HƯỚNG DẪN THỰC HÀNH XÁC SUẤT THỐNG KÊ VỚI PYTHON

Yêu cầu: tự học 15 giờ, tổng thời gian thực hành mẫu trên lớp 3 tiết, thực hành trên phòng máy 3 tiết.

Download

Python	`python.org/downloads`
Anaconda	`anaconda.com/products/individual`
Tài liệu	`tinyurl.com/dhxd-xstk`

Python cũng có thể chạy trên nền web tại `colab.research.google.com`

Cách sử dụng

- Mỗi khối lệnh có thể gồm một hoặc nhiều lệnh, mỗi lệnh một dòng.
- Thực thi một khối lệnh bằng tổ hợp phím `Shift` + ↵, xem kết quả, soát và sửa lỗi.
- Python phân biệt **CHỮ HOA** và **chữ thường**. Lệnh thường có dạng

  <thư viện>**.**<môđun>**.**<môđun con>**.**<phương thức>**(** <đối số 1> **,** <đối số 2> **)**

  trong đó các đối số của lệnh, nếu có, được đặt trong dấu ( ) và ngăn cách bởi dấu **,**

1 Xác suất

VD1:

Kết quả	*Lệnh*
$C_{10}^4 = 210$	`from sympy import *` `binomial(10, 4)`

$\sum_{k=0}^{10} C_{800}^k 0.005^k \cdot 0.995^{800-k} = 0.9972$	`from sympy import *`
	`k = symbols('k')`
	`Sum( binomial(800, k) * 0.005**k * 0.995**(800-k) ,`
	`     (k, 0, 10) )`
	`_.doit()  # _ là biến lưu kết quả gần đây nhất`

$\sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \lambda$	`from sympy import *`
	`k, l = symbols('k lambda')`
	`Sum( k * l**k * E**-l / factorial(k) , (k, 0, oo) )`
	`_.doit()`

VD2: Đại lượng ngẫu nhiên X có hàm mật độ $f(x) = ae^{4x-x^2}$.

a) Tìm a. b) Tính EX, DX. c) Tính $P(1 < X < 3.5)$.

HD

```
# Khai báo hàm số f(x)
from sympy import *
x, a = symbols('x a')
f = lambda x: a * E**(4*x - x**2)
```

a) $\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow ae^4\sqrt{\pi} = 1 \Rightarrow a = \frac{1}{e^4\sqrt{\pi}} = 0.01033$

```
f(x).integrate((x, -oo, oo))  # tính ∫_{-∞}^{∞} f(x) dx
```
```
_.simplify()                  # rút gọn kết quả trên → √π a e^4
solve(_ - 1, a)               # kết quả trên bằng 1, giải theo biến a
```
```
a = _[0]                      # nghiệm đầu tiên trong dãy các nghiệm → a = 1/(√π e^4)
```
```
N(a, 4)                       # đưa về số thập phân với 4 chữ số có nghĩa
```

b) $EX = \int_{-\infty}^{\infty} xf(x)\,dx = 2,\quad E\left(X^2\right) = \int_{-\infty}^{\infty} x^2 f(x)\,dx = \frac{9}{2},\quad DX = E\left(X^2\right) - (EX)^2 = \frac{1}{2}.$

```
e = (x * f(x)).integrate((x, -oo, oo))
e                 # xem kết quả
e.simplify()      # rút gọn e
e = _             # đặt e là kết quả rút gọn được ở trên

e2 = (x**2 * f(x)).integrate((x, -oo, oo))
e2
e2.simplify()
e2 = _

e2 - e**2
```

c) $P(1 < X < 3.5) = \int_{1}^{3.5} f(x)\,dx = 0.9044$

```
f(x).integrate((x, 1, 3.5))
N(_, 4)   # đưa kết quả trên về số thập phân
```

VD3: Đại lượng ngẫu nhiên X có hàm mật độ $f(x) = \begin{cases} kx, & x \in [0,2] \\ 0, & x \notin [0,2]. \end{cases}$

a) Xác định k.

b) Tìm hàm phân bố $F(x)$.

c) Tính $P(0 < X < 1)$.

HD

```
# Khai báo f(x)
from sympy import *
x, k = symbols('x k')
f = lambda x: Piecewise( (k*x, (x>=0) & (x<=2)) , (0, True) )
```

a) $\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow 2k = 1 \Rightarrow k = \frac{1}{2}.$

```
f(x).integrate((x, -oo, oo))  # 2k
```
```
k = Rational(1, 2)            # 1/2
```

b) $F(x) = \int_{-\infty}^{x} f(t)\,dt = -\frac{\min(0,x)^2}{4} + \frac{\min(2,x)^2}{4} = \begin{cases} 0 & \text{nếu } x \le 0 \\ \frac{x^2}{4} & \text{nếu } 0 < x \le 2 \\ 1 & \text{nếu } x > 2. \end{cases}$

```python
t = symbols('t')
F = f(t).integrate((t, -oo, x))
F

F.subs(Min(0, x), x).subs(Min(2, x), x)   # x ≤ 0
F.subs(Min(0, x), 0).subs(Min(2, x), x)   # 0 < x ≤ 2
F.subs(Min(0, x), 0).subs(Min(2, x), 2)   # x > 2
```

c) $P(0 < X < 1) = \int_0^1 f(x)\,dx = \frac{1}{4}$.

```python
f(x).integrate((x, 0, 1))
```

Chú ý: Trong Python, khi tính tích phân bội $I = \int_D f(x)\,dx$, với $f(x)$ có công thức nhánh, hoặc $D \subset \mathbb{R}^n$ là miền phức tạp hoặc phụ thuộc tham số, để tránh vẽ hình xác định cận lấy tích phân của từng biến, ta thực hiện hai bước:

1) Đặt $g(x) = \begin{cases} f(x), & \text{nếu } x \in D \\ 0 & \text{nếu } x \notin D \end{cases}$

2) Khi đó $\int_{\mathbb{R}^n} g(x)\,dx = \int_D g(x)\,dx + \int_{\overline{D}} g(x)\,dx = \int_D f(x)\,dx + \int_{\overline{D}} 0dx = I + 0 = I$, tức là lấy tích phân của $g(x)$ theo cận của các biến đều từ $-\infty$ đến ∞.

VD4: Véctơ ngẫu nhiên (X, Y) có hàm mật độ xác suất đồng thời

$$f(x,y) = \begin{cases} A\left(2x^2 + xy + y^2\right) & \text{nếu } (x,y) \in [0,1] \times [0,1] \\ 0 & \text{nếu } (x,y) \notin [0,1] \times [0,1]. \end{cases}$$

Tìm

a) A

b) EX

c) $P\left(X < 0.5 \mid Y > 0.5\right)$

HD

a) $\iint_{\mathbb{R}^2} f(x,y)\,dxdy = 1 \Rightarrow \frac{5A}{4} = 1 \Rightarrow A = \frac{4}{5}$

```python
from sympy import *
f(x, y).integrate((x, -oo, oo), (y, -oo, oo))  # 5A/4
A = Rational(4, 5)                             # 4/5
```

b) $EX = \iint_{\mathbb{R}^2} xf(x,y)\,dxdy = \frac{2}{3}$

```python
( x * f(x, y) ).integrate((x, -oo, oo), (y, -oo, oo))
```

c) $P\left(X < 0.5 \mid Y > 0.5\right) = \frac{P(X < 0.5, Y > 0.5)}{P(Y > 0.5)} = \frac{t}{m}$

$m = \iint\limits_{y > 0.5} f(x,y)\,dxdy = 0.65,\quad t = \iint\limits_{x<0.5, y>0.5} f(x,y)\,dxdy = 0.1875$

$\Rightarrow P\left(X < 0.5 \mid Y > 0.5\right) = 0.2885$

```
g = Piecewise( (f(x, y), y>0.5) , (0, True) )
m = g.integrate((x, -oo, oo), (y, -oo, oo))

g = Piecewise( (f(x, y), (x<0.5) & (y>0.5)) , (0, True) )
t = g.integrate((x, -oo, oo), (y, -oo, oo))

t/m
```

2 Thống kê

Giá trị	*Lệnh*

$\Phi(2) = 0.9772$	`from sympy.stats import Normal, P`
	`X = Normal('x', 0, 1)`
	`P(X < 2)`
	`from sympy import *`
	`N(_, 4)  # _ là kết quả của P(X < 2)`

$\Phi(z_0) = 0.95 \Rightarrow z_0 = 1.6449$	`from scipy.stats import norm` `norm.ppf(0.95)`
$t_{0.1}^{29} = 1.6991$	`from scipy.stats import t` `t.isf(0.1 / 2, 29)`
$\chi^2(0.05, 25) = 37.6525$	`from scipy.stats import chi2` `chi2.isf(0.05, 25)`
	`import numpy as np` `X = np.array([1, 2, 3, 4, 5])`
$\overline{x}$	`X.mean()`
s^2	`X.var()`
s	`X.std()`

	`X = [1, 2, 3, 4, 5]`
	`Y = [6, 7, 8, 9, 10]`
	`from scipy.stats import linregress`

a, b, r	`linregress(X, Y)`

VD5: Mẫu cỡ 50 từ $X \sim N\left(a, \sigma^2\right)$ với $\sigma = 2$ cho số liệu theo bảng sau:

x_i	10 – 12	12 – 14	14 – 16	16 – 18
n_i	9	18	17	6

a) Tìm ước lượng không chệch của a.

b) Tìm khoảng tin cậy của a với độ tin cậy 97%.

c) Kiểm định ở mức ý nghĩa 7% xem $EX = 13$ hay $EX > 13$.

d) Kiểm định ở mức ý nghĩa 6% xem có phải $EX = 13$ hay không.

HD

```
import numpy as np
X = np.array( [11]*9 + [13]*18 + [15]*17 + [17]*6 )
```

a) $X \sim N\left(a, \sigma^2\right) \Rightarrow a = EX$ có ước lượng không chệch $\overline{x} = 13.8$.

```python
X.mean()
```

b) $\sigma = 2 \Rightarrow$ khoảng tin cậy của a: $\left(\overline{x} - z_0\frac{\sigma}{\sqrt{n}}, \overline{x} + z_0\frac{\sigma}{\sqrt{n}}\right)$

$\Phi\left(z_0\right) = \frac{1+\gamma}{2} = \frac{1+0.97}{2} = 0.985 \Rightarrow z_0 = 2.1701$

Khoảng tin cậy của a là $(13.1862, 14.4138)$.

```python
from scipy.stats import norm
z0 = norm.ppf( (1 + 0.97) / 2 )
X.mean() - z0 * 2 / np.sqrt(50)
X.mean() + z0 * 2 / np.sqrt(50)
```

c) $H_0 : EX = 13$, $H_1 : EX > 13$, $\alpha = 7\%$.

$z_{qs} = \frac{\overline{x} - a_0}{\sigma}\sqrt{n} = 2.8284$

$\Phi\left(z_0\right) = 1 - \alpha = 0.93 \Rightarrow z_0 = 1.4758$

$z_{qs} > z_0 \Rightarrow$ bác bỏ $EX = 13$ (chấp nhận $EX > 13$).

```python
(X.mean() - 13) / 2 * np.sqrt(50)
norm.ppf(1 - 0.07)   # z0
```

d) $H_0 : EX = 13$, $H_1 : EX \neq 13$, $\alpha = 6\%$.

$z_{qs} = 2.8284$ [như ý (c)]

$\Phi\left(z_0\right) = 1 - \frac{\alpha}{2} = 0.97 \Rightarrow z_0 = 1.8808$

$\left|z_{qs}\right| > z_0 \Rightarrow$ bác bỏ $EX = 13$ (chấp nhận $EX \neq 13$).

```python
norm.ppf(1 - 0.06/2)
```

VD6: Mẫu cỡ $n = 31$ từ $X \sim N\left(a, \sigma^2\right)$ cho số liệu theo bảng sau

x_i	58 – 60	60 – 62	62 – 64	64 – 66
n_i	3	12	13	3

a) Tìm ước lượng không chệch của a và σ^2.

b) Tìm khoảng tin cậy của a với độ tin cậy 92%.

c) Kiểm định ở mức ý nghĩa 4% xem $EX = 64$ hay $EX < 64$.

HD

```python
import numpy as np
X = np.array( [59]*3 + [61]*12 + [63]*13 + [65]*3 )
```

a) $X \sim N\left(a, \sigma^2\right) \Rightarrow EX = a$, $DX = \sigma^2$. Ước lượng không chệch của a là $\overline{x} = 62.0322$, và của σ^2 là $s'^2 = \frac{n}{n-1}s^2 = 2.6323$

```python
X.mean()
31 / 30 * X.var()
```

b) Khoảng tin cậy của a là $\left(\overline{x} - t_0 \dfrac{s}{\sqrt{n-1}}, \overline{x} + t_0 \dfrac{s}{\sqrt{n-1}}\right)$

$s = 1.5960$, $t_0 = t^{n-1}_{1-\gamma} = t^{31-1}_{1-0.92} = 1.8120$, nên khoảng tin cậy của a là $(61.5042, 62.5603)$

```
s = X.std()
t0 = t.isf( (1 - 0.92) / 2 , 31 - 1 )
X.mean() - t0 * s / np.sqrt(30)
X.mean() + t0 * s / np.sqrt(30)
```

c) $t_{qs} = \dfrac{\overline{x} - a_0}{s}\sqrt{n-1} = -6.7528,$

$t_0 = t^{n-1}_{2\alpha} = t^{30}_{0.08} = 1.8120.$

$t_{qs} < -t_0 \Rightarrow$ bác bỏ $EX = 64$ (chấp nhận $EX < 64$).

```
(X.mean() - 64) / s * np.sqrt(30)
t.isf(0.08 / 2, 30)
```

VD7: Phương pháp thứ nhất cho tỷ lệ sản phẩm tốt là 85%. Kiểm tra 300 sản phẩm sản xuất theo phương pháp thứ hai thì thấy có 30 phế phẩm. Hãy kiểm định ở mức ý nghĩa 8% xem có phải phương pháp thứ hai tốt hơn phương pháp thứ nhất không.

HD

p = tỷ lệ sản phẩm tốt sản xuất theo phương pháp thứ hai. Xét bài toán

$$H_0 : p = 0.85, H_1 : p > 0.85, \alpha = 8\%.$$

m = số sản phẩm tốt sản xuất theo phương pháp thứ hai $= 300 - 30 = 270$.

$$z_{qs} = \frac{\frac{m}{n} - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n} = \frac{\frac{270}{300} - 0.85}{\sqrt{0.85 \cdot 0.15}}\sqrt{300} = 2.4254.$$

$\Phi(z_0) = 1 - \alpha = 0.92 \Rightarrow z_0 = 1.4051$

$z_{qs} > z_0 \Rightarrow$ bác bỏ H_0, tức là phương pháp thứ hai tốt hơn.

```
p0 = 0.85
from math import sqrt
(270/300 - p0) / sqrt(p0 * (1-p0)) * sqrt(300)
from scipy.stats import norm
norm.ppf(1 - 0.08)
```

VD8: Với mức ý nghĩa 0.06 hãy kiểm định $H_0 : EX = EY$ với đối thuyết $K_1 : EX > EY$ trong đó X, Y là 2 đại lượng ngẫu nhiên có phân bố chuẩn. Biết rằng 2 mẫu độc lập cỡ $n = 17$ và từ X và $m = 13$ từ Y cho ta số liệu sau:

x_i	21.7	21.9	22.1	22.3	22.5
n_i	1	4	6	5	1

y_i	21.8	22	22.2
m_i	5	7	1

Cho biết $DX = 0.03$, $DY = 0.02$.

HD:

$\overline{x} = 22.1118$, $\overline{y} = 21.9385 \Rightarrow z_{qs} = \dfrac{\overline{x} - \overline{y}}{\sqrt{\dfrac{DX}{n} + \dfrac{DY}{m}}} = 3.0153$

$\Phi(z_0) = 1 - \alpha = 0.94 \Rightarrow z_0 = 1.5548$

$z_{qs} > z_0 \Rightarrow$ bác bỏ H_0.

```
import numpy as np
X = np.array( [21.7]*1 + [21.9]*4 + [22.1]*6 + [22.3]*5 + [22.5]*1 )
Y = np.array( [21.8]*5 + [22]*7 + [22.2]*1 )
(X.mean() - Y.mean()) / np.sqrt(0.03/17 + 0.02/13)

from scipy.stats import norm
norm.ppf(1 - 0.06)
```

VD9: Khảo sát thu nhập (triệu đồng) trong 1 tháng của 10 công nhân ngành A và 15 công nhân ngành B ta thu được số liệu sau:

Ngành A	Ngành B
1, 1.2, 1.3, 1.3, 1.2, 1.3, 1.4, 1.2, 1.3, 1.4	1.1, 1.3, 1.3, 1.3, 1.4, 1.4, 1.4, 1.3, 1.4, 1.2, 1.5, 1.5, 1.5, 1.2, 1.2

Giả sử thu nhập trong 1 tháng của 1 công nhân ngành A và B là những đại lượng ngẫu nhiên X, Y có phân bố chuẩn với phương sai bằng nhau. Hãy kiểm định ở mức ý nghĩa 3% giả thuyết nói rằng thu nhập trung bình của công nhân 2 ngành trên như nhau với đối thuyết cho rằng thu nhập trung bình của công nhân ngành B cao hơn ngành A.

HD:

$H_0 : EX = EY, H_1 : EX < EY, \alpha = 0.03.$

$\overline{x} = 1.26, \overline{y} = 1.3333, s_X^2 = 0.0124, s_Y^2 = 0.01422$

$$t_{qs} = \frac{\overline{x} - \overline{y}}{\sqrt{ns_X^2 + ms_Y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} = -1.4832.$$

$t_0 = t_{2\alpha}^{n+m-2} = t_{0.06}^{23} = 1.9782$

$t_{qs} > -t_0 \Rightarrow$ chấp nhận H_0, hay bác bỏ H_1, tức là thu nhập trung bình của công nhân ngành A không thấp hơn ngành B.

```
import numpy as np
X = np.array( [1, 1.2, 1.3, 1.3, 1.2, 1.3, 1.4, 1.2, 1.3, 1.4] )
Y = np.array( [1.1, 1.3, 1.3, 1.3, 1.4, 1.4, 1.4, 1.3, 1.4, 1.2, 1.5, 1.5, 1.5, 1.2,
    1.2] )

X.mean(), X.var(), Y.mean(), Y.var()
(X.mean() - Y.mean()) / np.sqrt(10*X.var() + 15*Y.var()) * np.sqrt( 10*15*(10+15-2) /
    (10+15) )

from scipy.stats import t
t.isf(0.06 / 2, 23)
```

VD10: Mẫu cỡ $n = 100$ từ ĐLNN X cho ta số liệu sau:

x_i	$1-2$	$2-3$	$3-4$	$4-5$	$5-6$	$6-7$	$7-8$
n_i	4	12	27	30	20	5	2

Hãy kiểm định ở mức ý nghĩa 7% xem có phải X có phân bố chuẩn $N\left(4, 1.3^2\right)$.

HD: $H_0 : X \sim N\left(4, 1.3^2\right), H_1 : X \nsim N\left(4, 1.3^2\right), \alpha = 0.07$

S_i	n_i	p_{i0}	E_i	$\frac{(n_i - E_i)^2}{E_i}$
$(-\infty, 2]$	4	0.06197	6.1968	0.7788
$(2, 3]$	12	0.1589	15.891	0.9527

$(3,4]$	27	0.2791	27.9122	0.02981
$(4,5]$	30	0.2791	27.9122	0.1561
$(5,6]$	20	0.1589	15.891	1.0625
$(6,7]$	5	0.05146	5.146	0.004141
$(7,\infty)$	2	0.01051	1.0508	0.8574
			χ^2_{qs}	3.8415

$\chi_0^2 = \chi^2(0.07, 7-1) = 11.6599$. $\chi^2_{qs} < \chi_0^2 \Rightarrow$ chấp nhận $H_0 : X \sim N\left(4, 1.3^2\right)$.

```python
from sympy import oo
s = [-oo, 2, 3, 4, 5, 6, 7, oo]
n = [4, 12, 27, 30, 20, 5, 2]

from sympy.stats import Normal, P
X = Normal('x', 4, 1.3)
P((X > s[0]) & (X <= s[1]))  # tính thử p_10

from sympy import N
import numpy as np
p0 = np.array([ N(P((X > s[i]) & (X <= s[i+1])), 4) for i in range(7) ])
e = 100 * p0
(n - e)**2 / e
sum(_)

from scipy.stats import chi2
chi2.isf(0.07, 7-1)
```

VD11: Mẫu cỡ $n = 60$ từ ĐLNN X cho ta số liệu dưới đây:

x_i	$10-11$	$11-12$	$12-13$	$13-14$	$14-15$	$15-16$	$16-18$
n_i	9	6	7	8	6	7	17

Kiểm định ở mức ý nghĩa 8% xem X có phân bố đều không?

HD:

Bước 1: $H_0 : X \sim U(a,b)$, $H_1 : X \nsim U(a,b)$, $\alpha = 0.08$.

$\overline{x} = 14.0583$, $s = 2.3648$. Ước lượng theo phương pháp bình phương tối thiểu của a là $a^* = \overline{x} - s\sqrt{3} = 9.9623$, của b là $b^* = \overline{x} + s\sqrt{3} = 18.1543$ (số tham số chưa biết $r = 2$).

```python
import numpy as np
X = np.array( [10.5]*9 + [11.5]*6 + [12.5]*7 + [13.5]*8 + [14.5]*6 + [15.5]*7 +
    [17]*17 )
X.mean(), X.std()
a = X.mean() - X.std() * np.sqrt(3)  # → a*
b = X.mean() + X.std() * np.sqrt(3)  # → b*
```

Bước 2: $H_0^* : X \sim U\left(a^*, b^*\right)$, $H_1^* : X \notin U\left(a^*, b^*\right)$, $\alpha = 0.08$.

S_i	n_i	p_{i0}	E_i	$\dfrac{(n_i-E_i)^2}{E_i}$
$(-\infty, 11]$	9	0.1267	7.6001	0.2578

(11, 12]	6	0.1221	7.3242	0.2394
(12, 13]	7	0.1221	7.3242	0.01435
(13, 14]	8	0.1221	7.3242	0.06236
(14, 15]	6	0.1221	7.3242	0.2394
(15, 16]	7	0.1221	7.3242	0.01435
$(16, \infty)$	17	0.263	15.7788	0.09451
			χ^2_{qs}	0.9222

$\chi^2_0 = \chi^2(\alpha, h - r - 1) = \chi^2(0.08, 7 - 2 - 1) = 8.3365$.

$\chi^2_{qs} < \chi^2_0 \Rightarrow$ chấp nhận H_0^*, nên chấp nhận H_0, tức là X có phân bố đều.

```python
from sympy import oo
s = [-oo, 11, 12, 13, 14, 15, 16, oo]
n = [9, 6, 7, 8, 6, 7, 17]

from sympy.stats import Uniform, P
X = Uniform('x', a, b)
P((X > s[0]) & (X <= s[1]))

import numpy as np
p0 = np.array([ P((X > s[i]) & (X <= s[i+1])) for i in range(7) ])

e = 60 * p0
(n - e)**2 / e
sum(_)

from scipy.stats import chi2
chi2.isf(0.08, 7-2-1)
```

VD12: Khảo sát nhu cầu X (C = có, K = không) đối với một sản phẩm theo vùng miền Y (T = thành thị, N = nông thôn, M = miền núi), ta có mẫu cỡ 200 như sau:

$X \backslash Y$	T	N	M
C	26	48	24
K	51	43	8

Kiểm định ở mức ý nghĩa 5% xem X và Y độc lập nhau không?

HD:

$X \backslash Y$	T	N	M	$\sum$
C	26 37.73	48 44.59	24 15.68	98
K	51 39.27	43 46.41	8 16.32	102
$\sum$	77	91	32	

$$\chi^2_{qs} = \frac{(26-37.73)^2}{37.73} + \frac{(48-44.59)^2}{44.59} + \cdots + \frac{(8-16.32)^2}{16.32} = 16.3181.$$

$\chi^2_0 = \chi^2\left[\alpha, (h-1)(k-1)\right] = \chi^2\left[0.05, (2-1)(3-1)\right] = 5.9915.$

$\chi^2_{qs} > \chi^2_0 \Rightarrow$ nhu cầu về sản phẩm đó phụ thuộc vào vùng miền.

```
import numpy as np
n = np.array( [[26, 48, 24], [51, 43, 8]] )

nx = n.sum(axis=1)  # thành phần của Y chạy, X cố định
ny = n.sum(axis=0)

e = [ [i * j / 200 for j in ny] for i in nx ]
(n - e)**2 / e
_.sum()

from scipy.stats import chi2
chi2.isf(0.05, (2-1)*(3-1))
```

VD13: Mẫu cỡ $n = 12$ từ véctơ ngẫu nhiên (X, Y) cho ta số liệu sau:

x_i	3	3.5	2.5	3.5	3	3.1	2	4	4.5	3	3.5	3.1
y_i	13	14	10	14.5	14	13.5	9.5	16	18	12.5	15	14.5

a) Tìm hệ số tương quan mẫu giữa X và Y. Có thể dùng hồi quy bình phương trung bình tuyến tính của Y đối với X để dự báo giá trị của Y được không, vì sao?

b) Tìm hàm hồi quy bình phương trung bình tuyến tính thực nghiệm của Y đối với X và ước lượng sai số bình phương trung bình.

c) Dự báo giá trị của Y khi biết $X = 2.3$.

HD

a) $r = 0.957$. $|r|$ khá lớn (≥ 0.8) $\Rightarrow$ có thể dùng hồi quy bình phương trung bình tuyến tính của Y theo X để dự báo giá trị của Y.

b) Hàm hồi quy tuyến tính thực nghiệm của Y theo X là $\widetilde{\varphi}(X) = aX + b$, trong đó $a = 3.4397$, $b = 2.6154$.

Sai số bình phương trung bình $s^2_{Y/X} = s^2_Y\left(1 - r^2\right) = 0.4221$.

c) Khi $X = 2.3$ ta dự báo $Y = a \cdot 2.3 + b = 10.5266$.

```
import numpy as np
X = np.array( [3, 3.5, 2.5, 3.5, 3, 3.1, 2, 4, 4.5, 3, 3.5, 3.1] )
Y = np.array( [13, 14, 10, 14.5, 14, 13.5, 9.5, 16, 18, 12.5, 15, 14.5] )

from scipy.stats import linregress
a, b, r, _, _ = linregress(X, Y)

Y.var() * (1 - r**2)  # sai số
a * 2.3 + b
```

Thống kê thời gian thực hành:

VD	1	2	3	4	5	6	7	8	9	10	11	12	13	$\sum$
Time	1'2"	1'21"	1'1"	1'37"	3'9"	2'33"	47"	2'5"	2'24"	3'9"	4'26"	1'55"	2'45"	31'56"