

03

일원 배치 분산분석 실습

[일원 배치 분산분석 실습 - 전처리]

00:28

#독립변수(X)는 target, 종속변수(y)는 sepal_width입니다

- 독립변수는 범주형, 종속변수는 수치형이죠!

함수사용예제

- Python 에 내장되어 있는 iris 데이터를 이용하여 종(Species)별로 꽃받침의 폭(Sepal.Width)의 평균이 같은지 혹은 차이가 있는지를 확인하기 위해 일원 배치 분산분석을 수행해보자
- 검정을 수행하기에 앞서 설정할 수 있는 가설은 아래와 같다.
- 귀무가설(H0): 세 가지 종에 대해 Sepal.Width의 평균은 모두 같다.
- 대립가설(H1): 적어도 하나의 종에 대한 Sepal.Width의 평균값에는 차이가 있다.

1. 분산분석

```
In [2]: import scipy.stats as stats
import pandas as pd

iris_data = pd.read_csv("../data/iris.csv")
iris_data.head(100)
```

```
Out[2]:
```

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...
95	5.7	3.0	4.2	1.2	Iris-versicolor

[일원 배치 분산분석 실습 - 검정 및 해석]

03:34

```
In [20]: target_list
Out[20]: array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

```
In [24]: setosa = iris_data[iris_data["target"]==target_list[0]]["sepal width"]
versicolor = iris_data[iris_data["target"]==target_list[1]]["sepal width"]
virginica = iris_data[iris_data["target"]==target_list[2]]["sepal width"]
```

```
In [25]: type(setosa)
Out[25]: pandas.core.series.Series
```

```
In [9]: ## 결과에서 첫번째 값은 관찰통계치이고, 두번째 값은 p-value입니다.
## 정규성 검정에서는 p-value가 유의수준 0.05보다 클 경우 표본이 정규분포를 따른다고 판단할 수 있습니다
##a = df_iris[(df_iris['target'] == 'setosa') & (df_iris['sepal length (cm)'] > 5.5)]

print(stats.shapiro(setosa))
print(stats.shapiro(versicolor))
print(stats.shapiro(virginica))

ShapiroResult(statistic=0.96991895471344, pvalue=0.20465604960918427)
ShapiroResult(statistic=0.9741390742636999, pvalue=0.33798978384094507)
ShapiroResult(statistic=0.9673910140991211, pvalue=0.1809043289230896)
```

```
In [10]: ## Shapiro test 결과 0.07>0.05 이므로 정규성을 만족함으로써 등분산 검정 시행
## 귀무가설 : 집단간 분산이 같다.
## 대립가설 : 적어도 두 집단간 분산이 다르다.

stats.levene(setosa, versicolor, virginica)
```

```
Out[10]: LeveneResult(statistic=0.8475222363405327, pvalue=0.5248269975064537)
```