THỐNG KÊ MÁY TÍNH
(Computational Statistics)

Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giảng viên: TS.Nguyễn Khắc Cường

CHƯƠNG 5

BIẾN NGẪU NHIÊN
VÀ
PHÂN PHỐI XÁC SUẤT

5.1.Biến ngẫu nhiên (Random Variable)

- 5.1.1.Giới thiệu
  - Là qui tắc/cách biểu diễn các kết quả của phép thử ngẫu nhiên nào đó dưới dạng số
- 5.1.2.Định nghĩa toán:
  - Biến ngẫu nhiên X là một ánh xạ từ không gian các biến cố sơ cấp Ω vào tập số thực R

$$X : \Omega \rightarrow \Re \qquad \omega \mapsto X(\omega)$$

- 5.1.3.VD:
  - Thực hiện phép thử "thảy đồng thời 3 đồng xu cân đối, đồng chất" → các biến cố sơ cấp là:

ϖ_1=(HHH)	ϖ_2=(HHT)	ϖ_3=(HTT)	ϖ_4=(HTH)
ϖ_5=(TTT)	ϖ_6=(TTH)	ϖ_7=(THH)	ϖ_8=(THT)

5.1.Biến ngẫu nhiên (Random Variable)

ϖ_1=(HHH)	ϖ_2=(HHT)	ϖ_3=(HTT)	ϖ_4=(HTH)
ϖ_5=(TTT)	ϖ_6=(TTH)	ϖ_7=(THH)	ϖ_8=(THT)

- 5.1.3.VD:
  - Gọi X là biến cố: "cho biết số mặt T xảy ra khi thực hiện phép thử giao 3 đồng xu" → các giá trị của các phần tử của biến ngẫu nhiên X là:

$X(\varpi_1)=0$	$X(\varpi_2)=1$	$X(\varpi_3)=2$	$X(\varpi_4)=1$
$X(\varpi_5)=3$	$X(\varpi_6)=2$	$X(\varpi_7)=1$	$X(\varpi_8)=2$

5.2.Phân loại biến ngẫu nhiên

- 5.2.1.Cơ sở để phân loại
  - Giá trị mà biến ngẫu nhiên nhận được
- 5.2.2.Phân loại:
  - biến ngẫu nhiên rời rạc (discrete random variable)
    - Tập giá trị nhận được là hữu hạn / vô hạn đếm được
    - VD: số sản phẩm lỗi, số bit
  - biến ngẫu nhiên liên tục (continuous random variable)
    - Tập giá trị nhận được là một khoảng (a, b), (a, b], [a, b), [a, b], hoặc toàn bộ R
    - VD: nhiệt độ, thời gian, độ dài, …

5.3.Phân phối xác suất

- 5.3.1.Qui luật phân phối xác suất
  - biểu thức biểu diễn mối quan hệ
    - giữa các giá trị có thể có của biến ngẫu nhiên
    - với xác suất tương ứng của các giá trị đó
- 5.3.2.Hàm phân phối xác suất (CDF - Cumulative distribution function) của biến ngẫu nhiên X
  - xác định trên không gian các biến cố sơ cấp
  - là hàm F(x) được định nghĩa:

$$F(x) = P(X \le x), \forall x \in (-\infty, +\infty)$$

5.3.Phân phối xác suất

- 5.3.3.Phân phối xác suất của biến ngẫu nhiên rời rạc
  - Hàm giá trị xác suất (PMF - Probability Mass Function)
    - gọi tắt là hàm xác suất PMF
    - PMF của một biến ngẫu nhiên rời rạc X có thể nhận các giá trị x_1, x_2, . . . , x_n là hàm thỏa:

$$f(x_i) \geq 0, \forall i = \overline{1,n}$$

$$\sum_{i=1}^{n} f(x_i) = 1$$

$$f(x_i) = P(X = x_i), \forall i = \overline{1,n}$$

  - Bảng phân phối xác suất (của biến ngẫu nhiên X)

X	x_1	x_2	..	x_n	..
P	$f(x_1)$	$f(x_2)$	..	$f(x_n)$	..

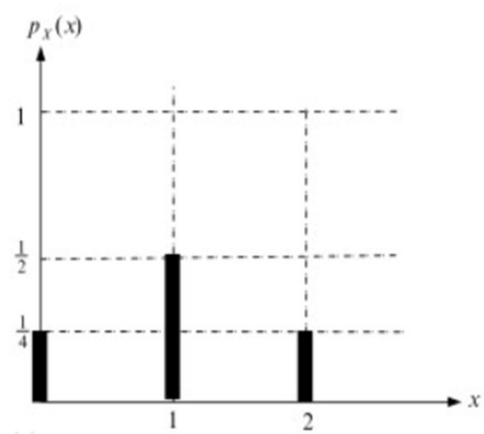
5.3.Phân phối xác suất

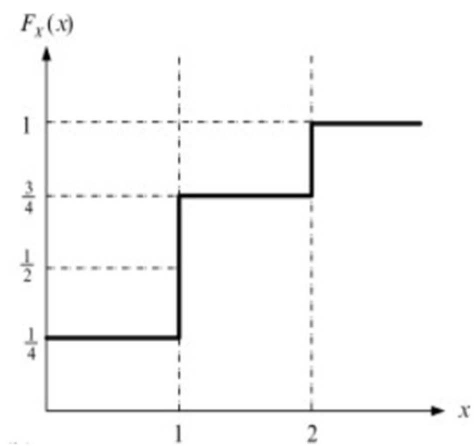
- 5.3.3.Phân phối xác suất của biến ngẫu nhiên rời rạc
  - Hàm phân phối xác suất CDF (*của biến ngẫu nhiên rời rạc*)
    - Là hàm được định nghĩa:

$$F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$$

    - Cụ thể:

$$F(x) = P(X \le x) = \begin{cases} 0, x < x_1 \\ f(x_1), x_1 \le x < x_2 \\ f(x_1) + f(x_2), x_2 \le x < x_3 \\ f(x_1) + f(x_2) + f(x_3), x_3 \le x < x_4 \\ .. \\ f(x_1) + .. + f(x_{n-1}), x_{n-1} \le x < x_n \\ 1, x \ge x_n \end{cases}$$





5.3.Phân phối xác suất

- 5.3.4.Phân phối xác suất của biến ngẫu nhiên liên tục
  - Hàm mật độ xác suất (PDF - Probability Density Function)
    - PDF của biến ngẫu nhiên liên tục X được định nghĩa là hàm f(x) thỏa

$$P(X \in I) = \int_I f(x)d(x), \forall I \subset \Re$$

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

5.4.Các đặc trưng số của biến ngẫu nhiên

- 5.4.1.Kỳ vọng (Expectation) của biến ngẫu nhiên
  - Kỳ vọng của biến ngẫu nhiên rời rạc

$$E(X)=\sum_{i=1}^{+\infty} x_i P(X=x_i)=\sum_{i=1}^{+\infty} x_i f(x_i)$$

$$E(X)=\sum_{x\in S} xf(x)$$

Trong đó:

X	x_1	x_2	..	x_n	..
P	$f(x_1)$	$f(x_2)$	..	$f(x_n)$	..

S : không gian mẫu

5.4.Các đặc trưng số của biến ngẫu nhiên

- 5.4.1.Kỳ vọng (Expectation) của biến ngẫu nhiên
  - Kỳ vọng của biến ngẫu nhiên liên tục

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

trong đó:

f(x) là hàm mật độ xác suất (PDF) của biến ngẫu nhiên liên tục X

  - Ý nghĩa của E(X)
    - Là giá trị trung bình theo xác suất của tất cả giá trị có thể có của biến ngẫu nhiên
    - Là giá trung bình của phân phối xác suất.

5.4.Các đặc trưng số của biến ngẫu nhiên

- 5.4.2.Phương sai (Variance) của biến ngẫu nhiên

$$Var(X) = E(X - E(X))^2$$

$$Var(X) = E(X^2) - (E(X))^2$$

- 5.4.3.Độ lệch chuẩn (Standard deviation)

$$\sigma(X) = \sqrt{Var(X)}$$

- Ý nghĩa của phương sai:
  - là trung bình bình phương sai lệch
  - phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình
  - Ứng dụng:
    - Trong công nghiệp: phương sai biểu thị độ chính xác trong sản xuất
    - Trong canh tác: phương sai biểu thị mức độ ổn định của năng suất
    - Trong đo lường: phương sai thể hiện độ "ổn định" của phép đo
    - . . .

5.4.Các đặc trưng số của biến ngẫu nhiên

- Tính các đặc trưng trong R (tính thủ công)
  - Bảng phân phối xác suất của biến ngẫu nhiên rời rạc X

```
> X<-c(0,1,2,3)
> f_X<-c(1/8,3/8,3/8,1/8)
```

  - Kỳ vọng

```
> E_X<-sum(X*f_X)
> E_X
[1] 1.5
```

  - Phương sai

```
> Var_X<-sum((X-E_X)^2*f_X)
> Var_X
[1] 0.75
```

  - Độ lệch chuẩn

```
> std_X<-sqrt(Var_X)
```

  - Phân phối xác suất CDF

```
> CDF_X<-cumsum(f_X)
> CDF_X
[1] 0.125 0.500 0.875 1.000
```

5.4.Các đặc trưng số của biến ngẫu nhiên

- Tính các đặc trưng trong R (dùng gói distrEx)
  - Cài đặt gói distrEx

```
> install.packages("distrEx")
> library(distrEx)
```

  - Tính các đặc trưng của biến ngẫu nhiên rời rạc X

```
> X<-DiscreteDistribution(supp=0:3,prob=c(1,3,3,1)/8)
> E(X)
[1] 1.5
> E(X)
[1] 1.5
> var(X)
[1] 0.75
> sd(X)
[1] 0.8660254
```

5.5.Ứng dụng

- Bài toán
  - Một cửa hàng thống kê số lượng xe bán trong 500 ngày:
    - a. Gọi X là biến ngẫu nhiên chỉ số xe bán trong một ngày. Hãy lập bảng phân phối xác suất cho biến ngẫu nhiên X
    - b. Tính kỳ vọng của X (số xe hy vọng bán được trong một ngày)
    - c. Tính độ lệch chuẩn
    - d. Tính xác suất để trong 1 ngày:
      - 1) Có ít hơn 4 xe được bán
      - 2) Ít nhất 4 xe được bán
      - 3) Nhiều hơn 4 xe được bán
      - 4) Tối đa 4 xe được bán
      - 5) Đúng 4 xe được bán
      - . . .

Số xe máy bán trong 1 ngày	Tần số
0	40
1	100
2	142
3	66
4	36
5	30
6	26
7	20
8	16
9	14
10	8
11	2

5.5.Ứng dụng

- Bài toán
  - a. Bảng phân phối xác suất của X

Số xe máy bán trong 1 ngày	Tần số
0	40
1	100
2	142
3	66
4	36
5	30
6	26
7	20
8	16
9	14
10	8
11	2

$X=x_i$	Tần số	Xác suất
0	40	0.080
1	100	0.200
2	142	0.284
3	66	0.132
4	36	0.072
5	30	0.060
6	26	0.052
7	20	0.040
8	16	0.032
9	14	0.028
10	8	0.016
11	2	0.004
500		1.000

= 40/500

5.5.Ứng dụng

Số xe máy bán trong 1 ngày	Tần số
0	40
1	100
2	142
3	66
4	36
5	30
6	26
7	20
8	16
9	14
10	8
11	2

- Bài toán
  - b. Kỳ vọng của X
    - E(X) = số xe hy vọng bán được trong 1 ngày

$$E(X) = \sum_{i=0}^{11} x_i p_i = 3.056$$

  - c. Độ lệch chuẩn

$$\sigma^2 = E((X - E(X))^2) = \sum_{i=0}^{11} (x_i - E(X))^2 p_i = 6.0689$$

$$\sigma = \sqrt{\sigma^2} = 3.056$$

5.5.Ứng dụng

Số xe máy bán trong 1 ngày	Tần số
0	40
1	100
2	142
3	66
4	36
5	30
6	26
7	20
8	16
9	14
10	8
11	2

- Bài toán
  - d. Tính xác suất để trong 1 ngày
    - có ít hơn 4 xe được bán

      $$P[X<4]=P[X=0]+P[X=1]+P[X=2]+P[X=3]=0.696$$
    - tối đa 4 xe được bán

      $$P[X\leq 4]=P[X=0]+P[X=1]+P[X=2]+P[X=3]+P[X=4]=0.768$$
    - Ít nhất 4 xe được bán

      $$P[X\geq 4]=1-P[X<4]=1-0.696=0.304$$
    - Đúng 4 xe được bán

      $$P[X=4]=0.072$$
    - Nhiều hơn 4 xe được bán

      $$P[X>4]=1-P[\leq 4]=1-0.768=0.232$$

Q / A