Xử lý số liệu thống kê

Assignment 1

Nhóm E:

Trần Tiến Đạt
Nguyễn Thị Ngọc Anh
Nguyễn Thái Hưng Thịnh

Ngày nộp: October 6, 2025

Contents

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):

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

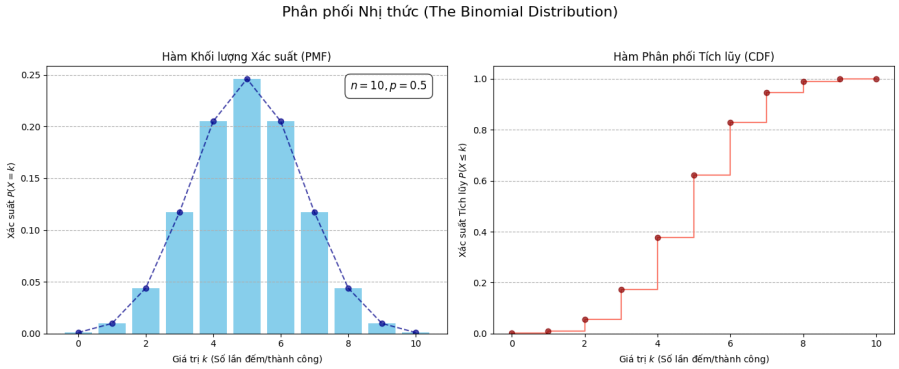

Figure 1: Biểu đồ Hàm Khối lượng Xác suất (PMF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Nhị thức. Phân phối này mô tả số lần thành công trong n phép thử độc lập.

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 $n \to \infty$.

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:

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

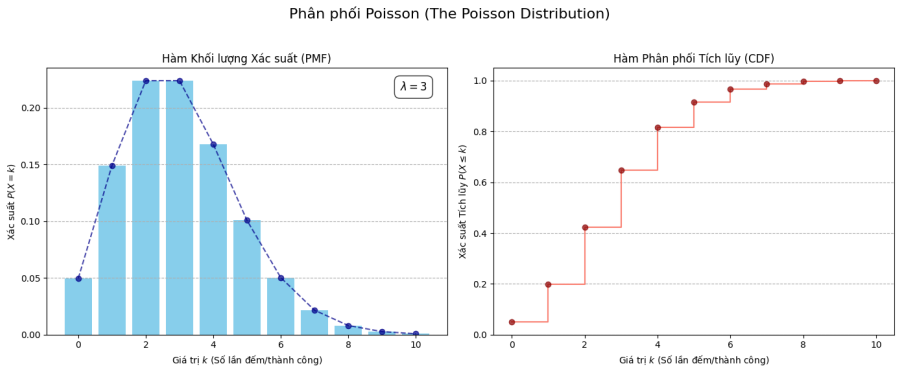

Figure 2: Biểu đồ PMF và CDF của Phân phối Poisson. Phân phối này mô hình hóa số sự kiện xảy ra trong một khoảng thời gian/không gian cố định, với tốc độ trung bình (λ) đã biết.

2.4 Các đặc trưng thống kê

- **Kỳ vọng (Mean):**
$$\mathbb{E}[X] = \lambda$$
- **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.

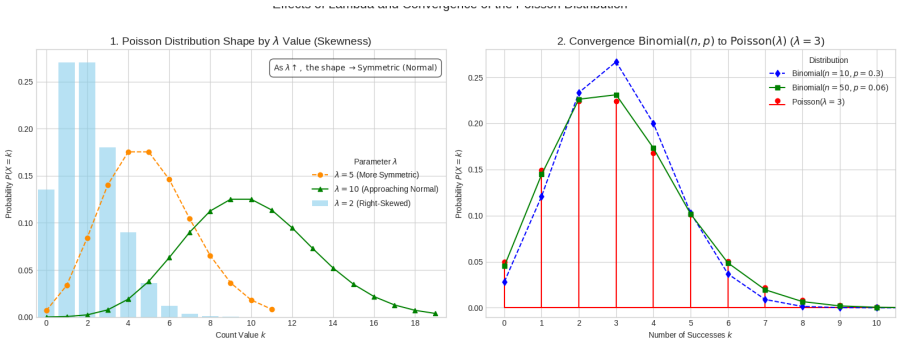

Figure 3: Hình dạng của phân phối Possion với các tham số khác nhau.

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. Mà ở đây ta đề cập đến việc sử dụng để mô hình hóa Tần suất Nhấp chuột (Click Frequency)

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:

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

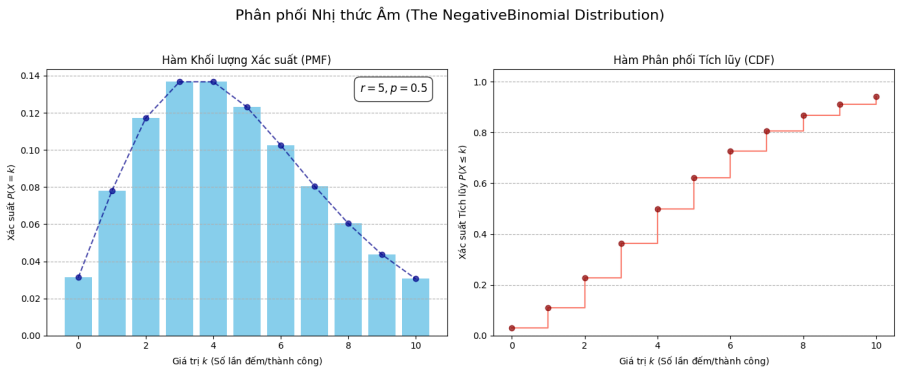

Figure 4: Biểu đồ Hàm Khối lượng Xác suất (PMF) và Hàm Phân phối Tích lũy (CDF) của Phân phối Nhị thức Âm. Phân phối này mô tả số lần thất bại trước khi đạt r thành công cố định.

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).

- **Mô hình hóa Overdispersion:** Phân phối Nhị thức Âm thường được xây dựng bằng cách giả định rằng λ (tốc độ trung bình) của phân phối Poisson không phải là hằng số mà tuân theo một **Phân phối Gamma**. Phân phối hỗn hợp (Compound Distribution) này tạo ra Phân phối Nhị thức Âm (Gamma-Poisson Mixture).

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 để 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). Dữ liệu giao dịch lịch sử của mỗi khách hàng, bao gồm:

- **ID Khách hàng:** Nhận dạng duy nhất.

- **Số lần Mua hàng (k):** Tổng số giao dịch trong một khoảng thời gian quan sát.

- **Các biến giải thích (Covariates):** Các đặc điểm của khách hàng ảnh hưởng đến tần suất mua (ví dụ: Giá trị đơn hàng trung bình, thời gian từ lần mua cuối cùng, nguồn gốc khách hàng).

Vấn đề của Phân phối Poisson

Khi cố gắng giải quyết bài toán này bằng **Hồi quy Poisson (Poisson Regression)**, họ thường gặp phải hiện tượng **Phân tán Quá mức (Overdispersion)**. Nguyên nhân có thể do:

- Giả định Poisson Yêu cầu Mean(μ) = Variance(σ^2).
- Chênh lệch giữa người "mua sắm nhiều" và nhóm người "thi thoảng mua sắm" làm cho phương sai của dữ liệu lớn hơn rất nhiều so với trung bình

Giải pháp: Sử dụng Phân phối Nhị thức Âm

Phân phối Nhị thức Âm là giải pháp hoàn hảo vì nó có khả năng mô hình hóa sự phân tán quá mức. Đây có thể được coi là sự kết hợp của hai phân phối:

- **Phân phối Poisson:** Mô tả số lần mua hàng của một cá nhân tại một tốc độ λ nhất định.
- **Phân phối Gamma:** Mô tả sự thay đổi của tốc độ λ giữa các cá nhân (tức là sự không đồng nhất trong hành vi mua hàng).

Bằng cách tích hợp (compound) hai phân phối này, Phân phối nhị thức âm (gọi tắc là NBD) có thêm **Tham số Phân tán** (thường ký hiệu là k hoặc α) cho phép Phương sai lớn hơn Trung bình:

$$\text{Var}(X) = \mu + \frac{\mu^2}{k}$$

(Trong đó k là tham số phân tán. Khi $k \to \infty$, Phương sai $\to \mu$, và NBD hội tụ về Poisson).

4 methods

Đây là phần giưuax

5 Conclusion

Đây là phần kết của tài liệu [Einstein, 1905].

References

[Einstein, 1905] Einstein, A. (1905). On the electrodynamics of moving bodies. *Annalen der Physik*, 17:891–921.