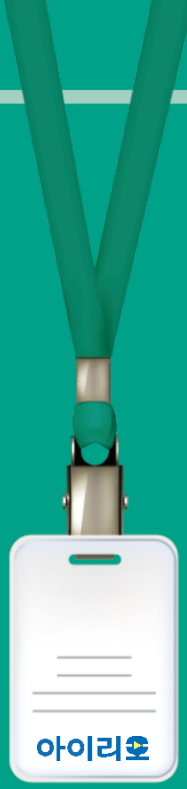


탐색적 데이터 분석 1강

- 데이터분석의 모든 것 -

강사명_이정인 저자

Contents

[01강] 학습 목표와 내용

학습 목표	탐색적 데이터 분석을 이해한다. 데이터의 대표값 탐색을 학습하고, R을 통해 실습한다.
학습 내용	mean(), median(), min(), max()
학습 자료	-
핵심 키워드	mean(), median(), min(), max()
참고 자료	데이터분석의 모든 것(출판 : 아이리포, 저자 : 이정인/장원중)

1. 데이터 대표값 탐색

1.1 평균값과 중앙값

1.2 절사평균

1.3 가중평균

1. 데이터 대표값 탐색

- 탐색적 자료 분석 : 존 튜키라는 통계학자가 창안한 것으로 가설 검정 등에 치우친 기존 통계학을 보완한 방법론
  - 평균, 중앙값, 분산, 표준편차, 사분위수 등의 기초 통계량을 활용하거나 그래프를 통한 시각화를 활용
  - 방대한 양의 데이터를 한눈에 볼 수 있도록 도표나 그래프로 시각화하면 즉각적인 상황 판단에 유리하고, 데이터를 기억하기 쉬우며, 사람들로부터 흥미를 유발시킴

- 신뢰성 있는 데이터 분석 수행 방법
  - 선입견 없이 객관적으로 진행되고 있는지 주의함
  - 분석 목표 수립 : 가설 수립, 트랜드 파악, 변수 간의 관계 파악
  - 데이터 분석 전, 데이터 제공처와 수집 과정의 신뢰성을 체크

1. 1 평균과 중앙값

- 데이터 요약을 통해서 데이터의 특징, 데이터 간 차이를 파악할 수 있으며, 대부분의 값이 어디쯤 위치하는지 추정
  - A기업과 B기업의 연봉 데이터가 있다. 어느 기업이 연봉이 높다고 말할 수 있을까?
    - 데이터 프레임 입력하기

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

```
A_salary   = pd.Series([25, 28, 50, 60, 30, 35, 40, 70, 40, 70, 40, 100, 30, 30 ])
B_salary   = pd.Series([20, 40, 25, 25, 35, 25 ,20, 10, 55, 65, 100, 100, 150, 300])
```

```
df = pd.DataFrame ( {'A_salary': A_salary,
                     'B_salary' :B_salary    }
)
```


1. 1 평균과 중앙값

◆ 평균값 - mean() : 두 기업 직원들의 평균 연봉을 비교하여 보자.

```
print("A 기업 연봉평균", A_salary.mean() , "  B 기업 연봉평균 ", B_salary.mean() )
```

```
A 기업 연봉평균 46.285714285714285   B 기업 연봉평균  69.28571428571429
```

```
A_salary.mean()
A_salary.sum()/A_salary.count()
```

```
46.285714285714285
```

```
df.mean(axis=0)  #열별 평균
```

```
A_salary    46.285714
B_salary    69.285714
dtype: float64
```

```
A_salary  = pd.Series([np.nan , 25, 28, 50, 60, 30, 35, 40, 70, 40, 70, 40, 100, 30, 30 ])
```

```
A_salary.mean( skipna=True )  #열별 평균 .  결측치 제외
```

```
46.285714285714285
```

1. 1 평균과 중앙값

◆ 중앙값 - median() : 중앙값을 구한다.

```
A_salary   = pd.Series([25, 28, 50, 60, 30, 35, 40, 70, 40, 70, 40, 100, 30, 30 ])
B_salary   = pd.Series([20, 40, 25, 25, 35, 25 ,20, 10, 55, 65, 100, 100, 150, 300])
```

```
print("A 기업 연봉 중앙값 ",  A_salary.median()  , "   B 기업 연봉 중앙값 ", B_salary.median()  )
```

```
A 기업 연봉 중앙값  40.0    B 기업 연봉 중앙값  37.5
```

```
df.median(axis=0)  #열별 중앙값
```

```
A_salary    40.0
B_salary    37.5
dtype: float64
```

- 평균 연봉은 B기업이 더 높고, 중앙값은 A기업이 더 높다.
- 과연 어느 기업이 연봉이 더 높다고 말할 수 있을까?
- 평균과 중앙값이 차이가 나는 이유가 무엇일까?

평균만으로는 대표값으로 충분하지 않은 경우가 많다

1. 2 절사평균

- 데이터 요약을 통해서 데이터의 특징, 데이터 간 차이를 파악할 수 있으며, 대부분의 값이 어디쯤 위치하는지 추정
  - 이상값에 민감한 평균의 특징을 보완한 것으로 절사평균

1. 3 가중평균

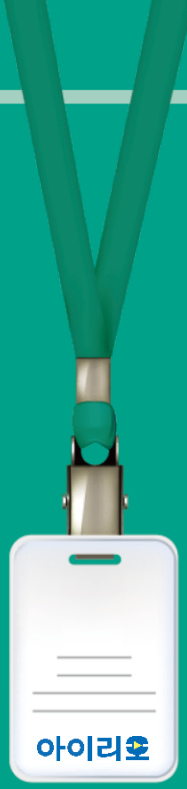
- 여러 모집단의 샘플이 똑같이 수집되지 않는 경우가 많음
- 이를 보정하기 위한 방법 → 데이터가 부족한 그룹에 더 높은 가중치를 적용
- 가중평균의 식 :

$$\overline{x_w} = \frac{\sum_{i=1}^{n} W_i x_i}{\sum_{i=1}^{n} W_i}$$

Summary

Summary

- mean() 함수로 평균을, median() 함수로 중앙값을 구할 수 있다.
- 최소값, 최대값을 각각 min() 함수, max() 함수로 구할 수 있다.



수고하셨습니다.