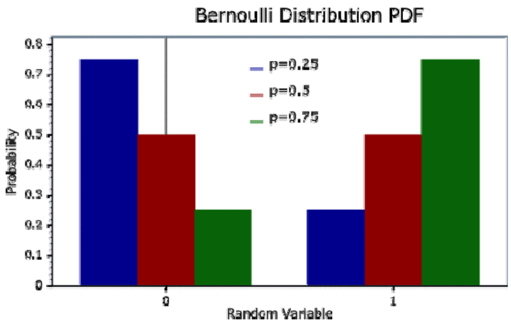
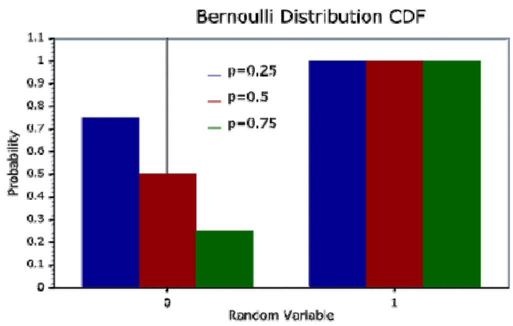
2. Một số phân phối xác suất (tt.)

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

X ~ *Bernoulli*(p)

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





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

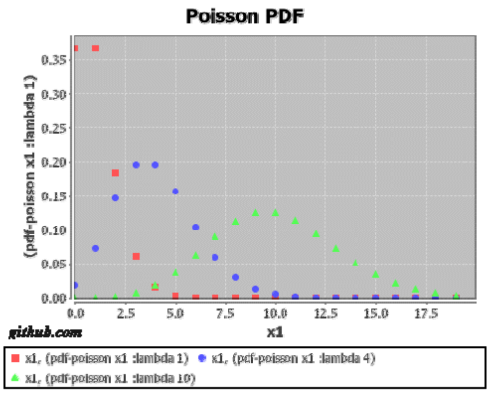
  X $\in$ N, rời rạc, ~ vô hạn đếm được (hữu hạn $\rightarrow$ bao nhiêu?)
  (phân phối nhị thức: X $\leq$ n lần thí nghiệm cố định)
- các sự kiện độc lập với nhau
- không có 2 sự kiện cùng xảy ra tại 1 thời điểm (hay tại 1 điểm)
- xs đều giống nhau trong những khoảng thời gian/không gian bằng nhau

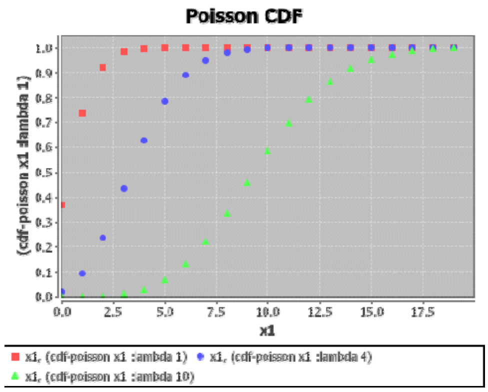
2. Một số phân phối xác suất (tt.)

❑ Phân phối Poisson

X ~ *Poisson*(λ)

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





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

- 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: $P(X = 5) = f(5) = \dfrac{10^5 e^{-10}}{5!} = 0.0378$

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 ⇒ $\mu_{3km} = 6$

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

  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) rất phức tạp → xấp xỉ phân phối chuẩn:

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

2. Một số phân phối xác suất (tt.)

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

  B1. Đưa bài toán về dạng $P(X < b)$, $P(X > b)$, $P(a < X < b)$

  B2. Chuyển các cận a, b sang Z

  B3. Tính xs nhỏ hơn hay bằng theo bảng phân phối chuẩn (Z)

  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 $X = \sigma Z + \mu$

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 nếu biết 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 $P(Z < 0.91) = 0.8186$

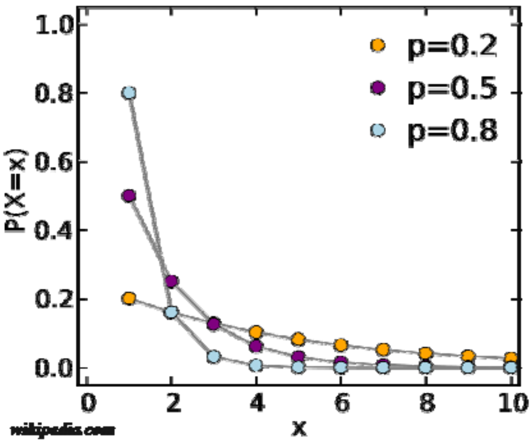
  B4. $P(Z > 0.91) = 1 – P(Z < 0.91) = 1 – 0.8186 = 0.1814$

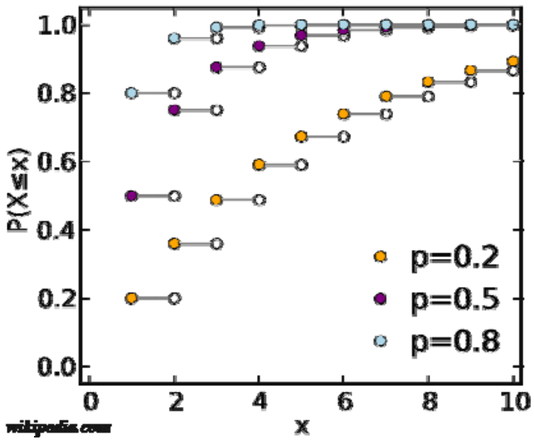
2. Một số phân phối xác suất (tt.)

❑ **Phân phối hình học** (*geometric distristribution*)**: X lần tiến trình Bernoulli cho đến khi <u>xuất hiện sự kiện mong muốn</u>**

X ~ *Geometric*(p), với 0 < p < 1

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





2. Một số phân phối xác suất (tt.)

❑ **Phân phối hình học** (*geometric distristribution*)**: n 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*)**: n 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: $P(X = 3) = p(1 – p)^2 = 0.063$

  $\Rightarrow$ 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.)

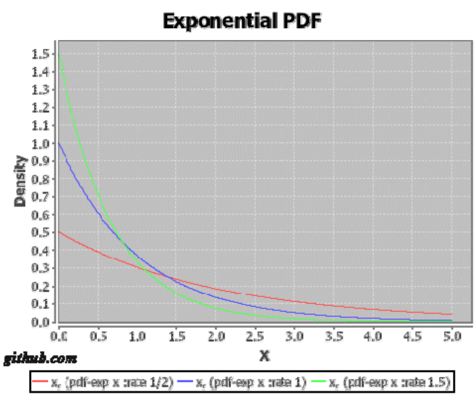
- **Phân phối lũy thừa/mũ** (*exopential distribution*)
  - thể hiện thời gian đối với các sự kiện
    - thời gian giữa các thời điểm trong quy trình Poisson
    - thời gian giữa các cuộc gọi cần tư vấn

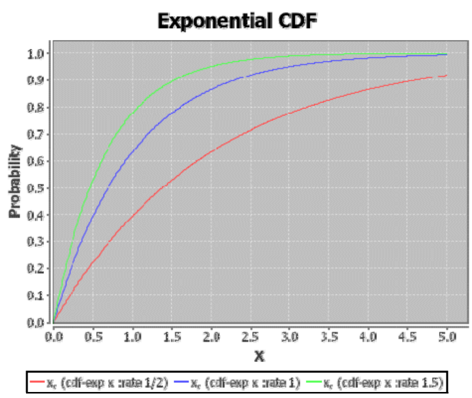
2. Một số phân phối xác suất (tt.)

- **Phân phối lũy thừa/mũ** (*exopential distribution*)

X ~ *Exponential*(λ)

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





2. Một số phân phối xác suất (tt.)

- **Phân phối lũy thừa/mũ** (*exopential 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** (*exopential distribution*)

<u>VD</u>: Thời gian chờ đợi trung bình là 1 giờ đồng hồ.

Xác suất chờ đợi tối đa 15 phút:

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

Thời gian chờ đợi trung bình 1 giờ $\Rightarrow \mu = 1 \Rightarrow \lambda = 1 / \mu = 1$

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

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

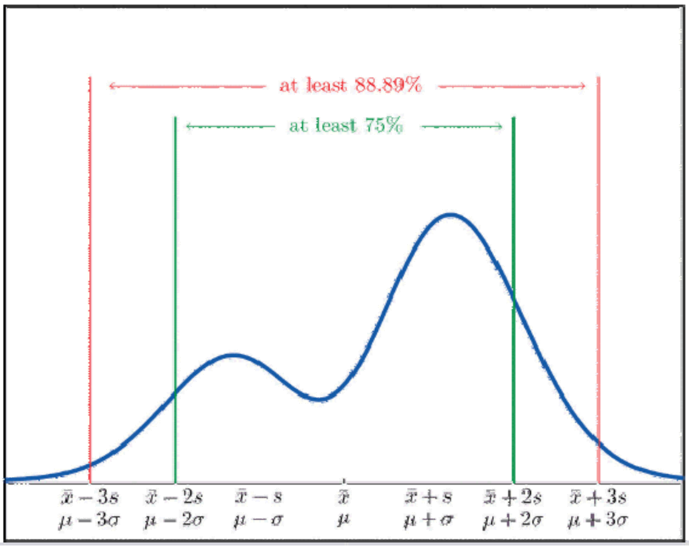
$$P(a \le X) \le \frac{\mu}{a}$$

- 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 [$\mu - z\sigma$, $\mu + z\sigma$], 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 rất nhiều lần: trị trung bình xấp xỉ 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*)
  - 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, …
    → X xấp xỉ phân phối chuẩn

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

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

Φ(x): *standard normal* CDF

4. Định lý giới hạn trung tâm (tt.)

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

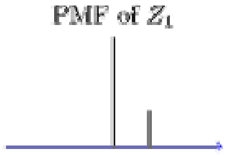
  $X_i \sim$ *Bernoulli*(p), $E(X_i) = p$, $Var(X_i) = p(1 - p)$

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

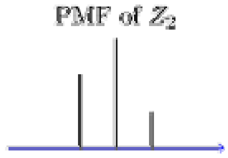
$$Z_n = \frac{Y - np}{\sqrt{np(1-p)}}$$

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

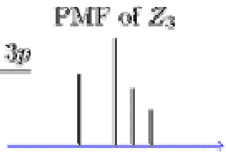
$$Z_1 = \frac{X_1 - p}{\sqrt{p(1-p)}}$$



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



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



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

4. Định lý giới hạn trung tâm (tt.)

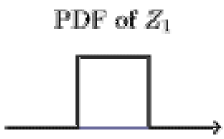
- **Central Limit Theorem** (*CLT*)

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

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

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

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



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

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

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

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:

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

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).