L3 Statistics Basics

By the end of this practical you will be able to perform basic statistics operation using Python

Descriptive Statistics

Recall from the lecture notes that we have discussed the following mean (average), median, standard deviation, skewness and kurtosis. We will use the python to do the implementation.

Mean

The mean is the numerical average of the entire data set.

$$nums = \{872, 432, 397, 427, 388, 782, 397\}$$

$$mean = \Sigma nums / |nums|$$

```
def mean(l):
    return sum(l) / len(l)
nums = [872, 432, 397, 427, 388, 782, 397]
print(mean(nums))
```

```
527.8571428571429
```

Median

Median is the center (position) value in the ordered list.

$$nums = \{872, 432, 397, 427, 388, 782, 397\} = \{388, 397, 397, \underline{427}, 432, 782, 872\}$$

$$median = 427$$

If the number of values in the data set is even, we take the average of the two center value.

```
from math import * # a library that provides math functions
def median(l):
    if (len(l) % 2 == 1):
        return sorted(l)[int(floor(len(l)/2))]
    else:
        i1 = int(floor(len(l)/2-1))
        i2 = int(floor(len(l)/2))
        sl = sorted(l)
        return (sl[i1]+sl[i2] ) / 2
        return (sl[i1]+sl[i2] ) / 2

nums = [872, 432, 397, 427, 388, 782, 397]
print(median(nums))

age=[21,20,21,23,23,19,30,60]
print(median(age))
```

```
427
22.0
```

Mode

Mode - the most frequent observation. If there is no repetition, no mode exists.

$$nums = \{872, \underline{432, 432, 432,} 388, 782, 388\}$$

$$mode = 432$$

```
def mode(l): # assuming l is non empty
    d = {}
    for x in l:
        if (x in d):
            d[x] +=1
        else:
            d[x] = 1
    print(d)
    print(d.items())

    return [g for g,l in d.items() if l==max(d.values())]

nums = [872, 432, 397, 427, 388, 782, 397]
print(mode(nums))
```

```
{872: 1, 432: 1, 397: 2, 427: 1, 388: 1, 782: 1}
dict_items([(872, 1), (432, 1), (397, 2), (427, 1), (388, 1), (782, 1)])
[397]
```

Variance

The variance measures how far each value in the data set is from the mean.

Let x denote the data set, n be the size of the data, $\bar{x}$ denote the mean, variance σ^2 is defined as

$$\sigma^2 = \frac{\Sigma(x - \bar{x})^2}{n}$$

```
def variance(l): # assuming l is non empty
    m = mean(l)
    diffsqsum = sum(map(lambda x:(x - m)**2, l))
    return diffsqsum/len(l)

nums = [872, 432, 397, 427, 388, 782, 397]
```

```
print(variance(nums))
```

```
    36598.69387755102
```

Standard deviation

The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. σ is the standard deviation Let x denote the data set, n be the size of the data, $\bar{x}$ denote the mean, standard deviation σ is defined as

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}$$

```
def std(l): # assuming l is non empty
    return sqrt(variance(l))

nums = [872, 432, 397, 427, 388, 782, 397]
print(std(nums))
```

```
    191.30785106093012
```

Skewness

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.skew.html

```
import pandas as pd

dataVal = [(10,20,30,40,50,60,70),

           (10,10,40,40,50,60,70),

           (10,20,30,50,50,60,80)]

dataFrame = pd.DataFrame(data=dataVal);

skewValue = dataFrame.skew(axis=1) #by row

print("DataFrame:")

print(dataFrame)

print("Skew:")

print(skewValue)
```

```
    DataFrame:
        0   1   2   3   4   5   6
```

```
0  10  20  30  40  50  60  70
1  10  10  40  40  50  60  70
2  10  20  30  50  50  60  80
Skew:
0    0.000000
1   -0.340998
2    0.121467
dtype: float64
```

kurtosis

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.kurtosis.html

[ ] ↳ *2 cells hidden*

Correlation

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html

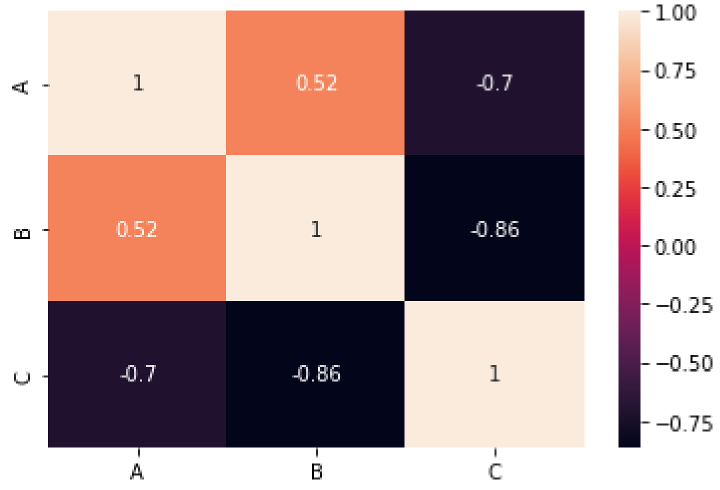
```
from pandas import DataFrame
import seaborn as sn

Data = {'A': [45,37,42,35,39],
        'B': [38,31,26,28,33],
        'C': [10,15,17,21,12]
        }

df = DataFrame(Data,columns=['A','B','C'])

corrMatrix = df.corr()
sn.heatmap(corrMatrix, annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0cacd1c450>
```



● ✕