TRƯỜNG ĐẠI HỌC
VĂN LANG
Đạo đức - Ý chí - Sáng tạo
KHOA CÔNG NGHỆ THÔNG TIN

XÁC SUẤT THỐNG KÊ ỨNG DỤNG

XÁC SUẤT CÓ ĐIỀU KIÊN – BIẾN NGẪU NHIÊN

TS. TRẦN NGỌC VIỆT
PGS. TS. NGUYỄN VĂN LỘC
TH.S LÊ CÔNG HIẾU

HỌC KỲ 2 – NĂM HỌC 2021-2022

KHÓA K26

1.Sự kiện phụ thuộc và sự kiện độc lập

Hai sự kiện (biến cố) E và F được gọi là phụ thuộc, nếu biết thông tin về sự kiện E (hay F) xảy ra thì chúng ta có thể suy ra thông tin khi sự kiện F (hay E) xảy ra.

Ngược lại, nếu biết thông tin về E (hay F) mà ta không biết thông tin gì về F (hay E) thì hai sự kiện E và F gọi là độc lập.

Ví dụ 1.1: tung đồng xu hai lần

-nếu gọi E là sự kiện tung đồng xu thứ nhất là mặt ngửa, F là sự kiện tung đồng xu thứ hai có mặt ngửa, thì E và F là hai sự kiện độc lập vì E hay F không ảnh hưởng đến kết quả xảy ra của nhau.

-nếu gọi G là sự kiện cả hai lần tung đồng xu là mặt sấp thì E và G là hai sự kiện phụ thuộc vì từ E ta có thể suy ra sự kiện G (trong trường hợp này chúng ta có thể suy ra là G sai).

Gọi:

P(E) là xác suất xảy ra sự kiện E;

P(F) là xác suất xảy ra sự kiện F;

P(E,F) là xác suất xảy ra cả hai sự kiện E và F, nếu E và F độc lập thì P(E,F) là tích của P(E) và P(F):

$$P(E,F) = P(E).P(F)$$

2.Xác suất có điều kiện

Xác suất có điều kiện là xác suất một biến cố E xảy ra, biết rằng biến cố F xảy ra.

ký hiệu: P(E|F).

$$P(E|F) = P(E,F)/P(F)$$

Suy ra:

$$P(E,F) = P(E|F).P(F)$$

Nếu E và F độc lập, ta có:

$$P(E|F) = P(E)$$

Ví dụ 2.1: một gia đình có hai người con biết rằng

-Giới tính mỗi người con có thể là nam hay nữ

-Giới tính người con thứ hai độc lập giới tính người con thứ nhất

Như vậy sẽ có 4 trường hợp giới tính của các người con như sau

Giới tính con thứ nhất	Giới tính con thứ hai
Nam	Nam
Nữ	Nữ
Nam	Nữ
Nữ	Nam

Gọi A là sự kiện giới tính cả hai người con không là Nữ

B là sự kiện giới tính một Nam và một Nữ (không quan tâm người con thứ nhất hay thứ hai)

C là sự kiện giới tính cả hai người con là Nữ

D là sự kiện giới tính người con thứ hai là Nữ

E là sự kiện ít nhất một người con là Nữ

P(A) = 3/4

P(B) = 2/4 = 1/2

P(C) = 1/4

P(D) = 2/4 = 1/2

P(E) = 3/4

P(C|D) = P(C,D)/P(D) = P(C)/P(D) = (1/4) / (1/2) = 1/2;

P(C,D) = P(D) , vì hai sự kiện C, D chỉ là một sự kiện C.

P(C|E) = P(C,E)/P(E) = P(C)/P(E) = (1/4) / (3/4) = 1/3

Đoạn mã Python minh họa tính xác suất các sự kiện A, B, C, D, E:

```
# THỰC HÀNH 01
# xác suất có điều kiện
>>> import math
>>> import random
>>> def chon_gioi_tinh():
        return random.choice(["Nam", "Nu"])
>>> A=0
>>> B=0
>>> C=0
>>> D=0
>>> E=0
>>> random.seed(0)
>>> for _ in range(10000): # chọn ngẫu nhiên trong 10000 lần
        thu_nhat = chon_gioi_tinh()
        thu_hai = chon_gioi_tinh()
        if not(thu_nhat == "Nu" and thu_hai == "Nu"):
                A += 1
        if (thu_nhat == "Nam" and thu_hai == "Nu") or
(thu_nhat == "Nu" and thu_hai == "Nam"):
                B += 1
        if thu_nhat == "Nu" and thu_hai == "Nu":
                C += 1
        if thu_hai == "Nu":
                D += 1
        if thu_nhat == "Nu" or thu_hai == "Nu":
                E += 1

>>> print("P(A) = ", A/10000)
P(A) =  0.7528                  # 3/4
>>> print("P(B) = ", B/10000)
P(B) =  0.4992                  # 1/2
>>> print("P(C) = ", C/10000)
P(C) =  0.2472                  # 1/4
>>> print("P(D) = ", D/10000)
P(D) =  0.4937                  # 1/2
>>> print("P(E) = ", E/10000)
P(E) =  0.7464                  # 3/4
>>> print("P(C|D) = ", C/D)
P(C|D) =  0.5007089325501317    # 1/2
>>> print("P(C|E) = ", C/E)
P(C|E) =  0.3311897106109325    # 1/3
```

#nhận xét: kết quả từng xác suất P(A), .., P(C/E) ?

3.Biến ngẫu nhiên

Trong một phép thử ngẫu nhiên, đầu ra của nó có thể là giá trị số hoặc không phải. Ví dụ phép thử ngẫu nhiên là tung một đồng xu lên và xét mặt nào của đồng xu ở phía trên, thì kết quả đầu ra có thể là {sấp, ngửa}.

Chẳng hạn như phép thử ngẫu nhiên là tung con súc sắc và xem mặt nằm phía trên là có mấy chấm, thì kết quả đầu ra có thể là {1,2,3,4,5,6}.

Tuy nhiên, trong các ứng dụng của thống kê, người ta muốn mỗi đầu ra đều gắn với một đại lượng đo đạc được, hay còn gọi là thuộc tính có giá trị là số. Để thực hiện điều này, người ta định ra biến ngẫu nhiên để ánh xạ mỗi đầu ra của một phép thử ngẫu nhiên với một giá trị số.

Cho không gian xác suất (Ω, A, P).
Một hàm X: Ω → R là một biến ngẫu nhiên giá trị thực nếu với mọi tập con
A_r = { ω: X(ω) ≤ r }, trong đó r ∈ R, ta cũng có A_r ∈ A.
Định nghĩa trên giúp ta xây dựng hàm phân bố của biến ngẫu nhiên.
Biến ngẫu nhiên có 2 dạng:

-Biến ngẫu nhiên rời rạc: tập giá trị nó là rời rạc, tức là đếm được. Ví dụ như mặt chấm xuất hiện của con xúc xắc.

-Biến ngẫu nhiên liên tục: tập giá trị là liên tục, tức là lấp đầy trên một khoảng trục số.

4.Phân phối xác suất

Phân phối xác suất hay thường gọi hơn là một hàm phân phối xác suất là quy luật cho biết cách gán mỗi xác suất cho mỗi khoảng giá trị của tập số thực, sao cho các tiên đề xác suất được thỏa mãn.

Có hai dạng phân phối xác suất là phân phối rời rạc và phân phối liên tục. Sự kiện tung đồng xu tương ứng với phân phối rời rạc vì xác suất của sự kiện này gắn liền với các giá trị rời rạc, ví dụ mặt sấp hay ngửa.

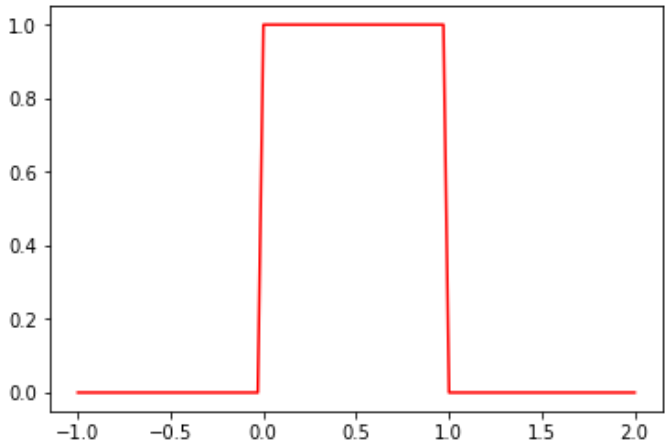
Tuy nhiên, trong thực tế chúng ta thường chỉ sử dụng phân phối liên tục trong khoảng (a, b) với a và b là các số thực.

Dạng đơn giản nhất của phân phối liên tục là phân phối đều liên tục trong khoảng giá trị là 0 và 1 với hàm mật độ xác suất có thể được định nghĩa bằng Python như sau:

THỰC HÀNH 02

```
1  import numpy as np
2  import matplotlib.pyplot as plt
3  x = np.linspace(-1, 2, 100)
4  def uniform_pdf(x):
5    return 1 if x >= 0 and x < 1 else 0
6  y = []
7  for i in range(100):
8    y.append(uniform_pdf(x[i]))
9  plt.plot(x, y, 'r')
10 plt.show()
```

Biểu đồ thể hiện cho phân phối liên tục đều bằng hàm pdf sau:



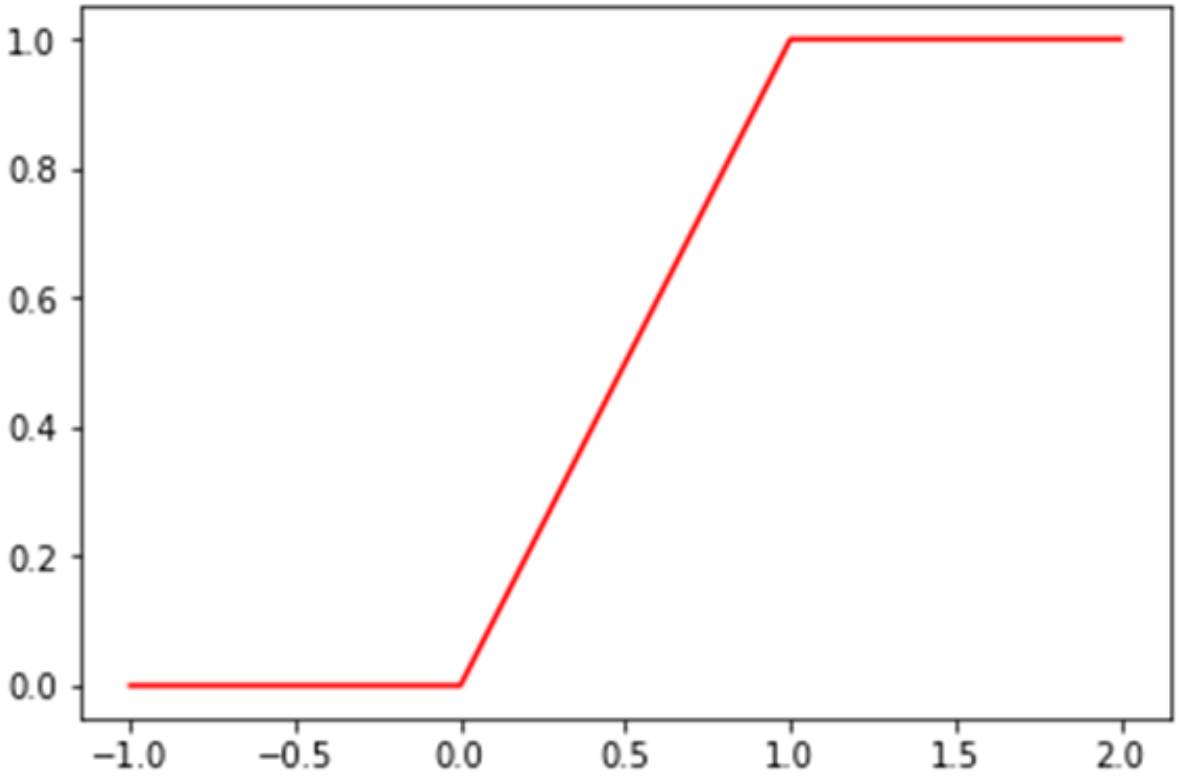
#nhận xét: đồ thị phân phối liên tục đều ?

Một dạng hàm phổ biến khác thường dùng cho phân phối liên tục là hàm phân phối tích lũy (cumulative distribution function – cdf). Hàm phân phối tích lũy cho phân phối liên tục đều có thể định nghĩa trong Python như sau:

 # THỰC HÀNH 03

```
1   import numpy as np
2   import matplotlib.pyplot as plt
3   x = np.linspace(-1, 2, 100)
4   def uniform_cdf(x):
5      if x < 0:
6        return 0
7      elif x < 1:
8        return x
9      else:
10       return 1
11  y = []
12  for i in range(100):
13    y.append(uniform_cdf(x[i]))
14  plt.plot(x, y, 'r')
15  plt.show()
```

Biểu đồ biểu diễn phân phối liên tục đều qua hàm cdf như sau



#nhận xét: đồ thị phân phối liên tục đều ?

5.Phân phối chuẩn

Định nghĩa: Biến ngẫu nhiên liên tục X nhận giá trị trên ℝ được gọi là có phân phối chuẩn với tham số μ, σ^2, ký hiệu là $X \sim N(\mu, \sigma^2)$, nếu hàm mật độ của nó có dạng:

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

Hay toán học có dạng: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $x \in R$

-Phân phối chuẩn là một dạng của phân phối liên tục.

-Phân phối chuẩn còn gọi là phân phối Gauss hay phân phối đường cong chuông. Biểu đồ của nó có dạng hình chuông.

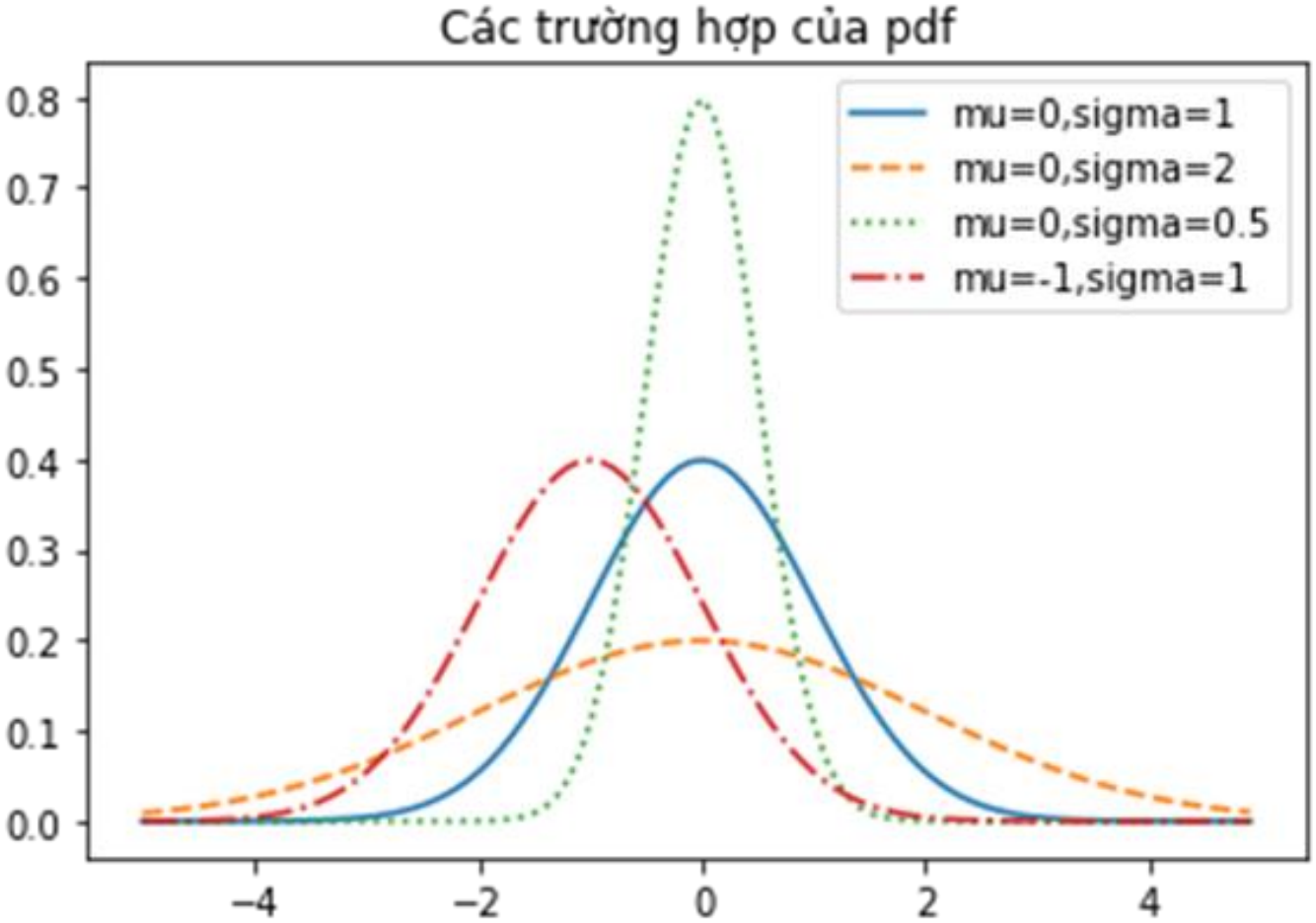
-Phân phối chuẩn phụ thuộc vào hai tham số là giá trị trung bình (mean), kí hiệu là μ (mu), và độ lệch chuẩn (standard deviation), kí hiệu σ (sigma). Giá trị mean là đỉnh của hình chuông, độ lệch chuẩn là độ rộng của chuông.

THỰC HÀNH 04

```
>>> def normal_pdf(x, mu=0, sigma=1):
        sqrt_two_pi = math.sqrt(2 * math.pi)
        return (math.exp(-(x-mu) ** 2 / 2 / sigma ** 2) / (sqrt_two_pi * sigma))

>>> xs = [x / 10.0 for x in range(-50, 50)]
>>> import numpy as np
>>> import matplotlib.pyplot as plt
>>> plt.plot(xs,[normal_pdf(x,sigma=1) for x in xs],'-',label='mu=0,sigma=1')
[<matplotlib.lines.Line2D object at 0x0B52F6F0>]
>>> plt.plot(xs,[normal_pdf(x,sigma=2) for x in xs],'--',label='mu=0,sigma=2')
[<matplotlib.lines.Line2D object at 0x0B52F910>]
>>> plt.plot(xs,[normal_pdf(x,sigma=0.5) for x in xs],':',label='mu=0,sigma=0.5')
[<matplotlib.lines.Line2D object at 0x0B52F7B0>]
>>> plt.plot(xs,[normal_pdf(x,mu=-1) for x in xs],'-.',label='mu=-1,sigma=1')
[<matplotlib.lines.Line2D object at 0x0B52FB90>]
>>> plt.legend()
<matplotlib.legend.Legend object at 0x0986FDF0>
>>> plt.title("Các trường hợp của pdf")
Text(0.5, 1.0, 'Các trường hợp của pdf')
>>> plt.show()
```

#nhận xét: hàm f(x) và đồ thị hình chuông ?

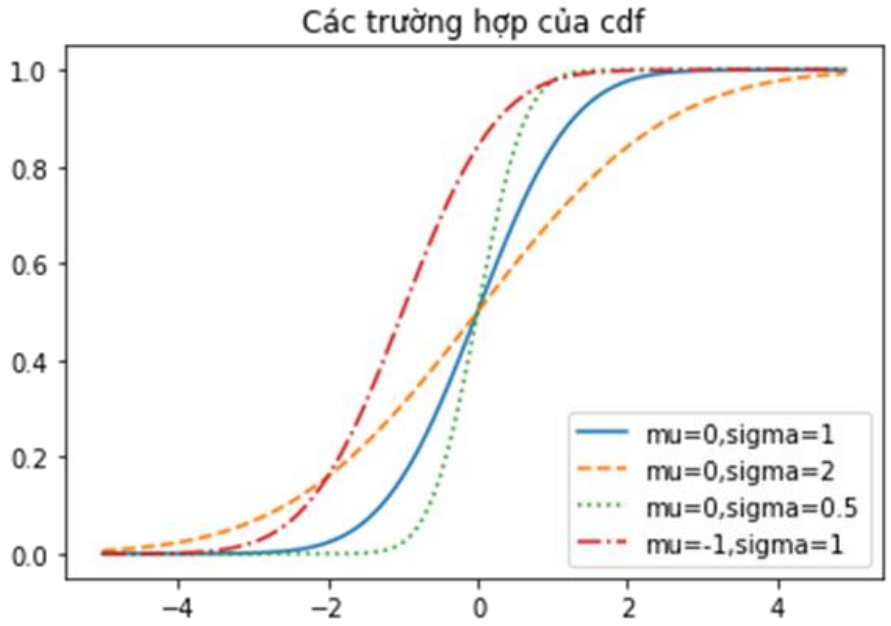

Các trường hợp của pdf
mu=0,sigma=1
mu=0,sigma=2
mu=0,sigma=0.5
mu=-1,sigma=1
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−4
−2
0
2
4

Nếu μ = 0 và σ = 1, chúng ta gọi là phân phối chuẩn tắc (normal standard distribution).

 # THỰC HÀNH 05

```
def normal_cdf(x, mu=0,sigma=1):
    return (1 + math.erf((x - mu) / math.sqrt(2) / sigma)) / 2

xs = [x / 10.0 for x in range(-50, 50)]
plt.plot(xs,[normal_cdf(x,sigma=1) for x in xs],'-',label='mu=0,sigma=1')
plt.plot(xs,[normal_cdf(x,sigma=2) for x in xs],'--',label='mu=0,sigma=2')
plt.plot(xs,[normal_cdf(x,sigma=0.5) for x in xs],':',label='mu=0,sigma=0.5')
plt.plot(xs,[normal_cdf(x,mu=-1) for x in xs],'-.',label='mu=-1,sigma=1')
plt.legend(loc=4)
plt.title("Các trường hợp của cdf")
plt.show()
```



#nhận xét: hàm f(x) và đường cong?

6.Định lý Bayes

$$P(A|B) = P(B|A).P(A) / P(B)$$

- P(A): Xác suất của sự kiện A xảy ra
- P(B): Xác suất của sự kiện B xảy ra
- P(B|A): Xác suất (có điều kiện) của sự kiện B xảy ra, nếu biết rằng sự kiện A đã xảy ra
- P(A|B): Xác suất (có điều kiện) của sự kiện A xảy ra, nếu biết rằng sự kiện B đã xảy ra

Các phương pháp suy diễn dựa trên xác suất sẽ sử dụng xác suất có điều kiện.

Giả sử chúng ta dự đoán và phân loại mail Spam hay Not Spam ? Dựa vào tập dữ liệu sau đây:

Ta có training data gồm 10 email đánh dấu 2 nhãn gồm Spam (S), Not Spam (N).

Bảng từ vựng như sau:

$W = [w1, w2, w3, w4, w5]$

Cần phân loại email E11 thuộc loại nào

	Email	w1	w2	w3	w4	w5	Label
Training data	E1	1	1	0	1	0	N
	E2	0	1	1	0	0	N
	E3	1	0	1	0	1	S
	E4	1	1	1	1	0	S
	E5	0	1	0	1	0	S
	E6	0	0	0	1	1	N
	E7	0	1	0	0	0	S
	E8	1	1	0	1	0	S
	E9	0	0	1	1	1	N
	E10	1	0	1	0	1	S
Test data	E11	1	0	0	1	1	?

THỰC HÀNH 06

```
>>> import math
>>> from sklearn.naive_bayes import
BernoulliNB
    #thư viện hàm Bayes-Bernoulli Naive Bayes
>>> import numpy as np
>>> e1 = [1, 1, 0, 1, 0]
    #training data;
>>> e2 = [0, 1, 1, 0, 0]
>>> e3 = [1, 0, 1, 0, 1]
>>> e4 = [1, 1, 1, 1, 0]
>>> e5 = [0, 1, 0, 1, 0]
>>> e6 = [0, 0, 0, 1, 1]
>>> e7 = [0, 1, 0, 0, 0]
>>> e8 = [1, 1, 0, 1, 0]
>>> e9 = [0, 0, 1, 1, 1]
>>> e10 = [1, 0, 1, 0, 1]
>>> train_data = np.array([e1, e2, e3, e4,
e5, e6, e7, e8, e9, e10])
>>> label = np.array(['N', 'N', 'S', 'S', 'S',
'N', 'S', 'S', 'N', 'S'])
>>> e11 = np.array([[1, 0, 0, 1, 1]])
    #test data;
>>> clf1 = BernoulliNB(alpha=1e-10)
>>> clf1.fit(train_data, label)
BernoulliNB(alpha=1e-10, binarize=0.0,
class_prior=None, fit_prior=True)
    #Huan luyen (phân tích, xử lý ..)
>>> print('Probability of e11 in each class:',
clf1.predict_proba(e11))
Probability of e11 in each class: [[0.45762712
0.54237288]]
    #xác suất
>>> print('Predicting class of e11:',
str(clf1.predict(e11)[0]))
Predicting class of e11: S
```

#nhận xét: kết quả đầu ra-output ; binary or classifier?

Tương tự như bài toán trên tiếp tục phân loại mail Spam (S) và Not Spam (N).

Bộ training data gồm E1, E2, E3. Cần phân loại E4.

Bảng từ vựng như sau:

W = [*w*1,*w*2,*w*3,*w*4,*w*5,*w*6,*w*7].

Số lần xuất hiện của từng từ trong từng email tương ứng như bảng dưới.

	Email	w1	w2	w3	w4	w5	w6	w7	Label
Training data	E1	1	2	1	0	1	0	0	N
	E2	0	2	0	0	1	1	1	N
	E3	1	0	1	1	0	2	0	S
Test data	E4	1	0	0	0	0	0	1	?

trong đó:

P(S) =1/3, P(N)=2/3.

THỰC HÀNH 07

```
    #pip install scikit-learn...
    #KeyboardInterrupt
    #naive_bayes_multinomial.py...
>>> from sklearn.naive_bayes import MultinomialNB
    #thư viện hàm Bayes-Multinomial Naive Bayes
>>> import numpy as np
    #train data-dữ liệu huấn luyện
>>> e1 = [1, 2, 1, 0, 1, 0, 0]
>>> e2 = [0, 2, 0, 0, 1, 1, 1]
>>> e3 = [1, 0, 1, 1, 0, 2, 0]
>>> train_data = np.array([e1, e2, e3])
>>> label = np.array(['N', 'N', 'S'])

>>> e4 = np.array([[1, 0, 0, 0, 0, 0, 1]]) #test data-dữ liệu kiểm tra

>>> clf1 = MultinomialNB(alpha=1)
>>> clf1.fit(train_data, label)
MultinomialNB(alpha=1, class_prior=None, fit_prior=True)
    #Huan luyen (phân tích, xử lý ..)
>>> print('Probability of e4 in each class:', clf1.predict_proba(e4))
Probability of e4 in each class: [[0.66589595 0.33410405]]
>>> print('Predicting class of e4:', str(clf1.predict(e4)[0]))
Predicting class of e4: N
```

#nhận xét: kết quả đầu ra-output ; binary or classifier?

7.Thống kê – Mẫu và các đặc trưng mẫu

+$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ là trung bình mẫu;

+$\bar{X} = \frac{1}{n}(X_1 + X_2 + .. + X_n)$, được gọi là trung bình mẫu;

+$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, gọi là phương sai mẫu đã hiệu chỉnh;

+ Nếu biến ngẫu nhiên X có phân phối chuẩn với **kỳ vọng** μ và **phương sai** σ^2 thì biến $Z = \frac{X-\mu}{\sigma}$ (theo phân phối chuẩn tắc);

+ Đặc trưng X ở tổng thể có E(X) = μ, D(X) = σ² ,

(E(X) = μ, Var =V(X)= D(X) = σ²) ;

STD (standard deviation-độ lệch chuẩn), được ký hiệu là: σ (đọc là -sigma) ;

Var (Variance), được ký hiệu là: σ² (đọc là -sigma bình phương) ;

THỰC HÀNH 08

```
>>> import math
>>> data = [0.6, 0.24, 0.36, 1.20, 1.8, 4.40, 2.2, 3.2, 60, 80]
>>> def mean(data):   #median —trung vị(giá trị ở giữa -middle value)
     return sum(data)/len(data)
>>> mu = mean(data)
>>> print(mu)

>>> import numpy  #sử dụng thư viện numpy
>>> numpy.mean(data)

>>> data = [0.6, 0.24, 0.36, 1.20, 1.8, 4.40, 2.2, 3.2, 60, 80]
>>> import statistics   #sử dụng thư viện thống kê
>>> print(statistics.mean(data))

    #Mode là giá trị xuất hiện nhiều lần nhất trong tập dữ liệu data
>>> def mode(data):
    dmax = data[5]
    for d in data:
        if data.count(d) > data.count(dmax):
            dmax = d
    return dmax
>>> mode(data)
```

THỰC HÀNH 09

```
    #Mode là giá trị xuất hiện nhiều lần nhất trong tập dữ liệu data

>>> data = [1.6, 52.8, 18.36, 5.2, 1.8, 4.9, 2.5, 3.2, 60, 90]
>>> def mode(data):
        dmax = data[7]
        for d in data:
                if data.count(d) > data.count(dmax):
                        dmax = d
        return dmax

>>> mode(data)


>>> import statistics   #sử dụng thư viện thống kê
>>> print(statistics.mean(data))
```

8.Mô tả tập dữ liệu

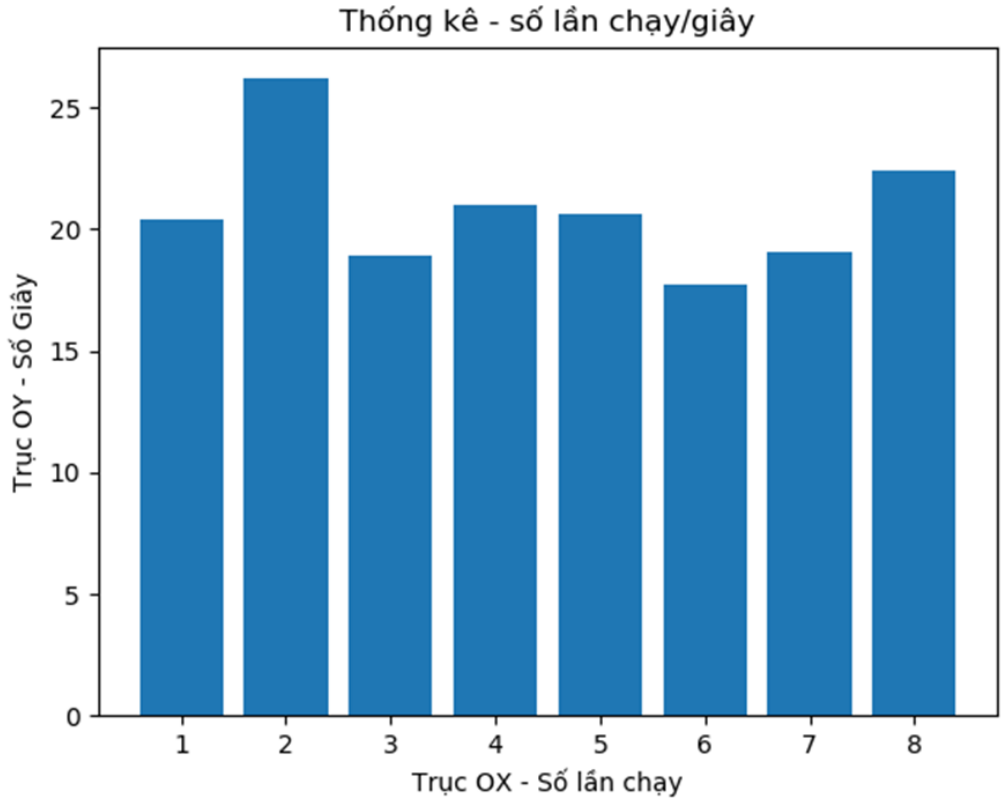
Giả sử rằng bạn chạy 100 m trong n lần, mỗi lần chạy bạn dùng đồng hồ đo lại thời gian chạy (đơn vị đo giây) và kết quả 8 lần chạy của bạn gồm tám giá trị (được gọi là quan sát).

Phương pháp được dùng trong thống kê bằng cách lập bảng thu thập dữ liệu sau

Số Lần Chạy	1	2	3	4	5	6	7	8
Giây	20.4	26.2	18.9	21.0	20.6	17.7	19.1	22.4

THỰC HÀNH 10

```
>>> import math
>>> from matplotlib import pyplot as plt
>>> SoGiay = [20.4, 26.2, 18.9, 21.0, 20.6, 17.7,
19.1,22.4]
>>> SoLanChay = [1,2,3,4,5,6,7,8]
>>> xs = [i + 0.3 for i, _ in enumerate(SoLanChay)]
>>> plt.bar(xs, SoGiay)
>>> plt.xticks([i + 0.3 for i, _ in
enumerate(SoLanChay)], SoLanChay)
>>> plt.show()#Ghi thêm title chính, title OX, title OY;
```



#Nhận xét: Qua biểu đồ hay số liệu thống kê ta thấy dễ dàng suy ra số lần chạy nào có thời gian số giây lớn nhất hay số giây nhỏ nhất. Phân tích dữ liệu bằng hình ảnh trực quan hơn từ tập dữ liệu.

9.Ước lượng khoảng

Ước lượng kì vọng của phân phối chuẩn: $X \sim N(\mu, \sigma^2)$ với mẫu ngẫu nhiên $(X_1, X_2, .., X_n)$ và với độ tin cậy cho ta tìm ước lượng khoảng của kì vọng μ.

Xét thống kê $Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n} \Rightarrow Z \sim N(0,1)$

Khoảng tin cậy của kì vọng μ chưa biết khi σ^2 đã biết là:

$(\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}};\ \bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}})$

Ví dụ: Giả sử biến X ở tổng thể có phân phối chuẩn với kì vọng chưa biết μ khi $\sigma^2 = 4$. Từ một mẫu có kích thước n = 16 ta tìm được $\bar{X} = 12$.

Với độ tin cậy P=0,95, hãy tìm ước lượng khoảng của μ.

Giải:

Ta có: 1-P = 0,05 = $\propto$; $U_{\frac{\propto}{2}} = 1,96$ (tra bảng);

$$\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 - 1,96.0,5 = 11,02;$$

$$\bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 + 1,96.0,5 = 12,98;$$

Vậy khoảng tin cậy của μ với độ tin cậy P=0,95 là

$(\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}};\ \bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}}) = (11.02,\ 12.98)$.

THỰC HÀNH 11

```
from scipy import stats
import numpy as np
from numpy import random
a = random.normal(0,1,10000)
mean, sigma = np.mean(a), np.std(a)
      # conf_int_a (confidence interval a- khoảng tin cậy của a);
conf_int_a = stats.norm.interval(0.68, loc=mean, scale=sigma)
conf_int_b = stats.norm.interval(0.68, loc=mean, scale=sigma /
np.sqrt(len(a)))
conf_int_a
(-1.0011149125527312, 1.0059797764202412)
conf_int_b
(-0.0076030415111100983, 0.01246790537861962 5)
```

#giải thích câu lệnh
#nhận xét ước lượng khoảng

THỰC HÀNH 12

```
>>> from scipy import stats
>>> import numpy as np
>>> from numpy import random
>>> N = 10000
#N tổng thể mẫu-số lượng
>>> a = np.random.normal(0, 1, N)
>>> mean, sigma = a.mean(), a.std(ddof=1)
#trung vị, độ lệch chuẩn
>>> conf_int_a = stats.norm.interval(0.68, loc=mean,
scale=sigma)
>>> conf_int_a
#khoảng tin cậy của a
(-0.979088025332332, 0.987566152992649)
```

#giải thích câu lệnh
#nhận xét ước lượng khoảng

THỰC HÀNH 13

Ví dụ: Giả sử biến X ở tổng thể có phân phối chuẩn với kì vọng chưa biết μ khi $\sigma^2 = 4$. Từ một mẫu có kích thước n = 16 ta tìm được $\bar{X} = 12$.

Với độ tin cậy P=0,95, hãy tìm ước lượng khoảng của μ.

Giải:

Ta có: 1-P = 0,05 = $\propto$; $U_{\frac{\propto}{2}} = 1{,}96$ (tra bảng);

$$\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 - 1{,}96.0{,}5 = 11{,}02;$$

$$\bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 + 1{,}96.0{,}5 = 12{,}98;$$

Vậy khoảng tin cậy của μ với độ tin cậy P=0,95 là

$(\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}}) = (11.02,\ 12.98)$.

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

THỰC HÀNH 14

Ví dụ: Khối lượng sản phẩm là đại lượng ngẫu nhiên X có luật phân phối chuẩn, biết rằng phương sai σ^2 = 4g. Kiểm tra n= 25 sản phẩm, tính được trung bình mẫu $\bar{x}$ = 20g. Hãy tìm ước lượng khoảng của μ với độ tin cậy P= 95%.

Giải

Đặt μ = E(X) chưa biết.

Chọn thống kê $Z = \dfrac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \in N(0,1)$ để ước lượng trung bình μ, trong đó:

$\sigma = 2g$, n = 25 , $\bar{x} = 20g$

1-P = 0,05 = $\propto$; $U_{\frac{\propto}{2}}$ = 1,96 (tra bảng);

$$\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 20 - 1{,}96.\frac{2}{\sqrt{25}} =$$

$$\bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 20 + 1{,}96.\frac{2}{\sqrt{25}} =$$

Vậy, khoảng ước lượng trung bình khối lượng sản phẩm với độ tin cậy 95% là: (19,216; 20,784) (gam)

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

BÀI KIỂM TRA GIỮA KÌ SỐ 1 – THỰC HÀNH XSTK UD

Bài toán: Khối lượng sản phẩm là đại lượng ngẫu nhiên X có luật phân phối chuẩn, biết rằng phương sai σ^2 = 9g. Kiểm tra 16 sản phẩm, tính được trung bình mẫu $\bar{x}$ = 30g.

Hãy tính khoảng ước lượng trung bình của khối lượng sản phẩm với độ tin cậy 95%.

Giải

Đặt μ = E(X) chưa biết.

Chọn thống kê $Z = \frac{(\overline{X} - \mu)\sqrt{n}}{\sigma} \in N(0,1)$ để ước lượng trung bình μ,

trong đó: σ=9g, n = 16 , $\bar{x} = 30\ g$ $\boldsymbol{\sigma^2 = 9 \rightarrow \sigma = 3}$

Độ tin cậy 1 – α = 95% = 0,95 → $\frac{1-\alpha}{2} = \frac{0,95}{2} = 0,475 \rightarrow U_{\frac{\alpha}{2}} = 1,96$ (tra bảng)

Do đó: $\varepsilon = U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1,96.\frac{3}{\sqrt{16}} =$

Suy ra: $\mu_1 = \bar{x} - \varepsilon$; $\mu_2 = \bar{x} + \varepsilon$;

Vậy: khoảng ước lượng trung bình khối lượng sản phẩm với độ tin cậy 95% là (..; ...) (gam).

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

BÀI KIỂM TRA GIỮA KÌ SỐ 1 – THỰC HÀNH XSTK UD

Thực hành: Giả sử biến X ở tổng thể có phân phối chuẩn với kì vọng chưa biết μ khi $\sigma^2 = 16$. Từ một mẫu có kích thước n = 25 ta tìm được $\bar{X} = 12$.

Với độ tin cậy P=0,95, hãy tìm ước lượng khoảng của μ.

Giải:

Ta có: 1-P = 0,05 = $\propto$; $U_{\frac{\propto}{2}} = 1{,}96$ (tra bảng);

$$\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 - 1{,}96.\frac{4}{5} \ldots$$

$$\bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}} = 12 + 1{,}96.\frac{4}{5} = \cdots;$$

Vậy khoảng tin cậy của μ với độ tin cậy P=0,95 là

$$(\bar{X} - U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + U_{\frac{\propto}{2}}\frac{\sigma}{\sqrt{n}}) = (\ldots;\ \ldots.).$$

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

10.Kiểm định giả thuyết thống kê về tỷ lệ

Khi kiểm định giả thuyết thống kê về tỷ lệ, độ chính xác dựa vào một số điều kiện. Trước khi kiểm định giả thuyết thống kê, nên kiểm tra các điều kiện như sau, để đảm bảo tính toán và kết luận chính xác.

10.1 Mẫu dữ liệu được lấy ngẫu nhiên, hoặc kết quả dựa vào thí nghiệm ngẫu nhiên

Nếu mẫu dữ liệu không được lấy ngẫu nhiên, thì dữ liệu mẫu bình thường hay có độ lệch so với tập hợp ta lấy dữ liệu. Sẽ có rủi ro, nếu sử dụng dữ liệu mẫu lệch với tập hợp, sau đó đưa ra kết luận về tập hợp.

10.2 Mỗi quan sát trong dữ liệu đều độc lập

-Trường hợp lấy mẫu có hoàn lại.

-Trường hợp lấy mẫu không hoàn lại thì mẫu phải nhỏ hơn 10% so với tập hợp.

-Độ lệch chuẩn đòi hỏi mỗi quan sát phải độc lập. Khi lấy mẫu không hoàn lại, các quan sát không độc lập. Vì khi lấy mỗi quan sát, thay đổi tập hợp ban đầu.

-Nếu lấy mẫu nhỏ hơn 10% thì ta có thể coi là mỗi quan sát độc lập vì loại bỏ mỗi quan sát từ tập hợp không thay đổi tập hợp quá nhiều.

-Ví dụ nếu số quan sát trong mẫu là n = 150, thì điều kiện số quan sát trong tập hợp là N >= 1500.

10.3 Tỷ lệ của mẫu có phân bố chuẩn

-Trong mẫu có ít nhất 10 lần thành công và 10 lần thất bại.

-Nếu p là tỷ lệ của giả thuyết không, và n là số dữ liệu trong mẫu.

10.Kiểm định giả thuyết thống kê về tỷ lệ

Tính tỷ lệ mẫu:

Để giải bài toán, đầu tiên, tính tỷ lệ mẫu là bao nhiêu. Trong trường hợp này, tỷ lệ mẫu sẽ là 10%

Tính z_score:

Tính z_score, để biết tỷ lệ lấy được trong dữ liệu mẫu các với tỷ lệ của giả thuyết không là bao nhiêu độ lệch chuẩn.

Công thức tính độ lệch chuẩn của kiểm định giả thuyết thống kê về tỷ lệ.

Công thức tính độ lệch chuẩn của tập hợp, dựa theo công thức tính độ lệch chuẩn cho phân phối nhị phân.

10.Kiểm định giả thuyết thống kê về tỷ lệ

Công thức tính:

$$+\sigma_p = \sqrt{\frac{p.(1-p)}{n}}$$

Trong đó :
σ_p *là độ lệch chuẩn của tập hợp*
p *là tỷ lệ của giả thuyết không*
n *là số dữ liệu trong mẫu*

$$+Z\_score = \frac{\hat{p} - p}{\sigma_p}$$

Trong đó :
$\bar{x}$ *là trung bình mẫu*
μ *là trung bình tập hợp*
σ_p *là độ lệch chuẩn của tập hợp*

10. Kiểm định giả thuyết thống kê về tỷ lệ

Kiểm định tỷ lệ

Giả sử tổng thể có hai loại phần tử (phần tử có tính chất A và không có tính chất A). Gọi p là tỉ lệ phần tử có tính chất A của tổng thể.

Giả thuyết về kiểm định (H_0): $p = p_0$. *Khi đó,* (H_0) *sẽ nhận đối thuyết tương ứng là:*

(H_1): $p \neq p_0$

Chọn thống kê: $$Z = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0 q_0}}$$

Nếu (H_0) đúng thì Z có phân phối chuẩn tắc tức là $Z \sim N(0; 1)$.

Với mức ý nghĩa α cho trước, miền bác bỏ W_α tương ứng:

Nếu (H_1): $p \neq p_0$ thì $W_\alpha = (-\infty; -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}; +\infty)$

Giá trị quan sát là: $Z_0 = \frac{(f - p_0)\sqrt{n}}{\sqrt{p_0 q_0}}$, với f là tỉ lệ phần tử có tính chất A.

Kết luận: *+Nếu* $Z_0 \in W_\alpha$ *: bác bỏ giả thuyết* (H_0)*, chấp nhận đối thuyết* (H_1)

+Nếu $Z_0 \notin W_\alpha$ *: chấp nhận* (H_0)

10.Kiểm định giả thuyết thống kê về tỷ lệ

Ví dụ 1: Xét nghiệm 1000 mẫu máu của những người dân ở vùng Tây Nguyên ta thấy có 232 mẫu máu có ký sinh trùng sốt rét. Hãy kiểm định sau

$H_0: p = 0.2$

$H_1: p \neq 0.2$, mức ý nghĩa $\alpha = 0.05$

Giải:

Ta có f= 232/1000 = 0.232;

$$Z_{Score} = \left|\frac{f - p_0}{\sqrt{p_0 \cdot q_0}}\sqrt{n}\right| = \left|\frac{f - 0.2}{\sqrt{0.2x0.8}}\sqrt{1000}\right| = 2.53;$$

$U_{\alpha/2} = U_{0.025} = 1.96\ (tra\ bảng);\ \ Z_{Score} = 2.53 > 1.96.$

Vậy giả thuyết H_0 bị bác bỏ.

10.Kiểm định giả thuyết thống kê về tỷ lệ

***Note:**

Qui tắc 10: Nếu: $\left|\frac{f - p_0}{\sqrt{p_0 q_0}}\sqrt{n}\right| > U_{\frac{\alpha}{2}}$ ta bác bỏ H_0

Nếu: $\left|\frac{f - p_0}{\sqrt{p_0 q_0}}\sqrt{n}\right| \leq U_{\frac{\alpha}{2}}$ ta chấp nhận H_0.

10.Kiểm định giả thuyết thống kê về tỷ lệ

Ví dụ 2:

Một báo cáo năm ngoái nói rằng 6% người dân ở Hà Nội đi làm bằng xe đạp. Muốn kiểm định xem số liệu này có thay đổi hay không. Lấy mẫu ngẫu nhiên 240 người ở Hà Nội để kiểm tra giả thuyết này. Ta thấy trong mẫu có 24 người đi làm bằng xe đạp.

Hỏi tỷ lệ người Hà Nội đi xe đạp bây giờ, có thay đổi so với báo cáo năm ngoái không?

Giả thuyết:

Giả sử có các điều kiện để kiểm định thống kê về tỷ lệ đã được thỏa mãn. Giả thuyết rằng

-Giả thuyết không là "Tỷ lệ người lớn ở Hà Nội đi làm việc bằng xe đạp là 6%"

-Giả thuyết nghịch là "Tỷ lệ người lớn ở Hà Nội đi làm bằng xe đạp khác với 6%"

$H_0: p = 0.06$

$H_1: p \neq 0.06,$ với chọn mức ý nghĩa là 5%

THỰC HÀNH 15

```
>>> import math
>>> tong_mau=240
>>> so_mau=24
>>> p=0.06
>>> muc_y_nghia=0.05
>>> f=so_mau/tong_mau
>>> f
>>> std_dev= math.sqrt((p*(1 - p))/tong_mau)
>>> std_dev
>>> z_score = (f - p)/std_dev
>>> z_score
>>> from scipy import stats
                                          #sử dụng thư viên Thống kê gọi hàm cdf
>>> p_one_way_value = 1 - stats.norm.cdf(z_score)
>>> p_value = p_one_way_value * 2
>>> p_value                               #xấp xỉ 0.01
```

#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H_0?

-Ta có p-value gần bằng 0.01. Nó có ý nghĩa là giả sử 6% người lớn ở Hà Nội đi làm bằng xe đạp, khả năng ta lấy được một mẫu ngẫu nhiên có 10% người Hà Nội đi làm bằng xe đạp là 1%.

-Kiểm tra p-value với mức ý nghĩa là 0.05, như định nghĩa ở trên.

P_value < muc_y_nghia=0.05 (P_value ≈ 0.01)

-Từ đó có thể suy ra p-value của hai đuôi (0.01) nhỏ hơn mức ý nghĩa thống kê là 0.05.

-Vậy, bác bỏ giả thuyết không và chấp nhận giả thuyết nghịch (chấp nhận H1).

Kết luận: chứng cứ rằng "Tỷ lệ người dân ở Hà Nội đi làm bằng xe đạp **khác** với 6%".

THỰC HÀNH 16

```
>>> import math
>>> tong_mau=300
>>> so_mau=30
>>> p=0.06
>>> muc_y_nghia=0.05
>>> f=so_mau/tong_mau
>>> f
>>> std = math.sqrt((p*(1 - p))/tong_mau)
>>> std
>>> z_score = (f - p)/std
>>> z_score
>>> from scipy import stats
>>> p_one_way_value = 1 - stats.norm.cdf(z_score)
>>> p_value = p_one_way_value * 2
>>> p_value
```

```
#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H0 ?
```

THỰC HÀNH 17

Ví dụ 1: Xét nghiệm 1000 mẫu máu của những người dân ở vùng Tây Nguyên ta thấy có 232 mẫu máu có ký sinh trùng sốt rét. Hãy kiểm định sau

$H_0: p = 0.2$

$H_1: p \neq 0.2$, mức ý nghĩa $\alpha = 0.05$

Giải:

Ta có f= 232/1000 = 0.232;

$$Z_{Score} = \left| \frac{f - p_0}{\sqrt{p_0 \cdot q_0}} \sqrt{n} \right| = \left| \frac{f - 0.2}{\sqrt{0.2x0.8}} \sqrt{1000} \right| = 2.53;$$

$U_{\alpha/2} = U_{0.025} = 1.96\ (tra\ bảng)$; $Z_{Score} = 2.53 > 1.96$.

Vậy giả thuyết H_0 bị bác bỏ.

```
#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H0 ?
```

THỰC HÀNH 18

Ví dụ: *Tỉ lệ phế phẩm của máy là p = 5%. Sau khi cải tiến kỹ thuật, kiểm tra 400 sản phẩm có 12 phế phẩm. Với độ tin cậy 95%, có thể kết luận việc cải tiến kỹ thuật có chính xác hay không?*

Giải

Xét giả thuyết (H_0): $p = 0{,}05$;

(H_1): $p \neq 0{,}05$.

Chọn thống kê: $Z = \dfrac{(f-p)\sqrt{n}}{\sqrt{pq}}$ làm tiêu chuẩn kiểm định giả thuyết (H_0).

Trong đó: $p_0 = 0{,}05$, $q_0 = 1 - 0{,}05 = 0{,}95$, $n = 400$, f là thống kê nhận giá trị bằng tỉ lệ mẫu. Độ tin cậy 95% nên $1 - \alpha = 0{,}95$.

(Bảng tra $U_{\frac{\alpha}{2}} = U_{0.025} = 1{,}96$ – Bảng 5: giá trị tới hạn chuẩn U_α)

Miền bác bỏ : $W_\alpha = (-\infty; -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}; +\infty) = (-\infty; -1{,}96) \cup (1{,}96; +\infty)$

Với mẫu có kích thước $n = 400$ và tỉ lệ mẫu $f = \dfrac{12}{400} = 0{,}03$

$$Z_0 = \frac{(0{,}03 - 0{,}05)\sqrt{400}}{\sqrt{(0{,}05)(0{,}95)}} = -1{,}835$$

Kết luận: $Z_0 \notin W_\alpha \Rightarrow$ Chấp nhận giả thuyết (H_0).

```
#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H0 ?
```

11.Kiểm định trung bình

Kiểm định trung bình:

Cho đại lượng ngẫu nhiên X có trung bình E(X) = μ chưa biết.

Giả thuyết kiểm định là (H_0): $\mu = \mu_0$ và

(H_1): $\mu \neq \mu_0$

***Trường hợp:* Var(X) = σ^2 đã biết và $n \geq 30$ (hoặc $n < 30$, X có phân phối chuẩn)**

Chọn thống kê: $Z = \dfrac{(\overline{X} - \mu).\sqrt{n}}{\sigma}$

Nếu (H_0) đúng thì Z có phân phối chuẩn tắc, tức là $Z \sim N(0; 1)$.

Với mức ý nghĩa α cho trước, tìm được miền bác bỏ W_α :

Nếu (H_1): $\mu \neq \mu_0$ thì $W_\alpha = \left(-\infty; -U_{\frac{\alpha}{2}}\right) \cup \left(U_{\frac{\alpha}{2}}; +\infty\right)$

Với mẫu cụ thể, tính được giá trị quan sát là: $Z_0 = \dfrac{(\overline{X} - \mu_0).\sqrt{n}}{\sigma}$

Kết luận: *+Nếu $Z_0 \in W_\alpha$: bác bỏ giả thuyết (H_0), chấp nhận (H_1)*

+Nếu $Z_0 \notin W_\alpha$: chấp nhận giả thuyết (H_0)

THỰC HÀNH 19

KIỂM ĐỊNH TRUNG BÌNH

Khối lượng sản phẩm của đại lượng ngẫu nhiên X có trung bình theo qui định $\mu = 100g$, độ lệch chuẩn $\sigma = 0,8g$. Sau một thời gian sản xuất, người ta nghi ngờ khối lượng sản phẩm được sản xuất ra không ổn định. Kiểm tra 60 sản phẩm tính được trung bình mẫu $\bar{x} = 100,2g$. Với độ tin cậy 95%, hãy kiểm định về nghi ngờ trên (cho biết: $U_{\frac{\alpha}{2}} = U_{0.025} = 1,96$).

Giải

Xét giả thuyết (H_0): $\mu = 100g$;

(H_1): $\mu \neq 100g$

Chọn thống kê $Z = \dfrac{(\overline{X} - \mu).\sqrt{n}}{\sigma}$ tiêu chuẩn kiểm định cho giả thuyết (H_0).

Trong đó: $\sigma = 0,8g$, $\mu_0 = 100g$, $n = 60$; $\bar{x} = 100,2g$,

Giả thuyết (H_0) đúng thì $Z \in N(0; 1)$. Độ tin cậy 95% , nên $1 - \alpha = 0,95$

Miền bác bỏ: $W_\alpha = \left(-\infty; -U_{\frac{\alpha}{2}}\right) \cup \left(U_{\frac{\alpha}{2}}; +\infty\right) = (-\infty; -1,96) \cup (1,96; +\infty)$

(*cho biết:* $U_{\frac{\alpha}{2}} = U_{0.025} = 1,96$).

Với mẫu đã cho: $n = 60$, $\bar{x} = 100,2g$, giá trị quan sát thực tế của U là:

$$U_0 = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} = \frac{(100,2 - 100)\sqrt{60}}{0,8} = 1,93$$

Kết luận: $U_0 \notin W_\alpha \Rightarrow$ Chấp nhận (H_0).

THỰC HÀNH 20

KIỂM ĐỊNH TRUNG BÌNH

Khối lượng sản phẩm của đại lượng ngẫu nhiên X có trung bình theo qui định $\mu = 80g$, độ lệch chuẩn $\sigma = 0,15g$. Sau một thời gian sản xuất, người ta nghi ngờ khối lượng sản phẩm được sản xuất ra không ổn định. Kiểm tra 90 sản phẩm tính được trung bình mẫu $\bar{x} = 80,5g$. Với độ tin cậy 95%, hãy kiểm định về nghi ngờ trên (cho biết: $U_{\frac{\alpha}{2}} = U_{0.025} = 1,96$).

Khối lượng sản phẩm là đại lượng ngẫu nhiên X có luật phân phối chuẩn, biết rằng phương sai σ^2 = 16g. Kiểm tra 36 sản phẩm, tính được trung bình mẫu $\bar{x}$ = 70g. Hãy tìm khoảng ước lượng trung bình của khối lượng sản phẩm với độ tin cậy 99% (cho biết : $U_{\frac{\alpha}{2}} = 2{,}58$)

Giải :

Đặt μ = E(X) chưa biết.

Chọn thống kê $Z = \dfrac{\left(\overline{X} - \mu\right)\sqrt{n}}{\sigma} \in N(0,1)$ để ước lượng trung bình μ,

trong đó: σ=4g, n = 36 , $\bar{x} = 70\ g$

Độ tin cậy $1 - \alpha = 99\% = 0{,}99 \rightarrow U_{\frac{\alpha}{2}} = 2{,}58$ (tra bảng)

Do đó: $\varepsilon = U_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} =$

Suy ra: $\mu_1 = \bar{x} - \varepsilon$; $\mu_2 = \bar{x} + \varepsilon$;

Vậy, khoảng ước lượng trung bình khối lượng sản phẩm với độ tin cậy 99% là (..; ...) (gam).

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

THỰC HÀNH 22

ƯỚC LƯỢNG KHOẢNG VỚI PHƯƠNG SAI CHƯA BIẾT – ÔN TẬP

Khảo sát chiều cao của cây cùng độ tuổi thu được kết quả như sau :

Chiều cao (cm)	*Số cây*
< 180	*3*
180 – 190	*12*
190 – 200	*35*
200 – 210	*70*
210 – 220	*62*
220 – 230	*32*
> 230	*6*

Ước lượng trung bình chiều cao cây, với độ tin cậy 99%.

Giải: Ta có:

Bảng số liệu cho trên, khoảng chiều cao được thay thế bởi điểm giữa, khoảng < 180 được thay thế bởi 175cm, còn khoảng > 230 được thay thế bởi 235cm.

Ta được $\bar{x}$ = 208,455cm; *s'*=12,23;

Trong đó:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$ = (175*3+185*12+195*35+205*70+215*62+225*32+235*6)/ 220

= **208.45**;

+ $\hat{s}^2 = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}\right)^2 n_i$ = ((175-208.45)**2*3 + (185-208.45)**2*12 + (195-208.45)**2*35 +(205-208.45)**2*70 + (215-208.45)**2*62 + (225-208.45)**2*32 + (235-208.45)**2*6)/ 220 ≈ 149

+s' ≈ **12.23**;

THỰC HÀNH 22

ƯỚC LƯỢNG KHOẢNG VỚI PHƯƠNG SAI CHƯA BIẾT – ÔN TẬP

Chọn thống kê: $Z = \frac{(\overline{X} - \mu).\sqrt{n}}{S} \sim N(0,1)$ để ước lượng trung bình μ.

$\overline{X}$, S lần lượt là thống kê nhận giá trị trung bình mẫu và độ lệch tiêu chuẩn điều chỉnh mẫu.

Khoảng ước lượng trung bình μ là (μ_1, μ_2) trong đó : n=220

Độ tin cậy: $1 - \alpha = 99\% \Rightarrow \frac{1-\alpha}{2} = 0,495 \Rightarrow z_{\frac{\alpha}{2}} = 2,58$;

$$\varepsilon = z_{\frac{\alpha}{2}}.\frac{s}{\sqrt{n}} = 2,58.\frac{12,233}{\sqrt{220}} = 2,128$$

Suy ra: $\mu_1 = 208,455 - 2,128 = 206,327$ (cm);

$\mu_2 = 208,455 + 2,128 = 210,583$ (cm)

Vậy, khoảng UL trung bình với độ tin cậy 99% là: (206,327 cm; 210,583 cm).

```
#Cách 1 - chuyển bài toán trên sang thành dòng lệnh code Python
#Cách 2 - viết code tinh gọn tương tự như THỰC HÀNH 11,12
```

THỰC HÀNH 23

KIỂM ĐỊNH GIẢ THUYẾT – ÔN TẬP

Khối lượng sản phẩm của đại lượng ngẫu nhiên X có trung bình theo qui định μ = 160g, độ lệch chuẩn σ = 0,8g. Sau một thời gian sản xuất, người ta nghi ngờ khối lượng sản phẩm được sản xuất ra không ổn định. Kiểm tra 120 sản phẩm tính được trung bình mẫu $\bar{x}$ = 160,2g. Với độ tin cậy 95%, hãy kiểm định về nghi ngờ trên (cho biết: $U_{\frac{\alpha}{2}} = U_{0.025} = 1{,}96$).

Giải

Xét giả thuyết (H_0): $\mu = 160g$;

(H_1): $\mu \neq 160g$

Chọn thống kê $Z = \dfrac{(\overline{X} - \mu).\sqrt{n}}{\sigma}$ tiêu chuẩn kiểm định cho giả thuyết (H_0).

Trong đó: $\sigma = 0{,}8g$, $\mu_0 = 160g$, $n = 120$; $\bar{x}$ = *160,2g*,

Giả thuyết (H_0) đúng thì $Z \in N(0; 1)$. Độ tin cậy 95% , nên $1 - \alpha = 0{,}95$

Miền bác bỏ: $W_\alpha = \left(-\infty; -U_{\frac{\alpha}{2}}\right) \cup \left(U_{\frac{\alpha}{2}}; +\infty\right) = (-\infty; -1{,}96) \cup (1{,}96; +\infty)$

(*cho biết:* $U_{\frac{\alpha}{2}} = U_{0.025} = 1{,}96$).

Với mẫu đã cho: $n = 120$, $\bar{x}$ = 160,2g, giá trị quan sát thực tế của U là:

$$U_0 = \frac{(\bar{x} - \mu_0)\sqrt{n}}{\sigma} = \frac{(160{,}2 - 160)\sqrt{120}}{0{,}8} = 2.73$$

Kết luận: $U_0 \in W_\alpha \Rightarrow$ Bác bỏ (H_0), Chấp nhận (H_1).

```
#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H0 ?
```

THỰC HÀNH 24

KIỂM ĐỊNH GIẢ THUYẾT – ÔN TẬP

Xét nghiệm 850 mẫu máu của những người dân ở vùng Tây Nguyên ta thấy có 125 mẫu máu có ký sinh trùng sốt rét.

Hãy kiểm định :

H_0: p = 0,5

H_1: p ≠ 0,5, mức ý nghĩa α = 0,05 (*Bảng tra* $U_{\frac{\alpha}{2}} = U_{0.025} = 1{,}96$)

Giải:

Xét giả thuyết (H_0): $p = 0{,}5$;

(H_1): $p \neq 0{,}5$.

Chọn thống kê: $Z = \dfrac{(f-p)\sqrt{n}}{\sqrt{pq}}$ làm tiêu chuẩn kiểm định giả thuyết (H_0).

Trong đó: $p_0 = 0{,}5$, $q_0 = 1 - 0{,}5 = 0{,}5$, $n = 850$, ($f = \frac{n_A}{n} = 125/850$) f là thống kê nhận giá trị bằng tỉ lệ mẫu. Độ tin cậy 95% nên $1 - \alpha = 0{,}95$.

(*Bảng tra* $U_{\frac{\alpha}{2}} = U_{0.025} = 1{,}96$ – *Bảng 5: giá trị tới hạn chuẩn* U_α)

Miền bác bỏ : $W_\alpha = \left(-\infty; -U_{\frac{\alpha}{2}}\right) \cup \left(U_{\frac{\alpha}{2}}; +\infty\right) = (-\infty; -1{,}96) \cup (1{,}96; +\infty)$

Với mẫu có kích thước $n = 850$ và tỉ lệ mẫu f = 125/850 = 0.147

$Z_0 = \dfrac{(0{,}147-0{,}5)\sqrt{850}}{\sqrt{0{,}5x0{,}5}} = -20.58$

Kết luận: $Z_0 \in W_\alpha$: Bác bỏ giả thuyết (H_0), chấp nhận (H_1).

```
#so sánh giữa p_value với mức ý nghĩa
#kết luận: bác bỏ giả thuyết H0 ?
```

TRƯỜNG ĐẠI HỌC
VĂN LANG
Đạo đức - Ý chí - Sáng tạo

KHOA CÔNG NGHỆ THÔNG TIN