#khai báo các thư viện

import math

import statistics

import numpy as np

import scipy.stats

import pandas as pd

#tạo dữ liệu

x = [8.0, 1, 2.5, 4, 28.0]

x\_with\_nan = [8.0, 1, 2.5, math.nan, 4, 28.0]

x

* [8.0, 1, 2.5, 4, 28.0]

#Tạo x với nan

x\_with\_nan

* [8.0, 1, 2.5, nan, 4, 28.0]

#tạo np.ndarray và pd.Series tương ứng với x và x\_with\_nan

y, y\_with\_nan = np.array(x), np.array(x\_with\_nan)

z, z\_with\_nan = pd.Series(x), pd.Series(x\_with\_nan)

y

* array([ 8. , 1. , 2.5, 4. , 28. ])

#tạo y, z với nan

y\_with\_nan

* array([ 8. , 1. , 2.5, nan, 4. , 28. ])

z\_with\_nan

0 8.0

1 1.0

2 2.5

3 NaN

4 4.0

5 28.0

* dtype: float64

#tính giá trị trung bình

mean\_ = sum(x) / len(x)

mean\_

* 8.7

#sử dụng hàm statistics.mean(x) để thống kê

mean\_ = statistics.mean(x)

mean\_

#chỉ chạy trên python 3.8

mean\_ = statistics.fmean(x)

mean\_

#thống kê dữ liệu không phải số thì trả về giá trị không phải số(nan). kết quả này giống với kết quả của hàm sum

mean\_ = statistics.mean(x\_with\_nan)

mean\_

* Nan

#Hàm np.mean trả về giá trị trung bình

mean\_ = np.mean(y)

mean\_

8.7

mean\_ = y.mean()

mean\_

8.7

wmean = np.average(z, weights=w)

wmean

mean /miːn/

mean hay đầy đủ là arithmetic mean, là tổng đại số của tất cả các số trong data, chia cho số điểm dữ liệu (data point). Mean được ký hiệu bằng ký tự latin “muy”: μ

Xác suất thống kê tiếng Việt gọi đây là trung bình. Trong tiếng Anh, mean hay được gọi là the average (trung bình), mặc dù đây chỉ là một trong nhiều cách tính trung bình toán học khác.

Mean là một giá trị dùng để biểu diễn điểm trung tâm (central location) của data. Tính trên Python:

def mean(data):  
 return sum(data)/len(data)mu = mean(data) # mu - đọc là muy, là tên tiếng Anh của chữ latin: μ  
print(mu)# 12.333333333333334

mean rất dễ bị “lệch” khi dữ liệu có những giá trị rất khác biệt so với các giá trị còn lại, ví dụ số 100 trong data ở trên là một giá trị rất khác. Bởi thế, nếu chỉ dùng mean để đưa ra “cảm nhận" về điểm trung tâm của bộ dữ liệu sẽ không được chính xác. Người ta đưa ra thêm khái niệm median để tính điểm trung tâm theo một cách khác, mà giá trị thu được sẽ biểu diễn điểm trung tâm một cách “tốt hơn”.

median /ˈmiːdɪən/

median — giá trị ở giữa (middle value)

Xác suất thống kê tiếng Việt gọi đây là trung vị.  
Ta sắp xếp data theo chiều tăng dần, sau đó xác định giá trị ở giữa. Khi xác định giá trị ở giữa, xảy ra 2 trường hợp:  
- số data point là chẵn: lấy 2 điểm ở giữa và lấy giá trị trung bình (cách làm này gọi là tính “mean of middle two”).  
- số data point là lẻ: lấy luôn điểm ở giữa.

def median(data):  
 data = sorted(data)  
 n = len(data)  
 if n % 2 == 1:  
 return data[n//2]  
 else:  
 i = n//2  
 return (data[i - 1] + data[i])/2median(data)  
# 1.5

Rõ ràng, median (1.5) phản ánh điểm giữa của tập dữ liệu chính xác hơn so với mean (12.333333333333334).

median low và median high

median có thể là một giá trị không trùng với 1 data point cụ thể (trường hợp có chẵn data point), mà có lúc người ta lại muốn lấy một data point để làm điểm trung tâm của bộ dữ liệu. Median low là data point nhỏ hơn trong 2 điểm giữa, median high là data point lớn hơn trong 2 điểm ở giữa.

def median\_low(data):  
 data = sorted(data)  
 n = len(data)  
 if n % 2 == 1:  
 return data[n//2]  
 else:  
 return data[n//2-1]def median\_high(data):  
 data = sorted(data)  
 n = len(data)  
 return data[n//2]

mean có vẻ “toán học” hơn, median có vẻ trực quan hơn. Mỗi giá trị sẽ được sử dụng trong từng trường hợp với yêu cầu cụ thể.

mode

Trong xác suất tiếng Việt gọi là “mốt” 😂  
Giá trị xuất hiện nhiều lần nhất trong tập dữ liệu.

def mode(data):  
 dmax = data[0]  
 for d in data:  
 if data.count(d) > data.count(dmax):  
 dmax = d  
 return dmaxmode(data)  
# 0.25

Với mode, ta có thể xử lý cả tập dữ liệu không phải số (nominal )

mode(["red", "blue", "blue", "red", "green", "red", "red"])  
# 'red'

Vậy nếu một tập có nhiều giá trị cùng xuất hiện nhiều lần (số lần bằng nhau) thì mode sẽ là giá trị nào?

Deviation

Nếu chỉ đại diện 1 tập dữ liệu bằng giá trị mean hay median, thì 2 tập sau được coi là như nhau:

mean([1,5,9]) == mean([3,4,5,6,7])

Rõ ràng ta cần thêm một thước đo để thể hiện việc tập trung hay phân tán của các giá trị so với điểm ở giữa. Suy nghĩ đơn giản, có thể tính giá trị trung bình của độ lệch mỗi data point so với mean:

def mydev(data):  
 mu = mean(data)  
 return sum([point-mu for point in data])/len(data)mydev([1,5,9])# 0.0

Nhưng bởi mean là giá trị trung bình, nên tổng các “độ lệch” này sẽ luôn bằng 0. Cách tính tự nghĩ này không ổn, ta thử một cách khác, dùng trung bình các giá trị tuyệt đối của các độ lệch: [1]

def mydev(data): # this is MAD   
 mu = mean(data)  
 return sum([abs(point-mu) for point in data])/len(data)mydev([1,5,9])  
# 2.6666666666666665

Với giá trị 2.6666666666666665, có thể hiểu rằng các điểm trong tập dữ liệu được phân bố cách điểm trung tâm (ở ví dụ này là mean) một khoảng trung bình là ~2.6

Công thức tự nghĩ này, thực chất có một cái tên chính thống trong khoa học là: Mean absolute deviation — trung bình độ lệch tuyệt đối, viết tắt là MAD. Độ lệch ở đây có thể hiểu là độ lệch so với giá trị trung tâm, tính bằng mean HOẶC median (tuỳ trường hợp cụ thể sẽ ghi rõ). Giá trị MAD được cho là “tốt” để mô tả độ phân tán của bộ dữ liệu, ít bị ảnh hưởng bởi các giá trị “lệch”.

Khi deviation càng lớn thì dữ liệu càng nằm phân tán, deviation nhỏ tức là dữ liệu tập trung quanh điểm trung tâm.

Phương sai và độ lệch (tiêu) chuẩn

variance và standard deviation

Đây là hai khái niệm phổ biến trong mọi cuốn sách viết về thống kê.  
Dưới góc độ của một người mới bắt đầu học, tôi không hiểu lý do cặn kẽ tại sao chúng được sử dụng. Hầu hết, các cuốn sách sẽ chỉ giới thiệu cái tên, ký hiệu và công thức mà không giải thích tại sao. Có thể vì đây là các khái niệm đã có từ lâu, được phổ biến rộng rãi.  
Khái niệm variance (phương sai, thường viết tắt là var), theo tôi hiểu là “bình **phương** của **sai** lệch (độ lệch)”, được giới thiệu là một công thức tính: giá trị trung bình của tổng các bình phương độ lệch

def variance(data):  
 mu = mean(data)  
 return sum([(point-mu)\*\*2 for point in data])/len(data)variance([1,5,9])  
# 10.666666666666666

Khái niệm standard deviation (độ lệch CHUẨN, thường viết tắt là STD hay SD), không hiểu từ đâu ra mà được cho là “tiêu chuẩn” (standard). Khái niệm này được định nghĩa bằng công thức tính: căn bậc 2 của variance: