Bài 6
Kiểm định giả thuyết

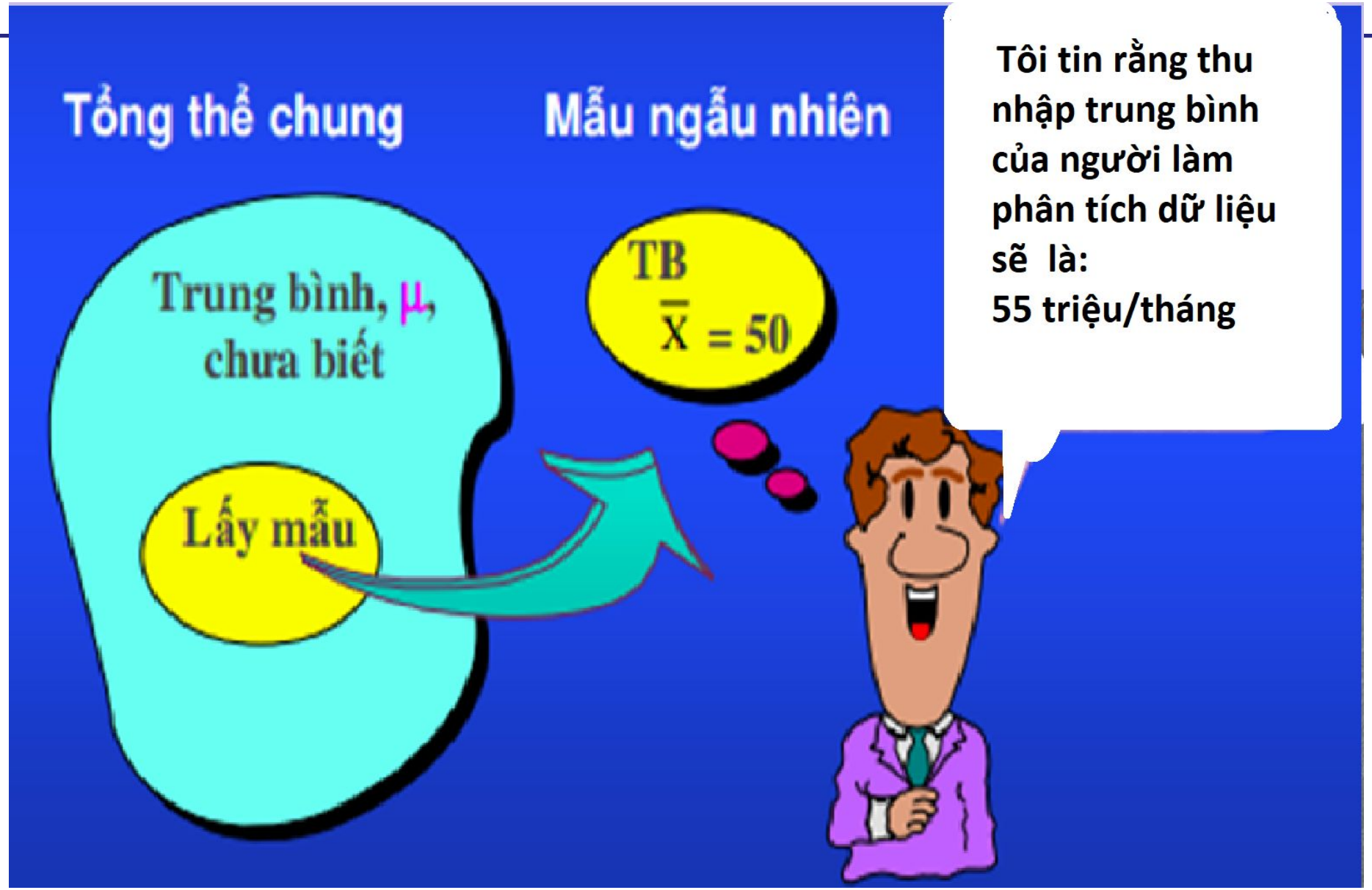
Khóa học: Phân tích dữ liệu với Python

Mục tiêu

- Biết cách viết giả thuyết thống kê
- Biết chọn phương pháp kiểm định hợp lý
- Kiểm định được giả thiết về giá trị trung bình của một thuộc
- So sánh hai giá trị trung bình
  - Mẫu độc lập
  - Mẫu phụ thuộc

Thảo luận

- Muốn đưa ra 1 nhận định về số đông nhưng lại chỉ có thông tin về thiểu số?

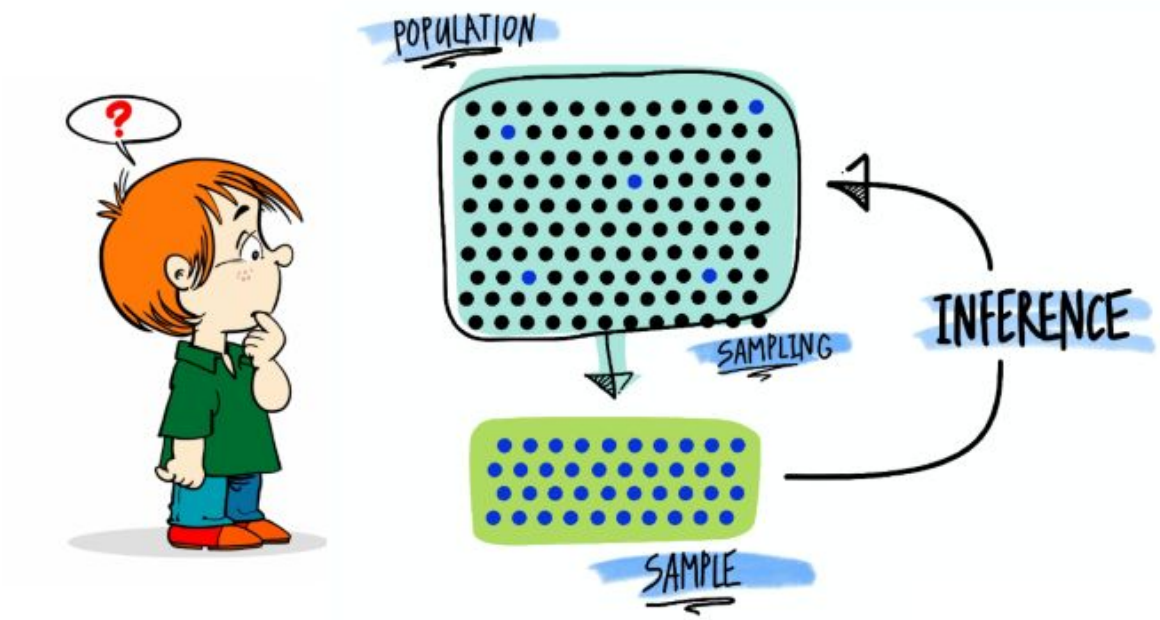

Tổng thể chung
Mẫu ngẫu nhiên
Trung bình, μ,
chưa biết
Lấy mẫu
TB
X = 50
Tôi tin rằng thu nhập trung bình của người làm phân tích dữ liệu sẽ là:
55 triệu/tháng

Nội dung

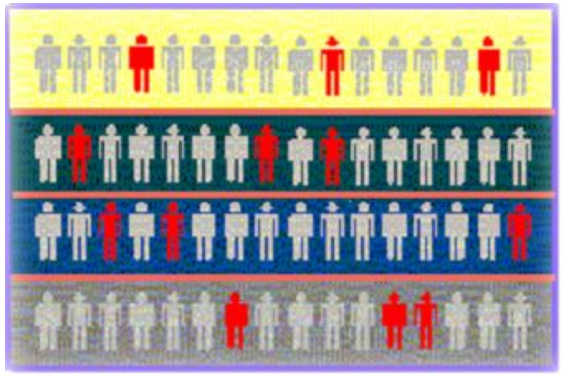
- Kiểm định giải thuyết
- Tổng thể chung ta tổng thể mẫu
- Giả thuyết thống kê
- Các loại sai lầm trong kiểm định
- Kiểm định giá trị trung bình của một thuộc tính
- So sánh giá trị trung bình: hai mẫu độc lập
- So sánh giá trị trung bình: hai mẫu phụ thuộc

Kiểm đinh giả thuyết

Kiểm định giả thuyết là việc chấp nhận hoặc bác bỏ một nhận định về tập hợp tất cả các đối tượng liên quan dựa trên một tập đối tượng mẫu



Tổng thể chung ta tổng thể mẫu



Tổng thể chung là: Tập hợp các đối tượng liên quan đến 1 vấn đề nghiên cứu.

- **Tổng thể bộc lộ**: Biết được tất cả các đối tượng

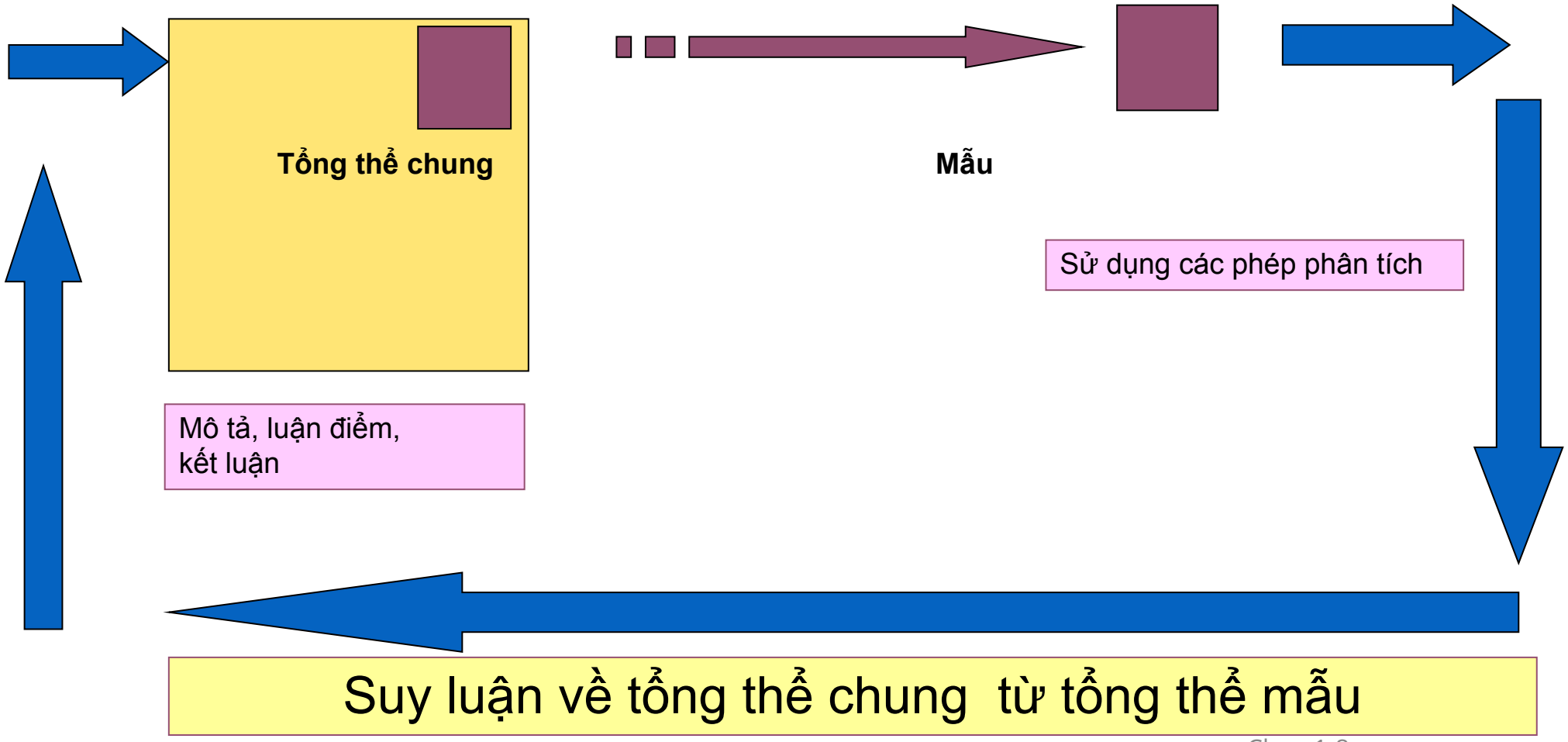
Ví dụ: nhân viên của công ty

- **Tổng thể tiềm ẩn**: Không biết được tất cả các đối tượng

Ví dụ: nghiên cứu sự hài lòng của tất cả các hành khách về dịch của hàng không. Tổng thể chung bao gồm khách hàng cũ, khách hàng đang sử dụng dịch vụ, khách hàng mới

Tổng thể mẫu

Tổng thể mẫu: là tập hợp một số đối tượng được lựa chọn từ tổng thể chung để phân tích



Giả thuyết thống kê

- Giả thuyết thống kê là giả thuyết về một vấn đề nào đó của tổng thế chung

Ví dụ:

Công ty Coca Cola cho rằng, trung bình một người dân ở Mỹ một năm sẽ tiêu thụ 2 lít Coca

Công ty Coca Cola cho rằng loại Coca màu xanh sẽ được khách hàng yêu thích hơn loại Caca màu nâu đen truyền thống

Công ty LifeBoy cho rằng phần lớn khách hàng thích xà bông mùi bạc hà hơn các mùi khác tiến hành sản xuất xà bông bạc hà phiên bản đại trà?

Cách viết giả thuyết

- Giả thuyết không (null hypothesis) – H0 là giả thuyết mà ta muốn kiểm định
- Giả thuyết đối (alternative hypothesis)-H1 là giả thuyết đối lập với giả thuyết không

Ví dụ: Gọi μ là giá trị trung bình của lượng coca mà 1 người dân Mỹ sẽ uống trong ví dụ trước

Loại kiểm định	H0	H1
Hai phía		
Phía trái		
Phía phải		

Bài tập

Hãy viết nốt các giả thuyết kiểm định cho các ví dụ ở trên

Các loại sai lầm khi kiểm định giả thuyết

Thực tế \ Kết luận		
	Kết luận đúng	Sai lầm loại 1
	Sai lầm loại 2	Kết luận đúng

$\dot{\alpha}$: Xác suất cho phép mắc sai lầm loại 1 – mức ý nghĩa

Pvalue: Xác suất phạm sai lầm lớn nhất nếu bác bỏ H0

→ P values $\geq \alpha$: chưa đủ cơ sở để bác bỏ giả thuyết H0

Các bước tiến hành một kiểm định giả thuyết

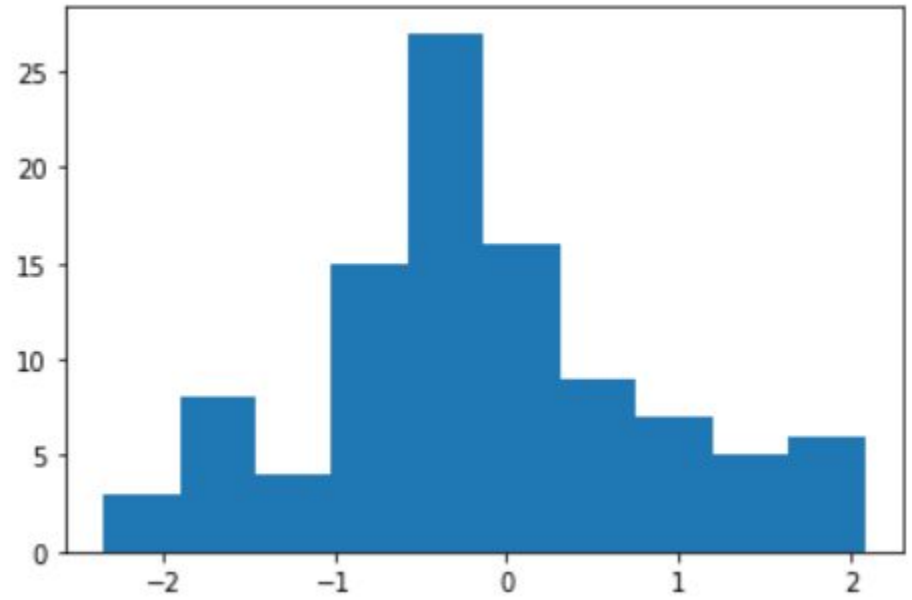
- Phát biểu giả thuyết không và giả thuyết đối
- Chọn mức ý nghĩa
- Chọn phương pháp kiểm định
- Tính toán các giá trị kiểm định theo phương pháp kiểm định
- Kết luận

Kiểm định giả thuyết cho giá trị trung bình

- Loại kiểm định: One Sample T test
- Điều kiện áp dụng: mẫu tuân theo phân phối chuẩn

```
import numpy as np
import matplotlib.pyplot as plt

data = np.random.normal(0,1,100)
plt.hist(data)
```



Kiểm định giả thuyết cho giá trị trung bình

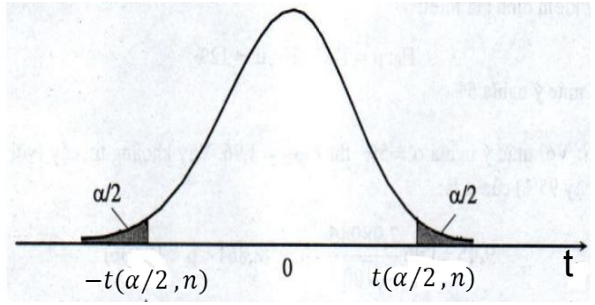
- Viết giả thuyết kiểm định

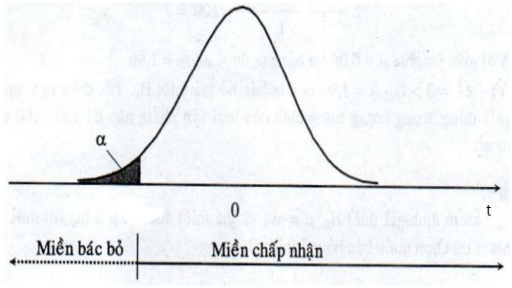
Loại kiểm định	H0	H1
Hai phía		
Phía trái		
Phía phải		

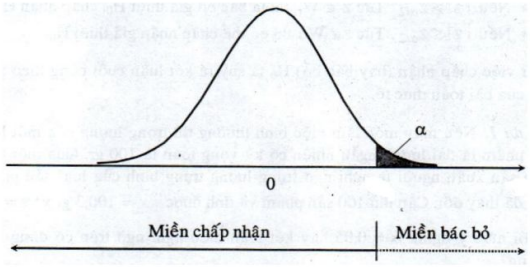
Thực hiện kiểm định giả thuyết cho giá trị trung bình

: Sử dụng giá trị thống kê t

$$t\_statistic = \frac{\bar{x} - \mu_0}{s}$$







t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Thực hiện kiểm định giả thuyết cho giá trị trung bình

Hàm sử dụng: **scipy.stats.ttest_1samp**

Parameters

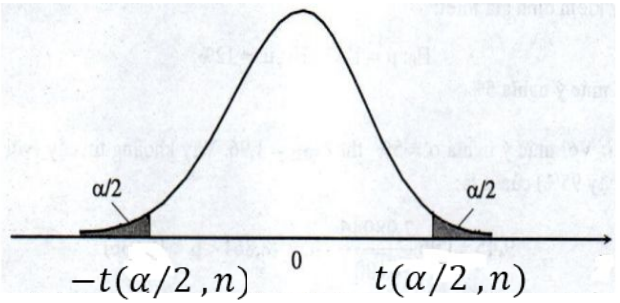
- a : array_like - dữ liệu của thuộc tính
- popmean: float hoặc array – giá trị trung bình muốn kiểm định
- axis: int or None, optional
- nan_policy : {'propagate', 'raise', 'omit'}, optional
- Định nghĩa cách xử lý giá trị nan, mặc định'propagate':

Returns

- statistic : float or array (t-statistic)
- pvalue : float or array (Two-sided p-value)

Đọc kết quả kiểm định

Do hàm **ttest_1samp** chỉ trả về giá trị Pvalue của kiểm định hai phía nên cần kết hợp cả (t- statistic, pvalue) để đưa ra kết luận



Loại kiểm định	H0	H1	Bác hỏ H0, chấp nhận H1 khi
Hai phía			
Phía trái			
Phía phải			

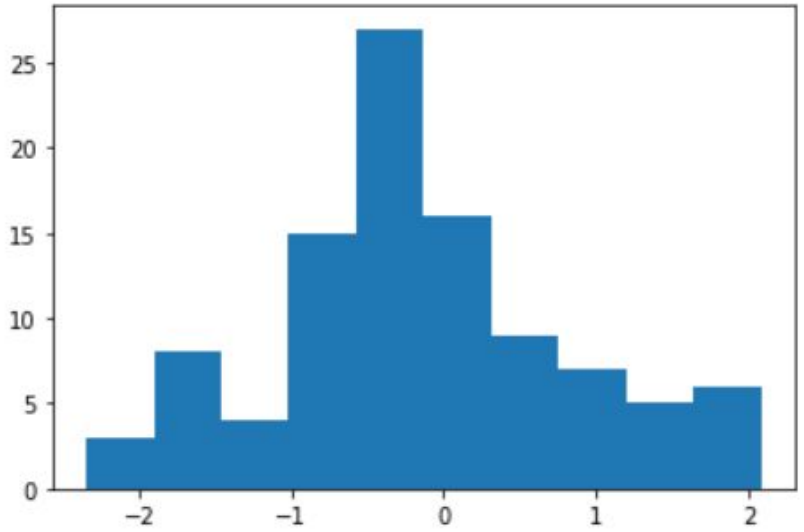
Demo

Ví dụ 1:

```
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
data = np.random.normal(0,1,100)
plt.hist(data)

stats.ttest_1samp(data, 0)
```

Out:




```
Statistic = 1.3922816321904066 pvalue=
0.1669566662536031 7
```

Kết luận: với mức ý nghĩa 5%, chấp nhận giả thuyết: giá trị trung bình của tổng thể chung bằng 0

So sánh giá trị trung bình: hai mẫu độc lập

- Hai mẫu độc lập: là hai tập đối tượng riêng biệt, từ tổng thể nghiên cứu và không có bất kỳ mối liên hệ gì với nhau trên các thuộc tính kiểm định
- Ví dụ:
  - So sánh thu nhập trung bình trên tháng giữa người dân ở Hà Nội và Hồ Chí Minh
  - So sánh năng suất chăn nuôi bằng loại thức thức ăn cũ so với thức ăn mới
  - So sánh chỉ số huyết áp giữa người tập thể dục và không tập thể dục
  - So sánh độ bền giữa hai loại vật liệu/ 2 loại sản phẩm
  - ...

So sánh giá trị trung bình: hai mẫu độc lập

- Điều kiện áp dụng
  - Dữ liệu định lượng, tuân theo phân phối chuẩn
- Viết giả thuyết thống kê

Loại kiểm định	H0	H1
Hai phía		
Phía trái		
Phía phải		

So sánh giá trị trung bình: hai mẫu độc lập

- Thực hiện kiểm định với Python

Hàm **scipy.stats.ttest_ind**

Parameter:

- *a, b: array_like*
- *equal_var: bool, optional (thường chọn equal_var = False khi không có thông tin)*
- nan_policy: *{'propagate', 'raise', 'omit'}, optional*

Return Values

- Statistic: *float or array (t – statistic)*
- Pvalue: *float or array* (giá trị p values 2 phía)

Lưu ý: trong hàm scipy.stats.ttest_ind, giá trị $\mu_0 = 0$, phép kiểm định hai phía

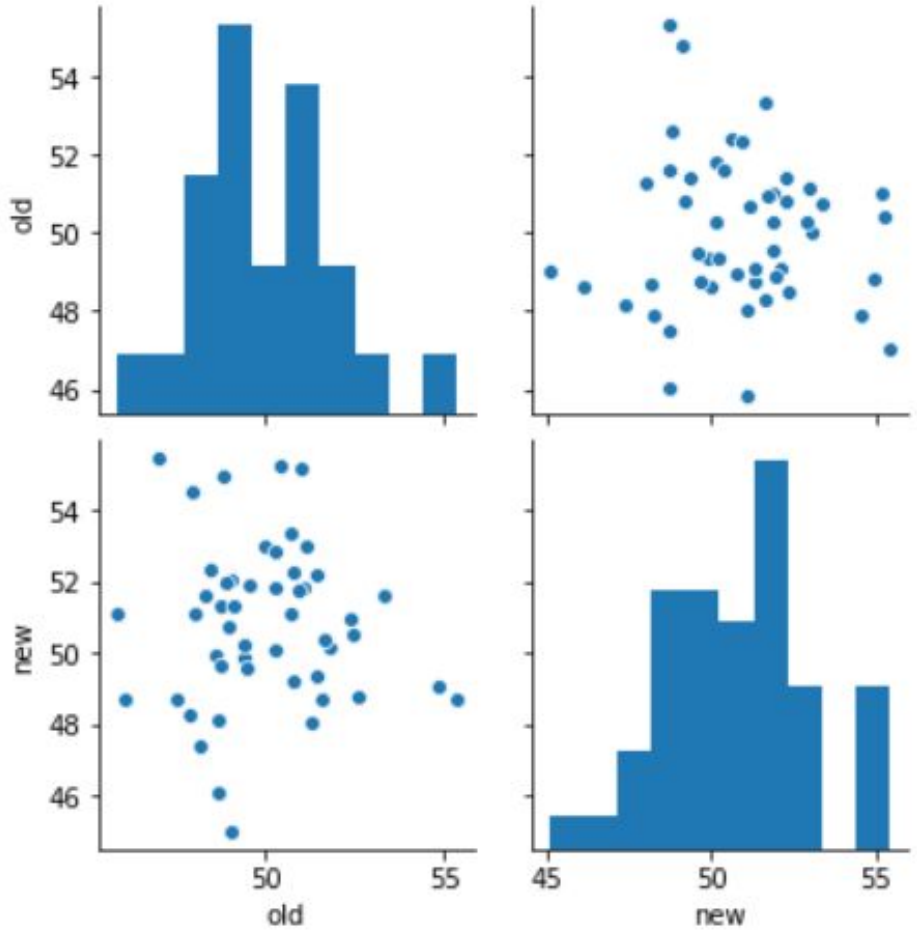
1. Để kiểm định cho giả thuyết H0: $\mu_1 - \mu_2 = \mu_0$, cần biến đổi giá trị của tham số đầu vào $b = b + \mu_0$ hoặc $(a = a + \mu_0)$ trước khi thực hiện kiểm định

2. Cách đọc kết quả kiểm định tương tự như với hàm scipy.stats.ttest_1samp

Demo

Ví dụ 1

Người ta cho rằng loại thức ăn chăn nuôi mới sẽ làm tăng trọng lượng trung bình của mỗi con lợn so với thức ăn cũ khi xuất chuồng thêm 2kg/con. Người ta tiến hành cân 50 con lợn của nhóm lợn được cho thức ăn cũ, và 50 con lợn được chăn nuôi bằng thức anh cũ. Với mức ý nghĩa bằng 10% hãy kiểm định giả thuyết trên



$H_0: \mu_1 - \mu_2 \geq 2$ ⟸ (trong đó 1~ new, 2~ old)

$H_1: \mu > 2$

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("thuc_An_Chan_Nuoi.csv")
# so sánh phân bố của 2 thuộc tính
sns.pairplot(df)
# biến đổi dữ liệu trước khi đưa vào kiểm định
print ("kiểm định hai phía với giả thuyết thức ăn mới không làm tăng năng suất")
print(stats.ttest_ind(df["new"], df["old"], equal_var=False))
df["new_processed"] = df["new"]-2
print ("kiểm định hai phía với giả thuyết làm tăng năng suất 2 kg")
stats.ttest_ind(df["new_processed"], df["old"], equal_var=False)
```

- Kiểm định hai phía với giả thuyết thức ăn mới không làm tăng năng suất
  - statistic=-1.9734203923573161,
  - pvalue=0.05132285569680214 4
- Kiểm định hai phía với giả thuyết thức ăn mới làm tăng năng suất với con số 2 kg/con
  - statistic=-6.683254264331812
  - pvalue=1.5334614691649278e-09

 Thức ăn mới có làm tăng năng suất nhưng không tới 2kg/con

So sánh giá trị trung bình: hai mẫu phụ thuộc

Ví dụ:

- Kiểm định hiệu quả của thuốc giảm mỡ máu trên bệnh nhân bằng cách so sánh chỉ số cholesterol trước và sau khi dùng thuốc
- Kiểm định giả thuyết rằng, nhìn chung trên thị trường: doanh số bán hàng của mặt hàng A sẽ tốt hơn mặt hàng B

STT bệnh nhân	Cholesterol trước	Cholesterol sau
1	5.2	5
2	4.8	4.9
3	6	5.5
4	5	5
..	…	…..

STT Cửa hàng	A	B
1	100	102
2	107	90
3	50	54
4	30	50
..	…	…..

Pair T test

- Trường hợp áp dụng: Dữ liệu định lượng
- Độ lệch giữa hai thuộc tính tuân theo phân phối chuẩn

Loại kiểm định	H0	H1
Hai phía		
Phía trái		
Phía phải		

Thực hiện kiểm định

- Hàm sử dụng: **scipy.stats.ttest_rel**

Parameter:

- *a, b: array_like (phải có cùng độ dài)*
- nan_policy: *{'propagate', 'raise', 'omit'}, optional*

Return Values

- Statistic: *float or array (t – statistic)*
- Pvalue: *float or array* (giá trị p values 2 phía)

Lưu ý: trong hàm scipy.stats.ttest_rel, giá trị $\mu_0 = 0$, phép kiểm định hai phía

1. Để kiểm định cho giả thuyết H0: $\mu_d = \mu_0$, cần biến đổi giá trị của tham số đầu vào $b = b + \mu_0$ hoặc $(a = a + \mu_0)$ trước khi thực hiện kiểm định

2. Cách đọc kết quả kiểm định tương tự như với hàm scipy.stats.ttest_1samp

Demo

Tình huống

Một nghiên cứu cho rằng, trẻ em sau 1 tháng đi nhà trẻ, do bị xáo trộn tâm lý nên bé nào cũng bị sụt cân. Trong chương trình chiêu sinh của mình, Trường mầm A đã tuyên bố rằng tại trường của họ trung bình cân nặng của học sinh khi đã đi học được 1 tháng so với ngày đầu tiên vào trường không hề bị giảm. Hội phụ huynh đã tiến hành kiểm định tuyên bố đó ở trường bằng cách ghi lại chỉ số cân nặng của một số bé trẻ trước và sau khi nhập học 1 tháng. Với mức ý nghĩa 5% hãy kiểm định nhận định của trường

- Gọi d là độ lệch giữa trước và sau về cân nặng của trẻ

$$H_0: \mu_d = 0$$

$$H_1: \mu > 0$$

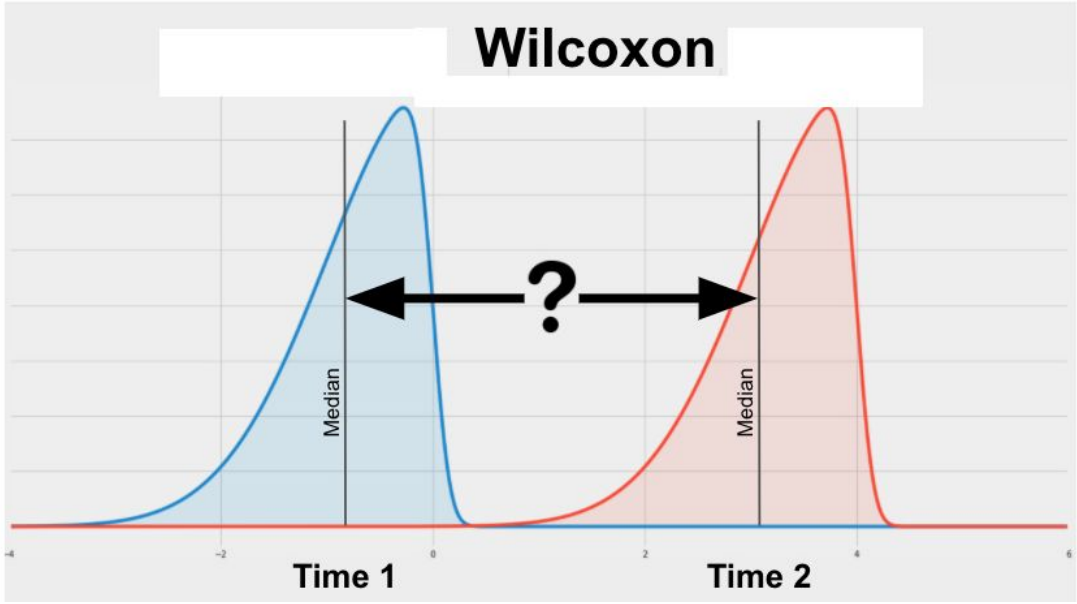
```
from scipy import stats
import pandas as pd
df = pd.read_csv("Man_Non.csv", index_col="STT")
print(stats.ttest_rel(df['before'],df['after'] ))
```

Kết quả:

- statistic=1.3430356655317932
- pvalue=0.1959510399396823

Wilcoxon test

- Điều kiện áp dụng:
  - Thuộc tính thuộc thang đo thứ bậc hoặc là thuộc tính định lượng
  - Dữ liệu tuân theo phân bố chuẩn



Thực hiện kiểm định

: Hàm sử dụng: **scipy.stats.Wilcoxon**

Parameters

- **X:** *array_like : thuộc tính thứ nhật, hoặc hiệu của 2 thuộc tính được tính theo cặp*
  - **Y:** *array_like, defaut = None* – ***dãy giá trị của thuộc tính thứ 2***
  - **alternative***{"two-sided", "greater", "less"}, optional'*

  Loại kiểm định: hai phía, 1 phía trái, 1 phía phải

Returns

Statistic: *float*

Pvalue: *float*

- *cách đọc kết quả:* $pvalue < \alpha$ -→Bác bỏ H0, chấp nhận H1

Nhà phân phối thời trang Alain Delon tại Hà Nội vừa tung ra thị trường một loại túi thời trang mới với hai màu đen và đỏ borheaux. Nhà phân phối muốn biết liệu có sự khác biệt về số lượng bán ra của hai loại túi nên đã lấy số lượng túi bán được tại 15 địa điểm bán hàng. Hãy giúp nhà phân phối có đủ cơ sở để đưa ra kết luận ở mức ý nghĩa 5%

```
from scipy import stats
# Độ lệch số lượng túi bán ra: màu đen – màu đỏ
d= [3, 4, -1, 2, 1, -2, 5, -1, -2, -2, 3,0,-2,5,-2]
print( stats.wilcoxon(d))
```

Kết quả

statistic=36.5, pvalue=0.31023779780459937

Tổng kết

Qua bài học này bạn đã nắm

- Viết giả thuyết kiểm định
- Thực hiện kiểm định cho giá trị trung bình của một thuộc tính (one sample T test)
- Thực hiện kiểm định so sánh hai giá trị trung bình
  - Hai mẫu độc lập (independent T test)
  - Hai mẫu phụ thuộc
    - Độ lệch tuân theo phân phối chuẩn (pair T test)
    - Độ lệch không tuân theo phân phối chuẩn (Wilcoxon)