Lecture16

March 28, 2024

[6]:
```
# ! pip install scipy
```

[7]:
```
from scipy import stats
```

0.0.1 Statistical Functions

[8]:
```
# stats.describe()
```

[27]:
```
# stats.mode(data)
# stats.skewTest(data)
# stats.kurtosisTest(data)
# stats.normalTest(data)
```

[9]:
```
values=[2,3,4,5,3,3,4,4,6]
```

[11]:
```
stats.describe(values)
```

[11]:
```
DescribeResult(nobs=9, minmax=(2, 6), mean=3.7777777777777777,
variance=1.4444444444444446, skewness=0.44314703612955064,
kurtosis=-0.45821005917159763)
```

0.0.2 Distributions

[13]:
```
# stats.norm(mean, variance)
my_norm=stats.norm(70000,10000)
my_norm
```

[13]:
```
<scipy.stats._distn_infrastructure.rv_continuous_frozen at 0x7f62c28a1ed0>
```

[14]:
```
my_norm.rvs(10)
```

[14]:
```
array([75419.8150853 , 66979.20858971, 66576.52700229, 82052.06614383,
       65618.01231255, 74014.35562598, 80715.8357428 , 76428.21793288,
       64725.86858689, 80848.23105858])
```

[15]:
```
my_norm.cdf(80000)
```

[15]:
```
0.8413447460685429
```

[16]:
```
my_norm.cdf(80000) - my_norm.cdf(60000)
```

[16]:
```
0.6826894921370859
```

[19]:
```
my_norm.pdf(700000)
```

[19]:
```
0.0
```

[20]:
```
# stats.binom(n_trails, probability)
my_binom=stats.binom(6,.5)
```

[21]:
```
my_binom.rvs(10)
```

[21]:
```
array([4, 1, 2, 3, 4, 4, 3, 3, 3, 1])
```

[22]:
```
my_binom.pmf(3)
```

[22]:
```
0.31249999999999983
```

0.0.3 Hypothesis Tests

[23]:
```
# One sample t-test
# stats.ttest_1samp(samples, mean)
```

[24]:
```
# Two samples t-test
# stats.ttest_ind(samples1, samples2)    #H0: u1 = u2
```

[25]:
```
# Anova Test:
# stats.f_oneway(samples1, samples3, samples3)   #H0: u1=u2=u3
```

[ ]:
```
# stats.chisquare
# stats.chi2_contingency
```

1 Pandas

[31]:
```
# !pip install pandas
```

[30]:
```
import pandas as pd
```

1.0.1 Series

[42]:
```
s=pd.Series([100, 500, 'hasan'], index=[555,44,900])
```

[43]:
```
s
```

[43]:
```
555      100
44       500
900    hasan
dtype: object
```

[41]:
```
s[44]
```

[41]:
```
500
```

[44]:
```
s.values
```

[44]:
```
array([100, 500, 'hasan'], dtype=object)
```

[45]:
```
s.index
```

[45]:
```
Index([555, 44, 900], dtype='int64')
```

1.0.2 DataFrame

[49]:
```
# Create   (list of lists) or (dict)
df=pd.DataFrame({'name':['hasan','alma','hala','shahd'],
                 'age':[40,18,16,8],
                 'salary':[2000,4000,5000,3000]})
df
```

[49]:

	name	age	salary
0	hasan	40	2000
1	alma	18	4000
2	hala	16	5000
3	shahd	8	3000

[80]:
```
pd.DataFrame([['hasan', 40, 2000],
              ['alma',18,4000],
              ['hala',16,5000],
              ['shahd',8,3000]], columns=['name','age', 'salary'])
```

[80]:

	name	age	salary
0	hasan	40	2000
1	alma	18	4000
2	hala	16	5000
3	shahd	8	3000

[55]:
```
cat datafiles/Employee.csv
```

```
,Name,Year,Department
0,Bob,1,IT
1,Sam,3,Trade
2,Peter,8,HR
3,Jake,2,IT
```

[60]:
```
# read Data from from file

pd.read_csv('datafiles/Employee.csv', index_col=0 )
```

[60]:

	Name	Year	Department
0	Bob	1	IT
1	Sam	3	Trade
2	Peter	8	HR
3	Jake	2	IT

[61]:
```
df
```

[61]:

	name	age	salary
0	hasan	40	2000
1	alma	18	4000
2	hala	16	5000
3	shahd	8	3000

[63]:
```
df.index
```

[63]:
```
RangeIndex(start=0, stop=4, step=1)
```

[64]:
```
df.columns
```

[64]:
```
Index(['name', 'age', 'salary'], dtype='object')
```

[65]:
```
df.values
```

[65]:
```
array([['hasan', 40, 2000],
       ['alma', 18, 4000],
       ['hala', 16, 5000],
       ['shahd', 8, 3000]], dtype=object)
```

[66]:
```
## Prepare
# set_index
# reset_index
# del df.[col_name], df.drop(col_name, axis=1)
```

[82]:
```
df=df.set_index('name')
```

[83]:
```
df
```

[83]:
```
        level_0  index  age  salary
name
hasan         0      0   40    2000
alma          1      1   18    4000
hala          2      2   16    5000
shahd         3      3    8    3000
```

[74]:
```
df.reset_index()
```

[74]:
```
    name  age  salary
0  hasan   40    2000
1   alma   18    4000
2   hala   16    5000
3  shahd    8    3000
```

[84]:
```
df=df.reset_index()
df
```

[84]:
```
    name  level_0  index  age  salary
0  hasan        0      0   40    2000
1   alma        1      1   18    4000
2   hala        2      2   16    5000
3  shahd        3      3    8    3000
```

[85]:
```
del df['index']
```

[86]:
```
df
```

[86]:
```
    name  level_0  age  salary
0  hasan        0   40    2000
1   alma        1   18    4000
2   hala        2   16    5000
3  shahd        3    8    3000
```

[94]:
```
df=df.drop('level_0', axis=1)
```

[95]:
```
df.columns=['first_name','age','Total Salary']
df
```

[95]:
```
  first_name  age  Total Salary
0      hasan   40          2000
1       alma   18          4000
2       hala   16          5000
3      shahd    8          3000
```

[96]:
```
df.columns=df.columns.str.replace('age', 'AGE')
```

[97]:
```
df
```

[97]:

	first_name	AGE	Total Salary
0	hasan	40	2000
1	alma	18	4000
2	hala	16	5000
3	shahd	8	3000

[ ]:
```

```