Lý thuyết xác suất

1. Một số khái niệm cơ bản trong xác suất

1.1. Biến ngẫu nhiên

Một ví dụ thực tế thường được dùng để giải thích các khái niệm trong xác suất là ví dụ tung xúc xắc, ở đây ta sẽ sử dụng ví dụ này để minh hoạ các khái niệm.

Biến ngẫu nhiên (random variable) là đại lượng dùng để đại diện cho những giá trị ngẫu nhiên không xác định. Trong ví dụ về tung xúc xắc, biến ngẫu nhiên là số chấm thu được ở mặt trên của xúc xắc.

Không gian mẫu (sample space) là tập hợp tất cả các giá trị mà một biến ngẫu nhiên có thể nhận. Trong ví dụ về tung xúc xắc, không gian mẫu là tập hợp tất cả các khả năng mà mặt trên của xúc xắc có thể nhận, cụ thể là 1 chấm, 2 chấm, 3 chấm, 4 chấm, 5 chấm, và 6 chấm.

Biến cố (event) là một sự kiện xảy ra giúp ta nhận được một kết quả của biến ngẫu nhiên. Trong ví dụ về tung xúc xắc, biến cố là sự kiện ta tung xúc xắc.

Kết quả (outcome) là giá trị mà biến ngẫu nhiên nhận được sau khi biến cố xảy ra. Trong ví dụ về tung xúc xắc, sau khi tung xúc xắc, ta thu được kết quả là 1 chấm chả hạn.

Khi ta thực hiện lặp đi lặp lại nhiều lần, ta có thể thu được nhiều kết quả khác nhau đối với cùng biến ngẫu nhiên, sẽ có những kết quả xuất hiện nhiều lần hơn các kết quả khác, sẽ có những kết quả xuất hiện ít lần hơn các lần khác, sẽ có những kết quả có số lần xuất hiện xấp xỉ nhau. Thông tin về đầu ra được đo bởi phân phối xác suất (probability distribution)

$$p(x)$$

của biến ngẫu nhiên.

1.2. Biến ngẫu nhiên rời rạc và biến ngẫu nhiên liên tục

Một biến ngẫu nhiên có thể là rời rạc (discrete) hoặc liên tục (continuous).

Trong ví dụ trên về việc tung xúc xắc, ta đang xét đến một biến ngẫu nhiên rời rạc. Một biến ngẫu nhiên rời rạc sẽ lấy giá trị trong một tập hợp cho trước, tập không gian mẫu ở ví dụ trên là {1 chấm, 2 chấm, 3 chấm, 4 chấm, 5 chấm, và 6 chấm}. Mỗi kết quả đầu ra sẽ có một giá trị xác suất tương ứng, giá trị xác suất này không âm và có tổng bằng 1.

$$\sum_x p(x) = 1$$

Khác với biến ngẫu nhiên rời rạc, biến ngẫu nhiên liên tục lấy giá trị trong một khoảng giá trị cho trước, hay nói cách khác, biến ngẫu nhiên liên tục lấy giá trị nằm trong các tập hợp con của tập số thực. Khoảng giá trị này có thể là hữu hạn (ví dụ: thời gian làm bài thi của một học sinh, nằm trong khoảng từ 0 phút đến 180 phút) hoặc là vô hạn (ví dụ: thời gian di chuyển từ điểm A đến điểm B của một chiếc xe). Xác suất để nhận biến ngẫu nhiên liên tục nhận đầu ra chính xác bằng một giá trị nào đó thường được coi là bằng 0. Thay vào đó, ta sẽ tính toán xác suất để biến ngẫu nhiên nhận đầu ra nằm trong một khoảng giá trị nào đó, và được mô tả bởi hàm mật đô xác suất (probability density function). Hàm mật độ xác suất luôn cho giá trị dương, và tích phân của nó trên toàn miền phải bằng 1.

$$\int p(x)dx = 1$$

Nhằm đơn giản hoá phần ký hiệu, hàm mật độ xác suất của một biến ngẫu nhiên liên tục x cũng được ký hiệu là $p(x)$.

2. Xác suất đồng thời, xác suất biên và xác suất điều kiện

Trong thực tế, thông thường ta không chỉ quan sát một biến ngẫu nhiên, ta thường quan sát cùng lúc hai hoặc nhiều biến ngẫu nhiên khác nhau, từ đó, xuất hiện các khái niệm mới về xác suất khi quan sát nhiều biến ngẫu nhiên.

2.1. Xác suất đồng thời (Joint probability)

Xét hai biến ngẫu nhiên x và y, khi ta quan sát hai biến ngẫu nhiên này, ta sẽ thấy những cặp giá trị x và y xuất hiện nhiều hơn các cặp giá trị khác, hoặc những cặp giá trị x và y xuất hiện ít hơn các cặp giá trị khác. Điều này được biển diễn thông qua xác suất xảy ra đồng thời của x và y, ký hiệu là $p(x, y)$.

Ví dụ: Ta xét x là biến ngẫu nhiên về học lực của học sinh, y là biến ngẫu nhiên về hạnh kiểm của học sinh. Ta có, $p(x = \text{học lực giỏi}, y = \text{hạnh kiểm khá})$ là xác suất mà học sinh vừa nhận học lực giỏi vừa nhận hạnh kiểm khá.

Biến ngẫu nhiên x và y có thể gồm hai biến ngẫu nhiên rời rạc, hai biến ngẫu nhiên liên tục hoặc một biến ngẫu nhiên rời rạc một biến ngẫu nhiên liên tục.

Với hai biến ngẫu nhiên rời rạc:

$$\sum_{x,y} p(x, y) = 1$$

Với hai biến ngẫu nhiên liên tục:

$$\int p(x, y)dxdy = 1$$

Với x là biến ngẫu nhiên rời rạc, y là biến ngẫu nhiên liên tục:

$$\sum_x \int p(x, y)dy = \int \left(\sum_x p(x, y)\right) dy = 1$$

2.2. Xác suất biên (Marginal probability)

Từ xác suất đồng thời của nhiều biến ngẫu nhiên, ta có thể xác định được phân bố xác suất của từng biến bằng cách lấy tổng (với biến ngẫu nhiên rời rạc) hoặc tích phân (với biến ngẫu nhiên liên tục) theo tất cả các biến còn lại. Quá trình này được gọi là marginalization, xác suất mà ta thu được từ quá trình này được gọi là marginal probability.

Với hai biến ngẫu nhiên rời rạc:

$$p(x) = \sum_y p(x, y), p(y) = \sum_x p(x, y)$$

Với hai biến ngẫu nhiên liên tục:

$$p(x) = \int p(x,y)dy, p(y) = \int p(x,y)dx$$

Với nhiều biến ngẫu nhiên rời rạc:

$$p(x) = \sum_{y,z,w} p(x,y,z,w)$$

$$p(x,y) = \sum_{z,w} p(x,y,z,w)$$

Trong một số tài liệu, để đơn giản hoá việc ký hiệu toán học, mặc dù sử dụng biến ngẫu nhiên liên tục, nhưng ta vẫn có thể sử dụng ký hiệu $\sum$ thay vì $\int$.

2.3. Xác suất điều kiện (Conditional probability)

Xác suất điều kiện là xác suất của một (hoặc nhiều) biến ngẫu nhiên với điều kiện một (hoặc nhiều) biến ngẫu nhiên khác nhận giá trị nào đó cụ thể. Xác suất có điều kiện của biến ngẫu nhiên x biết rằng biến ngẫu nhiên y có giá trị $y*$ được ký hiệu là $p(x|y = y^*)$.

Ta có thể tính $p(x|y = y^*)$ thông qua xác suất đồng thời như sau:

$$p(x|y = y^*) = \frac{p(x, y = y^*)}{\sum_x p(x, y = y^*)} = \frac{p(x, y = y^*)}{p(y = y^*)}$$

trong đó:

- $p(x, y = y^*)$ là xác suất đồng thời của mỗi giá trị của biến ngẫu nhiên x với giá trị của biến ngẫu nhiên $y = y^*$.
- $\sum_x p(x, y = y^*) = p(y = y^*)$ là xác suất biên của biến ngẫu nhiên y.

Ta có thể viết gọn công thức trên bằng việc không cần chỉ rõ giá trị $y = y^*$:

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

tương tự, ta có:

$$p(y|x) = \frac{p(y,x)}{p(x)}$$

Từ hai công thức rút gọn trên, ta có mối quan hệ giữa xác suất điều kiện, xác suất biên và xác suất đồng thời:

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

2.4. Quy tắc Bayes (Bayes' rule)

Từ công thức trên, ta có

$$p(y|x)p(x) = p(x|y)p(y)$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

Đây được gọi là Quy tắc Bayes, được sử dụng rất nhiều trong Machine Learning.

3. Một số phân phối xác suất phổ biến

3.1. Phân phối Bernoulli (Bernoulli distribution)

Phân phối Bernoulli là một phân phối rời rạc mô tả biến ngẫu nhiên nhị phân: nó mô tả trường hợp khi đầu ra chỉ nhận một trong hai giá trị $x \in 0, 1$.

Phân phối Bernoulli được mô tả bằng một tham số $\lambda \in [0, 1]$ và là xác suất để $x = 1$, ký hiệu là $p(x) = \text{Bern}_x[\lambda]$. Phân phối của mỗi giá trị đầu ra là:

- $p(x = 1) = \lambda$
- $p(x = 0) = 1 - p(x = 1) = 1 - \lambda$

Hai đẳng thức này thường được viết gọn lại trở thành hàm mật độ xác suất:

$$p(x) = \lambda^x(1 - \lambda)^{1-x}$$

Phân phối này được áp dụng trong các bài toán phân loại nhị phân như fraud/normal transaction, human/non-human in scence ...

3.2. Phân phối phân loại (Categorical distribution)

Phân phối phân loại là dạng tổng quát của phân phối Bernoulli. Thay vì biến ngẫu nhiên nhị phân của phân phối Bernoulli, biến ngẫu nhiên trong phân phối phân loại có thể nhận một trong nhiều giá trị từ một tập hợp $\{1, 2, \ldots, K\}$.

Nếu có K giá trị đầu ra có thể đạt được, phân phối phân loại sẽ được mô tả bởi K tham số và viết dưới dạng vector: $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_K]$ với các λ_K không âm và có tổng bằng 1. Mỗi giá trị λ_K thể hiện xác suất để biến ngẫu nhiên nhận giá trị K, ký hiệu là $p(x) = \text{Cat}_x[\lambda]$.

Trong thực tế, ta thường ký hiệu tập hợp các giá trị đầu ra $\{1, 2, \ldots, K\}$ dưới dạng one-hot $\mathbf{x} \in \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_K\}$ với $\mathbf{e}_K$ là vector gồm K phần tử, trong đó phần từ thứ K bằng 1 còn các phần tử còn lại bằng 0.

Từ đó, xác suất để biến ngẫu nhiên x nhận giá trị $\mathbf{e}_k$ được xác nhận như sau:

$$p(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k$$

Tích này sẽ gồm các thừa số $\lambda_j^{x_j} = 1$ ở các vị trí $x_j = 0, \forall j \neq k$ và thừa số $\lambda_j^{x_j} = \lambda_k$ ở vị trí $x_j = 1, j = k$.

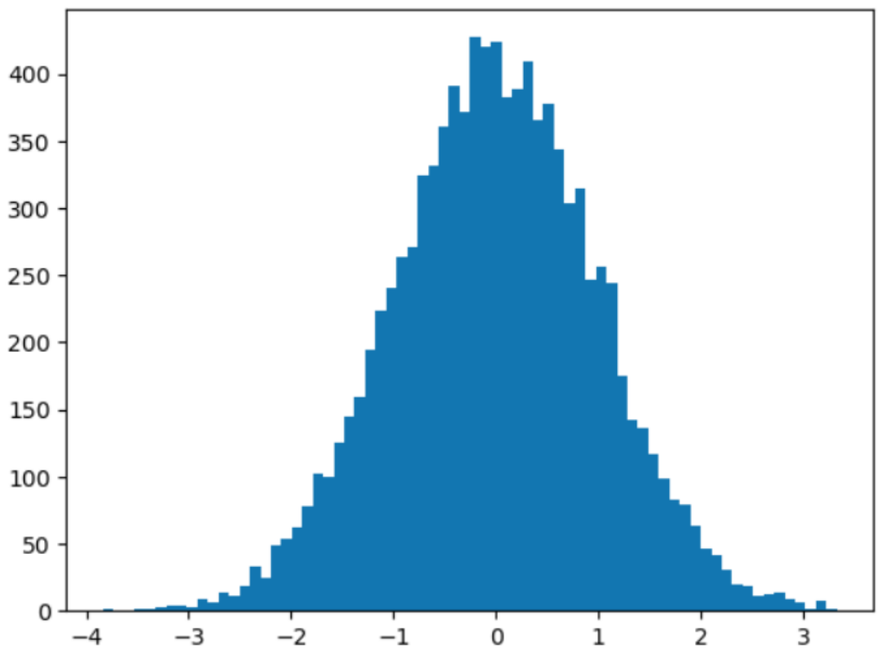
3.3. Phân phối chuẩn một biến (Univariate normal distribution)

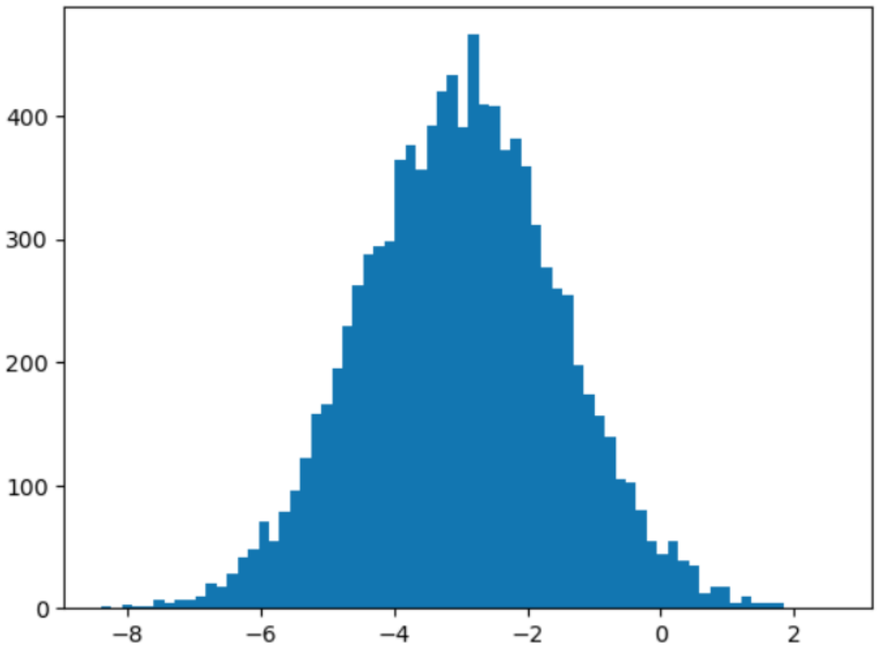
Phân phối chuẩn là phân phối mô tả biến ngẫu nhiên liên tục nhận giá trị $x \in (-\infty, \infty)$

Phân phối chuẩn được mô tả bởi hai tham số: kỳ vọng (mean) μ và phương sai (variance) σ^2 (hoặc độ lệch chuẩn (standard deviation) σ), được ký hiệu là $p(x) = \text{Norm}_x[\mu, \sigma^2]$. Trong đó, μ thể hiện vị trí đỉnh, nơi có mật độ xác suất cao nhất, σ thể hiện độ rộng của phân phối, σ lớn đồng nghĩa với phân phối có đầu ra biến đổi mạnh, σ nhỏ đồng nghĩa với phân phối có đầu ra ổn định.

Hàm mật độ xác suất của phân phối chuẩn một biến là:

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





3.4. Phân phối chuẩn nhiều biến (Multivariate normal distribution)

Phân phối chuẩn nhiều biến là dạng tổng quát của phân phối chuẩn một biến, được sử dụng để mô tả biến ngẫu nhiên liên tục nhiều chiều.

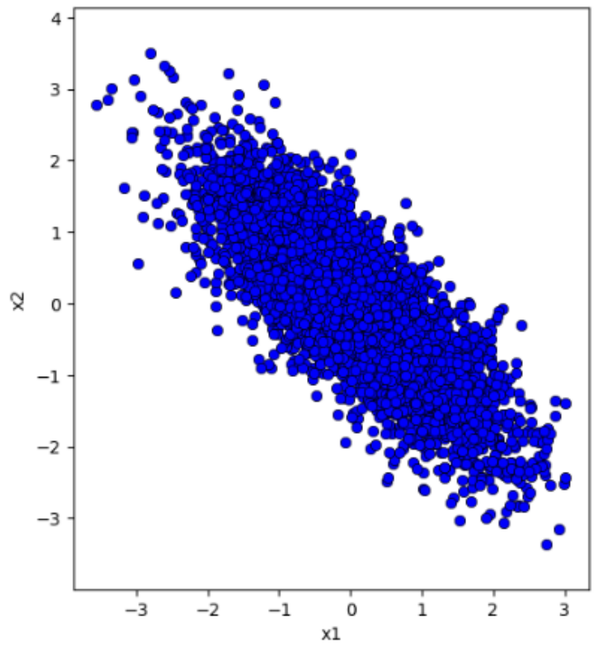
Giả sử ta xét biến ngẫu nhiên D chiều, phân phối chuẩn nhiều biến được mô tả bởi hai tham số vector kỳ vọng (mean vector) $\mu \in R^D$ và ma trận hiệp phương sai (covariance matrix) $\Sigma \in \mathbb{S}^D_{++}$ là một ma trận đối xứng xác định dương. Hàm mật độ xác suất có dạng:

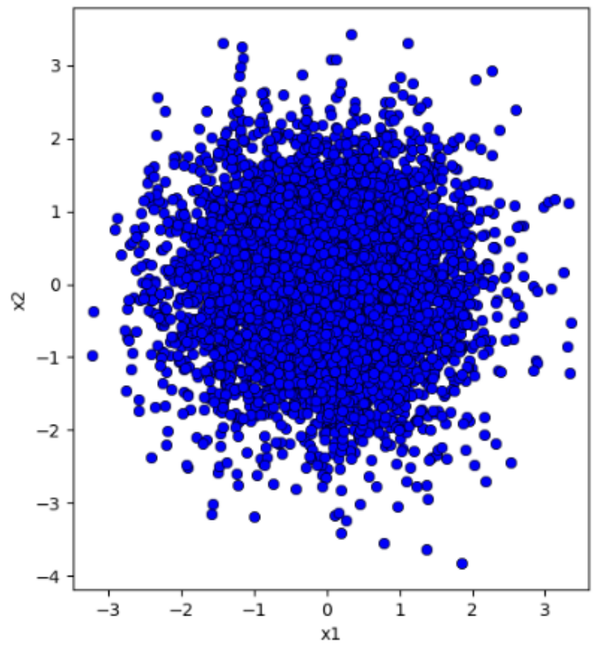
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right)$$

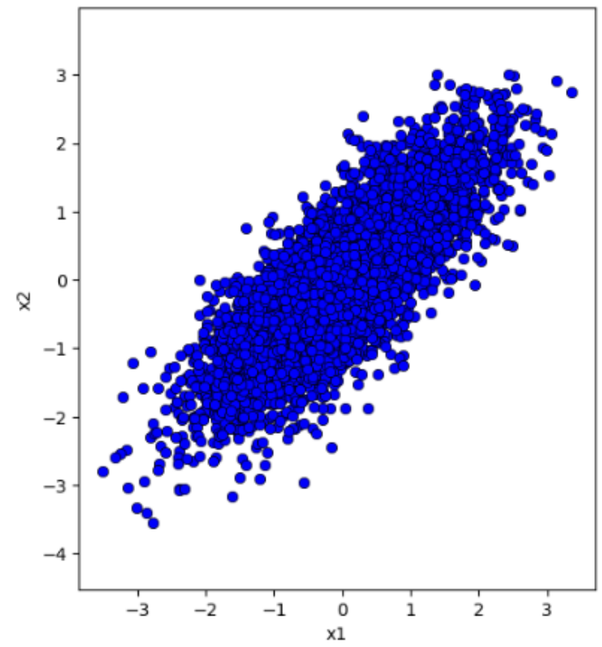
Vì sao ma trận hiệp phương sai lại là ma trận đối xứng xác định dương? Ma trận hiệp phương sai là một ma trận vuông, trong đó:

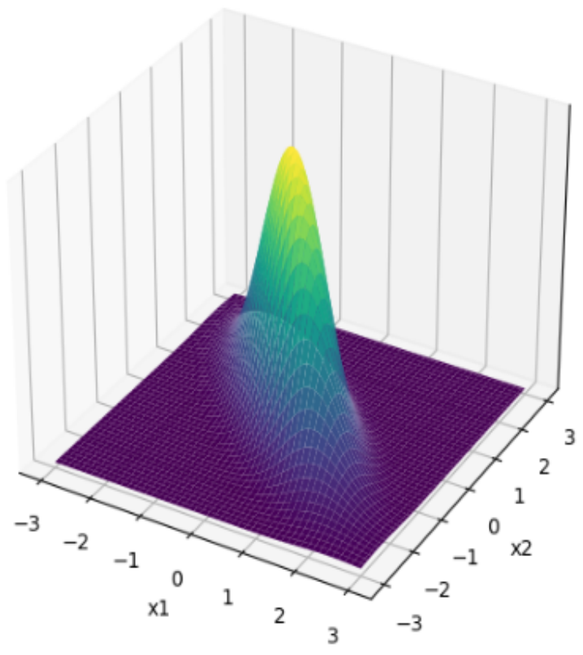
- các phần tử nằm trên đường chéo chính lần lượt là phương sai của từng biến.
- các phần từ còn lại (không nằm trên đường chéo) là các hiệp phương sai của đôi một hai biến ngẫu nhiên khác nhau trong tập hợp. Do đó, ma trận hiệp phương sai luôn đối xứng qua đường chéo chính và các phần từ trên đường chéo chính luôn dương nên các trị riêng chính của chúng luôn dương và ma trận xác định dương.

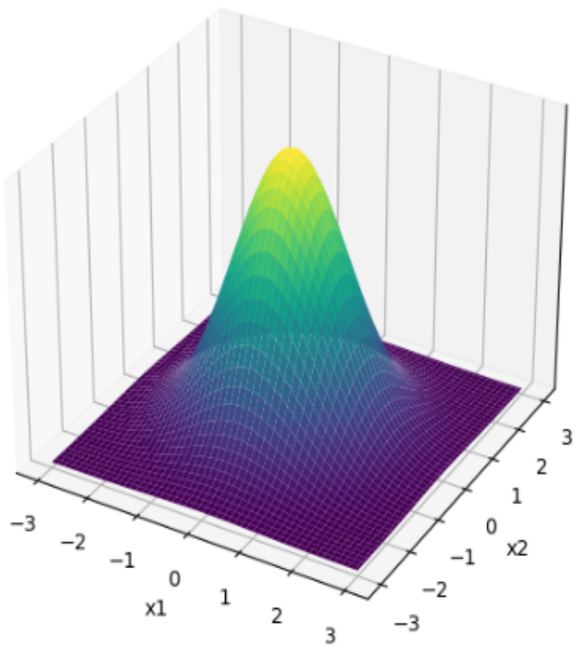
Hiệp phương sai là độ đo sự biến thiên cùng nhau của hai biến ngẫu nhiên (phân biệt với phương sai - đo mức độ biến thiên của một biến). Giá trị hiệp phương sai nằm trong khoảng từ $(-\infty, \infty)$ trong đó giá trị dương biểu thị rằng cả hai biến chuyển động theo cùng một hướng, giá trị âm biểu thị rằng cả hai biến chuyển động ngược chiều nhau, giá trị bằng không biểu thị hai biến không có tương quan với nhau.

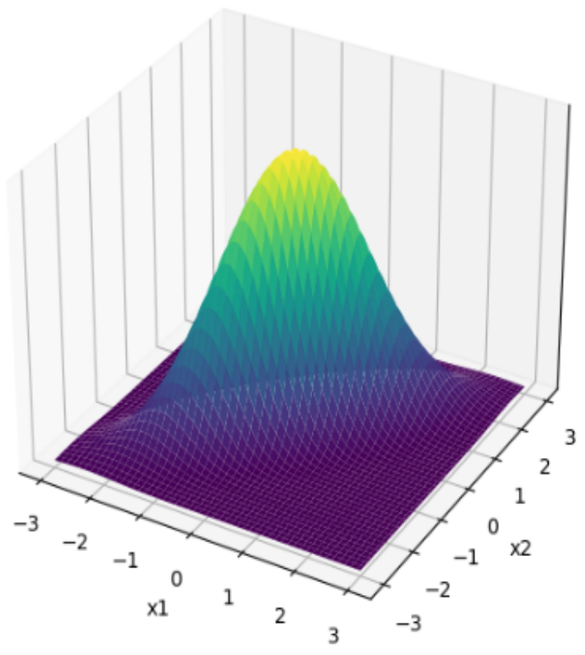






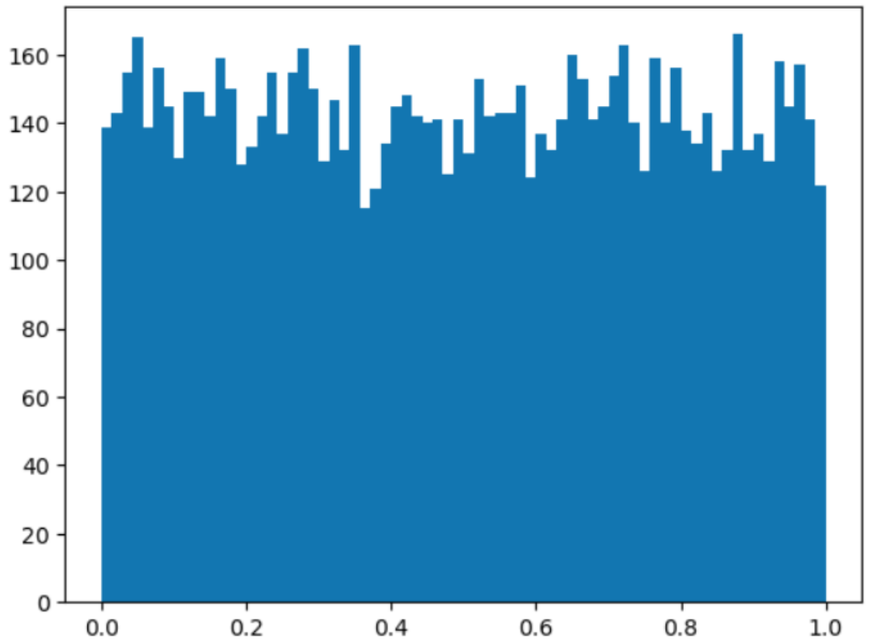






3.5. Phân phối đều (Uniform distribution)

Trái ngược với phân phối chuẩn, nơi mà có những giá trị có xác suất cao hơn các giá trị khác, phân phối đều được sử dụng để mô tả biến ngẫu nhiên liên tục có xác suất nhận các giá trị kết quả là như nhau.



Phân phối đều thường được sử dụng trong việc tạo mẫu dữ liệu trong quá trình thống kê và phân tích dữ liệu.