9과 [예제] 범주 vs 범주

1.환경준비

- 라이브러리 불러오기

In [ ]:

```
import pandas as pd
import numpy as np
import random as rd

import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.graphics.mosaicplot import mosaic      #mosaic plot!

import scipy.stats as spst
```

```
/usr/local/lib/python3.7/dist-packages/statsmodels/tools/_testing.py:19: Fut
ureWarning: pandas.util.testing is deprecated. Use the functions in the publ
ic API at pandas.testing instead.
  import pandas.util.testing as tm
```

- 데이터 불러오기 : 다음의 예제 데이터를 사용합니다.

> ① 타이타닉 생존자
> ② 보스톤 시, 타운별 집값
> ③ 아이리스 꽃 분류
> ④ 뉴욕 공기 오염도

In [ ]:

```
# 타이타닉 데이터
titanic = pd.read_csv('https://raw.githubusercontent.com/DA4BAM/dataset/master/titanic.
titanic.head()
```

Out[16]:

	PassengerId	Survived	Pclass	Title	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin
0	1	0	3	Mr	male	22.0	1	0	A/5 21171	7.2500	NaN
1	2	1	1	Mrs	female	38.0	1	0	PC 17599	71.2833	C85
2	3	1	3	Miss	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN
3	4	1	1	Mrs	female	35.0	1	0	113803	53.1000	C123
4	5	0	3	Mr	male	35.0	0	0	373450	8.0500	NaN

In [ ]:

```
# 아이리스 꽃 분류
iris = pd.read_csv('https://raw.githubusercontent.com/DA4BAM/dataset/master/iris.csv')
iris.head()
```

Out[3]:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

In [ ]:

```
# 보스톤 집값 데이터
boston = pd.read_csv('https://raw.githubusercontent.com/DA4BAM/dataset/master/boston2_N
boston.head()
```

Out[4]:

	crim	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv	zn2
0	0.00632	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0	1.(
1	0.02731	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6	0.(
2	0.02729	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7	0.(
3	0.03237	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4	0.(
4	0.06905	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2	0.(

In [ ]:

```
# 뉴욕시 공기 오염도 데이터
air = pd.read_csv('https://raw.githubusercontent.com/DA4BAM/dataset/master/air2.csv')
air['Date'] = pd.to_datetime(air['Date'])
air['Month'] = air.Date.dt.month
air['Weekday'] = air.Date.dt.weekday
air.head()
```

Out[5]:

	Ozone	Solar.R	Wind	Temp	Date	Month	Weekday
0	41	190.0	7.4	67	1973-05-01	5	1
1	36	118.0	8.0	72	1973-05-02	5	2
2	12	149.0	12.6	74	1973-05-03	5	3
3	18	313.0	11.5	62	1973-05-04	5	4
4	19	NaN	14.3	56	1973-05-05	5	5

2.교차표(pd.crosstab)

교차표를 연습해 봅시다.

① 타이타닉의 성별에 따른 생존여부의 관계를 교차표로 만들어 봅시다.

- 전체 갯수

In [ ]:

```
# 두 범주별 빈도수를 교차표로 만들어 봅시다.
pd.crosstab(titanic['Survived'], titanic['Sex'])
```

Out[6]:

Sex	female	male
Survived		
0	81	468
1	233	109

- 칼럼기준 비율

In [ ]:

```
pd.crosstab(titanic['Survived'], titanic['Sex'], normalize = 'columns')
```

Out[7]:

Sex	female	male
Survived		
0	0.257962	0.811092
1	0.742038	0.188908

- 행 기준 비율

In [ ]:

```
pd.crosstab(titanic['Survived'], titanic['Sex'], normalize = 'index')
```

Out[8]:

Sex	female	male
Survived		
0	0.147541	0.852459
1	0.681287	0.318713

- 전체 기준 비율

In [ ]:

```
1 pd.crosstab(titanic['Survived'], titanic['Sex'], normalize = 'all')
```

Out[9]:

Sex	female	male
Survived		
0	0.090909	0.525253
1	0.261504	0.122334

- 교차표를 통해 성별에 따라 생존여부가 관련 있다고 보이나요?
- 위 교차표 중 어떤 것이 성별-->생존여부 관련성을 확인하기에 적합한가요?

In [ ]:

```
1
```

② 타이타닉의 객실등급에 따른 생존여부의 관계를 교차표로 만들어 봅시다.

- 전체 갯수

In [ ]:

```
1 # 두 범주별 빈도수를 교차표로 만들어 봅시다.
2 pd.crosstab(titanic['Survived'], titanic['Pclass'])
```

Out[10]:

Pclass	1	2	3
Survived			
0	80	97	372
1	136	87	119

- 칼럼기준 비율

In [ ]:

```
1 pd.crosstab(titanic['Survived'], titanic['Pclass'], normalize = 'columns')
```

Out[11]:

Pclass	1	2	3
Survived			
0	0.37037	0.527174	0.757637
1	0.62963	0.472826	0.242363

- 행 기준 비율

In [ ]:

```
1 pd.crosstab(titanic['Survived'], titanic['Pclass'], normalize = 'index')
```

Out[12]:

Pclass	1	2	3
Survived			
0	0.145719	0.176685	0.677596
1	0.397661	0.254386	0.347953

- 전체 기준 비율

In [ ]:

```
1 pd.crosstab(titanic['Survived'], titanic['Pclass'], normalize = 'all')
```

Out[13]:

Pclass	1	2	3
Survived			
0	0.089787	0.108866	0.417508
1	0.152637	0.097643	0.133558

- 교차표를 통해 객실 등급에 따라 생존여부가 관련 있다고 보이나요?
- 위 교차표 중 어떤 것이 성별-->생존여부 관련성을 확인하기에 적합한가요?

In [ ]:

```
1
```

③ 성별과 객실등급 중 어떤 변수가 생존여부를 예측하는데 더 중요한 변수인가요?

In [ ]:

```
1
```

3.시각화 : bar chart, mosaic

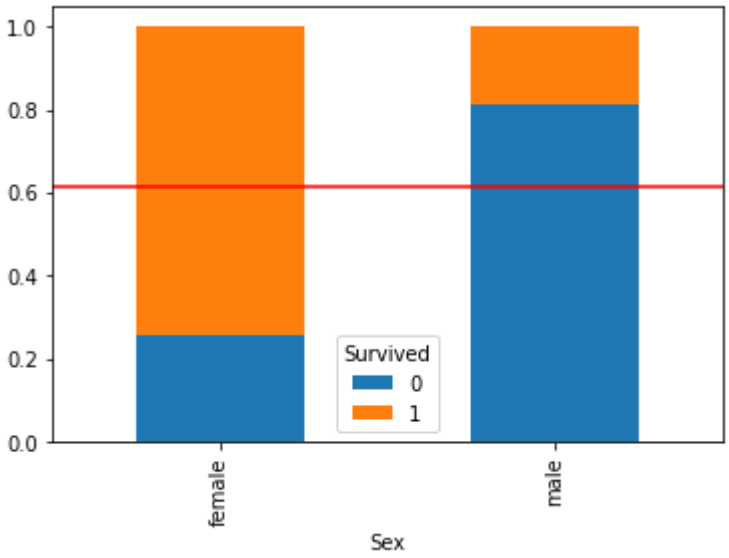
① Sex --> Survived

- 100% bar chart을 그려봅시다.

In [ ]:

```
temp = pd.crosstab(titanic['Sex'], titanic['Survived'], normalize = 'index')
print(temp)
temp.plot.bar(stacked=True)
plt.axhline(1-titanic['Survived'].mean(), color = 'r')
plt.show()
```

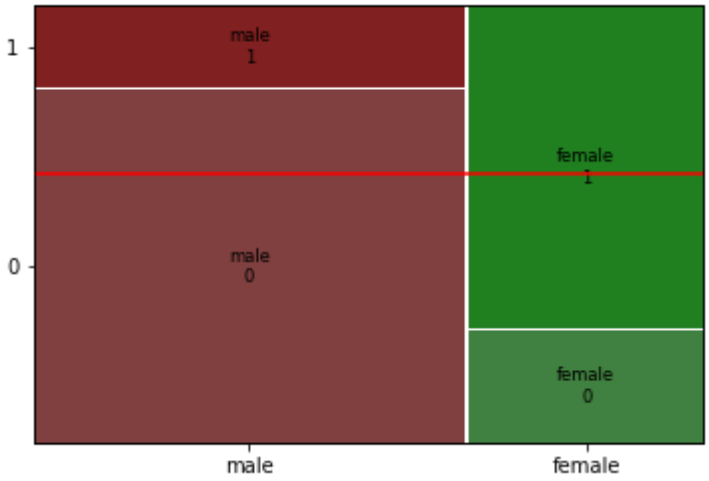
```
Survived         0         1
Sex
female    0.257962  0.742038
male      0.811092  0.188908
```



- 모자익 플롯을 그려봅시다.

In [ ]:

```
# Pclass별 생존여부를 mosaic plot으로 그려 봅시다.
mosaic(titanic, [ 'Sex','Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```



- 두 차트로 볼 때, 성별에 따라 생존여부가 달라지나요?

In [ ]:

```

```

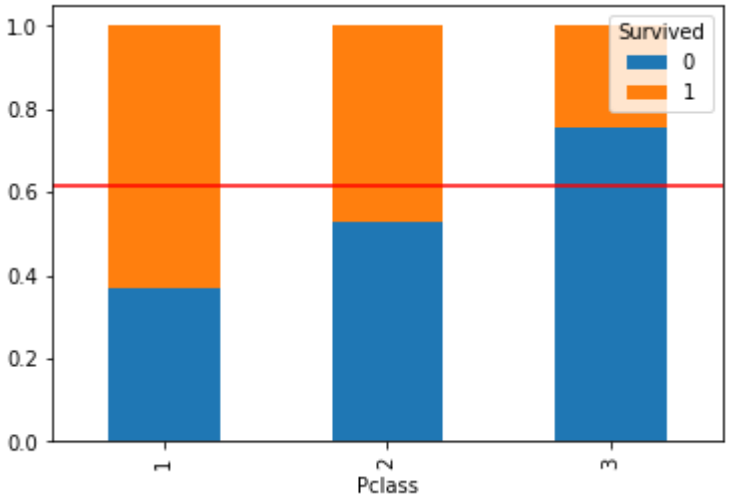
② Pclass --> Survived

- 100% bar chart을 그려봅시다.

In [ ]:

```
temp = pd.crosstab(titanic['Pclass'], titanic['Survived'], normalize = 'index')
print(temp)
temp.plot.bar(stacked=True)
plt.axhline(1-titanic['Survived'].mean(), color = 'r')
plt.show()
```

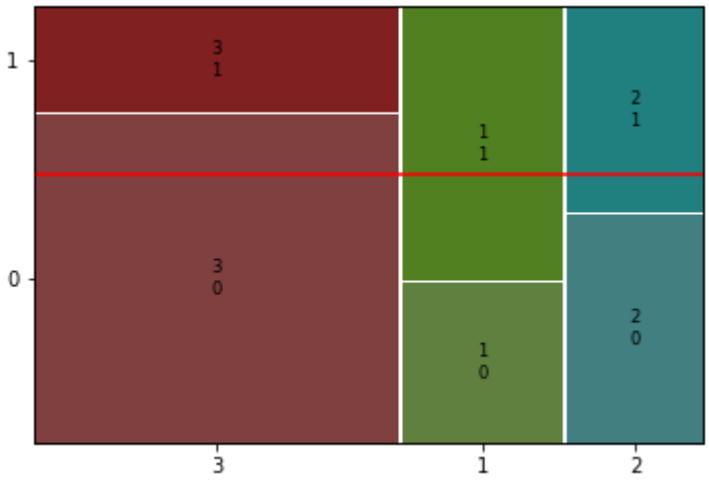
```
Survived         0         1
Pclass
1         0.370370  0.629630
2         0.527174  0.472826
3         0.757637  0.242363
```



- 모자익 플롯을 그려봅시다.

In [ ]:

```
# Pclass별 생존여부를 mosaic plot으로 그려 봅시다.
mosaic(titanic, [ 'Pclass','Survived'])
plt.axhline(1- titanic['Survived'].mean(), color = 'r')
plt.show()
```



- 두 차트로 볼 때, 성별에 따라 생존여부가 달라지나요?

In [ ]:

```
1
```

4.수치화 : 카이제곱검정

① Sex --> Survived

In [ ]:

```
# 먼저 집계
table = pd.crosstab(titanic['Survived'], titanic['Sex'])
print('교차표\n', table)
print('-' * 100)

# 카이제곱검정
result = spst.chi2_contingency(table)
print('카이제곱통계량', result[0])
print('p-value', result[1])
print('기대빈도\n',result[3])
```

② Pclass --> Survived

In [ ]:

```
# 먼저 집계
table = pd.crosstab(titanic['Survived'], titanic['Pclass'])
print('교차표\n', table)
print('-' * 100)

# 카이제곱검정
result = spst.chi2_contingency(table)
print('카이제곱통계량', result[0])
print('p-value', result[1])
print('기대빈도\n',result[3])
```