VẤN ĐỀ ƯỚC LƯỢNG (ESTIMATION)

Nguyễn Văn Hạnh

AI Academy Vietnam

Tháng 7 năm 2021

Nội dung

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.
- Tổng thể X có phân bố xác suất là $F(x; \theta)$ phụ thuộc vào tham số θ.

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.
- Tổng thể X có phân bố xác suất là $F(x; \theta)$ phụ thuộc vào tham số θ.
- Tham số θ là không biết vì chúng ta không quan sát được toàn bộ tổng thể.

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.
- Tổng thể X có phân bố xác suất là $F(x; \theta)$ phụ thuộc vào tham số θ.
- Tham số θ là không biết vì chúng ta không quan sát được toàn bộ tổng thể.
- Bài toán ước lượng: chúng ta cần ước lượng giá trị tham số θ từ mẫu quan sát $(X_1, X_2, \ldots, X_n)$.

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.
- Tổng thể X có phân bố xác suất là $F(x; \theta)$ phụ thuộc vào tham số θ.
- Tham số θ là không biết vì chúng ta không quan sát được toàn bộ tổng thể.
- Bài toán ước lượng: chúng ta cần ước lượng giá trị tham số θ từ mẫu quan sát $(X_1, X_2, \ldots, X_n)$.
- Ước lượng điểm (point estimation) của tham số θ là một thống kê $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$.

Đặt vấn đề

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X.
- Tổng thể X có phân bố xác suất là $F(x; \theta)$ phụ thuộc vào tham số θ.
- Tham số θ là không biết vì chúng ta không quan sát được toàn bộ tổng thể.
- Bài toán ước lượng: chúng ta cần ước lượng giá trị tham số θ từ mẫu quan sát $(X_1, X_2, \ldots, X_n)$.
- Ước lượng điểm (point estimation) của tham số θ là một thống kê $\hat{\theta} = h(X_1, X_2, \ldots, X_n)$.
- Ước lượng khoảng (interval estimation) của tham số θ với độ tin cậy (confidence level) $1 - \alpha$ là một khoảng $[\hat{\theta}_1; \hat{\theta}_2] = [h_1(X_1, X_2, \ldots, X_n); h_2(X_1, X_2, \ldots, X_n)]$ thoả mãn:

$$P(\hat{\theta}_1 \leq \theta \leq \hat{\theta}_2) = 1 - \alpha$$

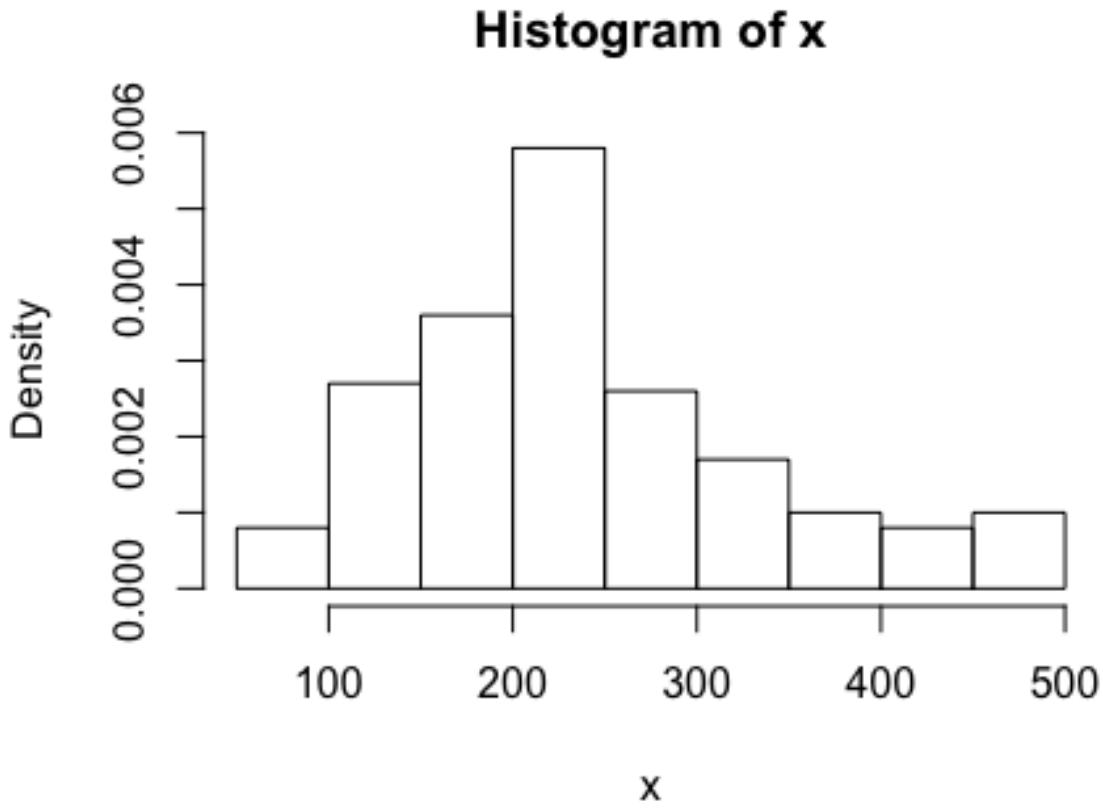
Ví dụ

Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực. Khảo sát tiền điện/tháng của 200 hộ ta thu được bảng số liệu như sau:

```
196.65 468.75 320.50 300.50 213.05 140.60 290.00 216.95 360.50 317.95 195.55
220.50 255.60 289.00 194.55 374.25 382.05 185.55 219.10 215.60 220.00 186.75
 97.80 340.50  88.50 209.50 234.04 333.00 291.10 108.50 245.00 184.00 153.50
219.50 214.15 155.20 140.40 108.50 410.00 125.50 220.30 160.00 300.50 310.20
244.40 194.50 210.20 360.00 456.50 237.40 235.00 203.25 109.20 240.15 260.50
275.50 101.55 455.50 246.25 291.55 262.00 378.65 194.50 248.00 262.92  85.75
248.00 204.75 310.70 213.10 320.50 125.60 110.25  77.35 119.50 313.50 222.00
388.10 110.50 160.00 210.00 310.30 380.10 281.00 105.35 280.15 188.80 272.50
103.40 213.50 280.50 119.50 166.10 180.50 212.00 154.75 100.50 452.60 436.35
225.00 124.30 170.00 127.35 107.90 140.00 195.00 315.10 241.05 168.00 120.50
223.95 237.05 285.45 100.50 228.55 248.70 175.80 466.05 219.00 216.00 425.50
390.00 176.85 240.50 226.00 108.70 160.00 470.50 225.00 440.00 265.00 162.80
260.50 175.80  73.05 460.50 263.60  59.50 198.00 416.50 315.50 155.00 190.00
158.50 225.00 266.70 153.60 238.00 297.60 201.75 240.50 270.90 196.65 299.20
 70.50 125.60 100.40 240.00 240.00 224.05 194.00 247.00 325.40 102.20 166.10
361.00 430.00 240.00 250.50 470.00 157.75  98.40 236.50 230.85 317.65 200.70
165.00 350.50 319.15 275.88 203.05 234.50 220.75 180.50 436.50 403.00 460.50
220.00 103.50 222.15 170.50 224.15 460.00 260.40 200.50 311.40 260.00 251.55
100.60 212.20
```

Ví dụ

Biểu đồ tần suất của tập dữ liệu này có dạng như sau:



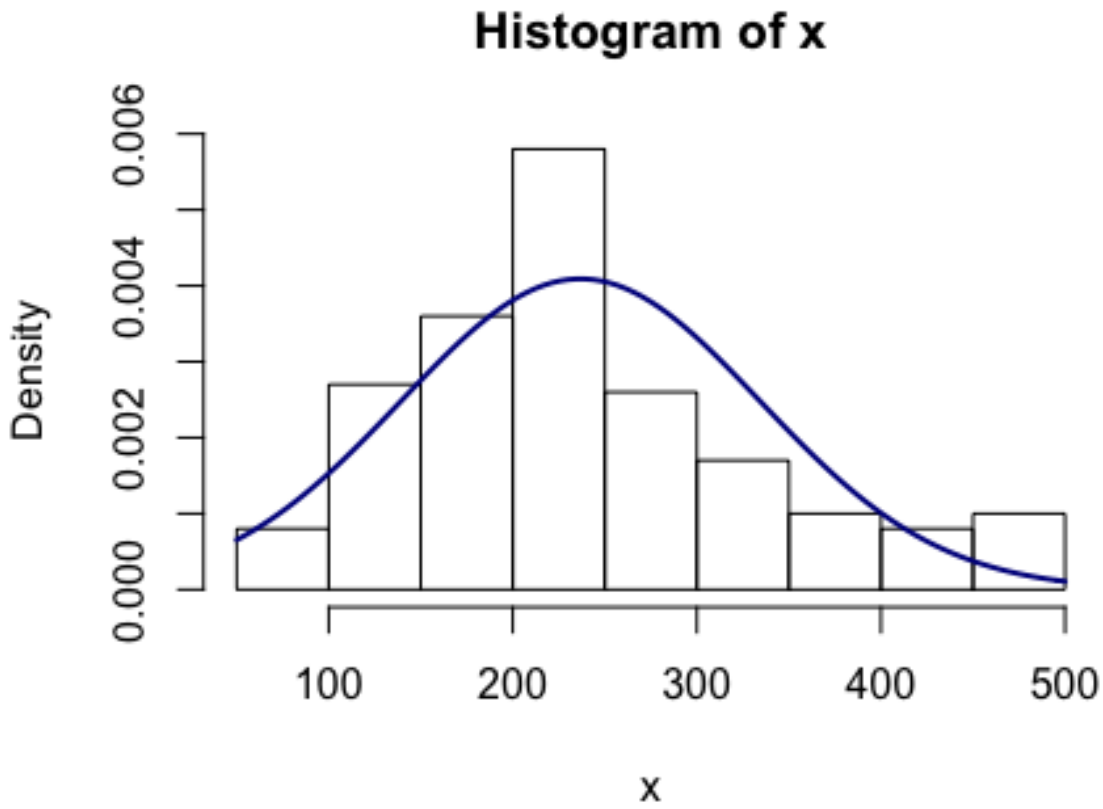
Ví dụ

Code Python:

```
import pandas as pd
import seaborn as sns
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x=df.values[:,0]
sns.distplot(x, hist=True, kde=False, color = 'blue',
             hist_kws={'edgecolor':'black'})
```

Ví dụ

Phân bố của dữ liệu có thể xấp xỉ bởi phân bố chuẩn:



Ví dụ

Code Python:

```
import pandas as pd
import seaborn as sns
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x=df.values[:,0]
sns.distplot(x, hist=True, kde=True, color = 'blue',
             hist_kws={'edgecolor':'black'})
```

Ví dụ

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$ với hàm mật độ xác suất:

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

Ví dụ

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$ với hàm mật độ xác suất:

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

- Tham số μ là trung bình tổng thể (là trung bình tiền điện/tháng của tất cả các hộ trong khu vực); tham số σ^2 là phương sai tổng thể.

Ví dụ

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$ với hàm mật độ xác suất:

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

- Tham số μ là trung bình tổng thể (là trung bình tiền điện/tháng của tất cả các hộ trong khu vực); tham số σ^2 là phương sai tổng thể.
- Một ước lượng của μ là trung bình mẫu

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = 236.78$$

Ví dụ

- Mô hình hoá: Ta có thể giả sử tiền điện X của các hộ gia đình cá nhân tại khu vực trên có phân bố chuẩn với tham số $\theta = (\mu, \sigma^2)$ với hàm mật độ xác suất:

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

- Tham số μ là trung bình tổng thể (là trung bình tiền điện/tháng của tất cả các hộ trong khu vực); tham số σ^2 là phương sai tổng thể.
- Một ước lượng của μ là trung bình mẫu

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = 236.78$$

- Một ước lượng điểm của σ^2 là phương sai mẫu

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 9460.12$$

Ước lượng hợp lý cực đại (maximum likelihood estimation)

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố là $f(x;\theta)$ (là hàm mật độ xác suất hoặc hàm khối xác suất).

Ước lượng hợp lý cực đại (maximum likelihood estimation)

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố là $f(x;\theta)$ (là hàm mật độ xác suất hoặc hàm khối xác suất).
- Hàm hợp lý (Likelihood function) là hàm như sau

$$L(\theta) = \Pi_{i=1}^{n} f(X_i;\theta)$$

Ước lượng hợp lý cực đại (maximum likelihood estimation)

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố là $f(x;\theta)$ (là hàm mật độ xác suất hoặc hàm khối xác suất).
- Hàm hợp lý (Likelihood function) là hàm như sau

$$L(\theta) = \Pi_{i=1}^{n} f(X_i;\theta)$$

- Hàm hợp lý đo xác suất đầu ra (xác suất để mẫu được quan sát). Đầu ra phụ thuộc vào tham số mô hình, mỗi tham số khác nhau sẽ cho đầu ra là khác nhau.

Ước lượng hợp lý cực đại (maximum likelihood estimation)

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố là $f(x;\theta)$ (là hàm mật độ xác suất hoặc hàm khối xác suất).
- Hàm hợp lý (Likelihood function) là hàm như sau

$$L(\theta) = \Pi_{i=1}^{n} f(X_i;\theta)$$

- Hàm hợp lý đo xác suất đầu ra (xác suất để mẫu được quan sát). Đầu ra phụ thuộc vào tham số mô hình, mỗi tham số khác nhau sẽ cho đầu ra là khác nhau.
- Ý tưởng của MLE là chọn tham số θ sao cho đầu ra của mô hình gần nhất với tập mẫu quan sát được.

Ước lượng hợp lý cực đại (maximum likelihood estimation)

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố là $f(x;\theta)$ (là hàm mật độ xác suất hoặc hàm khối xác suất).
- Hàm hợp lý (Likelihood function) là hàm như sau

$$L(\theta) = \Pi_{i=1}^{n} f(X_i;\theta)$$

- Hàm hợp lý đo xác suất đầu ra (xác suất để mẫu được quan sát). Đầu ra phụ thuộc vào tham số mô hình, mỗi tham số khác nhau sẽ cho đầu ra là khác nhau.
- Ý tưởng của MLE là chọn tham số θ sao cho đầu ra của mô hình gần nhất với tập mẫu quan sát được.
- Ước lượng hợp lý cực đại: Ta tìm tham số θ sao cho xác suất đầu ra là lớn nhất có thể, tức:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmax}} \log L(\theta)$$

Phương pháp ước lượng mô-men (Method of moment)

- Mô-men lý thuyết bậc k của biến ngẫu nhiên X là $\mathbb{E}(X^k)$.

Phương pháp ước lượng mô-men (Method of moment)

- Mô-men lý thuyết bậc k của biến ngẫu nhiên X là $\mathbb{E}(X^k)$.
- Mô-men thực nghiệm bậc k là $\frac{1}{n}\sum_{i=1}^{n} X_i^k$

Phương pháp ước lượng mô-men (Method of moment)

- Mô-men lý thuyết bậc k của biến ngẫu nhiên X là $\mathbb{E}(X^k)$.
- Mô-men thực nghiệm bậc k là $\frac{1}{n}\sum_{i=1}^{n} X_i^k$
- Phương pháp ước lượng mô-men: Nếu mô hình có r tham số thì ta tìm ước lượng của các tham số bằng cách giải hệ phương trình

$$\mathbb{E}(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k; \text{ với } k = 1, 2, \ldots, r$$

Ước lượng trung bình và phương sai: MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số là $\theta = (\mu, \sigma^2)$.

Ước lượng trung bình và phương sai: MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số là $\theta = (\mu, \sigma^2)$.
- Hàm hợp lý (Likelihood function) là

$$L(\theta) = \Pi_{i=1}^{n} f(X_i; \theta) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

Ước lượng trung bình và phương sai: MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số là $\theta = (\mu, \sigma^2)$.
- Hàm hợp lý (Likelihood function) là

$$L(\theta) = \Pi_{i=1}^{n} f(X_i; \theta) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

- Log-hàm hợp lý (Log-likelihood function) là

$$\log L(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

Ước lượng trung bình và phương sai: MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số là $\theta = (\mu, \sigma^2)$.
- Hàm hợp lý (Likelihood function) là

$$L(\theta) = \Pi_{i=1}^{n} f(X_i; \theta) = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(X_i - \mu)^2}{2\sigma^2} \right)$$

- Log-hàm hợp lý (Log-likelihood function) là

$$\log L(\theta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2$$

- Ước lượng hợp lý cực đại: Ta tìm tham số θ sao cho xác suất đầu ra là lớn nhất có thể, tức:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmax}} \log L(\theta)$$

Ước lượng trung bình và phương sai: MLE

- Ta giải hệ

$$\frac{\partial \log L(\theta)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\frac{\partial \log L(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

Ước lượng trung bình và phương sai: MLE

- Ta giải hệ

$$\frac{\partial \log L(\theta)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0$$

$$\frac{\partial \log L(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 = 0$$

- Ta thu được ước lượng

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ và } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Ước lượng trung bình và phương sai: Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với 2 tham số là μ và σ^2).

Ước lượng trung bình và phương sai: Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với 2 tham số là μ và σ^2).
- Mô-men lý thuyết bậc 1 và bậc 2 là:

$$\mathbb{E}(X) = \mu \text{ và } \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Ước lượng trung bình và phương sai: Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với 2 tham số là μ và σ^2).
- Mô-men lý thuyết bậc 1 và bậc 2 là:

$$\mathbb{E}(X) = \mu \text{ và } \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

- Khi đó ta giải hệ

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X) = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Ước lượng trung bình và phương sai: Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với 2 tham số là μ và σ^2).
- Mô-men lý thuyết bậc 1 và bậc 2 là:

$$\mathbb{E}(X) = \mu \text{ và } \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

- Khi đó ta giải hệ

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}(X) = \mu$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathbb{E}(X^2) = \sigma^2 + \mu^2$$

- Ta thu được ước lượng

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ và } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Ước lượng phương sai: Tại sao chia cho $n-1$?

- Ước lượng không chệch: Ước lượng $\hat{\theta}$ của tham số θ gọi là không chệch nếu $\mathbb{E}(\hat{\theta}) = \theta$.

Ước lượng phương sai: Tại sao chia cho $n-1$?

- Ước lượng không chệch: Ước lượng $\hat{\theta}$ của tham số θ gọi là không chệch nếu $\mathbb{E}(\hat{\theta}) = \theta$.
- Ước lượng $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ là ước lượng chệch của σ^2 vì

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big) = \frac{n-1}{n}\sigma^2$$

Ước lượng phương sai: Tại sao chia cho $n-1$?

- Ước lượng không chệch: Ước lượng $\hat{\theta}$ của tham số θ gọi là không chệch nếu $\mathbb{E}(\hat{\theta}) = \theta$.
- Ước lượng $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ là ước lượng chệch của σ^2 vì

$$\mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big) = \frac{n-1}{n}\sigma^2$$

- Khi đó ta hiệu chỉnh ước lượng trên để thu được ước lượng không chệch của σ^2 là phương sai mẫu:

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

Ước lượng khoảng của trung bình

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số trung bình là μ.

Ước lượng khoảng của trung bình

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số trung bình là μ.
- Thống kê $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ có phân phối Student với $n-1$ bậc tự do.

Ước lượng khoảng của trung bình

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số trung bình là μ.
- Thống kê $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ có phân phối Student với $n-1$ bậc tự do.
- Ta có

$$\mathbb{P}\Big(-t_{n-1;\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1;\alpha/2} \Big) = 1-\alpha$$

$$\mathbb{P}\Big(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \Big) = 1-\alpha$$

Ước lượng khoảng của trung bình

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố chuẩn với tham số trung bình là μ.
- Thống kê $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ có phân phối Student với $n-1$ bậc tự do.
- Ta có

$$\mathbb{P}\Big(-t_{n-1;\alpha/2} \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq t_{n-1;\alpha/2} \Big) = 1-\alpha$$

$$\mathbb{P}\Big(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \Big) = 1-\alpha$$

- Khi đó khoảng ước lượng của μ với độ tin cậy $1-\alpha$ là

$$\Big[\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}} \Big]$$

Ước lượng khoảng của trung bình: kích thước mẫu lớn

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).

Ước lượng khoảng của trung bình: kích thước mẫu lớn

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Theo định lý giới hạn trung tâm và định lý Slutky

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to N(0, 1)$$

Ước lượng khoảng của trung bình: kích thước mẫu lớn

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Theo định lý giới hạn trung tâm và định lý Slutky

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to N(0,1)$$

- Ta có

$$\mathbb{P}\Big(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq Z_{\alpha/2} \Big) \to 1 - \alpha$$

Ước lượng khoảng của trung bình: kích thước mẫu lớn

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Theo định lý giới hạn trung tâm và định lý Slutky

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \to N(0,1)$$

- Ta có

$$\mathbb{P}\Big(-Z_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq Z_{\alpha/2} \Big) \to 1 - \alpha$$

- Khi đó khoảng ước lượng của μ với độ tin cậy tiệm cận $1 - \alpha$ (n đủ lớn) là

$$\left[\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}} \right]$$

Ước lượng xác suất (tỷ lệ): Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.

Ước lượng xác suất (tỷ lệ): Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Ví dụ: Giả sử p là xác suất xuất hiện mặt sấp khi ta tung một đồng xu. Ta thực hiện tung đồng xu n lần, kí hiệu $X_i = 1$ nếu lần tung thứ i kết quả là mặt sấp và $X_i = 0$ nếu lần tung thứ i kết quả là mặt ngửa. Khi đó mẫu $(X_1, X_2, \ldots, X_n)$ được lấy từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.

Ước lượng xác suất (tỷ lệ): Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Ví dụ: Giả sử p là xác suất xuất hiện mặt sấp khi ta tung một đồng xu. Ta thực hiện tung đồng xu n lần, kí hiệu $X_i = 1$ nếu lần tung thứ i kết quả là mặt sấp và $X_i = 0$ nếu lần tung thứ i kết quả là mặt ngửa. Khi đó mẫu $(X_1, X_2, \ldots, X_n)$ được lấy từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Mô-men lý thuyết bậc 1 của phân phối Bernoulli là $\mathbb{E}(X) = p$

Ước lượng xác suất (tỷ lệ): Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Ví dụ: Giả sử p là xác suất xuất hiện mặt sấp khi ta tung một đồng xu. Ta thực hiện tung đồng xu n lần, kí hiệu $X_i = 1$ nếu lần tung thứ i kết quả là mặt sấp và $X_i = 0$ nếu lần tung thứ i kết quả là mặt ngửa. Khi đó mẫu $(X_1, X_2, \ldots, X_n)$ được lấy từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Mô-men lý thuyết bậc 1 của phân phối Bernoulli là $\mathbb{E}(X) = p$
- Ta giải hệ phương trình

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Ước lượng xác suất (tỷ lệ): Phương pháp mô-men

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Ví dụ: Giả sử p là xác suất xuất hiện mặt sấp khi ta tung một đồng xu. Ta thực hiện tung đồng xu n lần, kí hiệu $X_i = 1$ nếu lần tung thứ i kết quả là mặt sấp và $X_i = 0$ nếu lần tung thứ i kết quả là mặt ngửa. Khi đó mẫu $(X_1, X_2, \ldots, X_n)$ được lấy từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Mô-men lý thuyết bậc 1 của phân phối Bernoulli là $\mathbb{E}(X) = p$
- Ta giải hệ phương trình

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Ta thu được ước lượng của xác suất p theo phương pháp mô-men là

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

Ước lượng xác suất (tỷ lệ): MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.

Ước lượng xác suất (tỷ lệ): MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Hàm khối xác suất (probabilitty mass function):

$$f(x; p) = p^x(1-p)^{1-x}, x \in \{0, 1\}$$

Ước lượng xác suất (tỷ lệ): MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Hàm khối xác suất (probabilitty mass function):

$$f(x; p) = p^x(1-p)^{1-x}, x \in \{0, 1\}$$

- Hàm hợp lý đồng thời là:

$$L(p) = \Pi_{i=1}^n f(X_i; p) = \Pi_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

Ước lượng xác suất (tỷ lệ): MLE

- Giả sử ta quan sát một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có phân bố Bernoulli với tham số xác suất p.
- Hàm khối xác suất (probabilitty mass function):

$$f(x; p) = p^x(1-p)^{1-x}, x \in \{0, 1\}$$

- Hàm hợp lý đồng thời là:

$$L(p) = \Pi_{i=1}^n f(X_i; p) = \Pi_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^{\sum X_i}(1-p)^{n-\sum X_i}$$

- Log - hàm hợp lý đồng thời là:

$$\log L(p) = \sum X_i \log p + (n - \sum X_i) \log(1-p)$$

Ước lượng xác suất (tỷ lệ): MLE

- Cực đại hoá log-hàm hợp lý:

$$\hat{p} = \underset{p}{\operatorname{argmax}} \log L(p)$$

Ước lượng xác suất (tỷ lệ): MLE

- Cực đại hoá log-hàm hợp lý:

$$\hat{p} = \underset{p}{\operatorname{argmax}} \log L(p)$$

- Giải hệ phương trình

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum X_i}{p} - \frac{n - \sum X_i}{1 - p} = 0$$

Ước lượng xác suất (tỷ lệ): MLE

- Cực đại hoá log-hàm hợp lý:

$$\hat{p} = \underset{p}{\operatorname{argmax}} \log L(p)$$

- Giải hệ phương trình

$$\frac{\partial \log L(p)}{\partial p} = \frac{\sum X_i}{p} - \frac{n - \sum X_i}{1 - p} = 0$$

- Ta thu được ước lượng ML của p là

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

Ước lượng khoảng của xác suất

- Ước lượng điểm của p là

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Ước lượng khoảng của xác suất

- Ước lượng điểm của p là

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Theo định lý giới hạn trung tâm và định lý Slutky

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \to N(0,1)$$

Ước lượng khoảng của xác suất

- Ước lượng điểm của p là

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Theo định lý giới hạn trung tâm và định lý Slutky

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \to N(0,1)$$

- Khi đó khoảng ước lượng của p với độ tin cậy tiệm cận $1-\alpha$ (n đủ lớn) là

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Ước lượng hệ số tương quan

- Hệ số tương quan (correlation coefficient) giữa hai biến ngẫu nhiên X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{[\mathbb{E}(X^2) - \mathbb{E}(X)^2][\mathbb{E}(Y^2) - \mathbb{E}(Y)^2]}}$$

Ước lượng hệ số tương quan

- Hệ số tương quan (correlation coefficient) giữa hai biến ngẫu nhiên X và Y là

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sqrt{[\mathbb{E}(X^2) - \mathbb{E}(X)^2][\mathbb{E}(Y^2) - \mathbb{E}(Y)^2]}}$$

- $\rho(X,Y)$ được ước lượng bởi hệ số tương quan mẫu giữa 2 mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_n)$:

$$r(X,Y) = \frac{\overline{XY} - \overline{X}.\ \overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}$$

Phương pháp Bootstrap

- Phương pháp Bootstrap là phương pháp coi mẫu gốc ban đầu đóng vai trò tổng thể mà từ đó nó được rút ra.

Phương pháp Bootstrap

- Phương pháp Bootstrap là phương pháp coi mẫu gốc ban đầu đóng vai trò tổng thể mà từ đó nó được rút ra.
- Từ mẫu ban đầu lấy lại các mẫu ngẫu nhiên (resampling) cùng cỡ với mẫu gốc bằng phương pháp lấy mẫu có hoàn lại, gọi là mẫu Bootstrap.

Phương pháp Bootstrap

- Phương pháp Bootstrap là phương pháp coi mẫu gốc ban đầu đóng vai trò tổng thể mà từ đó nó được rút ra.
- Từ mẫu ban đầu lấy lại các mẫu ngẫu nhiên (resampling) cùng cỡ với mẫu gốc bằng phương pháp lấy mẫu có hoàn lại, gọi là mẫu Bootstrap.
- Với mỗi mẫu lấy lại ta tính được giá trị tham số thống kê quan tâm gọi là tham số Bootstrap

Phương pháp Bootstrap

- Phương pháp Bootstrap là phương pháp coi mẫu gốc ban đầu đóng vai trò tổng thể mà từ đó nó được rút ra.
- Từ mẫu ban đầu lấy lại các mẫu ngẫu nhiên (resampling) cùng cỡ với mẫu gốc bằng phương pháp lấy mẫu có hoàn lại, gọi là mẫu Bootstrap.
- Với mỗi mẫu lấy lại ta tính được giá trị tham số thống kê quan tâm gọi là tham số Bootstrap
- Sự phân bố của các tham số thống kê mẫu Bootstrap là phân phối Bootstrap

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1 - \alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1 - \alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1-\alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1-\alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các thống kê Bootstrap $\hat{\theta}_b^*$, với $b = 1, \ldots, B$.

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1-\alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các thống kê Bootstrap $\hat{\theta}_b^*$, với $b = 1, \ldots, B$.
- Tính các sai khác Bootstrap $\sigma_b^* = \hat{\theta}_b^* - \hat{\theta}$, với $b = 1, \ldots, B$.

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1-\alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các thống kê Bootstrap $\hat{\theta}_b^*$, với $b = 1, \ldots, B$.
- Tính các sai khác Bootstrap $\sigma_b^* = \hat{\theta}_b^* - \hat{\theta}$, với $b = 1, \ldots, B$.
- Tính 2 giá trị phân vị mức $\alpha/2$ và $1-\alpha/2$ của dãy sai khác σ_b^* là $\sigma_{\alpha/2}^*$ và $\sigma_{1-\alpha/2}^*$.

Phương pháp Bootstrap

Phương pháp Bootstrap để tính khoảng ước lượng với độ tin cậy $1-\alpha$:

- Giả sử có một mẫu $(X_1, X_2, \ldots, X_n)$ lấy từ tổng thể có phân bố $F(x, \theta)$ với tham số cần ước lượng θ.
- Tính thống kê $\hat{\theta}$ ước lượng cho tham số θ từ mẫu $(X_1, X_2, \ldots, X_n)$.
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các thống kê Bootstrap $\hat{\theta}_b^*$, với $b = 1, \ldots, B$.
- Tính các sai khác Bootstrap $\sigma_b^* = \hat{\theta}_b^* - \hat{\theta}$, với $b = 1, \ldots, B$.
- Tính 2 giá trị phân vị mức $\alpha/2$ và $1-\alpha/2$ của dãy sai khác σ_b^* là $\sigma_{\alpha/2}^*$ và $\sigma_{1-\alpha/2}^*$.
- Khoảng ước lượng của θ là $[\hat{\theta} - \sigma_{1-\alpha/2}^*; \hat{\theta} - \sigma_{\alpha/2}^*]$.

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$.

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$.
- Tính độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$.
- Tính độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$
- Tính 2 giá trị phân vị $\sigma_{\alpha/2}^*$ và $\sigma_{1-\alpha/2}^*$ của dãy sai khác σ_b^*.

Ước lượng giá trị trung bình: phương pháp Bootstrap

- Xét một mẫu $(X_1, X_2, \ldots, X_n)$ từ một tổng thể X có giá trị trung bình là μ (có phân bố bất kỳ).
- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n$
- Lấy B mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$.
- Tính độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$
- Tính 2 giá trị phân vị $\sigma_{\alpha/2}^*$ và $\sigma_{1-\alpha/2}^*$ của dãy sai khác σ_b^*.
- Khoảng ước lượng của μ là $[\bar{X} - \sigma_{1-\alpha/2}^*; \bar{X} - \sigma_{\alpha/2}^*]$.

Ví dụ minh hoạ

Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực. Khảo sát tiền điện/tháng của 200 hộ ta thu được bảng số liệu như sau:

```
196.65 468.75 320.50 300.50 213.05 140.60 290.00 216.95 360.50 317.95 195.55
220.50 255.60 289.00 194.55 374.25 382.05 185.55 219.10 215.60 220.00 186.75
 97.80 340.50  88.50 209.50 234.04 333.00 291.10 108.50 245.00 184.00 153.50
219.50 214.15 155.20 140.40 108.50 410.00 125.50 220.30 160.00 300.50 310.20
244.40 194.50 210.20 360.00 456.50 237.40 235.00 203.25 109.20 240.15 260.50
275.50 101.55 455.50 246.25 291.55 262.00 378.65 194.50 248.00 262.92  85.75
248.00 204.75 310.70 213.10 320.50 125.60 110.25  77.35 119.50 313.50 222.00
388.10 110.50 160.00 210.00 310.30 380.10 281.00 105.35 280.15 188.80 272.50
103.40 213.50 280.50 119.50 166.10 180.50 212.00 154.75 100.50 452.60 436.35
225.00 124.30 170.00 127.35 107.90 140.00 195.00 315.10 241.05 168.00 120.50
223.95 237.05 285.45 100.50 228.55 248.70 175.80 466.05 219.00 216.00 425.50
390.00 176.85 240.50 226.00 108.70 160.00 470.50 225.00 440.00 265.00 162.80
260.50 175.80  73.05 460.50 263.60  59.50 198.00 416.50 315.50 155.00 190.00
158.50 225.00 266.70 153.60 238.00 297.60 201.75 240.50 270.90 196.65 299.20
 70.50 125.60 100.40 240.00 240.00 224.05 194.00 247.00 325.40 102.20 166.10
361.00 430.00 240.00 250.50 470.00 157.75  98.40 236.50 230.85 317.65 200.70
165.00 350.50 319.15 275.88 203.05 234.50 220.75 180.50 436.50 403.00 460.50
220.00 103.50 222.15 170.50 224.15 460.00 260.40 200.50 311.40 260.00 251.55
100.60 212.20
```

Ví dụ minh hoạ

1. Tính ước lượng điểm của kỳ vọng và phương sai của X
2. Tính khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% trong các trường hợp:
   (a) Dữ liệu có phân bố chuẩn
   (b) Kích thước mẫu lớn
   (c) Xét bộ dữ liệu gồm 15 quan sát đầu tiên và áp dụng phương pháp Bootstrap với $B = 1000$
3. Gọi p là tỷ lệ các hộ gia đình tại vùng trên có tiền điện tháng 6/2020 lớn hơn 200 nghìn đồng.
   (a) Tính một ước lượng điểm của p
   (b) Tính khoảng ước lượng của p với độ tin cậy 90%
   (c) Giả sử chỉ có 25 quan sát đầu tiên trong tập dữ liệu. Hãy tính khoảng ước lượng Bootstrap của p với độ tin cậy 90% và số mẫu Bootstrap là 500.

Ví dụ minh hoạ

- Ước lượng điểm của kỳ vọng là trung bình mẫu:

$$\bar{x} = \frac{1}{200}(196.65 + \ldots + 212.20) = 236.78$$

Ví dụ minh hoạ

- Ước lượng điểm của kỳ vọng là trung bình mẫu:

$$\bar{x} = \frac{1}{200}(196.65 + \ldots + 212.20) = 236.78$$

Và ước lượng điểm của phương sai của X là phương sai mẫu:

$$s^2 = \frac{1}{199}[(196.65 - 236.78)^2 + \ldots + (212.2 - 236.78)^2] = 9460.12$$

- Code Python:

```
import numpy as np
import pandas as pd
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x=df.values[:,0]
print(np.mean(x))
print(np.var(x))
```

Ví dụ minh hoạ

- Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% trong các trường hợp dữ liệu có phân bố chuẩn là:

$$\left[\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right] = [225.41; 248.15]$$

- Code Python:

```
import numpy as np
import pandas as pd
from scipy.stats import t
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x = df.values[:,0]
n = len(x)
xbar =np.mean(x)
s = np.std(x)
t = t.ppf(1-(1-0.9)/2,n-1)
L = xbar - t*s/np.sqrt(n)
U = xbar + t*s/np.sqrt(n)
print(L, U)
```

Ví dụ minh hoạ

- Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% trong các trường hợp kích thước mẫu lớn là:

$$\left[\bar{X} - Z_{\alpha/2}\frac{S}{\sqrt{n}}; \bar{X} + Z_{\alpha/2}\frac{S}{\sqrt{n}}\right] = [225.47; 248.093]$$

- Code Python:

```
import numpy as np
import pandas as pd
from scipy.stats import norm
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x = df.values[:,0]
n = len(x)
xbar =np.mean(x)
s = np.std(x)
Z = norm.ppf(1-(1-0.9)/2)
L = xbar - Z*s/np.sqrt(n)
U = xbar + Z*s/np.sqrt(n)
print(L, U)
```

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$
- Lấy $B = 1000$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$
- Lấy $B = 1000$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$
- Lấy $B = 1000$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$
- Độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$
- Lấy $B = 1000$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$
- Độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$
- Hai giá trị phân vị $\sigma_{\alpha/2}^* = -33.34$ và $\sigma_{1-\alpha/2}^* = 35.7$ của dãy sai khác σ_b^*.

Ví dụ minh hoạ

Khoảng ước lượng của tiền điện trung bình tháng 6/2020 của các hộ gia đình ở khu vực trên với độ tin cậy 90% khi sử dụng phương pháp Bootstrap với $B = 1000; n = 15$:

- Trung bình mẫu là $\bar{X} = (X_1 + \ldots + X_n)/n = 265.38$
- Lấy $B = 1000$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Các trung bình mẫu Bootstrap $\bar{X}_b^* = (X_{b1}^* + \ldots + X_{bn}^*)/n$
- Độ lệch của các trung bình mẫu Bootstrap: $\sigma_b = \bar{X}_b^* - \bar{X}$
- Hai giá trị phân vị $\sigma_{\alpha/2}^* = -33.34$ và $\sigma_{1-\alpha/2}^* = 35.7$ của dãy sai khác σ_b^*.
- Khoảng ước lượng của μ là $[\bar{X} - \sigma_{1-\alpha/2}^*; \bar{X} - \sigma_{\alpha/2}^*] = [229.7; 298.7]$

Ví dụ minh hoạ

Code Python:

```
import numpy as np
import pandas as pd
from sklearn.utils import resample
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x = df.values[:,0]
xnew=x[0:15]
xbarnew = np.mean(xnew)
print(xbarnew)
n_iterations = 1000
n_size = len(xnew)
# run bootstrap
stats = []
for i in range(n_iterations):
  Xb = resample(xnew,n_samples=n_size)
  stats.append(np.mean(Xb)-xbarnew)
# confidence interval
alpha = 1 - 0.9
Z1 = np.percentile(stats, 100*alpha/2)
print(Z1)
Z2 = np.percentile(stats, 100*(1-alpha/2))
print(Z2)
L = xbarnew - Z2
U = xbarnew - Z1
print(L, U)
```

Ví dụ minh hoạ

Gọi p là tỷ lệ các hộ gia đình tại vùng trên có tiền điện tháng 6/2020 lớn hơn 200 nghìn đồng.

- Một ước lượng điểm của p là $\hat{p} = 0.645$ (tỷ lệ hộ có tiền điện tháng 6/2020 lớn hơn 200 nghìn đồng trong 200 hộ được khảo sát).
- Khoảng ước lượng của p với độ tin cậy 90% là

$$\left[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = [0.59; 0.70]$$

Ví dụ minh hoạ

Code Python:

```
import numpy as np
import pandas as pd
from scipy.stats import norm
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x = df.values[:,0]
n = len(x)
k = 200
count = sum(i > k for i in x)
pmu = count/n
print(pmu)
Z = norm.ppf(1-(1-0.9)/2)
L = pmu - Z*np.sqrt(pmu*(1-pmu)/n)
U = pmu + Z*np.sqrt(pmu*(1-pmu)/n)
print(L, U)
```

Ví dụ minh hoạ

Khoảng ước lượng Bootstrap của p với độ tin cậy 90%:

- Tỷ lệ trong mẫu là $\hat{p} = 0.68$

Ví dụ minh hoạ

Khoảng ước lượng Bootstrap của p với độ tin cậy 90%:

- Tỷ lệ trong mẫu là $\hat{p} = 0.68$
- Lấy $B = 500$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$

Ví dụ minh hoạ

Khoảng ước lượng Bootstrap của p với độ tin cậy 90%:

- Tỷ lệ trong mẫu là $\hat{p} = 0.68$
- Lấy $B = 500$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các tỷ lệ trong mẫu Bootstrap $\hat{p}_b^*$, với $b = 1, \ldots, B$.

Ví dụ minh hoạ

Khoảng ước lượng Bootstrap của p với độ tin cậy 90%:

- Tỷ lệ trong mẫu là $\hat{p} = 0.68$
- Lấy $B = 500$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các tỷ lệ trong mẫu Bootstrap $\hat{p}_b^*$, với $b = 1, \ldots, B$.
- Tính độ lệch của các tỷ lệ trong mẫu Bootstrap: $\sigma_b^* = \hat{p}_b^* - \hat{p}$

Ví dụ minh hoạ

Khoảng ước lượng Bootstrap của p với độ tin cậy 90%:

- Tỷ lệ trong mẫu là $\hat{p} = 0.68$
- Lấy $B = 500$ mẫu Bootstrap $X_b^* = (X_{b1}^*, \ldots, X_{bn}^*)$, với $b = 1, \ldots, B$
- Tính các tỷ lệ trong mẫu Bootstrap $\hat{p}_b^*$, với $b = 1, \ldots, B$.
- Tính độ lệch của các tỷ lệ trong mẫu Bootstrap: $\sigma_b^* = \hat{p}_b^* - \hat{p}$
- Hai giá trị phân vị $\sigma_{\alpha/2}^* = -0.16$ và $\sigma_{1-\alpha/2}^* = 0.12$ của dãy sai khác σ_b^*.
- Khi đó khoảng ước lượng Bootstrap của p với độ tin cậy 0.9 là $\left[\hat{p} - \sigma_{1-\alpha/2}^*; \hat{p} - \sigma_{\alpha/2}^*\right] = [0.56; 0.84]$.

Ví dụ minh hoạ

Code Python:

```
import numpy as np
import pandas as pd
from sklearn.utils import resample
df = pd.read_csv("/content/drive/My Drive/Dataset/data1_lecture7.csv")
x = df.values[:,0]
xnew=x[0:25]
k=200
n_size = len(xnew)
count = sum(i > k for i in xnew)
pmu = count/n_size
print(pmu)
n_iterations = 500
# run bootstrap
sigma = []
for i in range(n_iterations):
  Xb = resample(xnew,n_samples=n_size)
  sigma.append(sum(j > 200 for j in Xb)/n_size - pmu)
# confidence interval
alpha = 1 - 0.9
Z1 = np.percentile(sigma, 100*alpha/2)
print(Z1)
Z2 = np.percentile(sigma, 100*(1-alpha/2))
print(Z2)
L = pmu - Z2
U = pmu - Z1
print(L, U)
```

Bài tập

Quan sát thời gian sống sót X (tính bằng ngày) của 72 con chuột lang bị nhiễm trực khuẩn lao, ta thu được bảng số liệu như sau:

12	15	22	24	24	32	32	33	34	38	38	43	44
48	52	53	54	54	55	56	57	58	58	59	60	60
60	60	61	62	63	65	65	67	68	70	70	72	73
75	76	76	81	83	84	85	87	91	95	96	98	99
109	110	121	127	129	131	143	146	146	175	175	211	233
258	258	263	297	341	341	376						

Bài tập

- Hãy vẽ biểu đồ tần suất cho tập dữ liệu trên.
- Từ biểu đồ tần suất, ta có thể giả sử X có phân phối mũ (exponential distribution) với tham số $\theta > 0$ và hàm mật độ xác suất như sau: $f(x;\theta) = \theta e^{-\theta x}, x > 0$
  - Hãy tính ước lượng điểm của θ bằng phương pháp mô-men.
  - Hãy tính ước lượng điểm của θ bằng phương pháp hợp lý cực đại.
  - Áp dụng định lý giới hạn trung tâm, hãy tính ước lượng khoảng của θ với độ tin cậy 90%.
  - Giả sử chỉ có 10 quan sát đầu trong tập số liệu trên. Hãy tính ước lượng khoảng Bootstrap của θ với độ tin cậy 90%.