이공학도를 위한
확률과 통계학

- 1장-

Ki woon Moon

Protocol **E**ngineering **L**ab. **Sangmyung** University

2016. 1. 6

Content

- 통계학 기본 개념
- 표본추출 – 자료의 수집
- 위치의 측도 – 표본평균과 중앙값
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단

통계학

자료를 조사하여 유의미한 정보를 이끌어 내기 위하여
통계학적인 기법을 이용한 **데이터 분석 과정**을 거쳐야함

통계학

기본 개념

모집단
연구 대상이 되거나 조사의 대산이 되는 집단 전체

표본 집단
표본 조사를 통해 뽑아낸 모집단의 일부

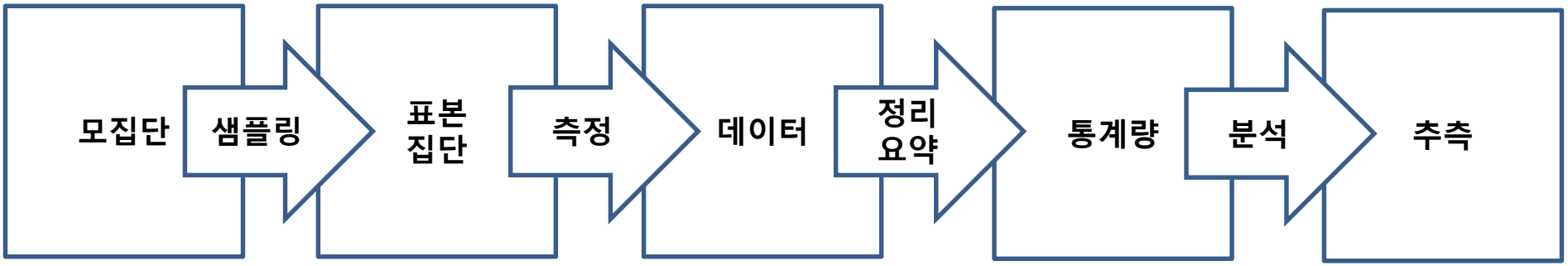
데이터
표본집단을 조사해서 얻은 수치 또는 자료

통계량
모집단의 특성을 추정하기 위하여 표본에서 계산한 주정량의 값

추측
통계학에서 임의 표본에 의하여 모집단을 규정하는 평균값, 분산 따위의 여러 수치를 추측하는 일

통계학

데이터 분석 과정



=한국인 전체

=한국인 일부

=표본집단에 속한 일부 한국인의 취침시간

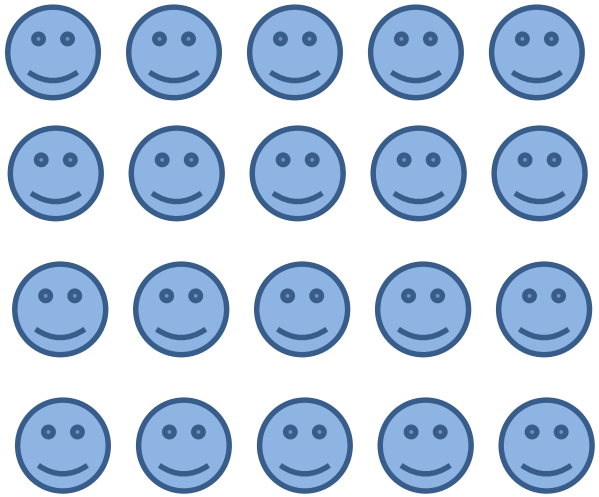
=표본 집단의 평균 취침 시간

='한국인의 취침시간은 23시 16분이 타당하다.'

표본 추출 – 자료의 수집

전수 조사

관심의 대상이 되는 집단을 이루는 모든 개체들을 조사하여 모집단 (조사하고자 하는 대상이 되는 집단 전체)의 특성을 측정하는 방법

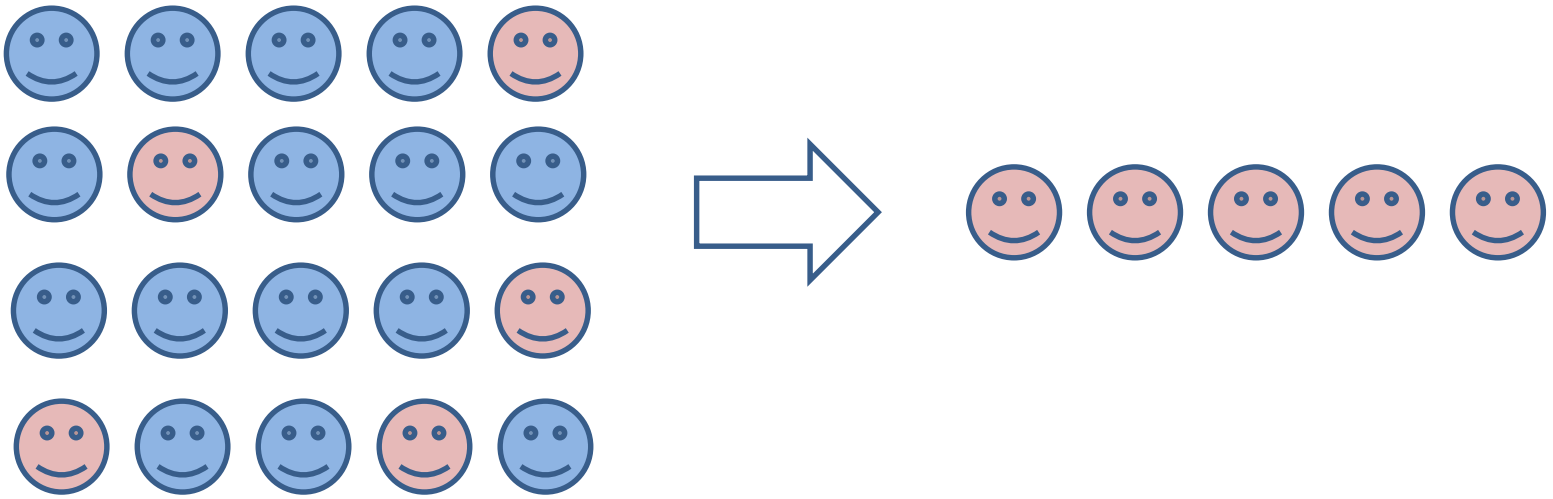


집단 내 모든 개체를 다 조사한다는 것은 현실적으로 불가능
많은 시간과 비용이 필요하며, 모든 조사를 전수조사로 할 수 없음

표본 추출 – 자료의 수집

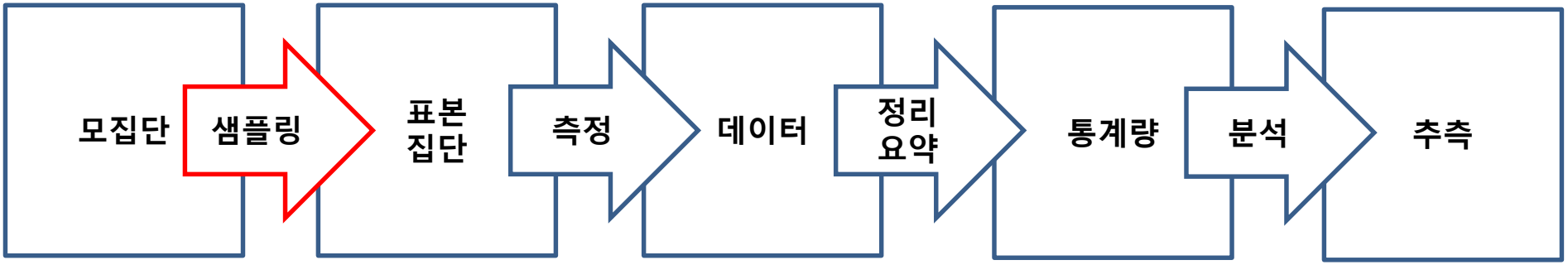
표본 조사

관심의 대상이 되는 전체 모집단 중 일부(표본)를 선택하고, 선택된 일부만을 대상으로 조사를 실시하여 이로부터 전체 모집단의 특성을 추정하는 방법



시간과 비용이 절감되고, 심도 있는 조사가 가능
모집단에서 추출한 소수의 표본이 전체 모집단의 특성을 잘 대표해야함

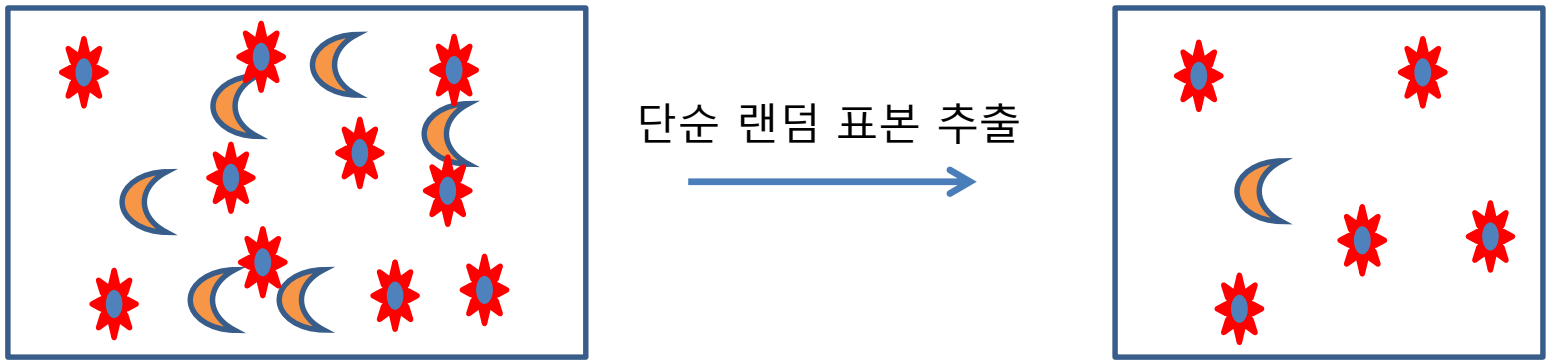
통계학



표본 추출 – 자료의 수집

단순 랜덤 표본 추출

모집단 전체의 일련 번호를 부여하여 표본조사 틀을 만든 후, 난수표를 활용하여 각 개체가 뽑힐 가능성이 동일하게 되게끔 표본을 추출하는 방법



1. 조사 대상자 전체에 일련번호를 부여
2. 난수표 E는 컴퓨터를 활용하여 필요한 표본수 만큼 난수를 생성
3. 생성된 난수에 해당하는 일련번호를 가진 것을 표본으로 선정

표본 추출 – 자료의 수집

단순 랜덤 표본 추출

- 장점

  - 모집단에 대한 사전 지식 불필요
  - 추출기회가 동등하고 독립적 이므로 추출된 표본의 대표성이 높음
  - 자료의 분류에 있어 오차의 개입이 적음

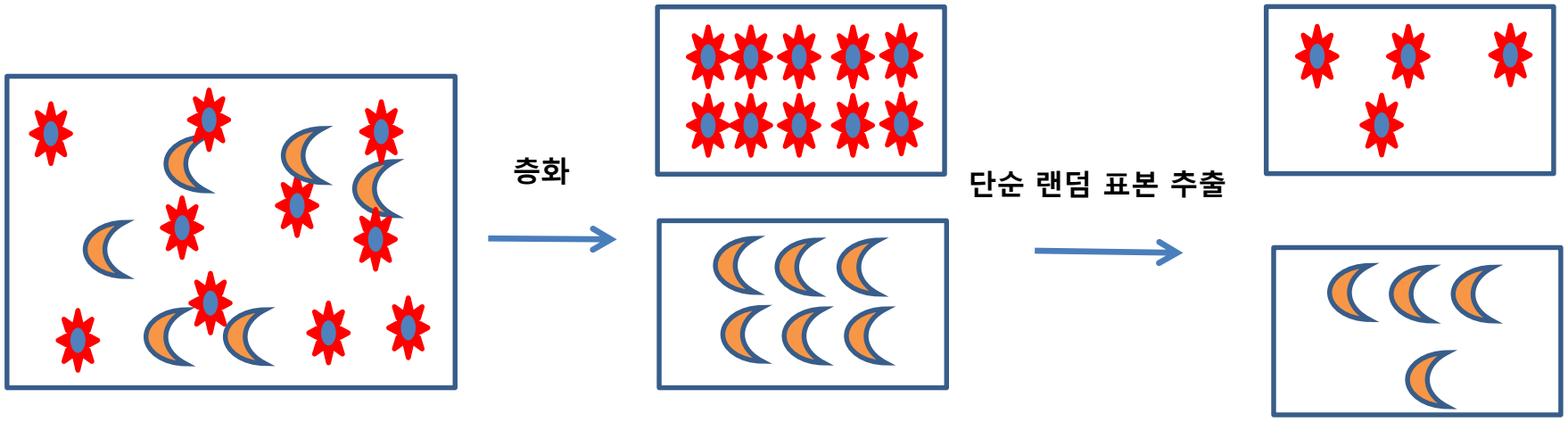
- 장점

  - 비교적 표본의 규모가 커야 함
  - 표본 프레임 작성이 어려움

표본 추출 – 자료의 수집

층화 랜덤 표본 추출

모집단을 먼저 서로 겹치지 않는 여러 개의 층으로 분할한 후, 각 층에서 단순 랜덤 표본 추출에 따라 표본을 추출하는 방법



표본 추출 – 자료의 수집

단순 랜덤 표본 추출

- 장점

  - 중요 집단은 빼놓지 않고 표본을 포함 시킬 수 있음
  - 표본의 수가 적어도 대표성이 높음
  - 각 층의 특성에 대한 추정과 비교가 가능

- 장점

  - 원형으로 복귀가 어려움
  - 층화시 시간과 노력이 소요
  - 모집단에 대한 지식이 필요

표본 추출 – 자료의 수집

실험 계획법

해결하고자 하는 문제에 대하여, 실험을 어떻게 행하고, 데이터를 어떻게 취하며, 어떠한 통계적 방법으로 데이터를 분석하면 최소 의 실험회수에서 최대의 정보를 얻을 수 있는가를 계획하는 것

실험 계획법의 목적

- ✓ 어떤 요인이 특성치의 변화에 유의한 영향을 주고있는가를 파 악하고, 그 영향이 양적으로 어느정도 큰가를 알아보기 위하 여 (검정과 추정의 문제)
- ✓ 취급된 인자들로는 설명이 되지 않는 오차변동은 어느정도 큰가를 알아내기 위하여 (오차항 추정의 문제)
- ✓ 인자들의 어떠한 수준에서 가장 바람직한 특성치를 얻을 수 있는가를 발견하기 위하여 (최적조건의 결정문제)

위치의 측도: 표본평균과 중앙값

위치 측도

자료들이 어떤 값을 기준으로 하여 어떤 형태로 분포되어 있는지를 나타내는 측도

표본 평균

평균은 단순히 산술 평균을 의미

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

위치의 측도: 표본평균과 중앙값

위치 측도

자료들이 어떤 값을 기준으로 하여 어떤 형태로 분포되어 있는지를 나타내는 측도

중앙값

극단값 이나 특이점에 영향을 받지 않고 자료의 중심경향을 측정하기 위한 값

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{n이 홀수 일 때} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{,n이 짝수 일 때} \end{cases}$$

위치의 측도: 표본평균과 중앙값

위치 측도

자료들이 어떤 값을 기준으로 하여 어떤 형태로 분포되어 있는지를 나타내는 측도

절사 평균

자료 중의 양쪽 끝에 있는 값을 제거하고 평균을 구한 값

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{n이 홀수 일 때} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{,n이 짝수 일 때} \end{cases}$$

산포의 측도

산포 측도

자료들이 중신위치에서 얼마나 떨어져 있는지를 나타내는 측도

범위

자료 중 가장 큰 값과 가장 작은 값의 차이

$$r = x_{(n)} - x_{(1)}$$

산포의 측도

산포 측도

자료들이 중신위치에서 얼마나 떨어져 있는지를 나타내는 측도

표본 분산

평균을 중심으로 자료들이 흩어진 정도가 어느 정도인지를 측정하는 것

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

산포의 측도

산포 측도

자료들이 중신위치에서 얼마나 떨어져 있는지를 나타내는 측도

표본 표준 편차

분산과 별 차이는 없으나, 자료와 같은 단위를 맞춰주기 위해 양의 제곱근을 취함

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

통계적 모형화, 과학적 조사, 그래프 진단

줄기-잎 그림

줄기와 잎을 이용하여 자료를 나타낸 그림

줄기는 자료값에서 자릿값이 높은 것이 위치
잎은 자료값에서 자릿값이 낮은 것이 위치

예) 점수가 34점 일 때

3 | 4

예) 횟수가 16회 일 때

1 | 6

통계적 모형화, 과학적 조사, 그래프 진단

줄기-잎 그림

줄기와 잎을 이용하여 자료를 나타낸 그림

줄기	잎
6	3 5
7	5 4 7
8	2 6 0 9 8 0 3
9	2 3 5 1 4

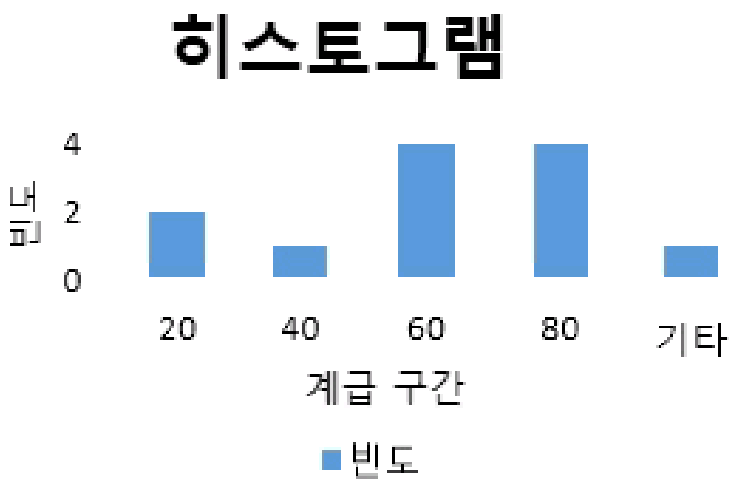
조사한 자료 수, 가장 높은 값, 가장 낮은 값과 같이 여러 가지 통계적 자료들을 쉽게 구할 수 있음

자료의 개수가 많은 경우에는 줄기와 잎 그림을 활용 하지 않음

통계적 모형화, 과학적 조사, 그래프 진단

히스토그램

도수 분포를 그래프로 나타낸 것



히스토그램은 자료의 분포 상태를 그래프 (그림) 으로 나타내어 쉽게 알아 볼 수 있음

통계적 모형화, 과학적 조사, 그래프 진단

히스토그램

도수 분포를 그래프로 나타낸 것

용어

- 도수(frequency) : 각 자료값이 나타나는 빈도
- 상대도수(relative frequency) : 도수를 전체 자료의 수로 나눈 것
- 도수분포(frequency distribution) : 각 자료값의 도수 또는 상대도수를 나열해 놓은 것
- 도수분포표(frequency table) : 도수분포를 표로 나타낸 것
- 히스토그램(histogram) : 상대도수 막대그래프

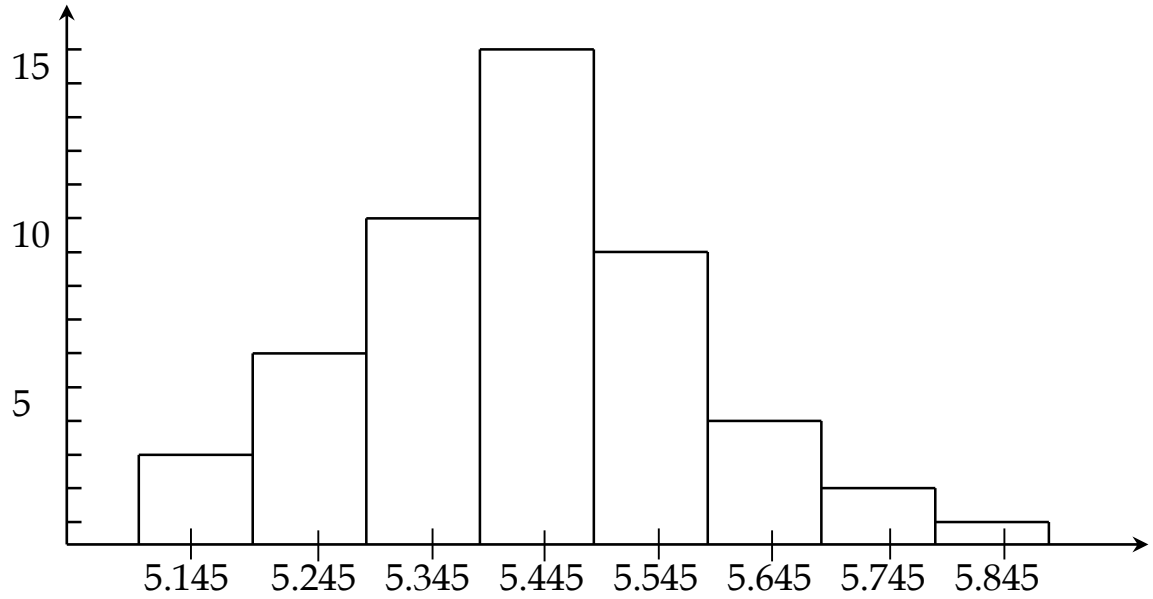
통계적 모형화, 과학적 조사, 그래프 진단

히스토그램

계급번호	계급구간	중간점	도수	상대도수
1	5.095-5.195	5.145	3	0.06
2	5.195-5.295	5.245	6	0.12
3	5.295-5.395	5.345	10	0.20
4	5.395-5.495	5.445	15	0.30
5	5.495-5.595	5.545	9	0.18
6	5.595-5.695	5.645	4	0.08
7	5.695-5.795	5.745	2	0.04
8	5.795-5.895	5.845	1	0.02

통계적 모형화, 과학적 조사, 그래프 진단

히스토그램



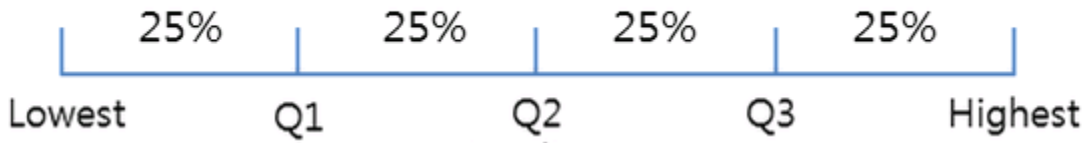
통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자

수치적 자료를 다섯 숫자 요약값을 사용하여 그래프로 표현하는 방법

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자



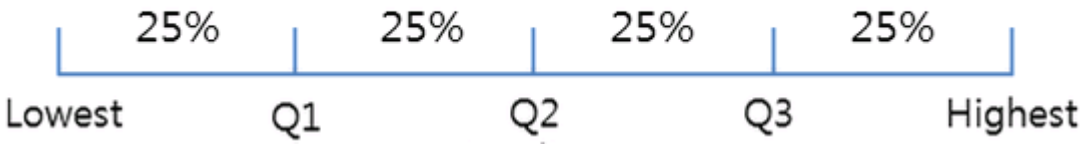
사분위수

Q3-Q1의 값을 **사분위수 범위(Interquartile Range : IQR) 라고 부름**

- Q1은 제 1사분위수
- Q2는 제 2사분위수
  - 50% 지점인 Q2는 자료의 정중앙
  - 중앙값(median)과 백분위수 50번째 값과 동일
- Q3은 제 3사분위수

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자



다섯숫자 요약(Five number summary)

Lowest, Q1, Q2, Q3, Highest의 5가지 값을 다섯 숫자 요약이라고 부름

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자

수치적 자료를 다섯 숫자 요약값을 사용하여 그래프로 표현하는 방법

15	32	8	4	15	30	37	19	3	17
4	24	27	29	16	12	16	8	31	37
24	28	41	26	13	43	34	6	45	26
14	15	46	16	15	40	20	19	21	8
25	7	16	11	18	23	12	18	9	27
46	16	16	47	13	12	45	40	17	17

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자

수치적 자료를 다섯 숫자 요약값을 사용하여 그래프로 표현하는 방법

3	4	4	6	7	8	8	8	9	11
12	12	12	13	13	14	15	15	15	15
16	16	16	16	16	16	17	17	17	18
18	19	19	20	21	23	24	24	25	26
26	27	27	28	29	30	31	32	34	37
37	40	40	41	43	45	45	46	46	47

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자

중앙 값 구하는 공식

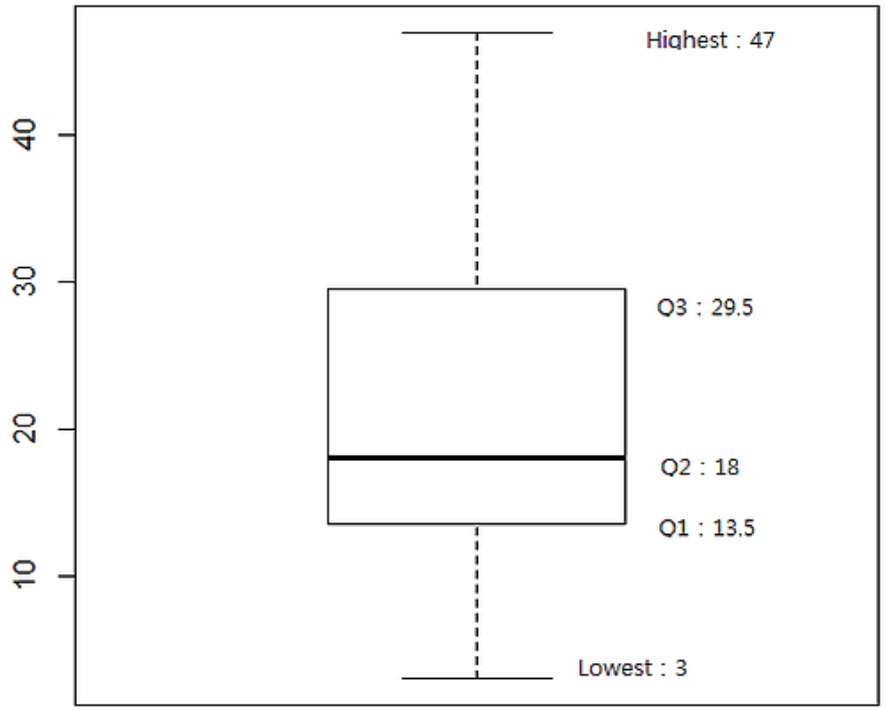
$$Median = \frac{n+1}{2}$$

$$Median = \frac{\frac{n}{2} + \frac{n+2}{2}}{2}$$

자료의 개수가 짝수 – 두번째 공식을 이용하여
Q2은 18
Q1은 Lowest value인 3과 30번째 자료의 중앙값으로 13.5
Q3는 Highest value인 47과 31번째 자료의 중앙값으로 29.5

통계적 모형화, 과학적 조사, 그래프 진단

상자-수염 그림 또는 그림상자



통계적 모형화, 과학적 조사, 그래프 진단

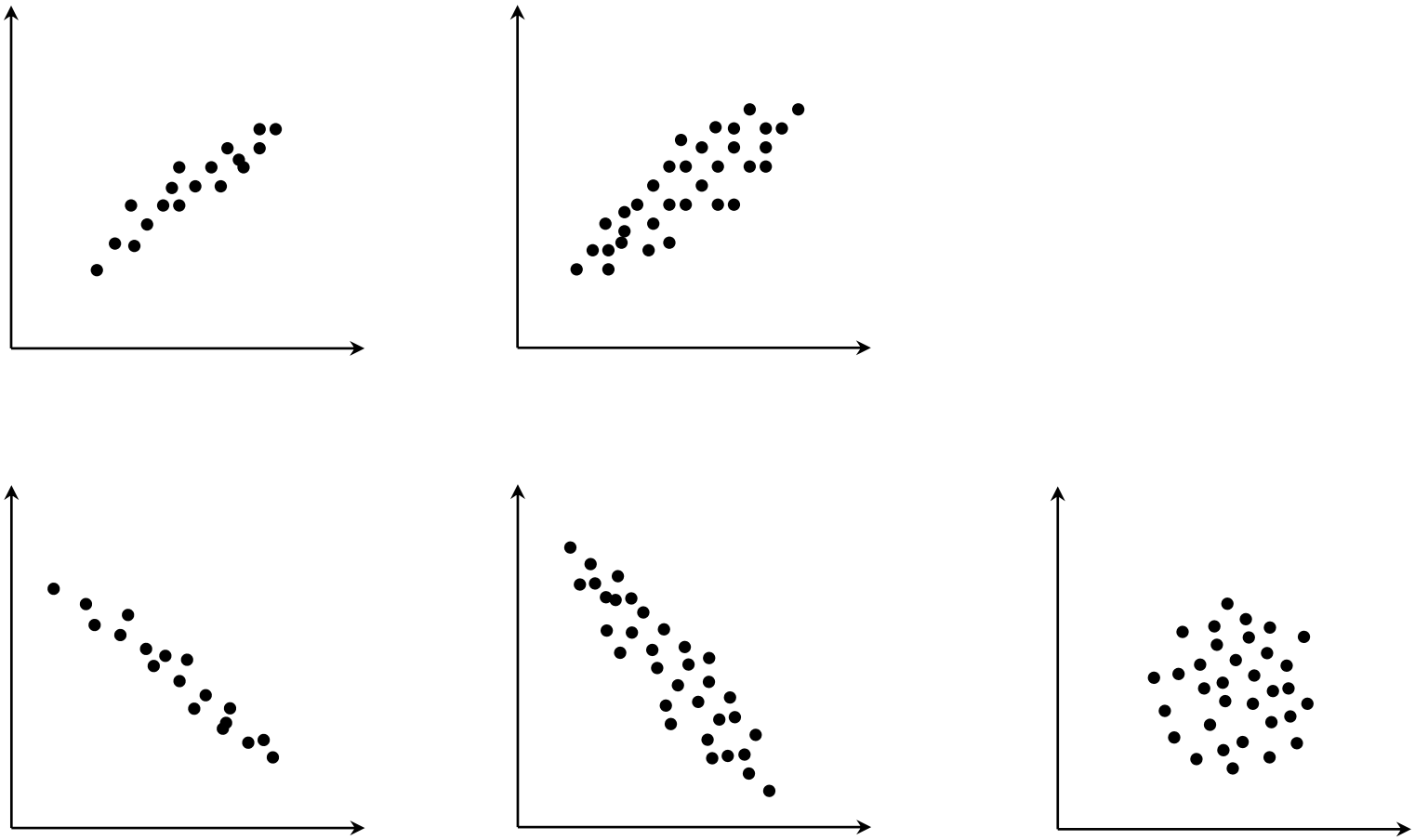
산점도

한 종류의 데이터만이 아닌 두 종류 이상의 데이터 사이의 관계를 고려해야 할 때 작성하는 분석 기법

두 변수에 대하여 특성(결과)과 요인(원인)의 관계를 규명하고, 시각적으로 표현

통계적 모형화, 과학적 조사, 그래프 진단

산점도



감사합니다.