



# 인공지능수학

## (확률과 통계1)

# 학습내용

---

## ■ 학습할 내용

1. 모집단과 표본집단의 평균
2. 산포도

# 1. 평균, 분산, 확률

## ■ 각 개념의 의미

- ➡ 평균 은 산술평균 또는 기대값
- ➡ 분산  $\sigma^2$  은 평균  $m$  으로 부터 평균 제곱거리를 측정
- ➡  $n$  개의 서로 다른 사건의 확률은 그 합이 1이 되는 양수

## ■ 집단

### ➡ 모집단

- ◆ 연구 또는 관찰의 대상이 되는 총체
- ◆ 연구자가 알고 싶어 하는 대상 그 자체
- ◆ 집단 전부

### ➡ 표본집단

- ◆ 모집단에서 추출한 일부 (부분집합)
- ◆ 모집단의 특성을 반영할 수 있는 표본 추출이 중요

# 산술평균

## ★ 표본평균(sample mean)

- 변량  $X$ 에 대한  $n$  개의 자료가  $x_1, x_2, \dots, x_n$  으로 주어질 때, 변량  $X$ 의 산술평균을 **표본평균**이라 하고 다음과 같이 정의된다.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} (x_1 + x_2 + \dots + x_n)$$

## ★ 모평균(population mean)

- 변량  $X$ 가 모집단에서 얻은 관측 값이  $x_1, x_2, \dots, x_N$  으로 주어질 때, 변량  $X$ 의 산술평균을 **모평균**이라 하고 다음과 같이 정의된다.

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i = \frac{1}{N} (x_1 + x_2 + \dots + x_N)$$

# 산술평균

## ★ 가중산술평균

- 변량  $X$ 의 자료가 범주형 도수분포표로 주어질 때,
- 전체 도수를  $n$ ,  $i$ 번째 범주에 속한 도수를  $f_i$  라 할 때 다음과 같이 계산한다.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i = \frac{1}{n} (f_1 x_1 + f_2 x_2 + \cdots + f_k x_k)$$

$$(k : \text{범주의 수}, \sum_{i=1}^k f_i = n)$$

# 산술평균

## 예제 1

다음 범주형 도수분포표를 보고 표본평균을 구하라.  
이때의 표본평균은 가중산술평균이 된다.

변량( $x_i$ )	도수( $f_i$ )
5	5
15	8
25	3
35	2
45	2
합계	20

## 풀이

범주형 도수분포표로 정리된 경우 표본평균  $\bar{x}$  는 가중산술평균으로 구한다. 이때  $k=5$  이고  $n = \sum_{i=1}^k f_i = \sum_{i=1}^5 f_i = 20$  이므로 표본평균을 구하면 다음과 같다.

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^5 f_i x_i = \frac{1}{n} (f_1 x_1 + f_2 x_2 + f_3 x_3 + f_4 x_4 + f_5 x_5) \\ &= \frac{1}{20} (5 \times 5 + 8 \times 15 + 3 \times 25 + 2 \times 35 + 2 \times 45) = 19\end{aligned}$$

# 산술평균

## 예제 2

다음은 예제는 청소년의 일주일 동안  
스마트폰 사용 시간에 대한 계급형  
도수분포표의 일부이다.  
이 도수분포표를 이용하여  
표본평균을 구하라.

계급(시간)	계급값( $m_i$ )	도수( $f_i$ )
10 <sup>이상</sup> ~ 17 <sup>미만</sup>	13.5	7
17 <sup>이상</sup> ~ 24 <sup>미만</sup>	20.5	24
24 <sup>이상</sup> ~ 31 <sup>미만</sup>	27.5	13
31 <sup>이상</sup> ~ 38 <sup>미만</sup>	34.5	4
38 <sup>이상</sup> ~ 45 <sup>미만</sup>	41.5	1
45 <sup>이상</sup> ~ 52 <sup>미만</sup>	48.5	1
합계	—	50

$$n = \sum_{i=1}^6 f_i = 50$$

$$\begin{aligned}\bar{x} &= \sum_{i=1}^6 f_i x_i = \frac{1}{n} (f_1 x_1 + f_2 x_2 + f_3 x_3 + f_4 x_4 + f_5 x_5 + f_6 x_6) \\ &= \frac{1}{50} (13.5 \times 7 + 20.5 \times 24 + 27.5 \times 13 + 34.5 \times 4 + 41.5 \times 1 + 48.5 \times 1) = 23.44\end{aligned}$$



# 산술평균

## ★ 산술평균의 성질

(1) 산술평균에 대한 편차<sup>5</sup>의 합은 0 이다.

$$\sum_{i=1}^n (x_i - \bar{x}) = (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) = 0$$

(2) 산술평균은 편차의 제곱의 합을 최소로 한다. 즉, 산술평균에 대한 편차의 제곱의 합은 임의의 수에 대한 편차의 제곱의 합보다 크지 않다.

$$\sum_{i=1}^n (x_i - \bar{x})^2 \leq \sum_{i=1}^n (x_i - a)^2 \quad (\text{단, } a \text{는 상수})$$

(3) 산술평균은 주어진 자료를 모두 사용하므로 정보 손실이 없고, 특히 표본들의 평균인 표본평균은 모집단을 추론할 때 유용하게 사용된다.

(4) 산술평균은 양적자료에 대해서만 구할 수 있으며, 대다수의 자료와 멀리 떨어져 있는 값인 극단값(outlier)에 매우 민감하게 작용한다(극단값은 이상점이라고도 한다).

# 중앙값 (Median)

★ 변량  $X$ 의  $n$  개의 자료  $x_1, x_2, \dots, x_n$  을 작은 값부터 크기 순으로 배열했을 때, 한가운데에 위치한 값을 **중앙값(median)** 또는 **중위수** 라고  $Me$ ,

로 나타낸다.  
[표 1-5] 중앙값 계산

$n$	중앙값	설명
홀수	$Me = x\left(\frac{n+1}{2}\right)$	$\frac{n+1}{2}$ 번째 값
짝수	$Me = \frac{x\left(\frac{n}{2}\right) + x\left(\frac{n}{2} + 1\right)}{2}$	$\frac{n}{2}$ 번째 값과 $\frac{n}{2} + 1$ 번째 값의 평균

- 중앙값은 다음과 같이 **편차의 절댓값의 합**을 최소로 하는 성질이 있다.

$$\sum_{i=1}^n |x_i - Me| \leq \sum_{i=1}^n |x_i - a| \quad (\text{단, } a \text{ 는 상수})$$

# 중앙값

## 예제 3

다음 표는 A공장의 작업자 5명이 하루 동안 제작하는 물품 수량을 나타낸다.  
이때 중앙값을 구하라.

작업자	1	2	3	4	5	6
물품 수량	250	285	230	265	290	800

## 풀이

먼저 자료(물품 수량)를 크기순으로 배열하면 230, 250, 265, 285, 290, 800이다.  $n=6$ , 즉 자료의 수가 짝수이므로 3번째와 4번째 자료의 값의 평균이 중앙값이다. 따라서 중앙값은 다음과 같이 구할 수 있다.

$$Me = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(265 + 285) = \frac{550}{2} = 275$$

# 중앙값

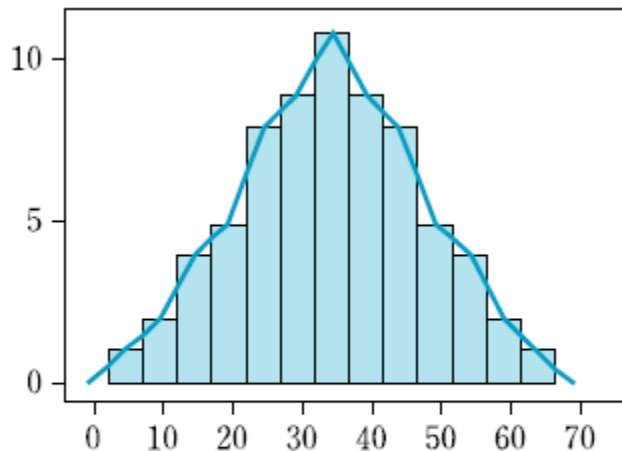
- ★ [예제 3] 에서 극단값 800을 제외하고 중앙값을 구하면  $n = 5$ , 즉 자료의 수가 홀수이므로 3번째 자료의 값이 중앙값이다.

따라서 중앙값은 다음과 같이 구할 수 있다.

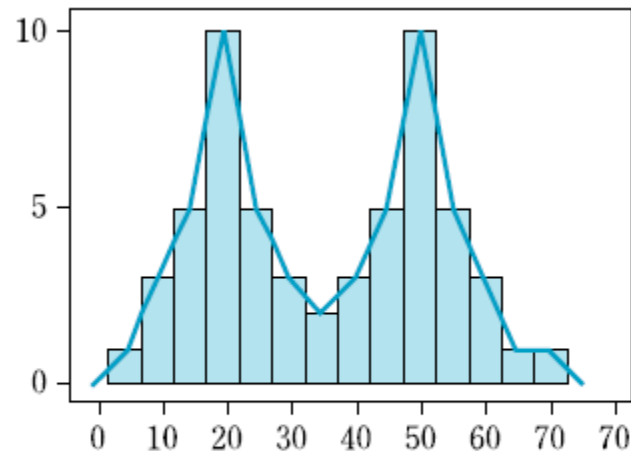
$$Me = x_3 = 265$$

# 최빈값

- ★ 변량  $X$ 의 자료 중에서 가장 많이 나타나는 값을 **최빈값 (mode)**이라 하고,  $Mo$ 로 나타낸다. 즉, 빈도수가 가장 많은 값을 나타낸다.



(a) 단봉형



(b) 쌍봉형

[그림 1-11] 도수분포곡선에서의 최빈값

- ✱ 값이 하나로 정해지는 평균이나 중앙값과는 달리 **최빈값은 자료에 따라 두 개 이상일 수도 있고, 없을 수도 있다.**

# 최빈값

## 예제 4

다음 50명의 통계학 성적에 대해 표본평균, 중앙값, 최빈값을 각각 구하라.

(단위 : 점)

83	90	60	25	50	94	60	62	97	43	67	84	79
62	78	48	85	52	77	90	25	84	41	65	58	75
83	71	74	68	89	88	76	69	77	89	73	98	77
58	77	69	75	69	65	67	69	79	85	45		

♣ 표본평균 =

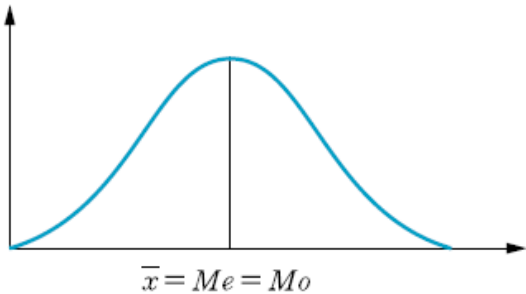
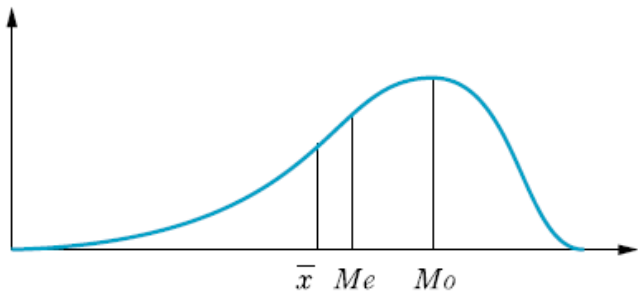
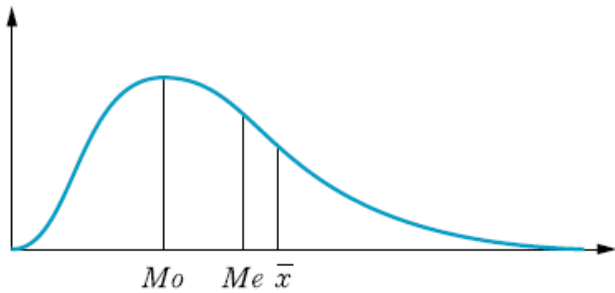
♣ 최빈값 =

♣ 중앙값 =

# 산술평균, 중앙값, 최빈값 사이의 관계

★ 피어슨의 실험 공식:  $\bar{x} - Mo \approx 3(\bar{x} - Me)$

[표 1] 도수분포곡선 모양에 따른 산술평균, 중앙값, 최빈값 사이의 관계

분류	도수분포곡선	관계
도수분포가 완전히 대칭인 경우		$\bar{x} = Me = Mo$
도수분포가 오른쪽으로 치우친 경우		$\bar{x} < Me < Mo$
도수분포가 왼쪽으로 치우친 경우		$Mo < Me < \bar{x}$

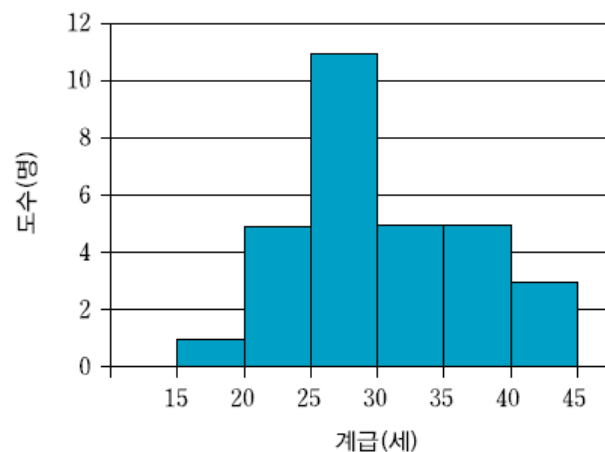
# 산술평균, 중앙값, 최빈값 사이의 관계

## 예제 5

벤처기업에 근무하는 직원 30명의 연령 자료에 대한 산술평균, 중앙값, 최빈값 사이의 관계를 설명하라.

### 풀이

먼저 직원의 연령에 대한 히스토그램을 그리면 다음과 같다.



이 히스토그램을 보면 약간 왼쪽으로 치우쳤음을 확인할 수 있다. 이 자료의 산술평균, 중앙값, 최빈값은 각각  $\bar{x} = 29.5$ ,  $Me = 28.5$ ,  $Mo = 26$ 이다. 따라서  $Mo < Me < \bar{x}$ 이고 피어슨의 실험 공식을 어느 정도 따른다고 볼 수 있다.

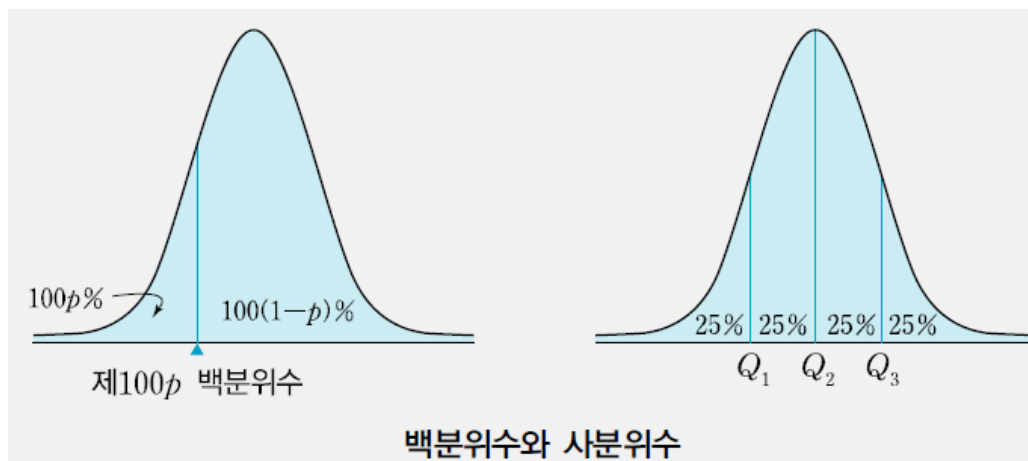


# 백분위수와 사분위수

## ★ 제 $100p$ 백분위수

- 변량  $X$ 의  $n$  개의 자료를 작은 값부터 크기 순으로 배열했을 때,  $0 \leq p \leq 1$  에 대하여 전체 자료를  $100p\%$  와  $100(1-p)\%$  로 나누는 값을 **제  $100p$  백분위수( $100p$ th percentile)**라 한다.

- 1 자료를 작은 값부터 크기순으로 배열한다.
- 2 자료 수  $n$ 에  $p$ 를 곱하여 다음과 같은 기준으로 제 $100p$  백분위수를 결정한다.
  - $np$ 가 정수이면,  $np$ 번째로 큰 자료와  $(np+1)$ 번째로 큰 자료의 평균을 택한다.
  - $np$ 가 정수가 아니면,  $np$ 의 정수 부분에 1을 더한 값  $m$ 을 구하고  $m$ 번째로 큰 자료를 택한다. 자료와 멀리 떨어진 값인 극단값에 매우 민감하게 작용한다.



# 백분위수와 사분위수

특히 제25, 50, 75 백분위수는 자료를 4등분하는 위치에 있는 값으로, 이 값을 **사분위수** (quartile)라고 한다. 이를 각각  $Q_1$ ,  $Q_2$ ,  $Q_3$ 로 표시하며,  $Q_1$ 을 제1사분위수,  $Q_2$ 를 제2사분위수(중앙값),  $Q_3$ 를 제3사분위수라고 한다.

## 예제 6

다음 자료에서 제50 백분위수와 제25 백분위수를 구하라.

16, 25, 4, 18, 11, 13, 20, 8, 11, 9

# 절사평균

★ **절사평균(trimmed mean)**은 평균의 장점과 중앙값의 장점을 모두 고려한 대푯값으로 극단값을 제외하고 구한 평균이다.

- 1 자료를 작은 값부터 크기순으로 배열한다.
- 2  $0 \leq \alpha \leq 0.5$ 인  $\alpha$ 에 대하여 자료 수  $n$ 에  $\alpha$ 를 곱하여 다음과 같은 기준으로 자료 수를 제거한다.
  - $\alpha n$ 이 정수이면, 이 정수에 해당하는 자료 수만큼 양 끝에서 제거한다.
  - $\alpha n$ 이 정수가 아니면,  $\alpha n$ 을 넘지 않는 최대 정수에 해당하는 자료 수만큼 양 끝에서 제거한다.
- 3 제거하고 남은 자료에 대하여 산술평균을 구한다.

- 절사평균을 계산하려면 **절사비율(%)**을 결정해야 하는데, 절사비율은 전체 데이터 개수에 대하여 상위 몇 퍼센트의 데이터와 하위 몇 퍼센트의 데이터를 배제할 것인가로 결정한다.

# 절사평균

## 예제 7

다음 자료에서 15% 절사평균을 구하라.

68, 70, 67, 10, 72, 68, 70, 71

**풀이** 자료를 다음과 같이 작은 값부터 크기순으로 배열한다.

10, 67, 68, 68, 70, 70, 71, 72

자료 수가  $n=8$ ,  $\alpha=0.15$ 이므로  $\alpha n=0.15 \times 8=1.2$ 이다. 이때  $\alpha n$ 이 정수가 아니므로, 1.2를 넘지 않는 최대 정수에 해당하는 1개의 자료를 각각 양 끝에서 제거한다.

67, 68, 68, 70, 70, 71

이들의 산술평균을 구하면  $\frac{67+68+68+70+70+71}{6}=69$  이므로, 주어진 자료에 대한 15% 절사평균은 69이다.

## ■ 산포도 : 자료의 흩어진 정도

# 범위

- ★ 변량  $X$ 의 자료가  $x_1, x_2, \dots, x_n$ 일 때,  $X$ 의 범위(range)는 이들 자료의 최댓값( $x_{\max}$ )과 최솟값( $x_{\min}$ )의 차를 의미하며, 보통  $R$ 로 표기한다.

$$R = x_{\max} - x_{\min}$$

## 예제 8

다음 자료의 범위를 구하라.

2, 3, 7, 12, 10, 14, 14, 9, 6

# 사분위수 범위

- ★ 범위는 자료의 두 극단값의 차이만을 나타내기 때문에 자료의 산포를 나타내기에 불충분하다.

이러한 단점을 일부 보완한 산포도가

**사분위수 범위(interquartile range, IQR)**이다.

- ★ 사분위수 범위는 다음과 같이 제3 사분위수와 제1 사분위수의 차이로 정의된다.

$$\text{IQR(사분위수 범위)} = Q_3 - Q_1$$

# 사분위수 범위

## 예제 9

[예제 8]를 참고하여, 다음 자료에 대한 제1 사분위수  $Q_1$ , 제3 사분위수  $Q_3$ , 사분위수 범위를 각각 구하라.

16, 25, 4, 18, 11, 13, 20, 8, 11, 9



# 분산과 표준편차

★ 평균을 중심으로 각 변량이 흩어진 정도를 알기 위하여 각 편차의 제곱의 합을 변량의 개수로 나눈 값, 즉 편차의 제곱의 평균인 **분산(variance)**을 이용한다.

- 모집단의 분산인 **모분산**은  $\sigma^2$  으로 나타내며, 다음과 같이 정의한다.  
여기에서  $\mu$  는 모평균이다

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

- 또한 변량  $X$ 의 자료가  $n$  개의 원소  $x_1, x_2, \dots, x_n$  으로 이루어진 모집단의 한 표본일 때, **표본분산**은  $S^2$  으로 나타내며, 다음과 같이 정의한다.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

# 분산과 표준편차

- $X$ 의 자료가 도수분포표로 주어질 때의 표본분산은 다음과 같이 구한다.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2, \quad \sum_{i=1}^k f_i = n$$

( $k$  : 계급의 수,  $f_i$  :  $i$ 번째 계급에 속한 도수,  $\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i$  : 표본평균)

## ★ 모분산과 표본분산의 간편식

[표 1-8] 모분산과 표본분산의 간편식

	모분산	표본분산
정의식	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$	$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$
간편식	$\sigma^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \mu^2$	$S^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$

# 분산과 표준편차

★ 분산의 양의 제곱근을 **표준편차(standard deviation)**라고 한다.

- 모분산의 양의 제곱근인  $\sigma$  를 **모표준편차(population standard deviation)**,  $S$  를 **표본표준편차(sample standard deviation)**라 하는데, 각각 다음과 같이 정의한다.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}, \quad S = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- $X$ 의 자료가 도수분포표로 주어질 때의 표준편차는 다음과 같다.

$$S = \sqrt{\frac{1}{n-1} \sum_{i=1}^k f_i (x_i - \bar{x})^2}, \quad \sum_{i=1}^k f_i = n$$

( $k$  : 계급의 수,  $f_i$  :  $i$  번째 계급에 속한 도수,  $\bar{x} = \frac{1}{n} \sum_{i=1}^k f_i x_i$  : 표본평균)

# 분산과 표준편차

## 예제 10

다음 자료에 대하여 물음에 답하라.

10, 11, 12, 13, 12, 14, 13, 11, 13, 12, 12, 11

- (a) 범위를 구하라.
- (b) 사분위수 범위를 구하라.
- (c) 표본분산과 표준편차를 각각 구하라.

# 변동계수

- ★ 측정 단위가 동일하지만 평균이 큰 차이로 다른 두 자료 집단 또는 측정 단위가 서로 다른 두 자료집단에 대한 산포도의 척도로 사용하는 것은 바람직하지 않다.

이러한 경우에 평균을 중심으로 상대적으로 흩어진 정도를 측정하는 척도를 사용하는데, 이를 **변동계수(coefficient of variation)**라 하고 보통 *CV*로 표기하며 다음과 같이 정의한다.

$$CV = \frac{\text{표준편차}}{\text{평균}} \times 100(\%)$$

# 변동계수

## 예제 11

다음 표는 저소득층과 고소득층의 하루 일당에 대한 변동계수를 구하고, 상대적으로 두 자료 집단의 흩어진 정도를 분석하라.

(단위 : 천 원)

저소득층	11.5	12.2	12.0	12.4	13.6	10.5
고소득층	171	164	167	156	159	164

# 5점 요약 표시

- ★ 주어진 자료의 중앙값  $Me$  , 제 1 사분위수  $Q_1$  , 제 3 사분위수  $Q_3$  , 최댓값  $x_{\max}$  , 최솟값  $x_{\min}$  을 구하여, 다음과 같이 5개의 통계량 조합으로 나타내는 방법을 **5점 요약 표시(5-number summary)**라고 한다.

$$[x_{\min}, Q_1, Me, Q_3, x_{\max}]$$

# 5점 요약 표시

## 예제 12

다음 자료를 5점 요약 표시로 나타내라.

60, 64, 72, 80, 92, 64, 68, 72, 76, 80, 84, 84, 76, 88, 88, 92, 96, 88, 92, 76



# 왜도와 첨도

## ★ 왜도(skewness)(또는 비대칭도)

- 분포의 대칭이나 비대칭의 정도를 표시하는 척도

$$\alpha = \frac{\sum_{i=1}^n \{(x_i - \bar{x})/S\}^3}{n-1} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^3}{S^3} = \frac{\mu_3}{S^3}$$

( $n$  : 표본의 수,  $S$  : 표본표준편차,

$\bar{x}$  : 표본평균,  $\mu_k = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^k$  :  $k$ 차 적률<sup>7)</sup>)

- 이때  $\alpha$  값에 따라 분포 형태를 알 수 있으며,  $\alpha$ 의 절댓값이 클수록 비대칭 정도가 심하다.
  - $\alpha = 0$ 이면 대칭분포이다.
  - $\alpha > 0$ 이면 왼쪽으로 치우친 분포이다.
  - $\alpha < 0$ 이면 오른쪽으로 치우친 분포이다.

# 왜도와 첨도

## ★ 첨도(kurtosis)

- 뾰족함의 정도를 나타내는 척도

$$\beta = \frac{\sum_{i=1}^n \{(x_i - \bar{x})/S\}^4}{n-1} = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4}{S^4} = \frac{\mu_4}{S^4}$$

( $n$  : 표본의 수,  $S$  : 표본표준편차,  $\bar{x}$  : 표본평균,  $\mu_4$  : 4차 적률)

- 이때  $\beta$  값에 따라 분포 형태를 알 수 있다.
  - $\beta = 3$ 이면 뾰족한 정도가 표준정규분포와 같다.
  - $\beta > 3$ 이면 표준정규분포 보다 정점이 높고 뾰족하다.
  - $\beta < 3$ 이면 표준정규분포보다 정점이 낮고 완만하다.

# 왜도와 첨도

## 예제 13

다음 자료에 대하여 왜도와 첨도를 각각 구하고, 이를 통해 자료의 분포 형태를 파악하라.

1, 3, 2, 0, 1, 1, 2, 3, 2, 4, 3