Xử lý số liệu thống kê

Assignment 1

Nhóm E:

Trần Tiến Đạt
Nguyễn Thị Ngọc Anh
Nguyễn Thái Hưng Thịnh

Ngày nộp: October 9, 2025

Contents

1 Phân phối Nhị thức (Binomial Distribution) 3
- 1.1 Định nghĩa . . . 3
- 1.2 Probability Mass Function - PMF . . . 3
- 1.3 Cumulative Distribution Function - CDF . . . 3
- 1.4 Các đặc trưng thống kê . . . 4
- 1.5 Tính chất hình dạng (Shape) . . . 4
- 1.6 Ví dụ dữ liệu và ứng dụng thực tế . . . 5

2 Phân phối Poisson (Poisson Distribution) 6
- 2.1 Định nghĩa . . . 6
- 2.2 Probability Mass Function – PMF . . . 6
- 2.3 Cumulative Distribution Function – CDF . . . 6
- 2.4 Các đặc trưng thống kê . . . 6
- 2.5 Mối liên hệ với phân phối nhị thức . . . 8

3 Phân phối Nhị thức Âm 8
- 3.1 Định nghĩa . . . 8
- 3.2 Probability Mass Function - PMF . . . 8
- 3.3 Cumulative Distribution Function - CDF . . . 9
- 3.4 Các đặc trưng thống kê . . . 9
- 3.5 Mối liên hệ với các phân phối khác . . . 10
- 3.6 Ví dụ dữ liệu và ứng dụng thực tế . . . 10

4 Phân phối Đa thức (Multinomial Distribution) 12
- 4.1 Định nghĩa . . . 12
- 4.2 Probability Mass Function – PMF . . . 12
- 4.3 Cumulative Distribution Function – CDF . . . 12
- 4.4 Các đặc trưng thống kê . . . 12

5 Phân phối Đều (Uniform Distribution) 13
- 5.1 1. Phân phối Đều Rời Rạc . . . 14
  - 5.1.1 Probability Mass Function - PMF . . . 14
  - 5.1.2 Cumulative Distribution Function - CDF . . . 14
- 5.2 2. Phân phối Đều Liên Tục . . . 14
  - 5.2.1 Probability Density Function - PDF . . . 14
  - 5.2.2 Cumulative Distribution Function - CDF . . . 15
  - 5.2.3 Các đặc trưng thống kê . . . 15
  - 5.2.4 Tính chất hình dạng (Shape) . . . 16
  - 5.2.5 Ví dụ dữ liệu và ứng dụng thực tế . . . 16

6 Phân phối Chuẩn (Normal Distribution) 16
- 6.1 Định nghĩa . . . 16
- 6.2 Probability Density Function – PDF . . . 17
- 6.3 Cumulative Distribution Function – CDF . . . 17
- 6.4 Các đặc trưng thống kê . . . 17

6.5 Tính chất hình dạng (Shape) 18
6.6 Ví dụ dữ liệu và ứng dụng thực tế 18

7 Phân phối Mũ **19**
7.1 Định nghĩa 19
7.2 Probability Density Function - PDF 19
7.3 Cumulative Distribution Function - CDF 20
7.4 Các đặc trưng thống kê 20
7.5 Tính chất hình dạng (Shape) 20
7.6 Ví dụ dữ liệu và ứng dụng thực tế 21

8 Phân phối Gamma (Gamma Distribution) **21**
8.1 Định nghĩa 21
8.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF) . . 21
8.3 Các đặc trưng thống kê 21
8.4 Biểu đồ minh họa 22
8.5 Ví dụ dữ liệu và bài toán thực tế 23

9 Phân phối Beta (Beta Distribution) **23**
9.1 Định nghĩa 23
9.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF) . . 24
9.3 Các đặc trưng thống kê 24
9.4 Biểu đồ minh họa 25
9.5 Cơ sở toán học: Phân phối Tiên nghiệm Liên hợp 26
9.6 Ví dụ thực tế: Ước lượng Tỷ lệ Phiếu ủng hộ 26

10 Phân phối Dirichlet (Dirichlet Distribution) **27**
10.1 Định nghĩa 27
10.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF) . . 27
10.3 Các đặc trưng thống kê 28
10.4 Biểu đồ minh họa 28
10.5 Ví dụ dữ liệu và bài toán thực tế 28

1 Phân phối Nhị thức (Binomial Distribution)

1.1 Định nghĩa

Phân phối nhị thức mô tả xác suất có chính xác k lần **thành công** trong n phép thử độc lập, trong đó mỗi phép thử có xác suất thành công p không đổi. Ta ký hiệu:

$$X \sim \text{Binomial}(n, p), \quad n \in \mathbb{N},\ 0 \leq p \leq 1$$

Đặt biệt, các điều kiện sau cần được thỏa:

- Số lượng phép thử n là cố định
- Các phép thử là độc lập nhau
- Xác suất thành công của từng phép thử là như nhau cho mỗi lần thử
- Mỗi phép thử, hoặc là thành công, hoặc là không thành công.

1.2 Probability Mass Function - PMF

Hàm trọng lượng xác suất của phân phối nhị thức được cho bởi:

$$P(X = k) = f(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

với:

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

là hệ số tổ hợp, biểu thị số cách chọn k thành công trong n phép thử.

1.3 Cumulative Distribution Function - CDF

Hàm xác suất tích lũy được định nghĩa là:

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i}$$

Không có công thức đóng cho $F(k; n, p)$, nhưng có thể tính xấp xỉ bằng hàm Beta không đều (incomplete Beta function):

$$F(k; n, p) = I_{1-p}(n-k, k+1)$$

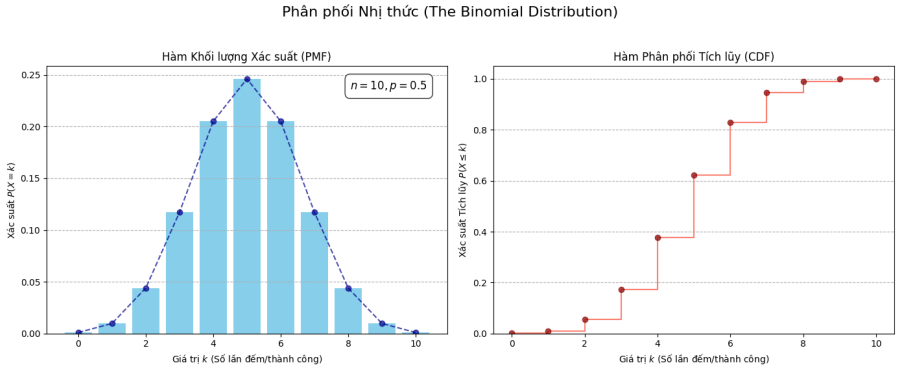

Figure 1: Biểu đồ Hàm Khối lượng Xác suất (PMF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Nhị thức. Phân phối này mô tả số lần thành công trong n phép thử độc lập.

1.4 Các đặc trưng thống kê

- **Giá trị kỳ vọng (Mean):**
$$\mathbb{E}[X] = np$$
- **Phương sai (Variance):**
$$\text{Var}(X) = np(1-p)$$
- **Mode (Giá trị có xác suất cao nhất):**
$$\text{mode} = \lfloor (n+1)p \rfloor$$
- **Median (Trung vị, xấp xỉ):**
$$\text{median} \approx \lfloor np + \tfrac{1}{2} \rfloor$$
- **Miền xác định:**
$$k \in \{0, 1, 2, \ldots, n\}$$

1.5 Tính chất hình dạng (Shape)

- Phân phối nhị thức là **đối xứng** nếu $p = 0.5$.
- **Lệch trái (left-skewed)** nếu $p > 0.5$.
- **Lệch phải (right-skewed)** nếu $p < 0.5$.
- Khi n lớn và p không quá gần 0 hoặc 1, phân phối nhị thức có thể được **xấp xỉ bằng phân phối chuẩn (Normal Distribution)** với:
$$X \approx \mathcal{N}(np,\ np(1-p))$$

1.6 Ví dụ dữ liệu và ứng dụng thực tế

Ứng dụng 1: Kiểm định chất lượng sản phẩm. Với số lượng các sản phẩm cho trước kết hợp với xác suất của một mặt hàng bị lỗi Phân phối nhị thức có thể giúp xây dựng mô hình và ước lượng số lượng mặt hàng bị lỗi, điều này giúp các nhà xây dựng sản phẩm cân nhắc về chất lượng sản phẩm cũng như việc quản lý hệ thống, thiết bị sản xuất.

Ứng dụng 2: Ứng dụng trong tài chính. Phân phối nhị thức đóng vai trò nền tảng trong *Binomial Option Pricing Model* – BOPM) Thay vì giả định giá tài sản biến thiên liên tục (như trong mô hình Black–Scholes), Mô hình này giả định rằng ở mỗi bước thời gian Δt, giá tài sản cơ sở S chỉ có thể:

$$S_u = S_0 u \quad \text{hoặc} \quad S_d = S_0 d$$

tức rằng tăng u lần hoặc là d lần Sau (n) bước (tức ta chia khoảng thời gian thành n windows và coi nó là rời rạc), giá cổ phiếu có thể đi qua nhiều đường khác nhau, ta có thể từ đó quan tâm đến số lần tăng k trong n bước.

Xác suất để cổ phiếu tăng đúng k lần tuân theo phân phối nhị thức:

$$P(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Từ đó, giá quyền chọn được tính bằng kỳ vọng có trọng số của các giá trị cuối cùng, với trọng số chính là xác suất nhị thức này.

$$C_0 = e^{-rT} \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}, \max(u^k d^{n-k} S_0 - K, 0)$$

Mô hình này cung cấp một cách tiếp cận rời rạc, trực quan và hiệu quả để ước lượng giá trị quyền chọn, đồng thời hội tụ về mô hình Black–Scholes khi $n \to \infty$.

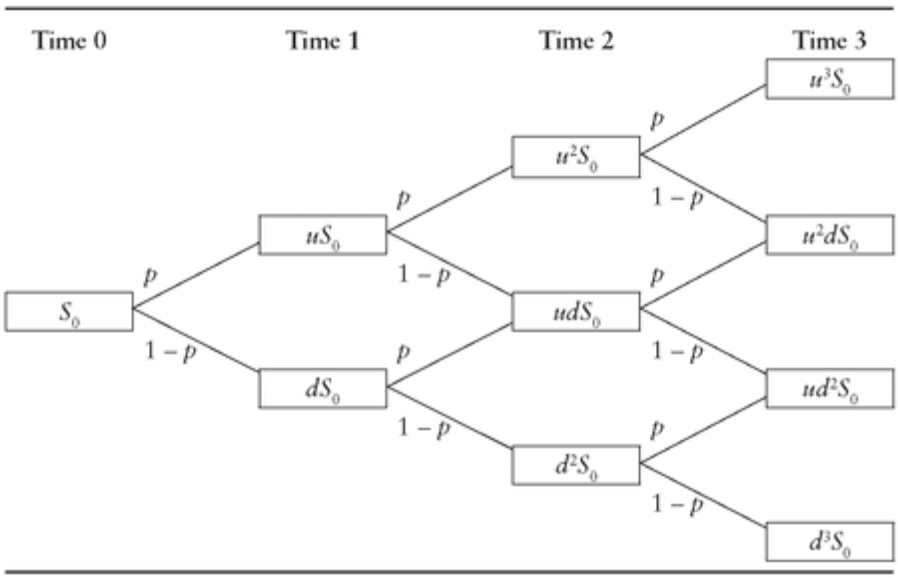

Figure 2: Cây nhị phân mô tả sự biến động của giá tài sản cơ sở qua ba giai đoạn thời gian trong mô hình định giá quyền chọn nhị thức

2 Phân phối Poisson (Poisson Distribution)

2.1 Định nghĩa

Phân phối Poisson mô tả xác suất của số sự kiện xảy ra trong một khoảng cố định (thời gian, không gian, v.v.), nếu các sự kiện xảy ra độc lập và với tốc độ trung bình λ không đổi. Ký hiệu:

$$X \sim \text{Poisson}(\lambda), \quad \lambda > 0$$

2.2 Probability Mass Function – PMF

Hàm trọng lượng xác suất của phân phối Poisson là:

$$P(X = k) = f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

2.3 Cumulative Distribution Function – CDF

Hàm xác suất tích lũy được định nghĩa là:

$$F(k; \lambda) = P(X \leq k) = \sum_{i=0}^{k} \frac{\lambda^i e^{-\lambda}}{i!}$$

2.4 Các đặc trưng thống kê

- **Kỳ vọng (Mean):**

$$\mathbb{E}[X] = \lambda$$

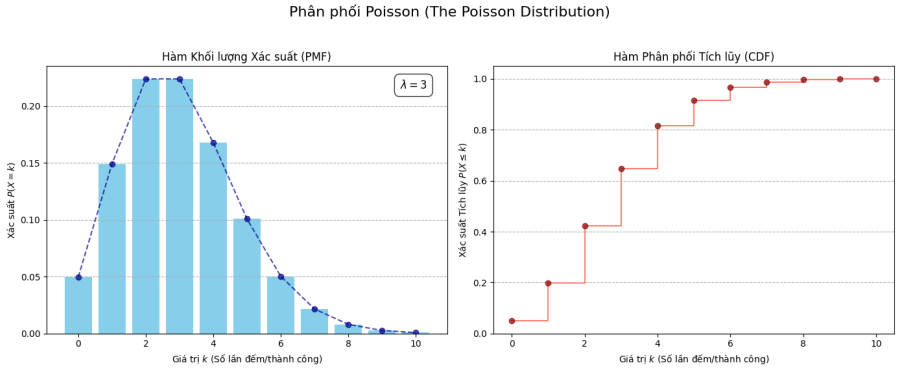

Figure 3: Biểu đồ PMF và CDF của Phân phối Poisson. Phân phối này mô hình hóa số sự kiện xảy ra trong một khoảng thời gian/không gian cố định, với tốc độ trung bình (λ) đã biết.

- **Phương sai (Variance):**
$$\mathrm{Var}(X) = \lambda$$
- **Mode (giá trị có xác suất cao nhất):** Nếu λ không phải số nguyên, mode $= \lfloor \lambda \rfloor$. Nếu λ là số nguyên, thì có hai mode là λ và $\lambda - 1$.
- **Median (Trung vị, xấp xỉ):** Không có công thức đóng chính xác; một xấp xỉ thường dùng là
$$\text{median} \approx \left\lfloor \lambda + \frac{1}{3} - \frac{1}{50\lambda} \right\rfloor$$
- **Miền xác định:**
$$k \in \{0, 1, 2, \dots\}$$
- **Hình dạng / Độ lệch:** - Phân phối Poisson thường mang lệch phải (right-skewed). - Khi λ lớn, phân phối gần đối xứng và có thể xấp xỉ bằng phân phối chuẩn.

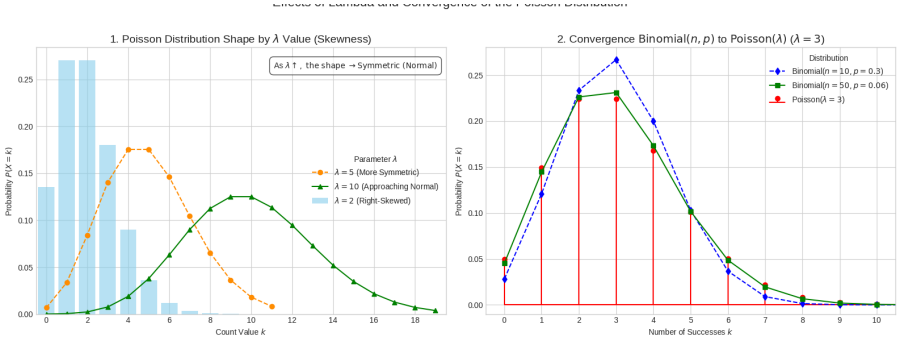

Figure 4: Hình dạng của phân phối Possion với các tham số khác nhau.

2.5 Mối liên hệ với phân phối nhị thức

Khi n rất lớn và p rất nhỏ sao cho $np = \lambda$ không đổi, phân phối nhị thức Binomial(n, p) hội tụ về phân phối Poisson Poisson(λ):

$$\lim_{n\to\infty,\, p\to 0,\, np=\lambda} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

Ví dụ dữ liệu và ứng dụng thực tế

Phân phối **Poisson** là công cụ tiêu chuẩn để mô hình hóa số lần xảy ra của các sự kiện hiếm và độc lập trong một khoảng thời gian cố định. Ví dụ, ta có thể xem xét việc số lần nhấp chuột vào một trang quảng cáo/ mặt hàng,...

Mục tiêu bài toán là dự đoán số lần (X) một khách hàng nhấp chuột vào quảng cáo trên trang web trong một khoảng thời gian cố định (ví dụ: 5 phút). Tỷ lệ nhấp chuột trung bình (mean click rate) trong khoảng thời gian đó. Giả định rằng các lần nhấp chuột xảy ra độc lập với một tốc độ không đổi. Xác suất để xảy ra chính xác k lần nhấp chuột được tính như sau:

$$P(X = k) = (e^{-\lambda} * \lambda^k)/k!$$

Điều kiện áp dụng: Phân phối Poisson chỉ phù hợp nếu **Trung bình gần bằng Phương sai** (Equidispersion). Nếu **Phương sai lớn hơn cả Trung bình** (Overdispersion), thì **Hồi quy Nhị thức Âm** cần được sử dụng để xử lý sự khác biệt hành vi lớn giữa các khách hàng.

3 Phân phối Nhị thức Âm

3.1 Định nghĩa

Phân phối Nhị thức Âm mô tả xác suất của số lần **thất bại** (k) xảy ra trước khi đạt được một số lượng **thành công** cố định là r. Phân phối này là một giải pháp quan trọng cho **dữ liệu đếm (count data)** khi có hiện tượng **phân tán quá mức (overdispersion)** so với mô hình Poisson.

Ta ký hiệu:

$$X \sim \text{NegativeBinomial}(r, p), \quad r \in \mathbb{N}^+,\ 0 < p \leq 1$$

Trong Data Science, nó thường được tham số hóa theo **kỳ vọng** (μ) và **tham số phân tán** (k hoặc α), nơi $\text{Var}(X) = \mu + \mu^2/k$.

3.2 Probability Mass Function - PMF

Hàm trọng lượng xác suất của phân phối nhị thức âm (số lần thất bại k trước r lần thành công) được cho bởi:

$$P(X = k) = f(k; r, p) = \binom{k+r-1}{k} p^r (1-p)^k, \quad k = 0, 1, 2, \ldots$$

với:

$$\binom{k+r-1}{k} = \frac{(k+r-1)!}{k!(r-1)!}$$

là hệ số tổ hợp, biểu thị số cách sắp xếp k thất bại và r thành công trong $(k+r)$ phép thử, với phép thử cuối cùng phải là thành công thứ r.

3.3 Cumulative Distribution Function - CDF

Hàm xác suất tích lũy được định nghĩa là:

$$F(k; r, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{i+r-1}{i} p^r (1-p)^i$$

Giống như phân phối Nhị thức, CDF của NBD có thể liên hệ với hàm Beta không đều:

$$F(k; r, p) = I_p(r, k+1)$$

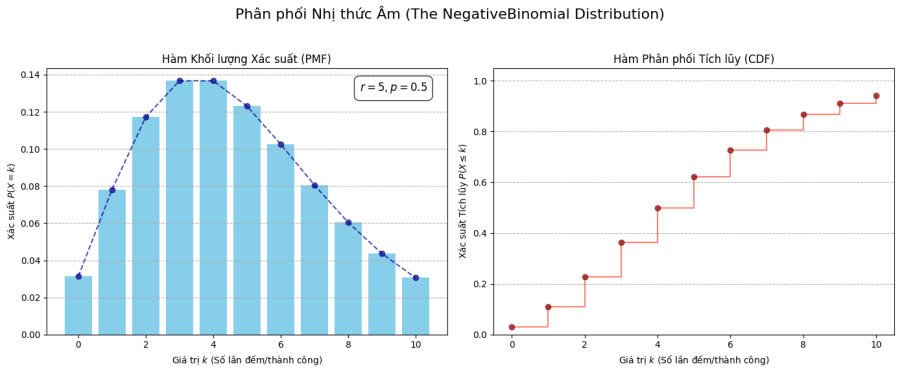

Figure 5: Biểu đồ Hàm Khối lượng Xác suất (PMF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Nhị thức Âm. Phân phối này mô tả số lần thất bại trước khi đạt r thành công cố định.

3.4 Các đặc trưng thống kê

- **Giá trị kỳ vọng (Mean):**

$$\mathbb{E}[X] = \frac{r(1-p)}{p} = \mu$$

- **Phương sai (Variance):**

$$\text{Var}(X) = \frac{r(1-p)}{p^2} = \mu + \frac{\mu^2}{r/(1-p)} = \mu + \frac{\mu^2}{k_{alt}}$$

  Lưu ý: Phương sai luôn **lớn hơn** giá trị kỳ vọng: $\text{Var}(X) > \mathbb{E}[X]$.

- **Miền xác định:**

$$k \in \{0, 1, 2, \dots\}$$

3.5 Mối liên hệ với các phân phối khác

- **Mở rộng của Hình học (Geometric):** Nếu $r = 1$, NBD trở thành Phân phối Hình học (Geometric Distribution), mô tả số lần thất bại trước *thành công đầu tiên*.
- **Xấp xỉ Poisson:** Nếu $r \to \infty$ và $p \to 1$ sao cho $\frac{r(1-p)}{p} = \lambda$ không đổi, NBD hội tụ về Poisson(λ).

3.6 Ví dụ dữ liệu và ứng dụng thực tế

Bài toán Dự đoán Tần suất Mua hàng

Mục tiêu bài toán là xây dựng mô hình dự đoán **số lần** một khách hàng cá nhân sẽ thực hiện giao dịch (mua hàng) trong một **khoảng thời gian cố định trong tương lai** (ví dụ: 6 tháng, 1 năm).

Để làm được điều đó, ta sử dụng dữ liệu giao dịch lịch sử của từng khách hàng, bao gồm:

- **Số lần mua hàng (k):** Tổng số giao dịch được ghi nhận trong khoảng thời gian quan sát đối với từng khách hàng.
- **Các biến giải thích (Covariates):** Đặc trưng hành vi và nhân khẩu học ảnh hưởng đến tần suất mua hàng — chẳng hạn như *giá trị đơn hàng trung bình*, *thời gian kể từ lần mua gần nhất*, hay *nguồn gốc khách hàng*.

Vấn đề của Phân phối Poisson

Khi áp dụng **Hồi quy Poisson (Poisson Regression)** để dự đoán tần suất mua hàng, người ta thường gặp phải hiện tượng **phân tán quá mức (Overdispersion)**, tức là phương sai của dữ liệu lớn hơn rất nhiều so với giá trị trung bình.

Nguyên nhân xuất phát từ hai giả định chính của mô hình Poisson:

- **Giả định về phương sai:** Phân phối Poisson yêu cầu Mean(μ) = Variance(σ^2).
- **Sự không đồng nhất hành vi:** Trong thực tế, có sự khác biệt rõ rệt giữa nhóm khách hàng "mua thường xuyên" và nhóm "mua ngẫu nhiên". Điều này khiến phương sai quan sát được trong dữ liệu lớn hơn nhiều so với trung bình, vi phạm giả định cơ bản của Poisson.

Giải pháp: Sử dụng Phân phối Nhị thức Âm (Negative Binomial Distribution)

Để xử lý hiện tượng *overdispersion*, ta sử dụng phân phối Nhị thức Âm (NBD), vốn là sự mở rộng của Poisson khi cho phép phương sai linh hoạt hơn.

Phân phối này có thể được hiểu như một **mô hình hợp (compound model)** giữa hai phân phối:

- **Phân phối Poisson:** Mô tả số lần mua hàng của một cá nhân, giả sử tốc độ mua hàng của họ là λ cố định.

- **Phân phối Gamma:** Mô hình hóa sự biến thiên ngẫu nhiên của λ giữa các cá nhân (tức là sự không đồng nhất trong hành vi mua hàng giữa khách hàng này với khách hàng khác).

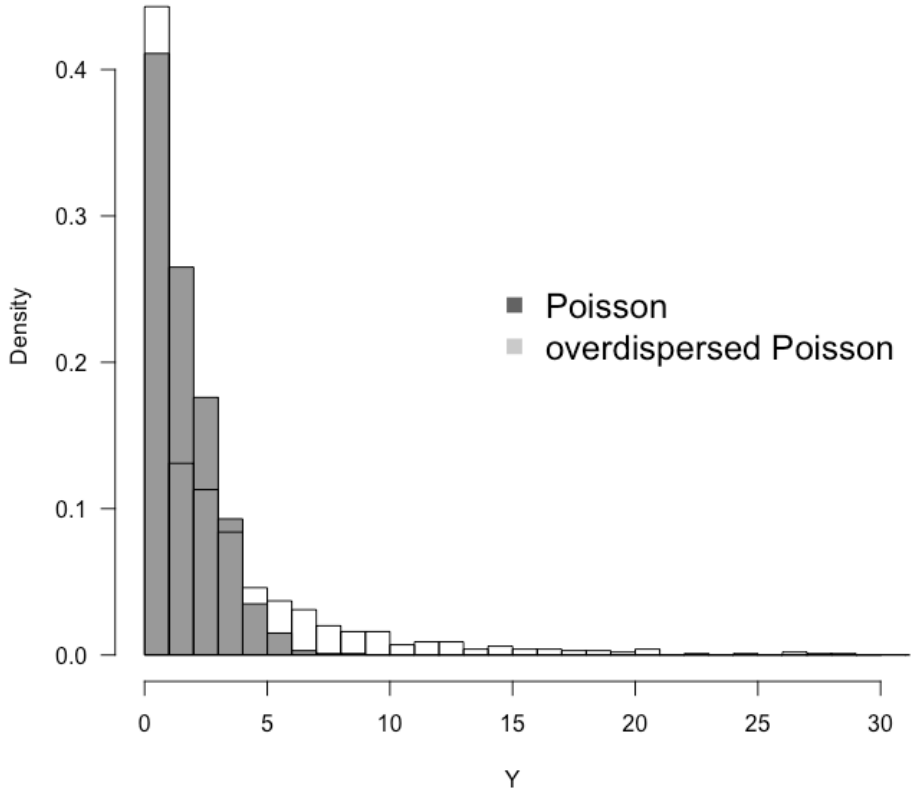

Figure 6: So sánh giữa phân phối Poisson chuẩn và Poisson có hiện tượng **overdispersion** (phương sai lớn hơn trung bình). Dữ liệu xám đậm thể hiện phân phối Poisson với tham số $\lambda = 2$, trong khi dữ liệu xám nhạt biểu diễn phân phối Poisson bị overdispersed với λ thay đổi ngẫu nhiên theo $2e^Z$, $Z \sim \mathcal{N}(0, 1)$.

Khi ta **tích hợp (marginalize)** λ theo phân phối Gamma, ta thu được phân phối Nhị thức Âm cho số lần mua hàng X:

$$X \sim \text{NegBinomial}(r, p)$$

với kỳ vọng và phương sai được biểu diễn là:

$$\mathbb{E}[X] = \mu, \qquad \text{Var}(X) = \mu + \frac{\mu^2}{k}$$

Trong đó:

- μ là giá trị trung bình kỳ vọng của số lần mua hàng.

- k (hoặc α) là **tham số phân tán (dispersion parameter)**. Khi $k \to \infty$, phương sai tiệm cận μ, và phân phối Nhị thức Âm hội tụ về phân phối Poisson.

Nhờ có tham số phân tán này, mô hình Negative Binomial linh hoạt hơn Poisson, cho phép mô hình hóa chính xác hành vi mua hàng thực tế, nơi phương sai thường vượt xa trung bình.

4 Phân phối Đa thức (Multinomial Distribution)

4.1 Định nghĩa

Phân phối **Đa thức** là sự mở rộng tự nhiên của phân phối Nhị thức (Binomial Distribution), được dùng để mô tả xác suất của các kết quả thuộc nhiều hơn hai loại, trong n lần thử độc lập, mỗi lần thử có k khả năng xảy ra tương ứng với các xác suất $p_1, p_2, \ldots, p_k$ (với $\sum_{i=1}^{k} p_i = 1$).

Ký hiệu:

$$(X_1, X_2, \ldots, X_k) \sim \text{Multinomial}(n; p_1, p_2, \ldots, p_k)$$

với điều kiện $\sum_{i=1}^{k} X_i = n$.

4.2 Probability Mass Function – PMF

Hàm trọng lượng xác suất của phân phối Đa thức là:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!} \prod_{i=1}^{k} p_i^{x_i}, \quad \text{với } \sum_{i=1}^{k} x_i = n$$

4.3 Cumulative Distribution Function – CDF

Không có biểu thức đóng cho CDF của phân phối Đa thức. Tuy nhiên, ta có thể tính xác suất cộng dồn bằng cách lấy tổng các giá trị PMF cho tất cả các tổ hợp $\{x_i\}$ thỏa $\sum_i x_i \leq m$:

$$F(m; n, \mathbf{p}) = P\left(\sum_{i=1}^{k} X_i \leq m\right) = \sum_{x_1 + \cdots + x_k \leq m} \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!} \prod_{i=1}^{k} p_i^{x_i}$$

4.4 Các đặc trưng thống kê

- **Kỳ vọng (Mean):**

$$\mathbb{E}[X_i] = np_i, \quad i = 1, \ldots, k$$

- **Phương sai (Variance):**

$$\text{Var}(X_i) = np_i(1 - p_i)$$

- **Hiệp phương sai (Covariance):**

$$\text{Cov}(X_i, X_j) = -np_ip_j, \quad i \neq j$$

- **Miền xác định:**

$$x_i \in \{0, 1, \ldots, n\}, \quad \sum_{i=1}^{k} x_i = n$$

- **Hình dạng:** - Phân phối này luôn nằm trong miền đơn hình (simplex). - Khi n lớn, phân phối Multinomial có thể được xấp xỉ bởi phân phối Chuẩn đa biến:

$$\mathbf{X} \sim \mathcal{N}\big(n\mathbf{p},\, n(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T)\big)$$

Ứng dụng: Mô hình Hóa Phân Bố Từ Ngữ trong Văn Bản

Một ứng dụng nổi bật của **phân phối Multinomial** là trong mô hình **Bag-of-Words (BoW)** trong xử lý ngôn ngữ tự nhiên (NLP). Giả sử ta có một từ điển gồm k từ khác nhau, và một văn bản chứa tổng cộng n từ. Nếu ta coi mỗi từ trong văn bản được chọn ngẫu nhiên và độc lập với xác suất p_i (xác suất xuất hiện của từ thứ i trong ngôn ngữ), thì số lần xuất hiện $(X_1, X_2, \ldots, X_k)$ của các từ đó trong văn bản tuân theo:

$$(X_1, X_2, \ldots, X_k) \sim \mathrm{Multinomial}(n; p_1, p_2, \ldots, p_k)$$

Ví dụ, trong mô hình **Naive Bayes phân loại văn bản**, ta giả định rằng xác suất xuất hiện của các từ trong một tài liệu thuộc lớp C_j được mô hình hóa bởi tham số $\mathbf{p}^{(j)} = (p_1^{(j)}, \ldots, p_k^{(j)})$. Với một tài liệu có vector tần suất từ $\mathbf{x} = (x_1, \ldots, x_k)$, ta có:

$$P(\mathbf{x} \mid C_j) = \frac{n!}{x_1! \cdots x_k!} \prod_{i=1}^{k} \left(p_i^{(j)}\right)^{x_i}$$

và xác suất hậu nghiệm theo định lý Bayes:

$$P(C_j \mid \mathbf{x}) = \frac{P(C_j)\, P(\mathbf{x} \mid C_j)}{\sum_l P(C_l)\, P(\mathbf{x} \mid C_l)}$$

Ảnh hưởng thực tế: Nhờ giả định Multinomial, mô hình Naive Bayes có thể ước lượng trực tiếp xác suất từ vựng của từng lớp dựa trên tần suất xuất hiện từ trong dữ liệu huấn luyện, cho phép phân loại tài liệu hiệu quả và có cơ sở xác suất chặt chẽ. Phân phối Multinomial đóng vai trò nền tảng trong việc định nghĩa hàm khả năng (likelihood) của văn bản trong nhiều mô hình ngôn ngữ thống kê.

5 Phân phối Đều (Uniform Distribution)

Phân phối đều là một trong những phân phối xác suất cơ bản nhất, đặc trưng bởi **xác suất hoặc mật độ phân bố đều trên toàn bộ miền xác định**. Phân phối đều có hai loại:

- **Phân phối đều rời rạc (Discrete Uniform Distribution)**.
- **Phân phối đều liên tục (Continuous Uniform Distribution)**.

Ta ký hiệu:

$$X \sim \mathrm{Uniform}(a, b), \quad -\infty < a < b < +\infty$$

5.1 1. Phân phối Đều Rời Rạc

5.1.1 Probability Mass Function - PMF

:

Hàm trọng lượng xác suất của phân phối Đều là: Giả sử biến ngẫu nhiên X có phân phối đều rời rạc trên tập $\{x_1, x_2, \ldots, x_n\}$. Khi đó:

$$P(X = x) = \begin{cases} \dfrac{1}{n}, & x \in \mathcal{S}, \\ 0, & \text{ngược lại.} \end{cases}$$

hay viết ngắn gọn:

$$\text{PMF}(x) = \frac{1}{n}, \quad \forall x \in \{x_1, x_2, \ldots, x_n\}.$$

5.1.2 Cumulative Distribution Function - CDF

Hàm xác suất tích lũy được định nghĩa là:

$$F(x) = P(X \leq x) = \begin{cases} 0, & x < x_1 \\ \dfrac{\text{số phần tử } \leq x}{n}, & x_1 \leq x \leq x_n \\ 1, & x \geq x_n \end{cases}$$

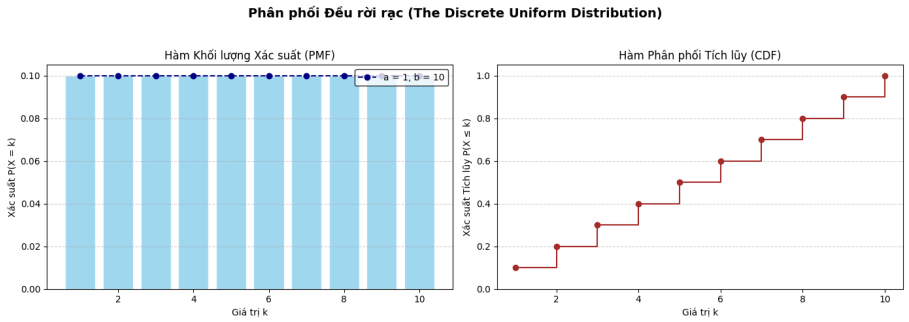

Figure 7: Biểu đồ Hàm Khối lượng Xác suất (PMF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Đều Rời Rạc.

5.2 2. Phân phối Đều Liên Tục

5.2.1 Probability Density Function - PDF

: Hàm mật độ xác suất của phân phối đều liên tục được định nghĩa như sau: Giả sử biến ngẫu nhiên X có phân phối đều liên tục trên đoạn $[a, b]$. Khi đó:

$$f(x;a,b) = \begin{cases} \dfrac{1}{b-a}, & a \le x \le b, \\ 0, & \text{ngược lại.} \end{cases}$$

Điều này phản ánh rằng mọi giá trị trong khoảng $[a,b]$ đều có **mật độ xác suất như nhau**.

5.2.2 Cumulative Distribution Function - CDF

Hàm phân phối tích lũy (CDF) của phân phối đều liên tục được xác định bằng cách tích phân hàm mật độ:

$$F(x;a,b) = P(X \le x) = \int_{-\infty}^{x} f(t;a,b)\,dt.$$

Do đó:

$$F(x;a,b) = \begin{cases} 0, & x < a, \\ \dfrac{x-a}{b-a}, & a \le x \le b, \\ 1, & x > b. \end{cases}$$

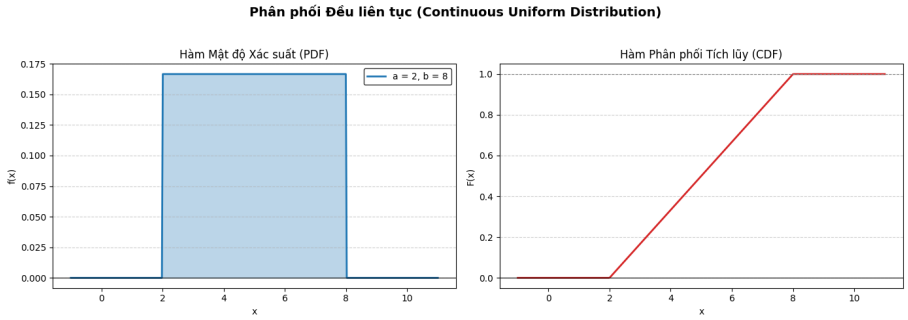

Figure 8: Biểu đồ Hàm Mật độ Xác suất (PDF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Đều Liên tục.

5.2.3 Các đặc trưng thống kê

- **Kỳ vọng (Mean):**

$$\mathbb{E}[X] = \frac{a+b}{2}$$

- **Phương sai (Variance):**

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

- **Mode:** Mọi giá trị trong $[a, b]$ đều có cùng mật độ nên không có mode duy nhất.
- **Median:**
$$\text{Median}(X) = \frac{a+b}{2}$$
- **Miền xác định:**
$$x \in [a, b]$$

5.2.4 Tính chất hình dạng (Shape)

- Phân phối đều liên tục có **dạng hình chữ nhật**, vì mật độ xác suất không đổi trong $[a, b]$.
- Phân phối này **đối xứng hoàn toàn** quanh trung điểm $\frac{a+b}{2}$.
- Không có độ lệch (skewness = 0) và hệ số nhọn (kurtosis excess = -1.2).

5.2.5 Ví dụ dữ liệu và ứng dụng thực tế

Ứng dụng 1: Kiểm soát chất lượng sản phẩm trong dây chuyền sản xuất. Trong nhiều quy trình sản xuất, các sai số gia công cơ khí (như độ lệch kích thước, độ sâu mũi khoan, hoặc vị trí khoan lỗ) thường nằm trong một khoảng xác định $[a, b]$. Khi không có yếu tố nào khiến sai số tập trung về phía nào trong khoảng, một giả định hợp lý là sai số này tuân theo phân phối đều liên tục $\mathcal{U}(a, b)$. Mô hình này giúp các kỹ sư chất lượng ước lượng xác suất sản phẩm vượt ngoài giới hạn kỹ thuật, ví dụ như:
$$P(X > b_{\text{chuẩn}}) = \frac{b - b_{\text{chuẩn}}}{b - a}.$$
Nhờ đó, doanh nghiệp có thể xác định tỷ lệ lỗi dự kiến, điều chỉnh khuôn mẫu, hoặc tối ưu quy trình để giảm chi phí kiểm tra.

Ứng dụng 2: Kiểm tra ngẫu nhiên trong quản lý kho và logistics. Trong công tác quản lý kho hoặc kiểm kê hàng hóa, nhân viên thường chọn ngẫu nhiên một vị trí trong khu vực lưu trữ để kiểm tra chất lượng hoặc số lượng hàng. Nếu mỗi điểm trong khu vực đều có khả năng được chọn như nhau, thì vị trí chọn kiểm tra tuân theo phân phối đều trên miền không gian của kho. Cách tiếp cận này giúp đảm bảo quá trình kiểm tra là khách quan, không thiên vị, từ đó phát hiện các sai lệch hoặc hàng lỗi rải rác trong kho mà không cần kiểm tra toàn bộ.

6 Phân phối Chuẩn (Normal Distribution)

6.1 Định nghĩa

Phân phối Chuẩn (hay còn gọi là phân phối Gauss) là một trong những phân phối xác suất quan trọng nhất trong thống kê, mô tả các hiện tượng ngẫu nhiên có xu hướng tập trung quanh một giá trị trung bình.

Phân phối chuẩn được đặc trưng bởi hai tham số:

- μ: giá trị kỳ vọng (mean) — thể hiện vị trí trung tâm của phân phối.
- σ^2: phương sai (variance) — thể hiện độ phân tán của phân phối.

Ký hiệu:

$$X \sim \mathcal{N}(\mu, \sigma^2), \quad -\infty < \mu < +\infty,\ \sigma > 0$$

6.2 Probability Density Function – PDF

Hàm mật độ xác suất (PDF) của phân phối Chuẩn được cho bởi:

$$f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in (-\infty, +\infty),$$

trong đó:

- $\mu \in \mathbb{R}$: giá trị kỳ vọng (trung tâm của phân phối),
- $\sigma > 0$: độ lệch chuẩn (độ lan rộng của phân phối).

Nếu $X \sim \mathcal{N}(\mu, \sigma^2)$ thì PDF đạt cực đại tại $x = \mu$, và đồ thị có dạng **chuông đối xứng**, gọi là *đường cong Gauss*.

6.3 Cumulative Distribution Function – CDF

Hàm phân phối tích lũy (CDF) của biến ngẫu nhiên $X \sim \mathcal{N}(\mu, \sigma^2)$ được định nghĩa bởi:

$$F(x; \mu, \sigma) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(t-\mu)^2}{2\sigma^2}\right] dt.$$

Tích phân này biểu diễn **xác suất để biến ngẫu nhiên X nhận giá trị nhỏ hơn hoặc bằng x**.

Không tồn tại công thức nguyên hàm dạng đóng cho tích phân này, do đó việc tính toán thường được thực hiện bằng:

- Bảng phân phối chuẩn (Z–table),
- Các hàm tích hợp trong phần mềm thống kê (R, Python, Excel, v.v.),
- Hoặc thông qua **chuẩn hóa** về phân phối chuẩn tắc.

6.4 Các đặc trưng thống kê

- **Kỳ vọng (Mean):**
$$\mathbb{E}[X] = \mu$$
- **Phương sai (Variance):**
$$\mathrm{Var}(X) = \sigma^2$$
- **Mode (Giá trị có xác suất cao nhất):** $\mathrm{mode} = \mu$.

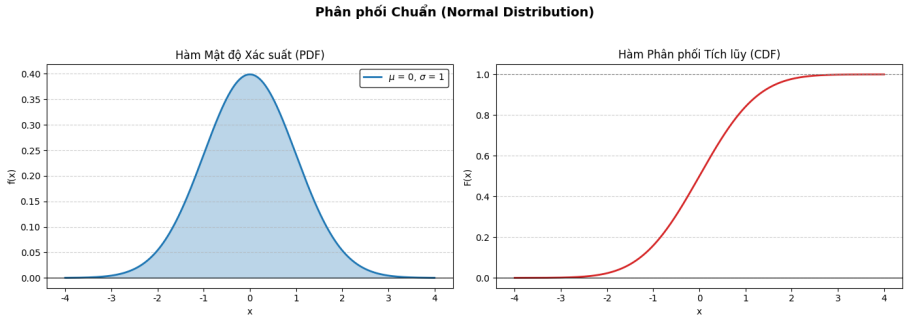

Figure 9: Biểu đồ PDF và CDF của Phân phối Chuẩn.

- **Median (Trung vị):** median $= \mu$.
- **Skewness (Độ lệch):** $\gamma_1 = 0$ (phân phối đối xứng).
- **Excess kurtosis (Hệ số nhọn dư):** $\gamma_2 = 0$ (kurtosis chuẩn bằng 3).
- **Miền xác định:** $x \in (-\infty, +\infty)$.

6.5 Tính chất hình dạng (Shape)

- Phân phối chuẩn có **dạng đường cong hình chuông đối xứng** (*bell curve*) quanh trung bình μ.
- **Đối xứng hoàn hảo**: trung bình, trung vị và mode đều trùng tại μ.
- **Độ lệch** (skewness) $= 0$; **kurtosis chuẩn** $= 3$ (excess kurtosis $= 0$).

Ví dụ dữ liệu và ứng dụng thực tế

6.6 Ví dụ dữ liệu và ứng dụng thực tế

Ứng dụng 1: Phân phối chuẩn trong điểm thi THPT Quốc gia. Điểm số của thí sinh chịu ảnh hưởng của nhiều yếu tố ngẫu nhiên nhỏ và độc lập (như kiến thức, tâm lý phòng thi, độ khó của đề, may mắn, v.v.).

Ví dụ, với dữ liệu phổ điểm môn Toán năm 2025 (hình 10), ta có các tham số thống kê đặc trưng như sau:

$$\text{Trung bình } (\mu) = 4.78, \quad \text{Độ lệch chuẩn } (\sigma) = 1.68.$$

Giả sử điểm của thí sinh được mô hình hoá bởi biến ngẫu nhiên:

$$X \sim \mathcal{N}(4.78,\ 1.68^2).$$

Ta có thể ước lượng xác suất đạt điểm cao, chẳng hạn điểm trên 8:

$$P(X > 8) = 1 - \Phi\Big(\frac{8 - 4.78}{1.68}\Big) = 1 - \Phi(1.92) \approx 0.0274,$$

tức chỉ khoảng **2.74%** thí sinh có điểm trên 8.

Kết quả này phù hợp với trực quan từ biểu đồ phổ điểm: phần lớn thí sinh tập trung quanh mức trung bình 4–6 điểm, và số lượng giảm dần ở hai đầu phổ điểm (dạng "chuông"). Điều này cho thấy phân phối chuẩn mô tả khá tốt xu hướng điểm thi thực tế.

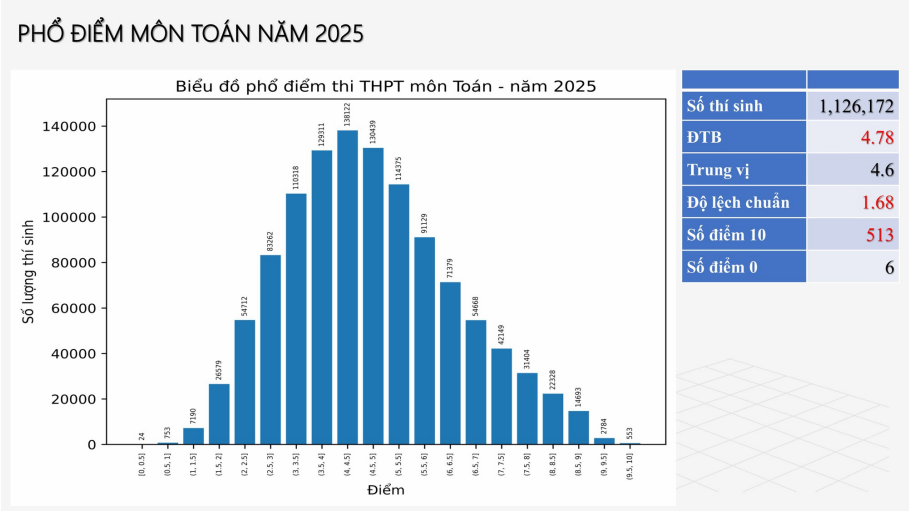

Figure 10: Phổ điểm môn Toán kỳ thi THPT Quốc gia năm 2025 ($\mu = 4.78$, $\sigma = 1.68$).

Ứng dụng 2: Kiểm định giả thuyết và khoảng tin cậy. Phân phối chuẩn là nền tảng cho nhiều phương pháp thống kê suy luận, đặc biệt là khi làm việc với mẫu lớn hoặc phương sai tổng thể đã biết.

7 Phân phối Mũ

7.1 Định nghĩa

Phân phối mũ mô tả thời gian giữa các sự kiện liên tiếp trong một quá trình Poisson với tốc độ (rate) không đổi. Nó là phiên bản liên tục tương ứng của phân phối hình học.

Ta ký hiệu:

$$X \sim \text{Exponential}(\lambda), \quad \lambda > 0.$$

7.2 Probability Density Function - PDF

Hàm mật độ xác suất của phân phối mũ được cho bởi:

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

7.3 Cumulative Distribution Function - CDF

Hàm xác suất tích lũy được định nghĩa là:

$$F(x;\lambda) = P(X \leq x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

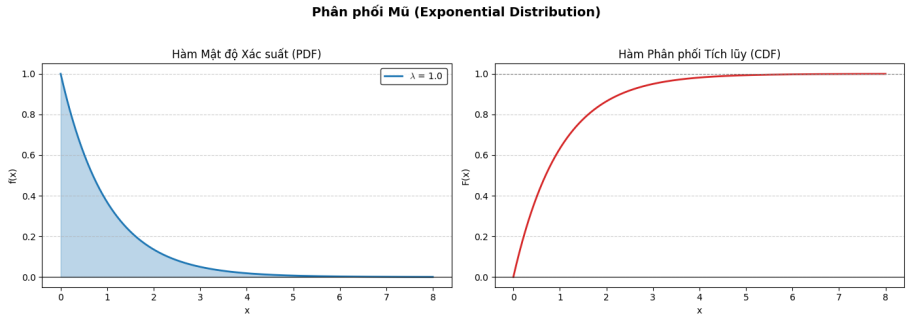

Figure 11: Biểu đồ Hàm Mật độ Xác suất (PDF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Mũ.

7.4 Các đặc trưng thống kê

- **Giá trị kỳ vọng (Mean):**
$$\mathbb{E}[X] = \frac{1}{\lambda}$$
- **Phương sai (Variance):**
$$\text{Var}(X) = \frac{1}{\lambda^2}$$
- **Mode:** 0
- **Median:**
$$\text{Median}(X) = \frac{\ln 2}{\lambda}$$
- **Miền xác định:** $x \in [0, +\infty)$

7.5 Tính chất hình dạng (Shape)

- Phân phối mũ có **dạng hàm mật độ giảm đơn điệu** trên $[0, \infty)$, với **đỉnh tại** $x = 0$ và giảm theo hàm mũ khi x tăng.
- **Lệch phải rõ rệt**: phần lớn dữ liệu tập trung ở gần số 0, và tần suất giảm dần theo hàm mũ khi giá trị tăng lên, tạo thành một cái "đuôi" dài về phía dương.
- **Không đối xứng**: chỉ có một đuôi phải, giảm theo $e^{-\lambda x}$.
- **Tính không nhớ (memoryless)**: $P(X > s + t \mid X > s) = P(X > t)$, đặc trưng riêng của phân phối mũ.

7.6 Ví dụ dữ liệu và ứng dụng thực tế

Ứng dụng 1: Độ tin cậy sản phẩm và thời gian hỏng hóc. Phân phối mũ thường được sử dụng để mô hình hóa *thời gian giữa các hỏng hóc* của linh kiện điện tử, thiết bị cơ khí hoặc hệ thống, đặc biệt khi **tốc độ hỏng hóc không thay đổi theo thời gian**. Đây là giả định thường gặp trong giai đoạn hoạt động ổn định của sản phẩm (không tính giai đoạn "hao mòn ban đầu" hoặc "lão hóa").

Ứng dụng 2: Quản lý lưu lượng và hàng đợi. Trong các hệ thống xếp hàng (ví dụ: khách đến cửa hàng, gói tin đến máy chủ), thời gian giữa các lượt đến thường được giả định là phân phối mũ.

8 Phân phối Gamma (Gamma Distribution)

Phân phối Gamma là một phân phối xác suất liên tục, dương và có hình dạng linh hoạt, thường được sử dụng để mô hình hóa thời gian chờ đợi hoặc các đại lượng dương có tính chất kéo dài.

8.1 Định nghĩa

Một biến ngẫu nhiên X tuân theo phân phối Gamma với tham số hình dạng (shape parameter) $k > 0$ và tham số tỷ lệ (scale parameter) $\theta > 0$, ký hiệu $X \sim \text{Gamma}(k, \theta)$, có hàm mật độ xác suất (PDF) được cho bởi:

$$f(x; k, \theta) = \frac{x^{k-1} e^{-x/\theta}}{\Gamma(k)\theta^k} \quad \text{với } x > 0$$

Trong đó, $\Gamma(k)$ là hàm Gamma, được định nghĩa là $\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt$.

8.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF)

Hàm xác suất tích lũy của phân phối Gamma không có dạng đóng (closed-form) đơn giản và thường được biểu diễn thông qua hàm Gamma không đầy đủ quy chuẩn (regularized lower incomplete gamma function) $P(k, x/\theta)$:

$$F(x; k, \theta) = P\left(k, \frac{x}{\theta}\right) = \frac{1}{\Gamma(k)} \int_0^x t^{k-1} e^{-t/\theta} dt$$

8.3 Các đặc trưng thống kê

- **Miền giá trị:** $x \in (0, \infty)$.
- **Giá trị kỳ vọng (Mean):** $E[X] = k\theta$.
- **Phương sai (Variance):** $\text{Var}[X] = k\theta^2$.

- **Mode:**
$$\text{Mode} = \begin{cases} (k-1)\theta & \text{nếu } k > 1 \\ \text{không xác định (tại 0)} & \text{nếu } k \leq 1 \end{cases}$$
- **Median:** Không có dạng đóng. Đối với các giá trị lớn của k, median xấp xỉ $k\theta - \frac{1}{3}\theta$.
- **Tính chất đối xứng/lệch:** Phân phối Gamma thường bị lệch phải. Khi k tăng, phân phối trở nên đối xứng hơn. Hệ số lệch (skewness) là $\frac{2}{\sqrt{k}}$.

8.4 Biểu đồ minh họa

Dưới đây là biểu đồ minh họa hàm mật độ xác suất của phân phối Gamma với các giá trị khác nhau của tham số hình dạng k và tham số tỷ lệ θ.

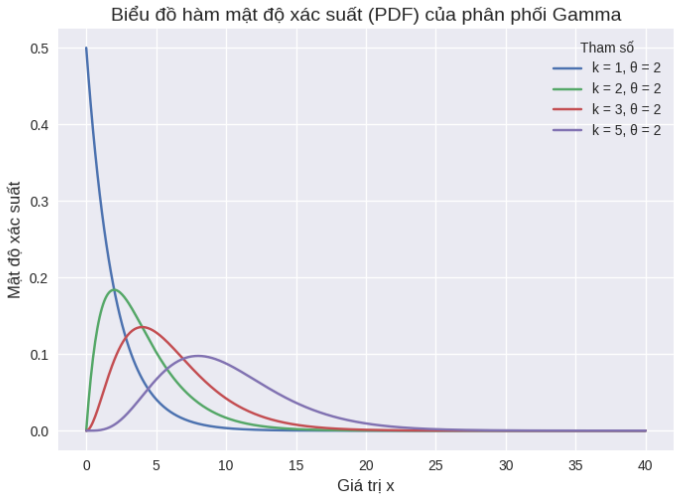

Figure 12: Hàm mật độ xác suất của Phân phối Gamma với các tham số khác nhau.

Dưới đây, là biểu đồ minh họa hàm phân phối tích lũy của phan phối Gamma

8.5 Ví dụ dữ liệu và bài toán thực tế

Dưới đây là biểu đồ minh họa hàm xác suất tích lũy của phân phối Gamma Nghịch đảo

8.6 Ví dụ dữ liệu và bài toán thực tế

Phân phối tiên nghiệm cho phương sai Trong thống kê Bayes, khi chúng ta cần đặt một phân phối tiên nghiệm cho tham số phương sai σ^2 của một phân phối chuẩn (normal distribution) hoặc các mô hình tuyến tính, phân phối Gamma Nghịch đảo là một lựa chọn phổ biến. Ví dụ, trong mô hình hồi quy tuyến tính, nếu chúng ta không có nhiều thông tin về phương sai của sai số, chúng ta có thể đặt một tiên nghiệm Gamma Nghịch đảo như $\text{InvGamma}(0.001, 0.001)$ để biểu thị sự không chắc chắn cao.

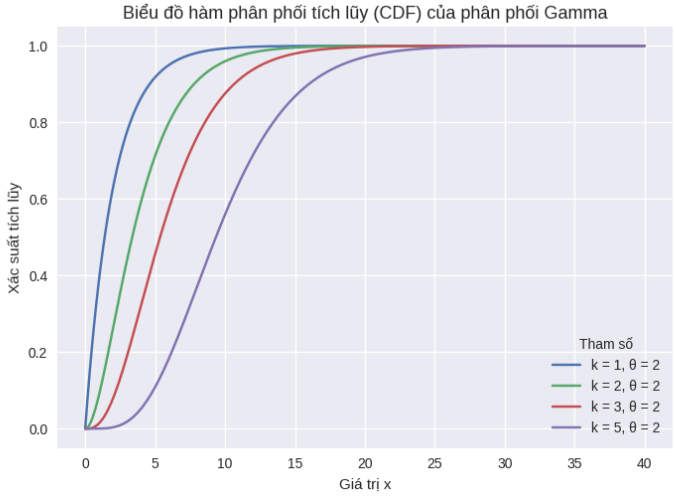

Figure 13: Hàm mật độ xác suất của Phân phối Gamma với các tham số khác nhau.

9 Phân phối Beta (Beta Distribution)

Phân phối Beta là một phân phối xác suất liên tục được định nghĩa trên khoảng $[0, 1]$. Nó rất hữu ích để mô hình hóa các đại lượng có giá trị giới hạn trong khoảng này, chẳng hạn như xác suất, tỷ lệ hoặc tỷ lệ phần trăm.

9.1 Định nghĩa

Một biến ngẫu nhiên X tuân theo phân phối Beta với hai tham số hình dạng $\alpha > 0$ và $\beta > 0$, ký hiệu $X \sim \text{Beta}(\alpha, \beta)$, có hàm mật độ xác suất được cho bởi:

$$f(x; \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)} \quad \text{với } 0 \leq x \leq 1$$

Trong đó, $B(\alpha, \beta)$ là hàm Beta, được định nghĩa là $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$.

9.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF)

Hàm xác suất tích lũy của phân phối Beta được biểu diễn thông qua hàm Beta không đầy đủ quy chuẩn (regularized incomplete beta function):

$$F(x; \alpha, \beta) = I_x(\alpha, \beta) = \frac{B_x(\alpha, \beta)}{B(\alpha, \beta)}$$

Trong đó, $B_x(\alpha, \beta) = \int_0^x t^{\alpha-1}(1-t)^{\beta-1} dt$ là hàm Beta không đầy đủ.

9.3 Các đặc trưng thống kê

- **Miền giá trị:** $x \in [0, 1]$.

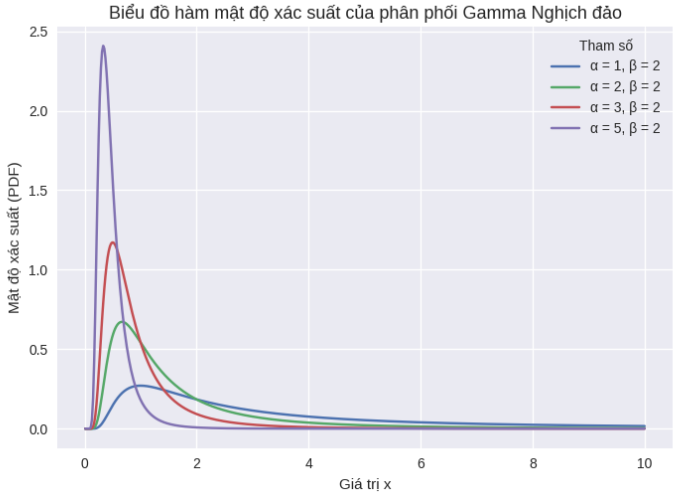

Figure 14: Hàm mật độ xác suất của Phân phối Gamma Nghịch đảo với các tham số khác nhau.

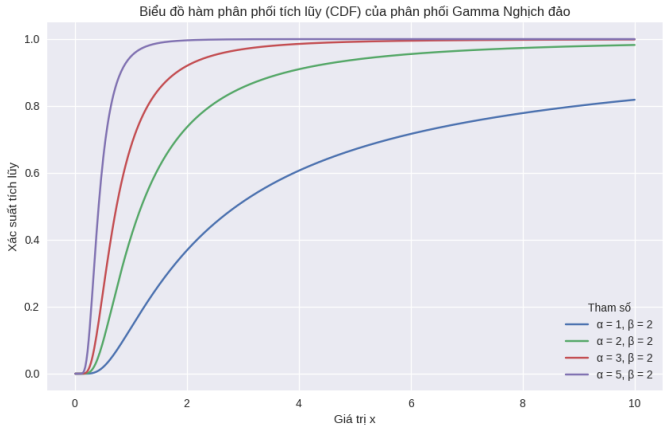

Figure 15: Hàm mật độ xác suất của Phân phối Gamma Nghịch đảo với các tham số khác nhau.

- **Giá trị kỳ vọng (Mean):** $E[X] = \frac{\alpha}{\alpha+\beta}$.
- **Phương sai (Variance):** $\text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.
- **Mode:**

$$\text{Mode} = \begin{cases} \frac{\alpha-1}{\alpha+\beta-2} & \text{nếu } \alpha > 1, \beta > 1 \\ 0 & \text{nếu } \alpha = 1, \beta > 1 \\ 1 & \text{nếu } \alpha > 1, \beta = 1 \\ \text{tất cả các giá trị trong (0,1)} & \text{nếu } \alpha = 1, \beta = 1 \text{ (phân phối đều)} \end{cases}$$

- **Median:** Không có dạng đóng. Đối với $\alpha = \beta$, median bằng 0.5.

- **Tính chất đối xứng/lệch:**
  - Nếu $\alpha = \beta$, phân phối đối xứng.
  - Nếu $\alpha > \beta$, phân phối lệch trái.
  - Nếu $\alpha < \beta$, phân phối lệch phải.

9.4 Biểu đồ minh họa

Dưới đây là biểu đồ minh họa hàm mật độ xác suất của phân phối Beta với các cặp tham số α, β khác nhau.

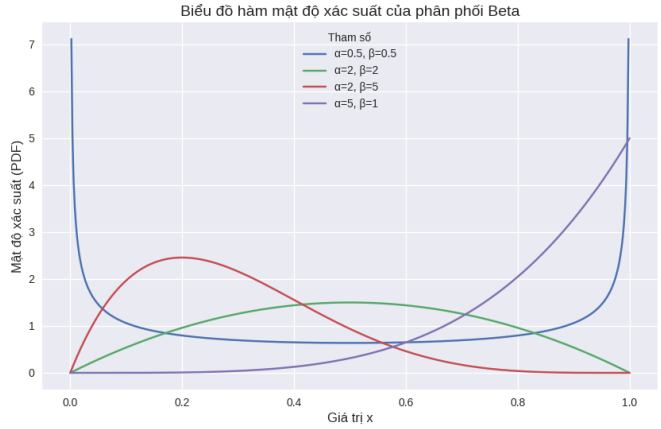

Figure 16: Hàm mật độ xác suất của Phân phối Beta với các tham số khác nhau.

C
Dưới đây là biểu đồ minh họa hàm xác suất tích lũy của phân phối Beta

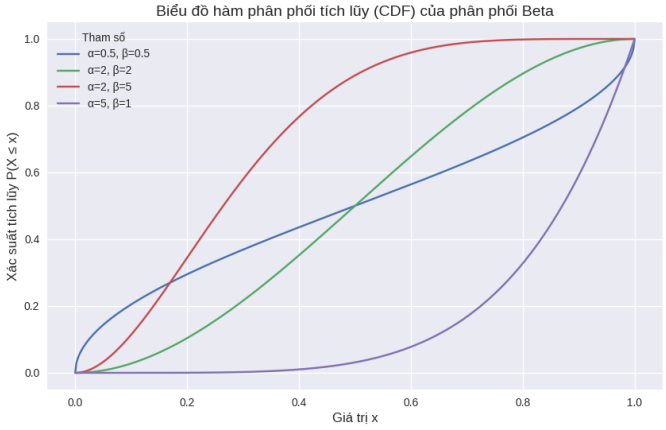

Figure 17: Hàm mật độ xác suất của Phân phối Beta với các tham số khác nhau.

9.5 Ví dụ dữ liệu và bài toán thực tế

Phân phối tiên nghiệm cho xác suất Bernoulli Trong thống kê Bayes, phân phối Beta là liên hợp tiên nghiệm (conjugate prior) cho tham số p của phân phối Bernoulli hoặc phân phối nhị thức (binomial distribution). Nếu chúng ta muốn ước lượng xác suất thành công p của một thí nghiệm, chúng ta có thể đặt một tiên nghiệm Beta trên p. Ví dụ, Beta$(1, 1)$ là tiên nghiệm không thông tin (uniform prior), trong khi Beta$(2, 2)$ cho thấy chúng ta tin rằng p có xu hướng ở gần 0.5.

Tỷ lệ phiếu bầu Giả sử chúng ta muốn mô hình hóa tỷ lệ phiếu bầu cho một ứng cử viên trong một cuộc bầu cử. Các giá trị tỷ lệ này nằm trong khoảng $[0, 1]$. Phân phối Beta có thể được sử dụng để biểu diễn sự không chắc chắn của chúng ta về tỷ lệ thực tế, dựa trên các cuộc thăm dò hoặc dữ liệu lịch sử.

10 Phân phối Dirichlet (Dirichlet Distribution)

Phân phối Dirichlet là một phân phối xác suất đa biến liên tục trên một simplex tiêu chuẩn. Nó là tổng quát hóa của phân phối Beta cho nhiều hơn hai danh mục và thường được sử dụng làm phân phối tiên nghiệm cho các tham số của phân phối Categorical hoặc phân phối đa thức (multinomial distribution).

10.1 Định nghĩa

Một vector ngẫu nhiên $X = (X_1, X_2, \ldots, X_K)$ tuân theo phân phối Dirichlet với tham số vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ với $\alpha_i > 0$ cho mọi i, ký hiệu $X \sim \text{Dirichlet}(\boldsymbol{\alpha})$, có hàm mật độ xác suất được cho bởi:

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

với $x_i > 0$, $\sum_{i=1}^{K} x_i = 1$.

Trong đó, $B(\boldsymbol{\alpha})$ là hàm Beta đa biến (multivariate beta function), được định nghĩa là:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$$

10.2 Hàm xác suất tích lũy (Cumulative Distribution Function - CDF)

Hàm xác suất tích lũy của phân phối Dirichlet không có dạng đóng đơn giản và phức tạp hơn nhiều so với các phân phối một biến. Thường không được sử dụng trực tiếp.

10.3 Các đặc trưng thống kê

- **Miền giá trị:** Simplex tiêu chuẩn

$$S_K = \left\{ (x_1, \ldots, x_K) \mid x_i > 0,\ \sum_{i=1}^{K} x_i = 1 \right\}.$$

- **Giá trị kỳ vọng (Mean):** Đối với từng thành phần X_i:

$$E[X_i] = \frac{\alpha_i}{\sum_{j=1}^{K} \alpha_j}.$$

- **Phương sai (Variance):** Đối với từng thành phần X_i:

$$\mathrm{Var}[X_i] = \frac{\alpha_i \left(\sum_{j=1}^{K} \alpha_j - \alpha_i \right)}{\left(\sum_{j=1}^{K} \alpha_j \right)^2 \left(\sum_{j=1}^{K} \alpha_j + 1 \right)}.$$

- **Mode:** Đối với từng thành phần X_i:

$$\mathrm{Mode}_i = \frac{\alpha_i - 1}{\sum_{j=1}^{K} \alpha_j - K}, \quad \text{nếu tất cả } \alpha_i > 1.$$

- **Median:** Không có dạng đóng.

- **Tính chất:** Phân phối Dirichlet là liên hợp tiên nghiệm cho phân phối Categorical và phân phối Đa thức.

10.4 Biểu đồ minh họa

Việc minh họa hàm mật độ xác suất của phân phối Dirichlet yêu cầu các biểu đồ trên simplex. Đối với $K = 3$, chúng ta có thể sử dụng biểu đồ tam giác.

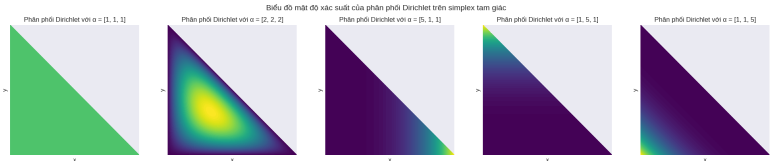

Figure 18: Hàm mật độ xác suất của Phân phối Dirichlet với $K = 3$ và các tham số khác nhau.

Biểu đồ minh hoạt hàm xuất sắc tích lũy của phân phối Dirichlet

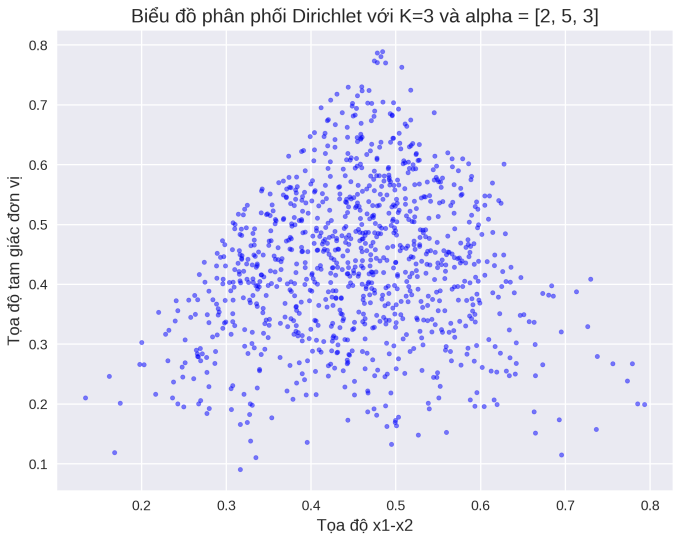

Figure 19: Hàm mật độ xác suất của Phân phối Dirichlet với $K = 3$ và các tham số khác nhau.

10.5 Ví dụ dữ liệu và bài toán thực tế

Phân phối tiên nghiệm cho phân loại văn bản Trong mô hình hóa chủ đề (topic modeling) như Latent Dirichlet Allocation (LDA), phân phối Dirichlet được sử dụng làm phân phối tiên nghiệm cho hai cấp độ:

- Phân phối các chủ đề trên một tài liệu (document-topic distribution).
- Phân phối các từ trên một chủ đề (topic-word distribution).

Ví dụ, Dirichlet($\boldsymbol{\alpha}$) với α_i nhỏ (ví dụ, $\alpha_i = 0.1$) sẽ khuyến khích các phân phối thưa, nghĩa là mỗi tài liệu có xu hướng chỉ nói về một vài chủ đề chính.

Phân phối tiên nghiệm cho xác suất đa thức Giả sử chúng ta có một thí nghiệm với K kết quả có thể xảy ra, mỗi kết quả có xác suất p_i. Vector $(p_1, \ldots, p_K)$ với $\sum p_i = 1$ có thể được mô hình hóa bằng phân phối Dirichlet. Ví dụ, tỷ lệ các loại sản phẩm khác nhau được bán trong một siêu thị.