기초 통계 I

- ADsP -

조상구
빅데이터과 경복대학교

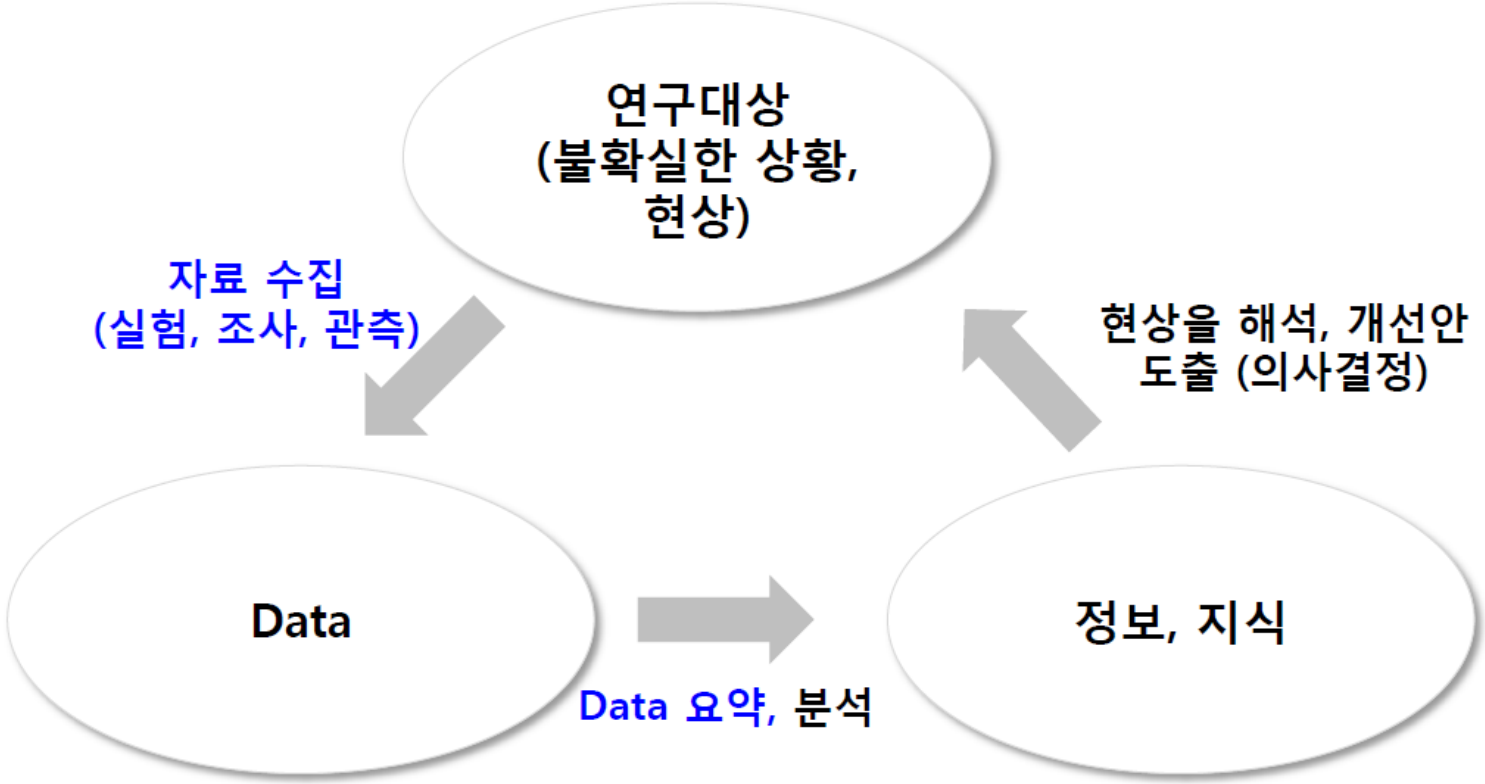
▪ 통계학 (Statistics) 이란?

- **자료의 수집 방법 (Sampling)**

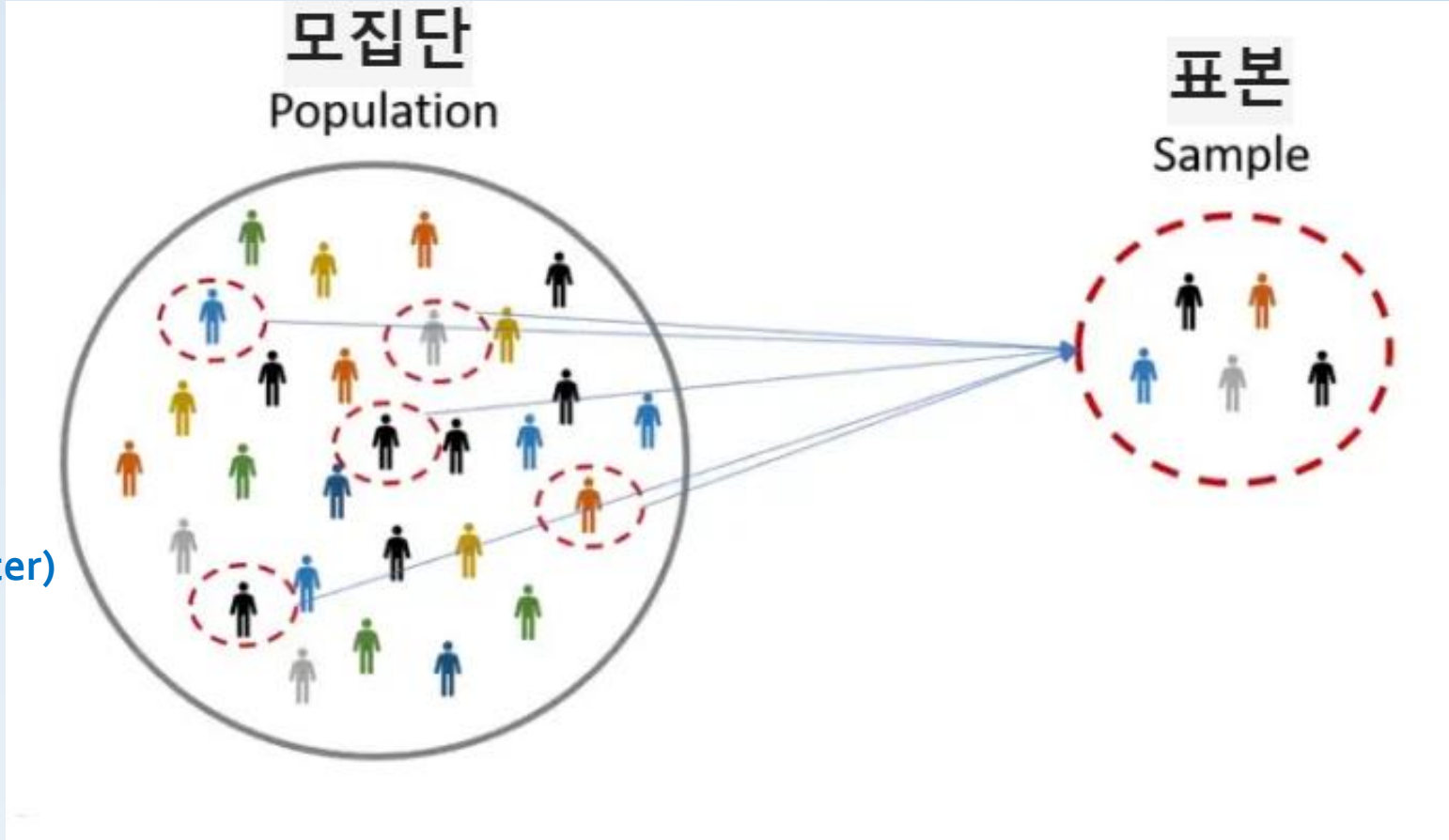
- **자료의 표현 및 요약 방법**
  (기술 통계학; descriptive statistics)

- 자료로부터 일반적인 성질을 끄집어 내는 방법
  (추론 통계학; inferential statistics)

Data 분석 과정


연구대상
(불확실한 상황,
현상)
자료 수집
(실험, 조사, 관측)
현상을 해석, 개선안
도출 (의사결정)
Data
Data 요약, 분석
정보, 지식

- 전수조사(Census)
- 표본조사(Sampling)



모수 (parameter)
- 모집단 평균
- 모집단 분산

모수 (parameter)
- 표본 평균
- 표본 분산 등

Sampling (1/3)

- **단순무작위추출법 (simple random sampling)**
  - 모집단의 모든 대상이 표본으로 선택될 확률이 동일한 추출방법
  - 다른 추출법의 기본이 되는 추출법
  - 모집단에 대한 사전 정보가 없는 경우 사용
  - 난수표 사용
  - **실습: 어떤 소비자 단체에서 담배의 니코틴 함유량을 조사하고자 한다. 100개의 모집단에서 임의로 10개의 sample을 단순무작위추출법을 이용해 sampling 해보자**

    ※ Tip) Excel 의 rand() 함수, roundup() 함수 사용

Sampling (2/3)

- **층화무작위추출법 (stratified random sampling)**
  - 모집단이 특정 기준에 따라 동질적인 몇 개의 집단으로 분류하고, 각 집단에서 무작위로 표본을 선택하는 방법
  - 집단의 특성을 모두 반영해야 하는 경우에 사용
  - 집단간의 성격이 이질적이나, 집단 내 대상간의 성격은 동질적이라고 가정
  - **실습: 어떤 소비자 단체에서 담배의 니코틴 함유량을 조사하고자 한다. A회사의 담배가 70개, B회사의 담배가 30개가 있다. 10개의 sample을 층화무작위추출법을 이용해 선택해보자**

Sampling (3/3)

- **집락추출법 (cluster sampling)**
  - 모집단이 여러 개의 집단으로 구성되어 있는 경우, 전체 집단 중 일부를 무작위로 선택한 뒤, 선택된 집단 모두를 표본으로 조사하는 방법
  - 집단간의 차이는 동질적, 집단내의 차이는 이질적이라고 가정함
  - 예) 고등학교 학생들의 몸무게 조사

통계 Data의 종류

명목형 (nominal)

- 어떤 대상에 숫자나 기호를 부여하여 구분하기 위해 사용
- 크기나 순서가 의미 없음
- 혈액형, 이름

서열형 (ordinal)

- 기준에 따라 대상을 서열화하여 숫자나 기호를 부여
- 순서가 의미 있으나 크기는 의미없음
- 품질등급, 학점, 올림픽 메달

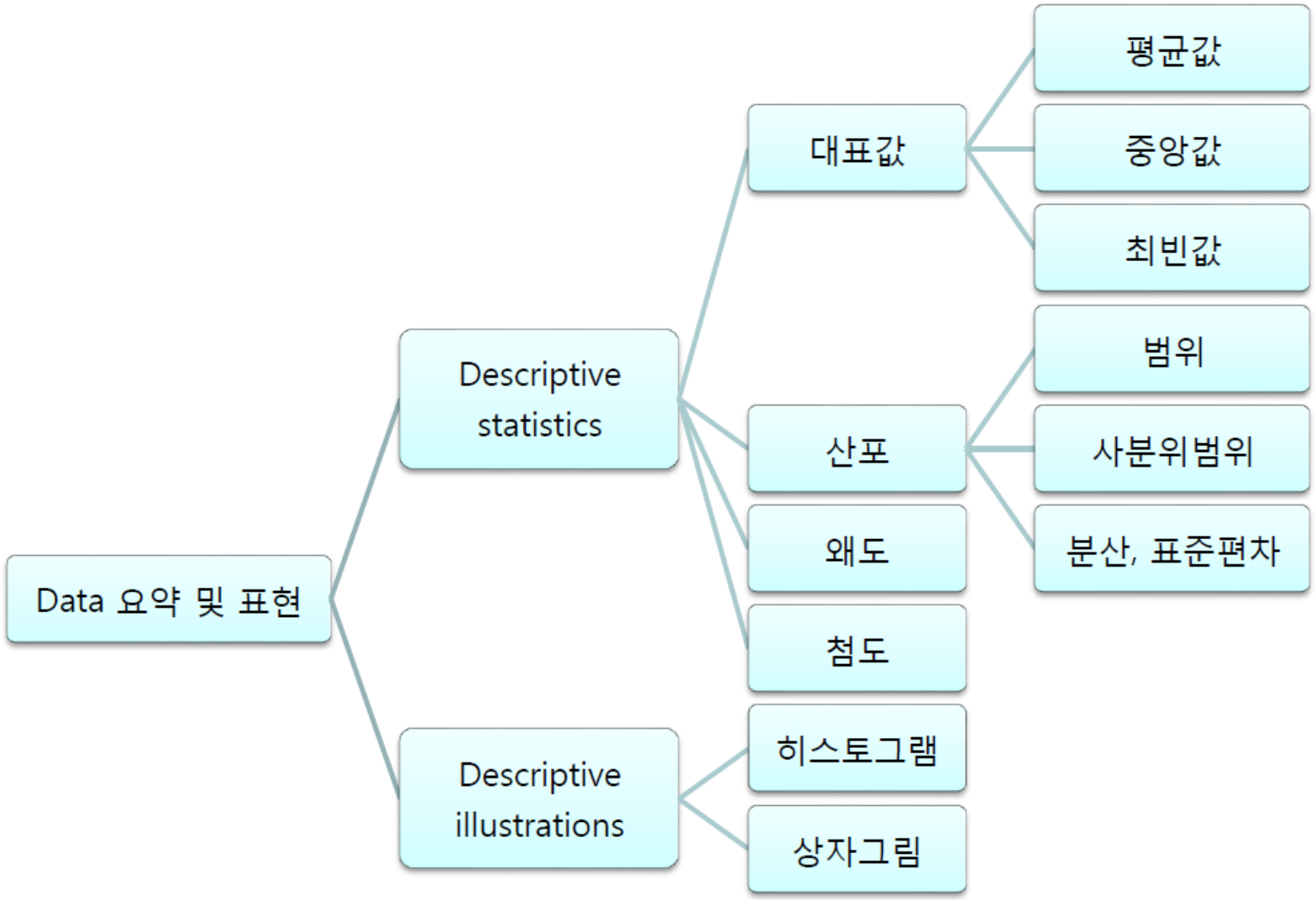
구간형 (interval)

- 대상들간의 상대적인 차이를 비교하는데 사용 (비율은 의미가 없음)
- 온도

비율형 (interval)

- 수치 자체가 실제적인 수량적 의미를 가짐
- 판매량, 무게

Data 요약 및 표현


Data 요약 및 표현
Descriptive statistics
Descriptive illustrations
대표값
산포
왜도
첨도
평균값
중앙값
최빈값
범위
사분위범위
분산, 표준편차
히스토그램
상자그림

Descriptive Statistics: 대표값

- **평균 (mean)**
  - 자료값들의 합(sum)을 표본 크기(관측치의 개수) n으로 나눈 것, 일반적으로 가장 널리 쓰이는 대표값
  - 극단적으로 크거나 작은 값 (outlier) 에 민감
  - Ex) 500, 489, 495, 493, 505, 248 → $\frac{500 + 489 + 495 + 493 + 505 + 248}{6} = 455$

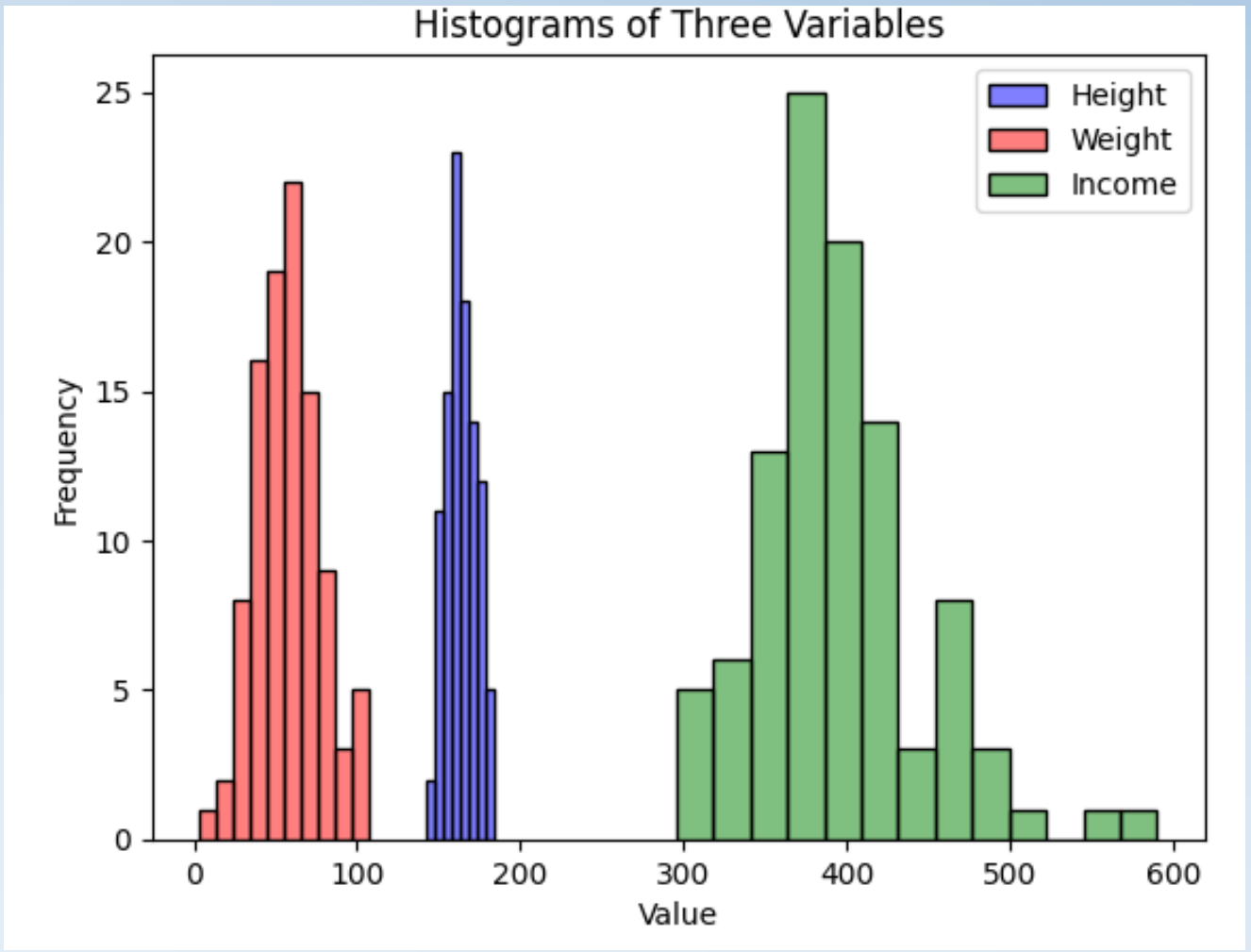
평균의 함정

- **중앙값 (median)**
  - 자료값들을 크기 순으로 정렬하였을 때 중앙에 위치하는 값, 평균에 비해 극단적으로 크거나 작은 값 (outlier) 에 민감하지 않음
  - Ex) 500, 489, 495, 493, 505, 248 → median = 494

- **최빈값 (mode)**
  - 자료값들 중 가장 빈도가 많은 자료값

	Height	Weight	Income
0	158.1	52.4	380.6
1	142.9	77.1	382.6
2	169.5	17.0	565.3
3	157.4	39.3	324.5
4	178.3	62.0	410.2
...	...	...	...
95	184.5	66.6	401.7
96	170.7	51.5	360.9
97	175.7	100.2	429.9
98	160.0	78.8	374.7
99	159.1	42.9	365.6

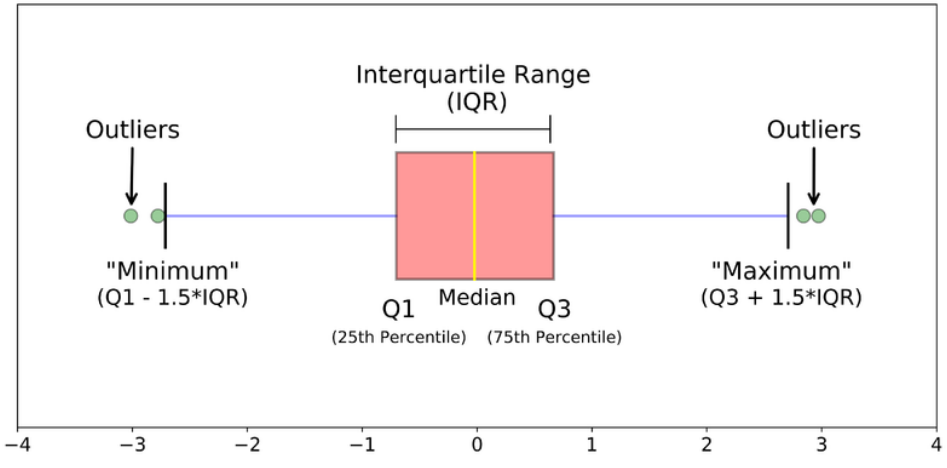
100 rows × 3 columns



Descriptive Statistics: 산포 (1/2)

- **범위 (Range)**
  - 최대값과 최소값의 차이
  - 극단적으로 크거나 작은 값 (outlier) 에 민감
  - Ex) 500, 489, 495, 493, 505, 248 → range = 257

- **사분위 범위 (Inter-quartile range; IQR)**
  - 사분위수 (quartile): 데이터를 크기순으로 나열하여 4등분 할 경우, 4등분되는 위치에 해당하는 값
    - Q1: 제1사분위수, 25%백분위수 (데이터의 25%에 해당하는 값)
    - Q2: 제2사분위수 (=중앙값), 50%백분위수 (데이터의 50%에 해당하는 값)
    - Q3: 제3사분위수, 75%백분위수 (데이터의 75%에 해당하는 값)
    - Ex) 1,2,3,4,5,6,7,8,9,10
      - Q1 = 3
      - Q2 = 5.5
      - Q3 = 8
    - 사분위 범위: Q3-Q1



Descriptive Statistics: 산포 (2/2)

- **분산 (Variance)**
  - 각 데이터들이 평균으로 부터 얼마나 떨어져 있는지 표현

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

  - Ex) 500, 489, 495, 493, 505, 248

$$s^2 = \frac{(500-455)^2 + (489-455)^2 + (495-455)^2 + (493-455)^2 + (505-455)^2 + (248-455)^2}{5} = 10314.8$$

- **표준편차 (Standard deviation)**
  - 분산에 제곱근을 취한 것으로 산포의 척도로 가장 널리 쓰임
  - $s = \sqrt{s^2}$
  - Ex) 500, 489, 495, 493, 505, 248 → s = 101.6

Descriptive Illustrations: Histogram (1/4)

- **히스토그램 (Histogram)**
  - 도수분포표에서 각 구간별 관측도수를 막대형태로 표현한 그래프
  - 도수분포표: 전체 data 범위를 구간으로 분할하고, 각 구간에 포함되는 데이터의 도수(개수)를 산출한 표
  - Histogram 작성 순서
    1. 구간 수 결정
    2. 구간 폭 결정 (구간폭 = 범위 / 구간수)
    3. 구간 경계치 결정
    4. 구간별 도수 산출
    5. Histogram 표현

Descriptive Illustrations: Histogram (2/4)

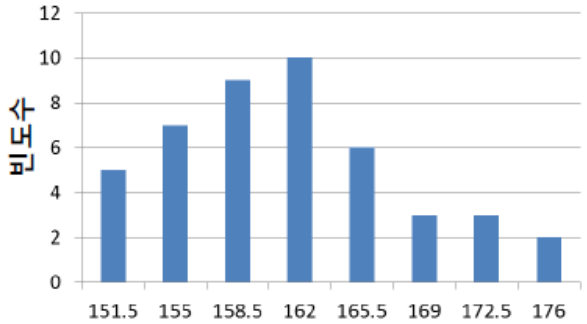
- **Example) 여대생 신장 자료**

170	151	154	160	158	154	171	156	160
157	160	157	148	165	158	159	155	151
152	161	156	164	156	163	174	153	170
149	166	154	166	160	160	161	154	163
164	160	148	162	167	165	158	158	176

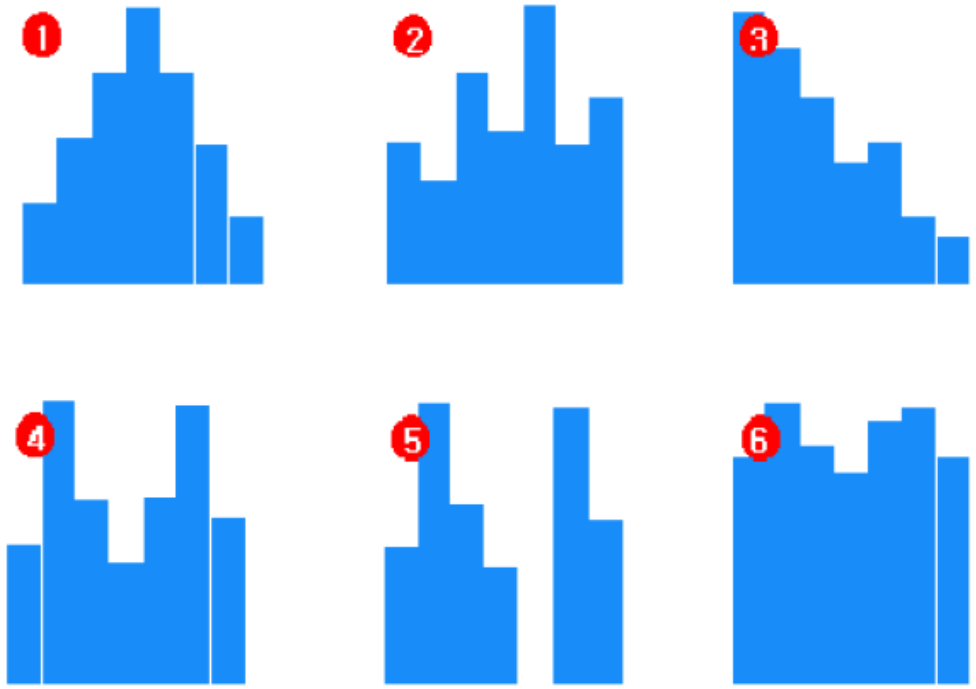
1. 구간 수 결정 — 8
2. 구간 폭 결정 — 3.5 (= (176-148)/8)
3. 구간 경계치 결정 — 151.5, 155, 158.5, …, 176
4. 구간별 도수 산출 — 5, 7, 9, …, 2
5. Histogram 표현

도수 분포표

구간	도수
151.5	5
155	7
158.5	9
162	10
165.5	6
169	3
172.5	3
176	2
Total	45



여대생 신장의 히스토그램

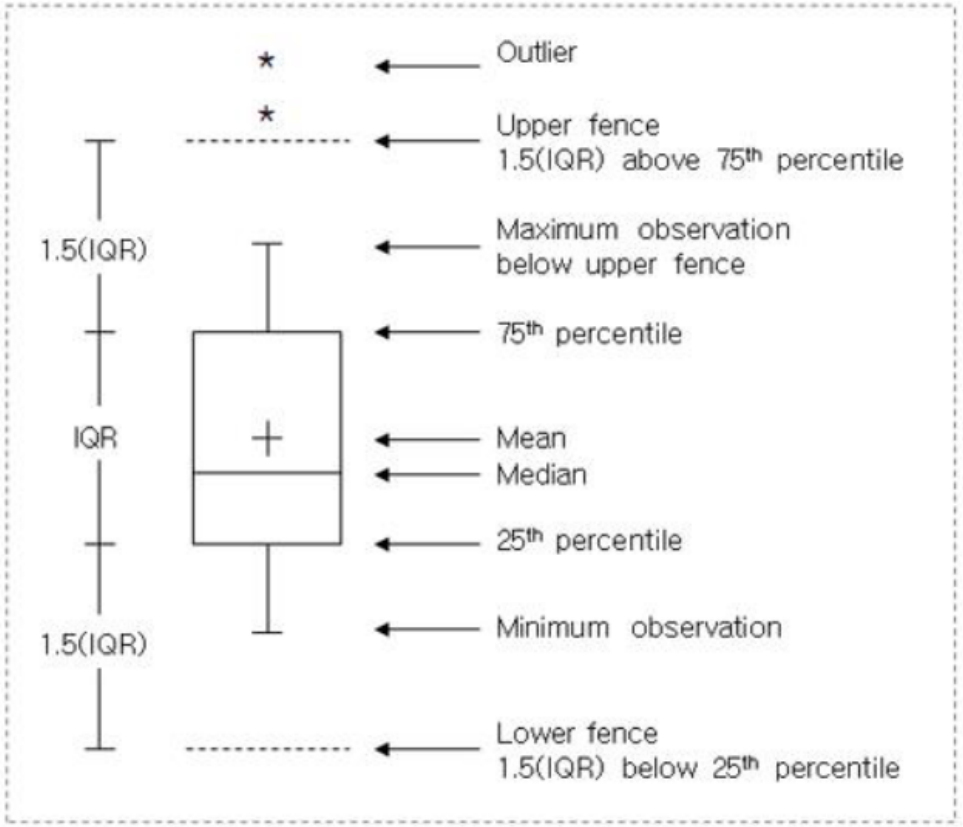


Source: http://www.allfortest.co.kr/customer.htm

(1) 일반형: 가장 흔하게 나타나는 분포로 도수가 중심 부근에 가장 많이 분포 되어있어 중심에서 멀어질수록 조금씩 작아진다. 거의 좌우 대칭이다.
(2) 이빠진형: 구간을 하나씩 걸러 도수가 작아진 분포의 모양이다. 이가 빠진 모양이 되고 있다. 이런 경우 구간 폭을 눈금의 정수 배로 한다든가, 측정자 읽는 법이 제대로 되었는가의 검토가 필요하다.
(3) 절벽형(왼쪽, 오른쪽): 평균값이 분포의 중심에서 극단적으로 한쪽에 치우쳐 있다. 이런 경우 측정속임수, 측정오차, 검사미스 등을 체크한다.
(4) 쌍봉우리형: 중심부근의 도수가 작아 산의 정상이 좌우로 나누어져 분포되어 있다. 이런 경우는 평균값이 다른 2개의 분포가 섞여 있을 경우에 나타나는데 층별한 히스토그램을 작성해보면 그 차이를 알 수 있다.
(5) 낙도형: 히스토그램의 왼쪽 끝이나 오른쪽 끝에 외딴 데이터가 나타난다. 이런 경우 데이터의 이력을 알아 보고 공정에 이상이 없는지, 다른 공정의 데이터가 들어와 있지 않은지 등을 조사한다.
(6) 고원형: 각 구간에 포함되어 있는 도수가 별 차이 없는 고원상태가 되고 있다. 이런 경우 층별한 히스토그램을 만들어 비교검토한다.

Descriptive Illustrations: Box plot (1/2)

- **상자 그림 (Box plot)**
  - Data의 분포에 대한 정보를 사분위수를 중심으로 나타내 주는 그림

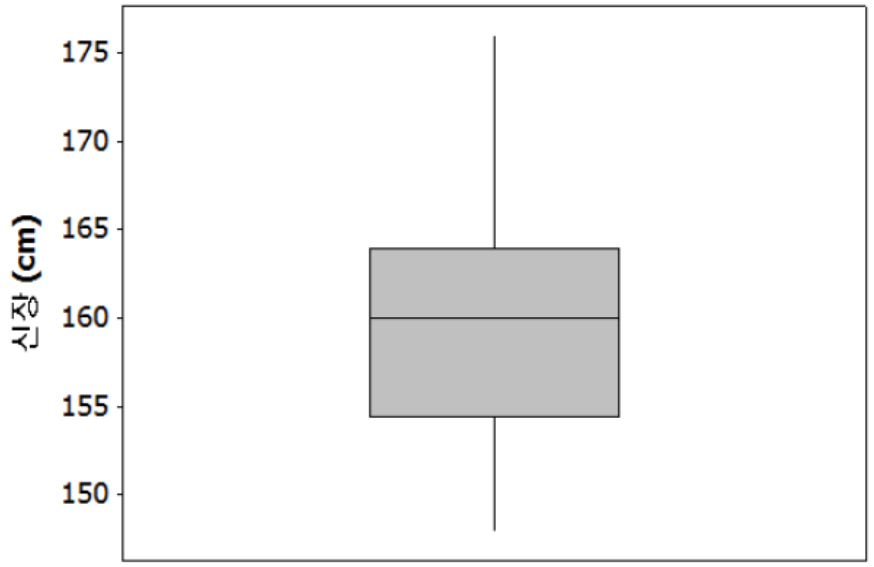


※ IQR : Inter Quantile Range

Descriptive Illustrations: Box plot (2/2)

- Example) 여대생 신장 자료

170	151	154	160	158	154	171	156	160
157	160	157	148	165	158	159	155	151
152	161	156	164	156	163	174	153	170
149	166	154	166	160	160	161	154	163
164	160	148	162	167	165	158	158	176



Summary

- **Sampling 방법**
  - 단순무작위추출법
  - 층화무작위추출법
  - 집락추출법

- **Descriptive statistics**
  - 대표값: 평균, 중앙값, 최빈값
  - 산포: 범위, 사분위범위, 분산, 표준편차
  - 왜도, 첨도

- **Descriptive illustration**
  - Histogram
  - Box plot