11장. 가설 검정			
학번	32233421	이름	이은지

※ 유의수준 0.05 로 검정한다. 검정의 전 과정과 검정 결과를 제시해야 함

1. iris 데이터셋에서 setosa 품종과 versicolor 품종의 꽃잎의 길이(Petal_Length)는 통계적으로 유의한 차이가 있는지 검정하시오

```
import pandas as pd
from scipy import stats
import matplotlib.pyplot as plt
df = pd.read_csv('iris.csv')
# 데이터 탐색
df.head()
df.groupby('Species').count() # 그룹별 표본 크기
df.groupby('Species').mean() # 그룹별 평균
df.groupby('Species').boxplot(grid=False)
plt.show()

group_1 = df.loc[df.Species=='setosa','Petal_Length']
group_2 = df.loc[df.Species=='versicolor','Petal_Length']
group_1
group_2
# 정규성 검정
stats.shapiro(group_1)
stats.shapiro(group_2)
# 등분산성 검정
stats.levene(group_1, group_2)
# 독립표본 T-검정
result = stats.ttest_ind(group_1, group_2, equal_var=True)
result
```

```
>>> stats.shapiro(group_2)
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
>>> stats.shapiro(group_2)
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
>>> result = stats.ttest_ind(group_1, group_2, equal_var=True)
>>> result
>>> stats.shapiro(group_2)
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
>>> result = stats.ttest_ind(group_1, group_2, equal_var=True)
>>> stats.shapiro(group_2)
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
>>> stats.shapiro(group_2)
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
ShapiroResult(statistic=np.float64(0.96600440254332), pvalue=np.float64(0.15847783815657573))
>>> stats.levene(group_1, group_2)
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
LeveneResult(statistic=np.float64(30.49950629474208), pvalue=np.float64(2.7443023022053677e-07))
>>> result = stats.ttest_ind(group_1, group_2, equal_var=True)
>>> result = stats.ttest_ind(group_1, group_2, equal_var=True)
>>> result
>>> result
TtestResult(statistic=np.float64(-39.492719391538095), pvalue=np.float64(5.404910513441677e-62), df=np.float64(98.0))
```

p-value가 약 5.40x10^62로 유의수준 0.05보다 작으므로 통계적으로 유의한 차이가 있다.

2. 다음은 두 지역의 정당 선호도를 조사한 분할 표이다. 지역에 따른 정당 선호도의 차이가 있는지 (지역과 정당 선호도는 관련이 있는지) 검정하시

	A정당	B정당
M지역	67	36
N지역	52	45

```python
from scipy import stats
M = [67,36]
N = [52,45]
# 카이제곱 검정
stats.chi2_contingency([M, N])
```

```
loat64(2.259215574336441), pvalue=np.float64(0.132821324245
92719), dof=1, expected_freq=array([[61.285, 41.715],
       [57.715, 39.285]]))
```

p-value가 약 0.13이다. 이는 유의 수준 0.05보다 크므로 귀무가설 채택한다,

→ 지역과 정당 선호도는 연관이 없다.