Mục lục

Chương 5

Mẫu và phân bố mẫu

5.1 Mẫu ngẫu nhiên đơn giản

Giả sử X là đặc trưng của một tập hợp. Để biết thông tin về X ta lần lượt lấy ngẫu nhiên từng phần tử của tập hợp (có hoàn lại) và đo đặc trưng X của phần tử đó. Nếu thực hiện n lần thì được dãy số liệu $(X_1, X_2, \dots, X_n)$ gọi là mẫu ngẫu nhiên đơn giản rút từ X.

$X_1, X_2, \dots$ là dãy đại lượng ngẫu nhiên độc lập, cùng phân bố với X; n gọi là cỡ mẫu.

Với mỗi thí nghiệm cụ thể, ta có một mẫu cụ thể $(x_1, x_2, \dots, x_n)$ là một giá trị cụ thể của véctơ ngẫu nhiên $(X_1, X_2, \dots, X_n)$.

5.2 Các đặc trưng mẫu

5.2.1 Kỳ vọng mẫu

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{5.1}$$

5.2.2 Phương sai mẫu

$$\begin{aligned} S^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \overline{X}^2. \end{aligned} \tag{5.2}$$

5.2.3 Phương sai mẫu điều chỉnh

$$\begin{aligned} S'^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \\ &= \frac{n}{n-1}S^2. \end{aligned} \tag{5.3}$$

S' gọi là độ lệch mẫu điều chỉnh hay độ phân tán tiêu chuẩn của mẫu.

Chú ý: a) Với mẫu cụ thể $(x_1, x_2, \dots, x_n)$, ta có giá trị cụ thể tương ứng của $\overline{X}, S^2, S'^2$ và được gọi cùng tên

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.4}$$

$$s^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \overline{x}^2 \tag{5.5}$$

$$s'^2 = \frac{n}{n-1}s^2. \tag{5.6}$$

b) Nếu mẫu cụ thể $(x_1, x_2, \dots, x_n)$ được rút gọn dưới dạng bảng

X	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k

(x_i xuất hiện n_i lần trong mẫu)

thì

$$n = \sum_{i=1}^{k} n_i \tag{5.7}$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{k} n_i x_i \tag{5.8}$$

$$s^2 = \left(\frac{1}{n} \sum_{i=1}^{k} n_i x_i^2\right) - \overline{x}^2. \tag{5.9}$$

Ví dụ 5.1. Để khảo sát cường độ chịu nén X của gạch đặc nung do một xí nghiệp sản xuất, người ta lấy mẫu gồm $n = 100$ viên gạch mang đi kiểm tra, được bảng sau

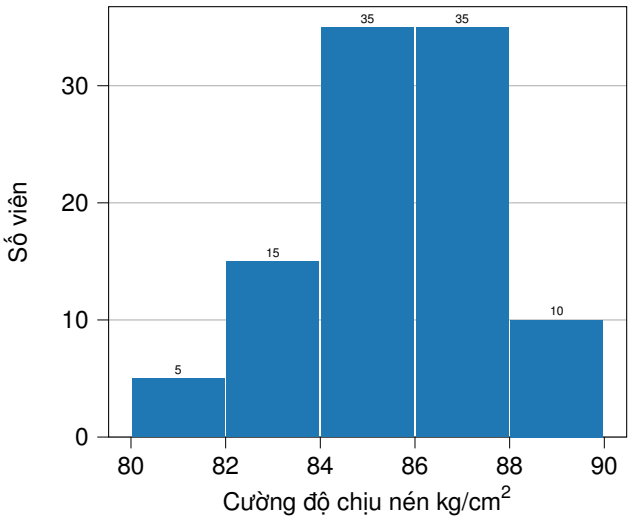
Cường độ kg/cm^2	80 – 82	82 – 84	84 – 86	86 – 88	88 – 100
Số viên	5	15	35	35	10

a) Mô tả dữ liêu bằng biểu đồ tần số, và biểu đồ tần suất (tổ chức đồ).

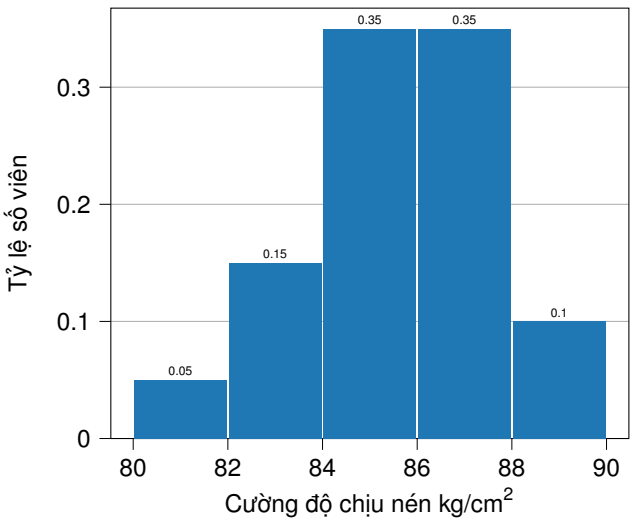
b) Tìm các đặc trưng số của mẫu trên, biết giá trị đại diện để tính toán là trung điểm của các khoảng chia trong bảng.

Giải. a)

```
1  import matplotlib.pyplot as plt

3  plt.bar(X, n, width=1.95)

5  for i in range(len(X)):
6    plt.text(X[i], n[i], n[i], horizontalalignment='center',
       verticalalignment='bottom')

8  plt.xlabel('Cường độ chịu nén kg/cm$^2$')
9  plt.ylabel('Số viên')
10 plt.grid(axis='y', alpha=0.75)
```




```
N = sum(n)

plt.bar(X, [x/N for x in n], width=1.95)

for i in range(len(X)):
  plt.text(X[i], n[i]/N, n[i]/N, horizontalalignment='center
    ', verticalalignment='bottom')
```



b) $\overline{x} = \dfrac{5 \cdot 81 + 15 \cdot 83 + 35 \cdot 85 + 35 \cdot 87 + 10 \cdot 89}{100} = 85.6$

$s^2 = \dfrac{5 \cdot 81^2 + 15 \cdot 83^2 + 35 \cdot 85^2 + 35 \cdot 87^2 + 10 \cdot 89^2}{100} - 85.6^2 = 4.04$

$s = \sqrt{4.04} = 2.0100$

$s'^2 = \dfrac{100}{99} 4.04 = 4.0808$, và $s' = \sqrt{4.0808} = 2.0201$.

```
L = []
for i in range(len(X)):
  L += [X[i]] * n[i]

X = np.array(L)
# hoặc X = np.array( [81]*5 + [83]*15 + [85]*35 + [87]*35 + [89]*10 )

X.mean(), X.var(), X.std()   # x̄, s² và s
```

□

Ví dụ 5.2. Số liệu về giới tính (gender), chiều cao (height, đơn vị: inch), cân nặng (weight, đơn vị: pound), tại địa chỉ https://raw.githubusercontent.com/Dataweekends/zero_to_deep_learning_video/master/data/weight-height.csv có dạng

	Gender	Height	Weight
0	Male	73.847017	241.893563
1	Male	68.781904	162.310473
2	Male	74.110105	212.740856
3	Male	71.730978	220.042470
4	Male	69.881796	206.349801
...	...	...	...
9995	Female	66.172652	136.777454
9996	Female	67.067155	170.867906
9997	Female	63.867992	128.475319
9998	Female	69.034243	163.852461
9999	Female	61.944246	113.649103

a) Đọc và hiển thị lại số liệu trên. Lọc ra các trường hợp nam có chiều cao trên 70 inch.

b) Vẽ biểu đồ tần số của chiều cao với 10 và 50 khoảng chia.

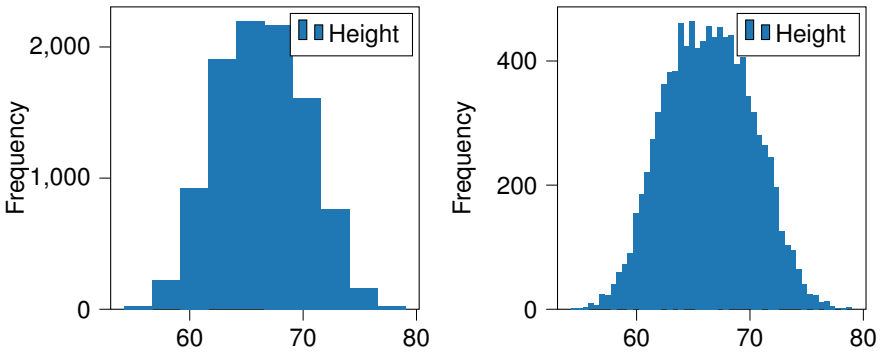
c) Tìm các đặc trưng mẫu của chiều cao.

Chứng minh. a)

```
import pandas as pd
df = pd.read_csv('https://raw.githubusercontent.com/
    Dataweekends/zero_to_deep_learning_video/master/data/
    weight-height.csv')

df          # 5 dòng đầu và 5 dòng cuối của số liệu
df.head()   # 5 dòng đầu
```

```
df[ (df['Gender'] == 'Male') & (df['Height'] > 70) ]
```

b)




```
df['Height'].plot(kind='hist', bins=10)
```

c) Đặc trưng mẫu của chiều cao là $\overline{x} = 66.367560$ inch, $s'^2 = 14.803473$ inch^2, $s' = 3.847528$ inch.

Cách 1: Bảng mô tả số liệu

	Height	Weight
count	10000.000000	10000.000000
mean	66.367560	161.440357
std	3.847528	32.108439
min	54.263133	64.700127
25%	63.505620	135.818051
50%	66.318070	161.212928
75%	69.174262	187.169525
max	78.998742	269.989699

```
df.describe()
```

Cách 2: Tính đặc trưng mẫu của mọi trường số liệu.

```
df.mean()
df.var()
df.std()
```

Cách 3: Tính đặc trưng mẫu của chiều cao

```
df['Height'].mean()
df['Height'].var()
df['Height'].std()
```

□

Bài tập 5.2

5.1. Bảng sau cho các giá trị quan sát được và số lần xuất hiện giá trị đó trong mẫu

X	0	2	3	4	6	8	9
n_i	1	2	4	1	1	5	1

a) Xây dựng biểu đồ tần suất của mẫu đã cho.

b) Tính các đặc trưng mẫu $\overline{x}, s^2, s, s'^2, s'$.

5.2. Kết quả thi môn xác suất thống kê của lớp gồm 30 sinh viên được cho như sau:

Điểm	10	9	8	7	6	5	4	3
Số sinh viên	3	4	7	4	4	3	3	2

a) Vẽ đường gấp khúc tần suất.

b) Tính số điểm trung bình mà lớp đạt được, độ phân tán tiêu chuẩn của số điểm của các sinh viên.

5.3. Bảng sau cho số liệu thống kê theo khoảng chiều cao của 30 người trưởng thành

Chiều cao (m)	1.55 – 1.65	1.65 – 1.75	1.75 – 1.85
Số người	6	15	9

a) Vẽ tổ chức đồ của mẫu đã cho.

b) Tính xác đặc trưng mẫu $\overline{x}, s^2, s, s'^2, s'$.

5.3 Các phân bố thường gặp trong thống kê

5.3.1 Phân bố chuẩn $N\left(a, \sigma^2\right)$

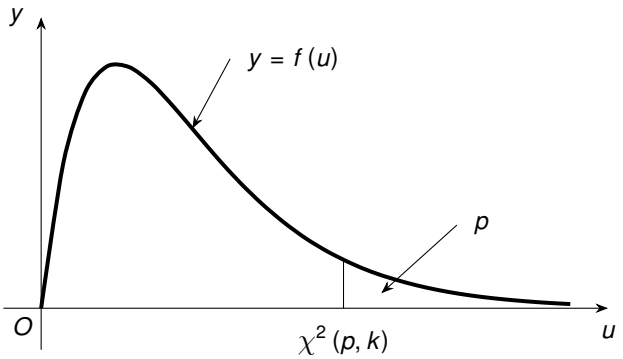
Xem **??**.

5.3.2 Phân bố χ^2

Đại lượng ngẫu nhiên U có phân bố χ^2 với k bậc tự do, ký hiệu $U \sim \chi^2_k$, nếu có hàm mật độ

$$f(u) = \begin{cases} \dfrac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} u^{\frac{k}{2}-1} e^{-\frac{u}{2}} & \text{nếu } u > 0 \\ 0 & \text{nếu } u \leq 0. \end{cases}$$

trong đó $\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x} dx, \ (a > 0)$.



$$P\left[U > \chi^2(p,k)\right] = p. \tag{5.10}$$

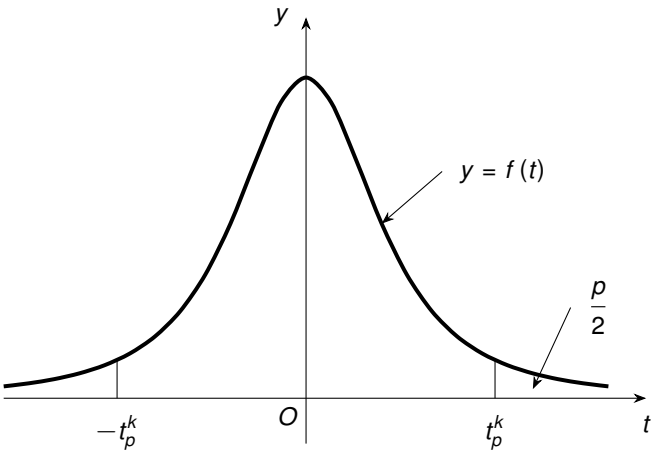
Ví dụ 5.3. $\chi^2(0.05, 30) = 43.773$.

```
from scipy.stats import chi2
chi2.isf(0.05, 30)
```

5.3.3 Phân bố Student

Đại lượng ngẫu nhiên T có phân bố Student với k bậc tự do, ký hiệu $T \sim t_k$, nếu có hàm mật độ

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-\frac{k+1}{2}}.$$



$$P\left(|T| > t_p^k\right) = p. \tag{5.11}$$

Ví dụ 5.4. $t_{0.02}^{25} = 2.4851$.

```
from scipy.stats import t
t.isf(0.02 / 2, 25)
```

5.4 Phân bố mẫu

Giả sử $(X_1, X_2, \dots, X_n)$ là mẫu ngẫu nhiên rút từ đại lượng ngẫu nhiên X.

a) Nếu $X \sim N\left(a, \sigma^2\right)$ thì

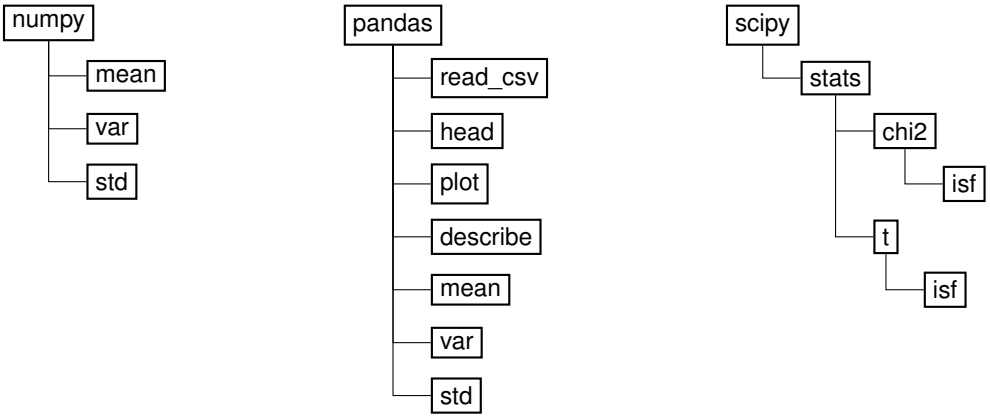
1) $\dfrac{\overline{X} - a}{\sigma}\sqrt{n} \sim N(0,1)$.

2) $\dfrac{\overline{X} - a}{S}\sqrt{n-1} \sim t_{n-1}$.

3) $\dfrac{nS^2}{\sigma^2} \sim \chi^2_{n-1}$.

b) Nếu $EX = a$, $DX = \sigma^2$, và n đủ lớn, thì $\dfrac{\overline{X} - a}{\sigma}\sqrt{n}$ có phân bố xấp xỉ chuẩn $N(0,1)$.

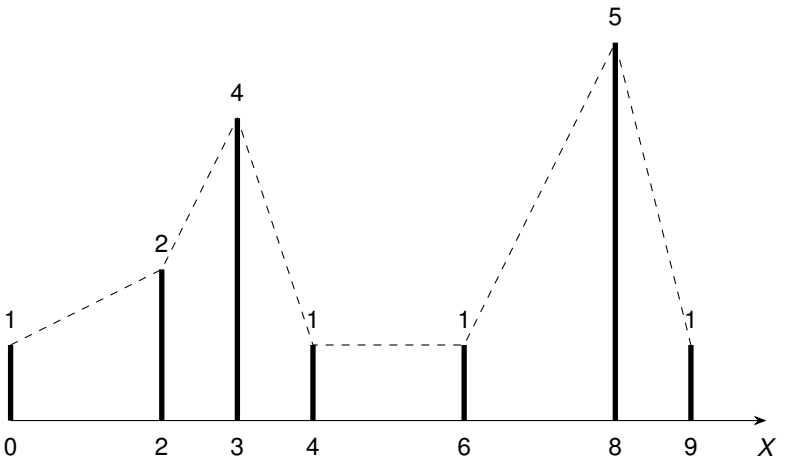
Tóm tắt về Python



Chương 5

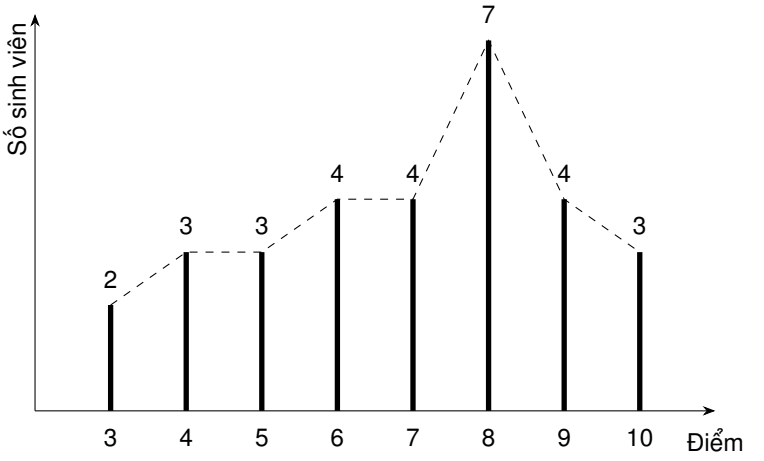
Mẫu và phân bố mẫu

5.1 a)



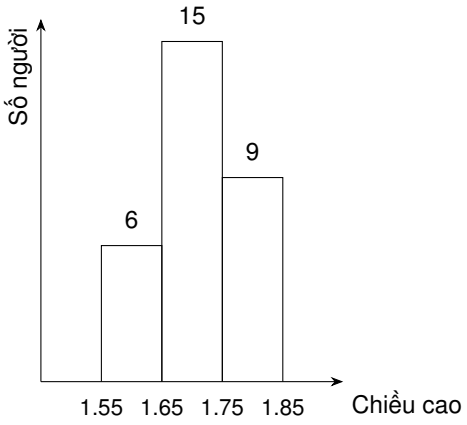
b) $\overline{x} = 5, s^2 = 8.1333, s'^2 = 8.7143, s = 2.8519, s' = 2.952.$

5.2 a)



b) $\overline{x} = 6.9, s^2 = 4.1567, s'^2 = 4.3, s = 2.0388, s' = 2.0736.$

5.3 a)



b) $\overline{x} = 1.71, s^2 = 0.0049, s'^2 = 0.005069, s = 0.07, s' = 0.07120.$