# [파이썬 통계분석 ; 2차-과제물]

2018015027 정보통계학과 김한탁

1. 특정 제품 A의 선호와 비선호를 알아보기 위해 1,000명을 대상으로 조사한 결과. A를 선호하는 사람들이 520명임을 알게 되었다, 제품 A의 선호율에 대해 물음에 답하여라. (주교재 5장 연습문제 3번)

(1) 신뢰수준 90%의 신뢰구간을 구하라.

solve)

In [1]:
```
#1-1
import numpy as np
import pandas as pd
from scipy import stats
from statsmodels.stats.proportion import proportions_ztest
import matplotlib.pyplot as plt
import math
```

In [3]:
```
def confidence_interval(x, n, ppf):
    phat=x/n
    interval=ppf*math.sqrt((phat*(1-phat))/n)
    return (phat-interval), (phat+interval)

confidence_interval(520, 1000, 1.645)
print("하한=%.5f, 상한=%.5f"%confidence_interval(520, 1000, 1.645))
```

```
하한=0.49401, 상한=0.54599
```

(2) 신뢰수준 95%의 신뢰구간을 구하라.

solve)

In [4]:
```
#1-2
confidence_interval(520, 1000, 1.96)
print("하한=%.5f, 상한=%.5f"%confidence_interval(520, 1000, 1.96))
```

```
하한=0.48903, 상한=0.55097
```

(3) 만일 4000명을 대상으로 조사하였을 때, A를 2,080명이 선호하였다면 신뢰수준 95%에서 A 선호율에 대한 신뢰구간을 구하고 (2)의 결과를 비교하라.

solve)

In [5]:
```
#1-3
confidence_interval(2080, 4000, 1.96)
print("하한=%.5f, 상한=%.5f"%confidence_interval(2080, 4000, 1.96))
```

```
하한=0.50452, 상한=0.53548
```

(2)의 결과와 비교하였을 때, 신뢰구간의 폭이 작아졌음을 알 수 있다, 이는 표본비율이 같을 때, 표본의 크기(n)가 클수록 더 정확한 결과를 얻을 수 있음을 보여주는 결과이다.

**2. 어느 금융기관의 일일 방문 고객 수는 정규분포한다. 이 금융기관에서 10일간의 방문 고객 수를 얻은 결과는 다음과 같다. (주교재 5장 연습문제 5번)**

{172, 169, 176, 170, 174, 173, 168, 172, 173, 170}

**(1) 이 금융기관의 일일 방문 고객 수 평균 $\mu$와 표준편차 $\sigma$을 추정하라.**

solve)

다음은 주어진 데이터를 입력한 결과이다.

In [5]:
```
#2-1
import numpy as np
import pandas as pd

x=np.array([172, 169, 176, 170, 174, 173, 168, 172, 173, 170])
d={'visitor':x}
data=pd.DataFrame(data=d)
data.describe().round(3)
```

Out[5]:

| | visitor |
|---|---|
| count | 10.000 |
| mean | 171.700 |
| std | 2.452 |
| min | 168.000 |
| 25% | 170.000 |
| 50% | 172.000 |
| 75% | 173.000 |
| max | 176.000 |

위의 describe()을 통해, 평균 $\mu$은 171.7, 표준편차 $\sigma$은 2.452임을 확인할 수 있다.

**(2) 평균 $\mu$에 대한 95% 신뢰구간을 구하라.**

solve)

In [6]:
```
#2-2
def confidence_interval(data, confidence=0.95):
    data=np.array(data)
    mean=np.mean(data)
    n=len(data)
    stderr = stats.sem(data)
           # standard error mean :표준오차
    interval=stderr*stats.t.ppf((1 + confidence) / 2, n-1)
            # ppf : inverse of cdf
    return (mean-interval, mean+interval)

print("하한=%.3f, 상한=%.3f"%confidence_interval(data))
```

```
하한=169.946, 상한=173.454
```

**3. 분산이 같은 정규분포하는 두 개의 모집단으로부터 다음과 같이 표본이 각각 얻어졌다고 할 때 유의수준 0.05에서**

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 \neq \mu_2$$

를 검정하여라. (주교재 6장 연습문제 3번)

| 모집단 1 | 모집단 2 |
|---|---|
| 4.8 | 5.0 |
| 5.2 | 4.7 |
| 5.0 | 4.9 |
| 4.9 | 4.8 |
| 5.1 | |

solve)
다음은 검정에 사용할 데이터를 입력한 결과이다.

```
In [20]: #3
import numpy as np
import pandas as pd
x=np.array([4.8, 5.2, 5.0, 4.9, 5.1, 5.0, 4.7, 4.9, 4.8])
g=np.array(['1', '1', '1', '1', '1', '2', '2', '2', '2'])
d={'x':x, 'group':g}
data1=pd.DataFrame(data=d)
data1.head()
```

Out[20]:

| | x | group |
|---|---|---|
| 0 | 4.8 | 1 |
| 1 | 5.2 | 1 |
| 2 | 5.0 | 1 |
| 3 | 4.9 | 1 |
| 4 | 5.1 | 1 |

다음은 group변수(1-모집단1, 2-모집단2)에 따른 x의 기술 통계량이다.

```
In [21]: data1.groupby('group').describe()
```

Out[21]:

| | x | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | count | mean | std | min | 25% | 50% | 75% | max |
| group | | | | | | | | |
| 1 | 5.0 | 5.00 | 0.158114 | 4.8 | 4.900 | 5.00 | 5.100 | 5.2 |
| 2 | 4.0 | 4.85 | 0.129099 | 4.7 | 4.775 | 4.85 | 4.925 | 5.0 |

모집단1의 표본평균$\overline{X_1}$는 5, 표본표준편차$S_1$는 약 0.158을 가지고, 모집단2는 표본평균$\overline{X_2}$ 4,

표본표준편차$S_2$는 약 0.129를 가짐을 알 수 있다. 여기서 모집단1, 2는 분산이 같은 모집단으로부터 얻어진 표본이므로, 가설검정에서 $S_1$, $S_2$이 아닌 합동분산추정량 $S_p^2$을 이용하여 검정통계량을 구하게 된다. 다음은 등분산 가정 하에, 두 모집단1, 2의 차이에 대한 $t$검정이다.

In [22]:
```
#등분산 가정
from scipy.stats import ttest_ind

group1=data1[data1.group=='1']
group2=data1[data1.group=='2']
result=ttest_ind(group1.x, group2.x, equal_var=True)
print("T-value=%.3f, P-value=%.3f"%result)  #귀무가설 기각
```

```
T-value=1.528, P-value=0.170
```

$t$값은 약 1.528이며, $p$값은 0.170이므로 0.05보다 크다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다.

**4. 조깅화 A와 B의 내구력에 차이가 있는가를 알아보기 위해 10명의 사람에게 A와 B를 각각 한 달씩 사용하게 한 후 운동화의 사용 잔여기간을 측정하였다. A의 내구력이 B보다 우수한지 유의수준 0.05에서 가설검정하라. (주교재 6장 연습문제 7번)**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 27 | 35 | 19 | 39 | 34 | 32 | 15 | 26 | 18 | 17 |
| B | 23 | 28 | 16 | 31 | 38 | 30 | 17 | 22 | 15 | 16 |

solve)
조깅화 A. B는 한 개의 모집단에서의 짝으로 된 표본이다. 따라서 조깅화 A의 내구력이 B보다 우수한지 확인하기 위해서 대응 T검정을 이용할 수 있다. 이에 따라 가설은 $H_0 : \mu_d = 0$ $vs$ $H_1 : \mu_d > 0$으로 설정될 수 있다. 여기서 d는 조깅화 A, B의 내구력 차이를 나타낸다.
다음은 검정에 사용할 데이터를 입력한 결과이다.

In [38]:
```
#4
import numpy as np
import pandas as pd
x=np.array([27, 35, 19, 39, 34, 32, 15, 26, 18, 17])
y=np.array([23, 28, 16, 31, 38, 30, 17, 22, 15, 16])
d={'A':x, 'B':y}
data2=pd.DataFrame(data=d)
data2.head(3)
```

Out[38]:

| | A | B |
|---|---|---|
| 0 | 27 | 23 |
| 1 | 35 | 28 |
| 2 | 19 | 16 |

다음은 조깅화 A, B 표본들의 내구력 차이를 나타내는 새로운 변수 d를 생성하여, 기술 통계량을 나타낸 것이다.

```
In [39]: data2['d']=data2.A-data2.B
         data2['d'].describe().round(3) #d=A-B , d의 평균, d의 표준편차

Out[39]: count    10.000
         mean      2.600
         std       3.658
         min      -4.000
         25%       1.250
         50%       3.000
         75%       4.000
         max       8.000
         Name: d, dtype: float64
```

이를 통해 d의 평균$\bar{d}$은 2.6, 표준편차$S_d$는 3.658임을 알 수 있다.
다음은 위의 결과를 토대로 하는, 대응 T-검정의 결과이다.

```
In [40]: # 대응T검정-단측검정
         from scipy.stats import ttest_rel
         tval=ttest_rel(x,y)[0]                          #t검정통계량 확인
         pval=ttest_rel(x,y)[1]/2                        #p-value/2 (단측이므로)
         print("T-value=%.3f, P-value=%.3f"%(tval, pval)) #귀무가설 기각

         T-value=2.248, P-value=0.026
```

$t$값은 약 2.248이며, $p$값은 0.026이므로 0.05보다 작다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각하게 된다. 그러므로 조깅화 A의 내구력이 B보다 우수하다고 할 통계적 근거가 충분하다.

**5. 맞벌이 부부들의 여가 시간 만족도 조사를 하기 위해 여성 근로자 100명, 남성 근로자 100명을 표본으로 얻었다. 남성 근로자들 100명 중에 여가 시간을 만족하는 사람은 56명, 여성 100명 중에는 40명이 만족한다고 할 때 맞벌이 여성 근로자가 남성에 비해 만족도가 낮은지 가설검정하라. (주교재 6장 연습문제 9번)**

solve)
여성 근로자의 만족도가 남성 근로자보다 낮은지 알아보고자 하는 것이므로, 가설은 다음과 같이 설정될 수 있다.

$$H_0 : p_M = p_W$$
$$H_1 : p_M > p_W$$

여기서 $p_M$은 남성 근로자의 여가 시간에 대한 만족도 , $p_W$은 여성 근로자의 여가 시간에 대한 만족도를 나타낸다.

다음은 검정에 사용할 데이터를 입력한 것이다.

```
In [45]:
#5
import pandas as pd
data3=pd.DataFrame([[56,44], [40, 60]], index=['남성', '여성'],
                   columns=['만족', '불만족'])
data3
```

Out[45]:

| | 만족 | 불만족 |
|---|---|---|
| 남성 | 56 | 44 |
| 여성 | 40 | 60 |

다음은 두 모비율 $p_M$, $p_W$의 차이에 대한 피셔의 정확검정(Fisher's exact test)이다.

```
In [46]:
#모비율 검정
from scipy.stats import fisher_exact
result=fisher_exact(data3, alternative='greater')
print("F-value=%.3f, P-value=%.3f"%result) #귀무가설 기각
```

```
F-value=1.909, P-value=0.017
```

$F$값은 약 1.909이며, $p$값은 0.017이므로 0.05보다 작다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각하게 된다. 그러므로 남성 근로자의 여가 시간에 대한 만족도가 여성보다 높다고 할 통계적 근거가 충분하다.

**6 지난 4년간의 과일 생산량 자료가 다음과 같을 때, 과일 종류별 생산량과 연도는 독립적인지 유의수준 0.05에서 가설검정하라. (주교재 7장 연습문제 4번)**

| 년도 \ 과일 | 사과 | 배 | 복숭아 | 딸기 |
|---|---|---|---|---|
| 4년 전 | 50 | 50 | 10 | 10 |
| 3년 전 | 55 | 50 | 8 | 12 |
| 2년 전 | 35 | 50 | 20 | 14 |
| 1년 전 | 60 | 50 | 12 | 14 |

solve)
자료를 구분하는 두 변수(과일 종류별 생산량, 연도) 간에 독립을 확인하기 위해서 독립성 검정(카이제곱 검정)을 이용할 수 있다. 이에 따라 가설은 다음과 같이 설정할 수 있다.

$H_0$: 과일 종류별 생산량과 연도는 독립이다.
$H_1$: 과일 종류별 생산량과 연도는 독립이 아니다.

다음은 검정에서 사용할 데이터를 입력한 것이다.

In [13]:
```
#6
import pandas as pd
years=np.tile(np.array(['4년 전', '3년 전', '2년 전', '1년 전']), 4)
A=np.repeat(np.array(['사과', '배', '복숭아', '딸기']),4)
B=np.array([50,55,35,60,50,50,50,50,10,8,20,12,10,12,14,14])
d={'Years':years, 'fruits':A, 'amounts':B}
data4=pd.DataFrame(data=d)
```

다음은 두 변수(과일 종류별 생산량, 연도)의 각 분류에 대한 분할표이다.

In [14]:
```
#빈도표
from scipy.stats import chi2_contingency
table=pd.crosstab(index=data4.Years, columns=data4.fruits,
                  values=data4.amounts, aggfunc=sum)
table
```

Out[14]:

| fruits | 딸기 | 배 | 복숭아 | 사과 |
|---|---|---|---|---|
| Years | | | | |
| 1년 전 | 14 | 50 | 12 | 60 |
| 2년 전 | 14 | 50 | 20 | 35 |
| 3년 전 | 12 | 50 | 8 | 55 |
| 4년 전 | 10 | 50 | 10 | 50 |

다음은 기대 도수표(기대 도수의 분할표)를 나타낸 것이다.

In [16]:
```
#기대도수의 분할표
chi, p, df, expected=chi2_contingency(table)
expected_table=pd.DataFrame(data=expected, index=table.index,
                            columns=table.columns)
expected_table
```

Out[16]:

| fruits | 딸기 | 배 | 복숭아 | 사과 |
|---|---|---|---|---|
| Years | | | | |
| 1년 전 | 13.6 | 54.4 | 13.6 | 54.4 |
| 2년 전 | 11.9 | 47.6 | 11.9 | 47.6 |
| 3년 전 | 12.5 | 50.0 | 12.5 | 50.0 |
| 4년 전 | 12.0 | 48.0 | 12.0 | 48.0 |

다음은 위의 결과를 토대로 하는, 카이제곱 검정의 결과이다.

In [17]:
```
#검정결과
chi, p, df, expected=chi2_contingency(table)
print(chi.round(3), p.round(3)) #귀무가설을 기각할 수 없다.
```

```
13.446 0.143
```

$\chi^2$값은 약 13.446이며, $p$값은 약 0.143이므로 0.05보다 크다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다. 그러므로 과일 종류별 생산량과 연도는 독립이라고 할 통계적 근거가 충분하다.

**7. 신제품에 대한 마케팅 전략을 수립하고자 전국 소비자를 대상으로 소비자 태도를 우편으로 조사하였다. 전국을 서울, 경기·충청, 영남, 호남, 강원, 제주로 크게 6개 권역으로 분류하고, '귀하의 가장 중요한 가치관은 무엇인가?'에 대한 질문의 응답 결과를 정리한 표이다. (주교재 7장 연습문제 5번)**

| 중요한 가치관 | 서울 | 경기·충정 | 영남 | 호남 | 강원 | 제주 |
|---|---|---|---|---|---|---|
| 자존심 | 154 | 147 | 55 | 34 | 8 | 12 |
| 안전 | 147 | 152 | 62 | 26 | 5 | 6 |
| 대인관계 | 125 | 90 | 63 | 27 | 3 | 2 |
| 성취감 | 88 | 65 | 38 | 17 | 3 | 3 |
| 책임감 | 74 | 55 | 23 | 24 | 1 | 2 |
| 체면 | 65 | 72 | 31 | 4 | 5 | 1 |
| 소속감 | 63 | 49 | 24 | 10 | 4 | 6 |
| 즐거움 | 34 | 23 | 11 | 8 | 3 | 2 |
| 합계 | 750 | 653 | 307 | 150 | 32 | 34 |

**(1) 지역과 중요가치관 간에는 독립적이라고 할 수 있는가를 가설검정하라.**
**이 경우, 크래머의 법칙에 대해 검토하고 가설검정하라.**

solve)
자료를 분류하는 두 변수 지역과 중요가치관의 독립여부를 확인하기 위해서, 독립성 검정(카이제곱 검정)을 이용할 수 있다. 이에 따라 가설은 다음과 같이 설정할 수 있다.

$H_0$: 지역과 중요가치관은 독립이다.

$H_1$: 지역과 중요가치관은 독립이 아니다.

다음은 검정에서 사용할 데이터를 입력한 것이다.

In [18]:
```
#7-1
import numpy as np
import pandas as pd

value=np.tile(np.array(['자존심', '안전', '대인관계', '성취감',
                        '책임감', '체면', '소속감', '즐거움']), 6)
region=np.repeat(np.array(['서울', '경기,충청', '영남',
                           '호남', '강원', '제주']), 8)
x=np.array([154,147,125,88,74,65,63,34,
            147,152,90,65,55,72,49,23,
            55,62,63,38,23,31,24,11,
            34,26,27,17,24,4,10,8,
            8,5,3,3,1,5,4,3,
            12,6,2,3,2,1,6,2])
d={'value':value, 'region':region, 'x':x}
data5=pd.DataFrame(data=d)
```

다음은 두 변수(지역, 중요가치관)의 각 분류에 대한 분할표이다.

In [26]:
```
#빈도표
from scipy.stats import chi2_contingency
table1=pd.crosstab(index=data5.value, columns=data5.region,
                   values=data5.x, aggfunc=sum)
table1
```

Out[26]:

| region / value | 강원 | 경기,충청 | 서울 | 영남 | 제주 | 호남 |
|---|---|---|---|---|---|---|
| 대인관계 | 3 | 90 | 125 | 63 | 2 | 27 |
| 성취감 | 3 | 65 | 88 | 38 | 3 | 17 |
| 소속감 | 4 | 49 | 63 | 24 | 6 | 10 |
| 안전 | 5 | 152 | 147 | 62 | 6 | 26 |
| 자존심 | 8 | 147 | 154 | 55 | 12 | 34 |
| 즐거움 | 3 | 23 | 34 | 11 | 2 | 8 |
| 책임감 | 1 | 55 | 74 | 23 | 2 | 24 |
| 체면 | 5 | 72 | 65 | 31 | 1 | 4 |

다음은 기대 도수표(기대 도수의 분할표)를 나타낸 것이다.

In [27]:
```
#기대도수의 분할표
chi, p, df, expected=chi2_contingency(table1)
expected_table1=pd.DataFrame(data=expected, index=table1.index,
                columns=table1.columns)
expected_table1.round(3)
```

Out[27]:

| region / value | 강원 | 경기,충청 | 서울 | 영남 | 제주 | 호남 |
|---|---|---|---|---|---|---|
| 대인관계 | 5.151 | 105.104 | 120.717 | 49.413 | 5.472 | 24.143 |
| 성취감 | 3.556 | 72.556 | 83.333 | 34.111 | 3.778 | 16.667 |
| 소속감 | 2.592 | 52.891 | 60.748 | 24.866 | 2.754 | 12.150 |
| 안전 | 6.613 | 134.940 | 154.984 | 63.440 | 7.026 | 30.997 |
| 자존심 | 6.812 | 139.008 | 159.657 | 65.353 | 7.238 | 31.931 |
| 즐거움 | 1.346 | 27.463 | 31.542 | 12.911 | 1.430 | 6.308 |
| 책임감 | 2.974 | 60.689 | 69.704 | 28.532 | 3.160 | 13.941 |
| 체면 | 2.957 | 60.350 | 69.315 | 28.373 | 3.142 | 13.863 |

위의 지역과 중요가치관이 독립이라는 가정하에 구해진 기대도수가 5보다 작은 항목이 10개로 전체 항목(48개의 항목)의 20%를 넘게 된다. 따라서 크래머 법칙에 따라 일부 항목을 통합하는 것이 합당하다고 판단된다. 이에 따라 두 변수의 각 분류에 따라 기대 도수가 가장 낮은 강원과 제주의 데이터를 통합하였다.

다음은 강원과 제주를 통합한 데이터를 재입력한 것이다.

In [34]:
```
# 강원,제주 통합
value=np.tile(np.array(['자존심', '안전', '대인관계', '성취감',
                        '책임감', '체면', '소속감', '즐거움']), 5)
region=np.repeat(np.array(['서울', '경기,충청', '영남',
                           '호남', '강원,제주']), 8)
x=np.array([154,147,125,88,74,65,63,34,
            147,152,90,65,55,72,49,23,
            55,62,63,38,23,31,24,11,
            34,26,27,17,24,4,10,8,
            20,11,5,6,3,6,10,5])
d={'value':value, 'region':region, 'x':x}
data6=pd.DataFrame(data=d)
```

다음은 두 변수(지역, 중요가치관)의 각 분류에 대한 분할표이다.

In [35]:
```
#빈도표
from scipy.stats import chi2_contingency
table2=pd.crosstab(index=data6.value, columns=data6.region,
                   values=data6.x, aggfunc=sum)
table2
```

Out[35]:

| region<br>value | 강원,제주 | 경기,충청 | 서울 | 영남 | 호남 |
|---|---|---|---|---|---|
| 대인관계 | 5 | 90 | 125 | 63 | 27 |
| 성취감 | 6 | 65 | 88 | 38 | 17 |
| 소속감 | 10 | 49 | 63 | 24 | 10 |
| 안전 | 11 | 152 | 147 | 62 | 26 |
| 자존심 | 20 | 147 | 154 | 55 | 34 |
| 즐거움 | 5 | 23 | 34 | 11 | 8 |
| 책임감 | 3 | 55 | 74 | 23 | 24 |
| 체면 | 6 | 72 | 65 | 31 | 4 |

다음은 기대 도수표(기대 도수의 분할표)를 나타낸 것이다.

In [39]:
```
#기대도수의 분할표
chi, p, df, expected=chi2_contingency(table2)
expected_table2=pd.DataFrame(data=expected, index=table2.index,
                columns=table2.columns)
expected_table2.round(3)
```

Out[39]:

| region<br>value | 강원,제주 | 경기,충청 | 서울 | 영남 | 호남 |
|---|---|---|---|---|---|
| 대인관계 | 10.623 | 105.104 | 120.717 | 49.413 | 24.143 |
| 성취감 | 7.333 | 72.556 | 83.333 | 34.111 | 16.667 |

| | | | | | |
|---|---|---|---|---|---|
| 소속감 | 5.346 | 52.891 | 60.748 | 24.866 | 12.150 |
| 안전 | 13.639 | 134.940 | 154.984 | 63.440 | 30.997 |
| 자존심 | 14.050 | 139.008 | 159.657 | 65.353 | 31.931 |
| 즐거움 | 2.776 | 27.463 | 31.542 | 12.911 | 6.308 |
| 책임감 | 6.134 | 60.689 | 69.704 | 28.532 | 13.941 |
| 체면 | 6.100 | 60.350 | 69.315 | 28.373 | 13.863 |

다음은 위의 결과를 토대로 하는, 카이제곱 검정의 결과이다.

```
In [40]:
#검정결과
chi, p, df, expected=chi2_contingency(table2)
print(chi.round(3), p.round(3)) #귀무가설을 기각
```

```
48.762 0.009
```

$\chi^2$값은 약 48.762이며, $p$값은 약 0.009이므로 0.05보다 작다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각하게 된다. 그러므로 지역과 중요가치관은 독립이 아니라고 할 통계적 근거가 충분하다.

**(2) 서울과 경기·충청 지역 간에 자존심에 대한 비율이 같은가를 가설검정하라.**

solve)
서울과 경기·충청 지역 간에 자존심에 대한 비율이 같은지 알아보기 위한 것이므로, 가설은 다음과 같이 설정될 수 있다.

$$H_0 : p_S = p_{G.C}$$
$$H_1 : p_S \neq p_{G.C}$$

여기서 $p_S$은 서울지역의 자존심에 대한 비율 , $p_{G.C}$은 경기·충청지역의 자존심에 대한 비율을 나타낸다.
다음은 검정에 사용할 데이터를 입력한 것이다.

```
In [44]:
#7-2
import pandas as pd
data7=pd.DataFrame([[154,596], [147,506]], index=['서울', '경기,충청'],
                   columns=['O', 'X'])
data7
```

Out[44]:

| | O | X |
|---|---|---|
| 서울 | 154 | 596 |
| 경기,충청 | 147 | 506 |

다음은 두 모비율 $p_S$, $p_{G.C}$의 차이에 대한 피셔의 정확검정(Fisher's exact test)이다.

```
In [45]:
#모비율 검정
from scipy.stats import fisher_exact
result=fisher_exact(data7, alternative='two-sided') #귀무가설 채택
print("F-value=%.3f, P-value=%.3f"%result) #귀무가설 기각

F-value=0.889, P-value=0.397
```

$F$값은 0.889이며, $p$값은 0.397이므로 0.05보다 크다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다. 그러므로 서울과 경기·충청 지역 간에 자존심에 대한 비율이 같다고 할 통계적 근거가 충분하다.

**(3) 만일 각 지역에 표본 수를 각각 (750, 653, …, 34)로 하여 면접조사된 결과라면 지역별로 중요한 가치관의 분포가 일차하는가를 가설검정할 수 있다, 이 경우, 검정의 결과에 대해 설명하고 위의 (1)과 어떤차이가 있는가를 논의해라.**

solve)
각 지역에 표본 수를 할당하여, 중요한 가치관의 분포가 일치하는가를 검정하는 것은 동일성 검정이다. 이에 따른 가설은 다음과 같이 설정할 수 있다.

$H_0$: 지역별 중요가치관은 동일하다.

$H_1$: 지역별 중요가치관은 동일하지 않다.

이에 대한 검정의 결과는 귀무가설 기각으로, 지역별 중요가치관은 동일하지 않다고 할 통계적 근거가 충분하다는 결론을 얻을 수 있다.
(1)의 문제를 해결하였다면 (3)은 동일성 검정 과정없이 결론을 구할 수 있는데, 이는 독립성 검정과 동일성 검정은 검정의 과정이 같기 때문이다.
다만 독립성 검정은 전체 표본에 대해서 두 가지 변수에 대한 분할표를 얻어 검정하는 것이고, 동질성검정은 각 집단에 대해 표본을 할당한 후 한 가지 변수에 대한 분포를 얻어 검정한다는 관점의 차이가 존재한다고 볼 수 있다.
따라서 (1)의 가설은 지역과 중요가치관을 변수로 두고 검정을 진행한 것이고, (3)의 가설은 지역을 집단으로 두고 중요가치관을 변수로 두어 검정을 진행했다는 점에서 관점의 차이가 존재한다고 생각한다.

**8. 실험군에 대한 건강관리프로그램 치료를 받기 전-후의 대사중후군 수치 중 bmi지수(사전-bmi1, 사후-bmi2)가 치료 이후에 더 감소하였다고 할 수 있는지를 검정하라. 또한 실험군 환자들만의 사전-사후 bmi지수의 차이를 하나의 그래프에 시각화(상자도표)하고, 해석하라**

solve)
변수 bmi1과 bmi2는 사전-사후 관계이므로 짝으로 된 자료이다. 따라서 bmi지수가 건강관리 프로그램 치료 이후에 감소하였는지 확인하기 위해서, 대응 T검정을 이용할 수 있다.
이에 따른 가설은 다음과 같이 설정될 수 있다.

$$H_0 : \mu_{bmi3} = 0$$
$$H_1 : \mu_{bmi3} > 0$$

여기서 bmi3는 bmi지수의 차이를 나타내는 새로운 변수이다.
다음은 검정에 사용할 데이터를 입력한 것이다.

In [1]:
```
#8
!python han-font.py
exec(open('han-font.py').read())

import numpy as np
import pandas as pd
desa=pd.read_csv("대사중후군.csv", encoding='CP949')
```

In [12]:
```
ex1ex2=desa.ex1ex2co[(desa.ex1ex2co ==1) | (desa.ex1ex2co == 2)]
bmi1=desa.bmi1[(desa.ex1ex2co ==1) | (desa.ex1ex2co == 2)]
bmi2=desa.bmi2[(desa.ex1ex2co ==1) | (desa.ex1ex2co == 2)]
bmi3=bmi1-bmi2
d={'ex1ex2':ex1ex2,  'pre_bmi':bmi1, 'post_bmi':bmi2, 'bmi감소량':bmi3}
data=pd.DataFrame(data=d)
data.head()
```

Out[12]:

| | ex1ex2 | pre_bmi | post_bmi | bmi감소량 |
|---|---|---|---|---|
| **1** | 2 | 22.7 | 21.1 | 1.6 |
| **3** | 2 | 27.6 | 27.4 | 0.2 |
| **5** | 2 | 26.9 | 26.0 | 0.9 |
| **6** | 1 | 33.0 | 30.7 | 2.3 |
| **7** | 2 | 31.0 | 34.1 | -3.1 |

위의 결과를 통해 변수ex1ex2co가 1,2(실험군)인 개체의 bmi1, bmi2만 추출하였고, 또한 추출한 bmi지수의 차이를 나타내는 새로운 변수bmi3를 생성하였음을 확인할 수 있다.
다음은 문제에서 실험군1, 2를 따로 구분하지 않았으므로, 새로운 변수 “실험군”에 실험군1, 2를 모두 1로 변환한 결과이다.

In [13]:
```
#실험군 1,2 통합
data.loc[(data.ex1ex2 ==1) | (data.ex1ex2 == 2), "실험군"] = '1'
data.head(3)
```

Out[13]:

| | ex1ex2 | pre_bmi | post_bmi | bmi감소량 | 실험군 |
|---|---|---|---|---|---|
| **1** | 2 | 22.7 | 21.1 | 1.6 | 1 |
| **3** | 2 | 27.6 | 27.4 | 0.2 | 1 |
| **5** | 2 | 26.9 | 26.0 | 0.9 | 1 |

다음은 bmi감소량(bmi3)의 기술 통계량이다.

```
In [14]: data.bmi감소량.describe().round(3) # bmi감소량=bmi1-bmi2, bmi감소량의 평균:0.95
                                            # bmi감소량의 표준편차:2.23
```

```
Out[14]: count    71.000
         mean      0.955
         std       2.230
         min      -9.600
         25%       0.200
         50%       0.900
         75%       1.550
         max      13.400
         Name: bmi감소량, dtype: float64
```

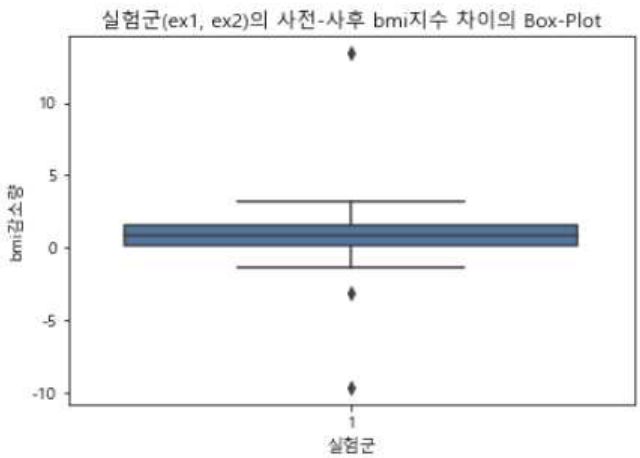
이를 통해 bmi감소량(bmi3)의 평균$\overline{bmi3}$은 0.955, 표준편차$S_{bmi3}$는 2.230임을 알 수 있다.
다음은 위의 결과를 토대로 하는, 대응 T-검정의 결과이다.

```
In [15]: #8-1대응표본 검정-단측검정
         #8-1대응표본 검정-단측검정
         from scipy.stats import ttest_rel
         tval=ttest_rel(bmi1, bmi2)[0]
         pval=ttest_rel(bmi1, bmi2)[1]/2
         print("T-value=%.3f, P-value=%.3f" %(tval, pval))#귀무가설 기각
```

```
T-value=3.608, P-value=0.000
```

$t$값은 약 3.608이며, $p$값은 0.000이므로 0.05보다 작다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각하게 된다. 그러므로 건강관리프로그램 치료 이후의 실험군의 bmi지수가 감소하였다고 할 통계적 근거가 충분하다.
다음은 실험군 환자들의 사전-사후 bmi지수의 차이를 하나의 그래프에 시각화한 것이다.

```
In [16]: #8-2 상자도표
         import seaborn as sns
         import matplotlib.pyplot as plt
         sns.boxplot(x='실험군', y='bmi감소량', data=data)
         plt.title("실험군(ex1, ex2)의 사전-사후 bmi지수 차이의 Box-Plot")
         plt.show()
```



사전-사후 bmi지수 차이의 평균이 0보다 큰 0.95를 나타내며, 분산이 작아 평균을 중심으로 자료들이 집중되어있다. 이를 통해 50% 이상의 사전-사후 bmi지수 차이가 0보다 크다고 할 수 있다. 따라서 건강관리프로그램 치료 이후의 실험군의 bmi지수가 감소하였다고 할 수 있다.

**9. 변수 group(1-실험군A, 2-실험군B, 3-대조군)에 따른 변수(삶의질)의 평균 점수에 차이가 있는지를 검정하라. group변수에 따른 (삶의질)의 상자도표를 그려 해석하고, 사후분석을 실행하고 결과를 해석하라.**

solve)
변수 group의 수준인 1(실험군A), 2(실험군B), 3(대조군)에 따른 변수(삶의질)의 평균 차이를 확인하기 위해서, 독립변수가 1개인 경우에 세 집단 이상의 평균을 비교하는 일원배치 분산분석을 적용할 수 있다. 이에 따른 가설은 다음과 같이 설정할 수 있다.

$$H_0 : \mu_{group1} = \mu_{group2} = \mu_{group3}$$
$$H_1 : not\ H_0$$

여기서 $\mu_{group1}, \mu_{group2}, \mu_{group3}$는 1(실험군A), 2(실험군B), 3(대조군)에 따른 삶의 질의 평균이다. 설정된 가설은 처리의 관점에서 변수 group(요인)의 효과 유무를 판단하는 것으로 표현될 수 있다. 다음은 검정에 사용될 데이터를 입력한 것이다.

In [27]:
```
#9
import pandas as pd
desa=pd.read_csv("대사중후군.csv", encoding='CP949')
group=desa.group
삶의질=desa.삶의질
d={'group':group, '삶의질':삶의질}
data1=pd.DataFrame(data=d)
data1.head()
```

Out[27]:

| | group | 삶의질 |
|---|---|---|
| 0 | 3 | 37.99 |
| 1 | 2 | 34.04 |
| 2 | 3 | 61.56 |
| 3 | 2 | 54.24 |
| 4 | 3 | 61.92 |

다음은 변수 group에 따른 변수 삶의 질의 기술 통계량이다.

In [2]:
```
data1.groupby(group).삶의질.describe()
```

Out[2]:

| group | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| 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 |

평균은 3(대조군)이 약55.54로 가장 높고, 분산은 2(실험군B)가 약19.68로 가장 작음을 알 수

있다. 그러나 세 집단의 평균과 분산을 비교하였을 때, 매우 약소한 차이가 나타나는 것을 통해 세 집단의 분포는 비슷할 것으로 예측할 수 있다. 이는 후에 group에 따른 삶의 질을 나타낸 상자도표를 통해 확인할 수 있다. 다음은 일원배치 분산분석의 결과이다.

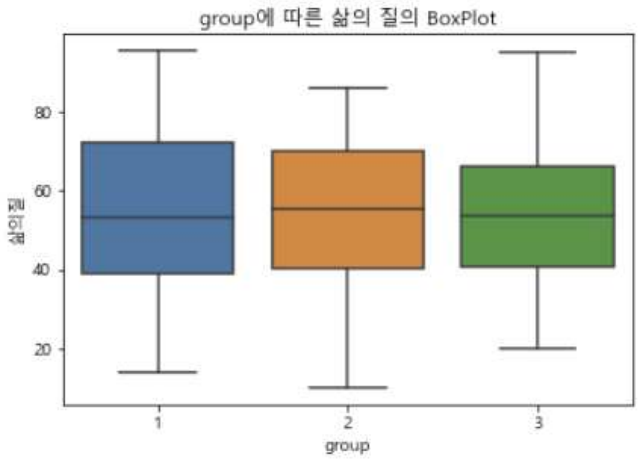
In [29]:
```
#9-1일원배치 분산분석
import statsmodels.api as sm
import statsmodels.formula.api as smf

fit=smf.ols('삶의질~C(group)', data1).fit()
sm.stats.anova_lm(fit, typ=1)
```

Out[29]:

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(group)** | 2.0 | 36.616930 | 18.308465 | 0.04504 | 0.955981 |
| **Residual** | 88.0 | 35771.137083 | 406.490194 | NaN | NaN |

$F$값은 약 0.045이며, $p$값은 0.956이므로 0.05보다 크다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다. 그러므로 변수 group에 따른 변수 삶의 질에 평균 차이가 없다고 할 통계적 근거가 충분하다. 같은 의미로 변수 group의 효과가 없다고 말할 수 있을 것이다.
다음은 group에 따른 삶의 질의 상자도표를 나타낸 것이다.

In [5]:
```
#9-2 상자도표
import seaborn as sns
import matplotlib.pyplot as plt

sns.boxplot(x='group', y='삶의질', data=data1)
plt.title("group에 따른 삶의 질의 BoxPlot")
plt.show()
```




사전에 변수 group에 따른 변수 삶의 질의 기술 통계량에서 확인하였듯 1(실험군A), 2(실험군B), 3(대조군)의 평균이 비슷하게 나타남을 알 수 있다. 분산 또한 세 집단 모두 비슷하여, 모든 박스도표가 비슷한 형태를 띠고 있다. 다음은 Tukey HSD의사후검정 결과이다.

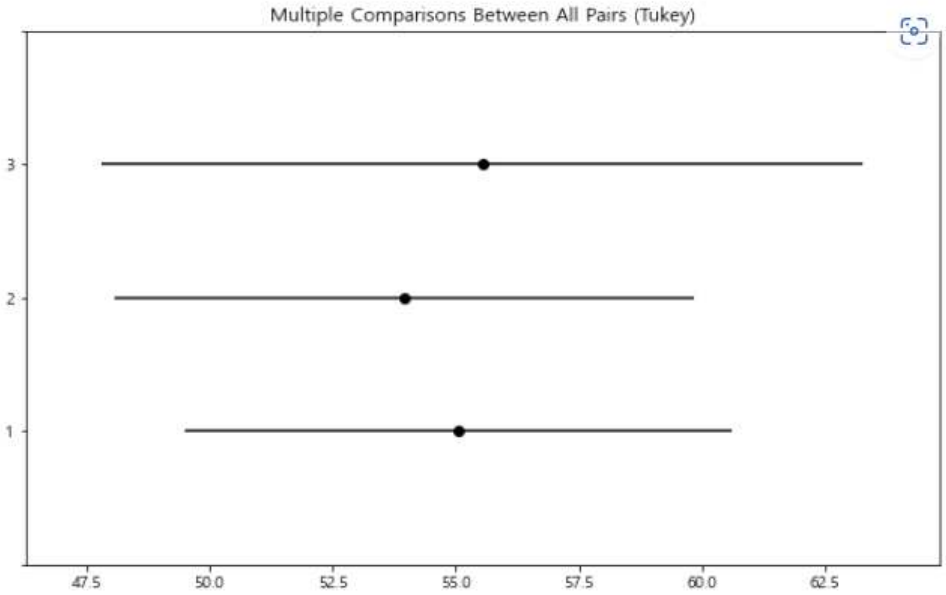
In [31]:
```
#9-3
#사후검정-Tukey HSD
from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey = pairwise_tukeyhsd(endog=data1.삶의질, groups=data1.group, alpha=0.05)
print(tukey) # ex1ex2의 수준은 모두 평균차이가 유의미하지않다.
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

| group1 | group2 | meandiff | p-adj | lower | upper | reject |
|---|---|---|---|---|---|---|
| 1 | 2 | -1.0979 | 0.9715 | -12.535 | 10.3391 | False |
| 1 | 3 | 0.4808 | 0.9959 | -12.7975 | 13.759 | False |
| 2 | 3 | 1.5787 | 0.9588 | -12.042 | 15.1994 | False |

사후검정 결과, 1(실험군A), 2(실험군B), 3(대조군)집단을 2개의 집단씩 묶어 비교하였을 때, 서로의 평균에 차이가 유의하지 않다는 점을 통해 세 집단이 구분되지 않는다는 것을 알 수 있다.
다음은 추가로 사후검정의 결과를 시각화하여 나타낸 것이다.

In [32]:
```
fig=tukey.plot_simultaneous() #마찬가지로 유의미하다고 보기 힘듬
```



**10. group변수와 건강변수를 요인(factor)변수로 설정하고, 삶의질(변수)를 반응변수로 하여 2원분산분석을 수행하고 결과를 해석하라. 또한 요인이 유의미하다면 사후분석을 수행하여 추가적인 결과를 해석하라.**

solve)
변수 group과 변수 건강의 교호작용이 없다는 가정하에, 세워지는 가설은 다음과 같다.

$H_0 : \mu_{group1} = \mu_{group2} = \mu_{group3}$ (변수 group(요인)의 효과가 없다.)
$H_1 : not\, H_0$ (변수 group(요인)의 효과가 있다.)

$H_o : \mu_{건강1} = \mu_{건강2} = \mu_{건강3} = \mu_{건강4} = \mu_{건강5}$ (변수 건강(요인)의 효과가 없다.)
$H_1 : not\, H_0$ (변수 건강(요인)의 효과가 있다.)

여기서 $\mu_{group1}, \mu_{group2}, \mu_{group3}$는 1(실험군A), 2(실험군B), 3(대조군)에 따른 삶의 질의 평균이고, $\mu_{건강1}, \mu_{건강2}, \mu_{건강3}, \mu_{건강4}, \mu_{건강5}$은 1(건강 매우 좋음), 2(건강 좋음), 3(보통), 4(건강 나쁨), 5(건강 매우 나쁨)에 따른 삶의 질의 평균이다,
설정된 각각의 가설은 처리의 관점에서 변수 group(요인), 변수 건강(요인)의 효과 유무를 판단하는 것으로 이해될 수 있다. 다음은 검정에 사용될 데이터를 입력한 것이다.

In [18]:
```
#10
import numpy as np
import pandas as pd
desa=pd.read_csv("대사증후군.csv", encoding='CP949')
```

In [27]:
```
group=desa.group
건강=desa.건강
삶의질=desa.삶의질
d={'group':group, '건강':건강, '삶의질':삶의질}
data2=pd.DataFrame(data=d)
data2.head()
```

Out[27]:

| | group | 건강 | 삶의질 |
|---|---|---|---|
| **0** | 3 | 5.0 | 37.99 |
| **1** | 2 | 3.0 | 34.04 |
| **2** | 3 | 2.0 | 61.56 |
| **3** | 2 | NaN | 54.24 |
| **4** | 3 | NaN | 61.92 |

다음은 변수 건강에 따른 변수 삶의 질의 기술 통계량이다. (변수 group에 따른 변수 삶의 질의 기술 통계량은 9번 문제에서 구한 결과와 같기 때문에 생략한다.)

In [28]:
```
data2.groupby(건강).삶의질.describe()
```

Out[28]:

| 건강 | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **2.0** | 27.0 | 66.709630 | 18.974242 | 27.31 | 59.6900 | 68.240 | 78.5650 | 95.38 |
| **3.0** | 24.0 | 59.984583 | 16.651554 | 34.04 | 42.3625 | 59.125 | 71.1025 | 88.63 |
| **4.0** | 28.0 | 43.148214 | 16.005628 | 13.88 | 33.0325 | 41.015 | 50.0425 | 78.24 |
| **5.0** | 9.0 | 37.208889 | 12.235118 | 10.19 | 33.9800 | 39.580 | 41.7400 | 53.44 |

평균은 2(건강 좋음)이 약 66.71으로 가장 높고, 분산은 5(건강 매우 나쁨)가 약 12.36 가장 작음을 알 수 있다. 특이하게 변수 건강은 5점 척도로 구성되었지만, 응답 결과에는 1(건강 매우 나쁨)이 나타나지 않았다.
따라서 집단 1(건강 매우 나쁨)을 제외한 네 집단의 평균과 분산을 비교하였을 때, 집단 1(건강 매우 나쁨)에 가까울수록 삶의 질의 평균이 높아짐을 확인할 수 있다. 이에 더하여 분산과 제 4분위 수를 함께 고려하였을 때. 건강이 좋을수록 삶의 질이 높다고 예측할 수 있다, 이는 건강의 상태가 삶의 질을 평가하는 데 효과가 있다는 예측과도 상통할 것이다.
다음은 이원배치 분산분석의 결과이다.

In [36]:
```
#10-1
import statsmodels.api as sm
import statsmodels.formula.api as smf

fit=smf.ols('삶의질~C(group)+C(건강)', data2).fit()
sm.stats.anova_lm(fit, typ=1)
```

Out[36]:

| | df | sum_sq | mean_sq | F | PR(>F) |
|---|---|---|---|---|---|
| **C(group)** | 2.0 | 19.081049 | 9.540524 | 0.033397 | 9.671677e-01 |
| **C(건강)** | 3.0 | 11452.553524 | 3817.517841 | 13.363365 | 3.496666e-07 |
| **Residual** | 82.0 | 23424.973605 | 285.670410 | NaN | NaN |

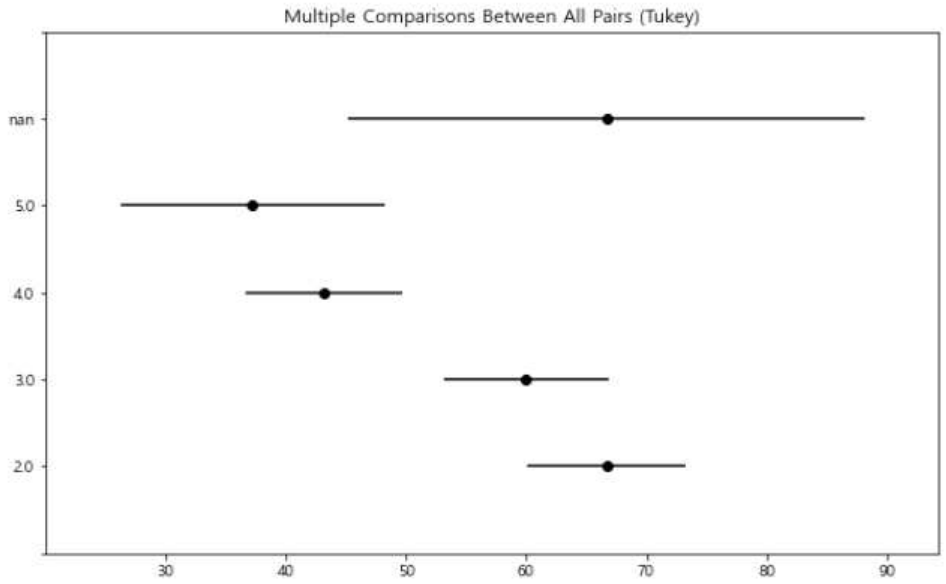
$F_{group}$값은 약 0.033이며, $p$값은 약 0.967로 0.05보다 크다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다. 그러므로 변수 group에 따른 변수 삶의 질에 평균 차이가 없다고 할 통계적 근거가 충분하다. 같은 의미로 변수 group의 효과가 없다고 말할 수 있을 것이다.
반면에 $F_{건강}$값은 약 13.363이며, $p$값은 0에 매우 가까운 값으로 0.05보다 작다. 따라서 유의수준 $\alpha = 0.05$하에서 $H_0$를 기각할 수 없다. 그러므로 변수 건강에 따른 변수 삶의 질에 평균 차이가 있다고 할 통계적 근거가 충분하며, 또한 변수 건강의 효과가 있다고 말할 수 있을 것이다.
다음은 변수 건강에 대한 Tukey HSD를 이용한 사후검정 결과이다. (변수 group에 대한 TuKey HSD 사후검정의 결과는 문제 10번에서 구한 결과와 같기 때문에 생략한다.)

In [30]:
```
#10-2
#사후검정-Tukey HSD
from statsmodels.stats.multicomp import pairwise_tukeyhsd

tukey1 = pairwise_tukeyhsd(endog=data2.삶의질, groups=data2.건강, alpha=0.05)
print(tukey1)
```

```
 Multiple Comparison of Means - Tukey HSD, FWER=0.05
====================================================
group1 group2 meandiff p-adj   lower    upper   reject
----------------------------------------------------
   2.0    3.0   -6.725 0.6132  -19.872   6.4219  False
   2.0    4.0 -23.5614    0.0 -36.2015 -10.9214   True
   2.0    5.0 -29.5007 0.0002 -47.5383 -11.4632   True
   2.0    nan   -0.043    1.0 -28.5628  28.4769  False
   3.0    4.0 -16.8364 0.0047 -29.8724  -3.8003   True
   3.0    5.0 -22.7757 0.0072 -41.0929  -4.4585   True
   3.0    nan   6.6821 0.9664 -22.0154  35.3796  False
   4.0    5.0  -5.9393 0.8878 -23.8961  12.0175  False
   4.0    nan  23.5185 0.1543  -4.9504  51.9873  False
   5.0    nan  29.4578 0.0743  -1.7841  60.6997  False
----------------------------------------------------
```

사후검정 결과, 2(건강 좋음)번과 4(건강 나쁨), 5(건강 매우 나쁨)의 집단이 구분되고, 3(건강 보통)과 4(건강 나쁨), 5(건강 매우 나쁨)의 집단이 구분된다는 것을 알 수 있다.
이는 2, 3번을 하나의 집단 그리고 4, 5번을 나머지 집단으로 나타내는 것과 같다. 다음의 사후검정 결과의 시각화된 그래프를 보면 확연하게 두 집단의 차이를 확인할 수 있다.

```
In [38]: fig=tukey1.plot_simultaneous()
```

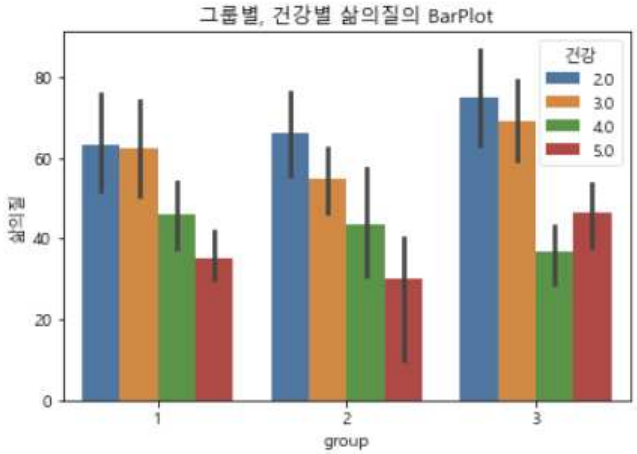



2, 3번 집단과 4, 5번 집단이 확연하게 구분되는 것뿐만 아니라 2, 3번 집단이 삶의 질 평균이 높다는 점을 확인할 수 있어, 예측했던 “건강이 좋을수록 삶의 질이 높다.”의 판단에 어느 정도 설득력을 더할 수 있는 것처럼 보이지만, 분류된 두 집단 내에서 각각의 집단의 평균의 95% 신뢰구간이 겹치는 점을 확인한다면, “건강이 좋을수록 삶의 질이 높다.”의 판단은 잘못되었다는 것을 찾을 수 있을 것이다. 그 이유를 예를 들어 설명하면 모집단에서 새로운 표본을 뽑았을 때, 그룹의 평균이 3(건강 보통) > 2(건강 좋음) > 5(건강 매우 나쁨) > 4(건강 나쁨)의 형태로 나타난다면 그것에 대해 “건강이 좋을수록 삶의 질이 높다.”라고 말할 수 없기 때문이다.

**11. 10의 결과를 하나의 그래프로 시각화하여 해석하기 위해 그룹별-건강상태에 따른 삶의질 점수에 대해 평균차트(막대도표 또는 상자도표)를 작성하고 해석하라.**

solve)
다음은 10의 결과를 그룹별-건강상태에 따른 삶의 질 점수의 막대도표 결과이다.

In [21]:
```
#11
#막대도표
import seaborn as sns
import matplotlib.pyplot as plt

sns.barplot(x="group", y="삶의질", hue="건강", data=data2)
plt.title("그룹별, 건강별 삶의질의 BarPlot")
plt.show()
```




위 95% 신뢰구간을 제외한 막대도표의 결과는 9, 10을 통해 “변수 group에 따른 삶의 질 평균에 차이가 없다.”라는 것과 “변수 건강에 따른 삶의 질 평균에 차이가 있다.”라는 가설이 채택된 이유를 더 직관적으로 이해하는 데 도움을 준다.
하지만 95% 신뢰구간을 고려한다면, 변수 group과 변수 건강, 각각의 효과에 대한 평균 차이에 관한 판단(10의 채택된 가설)이 항상 옳다고 하지 못할 것이다. 그 이유는 막대도표에 표시된 각 평균의 신뢰구간은 변수 group과 변수 건강에 따른 각각의 삶의 질 평균에 대한 신뢰구간이 아니라(9, 10의 tukey 사후검정의 결과 시각화로 확인) 2개의 변수의 교호작용의 효과에 따른 삶의 질 평균의 신뢰구간이기 때문이다. 따라서 이러한 결과는 변수 group과 변수 건강의 교호작용의 효과가 존재하여, 가설검정 과정에서 교호항을 포함한 모형으로 정의하여 진행함이 바람직하다는 것을 나타낸다.