

# 과제 보고서

제목 : 중간고사 문제 해석



과 목 명: 파이썬 통계분석

제출일자: 2022.11.03.

학 과: 정보통계학과

학 번: 2018015027

이 름: 김한택



**충북대학교**  
CHUNGBUK NATIONAL UNIVERSITY

1. 자료의 정리요약 및 결과해석. (36점)

자료) titanic

In [8]: titanic.head(6)

Out[8]:

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone	agegroup
0	0	3	male	22.0	1	0	7.2500	S	Third	man	True	NaN	Southampton	no	False	Young Adult
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False	Adult
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True	Young Adult
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False	Young Adult
4	0	3	male	35.0	0	0	8.0500	S	Third	man	True	NaN	Southampton	no	True	Young Adult
5	0	3	male	NaN	0	0	8.4583	Q	Third	man	True	NaN	Queenstown	no	True	NaN

1) python의 seaborn 모듈에 내장되어 있는 titanic자료를 불러와 연령(age)을 연령대(agegroup)로 코딩변경하고, 성별(sex)과 연령대(agegroup)에 대한 분할표를 작성하고, 결과를 해석하라.  
(단, 연령이 5세미만->Baby, 13세이하->Child, 19세이하->Teenager, 35세이하 -> YoungAdult, 35세초과->Adult로 변환하고, 피봇테이블 함수를 사용)-10점

답안)

1) 코드 및 결과

In [1]:

import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
titanic = sns.load\_dataset("titanic")

In [2]:

# 피봇테이블을 이용한 성별과 연령대에 따른 분할표 작성  
titanic.loc[(titanic.age<=5), 'agegroup']='Baby'  
titanic.loc[(titanic.age>6)&(titanic.age<=13), 'agegroup']='Child'  
titanic.loc[(titanic.age>14)&(titanic.age<=19), 'agegroup']='Teenager'  
titanic.loc[(titanic.age>20)&(titanic.age<=35), 'agegroup']='Young Adult'  
titanic.loc[(titanic.age>35), 'agegroup']='Adult'  
  
titanic\_size = titanic.pivot\_table(  
 index='sex', columns="agegroup", aggfunc="size")  
titanic\_size

Out[2]:

	agegroup	Adult	Baby	Child	Teenager	Young Adult
sex						
sex	female	73	21	11	37	111
	male	144	23	13	50	207

2) 결과 해석

성별(sex)과 연령대(agegroup)에 대한 분할표를 보면, 모든 연령대에 대해 여성(female)보다 남성(male) 탑승객이 많음을 확인할 수 있다. 표에서 성별에 따른 20세~35세 이하(YoungAdult), 35세 이상(Adult) 연령대 탑승객의 차이가 주요한 원인이라고 볼 수 있다.

또한 남성과 여성 모두 35세 이상(Adult) 연령대 탑승객이 가장 많고, 6세~13세 이하(Child) 연령대 탑승객이 가장 적음을 확인할 수 있다.

2) 성별(sex)과 연령대(agegroup)에 따른 분할표를 확률(비율)표로 변환하라. (단, crosstab 함수 사용)-6점  
 답안)

```
In [3]: # 성별과 연령대에 따른 교차표(확률표)작성
pd.crosstab(titanic.sex, titanic.agegroup, margins=True,
            margins_name='전체', normalize='all').round(2)
```

```
Out[3]:
```

agegroup	Adult	Baby	Child	Teenager	Young Adult	전체
sex						
female	0.11	0.03	0.02	0.05	0.16	0.37
male	0.21	0.03	0.02	0.07	0.30	0.63
전체	0.31	0.06	0.03	0.13	0.46	1.00

3) 성별(sex), 연령대(agegroup), 생존(survived)에 따른 3차원 빈도표를 아래와 같이 작성하고, 표에 생존율 변수(열)를 추가하여 작성하고 결과를 해석하라. -10점

답안)

1) 코드 및 결과

```
In [5]: # 성별, 연령대, 생존여부에 대한 3차원-빈도표를 작성하고, 생존율을 표에 추가하라.
titanic_df=titanic.loc[(titanic.survived==1),]
titanic_df=titanic_df.dropna()
titanic_df1=titanic_df.groupby(['sex', 'agegroup']).count()[['survived']]
titanic_df1['생존율']=titanic_df1.survived/sum(titanic_df1.survived)
titanic_df1
```

```
Out[5]:
```

		survived	생존율
sex	agegroup		
female	Adult	33	0.272727
	Baby	2	0.016529
	Teenager	12	0.099174
	Young Adult	34	0.280992
male	Adult	19	0.157025
	Baby	5	0.041322
	Child	1	0.008264
	Teenager	1	0.008264
	Young Adult	14	0.115702

2) 결과 해석

위의 빈도표의 생존율은 성별, 연령대에 대해 전체 생존자 인원 중 생존자(survived)의 비율을 나타낸다. 생존율을 확인해보았을 때, 여성(female)이면서 20세 이상~35세 이하(YoungAdult)인 그룹이 약 0.273으로 가장 높고, 여성(male)이면서 14세 이상~19세 이하(Child)인 그룹이 0으로 가장 낮게 나타났다. 결과적으로 여성의 생존율은 66.7%이고, 남성의 생존율은 33.3% 정도로 성별 대비 탑승률과 반대로 여성의 생존률이 높음을 확인할 수 있다.

4) 객실 등급별 탑승 인원이 배를 설계할 때 주어진 탑승률대로 적합하게 운영이 되고 있는지를 검정하고, 결과를 해석하라.-10점(단, 설계된 탑승률 : 1등석(15%), 2등석(25%), 3등석(60%) 이다.)

답안)

1) 코드 및 결과

```
In [9]: # 객실등급별 탑승률이 적절했는지 검토 : 적합도검정(0.15:0.25:0.6)
#객실별 탑승인원(관측빈도)
import numpy as np
titanic.groupby(['class']).size()
```

```
Out[9]: class
First      216
Second     184
Third      491
dtype: int64
```

```
In [10]: #관측빈도-기대빈도 : 카이제곱검정
from scipy.stats import chisquare
level=np.array([1,2,3])
x=np.array([216,184,491])
e_x=np.array([0.15,0.25,0.6])*np.sum(x)
chisquare(x,e_x)
```

```
Out[10]: Power_divergenceResult(statistic=61.037785260007475, pvalue=5.569463177983978e-14)
```

2) 결과 해석

객실 등급별 탑승 인원이 주어진 탑승률대로 운영되었는지 확인하기 위해서 다음과 같이 가설을 세울 수 있다.

$$H_0 : p_{First} = 0.15, p_{Second} = 0.25, p_{Third} = 0.6$$

위의 가설에 대하여, 적합도 검정을 수행할 수 있다.

객실 등급(class)에 따른 실제 탑승인원(관측빈도)는 First, Second, Third 순으로 216, 184, 491명이고, 설계된 탑승률에 따른 기대 탑승인원(기대빈도)는 133.65명, 222.75명, 534.6으로 나타났다.

이에 따른 카이제곱검정의 결과를 보면,  $\chi^2$ 검정통계량은 약 61.04이고 p-value가 0에 수렴하는 매우 작은 값을 가진다는 것을 알 수 있다. 따라서 귀무가설을 기각, 객실에 따른 탑승 인원이 각각 설계된 탑승률과 같다고 할 수 없다.

2. 대사중후군.csv 파일을 불러와 그룹(group), 건강, 연령(age) 자료에 대한 코딩 변경 후 다음 해당하는 가설 검정을 수행하시오. (단, group=1, 2, 3 : 실험군A, 실험군B, 대조군으로, 건강=1~5 : 건강매우 좋음, 건강 좋음, 보통, 건강나쁨, 건강매우나쁨으로, age : 50대미만, 50대, 60대, 70대이상으로 코딩변경) (46점)

자료) 대사중후군.csv

```
In [16]: import pandas as pd
data=pd.read_csv('대사중후군.csv',encoding='CP949')
desa=pd.DataFrame(data)

In [17]: # group, age 연속형변수를 범주형변수로 코딩변경
label={1: '실험군A', 2: '실험군B', 3: '대조군'}
desa['group']=desa['group'].map(label)
label1={1: '건강매우 좋음', 2: '건강 좋음', 3: '보통', 4: '건강나쁨', 5: '건강매우나쁨'}
desa['건강']=desa['건강'].map(label1)
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.head(6)
```

Out[17]:

	ex1ex2co	exco	site	wc1	wc2	tg1	tg2	hdl1	hdl2	ldl1	...	bp	ghp	psh	vitality	sof	RLE	mh	meh	삶의질	agegroup
0	3	2	5	98.2	96.6	150	86	65	46	248	...	24.44	5.0	21.11	5.0	62.5	100.00	52	54.88	37.99	60대
1	2	1	5	85.3	80.1	141	60	64	54	249	...	71.11	35.0	37.08	50.0	50.0	0.00	24	31.00	34.04	60대
2	3	2	1	139.3	139.7	313	199	50	63	114	...	34.44	45.0	53.61	30.0	100.0	100.00	48	69.50	61.56	60대
3	2	1	4	92.4	93.0	104	104	52	52	164	...	82.22	25.0	50.41	60.0	75.0	33.33	64	58.08	54.24	50대
4	3	2	1	103.7	105.5	216	130	43	51	117	...	82.22	40.0	63.06	50.0	62.5	66.67	64	60.79	61.92	NaN
5	2	1	5	93.8	92.0	85	189	62	68	140	...	45.56	5.0	18.89	25.0	37.5	0.00	40	25.63	22.26	60대

6 rows x 48 columns

1) 전체 환자 중에 실험군A가 대조군보다 사후(치료 이후) 체지방지수(bmi2)가 개선되었다고 할 수 있는지를 검정 절차에 맞게 상세히 수행하고, 결과를 해석하시오.

(단, bmi1-사전지수, bmi2-사후지수, 유의수준 5%, 정규분포, 등분산, 사전-동질성, 사후 평균비교) -10점

답안)

1) 체지방 사전지수(bmi1)의 정규성 검정

1-1) 코드 및 결과

```
In [20]: # 한글실행과 사전-사후 자료 분리
!python han-font.py
exec(open('han-font.py').read())
exa=desa.loc[(desa.group=='실험군A'),]
exb=desa.loc[(desa.group=='실험군B'),]
co=desa.loc[(desa.group=='대조군'),]

#체지방지수(bmi)의 정규성검정 : 정규성이 만족되지 않음
from scipy.stats import shapiro, ttest_ind, ttest_rel
shapiro(exa.bmi1)
```

Out[20]: ShapiroResult(statistic=0.9547072649002075, pvalue=0.12731009721755981)

```
In [21]: shapiro(co.bmi1)
```

Out[21]: ShapiroResult(statistic=0.7983894348144531, pvalue=0.000817552674561739)

1-2) 결과 해석

실험군 A의 bmi1은 정규성을 만족하는 반면, 대조군의 bmi1은 정규성을 만족하지 못하고 있다.



## 2) 실험군 A와 대조군의 bmi1의 등분산검정

### 2-1) 코드 및 결과

```
In [22]: # 실험군A와 대조군의 사전 동질성검정 : 등분산검정
from scipy import stats
test=stats.levene(exa.bmi1,co.bmi1)
print("statistic=%.3f, p-value=%.3f" % test)

statistic=0.294, p-value=0.590
```

### 2-2) 결과 해석

대조군의 bmi1 집단이 정규성을 만족하지 못하므로, levene의 등분산 검정을 실시한다.  
p-value가 0.590으로 0.5보다 크므로 귀무가설 채택, 등분산을 가정할 수 있다.

## 3) 실험군 A와 대조군의 bmi1 독립2표본검정

### 3-1) 코드 및 결과

```
In [23]: # 실험군A와 대조군의 사전 독립2표본검정(등분산가정)
from scipy.stats import ttest_ind
ttest=ttest_ind(exa.bmi1, co.bmi1)
print("statistic=%.3f, p-value=%.3f" % ttest)

statistic=-0.236, p-value=0.814
```

### 3-2) 결과 해석

실험군 A와 대조군의 bmi1에 대한 독립2표본검정은 bmi2(사후) 독립2표본검정을 진행하기 위한 필수적인 과정이다. 실험군 A와 대조군의 bmi1(사전)집단이 동질적으로 판단되어야 bmi2(사후)집단의 동질성을 논할 수 있기 때문이다. 이는 처리의 관점에서 보았을 때, 실험군과 대조군의 실험대상이 사전에 동질적이어야 사후에 실험군에 가해진 처리의 효과에 대해 알 수 있다는 점으로 설명할 수 있을 것이다. 따라서 bmi2(사후) 집단의 평균비교를 위해서 필수적으로 bmi1(사전)집단의 동질성을 확보해야 한다. 이를 확인하기 위한 가설은 다음과 같다.

$$H_0 : \mu_{exa.bmi1} = \mu_{co.bmi1}$$
$$H_1 : \mu_{exa.bmi1} \neq \mu_{co.bmi1}$$

등분산이 가정된 실험군 A와 대조군의 bmi1의 독립2표본검정의 결과는, 검정통계량이 -0.236이고, p-value가 0.814로 0.05보다 커 귀무가설을 채택하게 된다. 따라서 실험군 A와 대조군의 사전 bmi는 차이가 없다고 할 수 있을 것이다.

## 4) 실험군 A와 대조군의 bmi2의 정규성검정, 등분산검정

### 4-1) 코드 및 결과

```
In [24]: # 실험군A와 대조군의 사후 정규성검정, 등분산검정
shapiro(exa.bmi2)

Out[24]: ShapiroResult(statistic=0.8701222538948059, pvalue=0.0004051632131449878)

In [25]: shapiro(co.bmi2)

Out[25]: ShapiroResult(statistic=0.7910447716712952, pvalue=0.0006371996714733541)

In [26]: stats.levene(exa.bmi2,co.bmi2)

Out[26]: LeveneResult(statistic=0.36968568110614325, pvalue=0.5456337609146895)
```

#### 4-2) 결과 해석

실험군 A와 대조군의 bmi2이 정규성을 만족하고, 등분산 검정에서 p값이 0.05보다 커 등분산으로 가정된다.

#### 5) 실험군 A와 대조군의 bmi2 독립2표본검정

##### 5-1) 코드 및 결과

```
In [28]: # 실험군A와 대조군의 사후 독립2표본검정(등분산가정)
from statsmodels.stats.weightstats import ttest_ind
ttest=ttest_ind(exa.bmi2, co.bmi2, alternative='smaller', usevar='pooled')
print("T-statistic=%.3f, P-value=%.3f, unequal-variance=%.3f" % ttest)

T-statistic=-1.034, P-value=0.153, unequal-variance=56.000
```

##### 5-2) 결과 해석

실험군A가 대조군보다 bmi2가 개선되었다고 할 수 있는지 알아보려고 하였으므로, 다음과 같이 가설을 세울 수 있다.

$$H_0 : \mu_{exa.bmi2} = \mu_{co.bmi2}$$

$$H_1 : \mu_{exa.bmi2} < \mu_{co.bmi2}$$

따라서 등분산이 가정된 실험군 A와 대조군의 bmi2의 독립2표본검정의 결과는,검정통계량이 -1.034이고 p-value가 0.153로 0.05보다 작아 귀무가설을 기각하게 된다. 따라서 실험군 A의 bmi2가 대조군의 것보다 작으므로, 실험군 A가 대조군보다 bmi의 개선이 일어났다고 할 수 있다.

2) 대사중후군 환자 중에 보건소에서 새로운 치료를 받은 환자그룹인 실험군A의 경우 새로운 치료를 받고 난 후의 체지방지수(bmi2)가 많이 개선되었다고 할 수 있는가를 검정하고, 결과를 해석하시오.  
(단, 유의수준 5%, 정규분포 가정) -8점

답안)

##### 1) 코드 및 결과

```
In [32]: # 실험군A의 bmi지수에 대한 사전-사후 대응표본검정
from scipy.stats import ttest_rel
test=ttest_rel(exa.bmi1,exa.bmi2, alternative='greater')
print("statistic=%.3f, p-value=%.3f" % test)

statistic=3.082, p-value=0.002
```

##### 2) 결과 해석

실험군 A의 사전-사후 bmi차이를 통해, bmi의 개선 여부를 판단할 수 있다. 따라서 다음의 가설을 통해 검정을 진행할 수 있다.

$$H_0 : d = 0, d = \mu_{bmi1} - \mu_{bmi2}$$

$$H_1 : d > 0$$

따라서 실험군 A의 사전-사후 bmi에 대한 대응표본검정의 결과는, 검정통계량이 3.082이고 p-value가 0.002 0.05보다 작아 귀무가설을 기각하게 된다. 따라서 실험군 A의 bmi는 개선되었다고 할 수 있다.

3) 연령대(agegroup)에 따른 사후 체지방지수(bmi2)의 평균 차이가 있는지를 검정 절차에 따라 상세히 분석하고 결과를 해석하시오. (단, 유의수준 5%, 정규분포가정, 등분산이 아닌 경우 변수변환 후 실행, 사전-동질성검정과 사후-분산분석 결과를 비교설명)

답안)

1) 연령대(agegroup)에 따른 bmi1 등분산검정

1-1) 코드 및 결과

```
In [33]: # 연령대에 따른 평균 bmi지수에 차이가 있는지에 대한 분산분석
# 분산분석 : 사전 동질성검정 : 결측치 제거 후 분산분석
desa=desa.dropna()
from scipy import stats
stats.levene(desa.bmi1[desa.agegroup=='50대미만'],
             desa.bmi1[desa.agegroup=='50대'],
             desa.bmi1[desa.agegroup=='60대'],
             desa.bmi1[desa.agegroup=='70대이상'])

Out[33]: LeveneResult(statistic=0.728041904910836, pvalue=0.5384008210104763)
```

1-2) 결과 분석

등분산 검정 결과, p-value가 0.05보다 크므로 등분산으로 가정할 수 있다.

2) 연령대(agegroup)에 따른 bmi1 분산분석

2-1) 코드 및 결과

```
In [34]: # 등분산을 만족하는 분산분석 : bmi1(사전)
import statsmodels.api as sm
import statsmodels.formula.api as smf
fit1=smf.ols('bmi1~agegroup', desa).fit()
sm.stats.anova_lm(fit1, typ=1)

Out[34]:
```

	df	sum_sq	mean_sq	F	PR(>F)
agegroup	3.0	131.584202	43.861401	3.595089	0.017327
Residual	76.0	927.227798	12.200366	NaN	NaN

2-2) 결과 해석

연령대에 따른 bmi2의 평균 차이가 있는지 알아보기 위해선, 연령대에 따른 bmi1의 평균 차이가 없어야 한다. 즉 연령대에 따른 bmi1이 동질적이어야 한다. 연령대에 따른 bmi1(사전)이 동질적일 경우에, bmi2(사후)에 대해 순수한 연령(처리)의 효과로 평균차이를 비교할 수 있기 때문이다. 따라서 이를 확인하기 위한 가설은 다음과 같다.

$$H_0 : \mu_{50\text{대 미만 } bmi1} = \mu_{50\text{대 } bmi1} = \mu_{60\text{대 } bmi1} = \mu_{70\text{대 이상 } bmi1}$$

따라서 연령대에 대한 bmi1의 분산분석 결과는, 검정통계량이 약 3.96이고 p-value가 약 0.017로 0.05보다 작으므로 귀무가설을 기각하게 된다. 따라서 연령대에 따른 사전 bmi가 모두 같다고 할 수 없다.



### 3) 연령대(agegroup)에 따른 bmi2 등분산검정

#### 3-1) 코드 및 결과

```
In [35]: # 분산분석 : 사후 등분산검정
stats.levene(desa.bmi2[desa.agegroup=='50대미만'],
             desa.bmi2[desa.agegroup=='50대'],
             desa.bmi2[desa.agegroup=='60대'],
             desa.bmi2[desa.agegroup=='70대이상'])

Out[35]: LeveneResult(statistic=0.6107156711851179, pvalue=0.6101086130291871)
```

#### 3-2) 결과 분석

등분산 검정 결과, p-value가 0.05보다 크므로 등분산으로 가정할 수 있다.

### 4) 연령대(agegroup)에 따른 bmi2 분산분석

#### 4-1) 코드 및 결과

```
In [36]: #유의수준 5%에서 분산분석 : 사전차이가 있었고, 사후차이도 있다.
#사후 연령대별 순수한 차이라고 보기 어렵다. 추후 공분산분석이 필요.
import statsmodels.api as sm
import statsmodels.formula.api as smf
fit1=smf.ols('bmi2~agegroup', desa).fit()
sm.stats.anova_lm(fit1, typ=1)
```

```
Out[36]:
```

	df	sum_sq	mean_sq	F	PR(>F)
agegroup	3.0	100.051765	33.350588	2.870667	0.041847
Residual	76.0	882.946235	11.617714	NaN	NaN

#### 4-2) 결과 분석

연령대에 따른 bmi2의 평균 차이가 있는지 알아보기 위해서 다음과 같이 가설을 세울 수 있다.

$$H_0 : \mu_{50\text{대 미만 } bmi2} = \mu_{50\text{대 } bmi2} = \mu_{60\text{대 } bmi2} = \mu_{70\text{대 이상 } bmi2}$$

따라서 연령대에 대한 bmi2의 분산분석 결과는, 검정통계량이 약 2.87이고 p-value가 약 0.041로 0.05보다 작으므로 귀무가설을 기각하게 된다. 따라서 연령대에 따른 사후 bmi가 모두 같다고 할 수 없다.

연령대에 대한 사전, 사후 bmi에 대한 각각의 분산분석 결과에서 사전 차이, 사후 차이가 있다고 판단되었다. 하지만 연령대에 대한 사후 bmi차이가 유의하다고 할 수 없다. 그 이유는 먼저 언급한 연령대에 따른 사전 bmi차이와 모든 자료가 각각 다른 처리(실험군 A, 실험군 B, 대조군)로 인해 연령대에 의한 효과인지 다른 처리에 의한 효과인지 명확히 할 수 없기 때문이다.

따라서 순수한 연령대의 영향으로 사후 bmi차이를 확인하고자 한다면, 대조군을 대상으로 연령대에 대한 사후 bmi 평균차이를 확인하는 것이 좋다고 판단된다.

4) 3)의 결과가 유의하였다면, 다중비교(Tukey)방법을 이용해 결과를 해석하라. 또한 agegroup에 따른 상자도표를 정리하고, agegroup에 따른 사전(bmi1), 사후(bmi2) 지수의 평균을 하나의 표로 정리하여 상자도표와 함께 결과를 해석하시오. (단, 연령을 연령대로 변환 시 결측치가 있는 경우는 `desa=desa.dropna()`를 콘솔 창에 실행한 후 분석할 것) -10점

답안)

#### 1) 연령대(agegroup)에 대한 사후분석(Tukey)

##### 1-1) 코드 및 결과

```
In [37]: # agegroup에 대한 사후분석
#결측치 제거 후 다중비교 (유의수준=0.1)
from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey=pairwise_tukeyhsd(endog=desa.bmi2, groups=desa.agegroup, alpha=0.1)
print(tukey)
```

Multiple Comparison of Means - Tukey HSD, FWER=0.10

```
=====
group1 group2 meandiff p-adj lower upper reject
=====
```

```
50대 50대미만 1.5476 0.7886 -2.3294 5.4246 False
50대 60대 0.8924 0.8462 -1.6536 3.4384 False
50대 70대이상 -1.4679 0.5559 -4.0686 1.1329 False
50대미만 60대 -0.6552 0.9728 -4.19 2.8796 False
50대미만 70대이상 -3.0155 0.2098 -6.5899 0.559 False
60대 70대이상 -2.3603 0.0442 -4.4164 -0.3042 True
=====
```

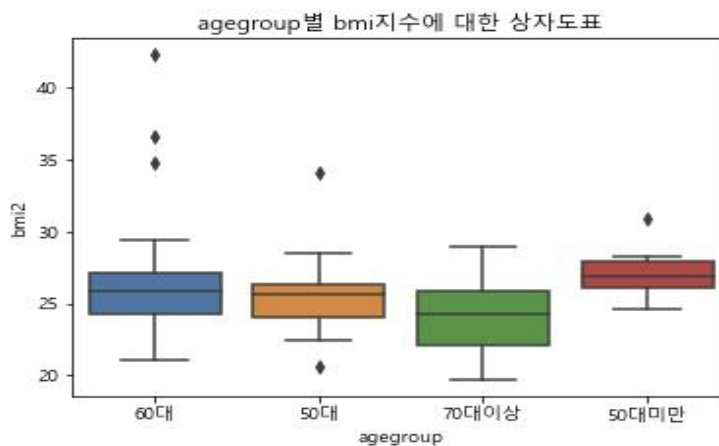
1-2) 결과 분석

60대와 70대 이상의 사후 bmi에서 차이가 있다고 나타났다. 그 외의 집단에서는 차이가 있다고 보기 힘들다.

2) 연령대(agegroup)에 따른 상자도표와 연령대(agegroup)에 따른 사전, 사후 bmi 평균

2-1) 코드 및 결과

```
In [38]: # agegroup에 따른 bmi지수의 상자도표 및 그룹별 평균비교표
import matplotlib.pyplot as plt
import seaborn as sns
sns.boxplot(x="agegroup", y="bmi2", data=desa)
plt.title("agegroup별 bmi지수에 대한 상자도표")
plt.show()
```



```
In [39]: # agegroup bmi지수의 사전-사후 평균수치변화 비교
desa.groupby(['agegroup']).mean()[['bmi1', 'bmi2']]
```

Out [39]:

	bmi1	bmi2
agegroup		
50대	27.271429	25.685714
50대미만	28.166667	27.233333
60대	27.256250	26.578125
70대이상	24.721429	24.217857

## 2-2) 결과 분석

50대 미만의 연령대에서 사후 bmi 평균이 27.233으로 가장 높게 나타났고, 70대 이상의 연령대에서 사후 bmi 평균이 24.218로 가장 낮게 나타났다. 또한 60대의 사후 bmi 상자도표를 통해 3개의 이상치를 확인할 수 있고, 50대에서는 2개, 50대 미만의 연령대에서는 1개의 이상치가 확인된다. 또한 사전 bmi가 동일하다는 가정하에, 50대의 사전-사후 bmi가 약 1.585로 가장 많이 감소했음을 알 수 있다. 반면에 70대 이상의 연령대에서는 약 0.502로 가장 적게 감소했음을 알 수 있다.

5) 실험군 A, 실험군 B, 대조군 3그룹 각각의 사전-체지방지수(bmi1)와 사후-체지방지수(bmi2)를 상자도표로 정리하여 사전-사후 bmi의 변화량(중심, 산포)을 해석하시오. -8점

답안)

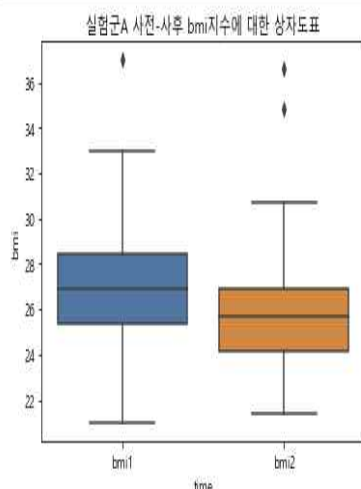
실험군 A, 실험군 B, 대조군의 사전-사후 bmi의 데이터셋

```
In [41]: # 추가문제 : 실험군a,b, 대조군 각각의 사전-사후 bmi자료의 상자도표해석
bmi=exa.iloc[:, [21,22]]
bmi=bmi.stack().reset_index()
df1=pd.DataFrame(bmi)
df1.columns=['id', 'time', 'bmi']
bmi=exb.iloc[:, [21,22]]
bmi=bmi.stack().reset_index()
df2=pd.DataFrame(bmi)
df2.columns=['id', 'time', 'bmi']
bmi=co.iloc[:, [21,22]]
bmi=bmi.stack().reset_index()
df3=pd.DataFrame(bmi)
df3.columns=['id', 'time', 'bmi']
```

### 1) 실험군 A의 사전-사후 bmi 상자도표

#### 1-1) 코드 및 결과

```
In [47]: # 한글실행 및 실험군A, B, 대조군 각각에 대한 사전-사후 BMI지수에 대한 상자도표
import matplotlib.pyplot as plt
import seaborn as sns
!python han-font.py
exec(open('han-font.py').read())
sns.boxplot(x="time", y="bmi", data=df1)
plt.title("실험군A 사전-사후 bmi지수에 대한 상자도표")
plt.show()
```



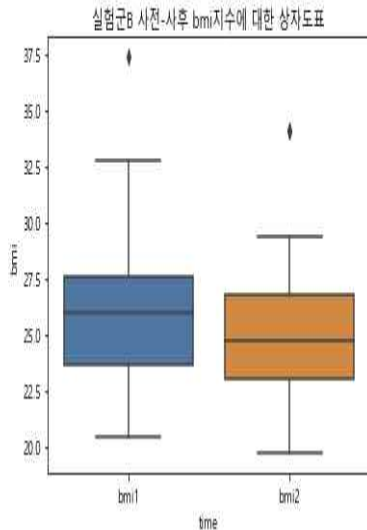
#### 1-2) 결과 해석

실험군 A의 사전 bmi 상자도표에 비해 사후 bmi 상자도표의 제 사분위수, 최댓값, 최소값이 작아졌음을 알 수 있다. 비록 기존에 나타나지 않았던 이상치가 하나 나타났지만, 대부분의 실험군 A의 사후 bmi가 감소하였으므로, 실험군 A의 bmi 평균이 감소했다고 판단된다. 또한 같은 이유로 산포 또한 감소했을 것으로 예측된다.

## 2) 실험군 B의 사전-사후 bmi 상자도표

### 2-1) 코드 및 결과

```
In [48]: sns.boxplot(x="time", y="bmi", data=df2)
plt.title("실험군B 사전-사후 bmi지수에 대한 상자도표")
plt.show()
```



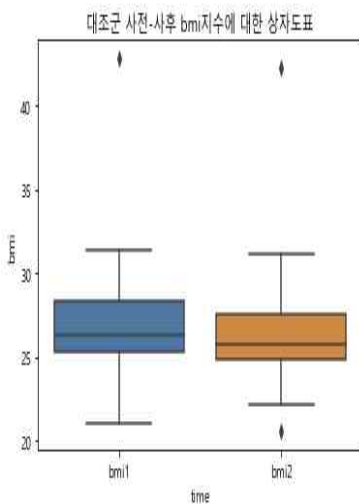
### 2-2) 결과 해석

실험군 B의 사전 bmi 상자도표에 비해 사후 bmi 상자도표의 제 사분위수, 최댓값, 이상값이 작아졌음을 알 수 있다. 이를 통해 실험군 B의 bmi 평균이 감소했음을 알 수 있다. 같은 이유로 상자도표의 이상값과 최댓값의 감소로 인해 산포 또한 감소했을 것이다.

## 3) 대조군의 사전-사후 bmi 상자도표

### 3-1) 코드 및 결과

```
In [49]: sns.boxplot(x="time", y="bmi", data=df3)
plt.title("대조군 사전-사후 bmi지수에 대한 상자도표")
plt.show()
```



### 3-2) 결과 해석

실험군 A, 실험군 B와 비교하여 사전-사후 bmi 차이가 미미하게 나타났다. 사전-사후 bmi의 상자도표의 차이가 미미하므로 평균과 산포가 거의 변하지 않았을 것으로 예측된다.

3. 대사증후군 자료에서 agegroup과 건강을 인자(factor)로, 반응변수를 '삶의질'로 고려한 분산분석과 사전-체지방지수(bmi1)을 공변량으로 고려하고, 인자(agegroup), 반응변수를 사후-체지방지수'bmi2'로 고려한 공분산 분석을 수행하시오

1) 반복 이원-분산분석의 통계적 모형을 쓰고, 분석을 수행하여 교호작용이 의미 없는 경우 이를 모형에서 제외한 후 유의한 인자에 대해서만 분산분석표를 정리하고, 유의한 인자의 사후분석(tukey)을 통해 결과를 해석하라. -10점

답안)

1) 연령대(agegroup)과 건강을 인자로, 삶의질을 반응변수로 하는 반복 이원-분산분석

1-1) 코드 및 결과

```
In [54]: # 반복 2원-분산분석 : 교호작용효과가 없으면 제거하고 2원배치분산분석을 수행
import statsmodels.api as sm
import statsmodels.formula.api as smf
fit=smf.ols('삶의질~agegroup+건강+agegroup*건강', desa).fit()
sm.stats.anova_lm(fit).round(3)
```

Out [54]:

	df	sum_sq	mean_sq	F	PR(>F)
agegroup	3.0	2662.038	887.346	3.354	0.024
건강	3.0	8422.629	2807.543	10.613	0.000
agegroup:건강	9.0	1845.228	205.025	0.775	0.640
Residual	65.0	17195.599	264.548	NaN	NaN

1-2) 결과 해석

연령대(agegroup)과 건강을 인자로, 삶의질을 반응변수로 하는 반복 이원-분산분석의 모형은 다음과 같다.

$$\text{삶의 질}_{ij} = \mu + \text{agegroup}_i + \text{건강}_j + \text{agegroup}_i * \text{건강}_j + \epsilon_{ij}$$

여기서  $i = 1, 2, 3, 4$ ;  $j = 1, 2, 3, 4, 5$

그러나 반복 2원 분산분석 결과, agegroup과 건강의 교호작용항에 대한 검정통계량이 0.775이고, p-value가 0.640으로 0.05보다 커 귀무가설 채택, 즉 agegroup과 건강의 교호효과가 없다는 결론이 얻게 되었다. 따라서 교호효과항을 제거한 새로운 모형으로 2원 배치 분산분석을 수행하는 것이 적절하다.

2) 연령대(agegroup)과 건강을 인자로, 삶의질을 반응변수로 하는 이원-분산분석

2-1) 코드 및 결과

```
In [56]: # 반복 2원-분산분석 : 교호작용효과가 없으면 제거하고 2원배치분산분석을 수행
import statsmodels.api as sm
import statsmodels.formula.api as smf
fit=smf.ols('삶의질~agegroup+건강', desa).fit()
sm.stats.anova_lm(fit, type=3).round(3)
```

Out [56]:

	df	sum_sq	mean_sq	F	PR(>F)
agegroup	3.0	2662.038	887.346	3.466	0.02
건강	3.0	8422.629	2807.543	10.967	0.00
Residual	73.0	18687.246	255.990	NaN	NaN

2-2) 결과 해석

분산분석 표를 통해, agegroup과 건강 두 인자 모두 각각의 p-value가 0.02, 0.00으로 0.05보다 작아 유의한 것으로 나타났다.



### 3) 연령대(agegroup)과 건강에 대한 사후분석(Tukey)

#### 3-1) 코드 및 결과

```
In [57]: # tukey방법의 다중비교(alpha=0.1)
from statsmodels.stats.multicomp import pairwise_tukeyhsd
tukey=pairwise_tukeyhsd(endog=desa.삶의질, groups=desa.agegroup, alpha=0.1)
print(tukey)
tukey=pairwise_tukeyhsd(endog=desa.삶의질, groups=desa.건강, alpha=0.1)
print(tukey)
```

```
Multiple Comparison of Means - Tukey HSD, FWER=0.10
=====
group1 group2 meandiff p-adj lower upper reject
-----
50대 50대미만 11.94 0.5686 -9.543 33.423 False
50대 60대 -7.6103 0.5926 -21.7181 6.4975 False
50대 70대이상 -8.6586 0.5029 -23.0698 5.7526 False
50대미만 60대 -19.5503 0.101 -39.1369 0.0363 False
50대미만 70대이상 -20.5986 0.0811 -40.4049 -0.7923 True
60대 70대이상 -1.0483 0.9965 -12.4413 10.3448 False
-----

Multiple Comparison of Means - Tukey HSD, FWER=0.10
=====
group1 group2 meandiff p-adj lower upper reject
-----
건강나쁨 건강매우나쁨 -7.8034 0.5946 -22.3067 6.6998 False
건강나쁨 건강 좋음 23.1606 0.0 12.5453 33.7759 True
건강나쁨 보통 14.8505 0.0124 3.848 25.8531 True
건강매우나쁨 건강 좋음 30.964 0.0 16.3062 45.6219 True
건강매우나쁨 보통 22.654 0.0038 7.7133 37.5946 True
건강 좋음 보통 -8.3101 0.3161 -19.5155 2.8954 False
-----
```

#### 3-2) 결과 해석

인자 agegroup에 대해서 50대 미만과 70대 이상의 삶의 질 평균에 차이가 있다고 할 수 있다.  
또 다른 인자 건강에 대해서 (건강나쁨, 건강 좋음), (건강나쁨, 보통), (건강매우나쁨, 건강 좋음), (건강매우나쁨, 보통)의 삶의 질의 평균에 차이가 있다고 할 수 있다.

2) 1)의 결과를 토대로 인자의 상대적 중요도를 해석하고, agegroup과 건강에 따른 삶의 질 평균 점수를 하나의 표로 정리하여 간략히 해석하시오. -8점

답안)  
1)의 분산분석 표에 따라 agegroup보다 건강의 p-value가 더 작음을 알 수 있다. 이것은 인자 건강이 agegroup보다 삶의 질의 평균을 더 확실하게 구분한다는 증거가 될 수 있다. 또한 같은 의미로 1)의 Tukey 사후 검정을 확인해보면, 각 인자의 수준에 대한 삶의 질 집단끼리 비교하였을 때, 건강의 수준에 대한 집단이 더 잘 구분되는 것을 알 수 있다. 따라서 삶의 질의 평균을 구분하는데 인자 건강이 agegroup보다 더 중요하다고 할 수 있다.

1) agegroup과 건강에 따른 삶의 질 평균 점수

1-1) 코드 및 결과

In [58]:

```
# 상대적 중요도(건강, 연령대가 삶의 질 점수에 미치는 영향력)
# agegroup, 건강에 따른 삶의 질 평균점수표 정리 및 최적조건
desa.groupby(['agegroup', '건강']).mean()[['삶의 질']]
```

Out [58]:

삶의 질

agegroup	건강	
50대	건강나쁨	44.360000
	건강매우나쁨	34.830000
	건강 좋음	63.804000
	보통	68.200000
50대미만	건강나쁨	41.490000
	건강 좋음	79.993333
	보통	75.565000
60대	건강나쁨	42.591538
	건강매우나쁨	35.985000
	건강 좋음	65.028182
	보통	56.770000
70대이상	건강나쁨	48.252727
	건강매우나쁨	38.650000
	건강 좋음	72.368000
	보통	50.881429

1-2) 결과 해석

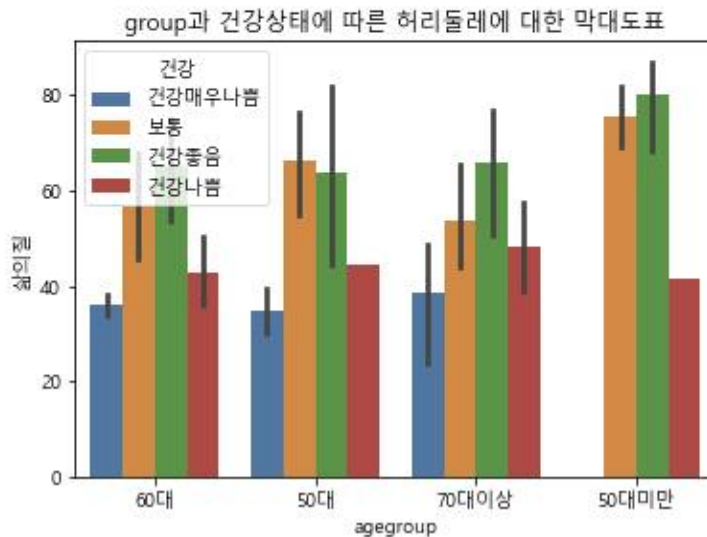
위의 표를 통해, 50대 미만 연령대의 '건강 좋음'에 해당하는 사람들의 삶의 질 평균 점수가 79.99로 가장 높고, 50대의 '건강매우나쁨'에 해당하는 사람들의 삶의 질 평균 점수가 34.83으로 가장 낮다. 전체적으로 모든 연령대에서 '건강매우나쁨' 다음 '건강나쁨'의 순서로 삶의 질 평균이 높고, 다음 순서로 '보통'과 '건강 좋음'이 비슷하게 평균이 높은 경향을 띄고 있음을 확인할 수 있다.

3) 인자 (agegroup, 건강)에 따른 삶의 질의 평균 막대도표를 작성하여 그래프를 해석하시오. -8점

답안)

1) 코드 및 결과

```
In [59]: # 연령대별 건강상태에 따른 삶의 질 점수에 대한 막대도표(3차원)
import matplotlib.pyplot as plt
import seaborn as sns
sns.barplot(x="agegroup", y="삶의 질", hue="건강", data=data)
plt.title("group과 건강상태에 따른 허리둘레에 대한 막대도표")
plt.show()
```



2. 결과 해석

50대 미만의 연령대에서는 ‘건강 좋음’ 그룹의 삶의 질 평균이 가장 높게 나타났고, ‘건강매우나쁨’의 그룹이 나타나지 않았다. 또한 특이한 점은 ‘건강 나쁨’ 그룹에 대한 신뢰구간이 나타나지 않은 것으로, 표본의 수가 적기 때문일 것으로 예측할 수 있다.

50대 연령대에서는 ‘보통’ 그룹의 삶의 질 평균이 가장 높게 나타났고, ‘건강 나쁨’ 그룹의 삶의 질 평균이 가장 낮게 나타났다. 50대 미만의 연령대와 비슷하게 ‘건강 나쁨’ 그룹에 대한 신뢰구간이 나타나지 않은 것으로 보아, 표본의 수가 적기 때문일 것으로 예측할 수 있다.

60대 연령대에서는 ‘건강 좋음’ 그룹의 삶의 질 평균이 가장 높게 나타났고, ‘건강매우나쁨’ 그룹의 삶의 질 평균이 가장 낮게 나타났다.

70대 이상의 연령대에서는 ‘건강 좋음’ 그룹의 삶의 질 평균이 가장 높게 나타났고, ‘건강매우나쁨’ 그룹의 삶의 질 평균이 가장 낮게 나타났다.

전체적으로 모든 연령대에서 ‘건강매우나쁨’ 다음 ‘건강 나쁨’의 순서로 삶의 질 평균이 높고, 다음 순서로 ‘보통’과 ‘건강 좋음’이 비슷하게 평균이 높은 경향을 띄고 있음을 확인할 수 있다.

4) 공분산분석의 결과를 해석하고, 추가로 공분산모형은 더미회귀모형과 같은 형태로 추정된 더미회귀모형을 쓰고 간략히 해석하시오. -10점

답안)

1) 공분산분석의 결과 해석

1-1) 코드 및 결과

```
In [29]: # 공분산분석 : agegroup 요인을 고려한 공분산분석~bmi1과 bmi2의 연관성이 높고, 이 효과를 배제한 후
# agegroup의 순수효과는 없으며, 연령대에 따른 bmi2(사후) 체지방지수의 순수한 차이로 보여지지 않는다.
fit2=smf.ols('bmi2~bmi1+C(agegroup)',desa).fit()
sm.stats.anova_lm(fit2, typ=3)
```

Out [29]:

	sum_sq	df	F	PR(>F)
Intercept	18.811108	1.0	4.709906	3.315543e-02
C(agegroup)	8.343772	3.0	0.696368	5.571564e-01
bmi1	583.400295	1.0	146.071158	2.785732e-19
Residual	299.545940	75.0	NaN	NaN

1-2) 결과 해석

위의 표는 agegroup을 인자로, bmi1을 공변량으로, 반응변수를 bmi2로 고려한 공분산 분석이다.

분산분석표의 결과에 따르면, agegroup(인자)의 검정통계량은 0.696이고 p-values가 0.56으로 0.05보다 크므로 bmi2(반응변수)에 유의하지 않은 것으로 보인다. 반면 bmi1(공변량)의 검정통계량은 146.071이고 p-value가 0에 수렴하여 0.05보다 작아 bmi2에 유의한 것으로 판단된다.

따라서 인자인 agegroup의 효과는 존재하지 않고 공변량인 bmi1의 효과만 존재하여, 사후 체지방 지수인 bmi2의 증감을 연령대인 agegroup의 차이로 설명할 수 없다.

2) 추정된 더미회귀모형

1-1) 코드 및 결과

```
In [31]: # 더미회귀모형 분석요약
print(fit2.summary())
fit2.params.round(1)
```

OLS Regression Results						
Dep. Variable:	bmi2	R-squared:	0.695			
Model:	OLS	Adj. R-squared:	0.679			
Method:	Least Squares	F-statistic:	42.78			
Date:	Thu, 03 Nov 2022	Prob (F-statistic):	1.20e-18			
Time:	10:19:25	Log-Likelihood:	-166.32			
No. Observations:	80	AIC:	342.6			
Df Residuals:	75	BIC:	354.6			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.0536	1.868	2.170	0.033	0.333	7.775
C(agegroup)[T.50대미만]	0.8375	0.977	0.857	0.394	-1.109	2.784
C(agegroup)[T.60대]	0.9045	0.640	1.412	0.162	-0.371	2.180
C(agegroup)[T.70대이상]	0.5548	0.675	0.822	0.414	-0.790	1.900
bmi1	0.7932	0.066	12.086	0.000	0.662	0.924
Omnibus:	38.045	Durbin-Watson:	2.002			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	830.013			
Skew:	0.555	Prob(JB):	5.82e-181			
Kurtosis:	18.741	Cond. No.	229.			

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Out [31]: Intercept 4.1  
C(agegroup)[T.50대미만] 0.8  
C(agegroup)[T.60대] 0.9  
C(agegroup)[T.70대이상] 0.6  
bmi1 0.8  
dtype: float64

## 2-2) 결과 해석

위의 요약을 토대로 공분산 분석 모형의 결정계수가 0.695로 자료들을 대강 잘 나타내고 있음을 알 수 있다. 또한 통계량이 42.78이고, p-value가 0에 수렴하여 0.05보다 작다는 것을 통해, 모형도 유의함을 확인할 수 있다. 추가로 Durbin-Watson 통계량이 2.002로 2에 매우 가까운 것으로 보아, 회귀분석의 주요한 가정인 오차의 등분산성도 만족한다는 것을 알 수 있다.

따라서 추정된 bmi1, agegroup의 계수, 상수항을 토대로 한 더미회귀모형은 다음과 같다.

$$(\text{agegroup}-50\text{대 미만}) : y = 4.1 + 0.8 + 0.8 \cdot \text{bmi1}$$

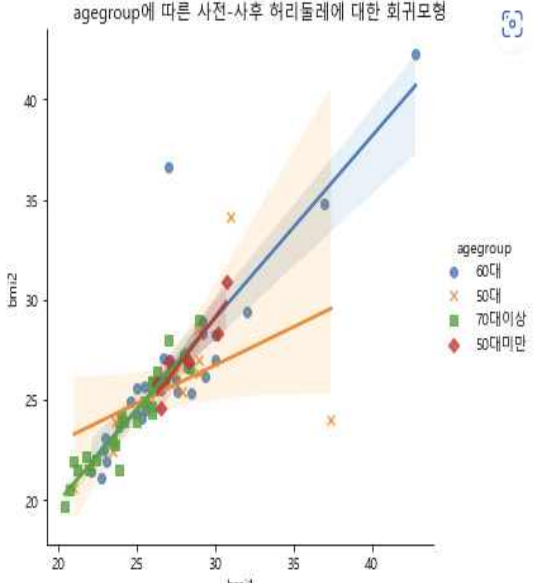
$$(\text{agegroup}-50\text{대}) : y = 4.1 + 0.8 \cdot \text{bmi1}$$

$$(\text{agegroup}-60\text{대}) : y = 4.1 + 0.9 + 0.8 \cdot \text{bmi1}$$

$$(\text{agegroup}-70\text{대 이상}) : y = 4.1 + 0.6 + 0.8 \cdot \text{bmi1}$$

그러나 위와 같은 더미 회귀모형식에서 요인 agegroup의 수준 50대미만, 60대, 70대이상의 효과는 p-value가 각각 0.394, 0.162, 0.414로 0.05보다 작으므로 유의하지 않음에 주의하여야 한다.

## (시험문제 제외) 공분산분석의 회귀모형 : 더미회귀모형의 시각화

1) 코드 및 결과	2) 결과 해석
<pre data-bbox="135 846 774 1041">In [32]: # 공분산분석의 회귀모형 : 더미회귀모형의 시각화(시험문제 제외) import matplotlib.pyplot as plt import seaborn as sns  sns.lmplot(x="bmi1", y="bmi2", hue="agegroup", data=desa, markers=["o", "x", "s", "D"]) plt.suptitle("agegroup에 따른 사전-사후 허리둘레에 대한 회귀모형", y=1.02) plt.show()</pre> 	<p>50대 미만, 60대, 70대의 회귀직선의 기울기가 비슷한 반면, 50대의 회귀직선의 기울기만 확연히 작은 것을 볼 수 있다.</p> <p>이는 50대 연령대에 존재하는 bmi1이 대략 37을 갖는 특이값의 영향으로 보여진다.</p> <p>이외에도 60대에서 1개의 특이값이 눈에 띄기는 하나 회귀직선의 기울기 변화에 영향을 미치지 않는 것으로 판단된다.</p>



4. 대사중후군 자료에서 두 범주형 자료인 agegroup과 건강의 연관성이 있는지를 검정하고 결과를 해석하시오. (22점)

1) 환자의 연령대(agegroup)과 환자의 건강(상태) 간의 연관성이 있는지를 교차분석표를 만들고, 가설검정 결과가 유의하다면 교차표를 근거로 결과를 해석하라. -10점

답안)

1) 연령대(agegroup)과 건강의 교차분석표 및 비율표

1-1) 코드 및 결과

```
In [60]: # 교차분석 : 독립성검정
# 크래머법칙(교차셀의 20%가 기대빈도 5이하의 경우 범주병합 후 검정)
# 연령대와 건강상태에 연관성 여부를 검정.
pd.crosstab(desa.agegroup, desa.건강, margins=True, margins_name='전체')
```

```
Out [60]:
```

건강	건강나쁨	건강매우나쁨	건강 좋음	보통	전체
agegroup					
50대	1	2	5	6	14
50대미만	1	0	3	2	6
60대	13	2	11	6	32
70대이상	11	5	5	7	28
전체	26	9	24	21	80

```
In [61]: # 비율표(propotion table) 작성
pd.crosstab(desa.agegroup, desa.건강, margins=True,
            margins_name='전체', normalize='all').round(4)
```

```
Out [61]:
```

건강	건강나쁨	건강매우나쁨	건강 좋음	보통	전체
agegroup					
50대	0.0125	0.0250	0.0625	0.0750	0.175
50대미만	0.0125	0.0000	0.0375	0.0250	0.075
60대	0.1625	0.0250	0.1375	0.0750	0.400
70대이상	0.1375	0.0625	0.0625	0.0875	0.350
전체	0.3250	0.1125	0.3000	0.2625	1.000

2) 연령대(agegroup)과 건강의 기대빈도표

2-1) 코드 및 결과

```
In [62]: # 검정결과 : 카이제곱통계량, p값, 자유도, 기대도수표
from scipy.stats import chi2_contingency
d_table=pd.crosstab(desa.agegroup, desa.건강, margins=True, margins_name='전체')
chi,p,df,expected=chi2_contingency(d_table)
expected_table=pd.DataFrame(data=expected, index=d_table.index,
                           columns=d_table.columns)
expected_table
```

Out [62]:

건강	건강나쁨	건강매우나쁨	건강 좋음	보통	전체
agegroup					
50대	4.55	1.575	4.2	3.675	14.0
50대미만	1.95	0.675	1.8	1.575	6.0
60대	10.40	3.600	9.6	8.400	32.0
70대이상	9.10	3.150	8.4	7.350	28.0
전체	26.00	9.000	24.0	21.000	80.0

```
In [63]: # 기대도수가 5미만인 샘플의 개수와 비율
sum(sum(expected<5))/16*100
```

Out [63]: 62.5

#### 2-2) 결과 해석

1)과 2)는 독립성검정에 검정통계량을 구하는 데 이용된다. 그런데 기대빈도표에서 5이하의 교차 셀이 62.5%로 전체 셀의 20%보다 매우 큰 것을 확인할 수 있다. 이러한 경우에 더 정확한 검정을 위해, 크래머법칙에 의거해 범주를 통합하여, 교차 셀의 기대빈도 수를 5보다 크게 변경해주는 것이 적절하다고 판단된다.

### 3) 연령대(agegroup)과 건강의 독립성 검정(카이제곱검정)

#### 3-1) 코드 및 결과

```
In [64]: # 카이제곱검정통계량과 유의확률
table=pd.crosstab(desa.agegroup, desa.건강)
print("chisquare=%.3f" % chi, "P-value=%.3f" % p)

chisquare=11.687 P-value=0.765
```

#### 3-2) 결과 해석

두 범주형 자료의 연관성을 알아보기 위해선 독립성 검정이 필요하다. 따라서 범주형 자료인 agegroup과 건강의 연관성을 알아보기 위한 가설은 다음과 같다.

$$H_0 : \text{agegroup과 건강은 서로 독립이다.}$$

따라서 두 자료에 대한 카이제곱검정 결과는, 검정통계량은 11.687이고 p-value가 0.765로 0.05보다 크므로 귀무가설을 기각할 수 없다. 따라서 두 범주형 자료 agegroup과 건강은 연관성이 없다(독립이다)고 할 수 있다.

그러나 위의 결과는 크래머의 법칙에 맞게 자료를 처리하지 않아, 적절한 통계적 판단으로 보기 어렵다.

### 2) 분석결과의 정확성에 문제가 있다면 확인하고, 해결 방법을 제시하고 정확성 있는 분석을 수행하여 결과를 해석하라. -12점

답안)

위의 1) 문제에서 확인하였듯이, 기대빈도표에서 5이하의 교차 셀이 62.5%로 전체 셀의 20%보다 매우 크기 때문에 크래머법칙에 의해 검정의 정확성에 문제가 생긴다. 따라서 적절하게 범주를 병합하여 기대빈도가 작은 셀의 수를 줄인 후에 독립성 검정을 하는 것이 적절하다고 판단한다.

## 1) 크래머법칙에 의한 범주통합

### 1-1) 코드 및 결과

```
In [65]: ▶ desa=pd.read_csv('대사중후군.csv',encoding='CP949')
label1={1:'건강매우 좋음', 2:'건강 좋음', 3:'보통', 4:'건강 나쁨', 5:'건강매우 나쁨'}
desa['건강']=desa['건강'].map(label1)
desa.loc[(desa.age<60), 'agegroup']='60대미만'
desa.loc[(desa.age>=60) & (desa.age<70), 'agegroup']='60대'
desa.loc[(desa.age>=70), 'agegroup']='70대이상'
```

## 2) 크래머 법칙에 의해 범주 통합 후, 연령대(agegroup)과 건강의 기대빈도표

### 2-1) 코드 및 결과

```
In [66]: ▶ # 통합후 기대도수표
pd.crosstab(desa.agegroup, desa.건강, margins=True, margins_name='전체')

d_table=pd.crosstab(desa.agegroup, desa.건강, margins=True, margins_name='전체')
chi,p,df,expected=chi2_contingency(d_table)
expected_table=pd.DataFrame(data=expected, index=d_table.index,
                             columns=d_table.columns)
expected_table
```

Out [66]:

건강	건강매우나쁨	건강 좋음	보통	전체
agegroup				
60대	10.467532	10.870130	9.662338	31.0
60대미만	6.415584	6.662338	5.922078	19.0
70대이상	9.116883	9.467532	8.415584	27.0
전체	26.000000	27.000000	24.000000	77.0

```
In [67]: ▶ # 기대도수가 5미만인 셀의 개수와 비율
sum(sum(expected<5))/9*100
```

Out [67]: 0.0

### 2-2) 결과 해석

기존의 agegroup의 50대 미만, 50대, 60대의 범주를 통합, 건강의 '건강나쁨', '건강매우나쁨'의 범주를 통합하면서, 변경된 기대 도수표에 5미만 셀의 수가 없어진 것을 확인할 수 있다.

## 3) 연령대(agegroup)과 건강의 독립성 검정(카이제곱검정)

### 3-1) 코드 및 결과

```
In [68]: ▶ # 카이제곱검정통계량과 유의확률
table=pd.crosstab(desa.agegroup, desa.건강)
print("chisquare=%.3f" % chi, "P-value=%.3f" % p)

chisquare=8.098 P-value=0.524
```

### 3-2) 결과 해석

문제 1)의 독립성검정과 동일하게 귀무가설 채택, 연령대와 건강은 연관성이 없다는 결론을 얻을 수 있었다. 크래머법칙에 의해 적절히 자료를 처리한 결과가 귀무가설의 기각여부를 바꾸진 못했지만, p-value를 변경하였다 것에 더 주목해야 할 것이다.