t3_p2

```
import pandas as pd
import numpy as np
```

```
df= pd.read_csv("data/high_blood_pressure.csv")
```

```
df.head(2)
```

	Id	sex	age	bp_pre	bp_post
0	p001	Male	33	149	129
1	p002	Male	39	168	168

```
from scipy import stats
stats.ttest_rel(df['bp_post'],df['bp_pre']) # 평균의 차이가 있다 : alternative= two-sided  기본값
```

```
TtestResult(statistic=-3.0002163948827545, pvalue=0.0032868948457022056, df=119)
```

문제 2

-생존 여부와 성별 독립성 검정 (Titanic.csv)

타이타닉 탑승자 데이터(Titanic.csv)를 이용하여, 생존 여부(Survived)와 성별(Gender) 간 관계를 검정하시오.

조건:

```
범주형 변수 두 개: Survived (0=사망, 1=생존), Gender (male/female)

유의수준 0.05
```

지시사항:

```
교차표를 작성하시오.

카이제곱 독립성 검정을 수행하고, 검정통계량과 p값을 구하시오.
```

제출 형식:
카이제곱 통계량, p값을 소수 셋째 자리까지 반올림하여 제출하시오.
예: 10.327, 0.001

```
df= pd.read_csv('data/Titanic.csv')
```

```
df.head(2)
```

	PassengerId	Survived	Pclass	Name	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S

	PassengerId	Survived	Pclass	Name	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

```
survived= df['Survived']
gender= df['Gender']
```

```
survived.shape, gender.shape
tab= pd.DataFrame({"survived":survived,"gender":gender})
tab
```

	survived	gender
0	0	male
1	1	female
2	1	female
3	1	female
4	0	male
...	...	...
886	0	male
887	1	female
888	0	female
889	1	male
890	0	male

891 rows × 2 columns

```
tab.melt(id_vars='gender', value_vars=
```

```
contingency_table = pd.crosstab(df['Gender'], df['Survived'])
```

```
contingency_table
```

Survived	0	1
Gender		
female	81	233
male	468	109

```
# 카이제곱 독립성 검정을 수행하고, 검정통계량과 p값을 구하시오.
stats.chi2_contingency(contingency_table)
```

```
Chi2ContingencyResult(statistic=260.71702016732104, pvalue=1.1973570627755645e-58, dof=1,
expected_freq=array([[193.47474747, 120.52525253],
```

```
       [355.52525253, 221.47474747]]))
```

문제 3

게시물 유형에 따른 평균 반응 수 차이 검정 (ANOVA) (fb.csv)

- 페이스북 게시물 반응 데이터(fb.csv)를 이용하여, 게시물 유형에 따른 평균 리액션 수(reactions)가 유의미하게 차이가 있는지 검정하시오.
- 조건:

  그룹 변수: type (video, photo 등)

  종속 변수: reactions

  유의수준 0.05
- 지시사항:

  type별 reactions 평균을 비교하는 **일원분산분석(ANOVA)**를 수행하시오.

  F값과 p값을 구하시오.

제출 형식:
F값, p값을 소수 셋째 자리까지 반올림하여 제출하시오.
예: 5.241, 0.007

```
fb = pd.read_csv("data/4th-t2/fb.csv")
```

```
fb.head(2)
```

	id	type	reactions	comments	shares	likes	loves	wows	hahas	sads	angrys
0	1	video	529	512	262	432	92	3	1	1	0
1	2	photo	150	0	0	150	0	0	0	0	0

```
fb['type'].value_counts()
```

```
type
photo     4288
video     2334
status     365
link        63
Name: count, dtype: int64
```

```
video= fb[(fb['type']=='video')]['reactions']
photo= fb[(fb['type']=='photo')]['reactions']
link= fb[(fb['type']=='link')]['reactions']
status= fb[(fb['type']=='status')]['reactions']
```

```
stats.f_oneway(video,photo,link,status)
```

```
F_onewayResult(statistic=54.118579714158855, pvalue=1.41210615543239e-34)
```

- F = 54.119: 그룹 간 분산이 그룹 내 분산보다 54배 크다는 의미
- p값 ≈ 0.000 → 유의수준 0.05에서 귀무가설 기각 → type에 따라 reactions 평균이 유의하게 다름

문제 5.

생존 여부에 대한 로지스틱 회귀 해석 (Titanic.csv)

```
타이타닉 생존 예측을 위해, 로지스틱 회귀모형을 아래와 같이 설정하시오.

모형:

    종속 변수: Survived (0=사망, 1=생존)

    독립 변수: Gender, Pclass (1~3등급)

지시사항:

    로지스틱 회귀를 수행하시오 (statsmodels 사용).

    Gender 변수의 **오즈비(odds ratio)**와 p-value를 계산하시오.

제출 형식:
오즈비, p값을 소수 셋째 자리까지 반올림하여 제출하시오.
예: 2.615, 0.000
```

```
import pandas as pd
import numpy as np
titanic= pd.read_csv('data/Titanic.csv')
```

```
titanic.head(2)
```

	PassengerId	Survived	Pclass	Name	Gender	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C

```
import statsmodels.formula.api as smf
```

```
formula= "Survived ~ Gender + C(Pclass)"
model = smf.logit(formula,data=titanic).fit()
```

```
Optimization terminated successfully.
         Current function value: 0.464023
         Iterations 6
```

```
model.summary()
```

Logit Regression Results

Dep. Variable:	Survived	**No. Observations:**	891
Model:	Logit	**Df Residuals:**	887

Method:	MLE	Df Model:	3
Date:	Fri, 20 Jun 2025	Pseudo R-squ.:	0.3032
Time:	17:50:01	Log-Likelihood:	-413.44
converged:	True	LL-Null:	-593.33
Covariance Type:	nonrobust	LLR p-value:	1.145e-77

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.2971	0.219	10.490	0.000	1.868	2.726
Gender[T.male]	-2.6419	0.184	-14.350	0.000	-3.003	-2.281
C(Pclass)[T.2]	-0.8380	0.245	-3.424	0.001	-1.318	-0.358
C(Pclass)[T.3]	-1.9055	0.214	-8.898	0.000	-2.325	-1.486

```
model.llf
```

```
-413.44418477650765
```

```
model.params
```

```
Intercept         2.297123
Gender[T.male]   -2.641875
C(Pclass)[T.2]   -0.837952
C(Pclass)[T.3]   -1.905495
dtype: float64
```

```
gender = np.exp(model.params[1])
```

```
C:\Users\pjjeo\AppData\Local\Temp\ipykernel_46536\4108129114.py:1: FutureWarning: Series.__getitem__ treating keys as positions is deprecated. In a future version, integer keys will always be treated as labels (consistent with DataFrame behavior). To access a value by position, use `ser.iloc[pos]`
  gender = np.exp(model.params[1])
```

```
gender
```

```
0.07122756433670245
```

해석

Gender 변수의 회귀계수는 약 -2.642이며, 오즈비는 약 0.071이다.

따라서 유의수준 0.05 기준으로 p값이 작다면, 남성은 여성보다 생존 가능성이 유의하게 낮다고 판단할 수 있다.

변수	계수(β)	p값	해석
Intercept	`+2.297`	0.000	기준값(여성 & 1등석 기준 생존 로그오즈)
Gender[T.male]	`-2.642`	0.000	남성은 여성보다 생존 오즈비 0.071 (92.9%↓)
Pclass=2	`-0.838`	0.001	2등석은 1등석보다 생존 오즈비 약 `exp(-0.838)=0.433`
Pclass=3	`-1.906`	0.000	3등석은 1등석보다 생존 오즈비 약 `0.149` (85.1%↓)

문제 6 pearson

좋아요 수와 댓글 수 간의 Pearson 상관분석 (fb.csv)

```
fb.csv 데이터에서 게시물의 likes와 comments 간 상관관계가 존재하는지 분석하시오.

조건:

    연속형 변수 두 개: likes, comments

    유의수준 0.05

지시사항:

    Pearson 상관계수와 p값을 계산하시오.

    통계적으로 유의한지 판단하시오.

제출 형식:
상관계수, p값을 소수 셋째 자리까지 반올림하여 제출하시오.
예: 0.784, 0.000
```

```
fb = pd.read_csv("data/4th-t2/fb.csv")
fb.head(2)
```

	id	type	reactions	comments	shares	likes	loves	wows	hahas	sads	angrys
0	1	video	529	512	262	432	92	3	1	1	0
1	2	photo	150	0	0	150	0	0	0	0	0

```
likes= fb['likes']
comments=fb['comments']
```

```
from scipy import stats
```

```
stats.pearsonr(likes,comments)
```

```
PearsonRResult(statistic=0.10168703564349908, pvalue=1.1409104185135603e-17)
```

```
# dir(stats)
```

문제 7. 평균 점수에 대한 95% 신뢰구간 직접 계산

```
scores = [71, 74, 70, 69, 75, 77, 78, 73, 74, 72]
```

```
mu= np.mean(scores)
from scipy import stats
dof= len(scores)-1
```

```
t= stats.t.ppf(0.975,dof)
std=  np.std(scores,ddof=1)
SEM = std/np.sqrt(len(scores))

upper = mu + t*SEM
lower = mu - t*SEM

print(f"구간은 아래 {lower}에서 위에는 {upper} 사이에 포집단의 평균점수가 위치한다. 신뢰구간 95%에서")
```

```
구간은 아래 71.21985385732074에서 위에는 75.38014614267925 사이에 포집단의 평균점수가 위치한다. 신뢰구간 95%에서
```

```
stats.t.interval(0.95 , df=dof, loc= mu, scale = SEM)
```

```
(71.21985385732074, 75.38014614267925)
```

```
```