

2022147034 박정현 기초통계 과제 Report

1. 데이터 개요 및 구조 확인

본 분석에서는 seaborn 라이브러리에서 제공하는 Iris 데이터셋을 사용하였다.

데이터는 총 150개의 관측치와 5개의 변수로 구성되어 있으며, 변수는 sepal_length, sepal_width, petal_length, petal_width, species로 이루어져 있다. head()와 info()를 통해 확인한 결과, 결측치는 존재하지 않았으며 모든 수치형 변수는 연속형 데이터임을 확인하였다.

2. 기술통계량 분석

Species별 Petal Length에 대한 기술통계량을 산출하였다. 각 종(setosa, versicolor, virginica)은 동일한 표본 수(각 50개)를 가지며, 평균 Petal Length는 setosa(1.462)가 가장 작고, versicolor(4.260)가 중간, virginica(5.552)가 가장 큰 값을 보였다. 표준편차 또한 virginica에서 가장 크게 나타나 종별 분포 차이가 존재함을 확인할 수 있었다.

3. 시각화 분석 (Boxplot)

Species별 Petal Length 분포를 Boxplot으로 시각화하였다. 시각적으로 setosa는 Petal Length가 매우 짧고 분산이 작았으며, virginica는 가장 긴 Petal Length와 비교적 넓은 분포를 보였다. versicolor는 두 종 사이에 위치하였고, 박스 간 중첩이 거의 없어 종별 차이가 명확하게 관찰되었다.

4. 정규성 검정 (Shapiro–Wilk Test)

Species별 Petal Length에 대해 Shapiro–Wilk 정규성 검정을 수행하였다. 귀무가설(H0): 해당 species의 Petal Length는 정규분포를 따른다. 대립가설(H1): 해당 species의 Petal Length는 정규분포를 따르지 않는다.

검정 결과, setosa($p=0.0548$), versicolor($p=0.1585$), virginica($p=0.1098$) 모두 유의수준 0.05보다 크게 나타나 귀무가설을 기각하지 못하였다. 따라서 각 species의 Petal Length는 정규성을 만족한다고 판단하였다.

5. 등분산성 검정 (Levene Test)

세 species 간 Petal Length 분산의 동일성을 검정하기 위해 Levene 검정을 수행하였다. 귀무가설(H0): 세 그룹의 분산은 동일하다. 대립가설(H1): 적어도 한 그룹의 분산은 다르다.

검정 결과 p-value가 0.000으로 나타나 귀무가설을 기각하였으며, 실제 데이터에서는 등분산성이 만족되지 않는 것으로 확인되었다. 다만 과제 지시에 따라, 이후 분석에서는 등분산성을 만족한다고 가정하였다.

6. ANOVA 가설 수립

귀무가설(H0): 세 species 간 Petal Length의 평균은 모두 같다. 대립가설(H1): 적어도 한 species의 Petal Length 평균은 다르다.

7. One-way ANOVA 결과

One-way ANOVA를 수행한 결과, F값은 1180.1612, p-value는 0.000으로 나타났다. 이에 따라 유의수준 0.05 기준에서 귀무가설을 기각할 수 있으며, species 간 Petal Length 평균에는 통계적으로 유의한 차이가 존재함을 확인하였다.

8. 사후검정 (Tukey HSD)

ANOVA 결과가 유의하므로 Tukey HSD 사후검정을 수행하였다. 검정 결과, setosa-versicolor, setosa-virginica, versicolor-virginica 모든 종 쌍에서 평균 차이가 통계적으로 유의하게 나타났다. 이는 세 종이 Petal Length 기준으로 명확히 구분되는 집단임을 의미한다.

9. 결과 요약 및 결론

Boxplot 시각화, ANOVA, Tukey HSD 사후검정 결과를 종합하면, setosa는 Petal Length가 통계적으로 유의하게 가장 짧고, virginica는 가장 길며, versicolor는 두 종 사이에 위치함을 확인하였다. 따라서 Petal Length는 Iris 종을 구분하는 데 매우 중요한 특성임을 알 수 있다.

10. 회귀 분석 결과

Petal Length를 타겟 변수로 설정하고, sepal_length, sepal_width, petal_width를 입력 변수로 하여 선형 회귀 모델을 구축하였다. Train/Test 데이터 분리를 통해 모델을 학습한 결과, MSE는 0.1300, R^2 값은 0.9603으로 나타나 모델의 설명력이 매우 우수함을 확인하였다.

회귀계수 분석 결과, petal_width(1.4675)가 Petal Length에 가장 큰 양의 영향을 미쳤으며, sepal_length(0.7228) 역시 양의 영향을 주는 변수로 나타났다. 반면 sepal_width(-0.6358)는 음의 영향을 주었으며, 상대적으로 영향력이 작았다.

최종 결론

본 과제에서는 정규성 및 등분산성 검정 결과를 확인한 후, 과제 지시에 따라 가정을 적용하여 평균 차이 검정과 회귀 분석을 수행하였다. 분석 결과, Iris 종별 Petal Length는 통계적으로 유의한 차이를 보였으며, Petal Length는 회귀 모델을 통해 높은 정확도로 예측 가능함을 확인하였다. 이는 Petal 관련 특성이 Iris 종 분석 및 예측 문제에서 핵심적인 역할을 수행함을 시사한다.