دوره جامع پایتون:

بخش علوم داده

جلسه دوازدهم

دکتر ذبیح اله ذبیحی

statistics ماژول

Import statistics

آمار و داده ها

- شاخص های تمرکز:میانگین، میانه، مد

❑در میانگین نقش داده ها(داده بزرگ، کوچک) اهیمت دارد، برای داده های کمی قابل استفاده است

❑در میانه تعداد داده اهمیت دادرد و برای داده های کمی و کیفی قابل استفاده است

اگر میانگین و میانه یکی باشند توزیع مقادیر کاملا متقارن خواهد بود

اگر میانگین بزرگتر از میانه باشد توزیع مقادیر دارای چولگی مثبت (به طرف راست) است.

اگر میانگین کوچکتر از میانه باشد توزیع مقادیر دارای چولگی منفی (به طرف چپ) است

- شاخص های پراکندگی: انحراف معیار، واریانس

تابع میانگین حسابی mean(data)

- اگر

$$data = [x_1, x_2, \dots, x_N]$$

آنگاه

$$mean = \frac{1}{N}\sum_{i=1}^{N} x_i$$

```
import statistics
data=[4,5,6]
s=statistics.mean(data)
print(s)
```

```
import statistics
s=statistics.mean([4,5,6])
print(s)
```

```
import statistics
print(statistics.mean([4,5,6]))
```

تابع میانگین fmean()

تابع fmean سریع تر از mean است و همواره یک عددی اعشاری باز می گرداند.

```
import statistics
data=[1.5,2,3]
s1=statistics.mean(data)
print(s1)
s2=statistics.fmean(data)
print(s2)
```

تابع میانگین هندسی geometric_mean()

اگر

$$data = [x_1, x_2, \dots, x_N]$$

آنگاه

$$geometric\ mean = (x_1 x_2 \dots x_N)^{\frac{1}{N}}$$

```
import statistics
data=[2,4,6]
s=statistics.geometric_mean(data)
print(s)
```

تابع میانگین همساز harmonic_mean(*data*)

- اگر

$$data = [x_1, x_2, \dots, x_N]$$

آنگاه

$$\text{harmonic\_mean} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}$$

نکته: برای استفاده از این تابع، داده ها نباید شامل مقدار منفی باشد.

مثال

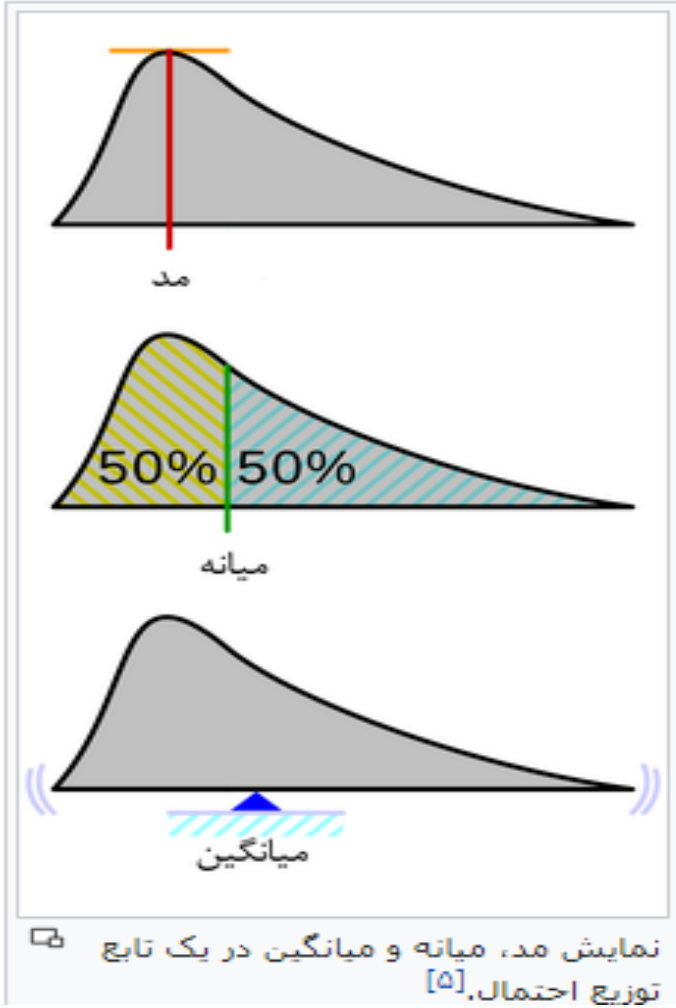
```
import statistics
data=[4,5,6]
s=statistics.harmonic_mean(data)
print(s)
```

```
import statistics
s=statistics.harmonic_mean([4,5,6])
print(s)
```

```
import statistics
print(statistics.harmonic_mean([4,5,6]))
```

تابع میانه median(data)

- میانه عددی است که یک جمعیت آماری یا یک توزیع احتمالی را به دو قسمت مساوی تقسیم می‌کند. یکی از مزیت‌های مهم میانه نسبت به میانگین این است که میانه از اعداد بسیار بزرگ و بسیار کوچک مجموعه اندازه‌ها متأثر نمی‌شود.
- برای تعداد فردی از اعضای لیست، ابتدا داده های لیست را مرتب میکند و بعد داده میانه لیست را چاپ می کند.
- برای تعداد زوجی از اعضای لیست، ابتدا داده های لیست را مرتب می کند و سپس میانگین دو داده میانی را چاپ می کند.



نمایش مد، میانه و میانگین در یک تابع توزیع احتمال.[۵]

مثال

```
import statistics
data=[4,5,6,8,12,22,9]
s=statistics.median(data)
print(s)
----------------------------------------
import statistics
data=[0,4,5,6,8, 12,22,9,0]
s=statistics.median(data)
print(s)
```

مثال

```
import statistics
data=[1,2,3,4,5,6]
s=statistics.median(data)
print(s)
-------------------------------
import statistics
data=[1,2,4,6,5,3]
s=statistics.median(data)
print(s)
```

تابع median_high(data)

- ابتدا داده های لیست مرتب می گردد. اگر تعداد داده ها فرد باشد دقیقا داده میانه لیست برگردانده می شود و اگر تعداد داده ها زوج باشد بین دو داده میانه مقدار بالاتر را بر می گرداند.

```
import statistics
data=[1,2,3,4,5]
s=statistics.median_high(data)
print(s)
-----------------------------------
import statistics
data=[1,2,3,4,5,6]
s=statistics.median_high(data)
print(s)
```

تابع median_low(data)

- ابتدا داده های لیست مرتب می گردد. اگر تعداد داده ها فرد باشد دقیقا داده میانه لیست برگردانده می شود و اگر تعداد داده ها زوج باشد بین دو داده میانه مقدار پایین تر را بر می گرداند.

```
import statistics
data=[1,2,3,4,5]
s=statistics.median_low(data)
print(s)
-----------------------------------
import statistics
data=[1,2,3,4,5,6]
s=statistics.median_low(data)
print(s)
```

تابع median_grouped()

محاسبه‌ی میانه در داده‌های طبقه بندی شده

ابتدا از فرمول $\frac{N}{2}$ محل میانه به دست می‌آید. سپس در ستون فراوانی تجمعی، اولین ستونی که فراوانی تجمعی آن بزرگ‌تر یا مساوی $\frac{N}{2}$ در نظر گرفته می‌شود.
میانه از رابطه زیر به دست می‌آید.

$$me = L_i + \frac{\frac{N}{2} - F_{i-1}}{f_i} C$$

که ستون iم ستون شامل میانه، F_{i-1} فراوانی تجمعی ستون پیشین ستون شامل میانه، L_i حد پایین طبقه میانه‌دار، f_i فراوانی ستون شامل میانه، C طول باره و N تعداد داده‌ها است.

به‌طور مثال در جدول توزیع فراوانی زیر:

بازه	فراوانی	فراوانی تجمعی
0.5-1.5	1	1
2.5-3.5	2	3
4.5-5.5	1	4
6.5-7.5	1	5

data=[1,3,3,5,7]

$\frac{N}{2} = \frac{5}{2} = 2.5 \rightarrow$ میانه در ستون دوم است.

$$me = L_2 + (\frac{\frac{N}{2} - F_1}{f_1})C = 2.5 + (\frac{2.5\text{-}1}{2}) \times 1 = 3.25$$

پس میانه در جدول توزیع فراوانی بالا برابر 3.25 است.

statistics.median_grouped(data, interval=1)

```
import statistics
data=[1,3,3,5,7]
s=statistics.median_grouped(data,1)
print(s)
```

تابع mode()

- داده ای که بیشترین تکرار(فراوانی) را دارد برمیگرداند.

```
import statistics
data=[0,0,3,4,5,6]
s=statistics.mode(data)
print(s)
-----------------------------------
import statistics
data=[1,0,3,4,1,5,6,0]
s=statistics.mode(data)
print(s)
```

```
import statistics
data=["red", "blue", "blue", "red", "green", "red", "red"]
s1=statistics.mode(data)
s2=statistics.median(data)
print(s1,s2)
```

برای محاسبه میانه، رشته ها را براساس طول هر رشته در لیست مرتبط می کند و بعد رشته میانه را تعیین می کند.

تابع multimode(data)

- لیستی از داده های پر تکرار (فراوانی بالاتر) را برمی گرداند

```
import statistics
data=["red", "blue", "blue", "red", "green", "red", "red"]
s=statistics.multimode(data)
print(s)
```

```
import statistics
data=[-1,0,0,2,3,4,-1]
s=statistics.multimode(data)
print(s)
```

انحراف معیار (standard deviation)

- **انحراف معیار** با نماد σ یکی از شاخص های پراکندگی است که نشان می‌دهد به‌طور میانگین داده‌ها چه مقدار از مقدار متوسط فاصله دارند. اگر انحراف معیار مجموعه‌ای از داده‌ها نزدیک به صفر باشد، نشانه آن است که داده‌ها نزدیک به میانگین هستند و پراکندگی اندکی دارند؛ در حالی که انحراف معیار بزرگ بیانگر پراکندگی قابل توجه داده‌ها می‌باشد. انحراف معیار برابر ریشه دوم واریانس است. خوبی آن نسبت به واریانس، این است که هم بعد با داده‌ها می‌باشد.
- انحراف معیار برای تعیین ضریب اطمینان در تحلیل‌های آماری نیز به کار می‌رود. در مطالعات علمی، معمولاً داده‌های با اختلاف بیشتر از دو انحراف معیار از مقدار میانگین به عنوان داده‌های پرت در نظر گرفته و از تحلیل، خارج می‌شوند.

انحراف معیار جمعیت

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

در پایتون *انحراف معیار جمیعت با تابع* pstdev() محاسبه می شود.

انحراف معیار نمونه

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

در پایتون *انحراف معیار جمیعت با تابع* stdev() محاسبه می شود.

```
import statistics
data=[1,2,3]
s1=statistics.mean(data)
print(s1)
s2=statistics.stdev(data)
print(s2)
s3=statistics.pstdev(data)
print(s3)
```

واریانس (Variance)

- **وردایی** یا **واریانس** درنظریه احتمال و آمار ، نوعی سنجش پراکندگی است.
- مقدار واریانس با میانگین‌گیری از مربع فاصله مقدار محتمل یا مشاهده شده با مقدار مورد انتظار محاسبه می‌شود. در مقایسه با میانگین می‌توان گفت که میانگین مکان توزیع را نشان می‌دهد، در حالی که واریانس مقیاسی است که نشان می‌دهد که داده‌ها حول میانگین چگونه پخش شده‌اند. واریانس کمتر بدین معنا است که انتظار می‌رود که اگر نمونه‌ای از توزیع مزبور انتخاب شود مقدار آن به میانگین نزدیک باشد. یکای واریانس مربع یکای کمیت اولیه می‌باشد. ریشه دوم واریانس که انحراف معیار نامیده می‌شود دارای واحدی یکسان با متغیر اولیه است.

واریانس جمعیت

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{var} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

در پایتون واریانس جمعیت با تابع pvariance() بدست می آید.

واریانس نمونه

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\text{var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

در پایتون واریانس جمعیت با تابع variance() بدست می آید.

```
import statistics
data=[1,2,3]
s1=statistics.mean(data)
print(s1)
s2=statistics.variance(data)
print(s2)
s3=statistics.pvariance(data)
print(s3)
```

- نحوه مختلف استفاده از ماژول

- Import statistics
- Import statistics as ss
- From statistics import mean,variance,…
- From statistics import *

مثال

```
import statistics
example_list = [5,2,5,6,1,2,6,7,2,6,3,5,5]
x = statistics.mean(example_list)
print("mean=",x)
y = statistics.median(example_list)
print("median=",y)
z = statistics.mode(example_list)
print("mode=",z)
a = statistics.stdev(example_list)
print("stdev=",a)
b = statistics.variance(example_list)
print("variance=",b)
```

```
import statistics as ss
example_list = [5,2,5,6,1,2,6,7,2,6,3,5,5]
x = ss.mean(example_list)
print("mean=",x)
y = ss.median(example_list)
print("median=",y)
z = ss.mode(example_list)
print("mode=",z)
a = ss.stdev(example_list)
print("stdev=",a)
b = ss.variance(example_list)
print("variance=",b)
```

```
from statistics import mean,median,mode,stdev,variance
example_list = [5,2,5,6,1,2,6,7,2,6,3,5,5]
x = mean(example_list)
print("mean=",x)
y = median(example_list)
print("median=",y)
z = mode(example_list)
print("mode=",z)
a = stdev(example_list)
print("stdev=",a)
b = variance(example_list)
print("variance=",b)
```

```
from statistics import mean,median,mode,stdev,variance
example_list = [5,2,5,6,1,2,6,7,2,6,3,5,5]
x = mean(example_list)
print("mean=",x)
y = median(example_list)
print("median=",y)
z = mode(example_list)
print("mode=",z)
a = stdev(example_list)
print("stdev=",a)
b = variance(example_list)
print("variance=",b)
```

```
import statistics as ss
def main():
  data=[]
  n=int(input("n="))
  for i in range(n):
    x=float(input("x="))
    data=data+[x]
  print("data=",data)
  s1=ss.mean(data)
  print("mean=",s1)
  s2=ss.median(data)
  print("median=",s2)
  s3=ss.multimode(data)
  print("mode=",s3)
  s4=ss.variance(data)
  print("variance=",s4)
  s5=ss.stdev(data)
  print("stdev=",s5)
  restart=input("do you want to .....?(y/n)")
  if restart=="y" or restart=="Y":
        main()
restart=input("do you want to .....?(y/n)")
if restart=="y" or restart=="Y":
   main()
```

فراخوانی لیست توابع موجود در ماژول

```
import statistics
print(dir(statistics))
```

راهنمایی و تعریف توابع یک ماژول

```
import statistics
help(statistics.multimode)
```

ماژول های داخلی پایتون

- Math
  - شامل توابع ریاضی مختلف
- Re
  - شامل الگوهای مختلف
- Random
  - توابع مختلف تولید اعداد رندوم ،انتخاب رندوم و توزیع های احتمال مختلف
- Statistics
  - انجام آنالیز اماری بر روی داده ها شامل میانگین، میانه،مد، انحراف معیار، واریانس

ماژول های نصبی

کتابخانه numpy

- Numpyمخفف Numerical Python به معنای پایتون عددی یا پایتون محاسباتی
- علم داده، هوش مصنوعی، ماشین لزنینگ
- آرایه ها و ماتریس ها، جبر خطی و حل معادلات خطی، رگرسیون، تبدیل فوریه، اعداد تصادفی، توابع آماری (کمینه، بیشینه، میانگین، میانه، چارک،انحراف معیار، واریانس)
- پایه بسیاری از کتاب خانه های دیگر است.
- آرایه های Numpyمحاسباتی سریع تر از لیست ها دارند و برای اجرای عملیات ریاضیاتی و منطقی بسیار کارآمدتر هستند.

کتابخانه pandas

- پانداس یک کتابخانه قدرتمند برای تجزیه و تحلیل داده‌ها، پیش‌پردازش (PreProcessing) و بصری‌سازی (Visualization) داده‌ها

- پیش‌پردازش داده‌ها مانند ادغام کردن، گروه‌بندی ،الحاق، تمیز کاری
- سری‌های زمانی

- هوش مصنوعی، یادگیری ماشین، یادگیری عمیق ، علم داده

کتابخانه scipy

- انجام محاسبات علمی و مهندسی، تحلیل ها و آنالیزهای یادگیری ماشینی و هوش مصنوعی
- آرایه ها و ماتریس، خوشه بندی داده ها،تبدیل فوریه (پردازش سیگنال و نویز، پردازش تصویر، پردازش سیگنال صوتی)،جبر خطی، پردازش تصویرو سیگنال،الگوریتم های بهینه سازی، درون یابی، حل کننده های معادلات دیفرانسیل معمولی ،

کتابخانه matplotpy

- رسم نمودار و مصور سازی
- رسم خطوط، اشکال دو بعدی، نمودار نقطه ای،ترسیم با رشته ها ، نمودار میله ای،نمودار هیستوگرام،رسم پراکندگی، نمودار دایره ای، نمودار توابع، نمودارهای سه بعدی، کانتور،پوسته های سه بعدی،

کتابخانه seaborn

- رسم انواع چارت هایی چون نمودارهای ماتریسی، نمودارهای شبکه ای (Grid)، نمودارهای رگرسیونی و غیره

کتابخانه pygame

- ساخت بازی

python –V

ورژن پایتون نصبی را نشان می دهد.

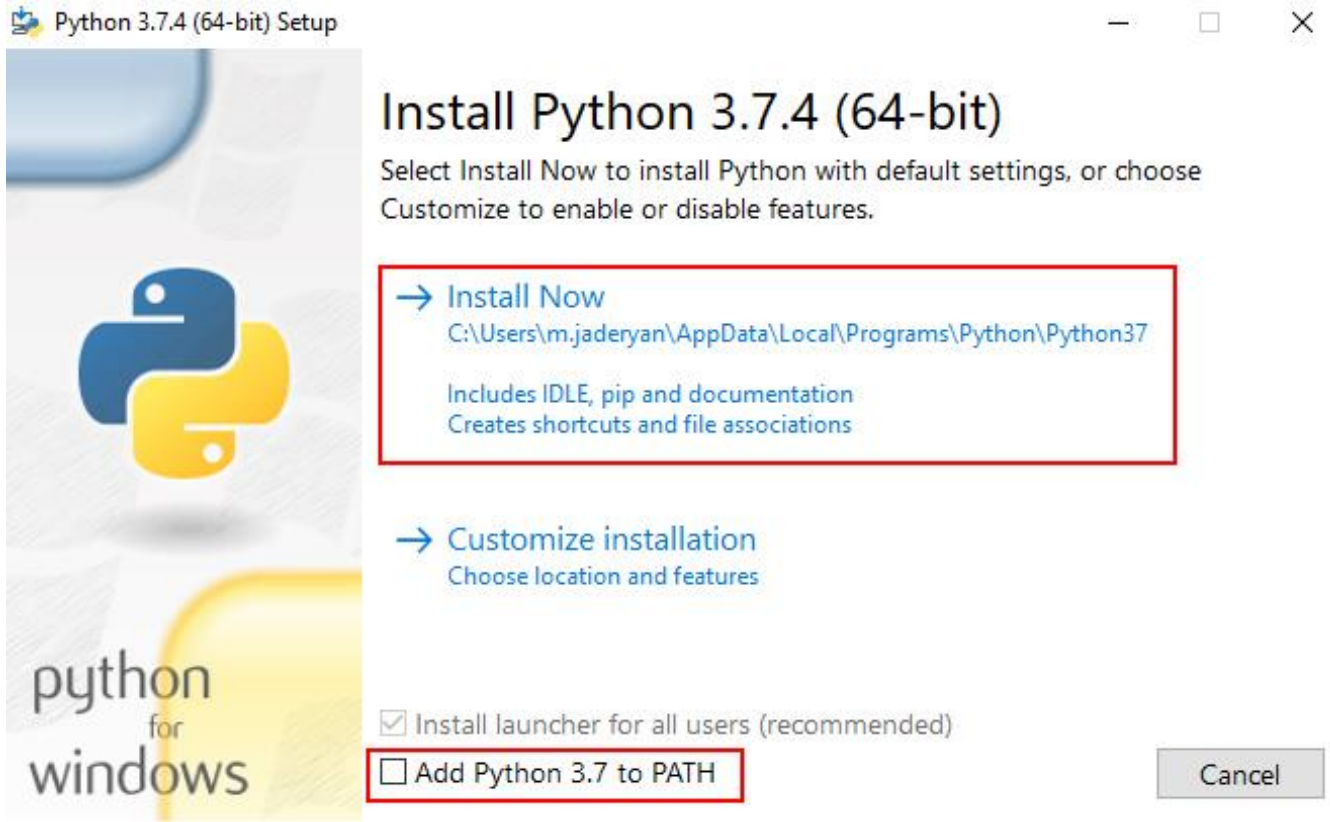
pip -V

محل نصب pip را نشان می دهد.

اگر به جای نشان دادن ورژن وادرس با ارور زیر برخوردید

'pip' is not recognized as an internal or external command,
operable program or batch file.

- اگر با ارور اسلاید قبل برخورد به این معنا هست که در مرحله نصب پایتون تیک مربوط به path را نزدید. بنابرین پایتون را لغو نصب کنید و مجدد نصب کنید و این بار تیک مربوطه را بزنید.



نصب پکیج های(ماژول/کتابخانه های پایتون)

در ویندوز

pip install packagename

در مک/لینوکس

sudo pip install packagename

چک کردن ماژول های نصب شده

```
pip freeze
```

فرخوانی اطلاعات کامل ماژول ها

```
pip show packagename
```

به روزرسانی ماژول ها

ویندوز

```
pip install --upgrade packagename
```

مک و لینوکس

```
sudo pip install --upgrade packagename
```

حذف و لغو نصب ماژول

ویندوز

```
pip uninstall packagename
```

مک و لینوکس

```
sudo pip uninstall packagename
```

نصب numpy

- در محیط خط فرمان (cmd) دستور زیر را تایپ کنید:

```
pip install numpy
```

فرخوانی ماژول

```
import numpy
import numpy as np
```

ورژن ماژول

```
import numpy as np
print(np.__version__)
```

تمرین

- کاربرد های میانگین هندسی و هارمونیک را بیان کنید.

تمرین

- کمیت های میانگین حسابی، میانگین هندسی، میانگین هارمونیک، میانه، واریانس، انحراف معیار داده های زیر را تعیین کنید.

Data1=[0,-1,3,4,3,4,0,5,8,9,-4,0,4,5]

Data2=[456.7,547.8,926.6,236.1,543,439]