연구문제 Keller“ManagerialStatistics”9thedition

시나리오

> 여성 CEO는 은행이 대출 승인율, 대출이자 측면에서 남성 CEO와 차별하고 있다고 주장한다.

> 하여, 남성 CEO 기업 1,050개, 여성 CEO 기업 115개 기업을 무작위 층화 추출하여 (승인 받은 CEO의 대출 이자율, 기업형태, 매출액, 기업설립연수)를 조사하였다.

> http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/bank.csv

> 데이터를 이용하여 여성 CEO의 주장을 검정하시오.

```
import pandas as pd
df=pd.read_csv('http://wolfpack.hnu.ac.kr/Stat_Notes/example_data/bank.csv')
df.head(3)
```

	gender	rate	biz_type	sales	found_yr
0	Female	1.65	private	473	12
1	Female	0.89	private	515	9
2	Female	1.13	private	772	11

[두모집단 모비율 차이 추론]

모비율 점추정

```
df_table=df.gender.value_counts()
df_table/[1050,115]
```

> > 여성 대출 승인율 87.8%, 남성 대출 승인율 90.7%

모비율 차이 신뢰구간

> 모수 (p1-p2) : p1=여성 CEO 기업 승인율, p2=남성 CEO 기업 승인율
>
> > 표본크기 n1=표본 여성ceo기업 수(1050), n2=표본 남성ceo기업 수(110)
>
> 추정치 (p1hat-p2hat) : p1hat=승인여성ceo기업수/n1, p2hat=승인남성ceo기업수/n2
>
> 샘플링분포 : (p1hat-p2hat) ~N(p1-p2, sqrt(p1*(1-p1)/n1+p2*(1-p2)/n2))
>
> 상한 : (p1hat-p2hat)+ z(1-alpha/2)*sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2))
>
> 하한 : (p1hat-p2hat)- z(1-alpha/2)*sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2))

```
import numpy as np
import scipy.stats
z=scipy.stats.norm.ppf(0.95)  #상한 단측가설이므로 95% 하한 신뢰구간 구하기 위한 것임
n1=1050;p1hat=0.906667
n2=115;p2hat=0.878261
lb=(p1hat-p2hat)-z*np.sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2)
ub=(p1hat-p2hat)+z*np.sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2)
print('하한=%.3f, 상한=%.3f'%(lb,ub))
```

```
하한=-0.024, 상한=0.081
```

> 신뢰수준 90%한 이유는 95% 단측 신뢰구간을 구하기 위함임
>
> > 하한 신뢰구간이 0을 포함하고 있으므로 남자승인율=여자 승인율

[통계량 주어진 경우]세미소사와 맥과이어 홈런 경쟁으로 인하여 여성 팬이 증가하였다고 주장한다. 이를 알아보기 위하여 CNN/ USA이 다음 조사를 하였다. 1995년 1008명 여성 중 413이 팬이라고 대답했고, 홈런 경쟁이 있은 1998년에는 1082 여성 중 681명이 팬이라고 답하였다. 이를 이용하여 주장에 대해 답하시오.

> 귀무가설 : 1995년 여성 야구팬 비율 = 1998 여성 야구팬 비율

> 대립가설 : 1995년 여성 야구팬 비율 < 1998 여성 야구팬 비율

```
import numpy as np
import scipy.stats
z=scipy.stats.norm.ppf(0.975)
n1=1008;p1hat=413/1008
n2=1082;p2hat=681/1082
lb=(p1hat-p2hat)-z*np.sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2)
ub=(p1hat-p2hat)+z*np.sqrt(p1hat*(1-p1hat)/n1+p2hat*(1-p2hat)/n2)
print('하한=%.3f, 상한=%.3f'%(lb,ub))
```

```
하한=-0.261, 상한=-0.178
```

모비율 차이 가설검정

> 모수 : p1=여성 CEO 기업 승인율, p2=남성 CEO 기업 승인율

> 표본크기 n1=표본 여성ceo기업 수(1050), n2=표본 남성ceo기업 수(110)

> 추정치 : p1hat=승인여성ceo기업수/n1, p2hat=승인남성ceo기업수/n2

> 귀무가설 : 여성 CEO 기업 승인율과 남성 CEO 기업 승인율은 동일하다. p1=p2

> 대립가설 : 여성 CEO 승인율이 남성 CEO 기업 승인율보다 낮다. p1 < p2 (성차별)

```
import numpy as np
import statsmodels.api as sm
population1 = np.random.binomial(1,0.906667,1050)
population2 = np.random.binomial(1,0.878261,115)
sm.stats.ttest_ind(population1, population2)
```

```
(1.2330288954083315, 0.21781415272906937, 1163.0)
```

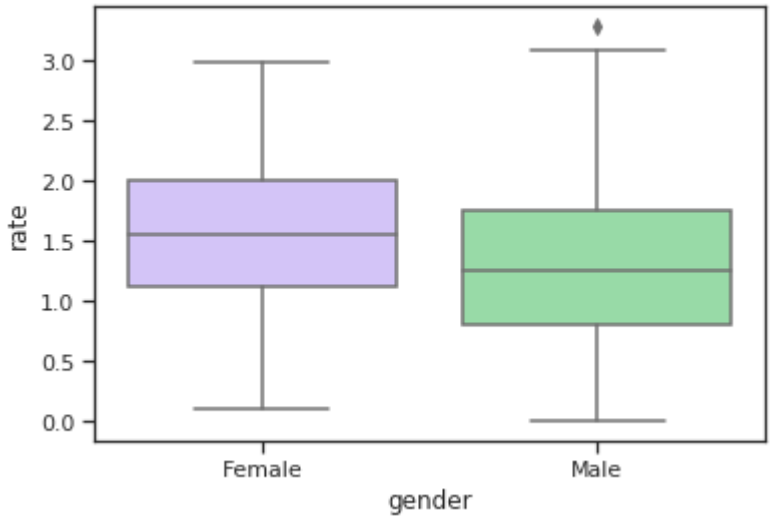
> > 유의확률 = 0.217/2 = 0.109(단측) 귀무가설 채택 : 여성, 남성 승인율 차이 없음 - 성차별 없음

[두 모집단 모평균 차이 추론]

[모평균 차이 그래프 요약]

```
1  import seaborn as sns
2  sns.set(style="ticks", palette="pastel")
3  sns.boxplot(x="gender", y="rate",palette=["m", "g"],data=df)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b78243d30>
```



> 남자 대출 이자율 상한 이상치 1개 존재

[이상치 진단 - 제외 필요]

```
1  df.shape
```

```
(1053, 5)
```

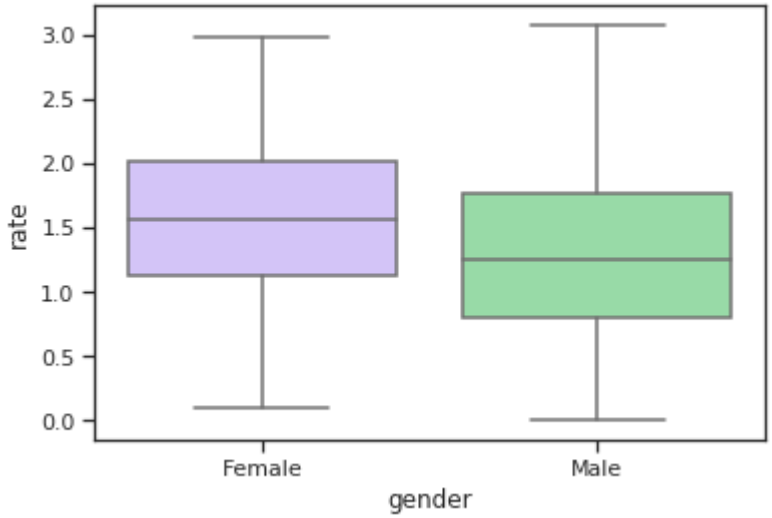
여자 데이터 모두, 남자 데이터 중 상한 이상치 1개 제외

```
indexNames=df[(df['gender']=='Male') & (df['rate']>3.1)].index
df_clean=df.drop(indexNames)
df_clean.shape
```

```
(1052, 5)
```

```
import seaborn as sns
sns.set(style="ticks", palette="pastel")
sns.boxplot(x="gender", y="rate",palette=["m", "g"],data=df_clean)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8b756c9c88>
```



집단별 평균, 표준편차 숫자요약

```
df_clean.groupby(['gender']).mean()
```

gender	rate	sales	found_yr
Female	1.545446	551.554455	9.237624
Male	1.275563	1183.388013	12.583596

```
df_clean.groupby(['gender']).std()
```

가설검정

귀무가설 : 남자기업 대출 이자 평균 = 여자기업 대출 이자 평균, mu1=mu2

대립가설 : 남자기업 대출 이자 평균이 여자기업 대출 이자 평균보다 낮다, mu1<mu2 (대출이자 측면 성차별)

두집단 모분산 동일성 검증

> 모평균 차이 검정 전에는 두 모집단 분산 차이 검정을 실시해야 한다.

> 귀무가설 : 두 집단의 모분산은 동일하다

```
from scipy import stats
scipy.stats.levene(df_clean.loc[df_clean.gender=='Male'].rate,df_clean.loc[df_clean.gender=='Female'].rate)
```

> 귀무가설 채택, 분산 동질성 귀무가설 채택 -> 아래 ttest_ind(equal_var=True) 사용

```
from scipy import stats
stats.ttest_ind(df_clean.loc[df_clean.gender=='Male'].rate,df_clean.loc[df_clean.gender=='Female'].rate,equal_var=True)
```

> 단측 가설이므로 유의확률(pvalue)/2 < 0.001 - 귀무가설 기각

> 남자 평균이자율 1.27%, 여자 평균이자율 =1.55% 여자 대출 이자율 유의적으로 높음 - 은행은 성 차별한다.

> 통계량이 주어진 경우 : scipy.stats.ttest_ind_from_stats(mean1, std1, nobs1, mean2, std2, nobs2, equal_var=True)

짝이룬 집단 평균 차이 검정 Keller“ManagerialStatistics”9thedition

[사례연구] MBA 전공 재무, 마케팅 연봉을 조사한 자료이다. 성적(GPA)에 따른 차이가 있을 가능성을 고려 하여 (4, 3.92)=1그룹, (3.92, 3,84)=2그룹2, ..., 총 25개 그룹에서 한 명씩 임의 추출하여 연봉을 조사하였다. 유의수준 5%에서 재무전공, 마케팅 전공의 연봉 차이가 있는지 검정하시오

```
import pandas as pd
df=pd.read_csv('http://wolfpack.hnu.ac.kr/Stat_Notes/elem_stat/Stat_methods/mba2.csv')
df.head(3)
```

숫자요약

```
df.describe()
```

이상치 진단

```
import seaborn as sns
sns.boxplot(df.Finance-df.Marketing)
```

가설검정

귀무가설 : 재무전공 평균 초봉 = 마케팅전공 평균초봉 <=> (x_i - y_i)=d_i 일집단 평균=0 검정과 동일

```
from scipy import stats
scipy.stats.ttest_rel(df.Finance,df.Marketing)
```

> 귀무가설 기각, 재무전공 평균연봉 6.54만불, 마케팅 전공 6.04만불로 재무 전공자의 초봉이 높다.

짝진 표본 비율 검증 : 동일한 개체로부터 이진형(성공, 실패) 변수를 서로 다른 기간(before - after)에 측정하여 프로그 램 효과가 있는지 알아보는 방법

Bland(2000)1319명어린이,12살에 독감에 걸릴 가능성은 나이가 14살이 되면 높아지는 지 낮아지는지 알아보기 위하여 조사한 결과 이다.

> 나이-14살 독감Yes | 독감No | 총합

12살

Yes : 212 | 144 | 356

No : 256 | 707 | 953

총합 : 468 | 851 | 1319

12살에 독감 걸린 사람 356명 중 14살에 212 걸림

> 만약 exact=True, 이항분포 사용

> false, chisquare distribution 분포 사용

```
print('12살 독감 확률=%.2f, 14살 독감 확률=%.2f'%(356/1319,468/1319))
```

```
from statsmodels.stats.contingency_tables import mcnemar
table = [[212, 144],[256, 707]]
# calculate mcnemar test
result = mcnemar(table,exact=False)
# summarize the finding
print('statistic=%.3f, p-value=%.3f' % (result.statistic, result.pvalue))
```

> [해석1] 귀무가설 기각, 12살에 비해 14살에 감기 걸릴 확률이 높아졌다.

> [해석2] 12살 감기 걸린 사람이 14살에 다시 걸릴 확률은 60%로 12살 미독감 사람이 14살에 걸릴 확률 27%보다 월등히 높다.

```
print('12살 독감 환자 14살 독감=%.2f, 12살 독감걸리지 않은 사람 14살 결릴 확률=%.2f'%(212/356,256/963))
```