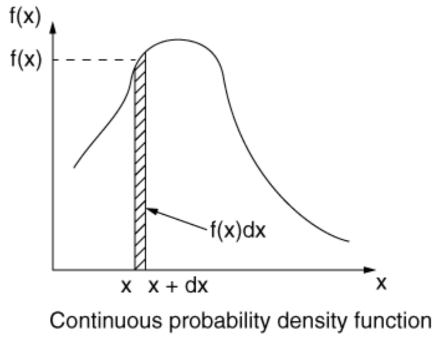
$$(ii) \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

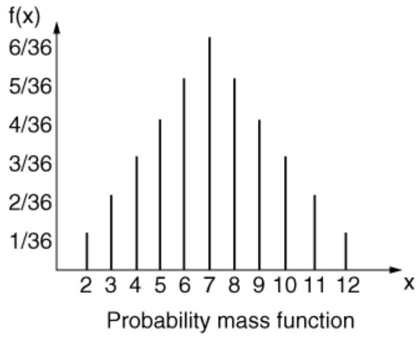
$$(iii) \quad P(X \in [a,b]) = P(X \in (a,b]) = P(X \in [a,b)) = P(X \in (a,b))$$

1. Xác suất (tt.)

❑ **Hàm mật độ** (*Probability Density Function – PDF*) **của X <u>liên tục</u>**

"A probability density function (PDF) describes the probability of the value of a continuous random variable falling within a range."





https://abaqus-docs.mit.edu/2017/English/SIMACAEMODRefMap/simamod-c-probdensityfunc.htm (07/2020)

1. Xác suất (tt.)

❑ **Hàm mật độ** (*Probability Density Function – PDF*) **của X <u>liên tục</u>**

- kỳ vọng, phương sai, độ lệch chuẩn (trên **<u>n lần</u>** thí nghiệm)

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

$$Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{+\infty} x^2 f(x) dx - \mu^2$$

1. Xác suất (tt.)

❑ Một số tính chất của kỳ vọng, phương sai

$(i)\quad E[a] = a$

$(ii)\quad E[aX] = aE[X]$

$(iii)\quad E[X+Y] = E[X] + E[Y]$

$(iv)\quad E[XY] = E[X]E[Y]$ X, Y độc lập

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

$(vi)\quad Var(aX+b) = a^2 Var(X)$

$(vii)\quad X = \sum_{i=1}^{n} X_i \quad \Rightarrow \quad Var(X) = \sum_{i=1}^{n} Var(X_i)$

1. Xác suất (tt.)

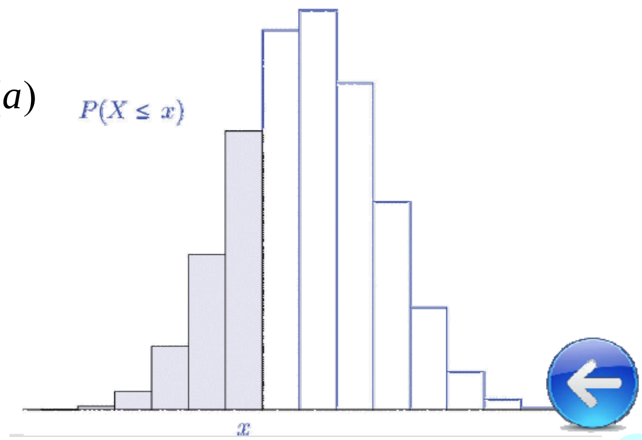
❑ Hàm phân phối tích lũy (*Cumulative Distribution Function – CDF*)

$$F(a) = P(X \le a)$$

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

$(i)\quad P(X < a) = F(a) - f(a)$

$(ii)\quad P(a < X \le b) = F(b) - F(a)$



Nội dung bổ sung

1. Xác suất
2. **Một số phân phối xác suất**
3. Quy tắc thực nghiệm
4. Định lý giới hạn trung tâm

2. Một số phân phối xác suất

- **Mô hình xác suất: biến ngẫu nhiên và phân phối xác suất**
  - mô hình hóa các tiến trình ngẫu nhiên
  - kết quả dự đoán gần với thực tế quan sát
  - cho trước 1 bài toán, cần *xác định phân phối xác suất* (hợp lý) của dữ liệu thu thập được → PMF/PDF, CDF, μ, σ, …

2. Một số phân phối xác suất (tt.)

- Phân phối đều
- Phân phối chuẩn
- Một số phân phối rời rạc
  - nhị thức, Bernoulli, hình học, Poisson, …
- Một số phân phối liên tục
  - lũy thừa (mũ), Gamma, Beta, Chi-bình phương, Student, …

2. Một số phân phối xác suất (tt.)

- **Phân phối đều** (*Uniform Distristribution*) – **Rời rạc**

X ~ *Uniform*(a, b), a < b

$$n = (b - a + 1)$$

$$f(x) = \begin{cases} \frac{1}{n}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} \qquad F(x) = \begin{cases} 0, & x < a \\ \frac{x-a+1}{n}, & x \in [a,b) \\ 1, & x \geq b \end{cases}$$

a = 0: *Rectangular Distristribution*

2. Một số phân phối xác suất (tt.)

- **Phân phối đều** (*Uniform Distristribution*) – **Rời rạc**

$$\text{(i) } \mu = \frac{(a+b)}{2} \qquad \text{(ii) } \sigma^2 = \frac{n^2-1}{12}$$

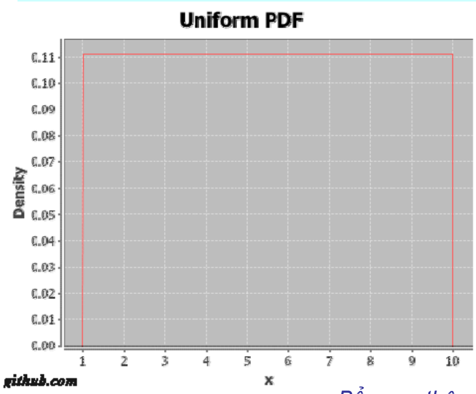
$$\text{(iii) } Skewness = 0 \qquad \text{(iv) } ExcessKurt = -\frac{6(n^2+1)}{5(n^2-1)}$$

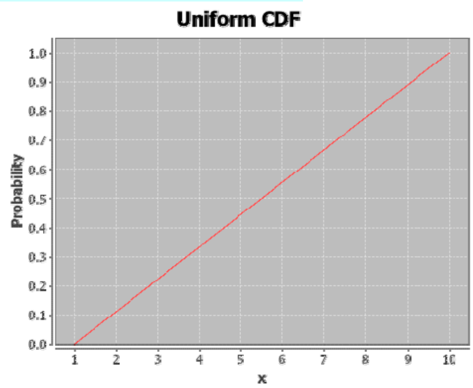
2. Một số phân phối xác suất (tt.)

- **Phân phối đều** (*Uniform Distristribution*) – **Liên tục**

X ~ *Uniform*(a, b)

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a,b] \\ 0, & x \notin [a,b] \end{cases} \qquad F(x) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & x \in [a,b) \\ 1, & x \geq b \end{cases}$$





2. Một số phân phối xác suất (tt.)

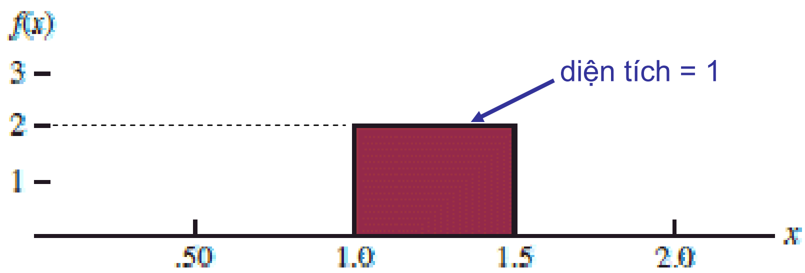
❑ **Phân phối đều** (*Uniform Distristribution*) – **Liên tục**

$$\text{(i) } \mu = \frac{(a+b)}{2} \qquad \text{(ii) } \sigma^2 = \frac{(b-a)^2}{12}$$

$$\text{(iii) } Skewness = 0 \qquad \text{(iv) } ExcessKurt = -\frac{6}{5}$$

2. Một số phân phối xác suất (tt.)

❑ **<u>VD</u>: Cho biến ngẫu nhiên X có phân phối đều**



a. P(X = 1.25) = 0

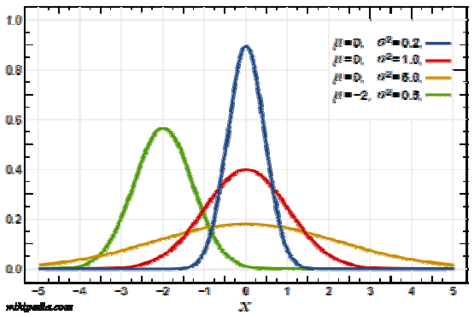
b. P(1.0 ≤ X ≤ 1.25) = 2(1.25 – 1.0) = 0.5

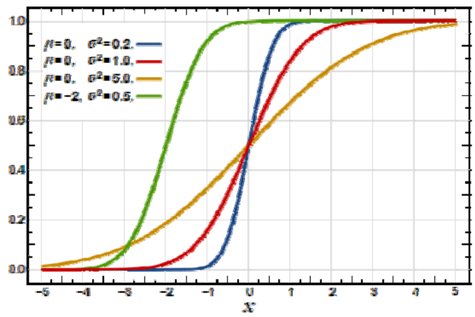
c. P(1.2 < X < 1.5) = 2(1.5 – 1.2) = 0.6

2. Một số phân phối xác suất (tt.)

❑ **Phân phối chuẩn** (*Normal Distribution* / *Gaussian Distribution*)

- phân phối hình chuông (*bell-shaped curve*)
- đặc trưng bởi “tâm” (μ) và “độ rộng” (σ)





2. Một số phân phối xác suất (tt.)

❑ **Phân phối chuẩn** (*Normal Distribution* / *Gaussian Distribution*)

$X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du$$

$$Skewness = ExcessKurt = 0$$

- Tính xs: $P(x \in [a, b]) \rightarrow$ độ phức tạp ?

2. Một số phân phối xác suất (tt.)

- ❑ **Phân phối chuẩn (chuẩn) tắc** (*Standard Normal Distribution* / *Z*)

$$Z = \frac{(X-\mu)}{\sigma} \qquad X = \sigma.Z + \mu$$

Z ~ N(0, 1)

hàm tích phân Laplace

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} \qquad \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{-u^2}{2}} du$$

$$\text{(i) } \mu = 0 \qquad \text{(ii) } \sigma^2 = 1$$

$$\text{(iii) } Skewness = 0 \qquad \text{(iv) } ExcessKurt = 0$$

2. Một số phân phối xác suất (tt.)

- ❑ **Standard normal table, Z table: P(Z < z)**

phần nguyên,
chữ số thập phân thứ 1:
n.d_

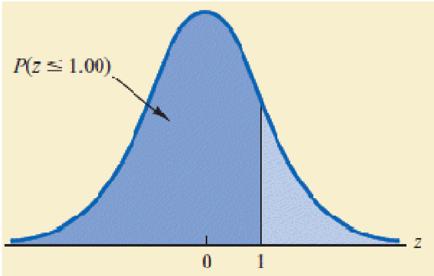
chữ số thập phân thứ 2:
._d

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831

http://www.z-table.com/

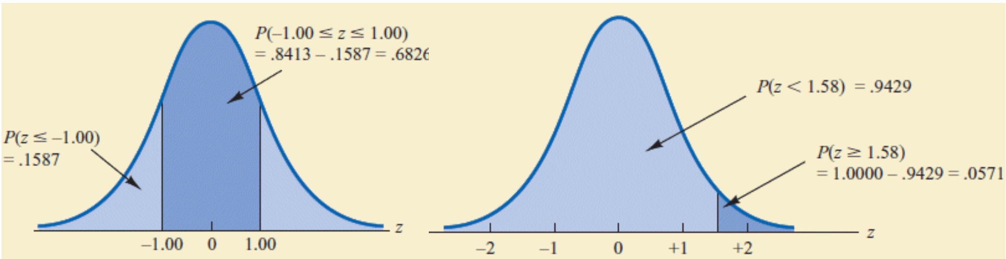
2. Một số phân phối xác suất (tt.)

- **Standard normal table, Z table: P(Z < z)**



z	.00	.01	.02
.			
.			
.			
.9	.8159	.8186	.8212
1.0	.8413	.8438	.8461
1.1	.8643	.8665	.8686
1.2	.8849	.8869	.8888
.			
.			
.			

P(z ≤ 1.00)



[*Anderson+*]

2. Một số phân phối xác suất (tt.)

- **Tính xác suất theo phân phối chuẩn**

B1. Mô hình hóa P(X < x)

B2. Chuyển về phân phối Z

B3. Tra bảng Z

- VD: X ~ N(μ = 16, σ = 4) → P(X < 8) ?

$X = 4Z + 16 < 8 \Rightarrow Z < -2$

$P(Z < -2) = 0.0228$

2. Một số phân phối xác suất (tt.)

- **VD: Tính xác suất của biến ngẫu nhiên có phân phối z**

  a. $P(z \leq 1.2) = 0.8849$

  b. $P(z \leq -0.71) = 0.2389$

  c. $P(0 \leq z \leq 0.83) = 0.2967$

  d. $P(-1.57 \leq z \leq 0) = 0.4418$

  e. $P(0.44 < z) = 0.3300$

  e. $P(-0.23 \leq z) = 0.5910$

2. Một số phân phối xác suất (tt.)

- **Tìm ngưỡng x tương ứng với xs đã biết**

  B1. Mô hình hóa $P(X < x)$

  B2. Tra bảng Z

  B3. Chuyển từ Z về $X = \sigma Z + \mu$

- VD: $X \sim N(\mu = 16, \sigma = 4)$, $P(Z > z) = 0.9834 \rightarrow x = ?$

  $P(Z < z) = 1 - (Z > z) = 0.0166$

  $z = -2.13$

  $x = 4.(-2.13) + 16$

2. Một số phân phối xác suất (tt.)

- **VD: Xác định giá trị của z khi biết:**

  a. Diện tích bên trái của z là 0.2119 → Python

  b. Diện tích bên trái của z là 0.9948

  c. Diện tích ở giữa –z và z là 0.9030

  d. Diện tích ở giữa –z và z là 0.2052

  e. Diện tích bên phải của z là 0.6915

2. Một số phân phối xác suất (tt.)

- **Phân phối nhị thức** (*Binominal Distristribution*)
  - tiến trình Bernoulli (*Bernoulli trial*) → { thành công, thất bại }
  - thí nghiệm: n (lần) Bernoulli trial(s) ĐỘC LẬP
  - xs để 1 Bernoulli trial thành công (p), hay thất bại q = (1 – p), giống nhau trong thí nghiệm
  - biến ngẫu nhiên X: *số lần thành công* ($0 \leq X \leq n$)

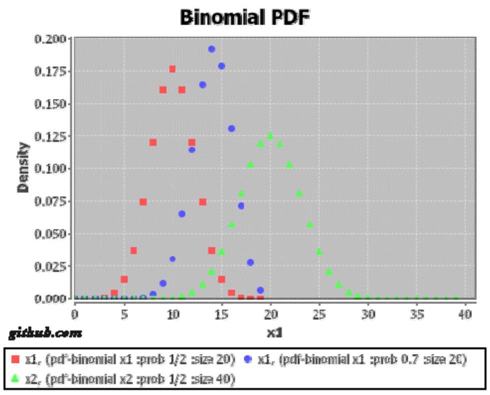
2. Một số phân phối xác suất (tt.)

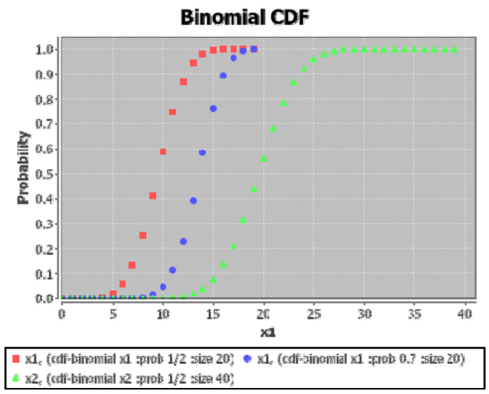
❑ **Phân phối nhị thức** (*Binominal Distristribution*)

X ~ *Binomial*(n, p)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad F(x) = \sum_{X \le x} \binom{n}{x} p^x (1-p)^{n-x}$$

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$





2. Một số phân phối xác suất (tt.)

❑ **Phân phối nhị thức** (*Binominal Distristribution*)

$$\text{(i) } \mu = np \qquad \text{(ii) } \sigma^2 = np(1-p)$$

$$\text{(iii) } Skewness = \frac{1-2p}{\sqrt{np(1-p)}} \qquad \text{(iv) } ExcessKurt = \frac{1-6p(1-p)}{np(1-p)}$$

- Có thể sử dụng phân phối chuẩn để xấp xỉ phân phối nhị thức với n đủ lớn và p gần 0.5

2. Một số phân phối xác suất (tt.)

- **VD: Xét thí nghiệm gồm 2 lần phép thử Bernoulli có p = 0.4**

  a. Xác suất 1 lần thành công:

$$f(1) = \binom{2}{1} 0.4^1 (1 - 0.4)^{2-1} = 0.48$$

  b. Xác suất không có lần nào thành công: f(0) = 0.36

  c. Xác suất tối thiểu 1 lần thành công: $P(1 \leq X) = f(1) + f(2) = 0.64$

  d. Tính kỳ vọng, phương sai, độ lệch chuẩn

  E[X] = n.p = 0.8

  Var(X) = n.p.(1 – p) = 0.48

2. Một số phân phối xác suất (tt.)

- **VD: Xét thí nghiệm gồm 10 lần phép thử Bernoulli có p = 0.1**

  a. Xác suất tối thiểu 1 lần thành công:

  b. Xác suất tối đa 2 lần thành công:

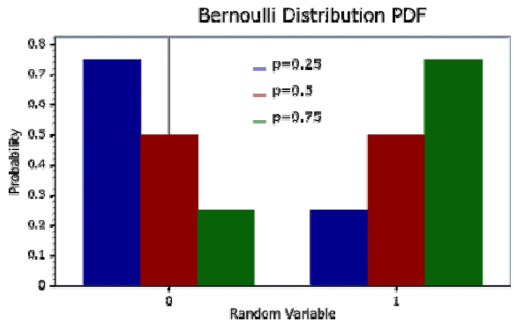
  c. Tính kỳ vọng, phương sai, độ lệch chuẩn:

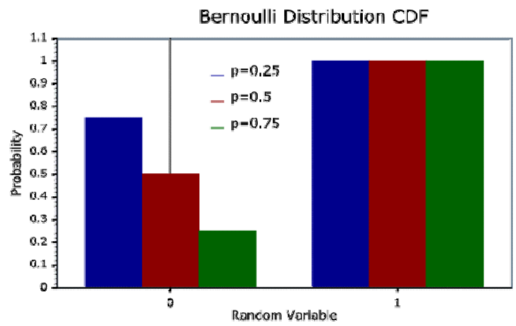
2. Một số phân phối xác suất (tt.)

❑ **Phân phối Bernoulli: phân phối nhị thức với n = 1**

X ~ *Bernoulli*(p)

$$f(x) = \begin{cases} p, & x = 1 \\ (1-p) & x = 0 \end{cases} \qquad F(x) = \begin{cases} 0, & x < 0 \\ (1-p), & 0 \le x < 1 \\ 1, & 1 < x \end{cases}$$





2. Một số phân phối xác suất (tt.)

❑ **Phân phối Poisson**

- số lần 1 sự kiện xảy ra trong một khoảng THỜI GIAN cố định
  - số lượng truy cập trang Web, cuộc gọi cần tư vấn trong 1 giờ
  - số tai nạn tại 1 giao lộ trong 1 ngày
  - …
- số lần 1 sự kiện xảy ra trong một vùng KHÔNG GIAN cố định
  - số lần gõ sai 2 từ trong một trang
  - số vết nứt trên 100m đường ống
  - …

2. Một số phân phối xác suất (tt.)

❑ Phân phối Poisson

- X: *số lần sự kiện xảy* ra trong 1 khoảng thời gian/không gian

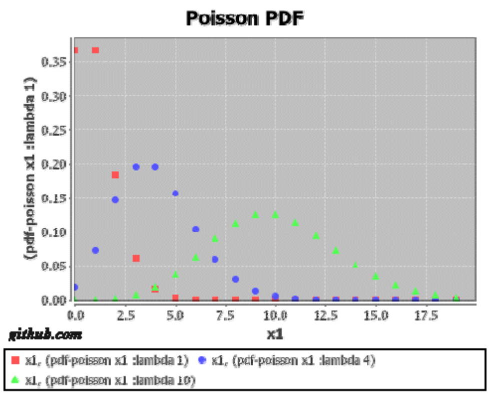
  X $\in$ N, *rời rạc*, ~ vô hạn đếm được (hữu hạn → bao nhiêu ?)
  (phân phối nhị thức: X ≤ n lần thí nghiệm cố định)
- các sự kiện độc lập với nhau
- không có 2 sự kiện cùng xảy ra tại 1 thời điểm (hay tại 1 điểm)
- xs như nhau trong những khoảng thời gian/không gian = nhau

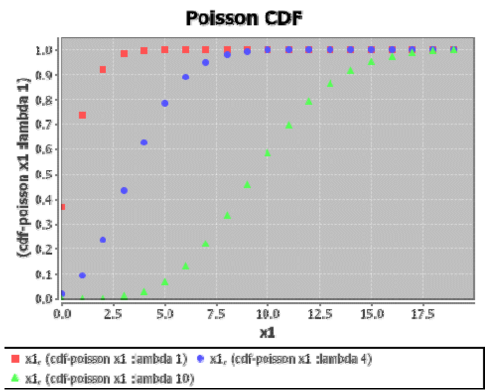
2. Một số phân phối xác suất (tt.)

❑ Phân phối Poisson

X ~ *Poisson*(λ)   *λ: số lần xảy ra trong 1 khoảng cho trước*

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \qquad F(x) = e^{-\lambda}\sum_{X \le x}\frac{\lambda^x}{x!}$$





2. Một số phân phối xác suất (tt.)

❑ **Phân phối Poisson**

$$\text{(i) } \mu = \lambda \qquad \text{(ii) } \sigma^2 = \lambda$$

$$\text{(iii) } Skewness = \lambda^{-1/2} \qquad \text{(iv) } ExcessKurt = \lambda^{-1}$$

2. Một số phân phối xác suất (tt.)

❑ **Phân phối Poisson**

VD: Quan sát lưu lượng xe tại 1 giao lộ vào giờ cao điểm.

- sự độc lập trong giao thông
- xs có xe là như nhau trong 2 khoảng thời gian dài bằng nhau
- thông tin quá khứ: 10 xe / 15 phút → X: số xe trong 15 phút

Xác suất có 5 xe trong 15 phút: $P(X = 5) = f(5) = e^{-10}.\dfrac{10^5}{5!} = 0.0378$

Xác suất có 1 xe trong 3 phút = 0.0378 ?

2. Một số phân phối xác suất (tt.)

❑ Phân phối Poisson

VD: Quan sát hư hại (lớn) trên mặt đường cao tốc.

- sự độc lập trong hư hại
- xs hư hại là như nhau trên 2 quảng đường dài bằng nhau
- thông tin quá khứ: 2 hư hại / km → X: số hư hại / km

Xác suất KHÔNG có hư hại trên 3km:

2 hư hại / km ⇒ kỳ vọng 6 hư hại / 3km ⇒ $\lambda_{3km} = 6$

$$P(X=0) = f(0) = e^{-6}.\frac{6^0}{0!} = 0.0025$$

Xác suất có tối thiểu 1 hư hại / 3km rất lớn: 1 – 0.0025 = 0.9975

2. Một số phân phối xác suất (tt.)

❑ Phân phối Poisson

Giả sử một phân phối Poisson có giá trị trung bình $\mu = 50 = \lambda$.
Tính P(X = 45).

$$P(X=45) = e^{-50}\frac{50^{45}}{45!}$$

→ xấp xỉ phân phối chuẩn: với λ đủ lớn

2. Một số phân phối xác suất (tt.)

- **Xấp xỉ phân phối Poisson bằng phân phối chuẩn: $\lambda \geq 20$**

  Left / Right / 2 sides

  B1. Đưa bài toán về dạng $P(X \leq b)$, $P(X \geq b)$, $P(a \leq X \leq b)$

  B2. Chuyển các cận a, b sang Z

  B3. Tính xs $P(Z \leq z)$ theo phân phối chuẩn tắc

  B4. Xét các trường hợp:

  Bài toán tính xs NHỎ hơn: NOP

  Bài toán tính xs LỚN hơn: kết quả = 1 – giá trị trong bảng tra

  Bài toán tính xs trong khoảng: thực hiện B1-B3 cho cận trên; sau đó trừ 2 kết quả nhận được

  B5. Chuyển từ Z trở về $X = \sigma Z + \mu$ (nếu cần thiết)

2. Một số phân phối xác suất (tt.)

- **<u>VD</u>: Một ngân hàng có trung bình 5 khách hàng trong 10 phút. Tính xs có nhiều hơn 35 khách hàng / giờ với λ = 30 / giờ.**

  B1. $p = P(X > 35) = ?$

  B2. $Z = (X - \mu) / \sigma = (35 - 30) / (30)^{1/2} = 0.91$; $p \approx P(Z > 0.91)$

  B3. Tra bảng phân phối Z ta được $P(Z < 0.91) = 0.8186$

  B4. $P(Z > 0.91) = 1 - P(Z < 0.91) = 1 - 0.8186 = 0.1814$

2. Một số phân phối xác suất (tt.)

- **Phân phối Poisson**
  - Có thể dùng phân phối nhị thức để xấp xỉ phân phối Poisson với n đủ lớn và p đủ nhỏ

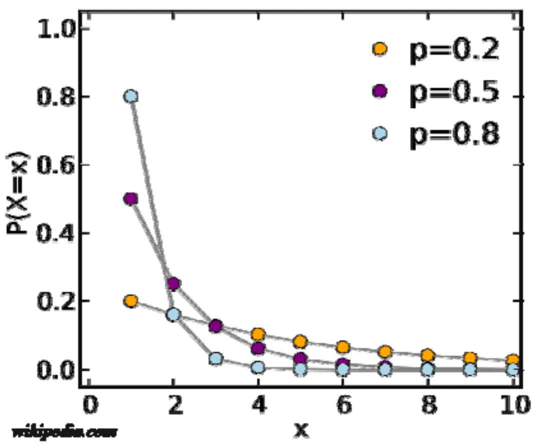
$$X \sim Binomial\left(n, p = \frac{\lambda}{n}\right) \Rightarrow \lim_{n \to \infty} P(x) = e^{-\lambda} . \frac{\lambda^x}{x!}$$

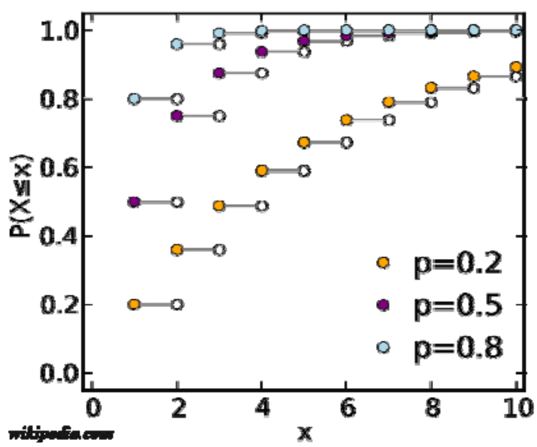
2. Một số phân phối xác suất (tt.)

- **Phân phối hình học (*Geometric Distristribution*): thực hiện X lần tiến trình Bernoulli cho đến khi xuất hiện sự kiện mong muốn**

X ~ *Geometric*(p), với 0 < p < 1

$$f(x) = p(1-p)^{x-1} \qquad F(x) = 1-(1-p)^x$$





2. Một số phân phối xác suất (tt.)

- **Phân phối hình học** (*Geometric Distristribution*)**: thực hiện X lần tiến trình Bernoulli cho đến khi xuất hiện sự kiện mong muốn**

$$\text{(i) } \mu = \frac{1}{p} \qquad \text{(ii) } \sigma^2 = \frac{(1-p)}{p^2}$$

$$\text{(iii) } Skewness = \frac{(2-p)}{\sqrt{(1-p)}} \qquad \text{(iv) } ExcessKurt = 6 + \frac{p^2}{(1-p)}$$

Nhận xét: X là số lần thực hiện thí nghiệm

- bắt đầu là (X – 1) lần thất bại, với xs thất bại là (1 – p)
- kế tiếp là lần thứ X thành công, với xs thành công là p

$$P(X = x) = (1-p)^{x-1}.p$$

2. Một số phân phối xác suất (tt.)

- **Phân phối hình học** (*Geometric Distristribution*)**: thực hiện X lần tiến trình Bernoulli cho đến khi xuất hiện sự kiện mong muốn**

VD: Trong 3 lần ném rổ, xs thành công là 0.7.

Xs ném rổ thành công ở lần thứ 2: P(X = 2) = p(1 – p) = 0.21

Xs ném rổ thành công ở lần thứ 3: P(X = 3) = p(1 – p)^2 = 0.063

⇒ Càng về sau, xs thành công càng giảm dần về 0

2. Một số phân phối xác suất (tt.)

- **So sánh 3 phân phối rời rạc: nhị thức, hình học và Poisson**
  - nhị thức: số lần thành công trong n (cố định) lần thực hiện
  - hình học: số lần thực hiện cho đến khi thành công
  - Poisson: số lần xảy ra trong 1 thời gian/không gian cố định

2. Một số phân phối xác suất (tt.)

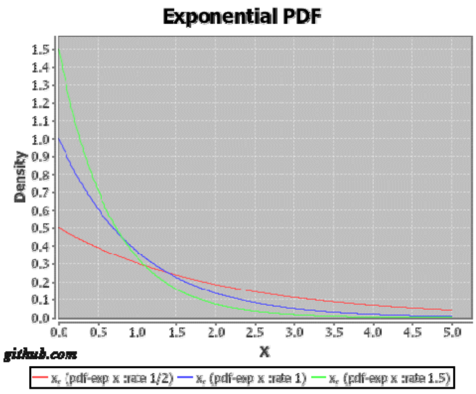
- **Phân phối lũy thừa/mũ** (*Exponential Distribution*)
  - thể hiện thời gian đối với các sự kiện
    - thời gian giữa các thời điểm trong quy trình Poisson
    - thời gian giữa các cuộc gọi cần tư vấn

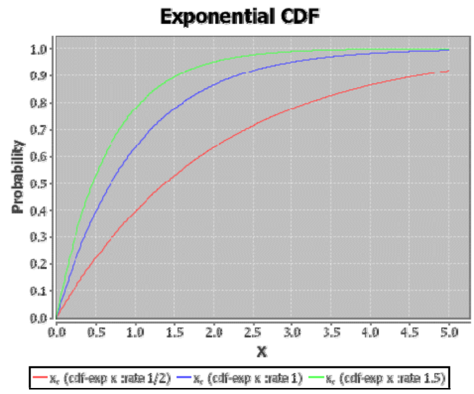
2. Một số phân phối xác suất (tt.)

❑ Phân phối lũy thừa/mũ (*Exponential Distribution*)

$\boxed{X \sim Exponential(\lambda)}$

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases} \qquad F(x) = \begin{cases} 1 - e^{-\lambda x}, & x > 0 \\ 0, & x \le 0 \end{cases}$$





2. Một số phân phối xác suất (tt.)

❑ Phân phối lũy thừa/mũ (*Exponential Distribution*)

$$\text{(i) } \mu = \frac{1}{\lambda} \qquad \text{(ii) } \sigma^2 = \frac{1}{\lambda^2}$$

$$\text{(iii) } Skewness = 2 \qquad \text{(iv) } ExcessKurt = 6$$

2. Một số phân phối xác suất (tt.)

❑ **Phân phối lũy thừa/mũ** (*Exponential Distribution*)

VD: Thời gian chờ đợi (xếp hàng) trung bình là 1 giờ đồng hồ.

Xác suất chờ đợi tối đa 15 phút:

Quy đổi về đơn vị giờ: 15 phút $\to$ 0.25

Thời gian chờ đợi trung bình 1 giờ $\Rightarrow \mu = 1 \Rightarrow \lambda = \mu^{-1} = 1$

$\Rightarrow P(X \leq 0.25) = (1 - e^{-0.25}) = 0.2211$

2. Một số phân phối xác suất (tt.)

❑ **Phân phối Pareto**

- xuất phát: xã hội học (dân số, thu nhập, …)
- mở rộng: thời gian sống $\to$ bệnh tật, hư hỏng, risk, …
- survival analysis: khoảng thời gian cho đến khi xảy ra sự kiện được quan tâm $\to$ thời gian tồn tại $\geq$ warranty period t
- quy luật 80-20: 20% nguyên nhân (I) $\to$ 80% kết quả

2. Một số phân phối xác suất (tt.)

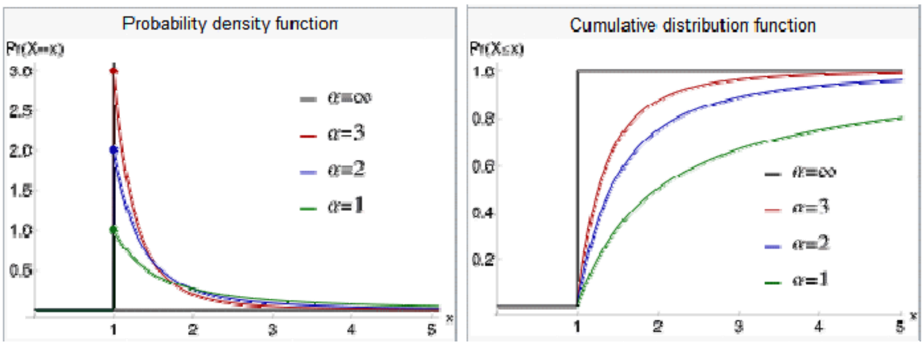
❑ Phân phối Pareto

X ~ *Pareto*(k, α)

k: giá trị chặn dưới

α: shape/slope parameter, tail/Pareto index

$$f(x) = \begin{cases} \dfrac{k^\alpha \alpha}{x^{(\alpha+1)}}, & x \ge k \\ 0, & x < k \end{cases} \qquad F(x) = \begin{cases} 1 - \left(\dfrac{k}{x}\right)^\alpha, & x \ge k \\ 0, & x < k \end{cases}$$



www.wikipedia.com

2. Một số phân phối xác suất (tt.)

❑ Phân phối Pareto

$$\text{(i) } \mu = \begin{cases} \infty, & \alpha \le 1 \\ \dfrac{k\alpha}{\alpha - 1}, & \alpha > 1 \end{cases} \qquad \text{(ii) } \sigma^2 = \begin{cases} \infty, & \alpha \le 2 \\ \dfrac{k^2\alpha}{(\alpha-1)^2(\alpha-2)}, & \alpha > 2 \end{cases}$$

$$\text{(iii) } Skewness = \frac{2(1+\alpha)}{\alpha - 3}\sqrt{\frac{\alpha-2}{\alpha}}, \alpha > 3$$

$$\text{(iv) } ExcessKurt = \frac{6(\alpha^3 + \alpha^2 - 6\alpha - 2)}{\alpha(\alpha-3)(\alpha-4)}, \alpha > 4$$

Nội dung bổ sung

1. Đạo hàm
2. Một số phân phối
3. Quy tắc thực nghiệm
4. Định lý giới hạn trung tâm

3. Quy tắc thực nghiệm

- Bất đẳng thức Markov cho X không âm

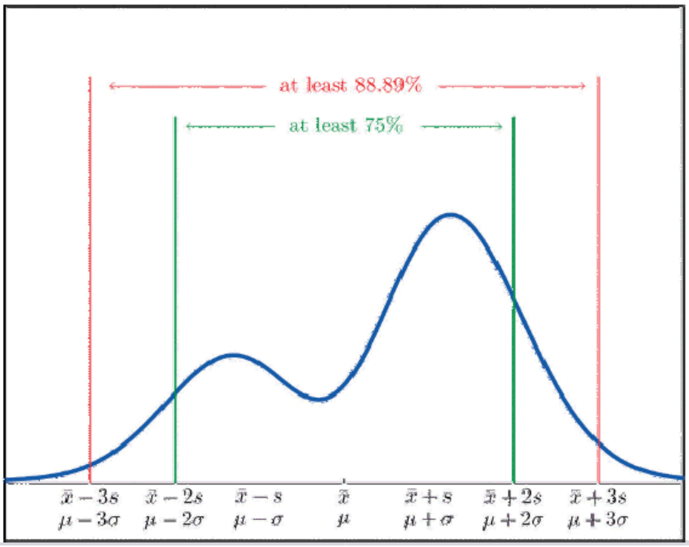
$$P(a \le X) \le \frac{\mu}{a}$$

- Bất đẳng thức Chebyshev

$$P(z\sigma \le (|X - \mu|) \le \frac{1}{z^2}$$

3. Quy tắc thực nghiệm (tt)

- **Định lý Chebyshev**
  - tối thiểu $\left(1-\frac{1}{z^2}\right)$ quan sát nằm trong [μ – zσ, μ + zσ], với k > 1



Nội dung bổ sung

1. Đạo hàm
2. Một số phân phối
3. Quy tắc thực nghiệm
4. **Định lý giới hạn trung tâm**

4. Định lý giới hạn trung tâm

- **Luật số lớn** (*Law of Large Numbers – LLN*)
  - thực hiện thí nghiệm n (rất nhiều) lần: trị trung bình ≈ kỳ vọng
- **Trung bình mẫu** (*sample mean*) **các biến độc lập, cùng ph. phối**

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

- **Luật số lớn YẾU** (*Weak Law of Large Numbers – WLLN*)

$$E(X_i) = \mu < \infty, \qquad \forall \varepsilon > 0, \qquad \lim_{n\to\infty} P(|\overline{X} - \mu| \geq \varepsilon) = 0$$

- **Luật số lớn MẠNH** (*Strong Law of Large Numbers – SLLN*)

$$E(X_i) = \mu < \infty, \qquad P(\lim_{n\to\infty} \overline{X} = \mu) = 1$$

4. Định lý giới hạn trung tâm (tt.)

- **Central Limit Theorem** (*CLT*)

  X_i: thời gian phục vụ khách i, với $E(X_i) = 2$ min, $Var(X_i) = 1$

  Xs phục vụ 50 người trong vòng 90 đến 110 phút là bao nhiêu ?

  - trong thực tế, một biến ngẫu nhiên X có thể được biểu diễn bằng tổng của một số lượng lớn các biến ngẫu nhiên X_1, X_2, …
    $\rightarrow$ X xấp xỉ phân phối chuẩn

$$E(X_i) = \mu < \infty, \qquad 0 < Var(X_i) = \sigma^2 < \infty, \qquad Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

$$\Rightarrow \qquad \lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

$\Phi(x)$: *standard normal* CDF

4. Định lý giới hạn trung tâm (tt.)

- **Central Limit Theorem** (*CLT*)

  X_i ~ *Bernoulli*(p), $E(X_i) = p$, $Var(X_i) = p(1 – p)$

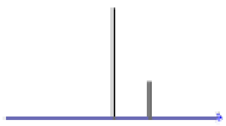
  Ta có: $Y = X_1 + X_2 + \ldots + X_n$ ~ *Binomial*(n, p)

$$Z_n = \frac{Y - np}{\sqrt{np(1-p)}}$$

$$\boxed{p = \tfrac{1}{3}}$$

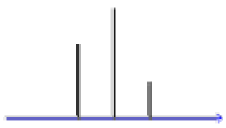
$$Z_1 = \frac{X_1 - p}{\sqrt{p(1-p)}}$$

PMF of Z_1



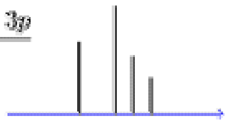
$$Z_2 = \frac{X_1 + X_2 - 2p}{\sqrt{2p(1-p)}}$$

PMF of Z_2



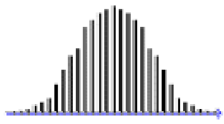
$$Z_3 = \frac{X_1 + X_2 + X_3 - 3p}{\sqrt{3p(1-p)}}$$

PMF of Z_3



$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - 30p}{\sqrt{30p(1-p)}}$$

PMF of Z_{30}



4. Định lý giới hạn trung tâm (tt.)

- **Central Limit Theorem** (*CLT*)

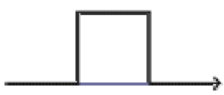
  X_i ~ *Uniform*(0, 1), $E(X_i) = 1/2$, $Var(X_i) = 1/12$

  Ta có: $Y = X_1 + X_2 + \ldots + X_n$ ~ Binomial(n, p)

$$Z_n = \frac{Y - \frac{n}{2}}{\sqrt{n/12}}$$

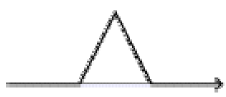
$$Z_1 = \frac{X_1 - \frac{1}{2}}{\sqrt{\frac{1}{12}}}$$

PDF of Z_1



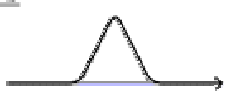
$$Z_2 = \frac{X_1 + X_2 - 1}{\sqrt{\frac{2}{12}}}$$

PDF of Z_2



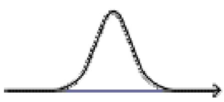
$$Z_3 = \frac{X_1 + X_2 + X_3 - \frac{2}{3}}{\sqrt{\frac{3}{12}}}$$

PDF of Z_3



$$Z_{30} = \frac{\sum_{i=1}^{30} X_i - \frac{30}{2}}{\sqrt{\frac{30}{12}}}$$

PDF of Z_{30}



4. Định lý giới hạn trung tâm (tt.)

- **Central Limit Theorem** (*CLT*)
  - áp dụng trong nhiều lãnh vực
  - đơn giản hóa quá trình tính toán: 1 biến ngẫu nhiên thay cho (SUM) rất nhiều biến ngẫu nhiên X_i khác (chỉ cần μ và σ của X_i)
  - ngưỡng giá trị của n phụ thuộc vào phân phối của X_i ($n \geq 30$)

4. Định lý giới hạn trung tâm (tt.)

- **Quy trình áp dụng CLT**

B1. Đặt: $Y = X_1 + X_2 + \ldots + X_n$, $E(X_i) = \mu$, $Var(X_i) = \sigma^2$

B2. Tính: $E(Y) = n\mu$, $Var(Y) = n\sigma^2$

B3. Tính xs:

$$P(y_1 \leq Y \leq y_2) = P\left(\frac{y_1 - n\mu}{\sigma\sqrt{n}} \leq \frac{Y - n\mu}{\sigma\sqrt{n}} \leq \frac{y_2 - n\mu}{\sigma\sqrt{n}}\right) \approx$$

$$\approx \Phi\left(\frac{y_2 - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{y_1 - n\mu}{\sigma\sqrt{n}}\right)$$

4. Định lý giới hạn trung tâm (tt.)

- **Quy trình áp dụng CLT**

  X_i: thời gian phục vụ khách i, với $E(X_i) = 2$ min, $Var(X_i) = 1$

  Xs phục vụ 50 người trong vòng 90 đến 110 phút là bao nhiêu ?

  B1. $Y = X_1 + X_2 + \ldots + X_n$, $n = 50$,

  B2. $E(Y) = 2 * 50 = 100$, $Var(Y) = 1 * 50 = 50$

  B3. Tính xs:

$$
\begin{aligned}
P(90 \le Y \le 110) &= P\left(\frac{90-100}{\sqrt{50}} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \frac{110-100}{\sqrt{50}}\right) = \\
&= P\left(-\sqrt{2} < \frac{Y - n\mu}{\sigma\sqrt{n}} < \sqrt{2}\right) \approx \\
&\approx \Phi\left(\sqrt{2}\right) - \Phi\left(-\sqrt{2}\right) = 0.8427
\end{aligned}
$$

Tài liệu tham khảo

Anderson et al., *Statistics for Business and Economics*, Cengage, 2016.

Nguyễn Đình Thúc và các tác giả, *Thống kê máy tính*, NXB Khoa học và kỹ thuật, 2010.

Pishro-Nik H., *Introduction to Probability, Statistics, and Random Processes*, Kappa Research LLC, 2014.

Schmitz Andy, *Introductory Statistics*, Saylor Academy, (https://saylordotorg.github.io/text_introductory-statistics/index.html, 09/2019).