탐색적 데이터 분석 1 : 기초 통계 처리 2

기술통계를 이용한 데이터 요약 방법

부산대학교 정보·의생명공학대학

정보컴퓨터공학부

부산대학교
PUSAN NATIONAL UNIVERSITY

Recap

❖ 3 main types of descriptive statistics:

- The distribution(분포) concerns

  the frequency(도수) of each value

  → Frequency Table (도수 분포표)

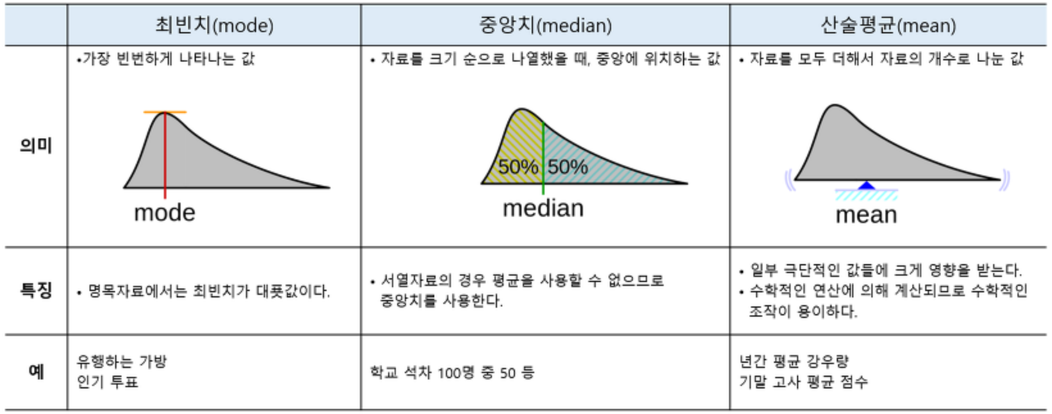
- The **central tendency(집중화 경향, 중심화 경향)**

  concerns the averages of the values

  → mode, median, mean

- The variability or **dispersion(분산, 산포도)** concerns

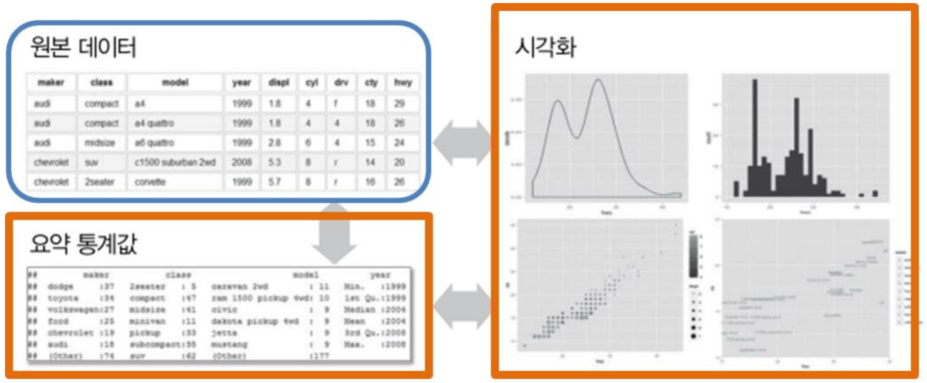
  how spread out the values are.



	최빈치(mode)	중앙치(median)	산술평균(mean)
의미	•가장 빈번하게 나타나는 값 mode	• 자료를 크기 순으로 나열했을 때, 중앙에 위치하는 값 50% 50% median	• 자료를 모두 더해서 자료의 개수로 나눈 값 mean
특징	• 명목자료에서는 최빈치가 대푯값이다.	• 서열자료의 경우 평균을 사용할 수 없으므로 중앙치를 사용한다.	• 일부 극단적인 값들에 크게 영향을 받는다. • 수학적인 연산에 의해 계산되므로 수학적인 조작이 용이하다.
예	유행하는 가방 인기 투표	학교 석차 100명 중 50 등	년간 평균 강우량 기말 고사 평균 점수

drhongdatanote.tistory.com

❖ 탐색적 데이터 분석 (Exploratory Data Analysis)



원본 데이터

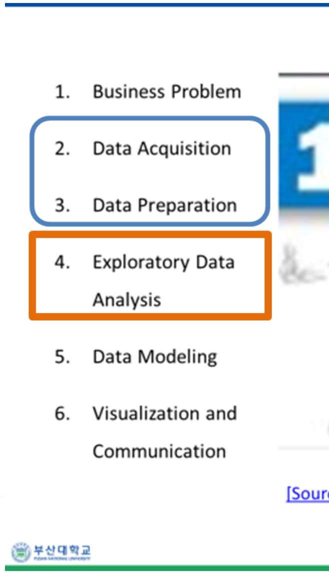
maker	class	model	year	displ	cyl	drv	cty	hwy
audi	compact	a4	1999	1.8	4	f	18	29
audi	compact	a4 quattro	1999	1.8	4	4	18	26
audi	midsize	a6 quattro	1999	2.8	6	4	15	24
chevrolet	suv	c1500 suburban 2wd	2008	5.3	8	r	14	20
chevrolet	2seater	corvette	1999	5.7	8	r	16	26

요약 통계값

```
##         maker            class                     model          year
##  dodge     :37   2seater    : 5   caravan 2wd          : 11   Min.   :1999
##  toyota    :34   compact    :47   ram 1500 pickup 4wd: 10   1st Qu.:1999
##  volkswagen:27   midsize    :41   civic                :  9   Median :2004
##  ford      :25   minivan    :11   dakota pickup 4wd    :  9   Mean   :2004
##  chevrolet :19   pickup     :33   jetta                :  9   3rd Qu.:2008
##  audi      :18   subcompact:35   mustang              :  9   Max.   :2008
##  (Other)   :74   suv        :62   (Other)              :177
```

▲ 탐색적 데이터 분석에는 원본 데이터, 요약 통계값, 시각화가 모두 필요하다.

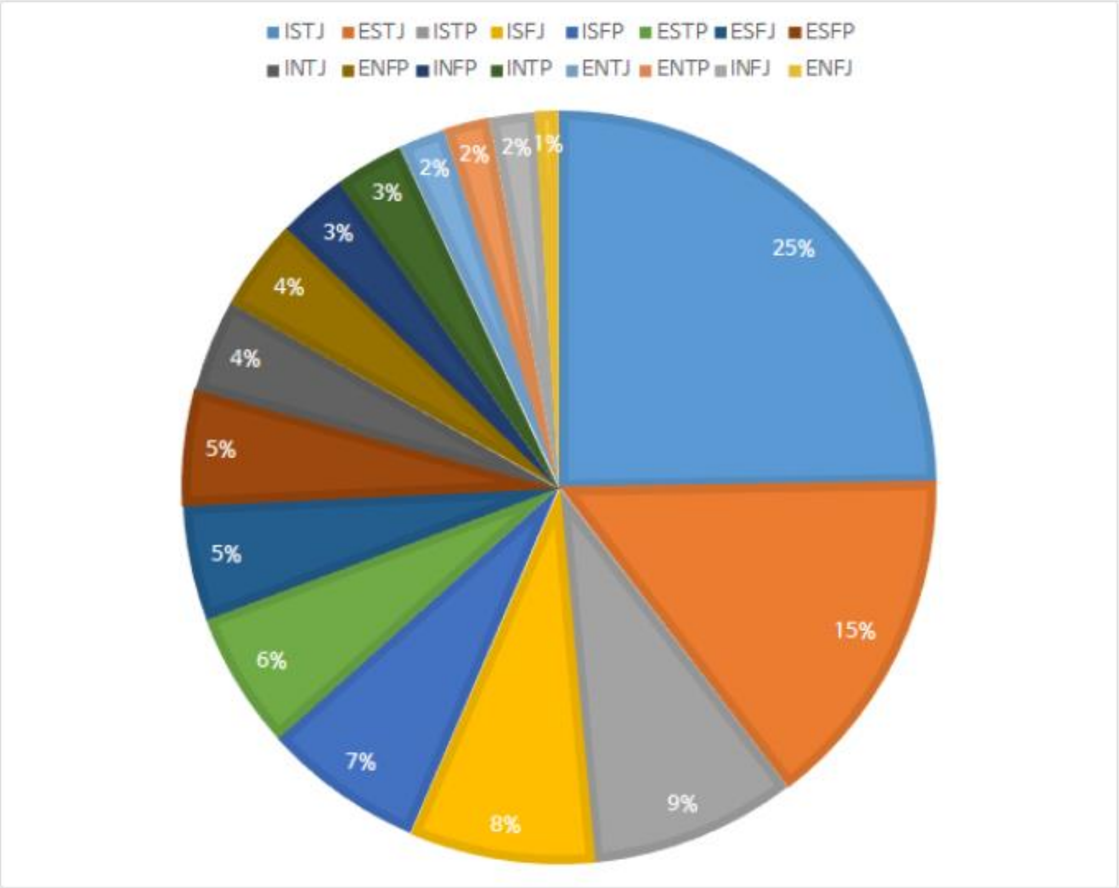
출처: 헬로 데이터 과학



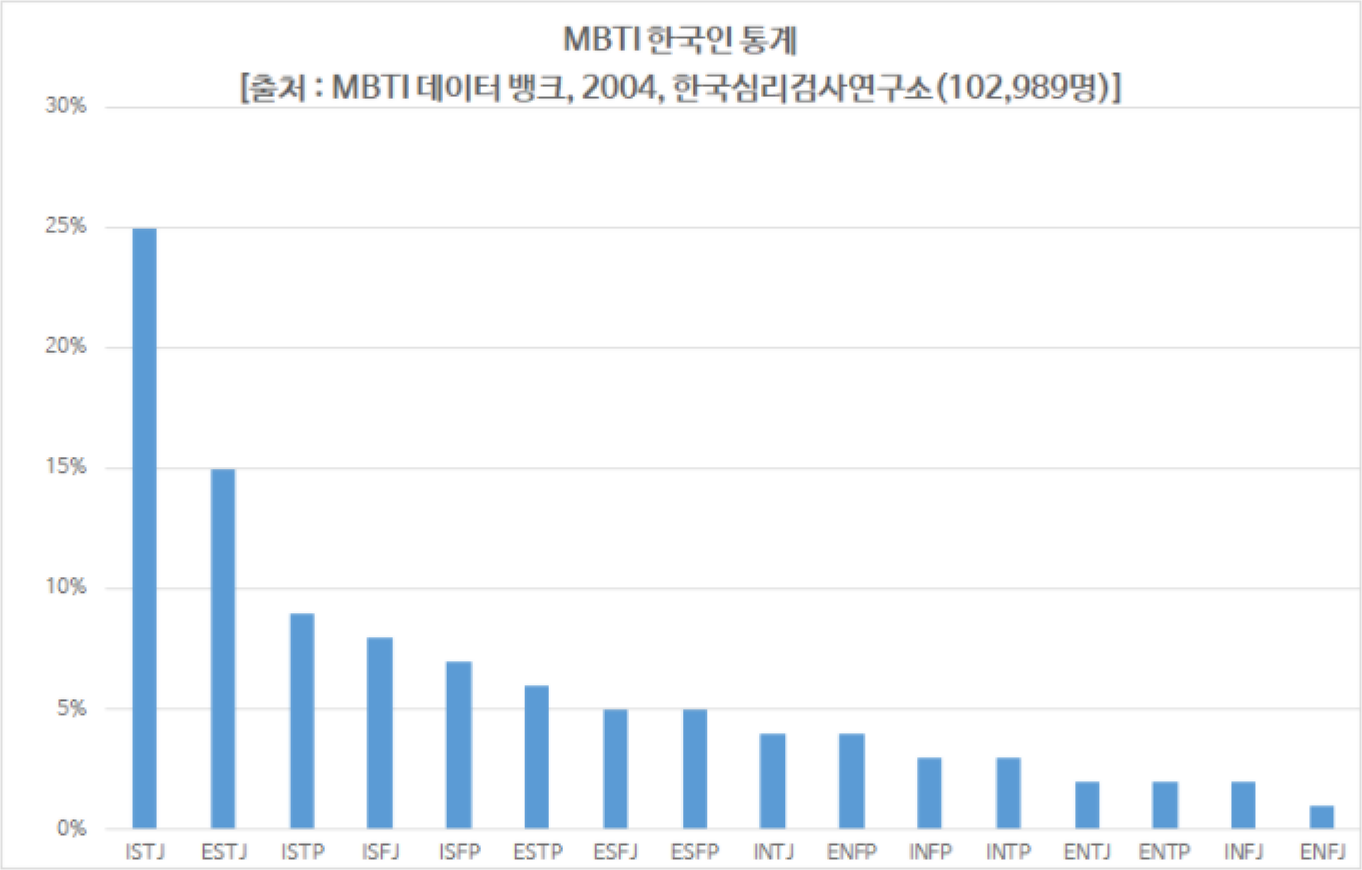
1. Business Problem
2. Data Acquisition
3. Data Preparation
4. Exploratory Data Analysis
5. Data Modeling
6. Visualization and Communication

[Sour

최빈값(Mode) 예



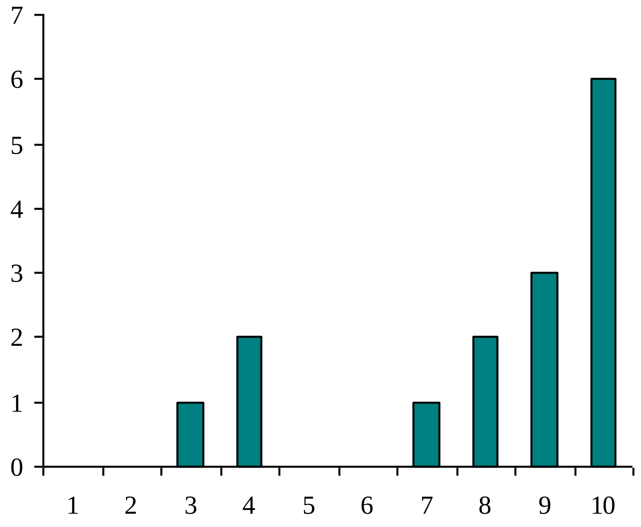
출처 : MBTI 데이터 뱅크, 2004, 한국심리검사연구소(102,989명)

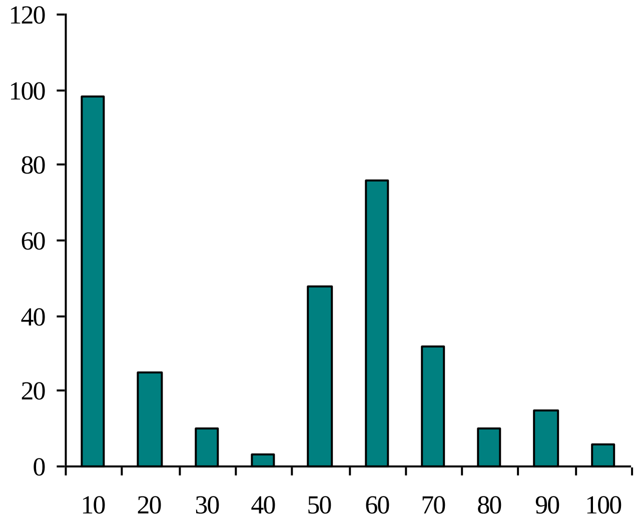


출처: https://m.blog.naver.com/taxandlove/221988052765

최빈값 단점

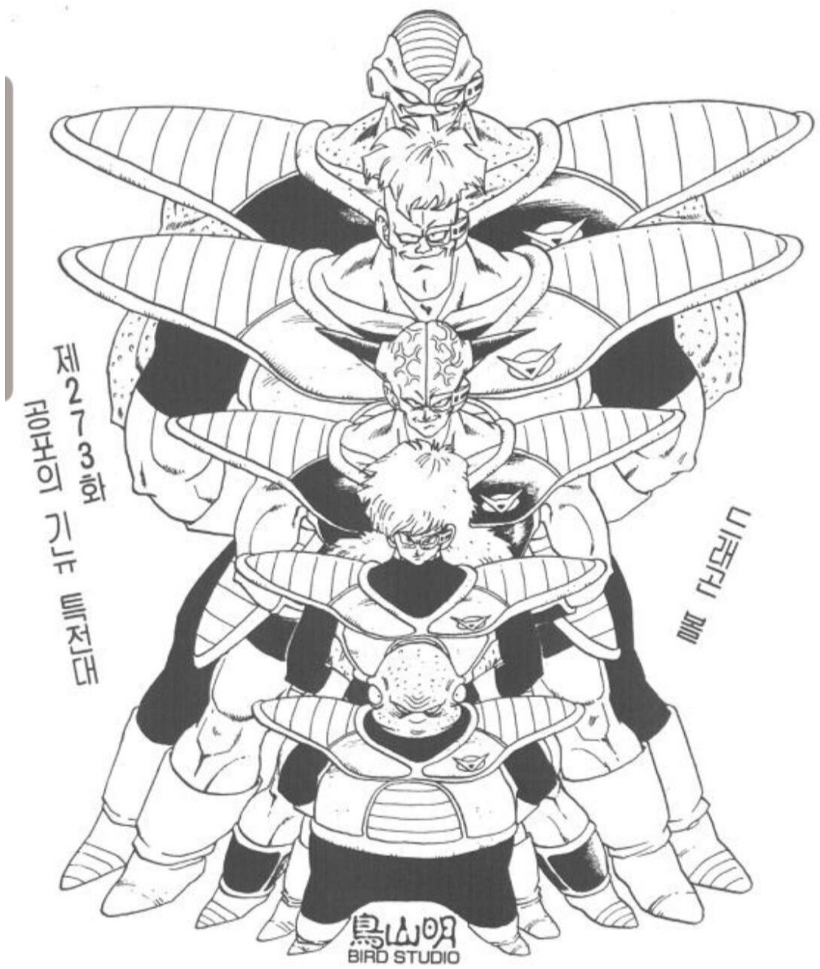
- The mode is not a very useful measure of central tendency
  - It is insensitive to large changes in the data set
    - That is, two data sets that are very different from each other can have the same mode





The Median (중앙값)

- ❖ the value that’s exactly in the middle of a data set when it is ordered.



How to find the median

1. Sort the data from highest to lowest
2. Find the score in the middle
   - If N is odd: middle = (N + 1) / 2
   - If N is even: the median is the average of the middle two scores

Finding the median with an odd-numbered data set

- What is the median of the following scores:

  10 8 14 15 7 3 3 8 12 10 9

- Sort the scores:

  15 14 12 10 10 9 8 8 7 3 3

- Determine the middle score:

  middle = (N + 1) / 2 = (11 + 1) / 2 = 6

- Middle score = median = 9

Finding the median with an even-numbered data set

- What is the median of the following scores:

  24 18 19 42 16 12

- Sort the scores:

  42 24 19 18 16 12

- Determine the middle score:

  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5

- Median = average of 3rd and 4th scores:

  (19 + 18) / 2 = 18.5

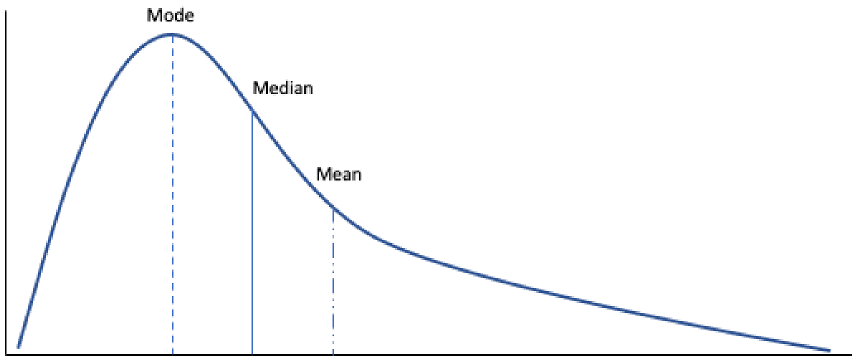
When to use the median

When should you use the median?

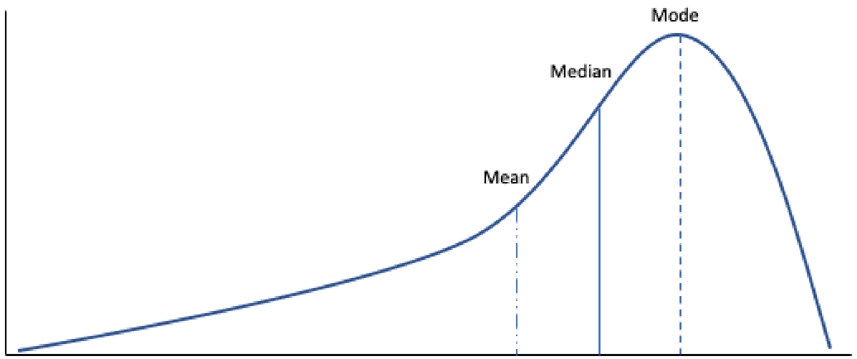
- The median is the most informative measure of central tendency for **skewed distributions or distributions with outliers**.
- In skewed distributions, more values fall on one side of the center than the other, and the mean, median and mode all differ from each other.
- In a positively skewed distribution, there's a cluster of lower scores and a spread out tail on the right.

The median can only be used on data that can be ordered – that is, from ordinal, interval and ratio levels of measurement.

Positively skewed distribution



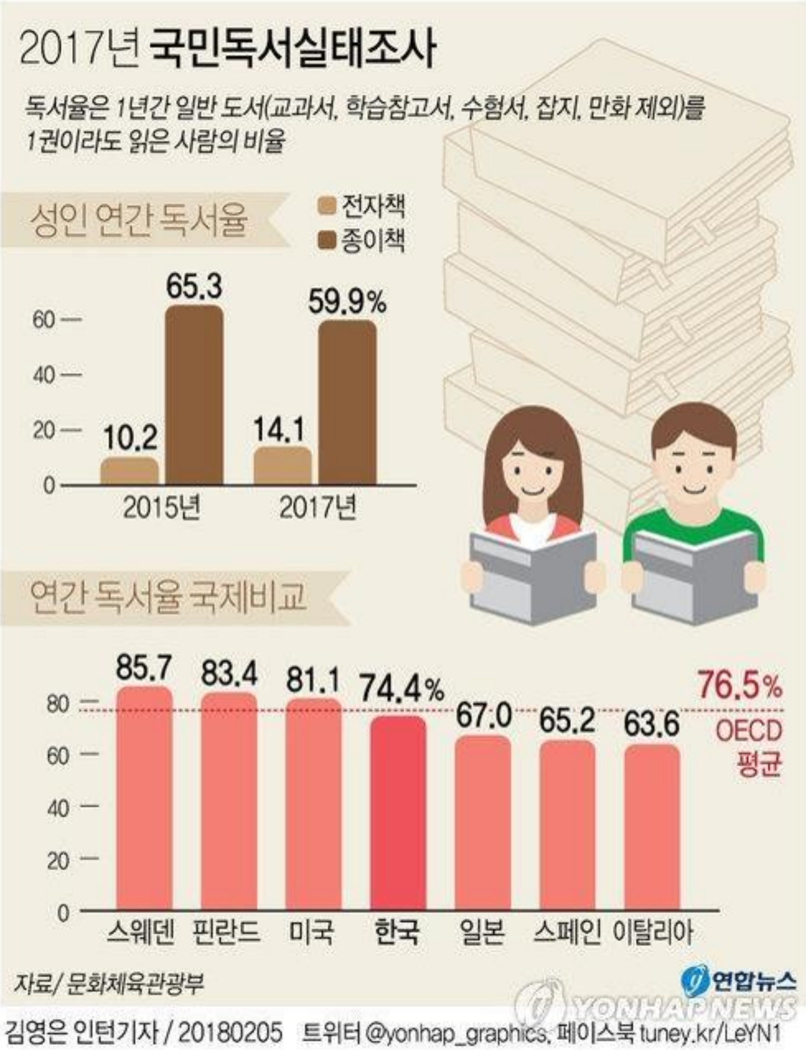
Negatively skewed distribution



The Mean (평균)

The mean, or arithmetic mean, of a data set is the sum of all values divided by the total number of values. It’s the most com monly used measure of central tendency and is often referred to as the “average.”

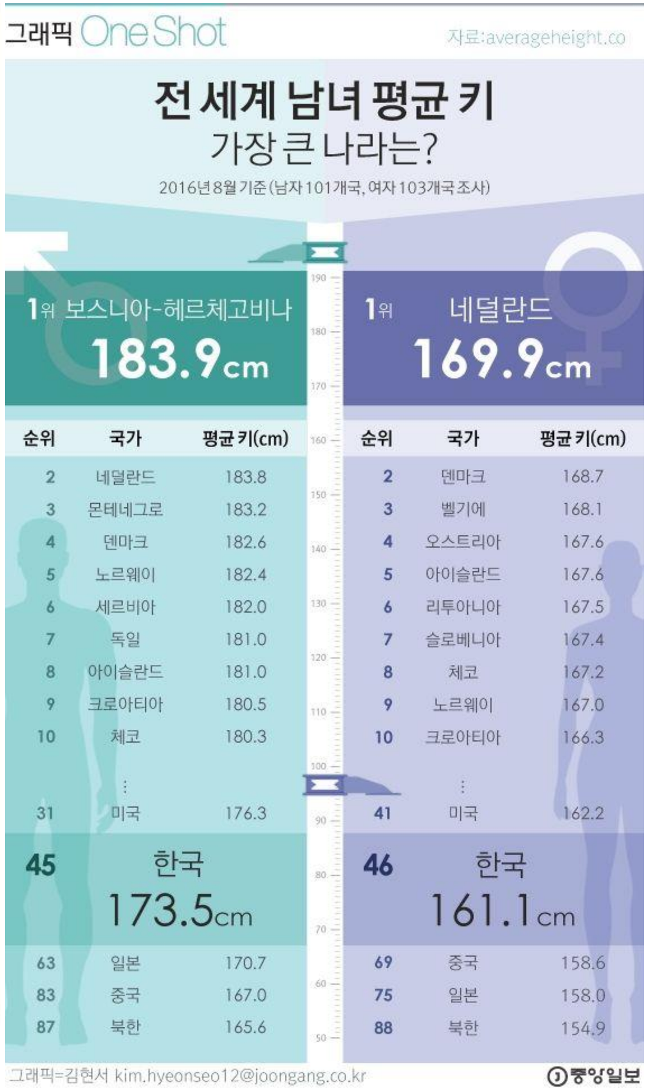
Population Mean	Sample Mean
$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$ *N* = number of items in the population	$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$ *n* = number of items in the sample


2017년 국민독서실태조사

독서율은 1년간 일반 도서(교과서, 학습참고서, 수험서, 잡지, 만화 제외)를 1권이라도 읽은 사람의 비율

자료/문화체육관광부

김영은 인턴기자 / 20180205 트위터 @yonhap_graphics, 페이스북 tuney.kr/LeYN1


그래픽 OneShot

자료:averageheight.co

전 세계 남녀 평균 키
가장 큰 나라는?

2016년 8월 기준 (남자 101개국, 여자 103개국 조사)

1위 보스니아-헤르체고비나 183.9cm

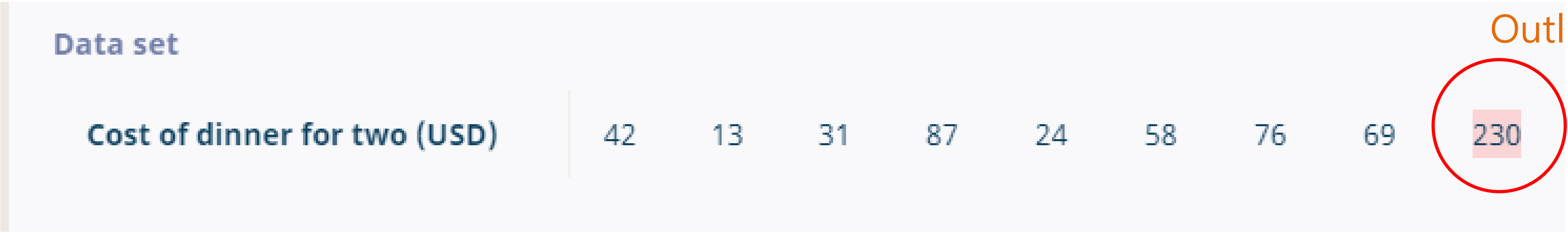
1위 네덜란드 169.9cm

순위	국가	평균 키(cm)	순위	국가	평균 키(cm)
2	네덜란드	183.8	2	덴마크	168.7
3	몬테네그로	183.2	3	벨기에	168.1
4	덴마크	182.6	4	오스트리아	167.6
5	노르웨이	182.4	5	아이슬란드	167.6
6	세르비아	182.0	6	리투아니아	167.5
7	독일	181.0	7	슬로베니아	167.4
8	아이슬란드	181.0	8	체코	167.2
9	크로아티아	180.5	9	노르웨이	167.0
10	체코	180.3	10	크로아티아	166.3
	⋮			⋮	
31	미국	176.3	41	미국	162.2
45	한국	173.5cm	46	한국	161.1cm
63	일본	170.7	69	중국	158.6
83	중국	167.0	75	일본	158.0
87	북한	165.6	88	북한	154.9

그래픽=김현서 kim.hyeonseo12@joongang.co.kr

중앙일보

Outlier(이상값) effect



Data set

Cost of dinner for two (USD)	42	13	31	87	24	58	76	69	230

Outlier

42 + 13 + 31 + 87 + 24 + 58 + 76 + 69 + 230 = **630**

$\bar{x} = 630 \div 9 = \mathbf{70}$

42 + 13 + 31 + 87 + 24 + 58 + 76 + 69 = **400**

$\bar{x} = 400 \div 8 = \mathbf{50}$

13, 24, 31, 42, **58**, 69, 76, 87, 230 Med = 58

13, 24, 31, **42, 58**, 69, 76, 87 Med = 50

최빈값, 중앙값, 평균 언제 사용?

- The mean can only be calculated for quantitative variables (e.g., height), and it can’t be found for categorical variables (e.g., gender).

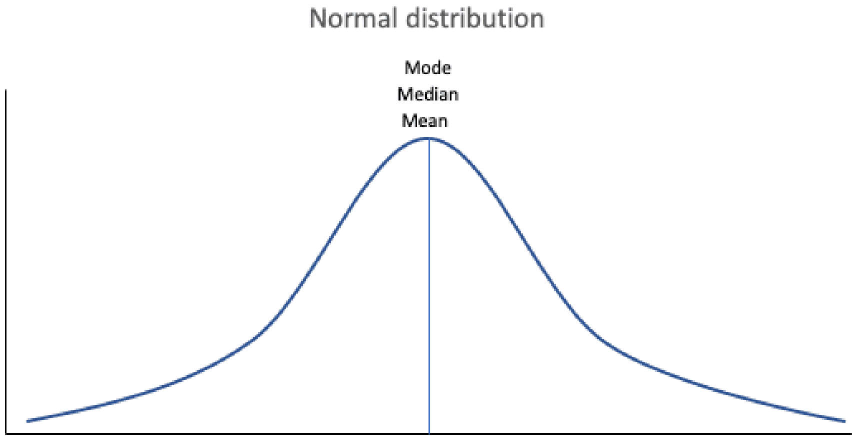
- In categorical variables, data is placed into groupings without exact numerical values, so the mean cannot be calculated. **For categorical variables, the mode is the best measure of central tendency** because it tells you the most common characteristic or popular choice for your sample.

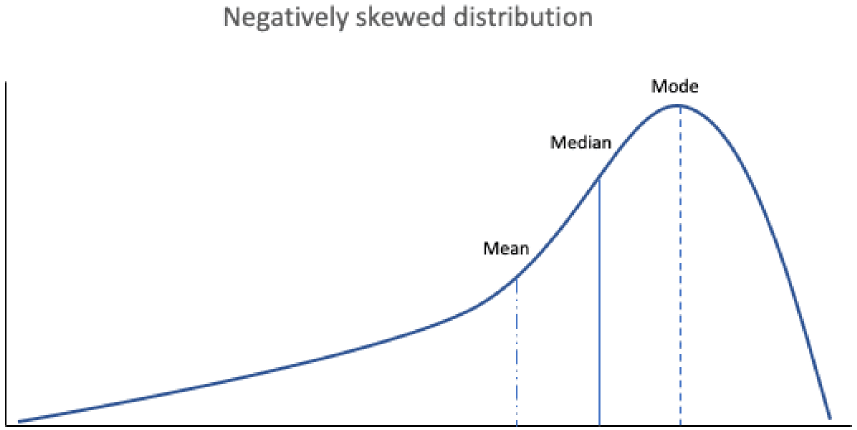
- But for continuous or discrete variables, you have exact numerical values. With these, you can easily calculate the mean or median.

최빈값, 중앙값, 평균 언제 사용?

- The mean is best for data sets with normal distributions. In a normal distribution, data is symmetrically distributed with no skew.

- For skewed distributions and distributions with outliers, the mean is easily influenced by extreme values and may not accurately represent the central tendency. The median is a better measure for these distributions as it takes a value from the middle of the distribution.

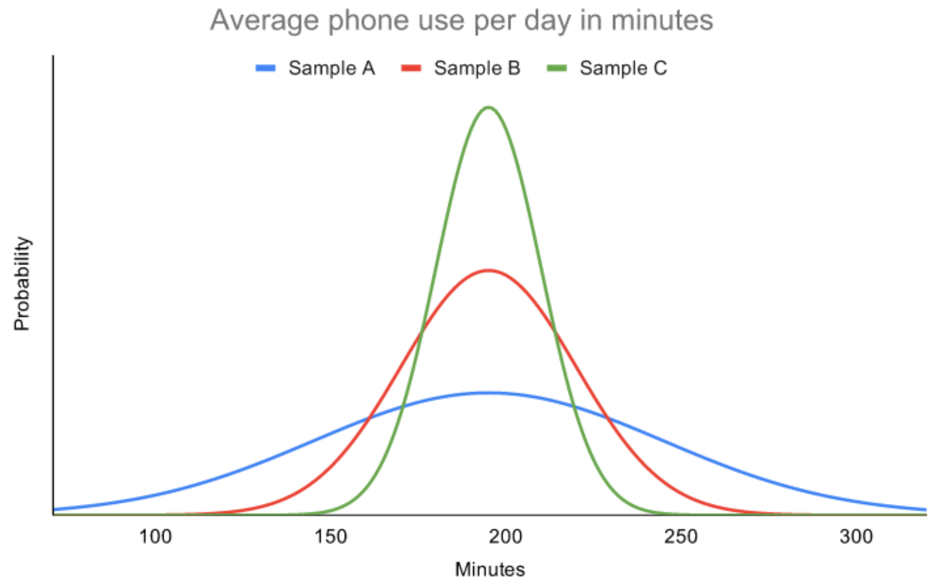




The variability or dispersion (산포도)

- 산포도: 자료가 퍼진 정도

- spread, scatter or dispersion.



- Measures of variability
  - Range(범위): the difference between the highest and lowest values
  - Interquartile range(사분위범위수): the range of the middle half of a distribution
  - Standard deviation(표준편차): average distance from the mean
  - Variance(분산): average of squared distances from the mean

Range (범위)

- = 최대값 – 최소값
- 이상값에 영향을 받음

Data (minutes)	72	110	134	190	238	287	305	324

The highest value (*H*) is **324** and the lowest (*L*) is **72**.

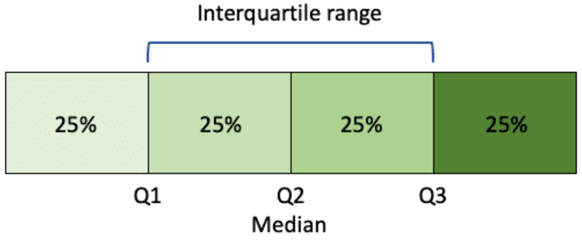
$R = H - L$

$R = 324 - 72 = \mathbf{252}$

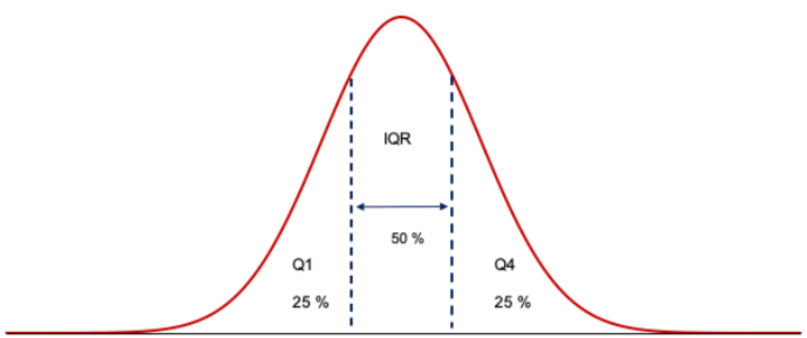
The range of your data is **252 minutes**.

Interquartile range (사분위수)

❖ Interquartile range = Q3 – Q1



Interquartile range on a normal distribution



Data	72	110	134	190	238	287	305	324

Q1 position: 0.25 x 8 = 2

Q3 position: 0.75 x 8 = 6

Q1 is the value in the 2nd position, which is **110**.

Q3 is the value in the 6th position, which is **287**.

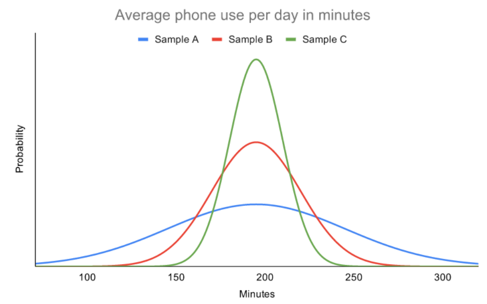
IQR = Q3 – Q1

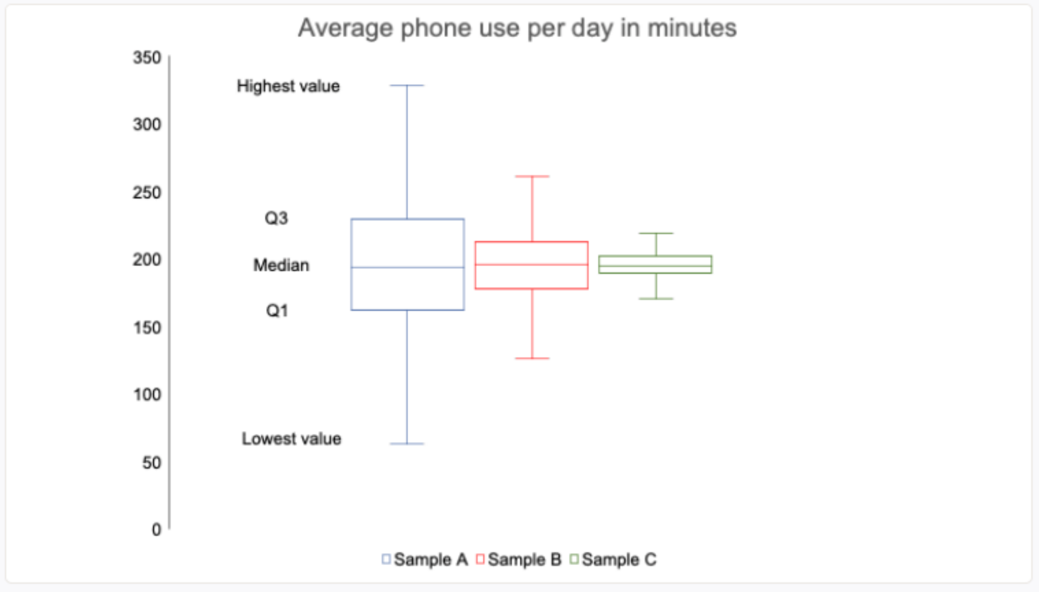
IQR = 287 – 110 = **177**

The interquartile range of your data is **177 minutes**.

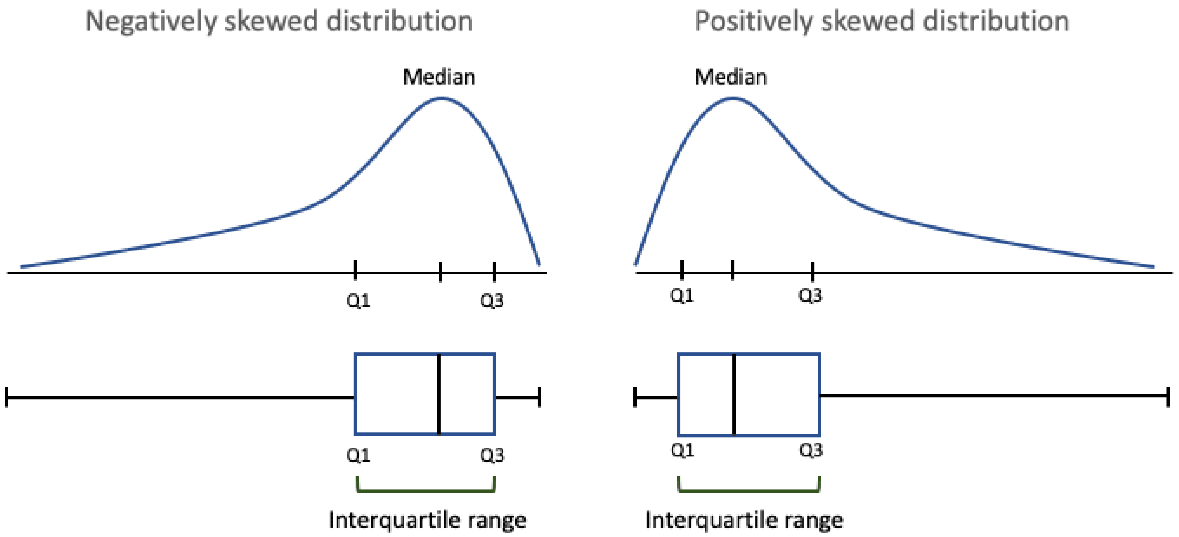
❖ Five-number summary

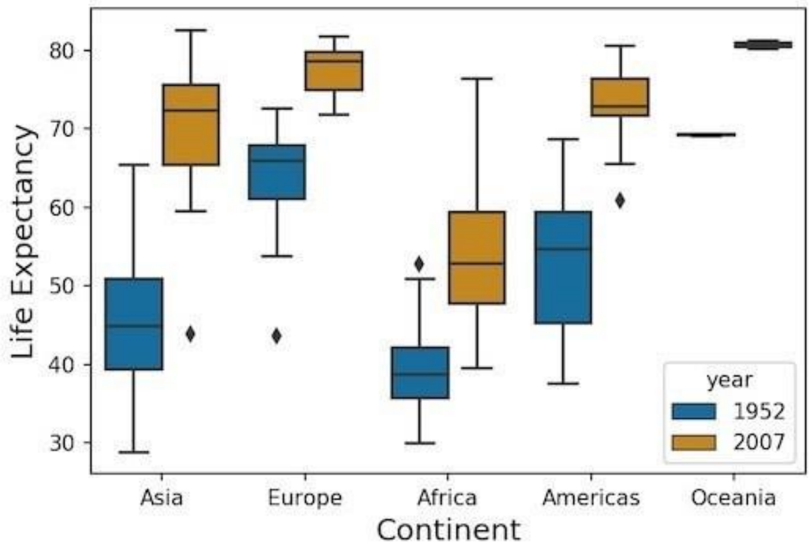
- Lowest value
- Q1: 25th percentile
- Q2: the median
- Q3: 75th percentile
- Highest value (Q4)





Skewed Distributions and their Boxplots (상자수염그래프)





Variance (분산) and Standard Deviation (표준편차)

- 분산: 대표적인 산포도의 측도
- 표준편차: 분산의 제곱근

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

모집단

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

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

샘플

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

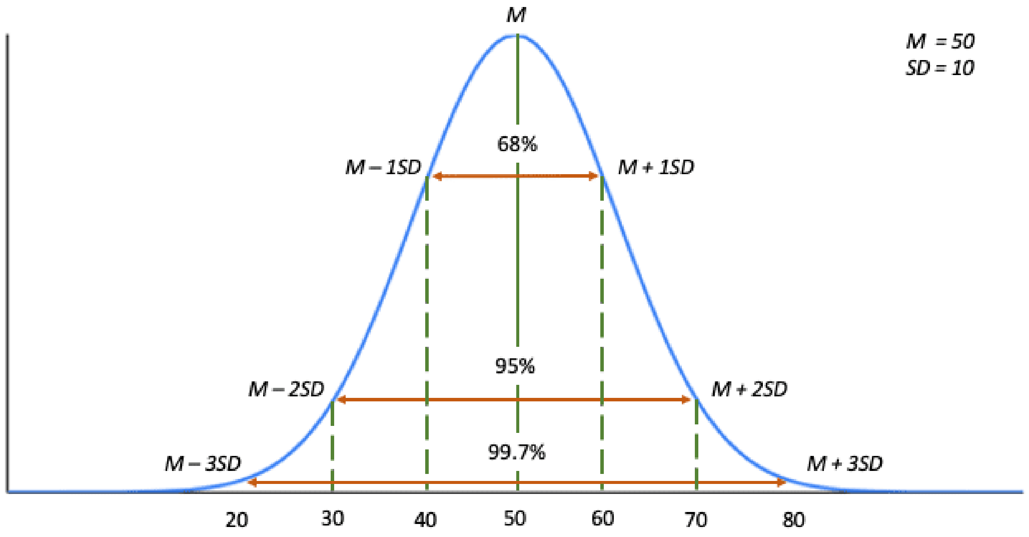
Both tells you, on average, how far each value lies from the mean.

A high standard deviation means that values are generally far from the mean, while a low standard deviation indicates that values are clustered close to the mean.

Since the units of variance are much larger than those of a typical value of a data set, it's harder to interpret the variance number intuitively. That's why standard deviatio n is often preferred as a main measure of variability.

표준편차와 자료의 범위

Standard deviations in a normal distribution



The empirical rule

The standard deviation and the mean together can tell you where most of the values in your distribution lie if they follow a normal distribution.

The **empirical rule,** or the 68-95-99.7 rule, tells you where your values lie:

- Around 68% of scores are within 2 standard deviations of th e mean,
- Around 95% of scores are within 4 standard deviations of th e mean,
- Around 99.7% of scores are within 6 standard deviations of the mean.

The best measure of variability?

- **The best measure of variability depends on your level of measurement and distribution.**
  - For ordinal data, the range and interquartile range are the only appropriate measures of variability.
  - For more complex interval and ratio levels, the standard deviation and variance are also applicable.

- **With regard to distribution**
  - For normal distributions, all measures can be used.
    The standard deviation and variance are preferred
    because they take your whole data set into account, but this also means that they are easily influenced by outliers.
  - For skewed distributions or data sets with outliers, the interquartile range is the best measure.
    It’s least affected by extreme values because it focuses on the spread in the middle of the data set.

THANK YOU!