PHÂN BỐ MẪU NGẪU NHIÊN (SAMPLING DISTRIBUTION)

Nguyễn Văn Hạnh

AI Academy Vietnam

December 04, 2020

Nội dung

Tổng thể (population) và mẫu (sample)

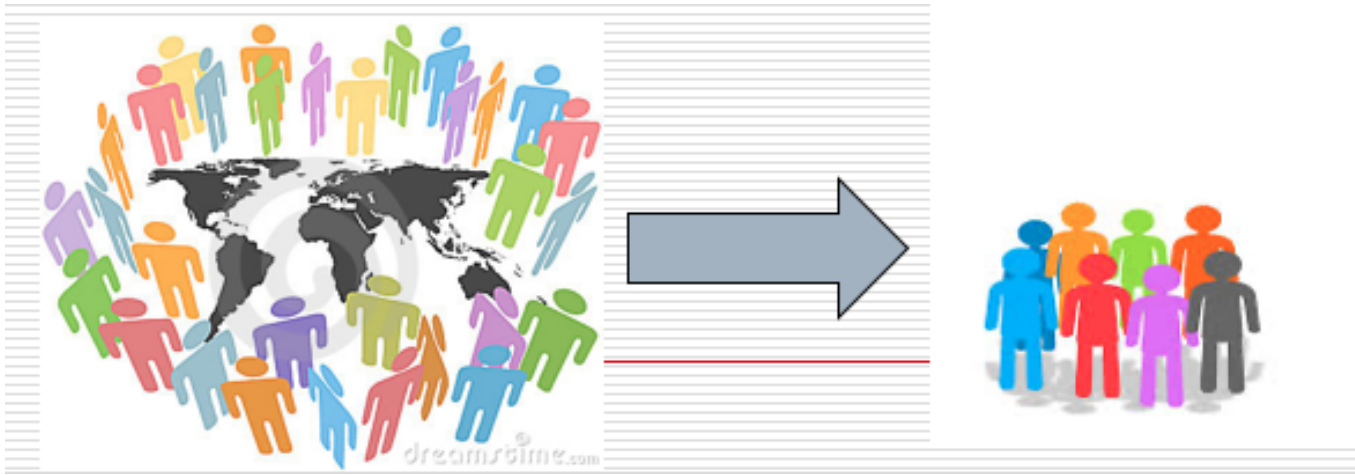
- Tổng thể là tập hợp tất cả cá thể mà chúng ta muốn nghiên cứu.

Tổng thể (population) và mẫu (sample)

- Tổng thể là tập hợp tất cả cá thể mà chúng ta muốn nghiên cứu.
- Kích thước N của tổng thể là số cá thể trong tổng thể (thường N rất lớn).

Tổng thể (population) và mẫu (sample)

- Tổng thể là tập hợp tất cả cá thể mà chúng ta muốn nghiên cứu.
- Kích thước N của tổng thể là số cá thể trong tổng thể (thường N rất lớn).
- Một mẫu kích thước n là một tập con n cá thể lấy ra từ tổng thể.



Mẫu ngẫu nhiên (random sample)

- Giả sử chúng ta nghiên cứu một tổng thể X có phân bố là $f(x)$ (là hàm mật độ xác suất (probability density function) hoặc hàm khối xác suất (probability mass function)).

Mẫu ngẫu nhiên (random sample)

- Giả sử chúng ta nghiên cứu một tổng thể X có phân bố là $f(x)$ (là hàm mật độ xác suất (probability density function) hoặc hàm khối xác suất (probability mass function)).
- Một véctơ ngẫu nhiên $(X_1, X_2, \ldots, X_n)$, với X_i là các biến ngẫu nhiên độc lập và có cùng phân phối $f(x)$, được gọi là một mẫu ngẫu nhiên kích thước n lấy từ tổng thể X.

Mẫu ngẫu nhiên (random sample)

- Giả sử chúng ta nghiên cứu một tổng thể X có phân bố là $f(x)$ (là hàm mật độ xác suất (probability density function) hoặc hàm khối xác suất (probability mass function)).
- Một véctơ ngẫu nhiên $(X_1, X_2, \ldots, X_n)$, với X_i là các biến ngẫu nhiên độc lập và có cùng phân phối $f(x)$, được gọi là một mẫu ngẫu nhiên kích thước n lấy từ tổng thể X.
- Phân bố xác suất đồng thời (joint probability distribution) của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$ là

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$$

Ví dụ

- Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực (gồm khoảng 1 triệu hộ). Tổng thể là tập hợp gồm tiền điện của 1 triệu hộ.

Ví dụ

- Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực (gồm khoảng 1 triệu hộ). Tổng thể là tập hợp gồm tiền điện của 1 triệu hộ.
- Giả sử X có phân bố chuẩn với kỳ vọng μ và phương sai σ^2, với hàm mật độ xác suất $f(x; \mu, \sigma^2)$.

Ví dụ

- Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực (gồm khoảng 1 triệu hộ). Tổng thể là tập hợp gồm tiền điện của 1 triệu hộ.
- Giả sử X có phân bố chuẩn với kỳ vọng μ và phương sai σ^2, với hàm mật độ xác suất $f(x; \mu, \sigma^2)$.
- Một véctơ ngẫu nhiên $(X_1, X_2, \ldots, X_{50})$, với X_i là các biến ngẫu nhiên độc lập và có cùng phân phối chuẩn $f(x; \mu, \sigma^2)$, được gọi là một mẫu ngẫu nhiên kích thước 50 lấy từ tổng thể X.

Ví dụ

- Gọi X là tiền điện trong tháng 6/2020 (nghìn đồng) của các hộ gia đình cá nhân tại một khu vực (gồm khoảng 1 triệu hộ). Tổng thể là tập hợp gồm tiền điện của 1 triệu hộ.
- Giả sử X có phân bố chuẩn với kỳ vọng μ và phương sai σ^2, với hàm mật độ xác suất $f(x; \mu, \sigma^2)$.
- Một véctơ ngẫu nhiên $(X_1, X_2, \ldots, X_{50})$, với X_i là các biến ngẫu nhiên độc lập và có cùng phân phối chuẩn $f(x; \mu, \sigma^2)$, được gọi là một mẫu ngẫu nhiên kích thước 50 lấy từ tổng thể X.
- Khảo sát tiền điện của 50 hộ ta có một mẫu $(x_1, x_2, \ldots, x_{50}) = (255, 367, \ldots, 423)$, đây là một giá trị của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_{50})$.

Thống kê (Statistic)

- Thống kê là một hàm $f(X_1, X_2, \ldots, X_n)$ của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.

Thống kê (Statistic)

- Thống kê là một hàm $f(X_1, X_2, \ldots, X_n)$ của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.
- Ví dụ trung bình mẫu ngẫu nhiên $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ là một thống kê.

Thống kê (Statistic)

- Thống kê là một hàm $f(X_1, X_2, \ldots, X_n)$ của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.
- Ví dụ trung bình mẫu ngẫu nhiên $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ là một thống kê.
- Thống kê $f(X_1, X_2, \ldots, X_n)$ là một biến ngẫu nhiên.

Thống kê (Statistic)

- Thống kê là một hàm $f(X_1, X_2, \ldots, X_n)$ của mẫu ngẫu nhiên $(X_1, X_2, \ldots, X_n)$.
- Ví dụ trung bình mẫu ngẫu nhiên $\bar{X} = \frac{X_1+X_2+\ldots+X_n}{n}$ là một thống kê.
- Thống kê $f(X_1, X_2, \ldots, X_n)$ là một biến ngẫu nhiên.
- Phân bố mẫu (sampling distribution) là phân bố xác suất của một thống kê.

Một số thống kê cơ bản

- Trung bình mẫu ngẫu nhiên (sample mean): $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$

Một số thống kê cơ bản

- Trung bình mẫu ngẫu nhiên (sample mean): $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$
- Phương sai mẫu ngẫu nhiên (sample variance)
  $S^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$

Một số thống kê cơ bản

- Trung bình mẫu ngẫu nhiên (sample mean): $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$
- Phương sai mẫu ngẫu nhiên (sample variance)
  $S^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$
- Độ lệch chuẩn mẫu ngẫu nhiên (sample standard deviation) $S = \sqrt{S^2}$

Một số thống kê cơ bản

- Trung bình mẫu ngẫu nhiên (sample mean): $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$
- Phương sai mẫu ngẫu nhiên (sample variance)
  $S^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$
- Độ lệch chuẩn mẫu ngẫu nhiên (sample standard deviation) $S = \sqrt{S^2}$
- Thống kê Z:

$$Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

Một số thống kê cơ bản

- Trung bình mẫu ngẫu nhiên (sample mean): $\bar{X} = \frac{X_1+X_2+...+X_n}{n}$
- Phương sai mẫu ngẫu nhiên (sample variance) $S^2 = \frac{(X_1-\bar{X})^2+...+(X_n-\bar{X})^2}{n-1}$
- Độ lệch chuẩn mẫu ngẫu nhiên (sample standard deviation) $S = \sqrt{S^2}$
- Thống kê Z:

$$Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

- Thống kê T:

$$T = \frac{\bar{X} - \mu}{S}\sqrt{n}$$

Phân bố của trung bình mẫu ngẫu nhiên

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.

Phân bố của trung bình mẫu ngẫu nhiên

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.
- Khi đó $\mathbb{E}(\bar{X}) = \mu$ và $\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}$

Phân bố của trung bình mẫu ngẫu nhiên

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.
- Khi đó $\mathbb{E}(\bar{X}) = \mu$ và $\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}$
- Nếu tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$ thì $\bar{X}$ có phân bố chuẩn $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Phân bố của trung bình mẫu ngẫu nhiên

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.
- Khi đó $\mathbb{E}(\bar{X}) = \mu$ và $\mathbb{V}(\bar{X}) = \frac{\sigma^2}{n}$
- Nếu tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$ thì $\bar{X}$ có phân bố chuẩn $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.
- Nếu tổng thể X không có phân bố chuẩn thì $\bar{X}$ xấp xỉ phân bố chuẩn $\mathcal{N}(\mu, \frac{\sigma^2}{n})$, với n đủ (định lý giới hạn trung tâm).

Định lý giới hạn trung tâm (Central limit theorem)

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.

Định lý giới hạn trung tâm (Central limit theorem)

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.

- Nếu tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$ thì
  - Thống kê $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ có phân bố chuẩn tắc $\mathbb{N}(0, 1)$.
  - Thống kê $T = \frac{\bar{X}-\mu}{S}\sqrt{n}$ có phân bố Student với $n-1$ bậc tự do.

Định lý giới hạn trung tâm (Central limit theorem)

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2.

- Nếu tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$ thì
  - Thống kê $Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n}$ có phân bố chuẩn tắc $\mathbb{N}(0,1)$.
  - Thống kê $T = \frac{\bar{X}-\mu}{S}\sqrt{n}$ có phân bố Student với $n-1$ bậc tự do.
- Tổng quát, với mọi phân bố của tổng thể X thì:
  - Định lý giới hạn trung tâm:

$$Z = \frac{\bar{X}-\mu}{\sigma}\sqrt{n} \to \mathbb{N}(0,1)$$

  - Định lý giới hạn trung tâm và định lý Slutky:

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

Bất đẳng thức Chebyshev (Chebyshev's inequality)

Cho (X là một biến ngẫu nhiên có kỳ vọng μ và phương sai σ^2.

- Bất đẳng thức Chebyshev: với mọi $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t}$$

Bất đẳng thức Chebyshev (Chebyshev's inequality)

Cho (X là một biến ngẫu nhiên có kỳ vọng μ và phương sai σ^2.

- Bất đẳng thức Chebyshev: với mọi $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t}$$

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có kỳ vọng μ và phương sai σ^2: khi đó với mọi $t > 0$

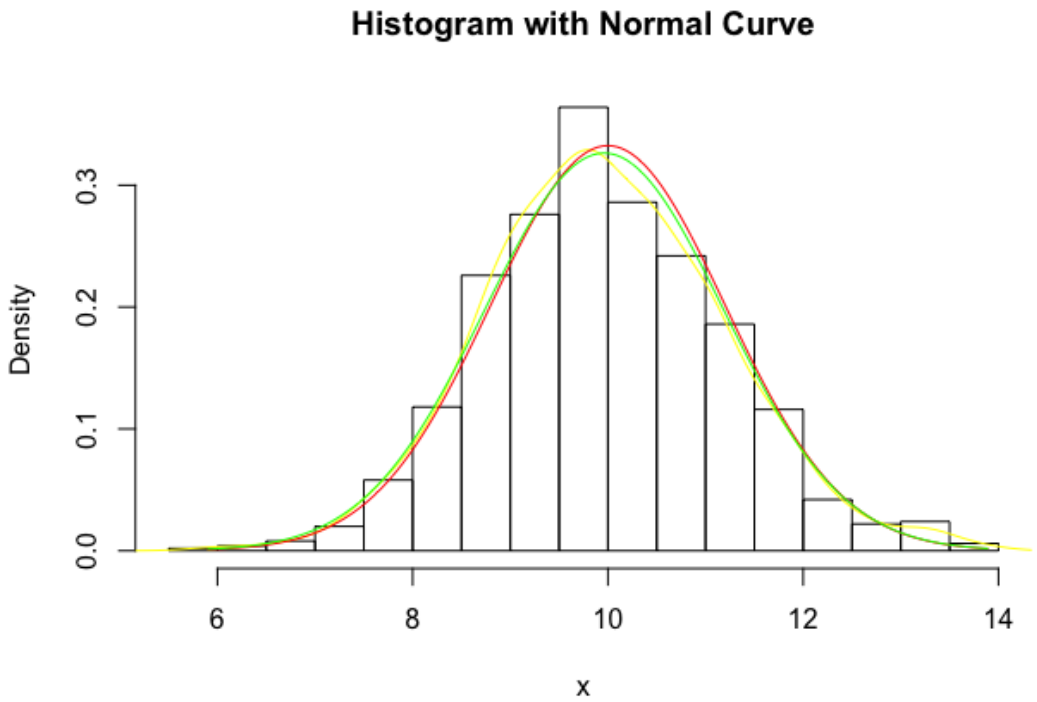
$$\mathbb{P}(|\bar{X} - \mu| \geq t) \leq \frac{\sigma^2}{nt}$$

Ví dụ mô phỏng phân bố của tổng thể (1)

- Mô phỏng một mẫu ngẫu nhiên có kích thước 1000 từ tổng thể có phân bố chuẩn $\mathcal{N}(10, 1.2^2)$

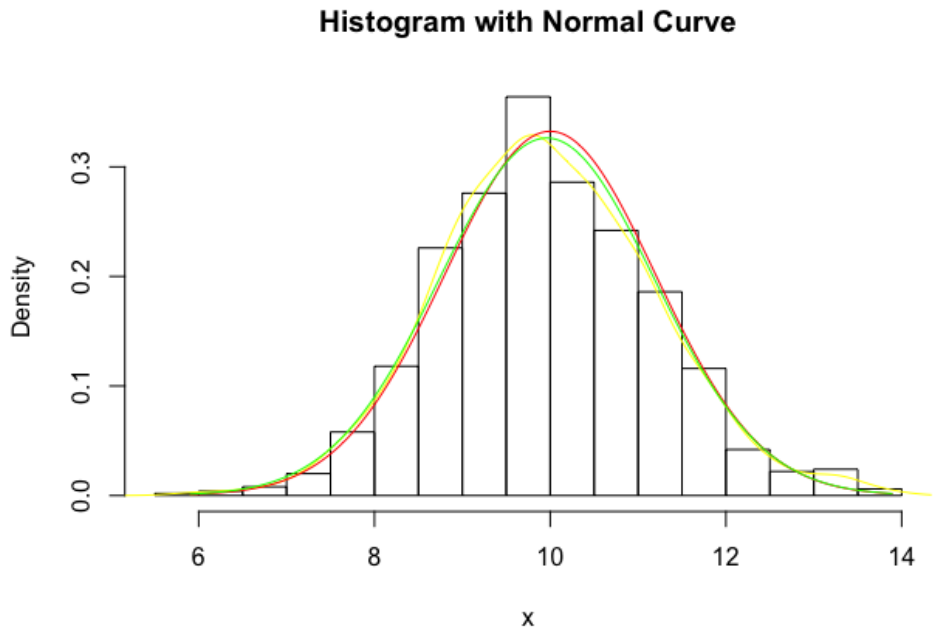
Ví dụ mô phỏng phân bố của tổng thể (1)

- Mô phỏng một mẫu ngẫu nhiên có kích thước 1000 từ tổng thể có phân bố chuẩn $\mathcal{N}(10, 1.2^2)$
- Biểu đồ tần suất, hàm mật độ của hai phân bố chuẩn $\mathcal{N}(10, 1.2^2)$ (đỏ) và $\mathcal{N}(\bar{x}, s^2)$ (xanh) và ước lượng hạt nhân của hàm mật độ (vàng) được vẽ đồng thời như sau



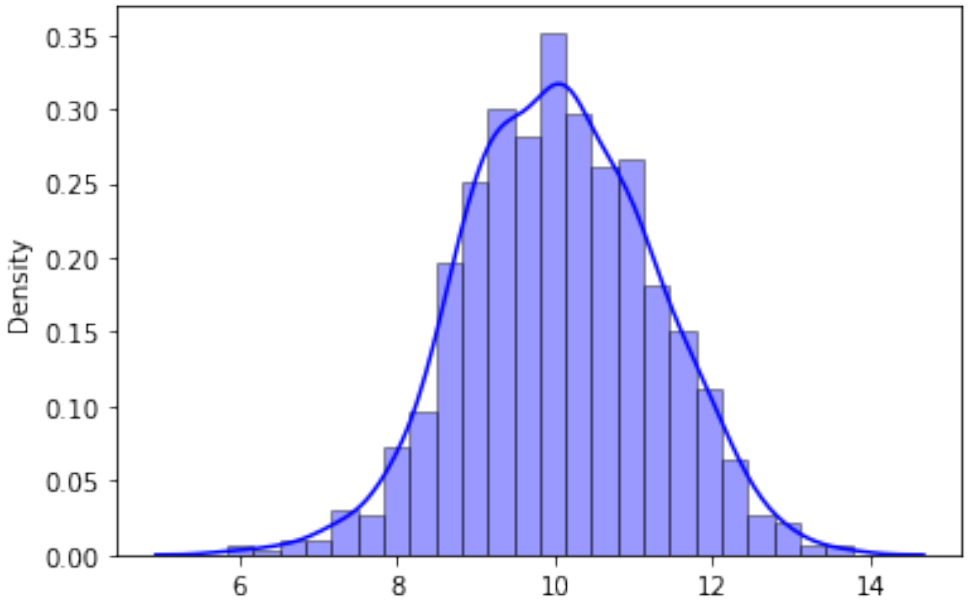
Ví dụ mô phỏng phân bố của tổng thể (1): Code trên R

```r
# phan bo cua tong the (phan bo chuan)
x=rnorm(1000,mean=10,sd=1.2)
hist(x, freq=F, breaks=12, main="Histogram with Normal Curve")
lines(density(x), col="yellow")
lines(seq(min(x), max(x), by=.05), dnorm(seq(min(x), max(x), by=.05),
                                         mean=mu,sd=sigma), col="red")
lines(seq(min(x), max(x), by=.05), dnorm(seq(min(x), max(x), by=.05),
                                         mean(x), sd(x)), col="green")
```



Ví dụ mô phỏng phân bố của tổng thể (1): Code trên Python

```
import seaborn as sns
from scipy.stats import norm
x = norm.rvs(size=1000,loc=10,scale=1.2)
sns.distplot(x, hist=True, kde=True, color = 'blue',
             hist_kws={'edgecolor':'black'})
```

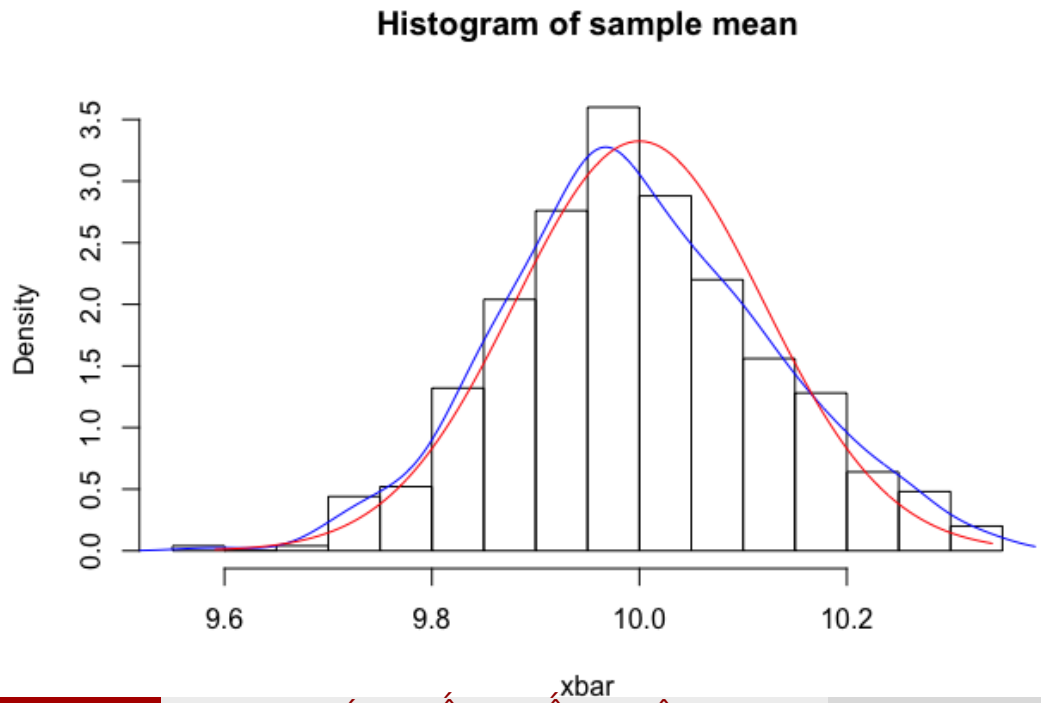


Ví dụ mô phỏng phân bố của trung bình mẫu (1)

- Mô phỏng 500 mẫu ngẫu nhiên có kích thước 100 từ tổng thể có phân bố chuẩn $\mathcal{N}(10, 1.2^2)$

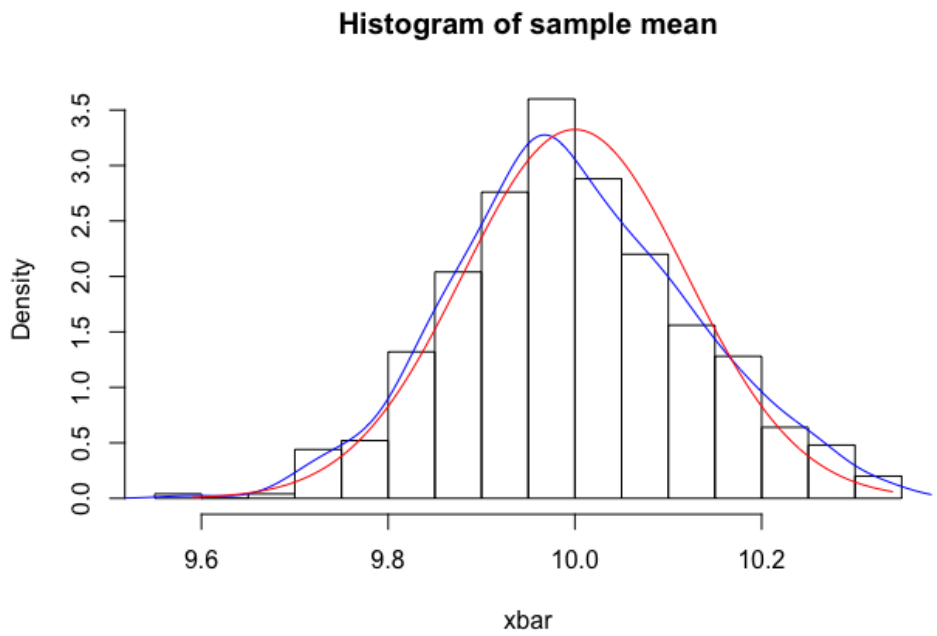
Ví dụ mô phỏng phân bố của trung bình mẫu (1)

- Mô phỏng 500 mẫu ngẫu nhiên có kích thước 100 từ tổng thể có phân bố chuẩn $\mathcal{N}(10, 1.2^2)$
- Tính giá trị trung bình mẫu cho từng mẫu được 500 giá trị của $\bar{X}$.
- Biểu đồ tần suất của 500 giá trị của $\bar{X}$, hàm mật độ của phân bố chuẩn $\mathcal{N}(10, \frac{1.2^2}{100})$ (đỏ) và ước lượng hạt nhân của hàm mật độ (xanh) được vẽ đồng thời như sau



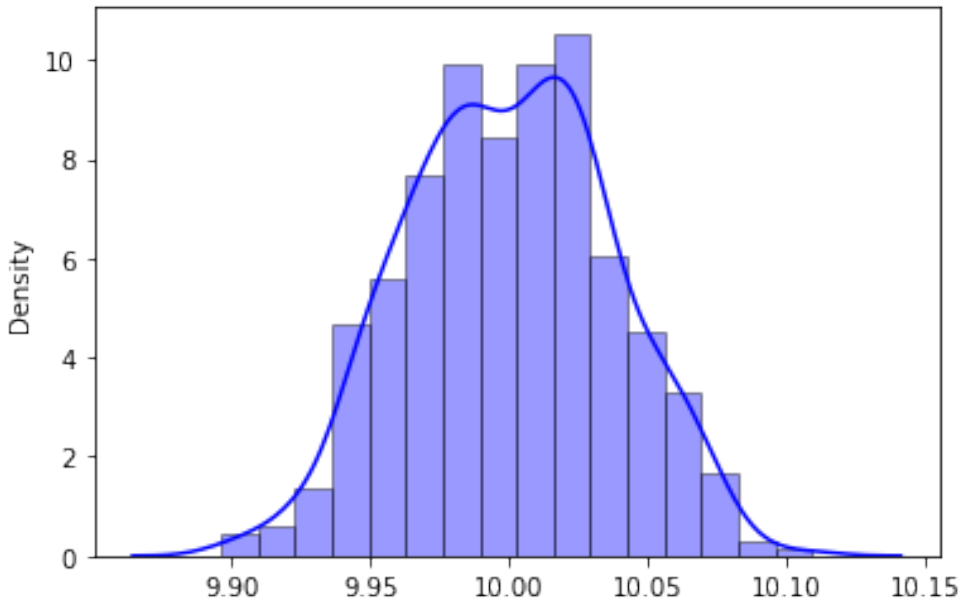
Ví dụ mô phỏng phân bố của trung bình mẫu (1): Code trên R

```
xbar = rep(0,500)
for (i in 1:500){
  u=rnorm(100,mean = 10,sd=1.2)
  xbar[i] = mean(u)
  }
hist(xbar, freq=F, breaks=12, main="Histogram of sample mean")
lines(density(xbar), col="blue")
lines(seq(min(xbar), max(xbar), by=.005), dnorm(seq(min(xbar),
      max(xbar), by=.005),mean=10,sd=1.2/sqrt(100)), col="red")
```



Ví dụ mô phỏng phân bố của trung bình mẫu (1): Code trên Python

```
import seaborn as sns
from scipy.stats import norm
import numpy as np
data = []
for i in range(0, 499):
    x = norm.rvs(size=1000,loc=10,scale=1.2)
    data.append(np.mean(x))
sns.distplot(data, hist=True, kde=True, color = 'blue',hist_kws={'edgecolor':'black'})
```

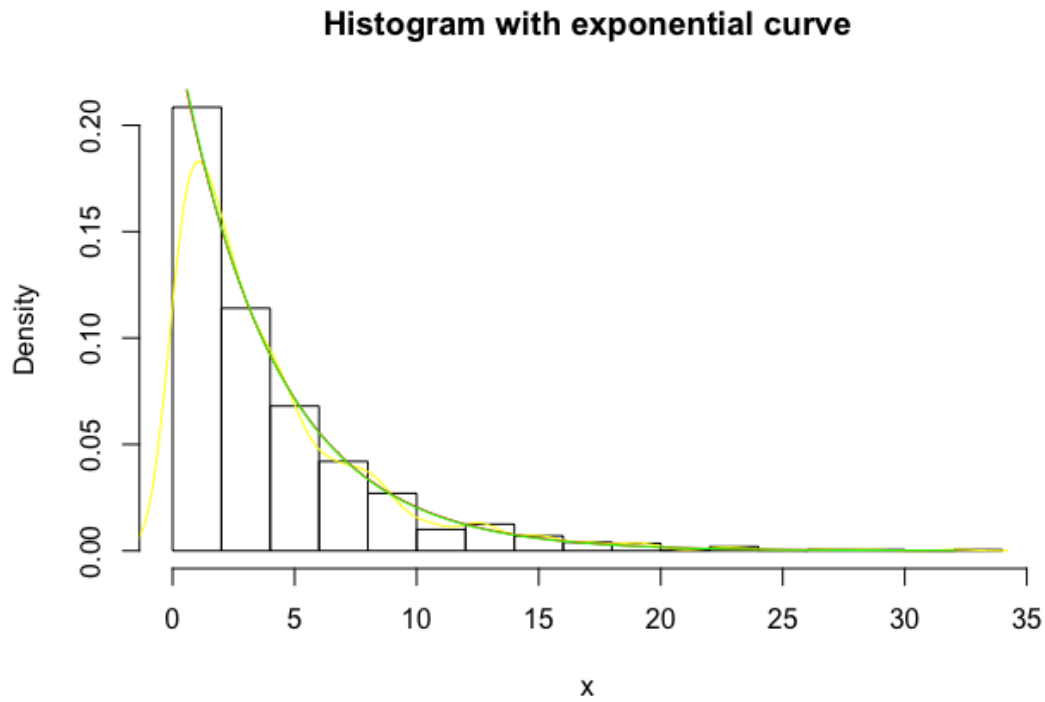


Ví dụ mô phỏng phân bố của tổng thể (2)

- Mô phỏng một mẫu ngẫu nhiên có kích thước 1000 từ tổng thể có phân bố mũ $\mathcal{E}(\frac{1}{4})$, tức là $\mathbb{E}(X) = 4$ và $\lambda = 1/4$.

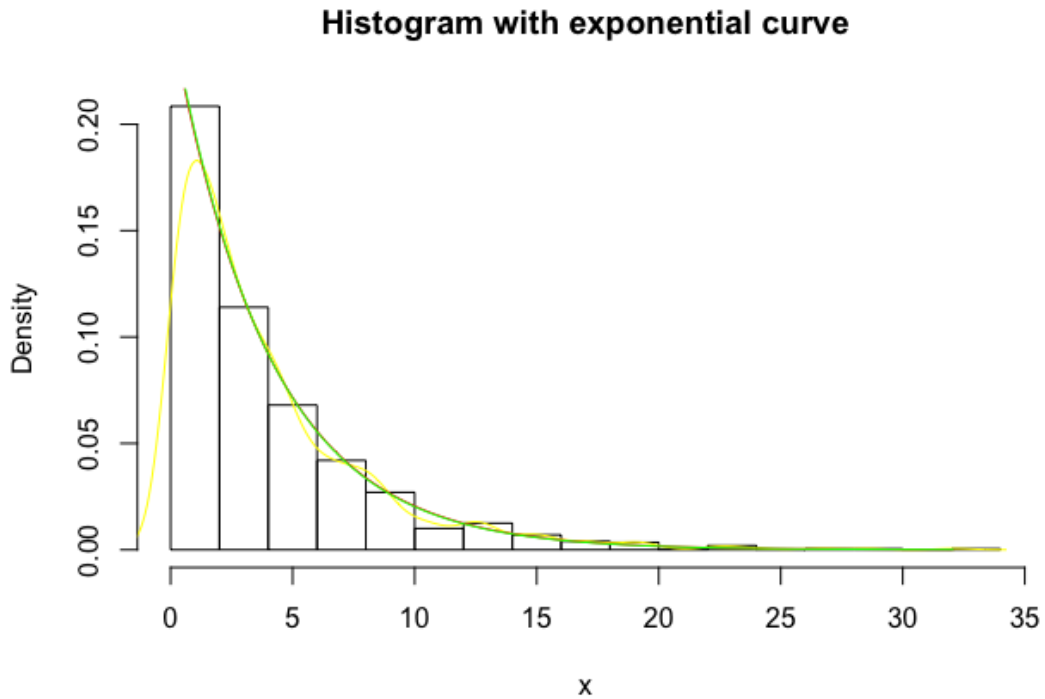
Ví dụ mô phỏng phân bố của tổng thể (2)

- Mô phỏng một mẫu ngẫu nhiên có kích thước 1000 từ tổng thể có phân bố mũ $\mathcal{E}(\frac{1}{4})$, tức là $\mathbb{E}(X) = 4$ và $\lambda = 1/4$.
- Biểu đồ tần suất, hàm mật độ của hai phân bố mũ $\mathcal{E}(\frac{1}{4})$ (đỏ) và $\mathcal{E}(\frac{1}{\bar{x}})$ (xanh) và ước lượng hạt nhân của hàm mật độ (vàng) được vẽ đồng thời như sau



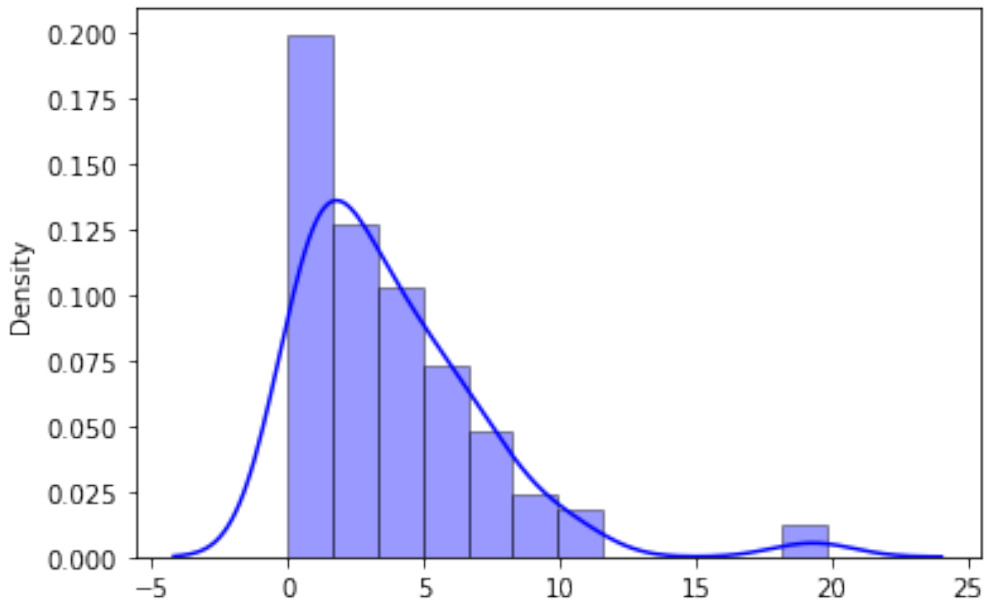
Ví dụ mô phỏng phân bố của tổng thể (2): Code trên R

```
x=rexp(1000,rate=1/4)
hist(x, freq=F, breaks=12, main="Histogram with exponential curve")
lines(density(x), col="yellow")
lines(seq(min(x), max(x), by=.05), dexp(seq(min(x), max(x), by=.05),
                                        rate=1/4), col="red")
lines(seq(min(x), max(x), by=.05), dexp(seq(min(x), max(x), by=.05),
                                        rate=1/mean(x)), col="green")
```



Ví dụ mô phỏng phân bố của tổng thể (2): Code trên Python

```
import seaborn as sns
from scipy.stats import expon
x = expon.rvs(size=100,scale= 4)
sns.distplot(x, hist=True, kde=True, color = 'blue',
             hist_kws={'edgecolor':'black'})
```

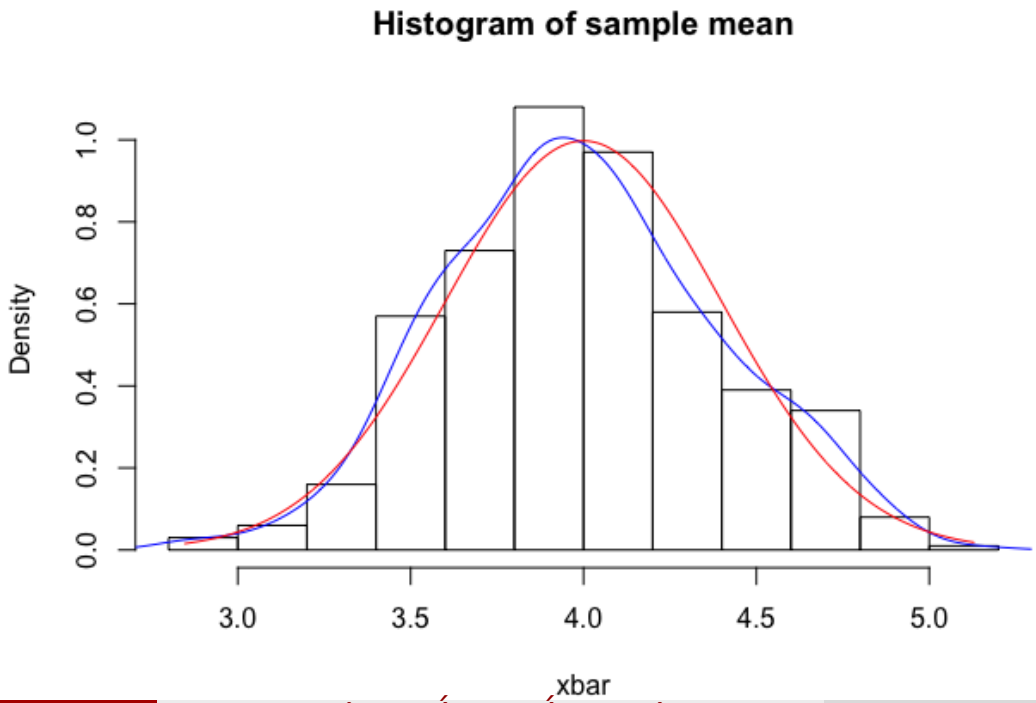


Ví dụ mô phỏng phân bố của trung bình mẫu (2)

- Mô phỏng 500 mẫu ngẫu nhiên có kích thước 100 từ tổng thể có phân bố mũ chuẩn $\mathcal{E}(\frac{1}{4})$. (kỳ vọng $\mathbb{E}(X) = 4$ và phương sai $\mathbb{V}(X) = 4^2$).

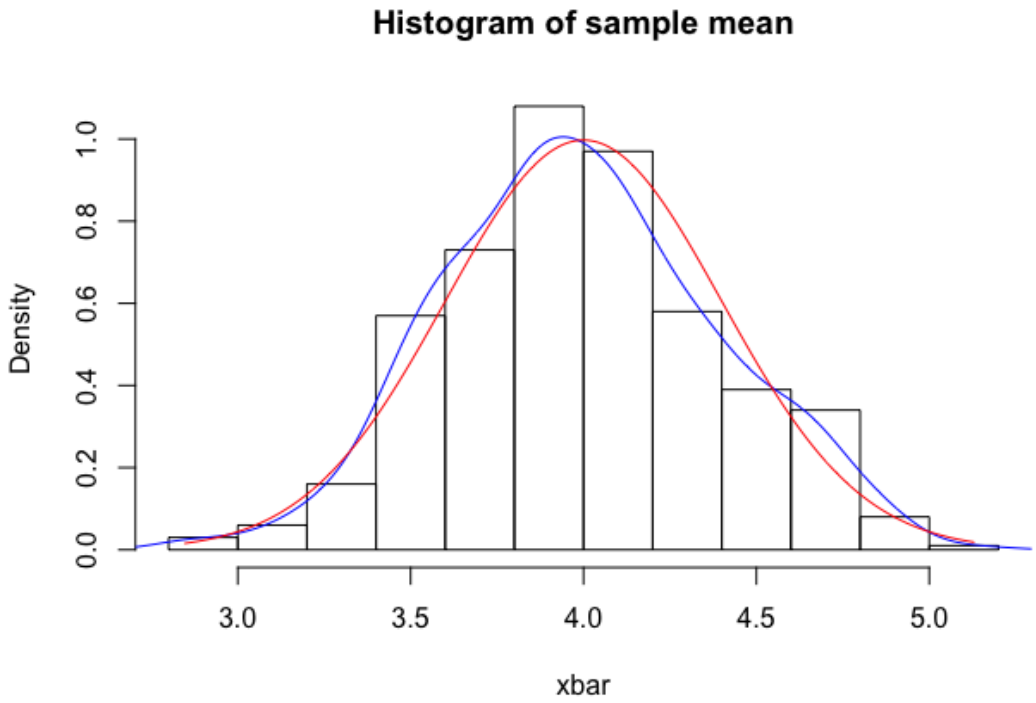
Ví dụ mô phỏng phân bố của trung bình mẫu (2)

- Mô phỏng 500 mẫu ngẫu nhiên có kích thước 100 từ tổng thể có phân bố mũ chuẩn $\mathcal{E}(\frac{1}{4})$. (kỳ vọng $\mathbb{E}(X) = 4$ và phương sai $\mathbb{V}(X) = 4^2$).
- Tính giá trị trung bình mẫu cho từng mẫu được 500 giá trị của $\bar{X}$.
- Biểu đồ tần suất của 500 giá trị của $\bar{X}$, hàm mật độ của phân bố chuẩn $\mathcal{N}(4, \frac{4^2}{100})$ (đỏ) và ước lượng hạt nhân của hàm mật độ (xanh) được vẽ đồng thời như sau



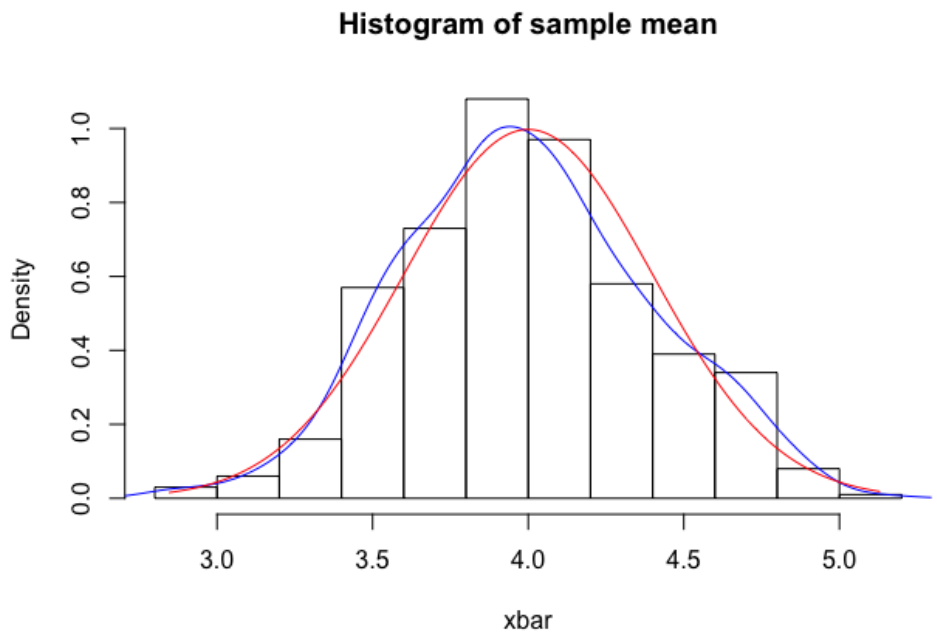
Ví dụ mô phỏng phân bố của trung bình mẫu (2)

- Ta thấy tổng thể X không có phân bố chuẩn nhưng trung bình mẫu $\bar{X}$ có phân bố xấp xỉ chuẩn.



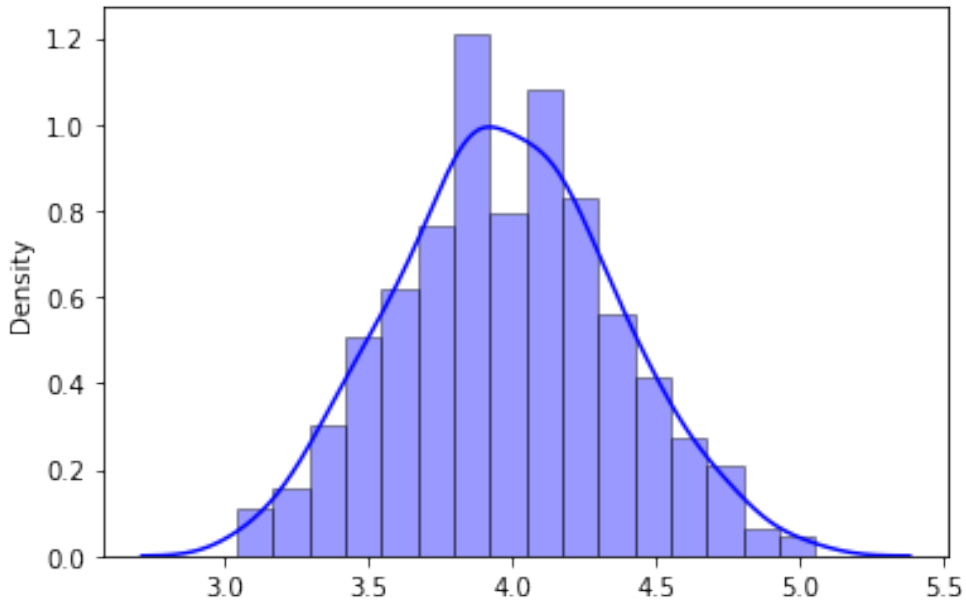
Ví dụ mô phỏng phân bố của trung bình mẫu (2): Code trên R

```
xbar = rep(0,500)
for (i in 1:500){
  u=rexp(100,rate=1/4)
  xbar[i] = mean(u)
  }
hist(xbar, freq=F, breaks=12, main="Histogram of sample mean")
lines(density(xbar), col="blue")
lines(seq(min(xbar), max(xbar), by=.005), dnorm(seq(min(xbar), max(xbar),
      by=.005),mean = 4,sd= 4/sqrt(100)), col="red")
```



Ví dụ mô phỏng phân bố của trung bình mẫu (2): Code trên Python

```
import seaborn as sns
from scipy.stats import expon
import numpy as np
data = []
for i in range(0, 499):
    x = expon.rvs(size=100,scale=4)
    data.append(np.mean(x))
sns.distplot(data, hist=True, kde=True, color = 'blue',
             hist_kws={'edgecolor':'black'})
```



Ví dụ minh hoạ 1

Ví dụ: Một nhà môi giới chứng khoán đã quan sát thấy rằng lợi nhuận một năm trong ngành xây dựng là một biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là 12.5% và độ lệch chuẩn là 2.5%.

Ví dụ minh hoạ 1

Ví dụ: Một nhà môi giới chứng khoán đã quan sát thấy rằng lợi nhuận một năm trong ngành xây dựng là một biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là 12.5% và độ lệch chuẩn là 2.5%.

- Tính xác suất để lợi nhuận một năm của một mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.

Ví dụ minh hoạ 1

Ví dụ: Một nhà môi giới chứng khoán đã quan sát thấy rằng lợi nhuận một năm trong ngành xây dựng là một biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là 12.5% và độ lệch chuẩn là 2.5%.

- Tính xác suất để lợi nhuận một năm của một mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.
  Lời giải: Ta có $X \sim \mathcal{N}(12.5; 2.5^2)$ nên $Z = \frac{X-12.5}{2.5} \sim \mathcal{N}(0; 1)$

Ví dụ minh hoạ 1

Ví dụ: Một nhà môi giới chứng khoán đã quan sát thấy rằng lợi nhuận một năm trong ngành xây dựng là một biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là 12.5% và độ lệch chuẩn là 2.5%.

- Tính xác suất để lợi nhuận một năm của một mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.
  Lời giải: Ta có $X \sim \mathcal{N}(12.5; 2.5^2)$ nên $Z = \frac{X-12.5}{2.5} \sim \mathcal{N}(0; 1)$

$$\mathbb{P}(X < 10.825) = \mathbb{P}\left(\frac{X - 12.5}{2.5} < -0.67\right) = \mathbb{P}(Z < -0.67) = 0.25$$

Ví dụ minh hoạ 1

Ví dụ: Một nhà môi giới chứng khoán đã quan sát thấy rằng lợi nhuận một năm trong ngành xây dựng là một biến ngẫu nhiên có phân phối chuẩn với giá trị trung bình là 12.5% và độ lệch chuẩn là 2.5%.

- Tính xác suất để lợi nhuận một năm của một mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.
  Lời giải: Ta có $X \sim \mathcal{N}(12.5; 2.5^2)$ nên $Z = \frac{X-12.5}{2.5} \sim \mathcal{N}(0; 1)$

$$\mathbb{P}(X < 10.825) = \mathbb{P}\left(\frac{X - 12.5}{2.5} < -0.67\right) = \mathbb{P}(Z < -0.67) = 0.25$$

- Code Python: `norm.cdf(10.825, 12.5, 2.5)`
- Code R: `pnorm(10.825,12.5,2.5)`

Ví dụ minh hoạ 1

- Tính xác suất để lợi nhuận trung bình một năm của 4 mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.
  Lời giải: Gọi $\bar{X}$ là lợi nhuận trung bình một năm của 4 mã chứng khoán được chọn. Ta có $\bar{X} \sim \mathcal{N}(12.5; 2.5^2/4)$ nên $Z = \frac{\bar{X}-12.5}{2.5/\sqrt{4}} \sim \mathcal{N}(0;1)$.

Ví dụ minh hoạ 1

- Tính xác suất để lợi nhuận trung bình một năm của 4 mã chứng khoán được chọn ngẫu nhiên trong ngành xây dựng nhỏ hơn 10.825%.
  Lời giải: Gọi $\bar{X}$ là lợi nhuận trung bình một năm của 4 mã chứng khoán được chọn. Ta có $\bar{X} \sim \mathcal{N}(12.5; 2.5^2/4)$ nên $Z = \frac{\bar{X}-12.5}{2.5/\sqrt{4}} \sim \mathcal{N}(0;1)$. Khi đó

$$\mathbb{P}(\bar{X} < 10.825) = \mathbb{P}\left(\frac{\bar{X} - 12.5}{2.5/2} < -1.34\right) = \mathbb{P}(Z < -1.34) = 0.09$$

- Code Python: `norm.cdf(10.825, 12.5, 2.5/2)`
- Code R: `pnorm(10.825,12.5,2.5/2)`

Ví dụ minh hoạ 2

Ví dụ: Gọi X là khoảng thời gian (tính bằng phút) mà nhân viên bưu điện dành cho khách hàng của mình. Giả sử X có phân phối mũ với lượng thời gian trung bình là 4 phút.

Ví dụ minh hoạ 2

Ví dụ: Gọi X là khoảng thời gian (tính bằng phút) mà nhân viên bưu điện dành cho khách hàng của mình. Giả sử X có phân phối mũ với lượng thời gian trung bình là 4 phút.

- Tính xác suất để thời gian mà nhân viên bưu điện dành cho một khách hàng nhiều hơn 4.5 phút.

Ví dụ minh hoạ 2

Ví dụ: Gọi X là khoảng thời gian (tính bằng phút) mà nhân viên bưu điện dành cho khách hàng của mình. Giả sử X có phân phối mũ với lượng thời gian trung bình là 4 phút.

- Tính xác suất để thời gian mà nhân viên bưu điện dành cho một khách hàng nhiều hơn 4.5 phút.
- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng nhiều hơn 4.5 phút.

Ví dụ minh hoạ 2

Ví dụ: Gọi X là khoảng thời gian (tính bằng phút) mà nhân viên bưu điện dành cho khách hàng của mình. Giả sử X có phân phối mũ với lượng thời gian trung bình là 4 phút.

- Tính xác suất để thời gian mà nhân viên bưu điện dành cho một khách hàng nhiều hơn 4.5 phút.
- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng nhiều hơn 4.5 phút.
- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 500 khách hàng nhiều hơn 4.5 phút.

Ví dụ minh hoạ 2

- Tính xác suất để thời gian mà nhân viên bưu điện dành cho một khách hàng nhiều hơn 4.5 phút.
  Lời giải: Tham số tỷ lệ của phân bố mũ của X là $\theta = 1/\mu = 1/4$.
  Hàm mật độ xác suất của X là

$$f(x;\theta) = \theta e^{-\theta x} = 0.25e^{-0.25x}, \text{ với } x \geq 0.$$

Ví dụ minh hoạ 2

- Tính xác suất để thời gian mà nhân viên bưu điện dành cho một khách hàng nhiều hơn 4.5 phút.
  Lời giải: Tham số tỷ lệ của phân bố mũ của X là $\theta = 1/\mu = 1/4$.
  Hàm mật độ xác suất của X là

$$f(x;\theta) = \theta e^{-\theta x} = 0.25e^{-0.25x}, \text{ với } x \geq 0.$$

  Khi đó

$$\mathbb{P}(X > 4.5) = \int_{4.5}^{+\infty} 0.25e^{-0.25x} dx = 0.325$$

- Code Python: `1-expon.cdf(4.5, scale =4)`
- Code R: `1-pexp(4.5,rate=0.25)`

Ví dụ minh hoạ 2

- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng nhiều hơn 4.5 phút.
  Lời giải: Gọi $\bar{X}$ là thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng. Ta có $\bar{X}$ xấp xỉ phân bố chuẩn $\mathcal{N}(\mu; \sigma^2/n) = \mathcal{N}(4; 4^2/50)$ nên $Z = \frac{\bar{X}-4}{4/\sqrt{50}} \approx \mathcal{N}(0; 1)$

Ví dụ minh hoạ 2

- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng nhiều hơn 4.5 phút.
  Lời giải: Gọi $\bar{X}$ là thời gian trung bình mà nhân viên bưu điện dành cho 50 khách hàng. Ta có $\bar{X}$ xấp xỉ phân bố chuẩn $\mathcal{N}(\mu; \sigma^2/n) = \mathcal{N}(4; 4^2/50)$ nên $Z = \frac{\bar{X}-4}{4/\sqrt{50}} \approx \mathcal{N}(0; 1)$
  Khi đó

$$\mathbb{P}(\bar{X} > 4.5) = \mathbb{P}\left(\frac{\bar{X}-4}{4/\sqrt{50}} > 0.884\right) = 1 - \mathbb{P}(Z < 0.884) = 0.188$$

- Code Python: `1- norm.cdf(4.5, 4, 4/np.sqrt(50))`
- Code R: `1- pnorm(4.5,mean = 4,sd = 4/sqrt(50))`

Ví dụ minh hoạ 2

- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 500 khách hàng nhiều hơn 4.5 phút.
  Lời giải: Gọi $\bar{X}$ là thời gian trung bình mà nhân viên bưu điện dành cho 500 khách hàng. Ta có $\bar{X}$ xấp xỉ phân bố chuẩn $\mathcal{N}(\mu; \sigma^2/n) = \mathcal{N}(4; 4^2/500)$ nên $Z = \frac{\bar{X}-4}{4/\sqrt{500}} \approx \mathcal{N}(0; 1)$

Ví dụ minh hoạ 2

- Tính xác suất để thời gian trung bình mà nhân viên bưu điện dành cho 500 khách hàng nhiều hơn 4.5 phút.
  Lời giải: Gọi $\bar{X}$ là thời gian trung bình mà nhân viên bưu điện dành cho 500 khách hàng. Ta có $\bar{X}$ xấp xỉ phân bố chuẩn $\mathcal{N}(\mu; \sigma^2/n) = \mathcal{N}(4; 4^2/500)$ nên $Z = \frac{\bar{X}-4}{4/\sqrt{500}} \approx \mathcal{N}(0; 1)$
  Khi đó

$$\mathbb{P}(\bar{X} > 4.5) = \mathbb{P}\left(\frac{\bar{X}-4}{4/\sqrt{500}} > 2.795\right) = 1 - \mathbb{P}(Z < 2.795) = 0.00259$$

- Code Python: `1- norm.cdf(4.5, 4, 4/np.sqrt(500))`
- Code R: `1- pnorm(4.5,mean = 4,sd = 4/sqrt(500))`

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X và Y có trung bình và phương sai lần lượt là μ_1, μ_2 và σ_1^2, σ_2^2.

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X và Y có trung bình và phương sai lần lượt là μ_1, μ_2 và σ_1^2, σ_2^2.
- Nếu X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$ thì thống kê $Z = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ có phân bố chuẩn tắc $\mathcal{N}(0, 1)$

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X và Y có trung bình và phương sai lần lượt là μ_1, μ_2 và σ_1^2, σ_2^2.
- Nếu X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$ thì thống kê $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ có phân bố chuẩn tắc $\mathcal{N}(0, 1)$
- Nếu X và Y không có phân bố chuẩn nhưng kích thước 2 mẫu đủ lớn $(n, m > 30)$ thì thống kê $Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$ xấp xỉ phân bố chuẩn tắc $\mathcal{N}(0, 1)$.

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và tổng thể Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$.

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và tổng thể Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Nếu $\sigma_1 = \sigma_2$ thì thống kê $T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$, với

  $S^2 = \dfrac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ có phân bố Student với $n + m - 2$ bậc tự do.

Phân bố mẫu của hiệu hai trung bình

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và tổng thể Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$.
- Nếu $\sigma_1 = \sigma_2$ thì thống kê $T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2\left(\frac{1}{n} + \frac{1}{m}\right)}}$, với

  $S^2 = \dfrac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ có phân bố Student với $n + m - 2$ bậc tự do.
- Nếu $\sigma_1 \neq \sigma_2$ thì thống kê $T = \dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{m}}}$ có phân bố Student với

  ν bậc tự do, với $\nu = \dfrac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\left(\frac{S_1^2}{n}\right)^2/(n-1) + \left(\frac{S_2^2}{m}\right)^2/(m-1)}$

Ví dụ minh hoạ 3

Ví dụ: Giả sử thu nhập trung bình (USD) của các sinh viên có bằng MBA tốt nghiệp từ 2 trường đại học A và B lần lượt là 62000 và 60000, độ lệch chuẩn lần lượt là 10000 và 9000.

Ví dụ minh hoạ 3

Ví dụ: Giả sử thu nhập trung bình (USD) của các sinh viên có bằng MBA tốt nghiệp từ 2 trường đại học A và B lần lượt là 62000 và 60000, độ lệch chuẩn lần lượt là 10000 và 9000.

- Tính xác suất để thu nhập trung bình của 50 sinh viên tốt nghiệp MBA từ trường đại học A cao hơn thu nhập trung bình của 60 sinh viên tốt nghiệp MBA từ trường đại học B.

Ví dụ minh hoạ 3

Lời giải:

- Ta có kích thước 2 mẫu đủ lớn ($n = 50, m = 60$) nên thống kê

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{\bar{X} - \bar{Y} - 2000}{\sqrt{\frac{10000^2}{50} + \frac{9000^2}{60}}} \approx N(0,1)$$

- Khi đó

$$\mathbb{P}(\bar{X} > \bar{Y}) = \mathbb{P}\left(Z > \frac{0 - 2000}{\sqrt{\frac{10000^2}{50} + \frac{9000^2}{60}}}\right) = \mathbb{P}(Z > -1.093) = 0.863$$

- Code Python: `1- norm.cdf(0, 2000, np.sqrt(pow(10000,2)/50+pow(9000,2)/60))`

Phân bố của tỷ lệ trong mẫu

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố Bernoulli với tham số xác suất là p.

- Ước lượng điểm của xác suất p là tỷ lệ $\hat{p} = (X_1 + \ldots + X_n)/n$

Phân bố của tỷ lệ trong mẫu

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố Bernoulli với tham số xác suất là p.

- Ước lượng điểm của xác suất p là tỷ lệ $\hat{p} = (X_1 + \ldots + X_n)/n$
- Định lý giới hạn trung tâm:

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

Phân bố của tỷ lệ trong mẫu

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố Bernoulli với tham số xác suất là p.

- Ước lượng điểm của xác suất p là tỷ lệ $\hat{p} = (X_1 + \ldots + X_n)/n$
- Định lý giới hạn trung tâm:

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

- Khi đó $\hat{p}$ xấp xỉ phân bố chuẩn $\mathbb{N}\left(p, \frac{p(1-p)}{n}\right)$

Phân bố của tỷ lệ trong mẫu

Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố Bernoulli với tham số xác suất là p.

- Ước lượng điểm của xác suất p là tỷ lệ $\hat{p} = (X_1 + \ldots + X_n)/n$
- Định lý giới hạn trung tâm:

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

- Khi đó $\hat{p}$ xấp xỉ phân bố chuẩn $\mathbb{N}\left(p, \frac{p(1-p)}{n}\right)$
- Định lý giới hạn trung tâm và định lý Slutky:

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

Ví dụ minh hoạ 4

Ví dụ: Thương hiệu của công ty Laurier có thị phần là $53,1\%$. Trong một cuộc khảo sát, 1000 người tiêu dùng được hỏi họ thích nhãn hiệu nào hơn.

Ví dụ minh hoạ 4

Ví dụ: Thương hiệu của công ty Laurier có thị phần là $53,1\%$. Trong một cuộc khảo sát, 1000 người tiêu dùng được hỏi họ thích nhãn hiệu nào hơn.

- Tính xác suất để có hơn 50% người được hỏi nói rằng họ thích nhãn hiệu Laurier.

Lời giải: Gọi $\hat{p}$ là tỷ lệ người thích nhãn hiệu Laurier trong số 1000 người được hỏi.

- Ta có $\hat{p} \approx \mathbb{N}\left(p, \frac{p(1-p)}{n}\right) = \mathbb{N}(0.531, 0.000249)$ hay $Z = \frac{\hat{p}-0.531}{0.01578} \approx N(0,1)$
- Khi đó

$$\mathbb{P}(\hat{p} > 50\%) = \mathbb{P}\left(\frac{\hat{p} - 0.531}{0.01578} > -1.96\right) = \mathbb{P}(Z < 1.96) = 0.975$$

Phân bố của phương sai mẫu

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$. Khi đó, thống kê

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

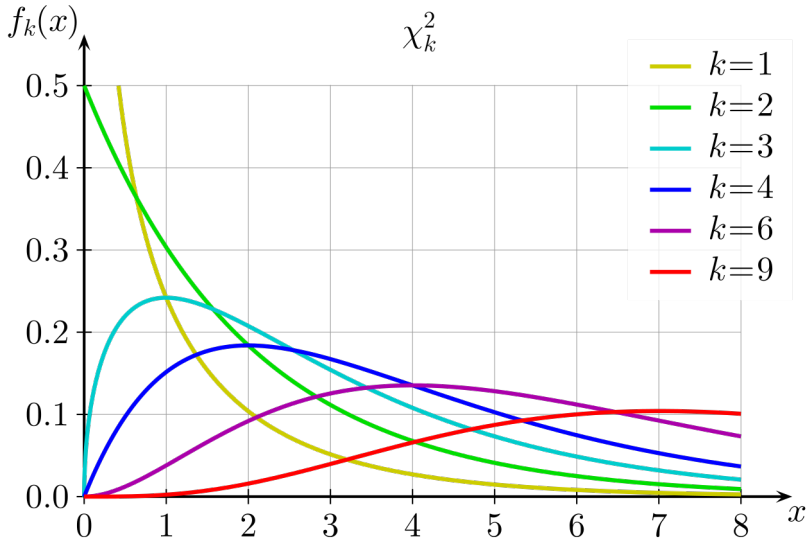
có phân bố Khi-bình phương với $n-1$ bậc tự do.

Phân bố của phương sai mẫu

- Cho $(X_1, X_2, \ldots, X_n)$ là một mẫu ngẫu nhiên lấy từ tổng thể X có phân bố chuẩn $\mathcal{N}(\mu, \sigma^2)$. Khi đó, thống kê

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

có phân bố Khi-bình phương với $n-1$ bậc tự do.



Phân bố của phương sai mẫu

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và tổng thể Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$. Khi đó thống kê

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

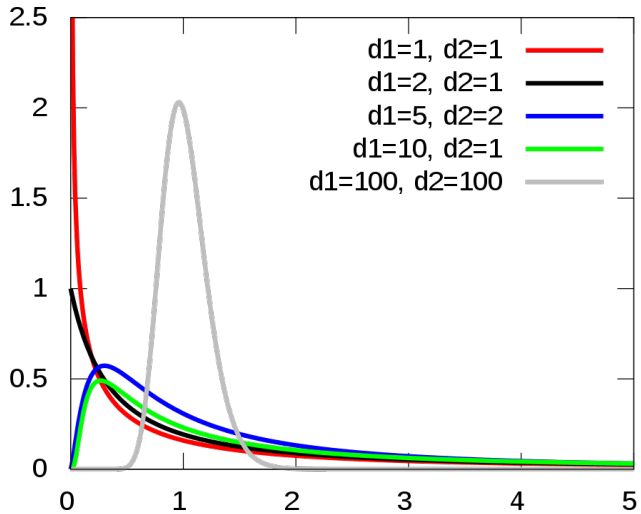
có phân bố Fisher với $n-1$ và $m-1$ bậc tự do.

Phân bố của phương sai mẫu

- Cho hai mẫu $(X_1, X_2, \ldots, X_n)$ và $(Y_1, Y_2, \ldots, Y_m)$ độc lập nhau và được lấy từ 2 tổng thể X có phân bố chuẩn $\mathcal{N}(\mu_1, \sigma_1^2)$ và tổng thể Y có phân bố chuẩn $\mathcal{N}(\mu_2, \sigma_2^2)$. Khi đó thống kê

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

có phân bố Fisher với $n-1$ và $m-1$ bậc tự do.



Bài tập 1

Biết rằng tuổi thọ của bóng đèn được sản xuất bởi một công ty có phân bố xấp xỉ chuẩn với giá trị trung bình bằng 800 giờ và độ lệch chuẩn là 40 giờ. Tìm xác suất để một mẫu ngẫu nhiên gồm 16 bóng đèn có tuổi thọ trung bình dưới 775 giờ.

- Tìm xác suất để một bóng đèn được chọn ngẫu nhiên từ nhà máy có tuổi thọ dưới 775 giờ.
- Tìm xác suất để một mẫu ngẫu nhiên gồm 16 bóng đèn có tuổi thọ trung bình dưới 775 giờ.
- Tìm xác suất để một mẫu ngẫu nhiên gồm 100 bóng đèn có tuổi thọ trung bình dưới 775 giờ.

Bài tập 2

Quan sát tuổi (X) của các học sinh trung học đăng ký tại một trường đại học lớn trong một học kỳ. Một mẫu ngẫu nhiên có kích thước 150 được lấy từ tổng thể này. Biết rằng tổng thể có trung bình là 21,1 tuổi và độ lệch chuẩn là 2,6. Phân bố của tổng thể không là phân bố chuẩn.

- Giả sử phân bố trung bình mẫu $\bar{X}$ là chuẩn có hợp lý không? Tại sao hoặc tại sao không?
- Tìm xác suất để trung bình mẫu $\bar{X}$ nằm trong khoảng từ 17,85 đến 25,65 tuổi.
- Tìm xác suất để trung bình mẫu $\bar{X}$ lớn hơn 25,91 tuổi.

Bài tập 3

Giả sử có hai loài sinh vật xanh trên sao Hỏa. Chiều cao trung bình của loài I là 32 cm trong khi chiều cao trung bình của loài II là 22 cm. Phương sai của hai loài lần lượt là 60 và 70 (cm^2) và chiều cao của cả hai loài đều phân bố chuẩn. Lấy mẫu ngẫu nhiên 10 cá thể của loài I và 14 cá thể của Loài II. Xác suất để giá trị trung bình của 10 cá thể của Loài I lớn hơn giá trị trung bình của 14 cá thể của loài II 5 cm là bao nhiêu? Không cần thực hiện bất kỳ phép tính nào, bạn có thể biết rằng xác suất này là khá cao vì sự khác biệt về trung bình của 2 tổng thể là 10 cm. Nhưng xác suất này chính xác là bao nhiêu?

Bài tập 4

Giả sử rằng có khoảng 2% tổng số các cuộc kết nối điện thoại di động của một nhà cung cấp bị ngắt đột ngột. Tìm xác suất để trong một mẫu ngẫu nhiên có 1500 cuộc gọi thì có nhiều nhất là 40 cuộc gọi bị ngắt đột ngột. Trước tiên hãy xác minh rằng cỡ mẫu đủ lớn để sử dụng phân phối chuẩn.