## 금융인을 위한

# 통계와 데이터 분석 입문

**기술통계량 이해하기**
**: 중심, 산포도**


한국금융연수원
KOREA BANKING INSTITUTE

# 학습 내용

1. 연속형/수치형 자료를 요약하는 기술통계량 (Descriptive Statistics)의 종류
   - 자료의 중심(center) : 평균, 중위수, 절사평균
   - 자료의 산포도(spread) : 표준편차, 분산, 범위, 사분위수 범위

2. (실습) 파이썬을 활용한 자료의 중심, 산포도 분석

# 자료의 중심(center)

## 평균(mean)

$x_i$ 들은 표본에 속한 자료라고 할 때, $\bar{x} = \frac{\sum_1^n x_i}{n}$

- 흔히 사용되는 산술평균은 모든 자료를 사용함
- 산술평균은 극단치(extreme value)에 영향을 받음

**예**

0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이
표본인 경우, 산술평균은 200

- 200을 자료의 중심이라 할 수 있을까?

# 자료의 중심(center)

## 중위수(median)

크기에 따라 재배열한 자료의 중간점

- 데이터의 순위에 관한 정보만을 이용함
- 극단치의 영향을 덜 받음(robust)
- 극단치 여부에 민감한 산술평균의 단점을 보완할 수 있음

**예**

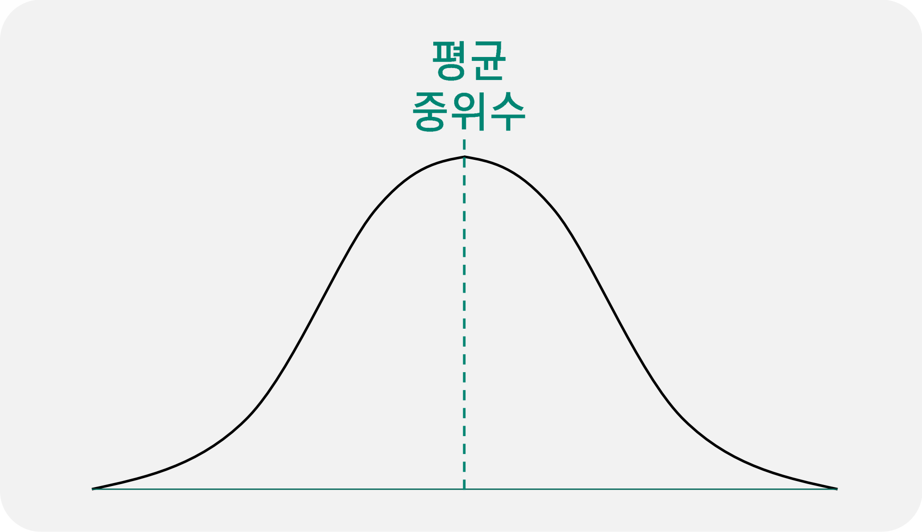
0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이
표본인 경우 중위수를 계산해보면?

- 표본의 크기가 10이므로 중위수는 크기 순서로 나열했을 때 중간, 즉 5번째, 6번째 위치한 값들의 평균인 20임

# 자료의 중심(center)

## 자료의 분포가 대칭일 때

평균 ≈ 중위수

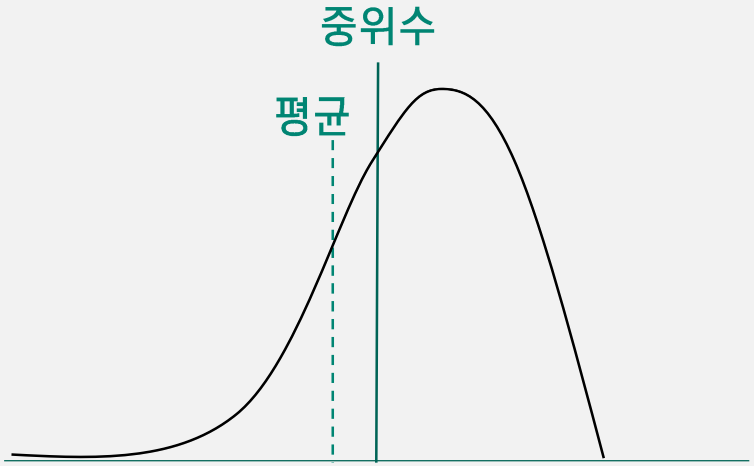


# 자료의 중심(center)

## 왼쪽으로 긴 꼬리가 있는 분포일 때

### 평균 < 중위수

- 극단치가 작은 쪽에 있는 경우 이런 분포를 가지게 됨

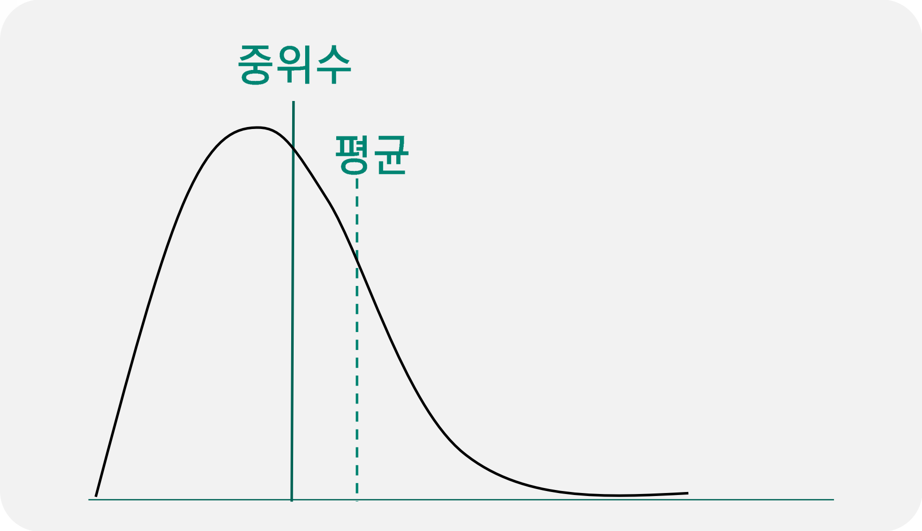


# 자료의 중심(center)

## 오른쪽으로 긴 꼬리가 있는 분포일 때

### 평균 > 중위수

- 극단치가 큰 쪽에 있는 경우 이런 분포를 가지게 됨



# 자료의 중심(center)

## 절사평균(trimmed mean)

자료를 크기 순으로 나열하여 크기가 작은 자료 일부와 큰 자료 일부를 제외하고 남은 자료들의 산술평균을 계산한 것

- 값이 큰 자료 10%와 작은 자료 10%를 제외한 나머지 자료들의 평균이 **10% 절사평균**임

예

0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이 표본인 경우, 10% 절사평균을 계산해보면?

- 표본의 크기가 10이므로 10% 해당하는 상위 1개, 하위 1개를 제외하고 8개의 평균이므로 절사평균은 25임

- 극단치가 존재하는 자료에서는 **중위수**와 **절사평균**이 산술평균보다 대표성을 갖음

# 자료의 산포도(spread)

## 범위(range)

- 최댓값에서 최솟값을 뺀 값 = 범위
- 자료의 흐트러짐 정도를 나타내는 가장 간단한 척도
- 극단치에 가장 **민감**하게 영향을 받음

예

0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이
표본인 경우, **범위**를 계산해보면?

- 최댓값 1800에서 최솟값 0을 뺀 **1800**임

# 자료의 산포도(spread)

## 분산(variance, $s^2$), 표준편차(standard deviation, $s$)

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$$

- 분산과 표준편차가 작을수록 자료가 평균에 근접하여 분포되어 있음
- 분산과 표준편차는 항상 0과 같거나 큰 값을 가짐
- 자료의 관찰값이 모두 동일할 때 분산과 표준편차는 0임
- 표준편차의 단위는 관찰값의 측정단위와 같지만
  분산의 단위는 관찰값 단위의 제곱임
- 평균으로부터의 거리를 재어 구하는 값이므로,
  평균과 마찬가지로 극단치들의 영향을 받음

# 자료의 산포도(spread)

## 분산(variance, $s^2$), 표준편차(standard deviation, $s$)

예

0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이
표본인 경우, 분산과 표준편차를 계산해보면?

| | x | $x-\bar{x}$ | $(x-\bar{x})^2$ | |
|---|---|---|---|---|
| 1 | 0 | -200 | 40000 | |
| 2 | 10 | -190 | 36100 | |
| 3 | 20 | -180 | 32400 | |
| 4 | 20 | -180 | 32400 | |
| 5 | 20 | -180 | 32400 | |
| 6 | 20 | -180 | 32400 | |
| 7 | 30 | -170 | 28900 | |
| 8 | 30 | -170 | 28900 | |
| 9 | 50 | -150 | 22500 | |
| 10 | 1800 | 1600 | 2560000 | |
| | | | | |
| $\bar{x}$ | 200 | | $\sum(x-\bar{x})^2=$ | 2846000 |
| | | | | |
| | | | 분산 = | 316222.2 |
| | | | 표준편차 = | 562.3364 |

# 자료의 산포도(spread)

## 사분위수 범위(interquartile range)

- $IQR = Q_3 - Q_1$
  - $Q_1$ = 1st quartile = 25th percentile
    : 자료를 크기 순으로 배열하였을 때 25%에 위치한 값(하위 25%)
  - $Q_2$ = median
  - $Q_3$ = 3rd quartile = 75th percentile
    : 자료를 크기 순으로 배열하였을 때 75%에 위치한 값(상위 25%)
  - 사분위수 범위는 중위수를 포함하는 가운데 50% 자료의 범위를 의미

# 자료의 산포도(spread)

## 사분위수 범위(interquartile range)

- 극단치의 영향을 덜 받으므로 다른 산포도를 나타내는 기술통계량을 보완할 수 있음

예

0, 10, 20, 20, 20, 20, 30, 30, 50, 1800 이 표본인 경우, **사분위수 범위**를 계산해보면?

- Q1이 20, Q3이 30으로 사분위수 범위는 10임

# 사분위수 범위를 활용한 이상치 감지

**이상치(outlier)**

자료에서 정상 범주를 **벗어나는** 것들

- [Q1 - 1.5 IQR, Q3 + 1.5 IQR] 범위가 이상치 판단 기준이 될 수 있음
  - 이 범위를 벗어나면 이상치라고 할 수 있음
- 이전 표본에서 **범위**를 계산해보면 [20 - 1.5 × 10, 30 + 1.5 × 10] = [5, 45]
  - 따라서 이 범위를 벗어나는 50, 1800은 이상치라 볼 수 있음

# 학습정리

- 기술통계량이란 데이터의 전반적인 특성을 요약해주는 값임

- 연속형/수치형 자료의 중심을 요약해주는 값으로 평균, 중위수, 절사평균이 있음
  > 중위수와 절사평균은 평균에 비해서 극단치의 영향을 덜 받으므로 평균을 보완할 수 있음

- 연속형/수치형 자료의 산포도를 요약해주는 값으로 범위, 표준편차(분산), 사분위수범위가 있음
  > 산포도는 그 값 자체 보다는 비슷한 그룹과의 비교를 통해서 크고 작음을 판단할 수 있음
  > 범위, 표준편차(분산), 사분위수범위는 각각 산포도를 재는 방식이 다르고 극단치가 미치는 영향이 다르므로 서로 보완해줄 수 있음