SALUTON DENOVE!

확률 및 통계학

동국대학교

2025년 2학기

문 동 휘

월 12:00 ~ 13:30

수 11:00 ~ 12:30

대푯값

- 대푯값: 자료 분포의 중심 또는 중심적 경향
- 자료의 분포를 정확히 파악하기는 어려움
- 그래서 대푯값을 사용하여 그 분포의 특징을 기술한다.
- 대푯값들: 평균, 중앙값, 최빈값, 중앙범위

평균

The **mean** is the sum of the values, divided by the total number of values.

The **sample mean**, denoted by $\overline{X}$ (pronounced "*X* bar"), is calculated by using sample data. The sample mean is a statistic.

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} = \frac{\Sigma X}{n}$$

- 표본평균
  - 일반적인 평균

⇒ 데이터 개수

where n represents the total number of values in the sample.

The **population mean**, denoted by μ (pronounced "mew"), is calculated by using all the values in the population. The population mean is a parameter.

$$\mu = \frac{X_1 + X_2 + X_3 + \cdots + X_N}{N} = \frac{\Sigma X}{N}$$

⇒ 모집단 전체 개수

where N represents the total number of values in the population.

평균

유한집단

모집단이 ∞ 이라면 모평균 불가능 ⇒ 표본평균으로 구해야 한다

- 유의사항: 무한 모집단의 모평균은 계산이 불가능. 그래서 표본을 추출하여 표본평균으로 그 값을 추정한다.

EXAMPLE 3–1 Police Incidents

The number of calls that a local police department responded to for a sample of 9 months is shown. Find the mean. (Data were obtained by the author.)

475, 447, 440, 761, 993, 1052, 783, 671, 621

SOLUTION

$$\overline{X} = \frac{\Sigma x}{n} = \frac{475 + 447 + 440 + 761 + 993 + 1052 + 783 + 671 + 621}{9}$$

$$= \frac{6243}{9} \approx 693.7$$

Hence, the mean number of incidents per month to which the police responded is 693.7.

평균

EXAMPLE 3–2 Hospital Infections

The data show the number of patients in a sample of six hospitals who acquired an infection while hospitalized. Find the mean.

110 76 29 38 105 31

Source: Pennsylvania Health Care Cost Containment Council.

SOLUTION

$$\overline{X} = \frac{\Sigma X}{n} = \frac{110 + 76 + 29 + 38 + 105 + 31}{6} = \frac{389}{6} = 64.8$$

The mean of the number of hospital infections for the six hospitals is 64.8.

도수분포표에서 평균 구하기

Procedure Table

Finding the Mean for Grouped Data

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$

Step 2 Find the midpoints of each class and place them in column C.

Step 3 Multiply the frequency by the midpoint for each class, and place the product in column D.

Step 4 Find the sum of column D.

Step 5 Divide the sum obtained in column D by the sum of the frequencies obtained in column B.

The formula for the mean is

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n}$$

[*Note:* The symbols $\Sigma f \cdot X_m$ mean to find the sum of the product of the frequency (f) and the midpoint (X_m) for each class.]

도수분표포에서의 평균

EXAMPLE 3–3 **Miles Run per Week**

Using the following frequency distribution (taken from Example 2–7), find the mean. The data represent the number of miles run during one week for a sample of 20 runners.

Class boundaries	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2
	Total 20

SOLUTION

The procedure for finding the mean for grouped data is given here.

Step 1 Make a table as shown.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

$\frac{10.5 + 15.5}{2}$

$\frac{30.5 + 35.5}{2}$

$\Rightarrow \bar{X} = \frac{490}{20}$

$= 24.5$

도수분표포에서의 평균

Step 2 Find the midpoints of each class and enter them in column C.

$$X_m = \frac{5.5 + 10.5}{2} = 8 \qquad \frac{10.5 + 15.5}{2} = 13 \quad \text{etc.}$$

Step 3 For each class, multiply the frequency by the midpoint, as shown, and place the product in column D.

$$1 \cdot 8 = 8 \qquad 2 \cdot 13 = 26 \quad \text{etc.}$$

The completed table is shown here.

A Class	B Frequency f	C Midpoint X_m	D $f \cdot X_m$
5.5–10.5	1	8	8
10.5–15.5	2	13	26
15.5–20.5	3	18	54
20.5–25.5	5	23	115
25.5–30.5	4	28	112
30.5–35.5	3	33	99
35.5–40.5	2	38	76
	$n = 20$		$\Sigma f \cdot X_m = 490$

Step 4 Find the sum of column D.

Step 5 Divide the sum by n to get the mean.

$$\overline{X} = \frac{\Sigma f \cdot X_m}{n} = \frac{490}{20} = 24.5 \text{ miles}$$

중앙값

- 구간척도나 비율척도로 관측된 자료에 사용되는 대푯값의 일종
- 자료를 크기 순서로 배열한다.
- 그 배열의 중간에 위치한 값 = 중앙값
- 자료의 개수는 n개. 자료의 개수 n이 홀수이면 (n + 1)/2번째 값이 중앙값, n이 짝수이면 n/2번째와 n/2 + 1 번째의 순위 값들의 산술평균이 중앙값.

중앙값

EXAMPLE 3–4 Police Officers Killed

The number of police officers killed in the line of duty over the last 11 years is shown. Find the median.

177 153 122 141 189 155 162 165 149 157 240

Source: National Law Enforcement Officers Memorial Fund.

SOLUTION

Step 1 Arrange the data in ascending order.

122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240

$\frac{11+1}{2} = 6$

Step 2 There are an odd number of data values, namely, 11.

Step 3 Select the middle data value.

122, 141, 149, 153, 155, 157, 162, 165, 177, 189, 240

↑

Median

The median number of police officers killed for the 11-year period is 157.

중앙값

EXAMPLE 3–5 Tornadoes in the United States

The number of tornadoes that have occurred in the United States over an 8-year period follows. Find the median.

684, 764, 656, 702, 856, 1133, 1132, 1303

SOLUTION

Step 1 Arrange the data values in ascending order.

656, 684, 702, 764, 856, 1132, 1133, 1303

Step 2 There are an even number of data values, namely, 8.

Step 3 The middle two data values are 764 and 856.

656, 684, 702, 764, 856, 1132, 1133, 1303

↑

Median

Since the middle point falls halfway between 764 and 856, find the median MD by adding the two values and dividing by 2.

$$\text{MD} = \frac{764 + 856}{2} = \frac{1620}{2} = 810$$

$\Rightarrow$ $\frac{n\text{번째} + (n+1)\text{번째}}{2}$

The median number of tornadoes is 810.

최빈값

단일최빈값: 가장 많이 나타나는 값 1
이중최빈값: 가장 많이 나타나는 값이 2개 ⇒ 모두 최빈값
다중최빈값: 가장 많이 나타나는 값이 3개 이상
⇒ 각 값을 최빈값
최빈값 없음 최빈값 없다고 ∅ 처리 X
⇒ 실제값이 ∅인 것 존재 : 온도 등

- 최빈값 = 자료의 측정값들 중에서 가장 높은 빈도로 나타나는 값. 질적자료 (명목형, 순서형)

최빈값

The value that occurs most often in a data set is called the **mode**.

A data set that has only one value that occurs with the greatest frequency is said to be **unimodal.**

If a data set has two values that occur with the same greatest frequency, both values are considered to be the mode and the data set is said to be **bimodal.** If a data set has more than two values that occur with the same greatest frequency, each value is used as the mode, and the data set is said to be **multimodal.** When no data value occurs more than once, the data set is said to have *no mode. Note: Do not say that the mode is zero.* That would be incorrect, because in some data, such as temperature, zero can be an actual value. A data set can have more than one mode or no mode at all. These situations will be

최빈값

- 장점 1: 이상값들의 영향을 받지 않는다.
  - 이상값: 자료를 크기 순서로 나열했을 때 그 크기가 다른 값들에 비해 훨씬 크거나 작은 값
- 장점 2: 질적 자료의 분석에서 유일하게 사용되는 대푯값
- 단점1: 자료의 측정값들이 모두 동일한 빈도수를 가진 경우, 그 자료에는 최빈값이 존재하지 않는다.
- 단점2: 자료에 따라 여러 개의 최빈값이 존재 가능.

최빈값

EXAMPLE 3–6 NFL Signing Bonuses

Find the mode of the signing bonuses of eight NFL players for a specific year. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

Source: *USA TODAY.*

단일최빈값

SOLUTION

It is helpful to arrange the data in order, although it is not necessary.

10, 10, 10, 11.3, 12.4, 14.0, 18.0, 34.5

Since $10 million occurred 3 times—a frequency larger than any other number—the mode is $10 million.

EXAMPLE 3–7 Licensed Nuclear Reactors

The data show the number of licensed nuclear reactors in the United States for a recent 15-year period. Find the mode.

Source: *The World Almanac and Book of Facts.*

104	104	104	104	104
107	109	109	109	110
109	111	112	111	109

이중최빈값

SOLUTION

Since the values 104 and 109 both occur 5 times, the modes are 104 and 109. The data set is said to be bimodal.

최빈값

The mode for grouped data is the modal class. The **modal class** is the class with the largest frequency.

EXAMPLE 3–9 Miles Run per Week

Find the modal class for the frequency distribution of miles that 20 runners ran in one week, used in Example 2–7.

Class	Frequency
5.5–10.5	1
10.5–15.5	2
15.5–20.5	3
20.5–25.5	5 ← Modal class
25.5–30.5	4
30.5–35.5	3
35.5–40.5	2

20.5 + 25.5
= 46/2 = 23

SOLUTION

이게 어떤 경우인데?

The modal class is 20.5–25.5, since it has the largest frequency. Sometimes the midpoint of the class is used rather than the boundaries; hence, the mode could also be given as 23 miles per week.

↳ 20.5와 25.5의 중앙값

A small company consists of the owner, the manager, the salesperson, and two technicians, all of whose annual salaries are listed here. (Assume that this is the entire population.)

Staff	Salary
Owner	$100,000
Manager	40,000
Salesperson	24,000
Technician	18,000
Technician	18,000

Total : 200,000
mean : 200,000/5 = 40,000
median : 24,000
mode : 18,000

Find the mean, median, and mode.

SOLUTION

$$\mu = \frac{\Sigma X}{N} = \frac{\$100{,}000 + 40{,}000 + 24{,}000 + 18{,}000 + 18{,}000}{5} = \frac{\$200{,}000}{5} = \$40{,}000$$

Hence, the mean is $40,000, the median is $24,000, and the mode is $18,000.

중앙범위

- 중앙범위: 자료의 측정값들 중에서 얻은 최댓값과 최솟값의 산술평균

The **midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2. The symbol MR is used for the midrange.

$$MR = \frac{\text{lowest value} + \text{highest value}}{2}$$

중앙범위

EXAMPLE 3–12 Bank Failures

The number of bank failures for a recent five-year period is shown. Find the midrange.

3, 30, 148, 157, 71

Source: Federal Deposit Insurance Corporation.

SOLUTION

The lowest data value is 3, and the highest data value is 157.

$$\text{MR} = \frac{3 + 157}{2} = \frac{160}{2} = 80$$

The midrange for the number of bank failures is 80.

EXAMPLE 3–13 NFL Signing Bonuses

Find the midrange of data for the NFL signing bonuses in Example 3–6. The bonuses in millions of dollars are

18.0, 14.0, 34.5, 10, 11.3, 10, 12.4, 10

SOLUTION

The lowest bonus is \$10 million, and the largest bonus is \$34.5 million.

$$\text{MR} = \frac{10 + 34.5}{2} = \frac{44.5}{2} = \$22.25 \text{ million}$$

Notice that this amount is larger than seven of the eight amounts and is not typical of the average of the bonuses. The reason is that there is one very high bonus, namely, \$34.5 million.

각 값에 가중치를 곱한 뒤, 곱들의 합/가중치들의 합 ⇒ 가중평균

가중평균

Find the **weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of the weights.

$$\overline{X} = \frac{w_1X_1 + w_2X_2 + \cdots + w_nX_n}{w_1 + w_2 + \cdots + w_n} = \frac{\Sigma wX}{\Sigma w}$$

where $w_1, w_2, \ldots, w_n$ are the weights and $X_1, X_2, \ldots, X_n$ are the values.

EXAMPLE 3–14 Grade Point Average

A student received an A in English Composition I (3 credits), a C in Introduction to Psychology (3 credits), a B in Biology I (4 credits), and a D in Physical Education (2 credits). Assuming A = 4 grade points, B = 3 grade points, C = 2 grade points, D = 1 grade point, and F = 0 grade points, find the student's grade point average.

SOLUTION

가중치

Course	Credits (w)	Grade (X)
English Composition I	3	A (4 points)
Introduction to Psychology	3	C (2 points)
Biology I	4	B (3 points)
Physical Education	2	D (1 point)

$$\frac{3\times4 + 3\times2 + 4\times3 + 2\times1}{12}$$

$$= \frac{12+6+12+2}{12} = \frac{32}{12} = 2.7\cdots$$

$$\overline{X} = \frac{\Sigma wX}{\Sigma w} = \frac{3\cdot4 + 3\cdot2 + 4\cdot3 + 2\cdot1}{3+3+4+2} = \frac{32}{12} \approx 2.7$$

The grade point average is 2.7.

자료의 위치점

- P% 위치점: 자료의 측정값들을 작은 것부터 크기 순서로 나열하였을 때 전체 측정값 중에서 주어진 비율 p 만큼의 측정값들이 그 값보다 작거나 같고 (1- p) 만큼의 측정값들이 그 값과 같거나 큰 값인 측정값/데이터 포인트.

- 예) 중앙값 = 50% 위치점 (자료에서 50%가 작거나 같고 50%가 크거나 같다)

백분위수 : 100개의 동일한 구간으로 나누는 값이다

Percentiles divide the data set into 100 equal groups.

Percentiles are symbolized by

$$P_1, P_2, P_3, \ldots, P_{99}$$

and divide the distribution into 100 groups.

백분위수

X 미만의 수의 개수

자료위치를 더 정확히 하기 위해 사용.

Percentile Formula

The percentile corresponding to a given value X is computed by using the following formula:

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

EXAMPLE 3–30 Test Scores

A teacher gives a 20-point test to 10 students. The scores are shown here. Find the percentile rank of a score of 12.

18, 15, 12, 6, 8, 2, 3, 5, 20, 10

SOLUTION

Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

$\frac{6+0.5}{10} \times 100$

$= 6.5 \times 10$

$= 65$

Then substitute into the formula.

$$\text{Percentile} = \frac{(\text{number of values below } X) + 0.5}{\text{total number of values}} \cdot 100$$

Since there are six values below a score of 12, the solution is

$$\text{Percentile} = \frac{6 + 0.5}{10} \cdot 100 = 65\text{th percentile}$$

Thus, a student whose score was 12 did better than 65% of the class.

: 12점을 받은 학생은 전체 학생 65%보다 높은 성적 기록한 것이다

: 12점은 상위 35% 이다

백분위수 (X - 데시보기)

Procedure Table

Finding a Data Value Corresponding to a Given Percentile

Step 1 Arrange the data in order from lowest to highest.

Step 2 Substitute into the formula

$$c = \frac{n \cdot p}{100}$$

where n = total number of values
p = percentile

Step 3A If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded-up value.

Step 3B If c is a whole number, use the value halfway between the cth and $(c + 1)$st values when counting up from the lowest value.

백분위수

EXAMPLE 3–32 **Test Scores**

Using the scores in Example 3–30, find the value corresponding to the 25th percentile.

25번째 백분위수에 해당하는 값 찾아라.

SOLUTION ①

Step 1 Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Compute

여기까 선 페이지의 X미만의 개수?

$$c = \frac{n \cdot p}{100}$$

where n = total number of values : 전체 데이터 개수
p = percentile : 구하고자하는 백분위수

Thus,

$$c = \frac{10 \cdot 25}{100} = 2.5$$

$$P = \frac{C\text{(X이만 개수)} + (0.5)}{\text{total number}} \times 100$$

또는 뽑힌 값이기에 C구할 때 제외

$$C = P \times \frac{\text{total number}}{100}$$

$$\therefore C = \frac{n \cdot P}{100}$$

$C = \frac{10 \times 25}{100} = \frac{250}{100} = 2.5$

⇒ 반올림 = 3

∴ 3번째 값 5가 P25이다

Step 3 Since c is not a whole number, round it up to the next whole number; in this case, $c = 3$. Start at the lowest value and count over to the third value, which is 5. Hence, the value 5 corresponds to the 25th percentile.

백분위수

EXAMPLE 3–33

Using the data set in Example 3–30, find the value that corresponds to the 60th percentile.

SOLUTION (2)

Step 1 Arrange the data in order from lowest to highest.

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

Step 2 Substitute in the formula.

$$c = \frac{n \cdot p}{100} = \frac{10 \cdot 60}{100} = 6$$

$C = \frac{10 \times 60}{100} = 6$

$\frac{6\text{번째} + 7\text{번째}(c+1)}{2} = \frac{22}{2} = 11$

$\therefore P_{60} \Rightarrow 11$이다 $\Rightarrow$ 11점 받은 학생이 반에서 60%학생보다 좋은 점수 받음

$\Rightarrow$ 11점 받은 학생은 상위 40%

Step 3 Since c is a whole number, use the value halfway between the c and $c + 1$ values when counting up from the lowest value—in this case, the 6th and 7th values.

→ C + C+1의 평균

2, 3, 5, 6, 8, 10, 12, 15, 18, 20

6th value 7th value

The value halfway between 10 and 12 is 11. Find it by adding the two values and dividing by 2.

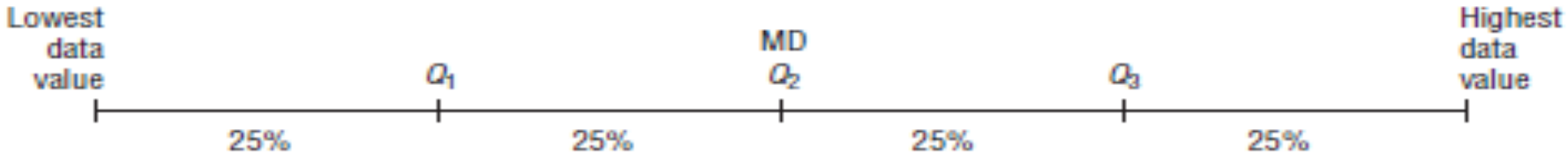
$$\frac{10 + 12}{2} = 11$$

Hence, 11 corresponds to the 60th percentile. Anyone scoring 11 would have done better than 60% of the class.

사분위수

Quartiles divide the distribution into four equal groups, denoted by Q_1, Q_2, Q_3.

Note that Q_1 is the same as the 25th percentile, Q_2 is the same as the 50th percentile, or the median; Q_3 corresponds to the 75th percentile, as shown:



Quartiles can be computed by using the formula given for computing percentiles on page 153. For Q_1 use $p = 25$. For Q_2 use $p = 50$. For Q_3 use $p = 75$. However, an easier method for finding quartiles is found in this Procedure Table.

Procedure Table

Finding Data Values Corresponding to Q_1, Q_2, and Q_3

Step 1	Arrange the data in order from lowest to highest.
Step 2	Find the median of the data values. This is the value for Q_2.
Step 3	Find the median of the data values that fall below Q_2. This is the value for Q_1.
Step 4	Find the median of the data values that fall above Q_2. This is the value for Q_3.

사분위수

EXAMPLE 3–34

Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

SOLUTION

Step 1 Arrange the data in order from lowest to highest.

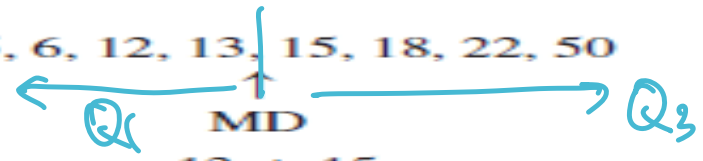
5, 6, 12, 13, 15, 18, 22, 50

Step 2 Find the median (Q_2).

5, 6, 12, 13, 15, 18, 22, 50

↑

MD



$$\text{MD} = \frac{13 + 15}{2} = 14$$

Step 3 Find the median of the data values less than 14.

5, 6, 12, 13

↑

Q_1

$$Q_1 = \frac{6 + 12}{2} = 9$$

So Q_1 is 9.

Step 4 Find the median of the data values greater than 14.

15, 18, 22, 50

↑

Q_3

$$Q_3 = \frac{18 + 22}{2} = 20$$

Here Q_3 is 20. Hence, $Q_1 = 9$, $Q_2 = 14$, and $Q_3 = 20$.

사분위수 범위

The **interquartile range (IQR)** is the difference between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

EXAMPLE 3–34

Find Q_1, Q_2, and Q_3 for the data set 15, 13, 6, 5, 12, 50, 22, 18.

EXAMPLE 3–35

Find the interquartile range for the data set in Example 3–34.

SOLUTION

First it is necessary to find the values of Q_1 and Q_3. These values were found in Example 3–34: $Q_1 = 9$ and $Q_3 = 20$. Next subtract the value of Q_1 from Q_3 to get the interquartile range.

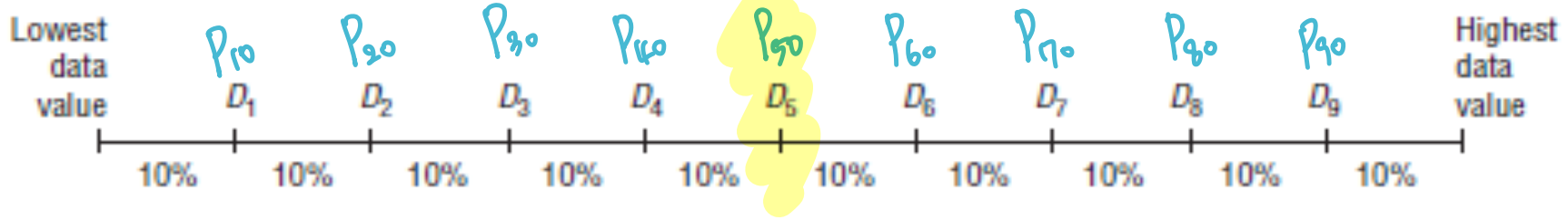
$$IQR = Q_3 - Q_1 = 20 - 9 = 11$$

The interquartile range is equal to 11.

십분위수

$D_5 = P_{50} = Q_2$

Deciles divide the distribution into 10 groups, as shown. They are denoted by D_1, D_2, etc.



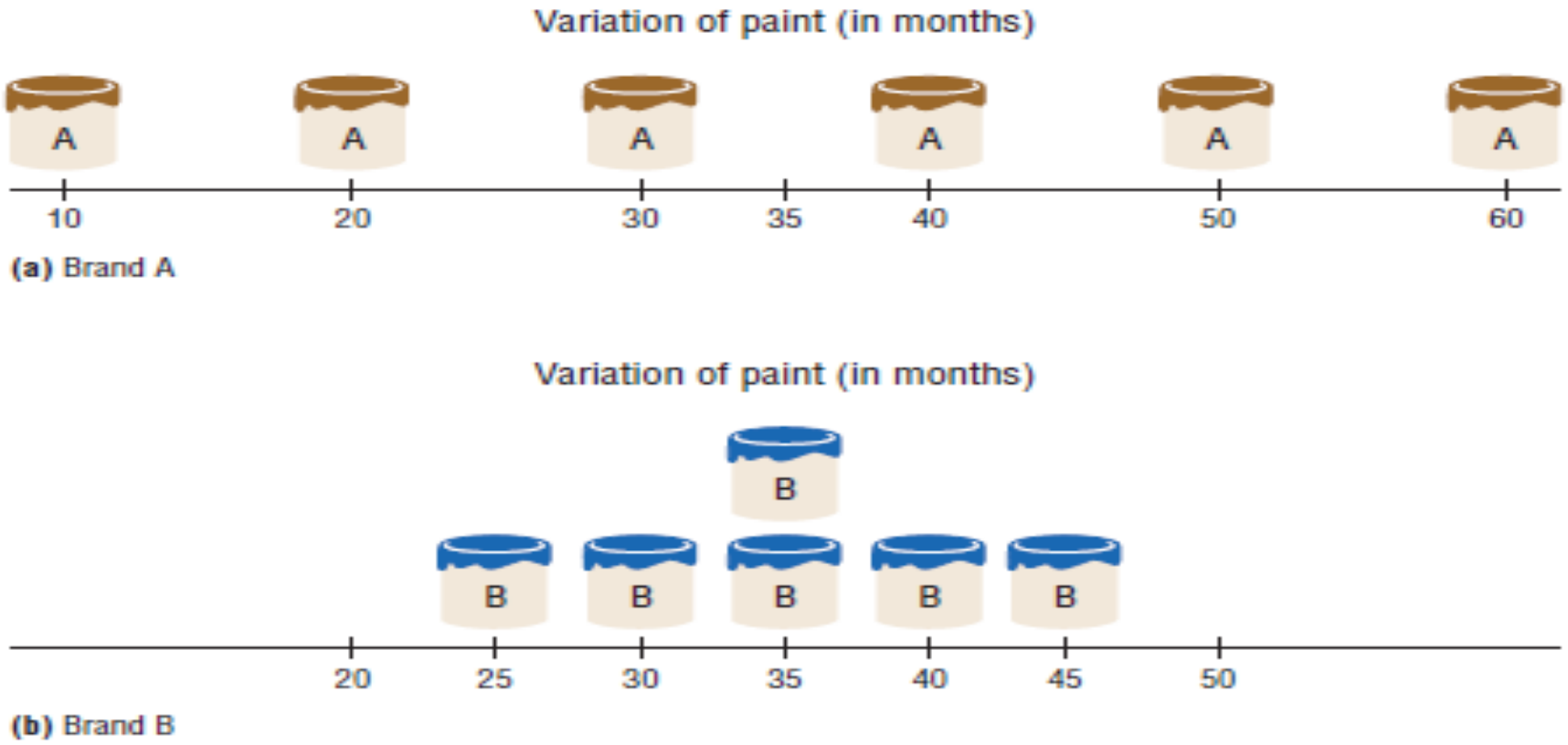
Note that D_1 corresponds to P_{10}; D_2 corresponds to P_{20}; etc. Deciles can be found by using the formulas given for percentiles. Taken altogether then, these are the relationships among percentiles, deciles, and quartiles.

Deciles are denoted by $D_1, D_2, D_3, \ldots, D_9$, and they correspond to $P_{10}, P_{20}, P_{30}, \ldots, P_{90}$.

Quartiles are denoted by Q_1, Q_2, Q_3 and they correspond to P_{25}, P_{50}, P_{75}.

The median is the same as P_{50} or Q_2 or D_5.

산포도 (퍼짐 정도) 통계값



- 산포도를 나타내는 통계값: 범위, 사분위수 범위, 분산, 표준편차, 변이계수

범위 : 데이터의 흩어짐 정도이기 때문에, [범위 큼 : 데이터 변동 큼 / 범위 작음 : 데이터 변동 작음]

The **range** is the highest value minus the lowest value. The symbol R is used for the range.

$$R = \text{highest value} - \text{lowest value}$$

EXAMPLE 3–16 Comparison of Outdoor Paint

Find the ranges for the paints in Example 3–15.

Brand A	Brand B
10	35
60	45
50	30
30	35
40	40
20	25

SOLUTION

For brand A, the range is

$$R = 60 - 10 = 50 \text{ months}$$

For brand B, the range is

$$R = 45 - 25 = 20 \text{ months}$$

Make sure the range is given as a single number.

The range for brand A shows that 50 months separate the largest data value from the smallest data value. For brand B, 20 months separate the largest data value from the smallest data value, which is less than one-half of brand A's range. : B 범위 < A의 절반. ⇒ B가 더 일관성있게 모임

범위

EXAMPLE 3–17 Employee Salaries

The salaries for the staff of the XYZ Manufacturing Co. are shown here. Find the range.

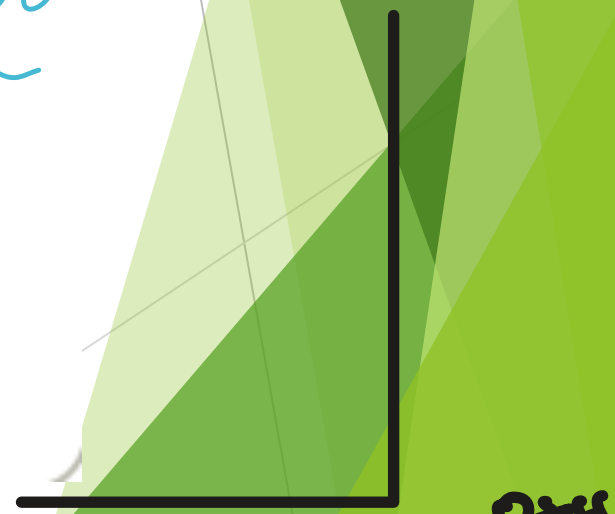
Staff	Salary
Owner	$100,000
Manager	40,000
Sales representative	30,000
Workers	25,000
	15,000
	18,000

100,000 − 15,000 = 85,000

R = 85,000

SOLUTION

The range is $R = \$100{,}000 - \$15{,}000 = \$85{,}000$.



모분산과 모표준편차

모분산

The **population variance** is the average of the squares of the distance each value is from the mean. The symbol for the population variance is σ^2 (σ is the Greek lower-case letter sigma).

The formula for the population variance is

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

where X = individual value (개별 데이터 값)

μ = population mean (모평균)

N = population size (모집단크기)

모표준편차

The **population standard deviation** is the square root of the variance. The symbol for the population standard deviation is σ.

The corresponding formula for the population standard deviation is

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

모분산과 모표준편차

Procedure Table

Finding the Population Variance and Population Standard Deviation

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N}$$ = 개별 데이터 합 / 모집단 크기

Step 2 Find the deviation for each data value.

$$X - \mu$$

Step 3 Square each of the deviations.

$$(X - \mu)^2$$

Step 4 Find the sum of the squares.

$$\Sigma(X - \mu)^2$$

Step 5 Divide by N to get the variance.

$$\sigma^2 = \frac{\Sigma(X - \mu)^2}{N}$$

Step 6 Take the square root of the variance to get the standard deviation.

$$\sigma = \sqrt{\frac{\Sigma(X - \mu)^2}{N}}$$

모분산과 모표준편차

Find the variance and standard deviation for the data set for brand A paint in Example 3–15. The number of months brand A lasted before fading was

$$10, 60, 50, 30, 40, 20$$

Step 1 Find the mean for the data.

$$\mu = \frac{\Sigma X}{N} = \frac{10 + 60 + 50 + 30 + 40 + 20}{6} = \frac{210}{6} = 35$$

Step 2 Subtract the mean from each data value $(X - \mu)$. ⇒ 각각의 값 − 모평균

$$10 - 35 = -25 \qquad 50 - 35 = +15 \qquad 40 - 35 = +5$$
$$60 - 35 = +25 \qquad 30 - 35 = -5 \qquad 20 - 35 = -15$$

Step 3 Square each result $(\overline{X} - \mu)^2$.

$$(-25)^2 = 625 \qquad (+15)^2 = 225 \qquad (+5)^2 = 25$$
$$(+25)^2 = 625 \qquad (-5)^2 = 25 \qquad (-15)^2 = 225$$

모분산과 모표준편차

Step 4 Find the sum of the squares $\Sigma(\overline{X} - \mu)^2$.

$$625 + 625 + 225 + 25 + 25 + 225 = 1750$$

Step 5 Divide the sum by N to get the variance $\dfrac{[\Sigma(\overline{X} - \mu)^2]}{N}$.

$$\text{Variance} = 1750 \div 6 \approx 291.7$$: 모분산

Step 6 Take the square root of the variance to get the standard deviation. Hence, the standard deviation equals $\sqrt{291.7}$, or 17.1. It is helpful to make a table.

→ 모표준편차.

A Values X	B $X - \mu$	C $(X - \mu)^2$
10	−25	625
60	+25	625
50	+15	225
30	−5	25
40	+5	25
20	−15	225
		1750

Column A contains the raw data X. Column B contains the differences $X - \mu$ obtained in step 2. Column C contains the squares of the differences obtained in step 3.

: 모분산 & 모표준편차와 비교.

표본분산과 표본표준편차

Formula for the Sample Variance

The formula for the sample variance (denoted by s^2) is

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

where X = individual value
$\overline{X}$ = sample mean
n = sample size

Formula for the Sample Standard Deviation

The formula for the sample standard deviation, denoted by s, is

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}}$$

where X = individual value
$\overline{X}$ = sample mean
n = sample size

표본분산과 표본표준편차

EXAMPLE 3–20 Teacher Strikes

The number of public school teacher strikes in Pennsylvania for a random sample of school years is shown. Find the sample variance and the sample standard deviation.

9 10 14 7 8 3

Source: Pennsylvania School Board Association.

Step 1 Find the mean of the data values.

$$\overline{X} = \frac{\Sigma X}{n} = \frac{9 + 10 + 14 + 7 + 8 + 3}{6} = \frac{51}{6} = 8.5$$

- 편차평균 구할.

Step 2 Find the deviation for each data value $(X - \overline{X})$.

$9 - 8.5 = 0.5$ $10 - 8.5 = 1.5$ $14 - 8.5 = 5.5$
$7 - 8.5 = -1.5$ $8 - 8.5 = -0.5$ $3 - 8.5 = -5.5$

Step 3 Square each of the deviations $(X - \overline{X})^2$.

$(0.5)^2 = 0.25$ $(1.5)^2 = 2.25$ $(5.5)^2 = 30.25$
$(-1.5)^2 = 2.25$ $(-0.5)^2 = 0.25$ $(-5.5)^2 = 30.25$

표본분산과 표본표준편차

Step 4 Find the sum of the squares.

$$\Sigma(X - \overline{X})^2 = 0.25 + 2.25 + 30.25 + 2.25 + 0.25 + 30.25 = 65.5$$

Step 5 Divide by $n - 1$ to get the variance.

$$s^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1} = \frac{65.5}{6 - 1} = \frac{65.5}{5} = 13.1$$

Step 6 Take the square root of the variance to get the standard deviation.

$$s = \sqrt{\frac{\Sigma(X - \overline{X})^2}{n - 1}} = \sqrt{13.1} \approx 3.6 \text{ (rounded)}$$

Here the sample variance is 13.1, and the sample standard deviation is 3.6.

Sum of Squares, 도수분포표의 분산

- $\Sigma(x_i - \bar{x})^2 = \Sigma {x_i}^2 - n\bar{x}^2$

- K의 계급으로 구성된 도수분포표의 계급값을 $x_1^*, x_2^*, \ldots, x_k^*$ 라 하고 이에 대응되는 도수를 각각 $f_1, f_2, \ldots, f_k$ 라고 할 때, 자료의 분산은 다음과 같이 근사된다.

- 모분산: $\sigma^2 = \Sigma f_i (x_i^* - \mu)^2$ / N = $\left[\Sigma f_i {x_i^*}^2 - N\mu^2\right]$ / N

- 표본분산: $s^2 = \Sigma f_i (x_i^* - \bar{x})^2$ / (n - 1) = $[\Sigma f_i {x_i^*}^2 - n\bar{x}^2]/(n - 1)$

도수분포표의 분산

표본분산: $\Sigma f_i(x_i^* - \bar{x})^2$

$\bar{x} = 17.8$

$\Sigma f_i(x_i^* - \bar{x})^2 = 167.8$

표본분산: 11.98

표본표준편차: 3.46…

근사값: 3

표본분산: $s^2 = \Sigma f_i(x_i^* - \bar{x})^2 / (n-1) = [\Sigma f_i x_i^{*2} - n\bar{x}^2]/(n-1)$

15명의 학생이 일주일간 지하철을 탄 횟수. 표본평균과 표본분산은?

계급값	10	15	17	20	22	
빈도수	1	3	5	2	4	

계급값	빈도수	계급값*빈도수	계급값 제곱	계급값 제곱*빈도		
10	1	10	100	100	표본평균	17.86667 ⇒ 17.8
15	3	45	225	675	제곱합	167.7333
17	5	85	289	1445	표본분산	11.98095
20	2	40	400	800		
22	4	88	484	1936		
	15	268	1498	4956		

표본표준편차의 근삿값

- 자료크기가 20이상, 종형분포 (정규분포형 분포)를 따르는 경우, 아래 공식으로 표본표준편차 근사 가능

- 표본표준편차 ≈ 범위 / 4
  - 근사값을 구할 때 반올림 X

체비셰프 정리 ⇒ 일변량 종류

└ 한 변수 X가 평균 μ, 표준편차 σ를 가질 때 — 일변량
변수 X 하나로 평균으로부터 얼마나 떨어졌는지 확률을 다루기 때문.

❒ **Chebyshev's inequality** — Let X be a random variable with expected value μ. For $k, \sigma > 0$, we have the following inequality:

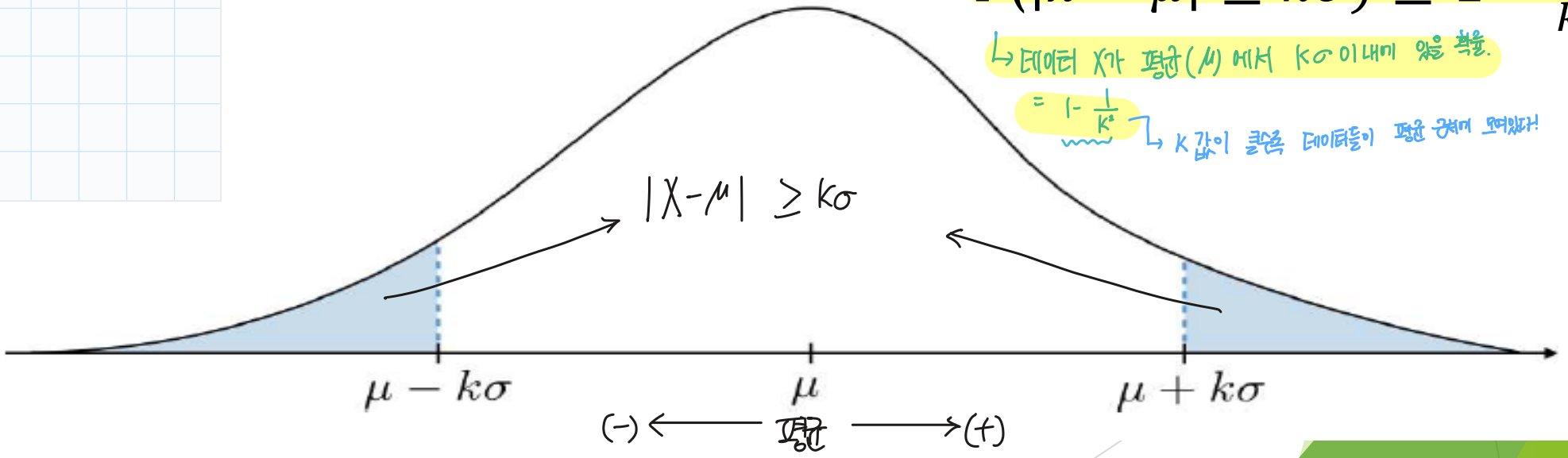
$$P(|X - \mu| \geqslant k\sigma) \leqslant \frac{1}{k^2}$$

$$P(|x - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$$

↳ 데이터 X가 평균(μ)에서 $k\sigma$ 이내에 있을 확률.

$= 1 - \frac{1}{k^2}$ ↳ k 값이 클수록 데이터들이 평균 근처에 모여있다!

X : 데이터 값 (확률 변수)
μ : 평균
σ : 표준편차
k : 표준편차로부터 얼마나 떨어져 있는지 ($k > \sigma$)



체비셰프 정리

$P(|X-\mu| \le k\sigma) \ge 1 - \frac{1}{k^2}$ → 75%

$|X-\mu| \le k\sigma$

$-k\sigma \le X-\mu \le k\sigma$

$\mu - k\sigma \le X \le k\sigma + \mu$

k=2

EXAMPLE 3–25 Prices of Homes

The mean price of houses in a certain neighborhood is $50,000, and the standard deviation is $10,000. Find the price range for which at least 75% of the houses will sell.

평균 집값

표준편차

75%의 집값이 포함되는 가격 범위 구하기

SOLUTION

Chebyshev's theorem states that three-fourths, or 75%, of the data values will fall within 2 standard deviations of the mean. Thus,

$$\$50{,}000 + 2(\$10{,}000) = \$50{,}000 + \$20{,}000 = \$70{,}000$$

and

$$\$50{,}000 - 2(\$10{,}000) = \$50{,}000 - \$20{,}000 = \$30{,}000$$

Hence, at least 75% of all homes sold in the area will have a price range from $30,000 to $70,000.

체비셰프 정리

$P(|X-\mu| \le K\sigma) \ge 1 - \frac{1}{K^2}$

① $\Rightarrow |X-\mu| \le K\sigma$

$-K\sigma \le X-\mu \le K\sigma$

$-K\sigma+\mu \le X \le K\sigma+\mu$

$-K\sigma+\mu = 0.2$, $K\sigma+\mu = 0.3$

② $K\sigma+\mu = 0.3$

$K \times 0.02 + 0.25 = 0.3$

$K \times 0.02 = 0.05$

$K = \frac{0.05}{0.02} = 2.5$

③ $1 - \frac{1}{(2.5)^2} = 1 - \frac{1}{6.25}$

$= 1 - 0.16 = 0.84$

$\therefore 84\%$

평균 / 표준 / 요 범위에 있다.

EXAMPLE 3–26 **Travel Allowances**

A survey of local companies found that the mean amount of travel allowance for couriers was \$0.25 per mile. The standard deviation was \$0.02. Using Chebyshev's theorem, find the minimum percentage of the data values that will fall between \$0.20 and \$0.30.

SOLUTION

Step 1 Subtract the mean from the larger value.

$$\$0.30 - \$0.25 = \$0.05$$

Step 2 Divide the difference by the standard deviation to get k.

$$k = \frac{0.05}{0.02} = 2.5$$

Step 3 Use Chebyshev's theorem to find the percentage.

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2.5^2} = 1 - \frac{1}{6.25} = 1 - 0.16 = 0.84 \quad \text{or} \quad 84\%$$

Hence, at least 84% of the data values will fall between \$0.20 and \$0.30.

1) 0.20 과 0.30 의 중간 : 0.25

2) 평균 : 0.25

⇒ 왼쪽이 평균으로부터 떨어진 거리 : 0.05

오른쪽이 평균으로부터 떨어진 거리 : 0.05

⇒ 대칭!!

↳ 만약 대칭이 아니라면 체비셰프 사용X

0.2와 0.3은 전체 84% 이내에 존재

· 단위 없애고 숫자끼리 보게 하려고.

변이계수

· 표본은 표본끼리 비교.

- 각각 다른 측정 단위로된 데이터 자료 사이에 산포도를 비교하기 위하여 사용한다.

단위 없애고 숫자만 보기 위해서 사용.

표본은 표본끼리 비교.

The **coefficient of variation**, denoted by CVar, is the standard deviation divided by the mean. The result is expressed as a percentage.

For samples,

$$CVar = \frac{s}{\overline{X}} \cdot 100$$

For populations,

$$CVar = \frac{\sigma}{\mu} \cdot 100$$

변이계수

EXAMPLE 3–23 Sales of Automobiles

The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two.

SOLUTION

The coefficients of variation are

$$\text{CVar} = \frac{s}{\bar{X}} = \frac{5}{87} \cdot 100 = 5.7\% \quad \text{sales}$$

$$\text{CVar} = \frac{773}{5225} \cdot 100 = 14.8\% \quad \text{commissions}$$

Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

변이계수

EXAMPLE 3–24 Pages in Women's Fitness Magazines

The mean for the number of pages of a sample of women's fitness magazines is 132, with a variance of 23; the mean for the number of advertisements of a sample of women's fitness magazines is 182, with a variance of 62. Compare the variations.

SOLUTION

The coefficients of variation are

$$\text{CVar} = \frac{\sqrt{23}}{132} \cdot 100 = 3.6\% \quad \text{pages}$$

$$\text{CVar} = \frac{\sqrt{62}}{182} \cdot 100 = 4.3\% \quad \text{advertisements}$$

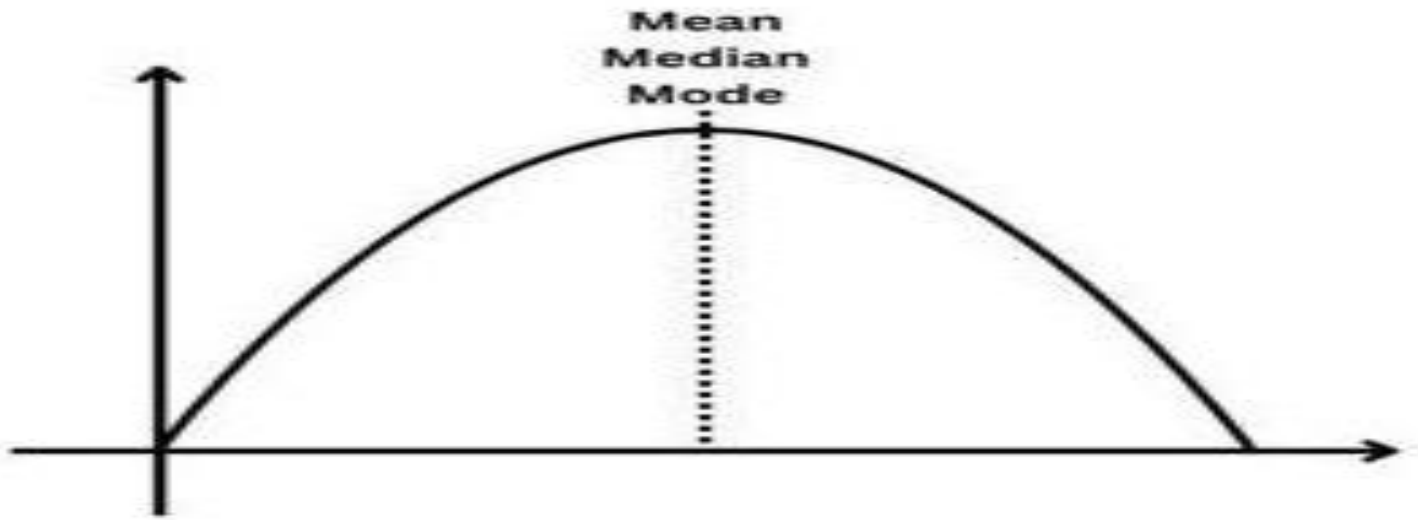
→ 얘가 더 흩어진다 == 변동가능성이 크다.

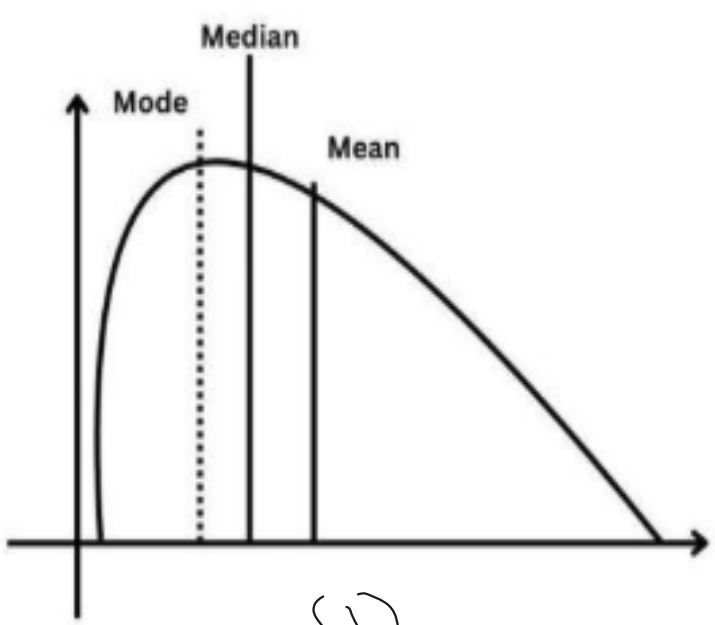
The number of advertisements is more variable than the number of pages since the coefficient of variation is larger for advertisements.

왜도 : 분포도가 얼마나 치우쳤나?

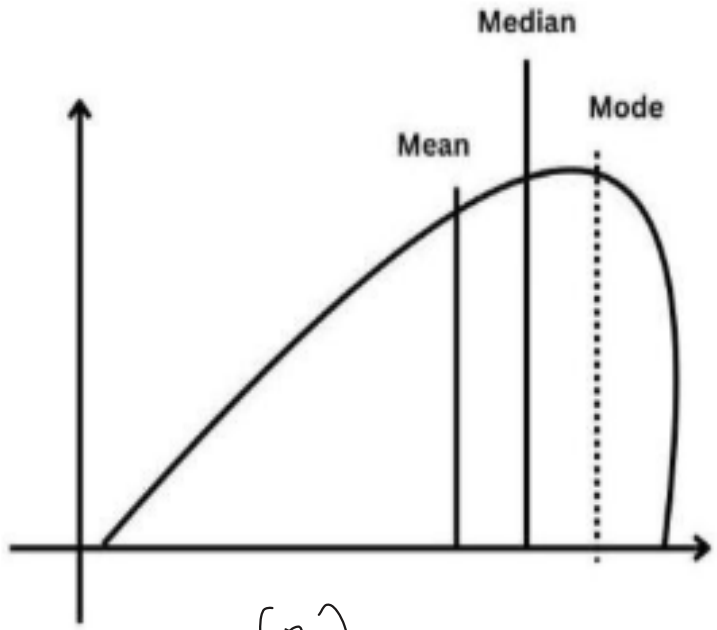
- 왜도 = 0 = 완벽한 대칭분포
- 왜도 < 0 = 좌 비대칭분포 = 좌측에 긴 꼬리
- 왜도 > 0 = 우 비대칭분포 = 우측에 긴 꼬리

정규분포 조건





(1)

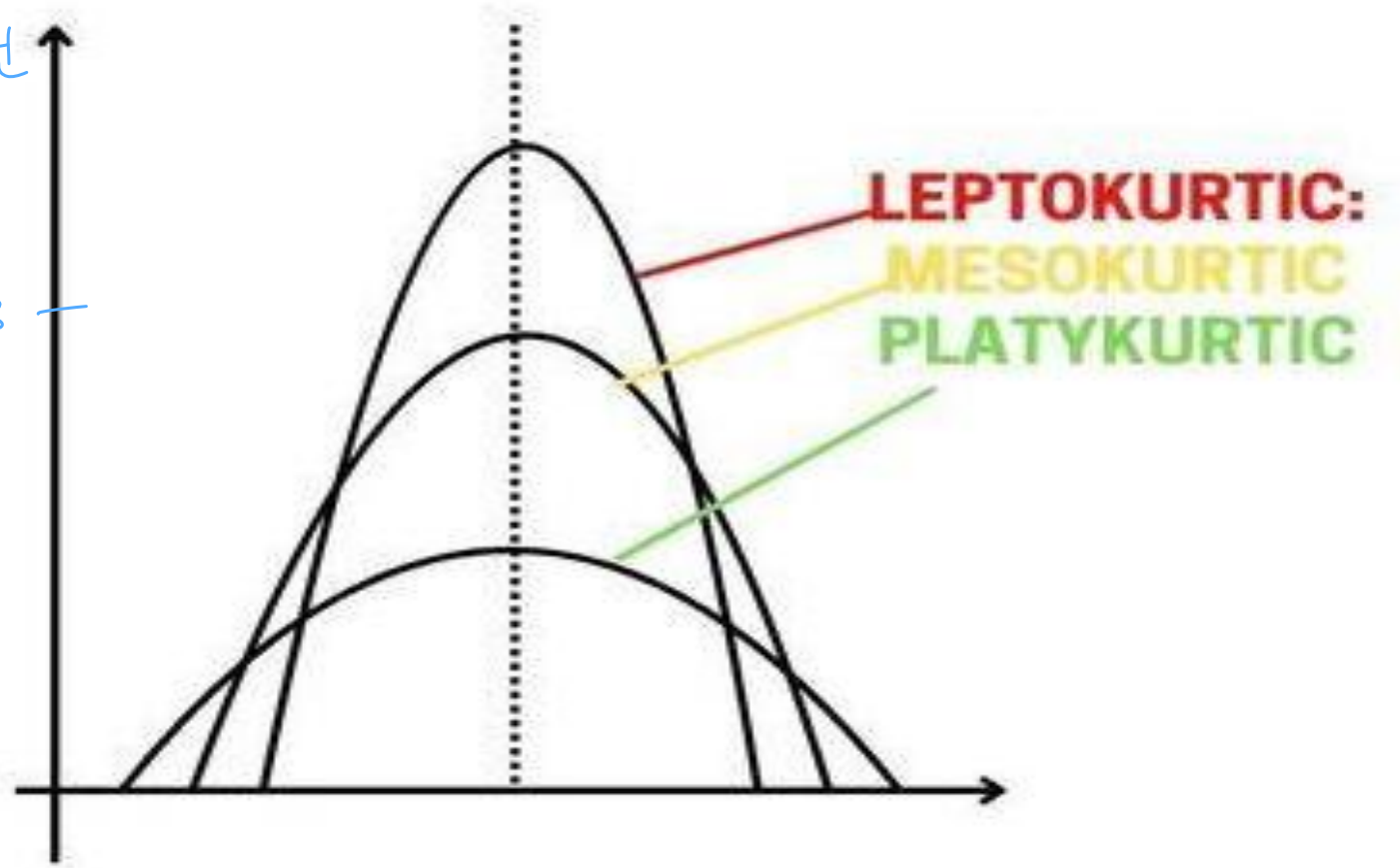


(2)

첨도

- 첨도 = 3 = 중첨 = Mesokurtic
- 첨도 < 3 = 완첨 = Platykurtic
- 첨도 > 3 = 급첨 = Leptokurtic

- 정규분포: 왜도 = 0, 첨도 = 3



VS 변이계수 : 전체 집단

표준점수 ⇒ 개별 데이터 보고 평균으로부터 몇 표준편차 떨어져 있는지

- 표준점수는 측정단위가 다르거나 분포가 다른 자료들 사이의 비교를 용이하게 한다.

A **z score** or **standard score** for a value is obtained by subtracting the mean from the value and dividing the result by the standard deviation. The symbol for a standard score is *z*. The formula is

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

For samples, the formula is

$$z = \frac{X - \bar{X}}{s}$$: 표본이 나오면 표본표준편차 사용

For populations, the formula is

$$z = \frac{X - \mu}{\sigma}$$: 모집단 나오면 모집단 표준편차 사용

모집단 알 수 없으면 표본 값으로 추정해서!

The *z* score represents the number of standard deviations that a data value falls above or below the mean.

표준점수

EXAMPLE 3–27 Test Scores

A student scored 65 on a calculus test that had a mean of 50 and a standard deviation of 10; she scored 30 on a history test with a mean of 25 and a standard deviation of 5. Compare her relative positions on the two tests.

SOLUTION

First, find the z scores. For calculus the z score is

$$z = \frac{X - \overline{X}}{s} = \frac{65 - 50}{10} = 1.5$$

For history the z score is

$$z = \frac{30 - 25}{5} = 1.0$$

→ 상대적으로 미적을 더 잘했다고 볼 수 O

Since the z score for calculus is larger, her relative position in the calculus class is higher than her relative position in the history class.

표준점수

점수의 범위, 표준편차, 평균이 다라 비교 불가 ⇒ 표준점수 사용 : $z = \frac{X - \bar{X}}{s}$

EXAMPLE 3–28 Test Scores

Find the z score for each test, and state which is higher.

Test A	$X = 38$	$\bar{X} = 40$	$s = 5$
Test B	$X = 94$	$\bar{X} = 100$	$s = 10$

→ $\frac{-2}{5} = -0.4$ ⇒ 더 높음 : 더 잘함

→ $\frac{-6}{10} = -0.6$

SOLUTION

For test A,

$$z = \frac{X - \bar{X}}{s} = \frac{38 - 40}{5} = -0.4$$

For test B,

$$z = \frac{94 - 100}{10} = -0.6$$

The score for test A is relatively higher than the score for test B.

상자그림과 이상값

The Five-Number Summary and Boxplots

A **boxplot** can be used to graphically represent the data set. These plots involve five specific values:

1. The lowest value of the data set (i.e., minimum) 최소값
2. Q_1 25%
3. The median Q2: 50%
4. Q_3 75%
5. The highest value of the data set (i.e., maximum) 최댓값

These values are called a **five-number summary** of the data set.

Procedure Table

Constructing a Boxplot

Step 1 Find the five-number summary for the data.

Step 2 Draw a horizontal axis and place the scale on the axis. The scale should start on or below the minimum data value and end on or above the maximum data value.

Step 3 Locate the lowest data value, Q_1, the median, Q_3, and the highest data value; then draw a box whose vertical sides go through Q_1 and Q_3. Draw a vertical line through the median. Finally, draw a line from the minimum data value to the left side of the box, and draw a line from the maximum data value to the right side of the box.

상자그림과 이상값

EXAMPLE 3–37 Number of Meteorites Found

The number of meteorites found in 10 states of the United States is 89, 47, 164, 296, 30, 215, 138, 78, 48, 39. Construct a boxplot for the data.

Source: Natural History Museum.

SOLUTION

Step 1 Find the five-number summary for the data.

Arrange the data in order:

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Find the median.

30, 39, 47, 48, 78, 89, 138, 164, 215, 296

Median

$$\text{Median} = \frac{78 + 89}{2} = 83.5$$

Find Q_1.

30, 39, 47, 48, 78

Q_1

Find Q_3.

89, 138, 164, 215, 296

Q_3

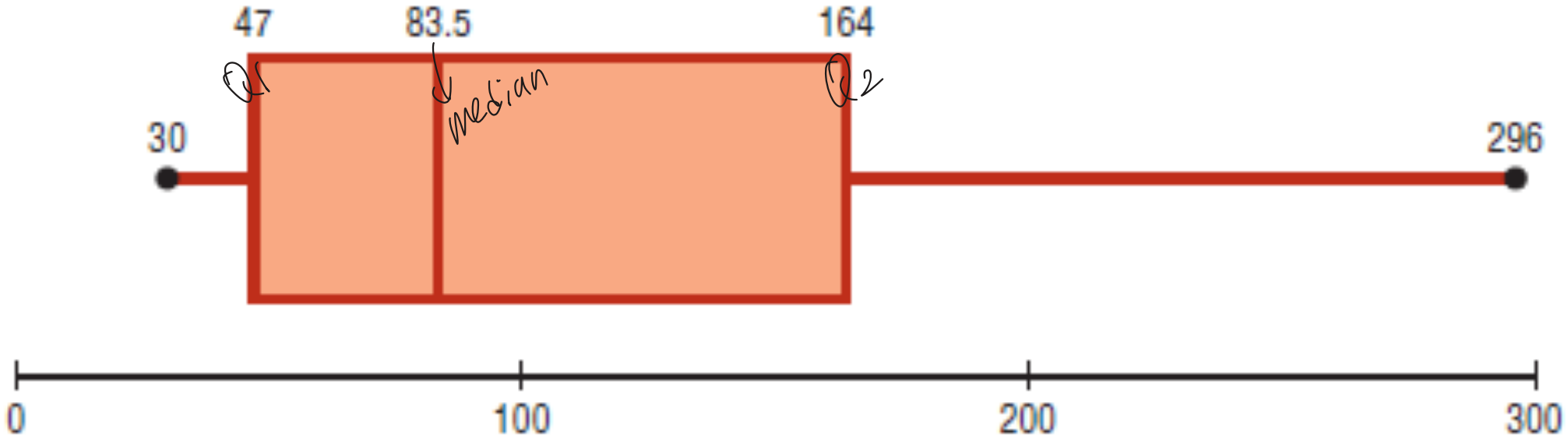
The minimum data value is 30, and the maximum data value is 296.

상자그림과 이상값

Step 2 Draw a horizontal axis and the scale.

Step 3 Draw the box above the scale using Q_1 and Q_3. Draw a vertical line through the median, and draw lines from the lowest data value to the box and from the highest data value to the box. See Figure 3–7.

FIGURE 3–7 Boxplot for Example 3–37

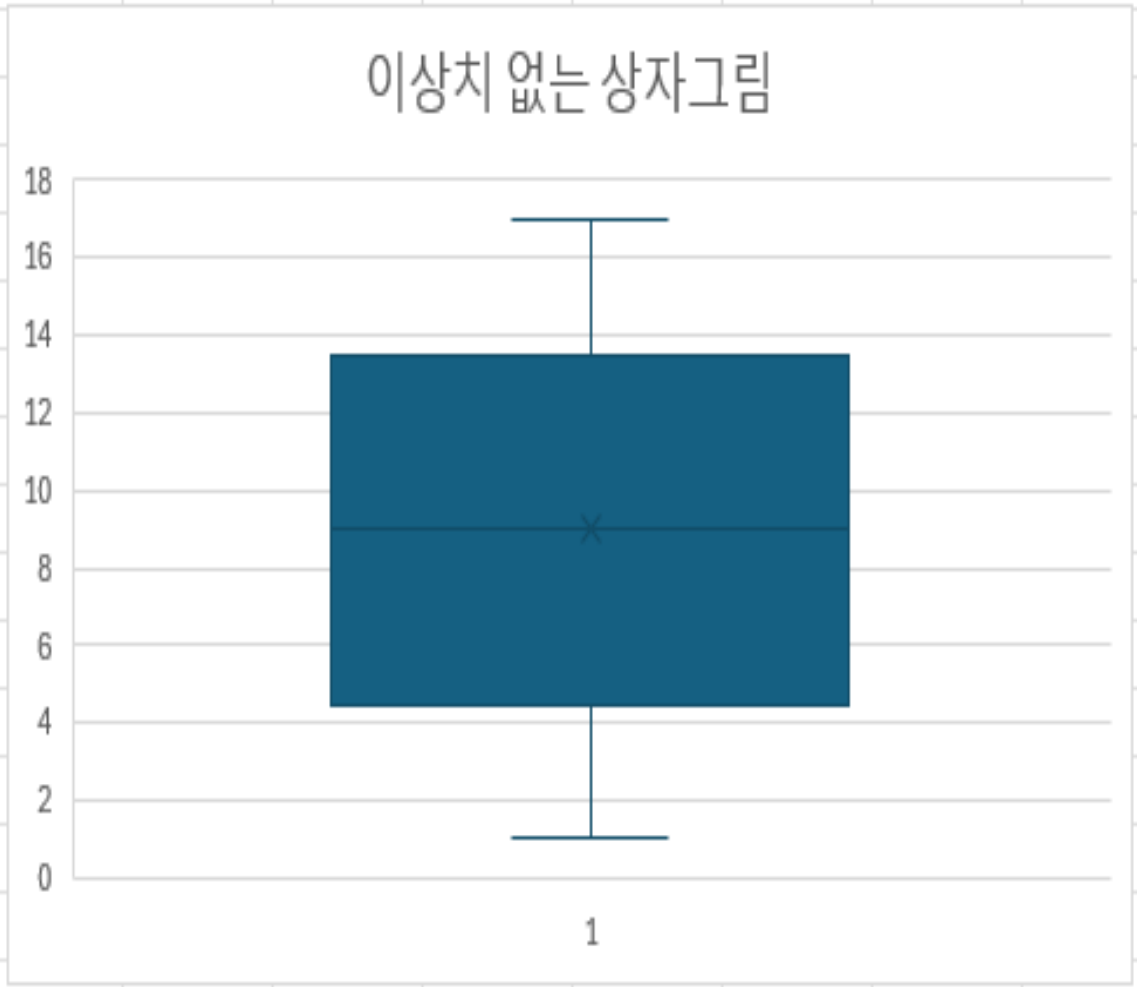


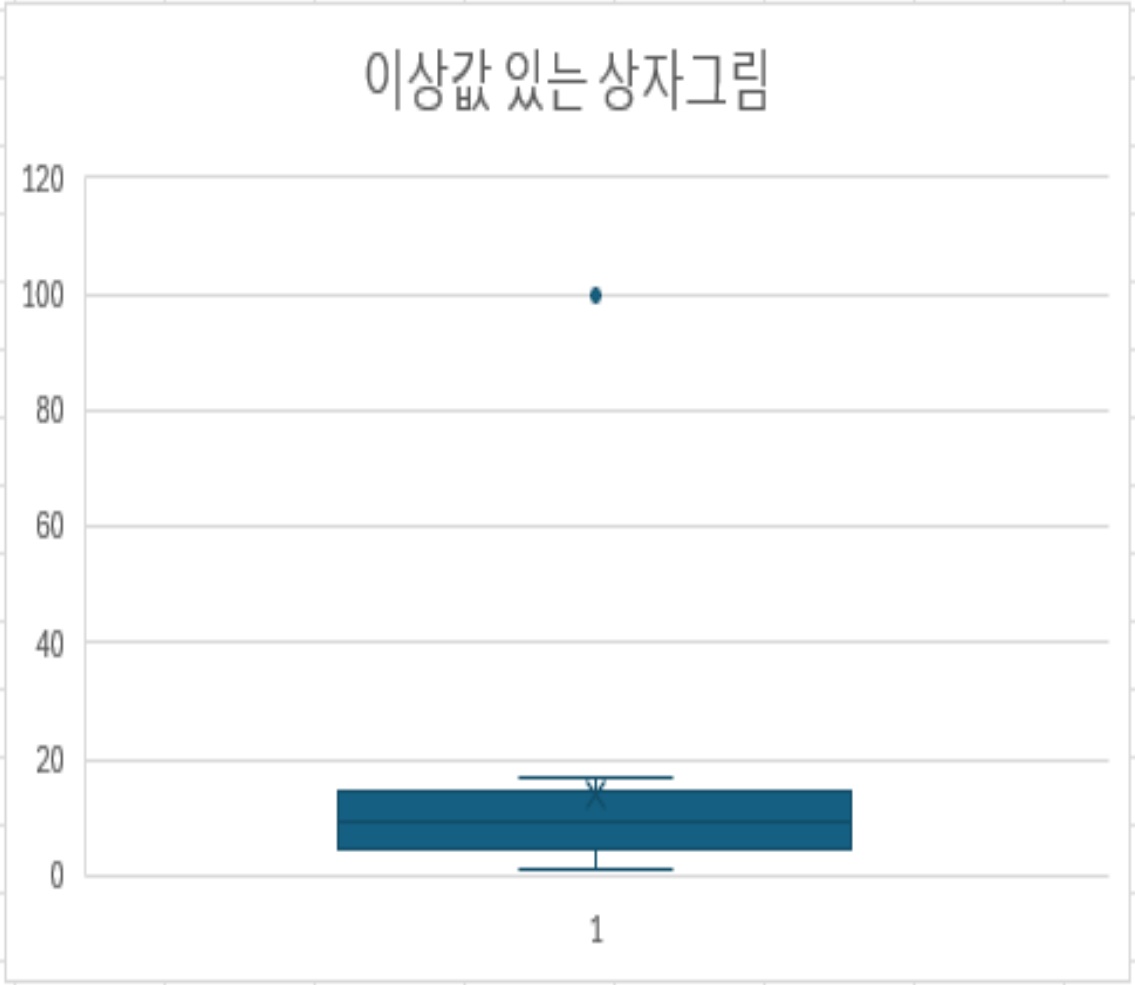
상자그림과 이상값

- 크기가 10 이상인 자료의 분포가 종형일 때 (정규분포와 같은 왜도와 첨도를 가질 때), 이 자료에서 표준점수가 3 이상 또는 -3 이하인 자료점

- 자료에서 **IQR = Q3(제3 사분위수) - Q1(제1 사분위수)**을 계산하고 측정값이 **Q1 - 1.5IQR** 보다 작거나 **Q3 + 1.5IQR**보다 크면 이상값으로 판단한다.

이상값과 상자그림


이상치 없는 상자그림
18
16
14
12
10
8
6
4
2
0
1


이상값 있는 상자그림
120
100
80
60
40
20
0
1

Ĝis revido!