Thống kê mô tả (Descriptive statistics)

Nguyen Thi Ngoc Anh

AI Academy Vietnam

September 19, 2020

Nội dung

1. Thu thập và phân loại dữ liệu
   - Tổng thể và tập mẫu
   - Biểu diễn dữ liệu
2. Mẫu ngẫu nhiên và các đặc trưng mẫu
   - Mẫu ngẫu nhiên
   - Các đặc trưng mẫu
3. Mô tả phân bố của dữ liệu bằng đồ thị
4. Mô tả phân bố của dữ liệu bằng các số đặc trưng
5. Bài tập thực hành

Tổng thể (population)

Khi nghiên cứu về một vấn đề người ta thường khảo sát trên một dấu hiệu nào đó, các dấu hiệu này được thể hiện trên nhiều phần tử.

Definition

Tập hợp các phần tử mang dấu hiệu ta quan tâm được gọi là tổng thể hay đám đông (population).

Ví dụ

- Nghiên cứu tập hợp gà trong một trại chăn nuôi, ta quan tâm đến dấu hiệu trọng lượng.
- Nghiên cứu chất lượng sinh viên trong 1 trường đại học, ta quan tâm đến dấu hiệu điểm.
- Nghiên cứu về giá của một loại sản phẩm A, đối tượng ta quan tâm là các sản phẩm loại A bán trên thị trường.

Một số lý do không thể khảo sát toàn bộ tổng thể

- **Giới hạn về thời gian, tài chính**: Ví dụ muốn khảo sát xem chiều cao của thanh niên Việt Nam hiện nay có tăng lên hay không ta phải khảo sát toàn bộ thanh niên VN (khoảng 40 triệu người). Để khảo sát hết sẽ tốn nhiều thời gian và kinh phí.
- **Phá vỡ tổng thể nghiên cứu**: Ví dụ ta cất vào kho $N = 10000$ hộp sản phẩm và muốn biết tỷ lệ hộp hư sau 1 năm bảo quản. Ta phải kiểm tra từng hộp để xác định số hộp hư. Một hộp sản phẩm sau khi kiểm tra thì không bán ra thị trường được.
- **Không xác định được chính xác tổng thể**: Ví dụ muốn khảo sát tỷ lệ người bị nhiễm HIV qua đường tiêm chích là bao nhiêu. Tổng thể lúc này là toàn bộ người bị nhiễm HIV, nhưng ta không thể xác định chính xác là bao nhiêu người. Ngoài ra số người bị nhiễm HIV mới và bị chết do HIV thay đổi liên tục nên tổng thể thay đổi liên tục.

Tập mẫu (sample)

Do đó người ta nghĩ ra cách thay vì khảo sát tổng thể, người ta chỉ cần chọn ra một tập nhỏ để khảo sát và đưa ra quyết định.

Definition

- Tập mẫu là tập con của tổng thể và có tính chất tương tự như tổng thể.
- Số phần tử của tập mẫu được gọi là *kích thước mẫu*.

Câu hỏi

Làm sao chọn được tập mẫu có tính chất tương tự như tổng thể để các kết luận của tập mẫu có thể dùng cho tổng thể ?

Một số cách chọn mẫu cơ bản

Một số cách chọn mẫu

- Chọn mẫu ngẫu nhiên có hoàn lại: Lấy ngẫu nhiên 1 phần tử từ tổng thể và khảo sát nó. Sau đó trả phần tử đó lại tổng thể trước khi lấy 1 phần tử khác. Tiếp tục như thế n lần ta thu được một mẫu có hoàn lại gồm n phần tử.
- Chọn mẫu ngẫu nhiên không hoàn lại: Lấy ngẫu nhiên 1 phần tử từ tổng thể và khảo sát nó rồi để qua một bên, không trả lại tổng thể. Sau đó lấy ngẫu nhiên 1 phần tử khác, tiếp tục như thế n lần ta thu được một mẫu không hoàn lại gồm n phần tử.
- Chọn mẫu phân nhóm: Đầu tiên ta chia tập nền thành các nhóm tương đối thuần nhất, từ mỗi nhóm đó chọn ra một mẫu ngẫu nhiên. Tập hợp tất cả mẫu đó cho ta một mẫu phân nhóm. Phương pháp này dùng khi trong tập nền có những sai khác lớn. Hạn chế là phụ thuộc vào việc chia nhóm.
- Chọn mẫu có suy luận: dựa trên ý kiến của chuyên gia về đối tượng

Biểu diễn dữ liệu

Từ tổng thể ta trích ra tập mẫu có n phần tử. Ta có n số liệu.

Dạng liệt kê

Các số liệu thu được ta ghi lại thành dãy số liệu:

$$x_1, x_2, \ldots, x_n$$

Dạng rút gọn

Số liệu thu được có sự lặp đi lặp lại một số giá trị thì ta có dạng rút gọn sau:

- Dạng tần số: $(n_1 + n_2 + \ldots + n_k = n)$

Giá trị	x_1	x_2	$\ldots$	x_k
Tần số	n_1	n_2	$\ldots$	n_k

- Dạng tần suất: $(p_k = n_k/n)$

Giá trị	x_1	x_2	$\ldots$	x_k

Biểu diễn dữ liệu

Dạng khoảng

Dữ liệu thu được nhận giá trị trong (a, b). Ta chia (a, b) thành k miền con bởi các điểm chia: $a_0 = a < a_1 < a_2 < ... < a_{k-1} < a_k = b$.

- Dạng tần số: $(n_1 + n_2 + \ldots + n_k = n)$

Giá trị	$(a_0 - a_1]$	$(a_1 - a_2]$	$\ldots$	$(a_{k-1} - a_k]$
Tần số	n_1	n_2	$\ldots$	n_k

- Dạng tần suất: $(p_k = n_k/n)$

Giá trị	$(a_0 - a_1]$	$(a_1 - a_2]$	$\ldots$	$(a_{k-1} - a_k]$
Tần suất	p_1	p_2	$\ldots$	p_k

- Một số vấn đề chú ý:
  - k bao nhiêu là hợp lý: nếu k nhỏ thì mất mát nhiều thông tin, k lớn thì tính toán mất nhiều công sức. Thông thường chọn $k = 5 \rightarrow 15$.
  - Độ dài các khoảng thường chia bằng nhau, một số trường hợp có thể chia độ dài khác nhau.

Biểu diễn dữ liệu

Dạng khoảng

- Nếu độ dài các khoảng bằng nhau ta có thể chuyển về dạng rút gọn.

Giá trị	x_1	x_2	...	x_k
Tần suất	p_1	p_2	...	p_k

Trong đó x_i là điểm đại diện cho $(a_{i-1}, a_i]$ thường được xác định là trung điểm của miền: $x_i = \frac{1}{2}(a_{i-1} + a_i)$

- Dạng rút gọn thường được thể hiện bằng đồ thị dạng đường hoặc dạng hình tròn.
- Dạng khoảng thường được thể hiện bằng đồ thị dạng hình cột.

Mẫu ngẫu nhiên

Tổng thể được đặc trưng bởi dấu hiệu nghiên cứu X là một biến ngẫu nhiên. Do đó khi nói về X là nói về tổng thể.
Từ tổng thể trích ra n phần tử làm một tập mẫu. Ta có 2 loại tập mẫu: *mẫu ngẫu nhiên* và *mẫu cụ thể*
Gọi X_i là biến ngẫu nhiên chỉ giá trị thu được của phần tử thứ $i, i = 1, 2, \ldots, n$. Ta có $X_1, X_2, \ldots, X_n$ là n biến ngẫu nhiên độc lập và có cùng phân phối với biến ngẫu nhiên X.

Definition

- *Mẫu ngẫu nhiên*: là véctơ $(X_1, X_2, \ldots, X_n)$, trong đó mỗi thành phần X_i là một biến ngẫu nhiên. Các biến ngẫu nhiên này độc lập và có cùng phân phối xác suất với X.
- *Mẫu cụ thể*: là véctơ $(x_1, x_2, \ldots, x_n)$, trong đó mỗi thành phần x_i là một giá trị cụ thể.
- Với một mẫu ngẫu nhiên thì có nhiều mẫu cụ thể ứng với các lần lấy mẫu khác nhau.

Mẫu ngẫu nhiên

Example

Một kệ chứa các đĩa nhạc với giá như sau:

Giá (ngàn đồng)	20	25	30	34	40
Số đĩa	35	10	25	17	13

Ta cần lấy 4 đĩa có hoàn lại để khảo sát.
Ta xét trong 2 trường hợp:

- Xét về mặt định lượng: giá của từng đĩa là bao nhiêu?
- (X_1, X_2, X_3, X_4) là một mẫu ngẫu nhiên
- $(x_1, x_2, x_3, x_4) = (20, 30, 20, 40)$, đây là mẫu cụ thể.

Các đặc trưng mẫu

Thống kê (statistics)

Cho $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên.
Biến ngẫu nhiên $Y = g(X_1, X_2, ..., X_n)$ (với g là một hàm nào đó) được gọi là một thống kê

Các đặc trưng mẫu

Thống kê (statistics)

Cho $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên.
Biến ngẫu nhiên $Y = g(X_1, X_2, ..., X_n)$ (với g là một hàm nào đó) được gọi là một thống kê

Các tham số đặc trưng

- *Xét tổng thể về mặt định lượng*: tổng thể được đặc trưng bởi dấu hiệu nghiên cứu X, (X là biến ngẫu nhiên). Ta có:
  - Trung bình tổng thể: $EX = \mu$
  - Phương sai tổng thể: $VX = \sigma^2$
  - Độ lệch chuẩn của tổng thể: σ.
- *Xét tổng thể về mặt định tính*: tổng thể có kích thướcN, trong đó có M phần tử có tính chất A. Khi đó $p = M/N$ gọi là tỷ lệ xảy ra A của tổng thể.

Các đặc trưng mẫu

Trung bình mẫu

Cho $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên.

- Thống kê - **Trung bình mẫu ngẫu nhiên**:

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

- Mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ có mẫu cụ thể $(x_1, x_2, ..., x_n)$ thì $\overline{X}$ nhận giá trị:

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

$\overline{x}$ được gọi là **trung bình mẫu**.

Nếu mẫu dạng rút gọn thì: $\overline{x} = \frac{1}{k}\sum_{i=1}^{n} x_i n_i$

Các đặc trưng mẫu

Phương sai mẫu(chưa hiệu chỉnh)

Cho $(X_1, X_2, ..., X_n)$ là một mẫu ngẫu nhiên.

- Thống kê - **Phương sai mẫu ngẫu nhiên**:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ có mẫu cụ thể $(x_1, x_2, ..., x_n)$ thì S^2 nhận giá trị:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

  S^2 được gọi là **Phương sai mẫu (chưa hiệu chỉnh)**.

- Vấn đề: $E(S^2) = \frac{n-1}{n}\sigma^2$

Các đặc trưng mẫu

Phương sai mẫu hiệu chỉnh

Ta phải hiệu chỉnh đi để thu được giá trị thay thế σ^2 tốt hơn.

- Thống kê - **Phương sai mẫu ngẫu nhiên hiệu chỉnh**:

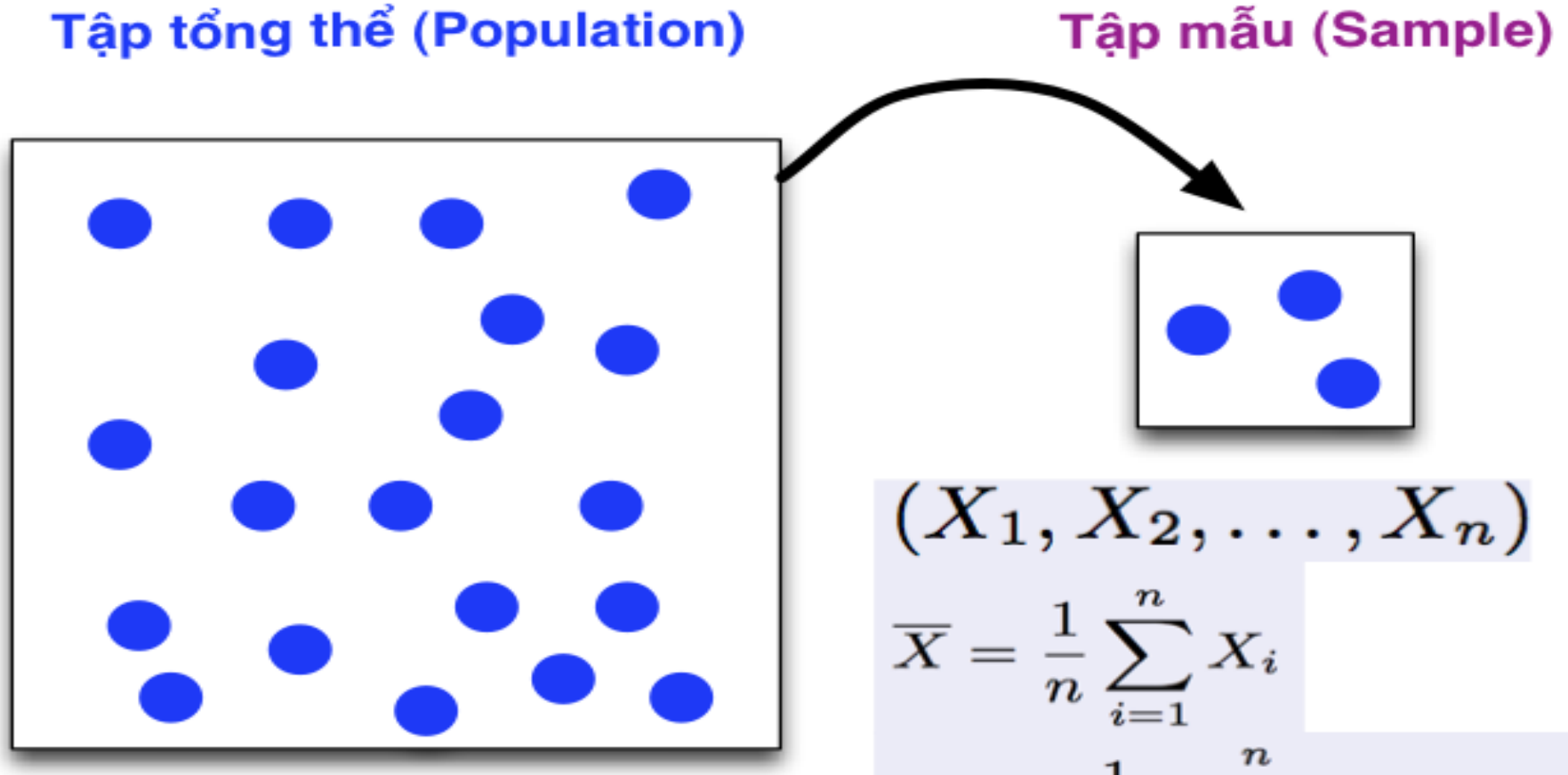
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

- Mẫu ngẫu nhiên $(X_1, X_2, ..., X_n)$ có mẫu cụ thể $(x_1, x_2, ..., x_n)$ thì s^2 nhận giá trị:

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

  s^2 được gọi là **Phương sai mẫu hiệu chỉnh**.
- s được gọi là **độ lệch chuẩn mẫu hiệu chỉnh**.

Các đặc trưng mẫu



$(X_1, X_2, \ldots, X_n)$

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

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

X Bnn gốc

$EX = \mu$ $VX = \sigma^2$

Thu thập dữ liệu từ file

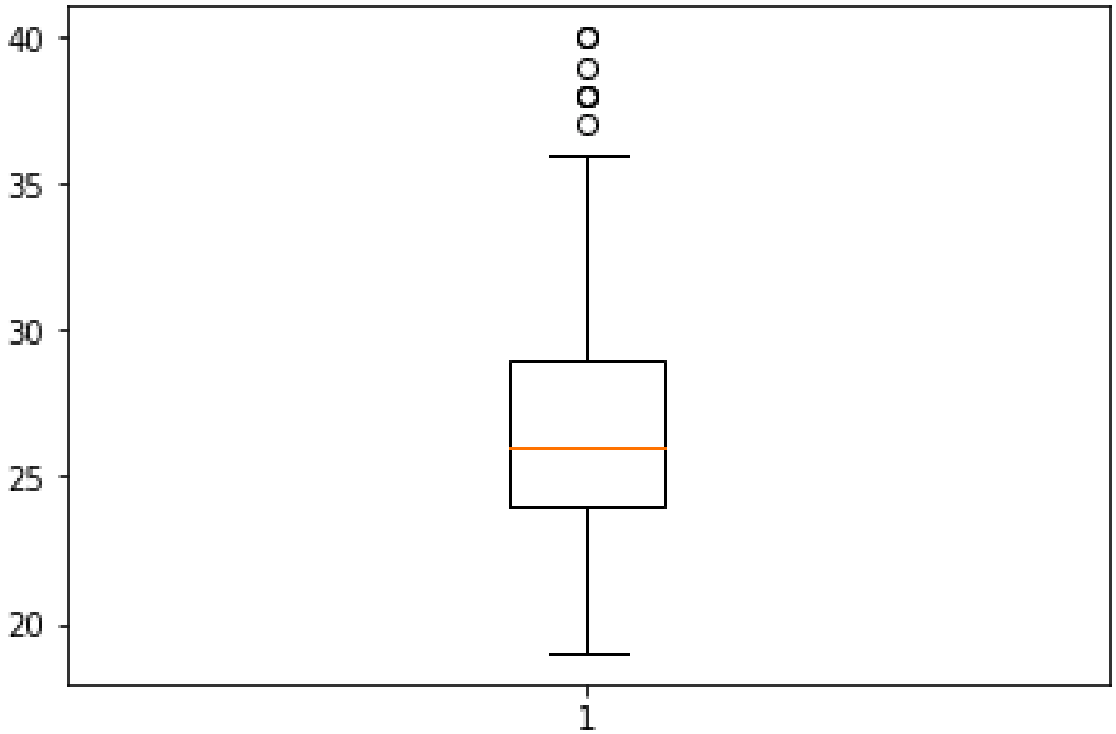
```
import pandas as pd  # gọi thư viện pandas
import re # Gọi modun regex

data = pd.read_csv("https://media.geeksforgeeks.org/wp-content/uploads/nba.csv")
# lấy một bảng dữ liệu từ trên mạng
data.dropna(inplace = True)  # Bỏ đi những trường nan
data.head(10)# Bảng dữ liệu lấy 10 bản ghi đầu tiên
```

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
0	Avery Bradley	Boston Celtics	0.0	PG	25.0	6-2	180.0	Texas	7730337.0
1	Jae Crowder	Boston Celtics	99.0	SF	25.0	6-6	235.0	Marquette	6796117.0
3	R.J. Hunter	Boston Celtics	28.0	SG	22.0	6-5	185.0	Georgia State	1148640.0
6	Jordan Mickey	Boston Celtics	55.0	PF	21.0	6-8	235.0	LSU	1170960.0
7	Kelly Olynyk	Boston Celtics	41.0	C	25.0	7-0	238.0	Gonzaga	2165160.0
8	Terry Rozier	Boston Celtics	12.0	PG	22.0	6-2	190.0	Louisville	1824360.0
9	Marcus Smart	Boston Celtics	36.0	PG	22.0	6-4	220.0	Oklahoma State	3431040.0
10	Jared Sullinger	Boston Celtics	7.0	C	24.0	6-9	260.0	Ohio State	2569260.0
11	Isaiah Thomas	Boston Celtics	4.0	PG	27.0	5-9	185.0	Washington	6912869.0
12	Evan Turner	Boston Celtics	11.0	SG	27.0	6-7	220.0	Ohio State	3425510.0

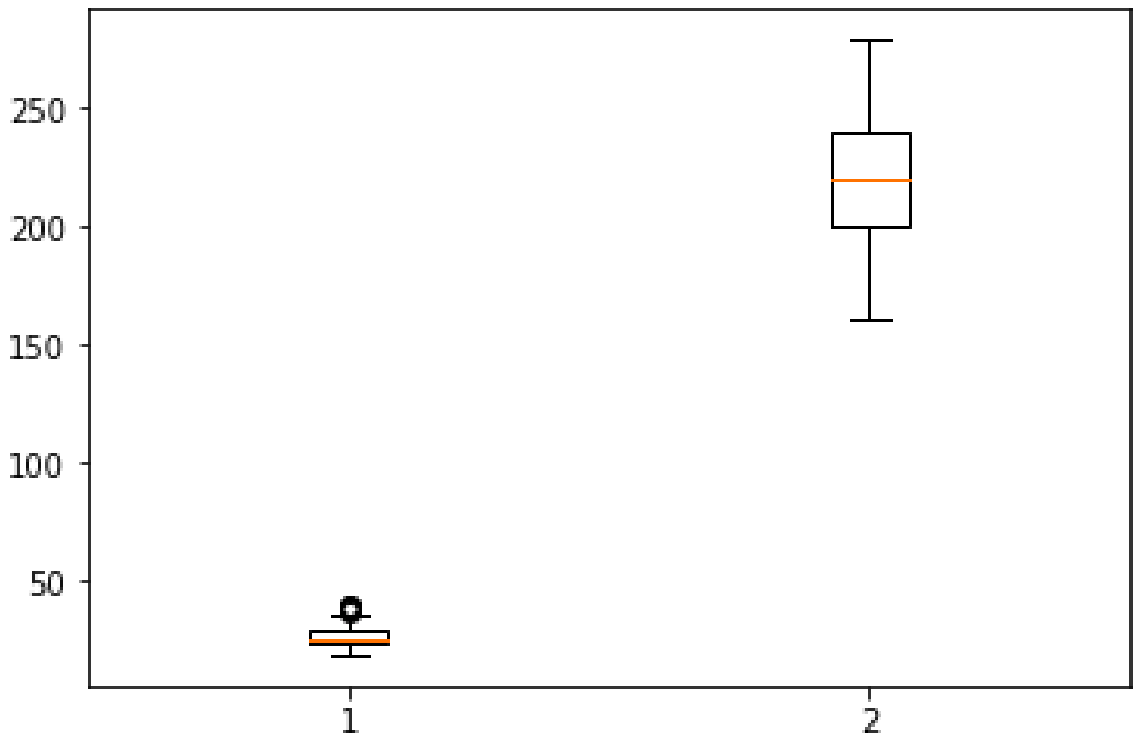
Đồ thị mô tả phân bố dữ liệu

```
plt.boxplot(data['Age'])
plt.show()
```



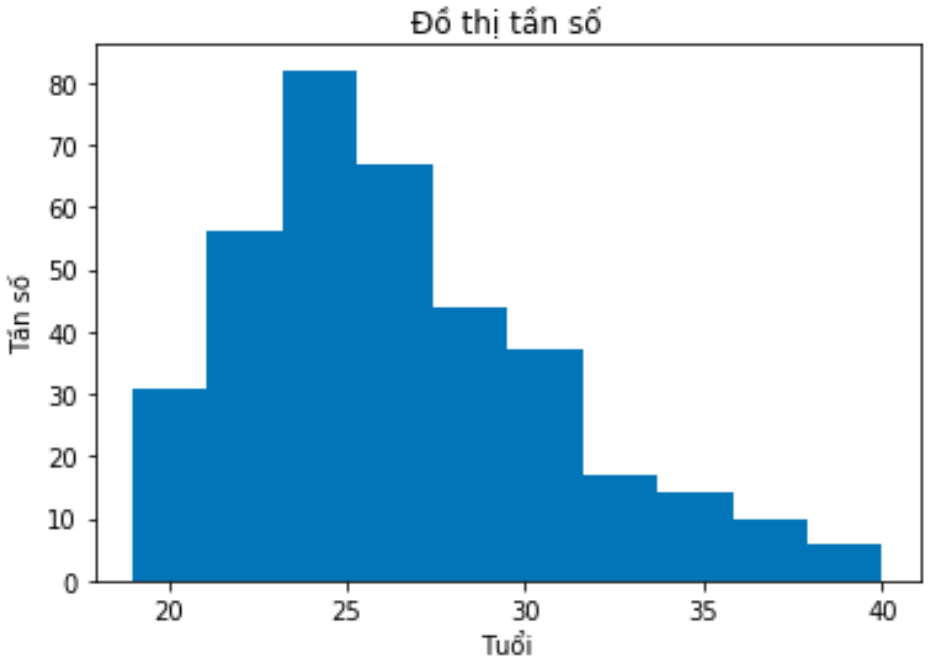
Đồ thị mô tả phân bố dữ liệu

```
datanew=[data['Age'], data['Weight']]
plt.boxplot(datanew)
plt.show()
```



Đồ thị mô tả phân bố dữ liệu

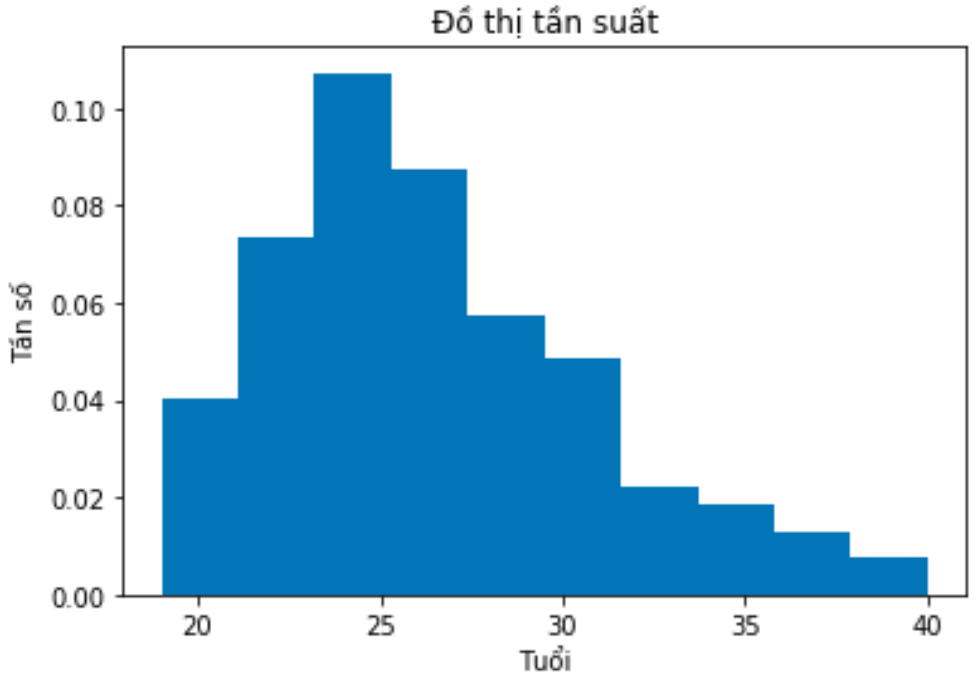
```
import matplotlib.pyplot as plt #Gọi thư viện
plt.hist(data['Age'], bins=10)
plt.title('Đồ thị tần số')
plt.ylabel('Tần số')
plt.xlabel('Tuổi')
# Vẽ đồ thị histogram với 10 khoảng
plt.show()#Hiển thị hình vẽ
```



Đồ thị mô tả phân bố dữ liệu

```
plt.hist(data['Age'], bins=10, density=True)
plt.title('Đồ thị tần suất')
plt.ylabel('Tần số')
plt.xlabel('Tuổi')
```

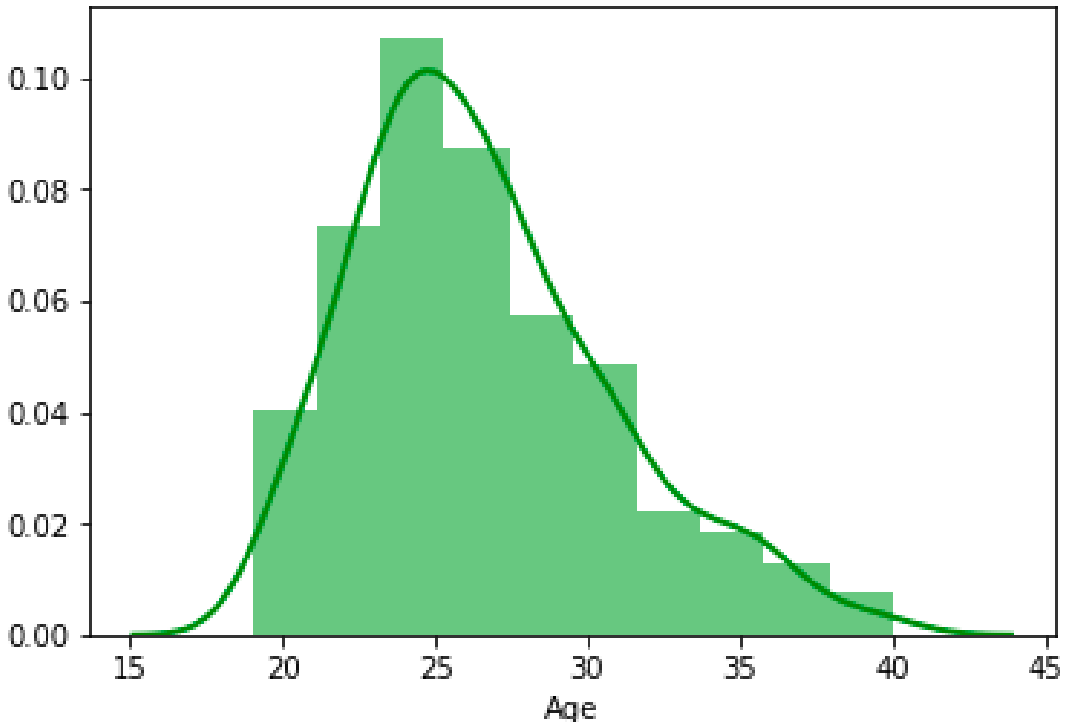
```
Text(0.5,0,'Tuổi')
```



Đồ thị mô tả phân bố dữ liệu

```
import seaborn as sns
kwargs = dict(hist_kws={'alpha':.5}, kde_kws={'linewidth':2})
sns.distplot(data['Age'], color="green", bins=10, label="Hàm mật độ thực nghiệm
```

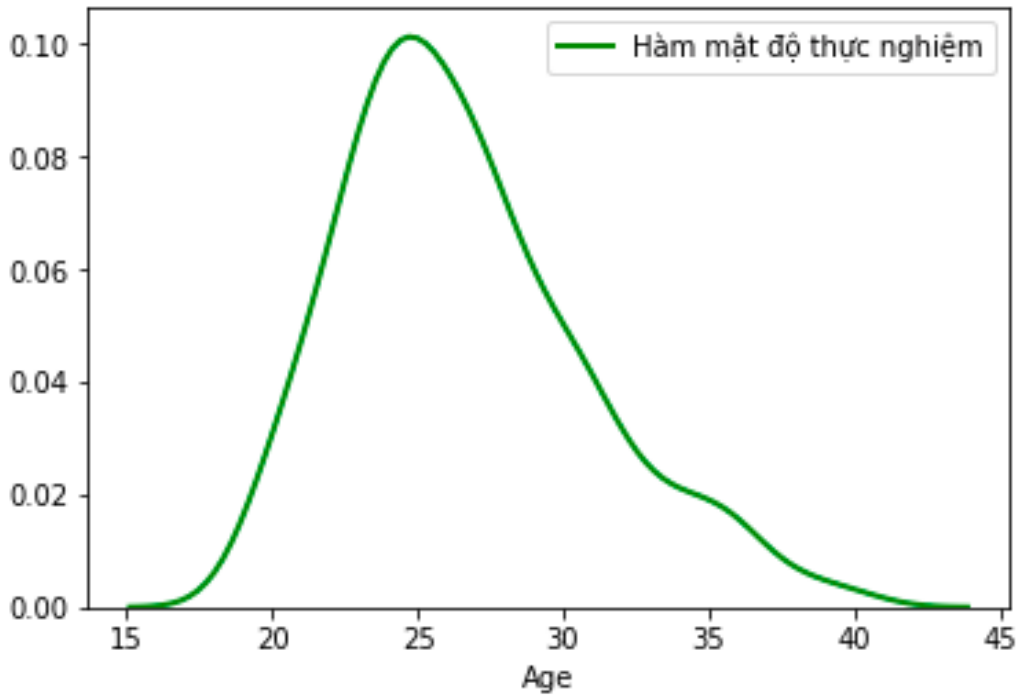
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f414080>
```



Đồ thị mô tả phân bố dữ liệu

```
import seaborn as sns
kwargs = dict(hist_kws={'alpha':.5}, kde_kws={'linewidth':2})
sns.distplot(data['Age'], color="green",hist = False, bins=10, label="Hàm mật độ th
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1f770828>
```



Mô tả dữ liệu có được thông qua các hàm thống kê

```
data.describe()#Thống kê mô tả dữ liệu
```

	Number	Age	Weight	Salary
count	364.000000	364.000000	364.000000	3.640000e+02
mean	16.829670	26.615385	219.785714	4.620311e+06
std	14.994162	4.233591	24.793099	5.119716e+06
min	0.000000	19.000000	161.000000	5.572200e+04
25%	5.000000	24.000000	200.000000	1.000000e+06
50%	12.000000	26.000000	220.000000	2.515440e+06
75%	25.000000	29.000000	240.000000	6.149694e+06
max	99.000000	40.000000	279.000000	2.287500e+07

Mô tả dữ liệu có được thông qua các hàm thống kê

```
phan_tram =[0.1,.20,0.3, .40,0.5, .60,0.7, .80, 0.9] #phân vị muốn lấy
include =['object', 'float', 'int'] # danh sách các loại dữ liệu
mo_ta = data.describe(percentiles = phan_tram, include = include)# thống kê mô tả
mo_ta
```

	Name	Team	Number	Position	Age	Height	Weight	College	Salary
count	364	364	364.000000	364	364.000000	364	364.000000	364	3.640000e+02
unique	364	30	NaN	5	NaN	17	NaN	115	NaN
top	E'Twaun Moore	New Orleans Pelicans	NaN	SG	NaN	6-9	NaN	Kentucky	NaN
freq	1	16	NaN	87	NaN	49	NaN	22	NaN
mean	NaN	NaN	16.829670	NaN	26.615385	NaN	219.785714	NaN	4.620311e+06
std	NaN	NaN	14.994162	NaN	4.233591	NaN	24.793099	NaN	5.119716e+06
min	NaN	NaN	0.000000	NaN	19.000000	NaN	161.000000	NaN	5.572200e+04
10%	NaN	NaN	1.300000	NaN	22.000000	NaN	186.000000	NaN	6.650000e+05
20%	NaN	NaN	4.000000	NaN	23.000000	NaN	195.000000	NaN	9.472760e+05
30%	NaN	NaN	6.900000	NaN	24.000000	NaN	205.000000	NaN	1.146836e+06
40%	NaN	NaN	9.000000	NaN	25.000000	NaN	212.000000	NaN	1.638754e+06
50%	NaN	NaN	12.000000	NaN	26.000000	NaN	220.000000	NaN	2.515440e+06
60%	NaN	NaN	17.000000	NaN	27.000000	NaN	228.000000	NaN	3.429934e+06
70%	NaN	NaN	22.000000	NaN	28.000000	NaN	235.000000	NaN	5.106651e+06
80%	NaN	NaN	30.000000	NaN	30.000000	NaN	242.400000	NaN	7.838202e+06
90%	NaN	NaN	35.700000	NaN	33.000000	NaN	250.000000	NaN	1.347000e+07
max	NaN	NaN	99.000000	NaN	40.000000	NaN	279.000000	NaN	2.287500e+07

Bài tập thực hành 1

- Tạo 5000 số ngẫu nhiên có phân phối nhị thức với n=50, p=0.7
- Tính các giá trị thống kê min, max, Q1, Q2, Q3
- Vẽ đồ thị boxplot để mô tả các giá trị đặc trưng
- Vẽ đồ thị tần số, tần xuất, hàm mật độ có histogram, chỉ có hàm mật độ

Bài tập thực hành 2

- Tạo 5000 số ngẫu nhiên (Chiều cao, cân nặng) có phân phối chuẩn 2 chiều với mean=[165, 55], cov=[[100, 0.4], [0.4, 36]]
- Tính các giá trị thống kê min, max, Q1, Q2, Q3 của chiều cao, cân nặng
- Vẽ đồ thị boxplot để mô tả các giá trị đặc trưng của chiều cao, cân nặng
- Vẽ đồ thị tần số, tần xuất, hàm mật độ có histogram, chỉ có hàm mật độ đồng thời của chiều cao, cân nặng