[파이썬 통계분석 ; 1차-과제물]

2018015027 정보통계학과 김한탁

1. 대사증후군 자료(엑셀)을 불러와 범주형 자료인데 숫자로 코딩된 자료에 대해 범주형(object)으로 변환하여 데이터프레임(desa)를 만들어라.(변수가 많은 관계로 4개 변수 ex1ex2co, exco, 건강(1~5), 수입(1,4)에 대해서만 범주형으로 변경, 단, 건강 : 1=50만원 미만, 2=50-100만, 3=100-200만, 4=200만이상)을 의미함)

풀이)
문제에 사용되는 기본적인 모듈과 한글폰트를 이용하기 위한 파일이다.
1번뿐만 아니라 나머지 문제에서도 사용되어 풀이에 반복적으로 등장하기 때문에, 1번 풀이에 대표로 코드를 삽입할 것이다.

In [1]:
```
!python han-font.py
exec(open('han-font.py').read())

import pandas as pd
import numpy as np
import matplotlib as plt
import seaborn as sns
```

다음은 기존 대사증후군에서 각 변수의 자료형을 보여준다. (0, 1, 27, 33번 확인)

In [2]:
```
#1
desa=pd.read_csv('대사증후군.csv', encoding = 'cp949')
desa.head(3)
desa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91 entries, 0 to 90
Data columns (total 47 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ex1ex2co  91 non-null     int64
 1   exco      91 non-null     int64
                   ⋮
 27  건강        88 non-null     float64
 28  결혼        89 non-null     float64
 29  동거        90 non-null     float64
 30  종교        90 non-null     float64
 31  직업        86 non-null     float64
 32  학력        90 non-null     float64
 33  수입        87 non-null     float64
                   ⋮
```

이후 astype()을 이용하여, ex1ex2co, exco, 건강, 수입 변수의 자료형이 범주형(object)으로 바뀌었음을 확인할 수 있다.

```
In [3]: desa['ex1ex2co'] = desa['ex1ex2co'].astype(object)
        desa['exco'] = desa['exco'].astype(object)
        desa['건강'] = desa['건강'].astype(object)
        desa['수입'] = desa['수입'].astype(object)
        desa.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 91 entries, 0 to 90
Data columns (total 47 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   ex1ex2co  91 non-null     object
 1   exco      91 non-null     object
                 ⋮
 27  건강        88 non-null     object
 28  결혼        89 non-null     float64
 29  동거        90 non-null     float64
 30  종교        90 non-null     float64
 31  직업        86 non-null     float64
 32  학력        90 non-null     float64
 33  수입        87 non-null     object
                 ⋮
```

2. 변수 exco(실험군-1, 대조군2)에 따른 대사중후군 수치(wc1, wc2, bmi1, bmi2)에 대한 평균값을 정리하라.

풀이)
다음은 groupby()를 이용하여, 변수 exco(실험군-1, 대조군2)에 따른 대사중후군 수치(wc1, wc2, bmi1, bmi2)에 대한 평균값을 정리한 것이다.

```
In [4]: #2
        desa.groupby(['exco']).mean()[['wc1', 'wc2', 'bmi1', 'bmi2']]
```

Out[4]:

exco	wc1	wc2	bmi1	bmi2
1	93.502817	89.401408	26.51831	25.56338
2	94.885000	95.200000	27.14000	26.93000

exco 1(실험군-1)의 wc1, wc2, bmi1, bmi2의 평균값이 각각 93.5, 89.4, 26.5, 25.6이고, exco 2(대조군-2)의 wc1, wc2, bmi1, bmi2의 평균값은 각각 94.9, 95.2, 27.1, 26.9이다.

3. 연령(age)자료를 범주형자료(50대미만, 50대, 60대, 70대이상)로 바꾸어 새로운 변수 연령대(agegroup)를 생성하고, 범주형 변수(건강)에 (1-건강매우좋음, 2- 건강좋음, 3-보통, 4-건강나쁨, 5-건강매우나쁨)로 숫자의 의미부여를 하여 guengang 변수를 생성하라.

풀이)
다음은 변수 age를 조건에 따라 범주형으로 나누어 agegroup(연령대)변수를 생성한 것이다. age와 agggroup의 자료를 비교하였을 때, age의 자료가 범주에 의해 잘 분류되었다고 볼 수 있다.

```
In [5]: desa.loc[desa['age']<50, 'agegroup']='50대미만'
        desa.loc[(desa['age']>=50) & (desa['age']<60), 'agegroup']='50대'
        desa.loc[(desa['age']>=60) & (desa['age']<70), 'agegroup']='60대'
        desa.loc[desa['age']>=70, 'agegroup']='70대이상'
        desa.loc[:, ['age', 'agegroup']]
```

Out[5]:

	age	agegroup
0	69.0	60대
1	63.0	60대
2	62.0	60대
3	59.0	50대
4	NaN	NaN
...	...	...
86	76.0	70대이상
87	41.0	50대미만
88	65.0	60대
89	71.0	70대이상
90	76.0	70대이상

91 rows × 2 columns

다음은 변수 '건강'을 조건에 따라 범주형으로 나누어 guengang 변수를 생성한 것이다. '건강'과 guengang의 자료를 비교하였을 때, '건강'의 자료가 범주에 의해 잘 분류되었다고 볼 수 있다.

```
In [6]: desa.loc[desa['건강']==1, 'guengang']='건강매우좋음'
        desa.loc[desa['건강']==2, 'guengang']='건강좋음'
        desa.loc[desa['건강']==3, 'guengang']='보통'
        desa.loc[desa['건강']==4, 'guengang']='건강나쁨'
        desa.loc[desa['건강']==5, 'guengang']='건강매우나쁨'
        desa.loc[:, ['건강', 'guengang']]
```

Out[6]:

	건강	guengang
0	5.0	건강매우나쁨

1	3.0	보통
2	2.0	건강좋음
3	NaN	NaN
4	NaN	NaN
...	...	...
86	2.0	건강좋음
87	2.0	건강좋음
88	2.0	건강좋음
89	5.0	건강매우나쁨
90	5.0	건강매우나쁨

91 rows × 2 columns

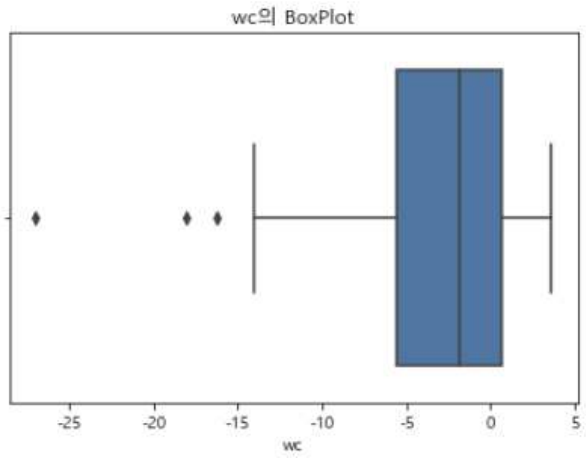
4. 실험군에 대한 건강관리프로그램 치료를 받기 전-후의 대사중후군 수치변화(wc=wc2-wc1, bmi=bmi2-bmi1)를 계산하여 출력하고, wc, bmi에 대한 박스플랏과 산점도를 구해서 결과를 해석하라.

풀이)
다음은 wc와 bmi의 수치변화 계산 결과이다.

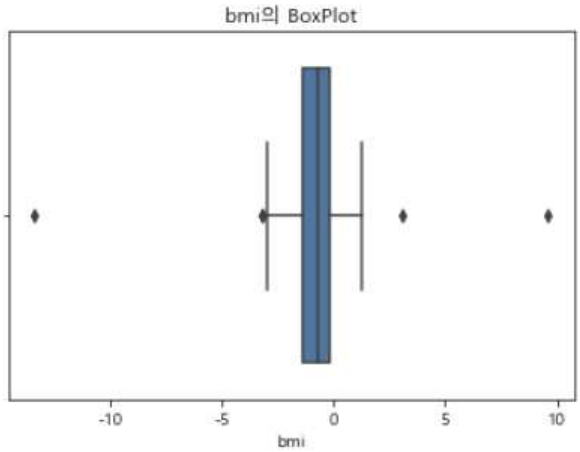
```
In [7]:  #4-1 wc, bmi의 수치변화 계산
         desa['wc']=desa['wc2']-desa['wc1']
         desa.wc.head(5)

Out[7]:  0   -1.6
         1   -5.2
         2    0.4
         3    0.6
         4    1.8

In [8]:  desa['bmi']=desa['bmi2']-desa['bmi1']
         desa.bmi.head(5)

Out[8]:  0   -0.8
         1   -1.6
         2   -0.5
         3   -0.2
         4   -0.2
         Name: bmi, dtype: float64
```

다음은 각각 wc와 bmi에 대한 박스플랏이다.

```
In [10]: #4-2 wc, bmi에 대한 박스플랏
         sns.boxplot(x='wc', data=desa)
         plt.title("wc의 BoxPlot")
         plt.show()
```



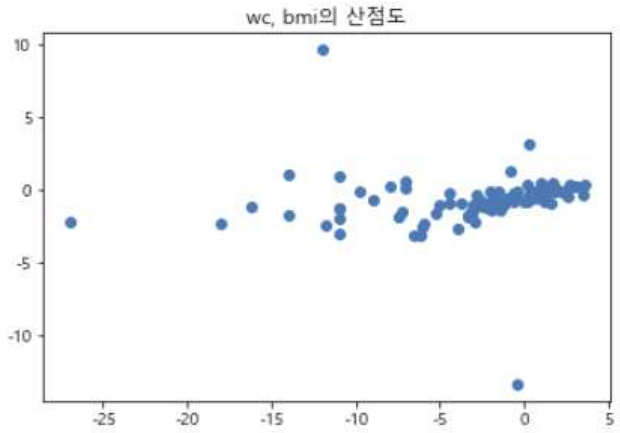
In [11]:
```
sns.boxplot(x='bmi', data=desa)
plt.title("bmi의 BoxPlot")
plt.show()
```



wc의 박스플랏은 0을 중심으로 약 70%의 자료가 왼쪽에 분포하는 것을 알 수 있다.
이를 통해 건강관리프로그램이 wc의 수치 감소에 영향을 미친다고 할 수 있다.
반면, bmi의 박스플랏은 0과 매우 가까운 지점에서 대칭적으로 자료들이 분포하고 있다. 더하여 박스플랏에서 중앙값 기준 오른쪽 자료들이 0과 매우 가까운 수치를 나타내는 것을 볼 수 있다. 이러한 점들이 bmi의 수치 전후 차이가 대체로 매우 작아, 건강관리프로그램이 bmi 변화에 영향을 미치지 못한다고 말할 수 있는 근거가 될 수 있다.

다음은 wc와 bmi의 산점도이다.

In [12]:
```
#4-3 wc, bmi의 산점도
plt.scatter(x='wc', y='bmi', data=desa)
plt.title("wc, bmi의 산점도")
plt.show()
```



x축은 wc, y축은 bmi이다.
위의 박스플랏에서 확인했듯이, wc 축에서 0을 기준으로 왼쪽에 골고루 자료가 분포하고 bmi 축에서는 0에 가깝게 많은 자료가 집중된 것을 볼 수 있다.
추가로 산점도에서 점들의 특징적인 패턴을 확인할 수 없어, wc와 bmi가 어떠한 관계를 갖지 않을 것으로 판단된다.

5. 환자의 변수(삶의질)에 대한 평가점수에 대해 도수분포표를 작성하고, seaborn모듈을 이용해 히스토그램을 작성하고 해석해라.

풀이)
다음은 변수 '삶의질'에 대한 기술통계량이다.

```
In [13]: desa.삶의질.describe()
Out[13]: count    91.000000
         mean     54.767253
         std      19.946527
         min      10.190000
         25%      39.865000
         50%      53.600000
         75%      70.120000
         max      95.380000
         Name: 삶의질, dtype: float64
```

다음은 변수 '삶의질'의 자료에 대한 도수분포표를 작성하는 과정이다.
먼저 계급의 수를 임의로 정하여, 계급구간을 출력하였다.

```
In [14]: freq,bins=np.histogram(desa['삶의질'], bins=7)
         bins
Out[14]: array([10.19, 22.36, 34.53, 46.7 , 58.87, 71.04, 83.21, 95.38])
```

다음은 정해진 계급구간을 토대로 만들어진 도수분포표이다.

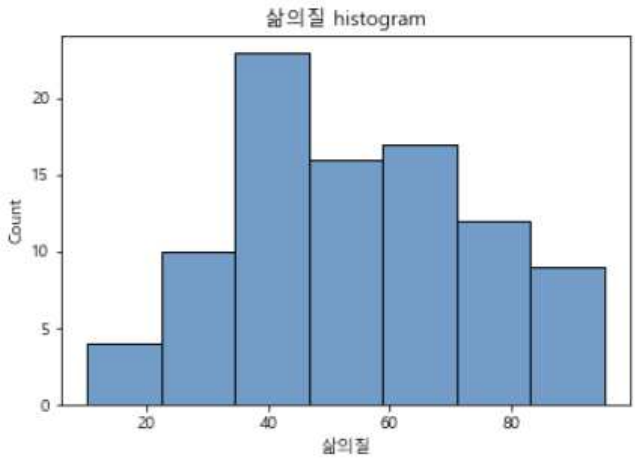
In [15]:
```
freq_class=['10.19~22.36', '22.36~34.53', '34.53~46.70', '46.70~58.87',
            '58.87~71.04', '71.04~83.21', '83.21~95.38',]
freq_table=pd.DataFrame({'frequency':freq},
                        index=pd.Index(freq_class, name='class'))
freq_table
```

Out[15]:

class	frequency
10.19~22.36	4
22.36~34.53	10
34.53~46.70	23
46.70~58.87	16
58.87~71.04	17
71.04~83.21	12
83.21~95.38	9

다음은 seaborn모듈을 이용해 만들어진 히스토그램이다.

In [16]:
```
sns.histplot(desa['삶의질'], bins=7)
plt.title('삶의질 histogram')
plt.show()
```



위의 히스토그램은 먼저 작성하였던 도수분포표와 계급의 수가 7로 같아. 각 계급구간이 동일하고, 계급구간에 해당하는 자료들의 수가 일치함을 알 수 있다.

6. 변수 ex1ex2co(실험군1, 실험군2, 대조군)에 따른 변수(삶의질)의 기초통계량을 표로 정리하여 결과를 해석하고, ex1ex2co에 따른 변수 (삶의질)의 상자도표를 그려 해석하라.

풀이)
다음은 변수 ex1ex2co에 따른 변수 '삶의질'의 기초통계량을 표로 나타낸 것이다.

In [19]:
```
#6-1 변수 ex1ex2co에 따른 삶의 질의 기초통계량
desa.groupby(['ex1ex2co']).describe()[['삶의질']]
```

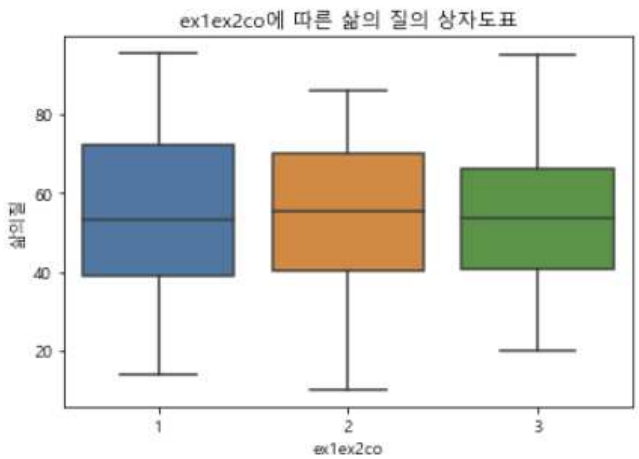
Out[19]:

	삶의질							
	count	mean	std	min	25%	50%	75%	max
ex1ex2co								
1	38.0	55.059737	20.620682	13.88	38.8925	53.19	72.005	95.38
2	33.0	53.961818	19.677531	10.19	40.1500	55.57	69.850	86.03
3	20.0	55.540500	20.062674	20.18	40.6150	53.52	66.340	95.10

ex1ex2co 1(실험군1)이 33으로, ex1ex2co 2(실험군2), ex1ex2co 3(대조군)에 비해 자료 수가 많음을 알 수 있다. 자료의 수와 별개로 평균과 표준편차, 사분위수가 비슷한 점을 통해 1(실험군1), 2(실험군2), 3(대조군)의 자료 분포 형태가 비슷할 것으로 예측된다.

다음은 변수 ex1ex2co에 따른 변수 '삶의질'의 상자도표를 나타낸 것이다.

In [20]:
```
#6-2 변수 ex1ex2co에 따른 삶의 질의 상자도표
sns.boxplot(x='ex1ex2co', y='삶의질', data=desa)
plt.title('ex1ex2co에 따른 삶의 질의 상자도표')
plt.show()
```

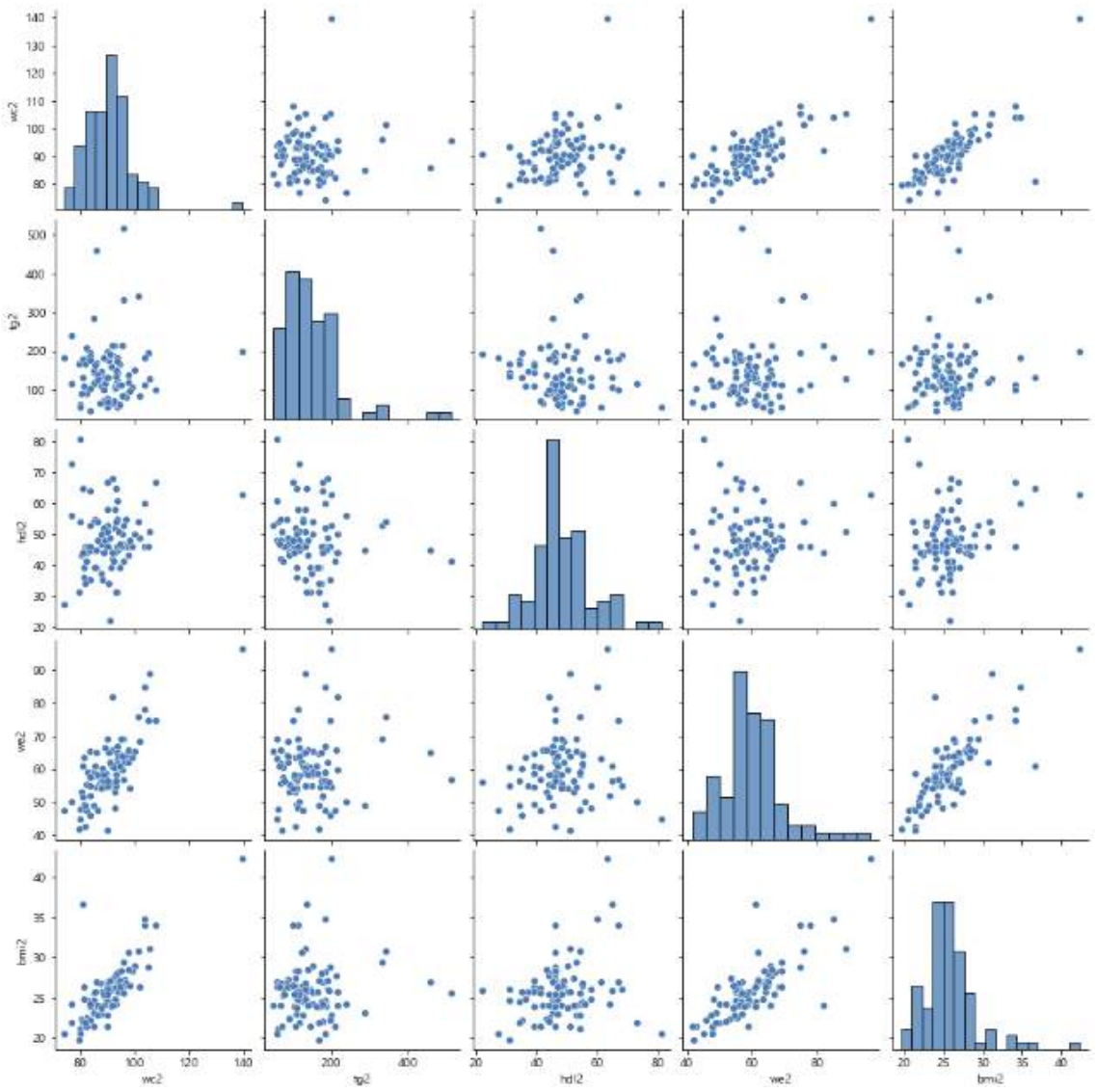


먼저 정리한 기초통계량 표를 통해 사분위수와 평균값을 통해 1, 2, 3의 자료 분포 형태가 비슷할 것이라는 예측과 같이 상자 도표가 나타났음을 알 수 있다. 세 분포 모두 중앙값과 평균이 거의 일치하기 때문에, 평균을 기준으로 대칭적이라 할 수 있다.

7. 대사중후군 환자의 건강프로그램 적용 후의 수치(wc2, tg2, hdl2, we2, bmi2)이 대해 행렬산점도(pairplot)를 그려 해석하고, 추가로 그룹변수(ex1ex2co)에 따른 대사중후군 수치 변수(wc2, tg2, hdl2, we2, bmi2)에 대한 행렬산점도를 작성하여 그룹별 결과를 해석하라.

풀이)
다음은 대사증후군 환자의 건강프로그램 적용 후의 수치(wc2, tg2, hdl2, we2, bmi2)의 행렬산점도(pairplot)이다.

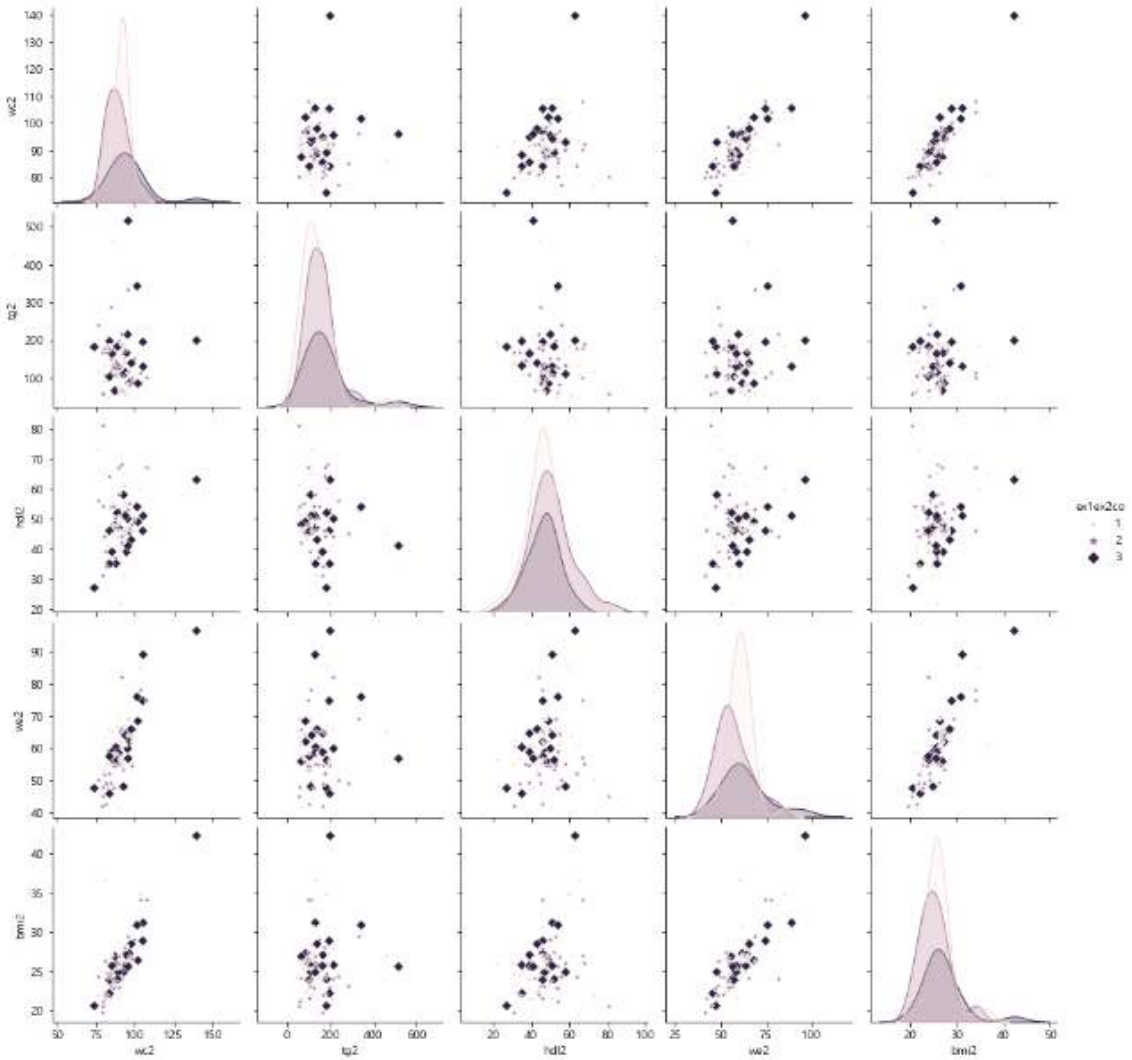
```
In [21]:  #7-1 대사중후군 환자의 건강프로그램 적용후 수치의 행렬산점도
          sns.pairplot(data=desa, vars=['wc2', 'tg2', 'hdl2', 'we2', 'bmi2'])
          plt.show()
```



위의 행렬산점도를 통해 (wc2, bmi2), (wc2, we2), (we2, bmi2)의 양의 관계를 나타내고 있는 것으로 보인다. 그 외의 변수 간 산점도에선 명확한 관계를 찾기 어렵다고 판단된다.

다음은 그룹변수(ex1ex2co)에 따른 대사중후군 수치변수(wc2, tg2, hdl2, we2, bmi2)의 행렬산점도이다.

```
In [22]:
#7-2 그룹변수(ex1ex2co)에 따른 대사중후군 수치의 행렬산점도
sns.pairplot(data=desa, vars=['wc2', 'tg2', 'hdl2', 'we2', 'bmi2'],
             hue='ex1ex2co', markers=['.', '*', 'D'])
plt.show()
```

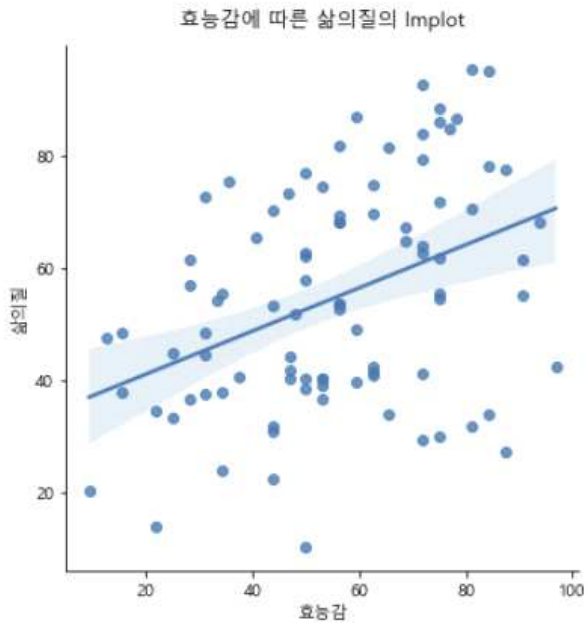


위의 행렬산점도에선 ex1ex2co 1(실험군1)이 33으로, ex1ex2co 2(실험군2), ex1ex2co 3(대조군)이 각각 다른 기호로 표현되었다.
그러나 행렬산점도의 어느 플롯에서도 실험군1, 실험군2, 대조군의 그룹으로 명확히 구분되지 않았다. 이는 그룹변수(ex1ex2co)의 실험군1, 실험군2, 대조군의 그룹이 건강프로그램 적용 후 수치(wc2, tg2, hdl2, we2, bmi2)들로 구분될만한 특성을 가지 않는다고 해석될 수 있다.

8. 변수(효능감(x), 삶의질(y)) 평가점수에 대해 lmplot을 작성하고 그래프를 해석하시오. 또한 (효능감, exco)을 원인변수, 삶의질을 결과변수로하여 lmplot을 작성하고 exco변수에 따라 lmplot의 결과를 해석하라.

풀이)
다음은 변수('효능감'(x), '삶의질'(y))에 대한 lmplot이다.

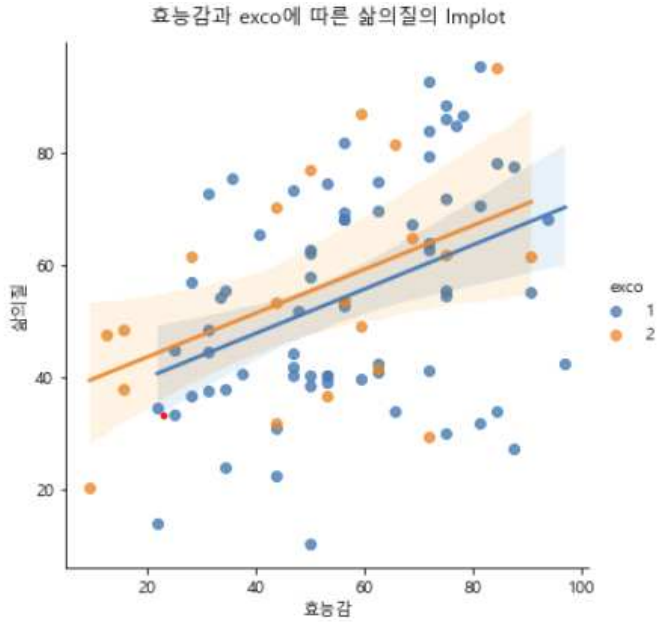
In [23]:
```
#8-1 효능감(x), 삶의질(y) 의 평가점수의 lmplot
sns.lmplot(x="효능감", y="삶의질", data=desa)
plt.title("효능감에 따른 삶의질의 lmplot", y=1.02)
plt.show()
```



위의 lmplot을 통해 직관적으로 자료(관측값)들을 추정된 회귀직선이 잘 설명하지 못하고 있다는 것을 알 수 있다. 그뿐만 아니라, 자료의 값(관측값)과 회귀직선 상의 적합값의 차이가 커, 결정계수가 작아진다. 따라서 추정된 회귀직선이 변수 '효능감'과 '삶의질'의 관계를 잘 나타내지 못한다는 할 또 하나의 근거가 될 수 있다.

다음은 (효능감, exco)을 원인변수, '삶의질'을 결과변수로 하여 작성한 lmplot이다.

In [24]:
```
#8-2 효능감(x1), exco(x2), 삶의질(y) 의 평가점수의 lmplot
sns.lmplot(x="효능감", y="삶의질", hue='exco', data=desa)
plt.title("효능감과 exco에 따른 삶의질의 lmplot", y=1.02)
plt.show()
```



범주형 변수인 exco(1, 2)로 인해 2개의 추정된 회귀직선이 표현되었다.
먼저 exco(1)이 대상이 되는 회귀직선은 자료의 값(관측값)과 회귀직선 상의 적합값의 차이가 큰 것을 통해, 결정계수가 작아져, exco(1) 내에서 '효능감'과 '삶의질'의 관계를 잘 나타내지 못한다고 할 수 있다.
다음 exco(2)이 대상이 되는 회귀직선 또한 같은 이유로, exco(2) 내에서 '효능감'과 '삶의질'의 관계를 잘 나타내지 못했다고 할 수 있을 것이다.