2020/08/14, 2020 확률 세미나

확률 및 통계학

- 1장 통계학과 자료 분석 -

박 재 형(jaehyoung@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 통계학과 자료
- 표본 추출 (Sample Sampling)
- 측도
- 자료의 표현

통계학과 자료

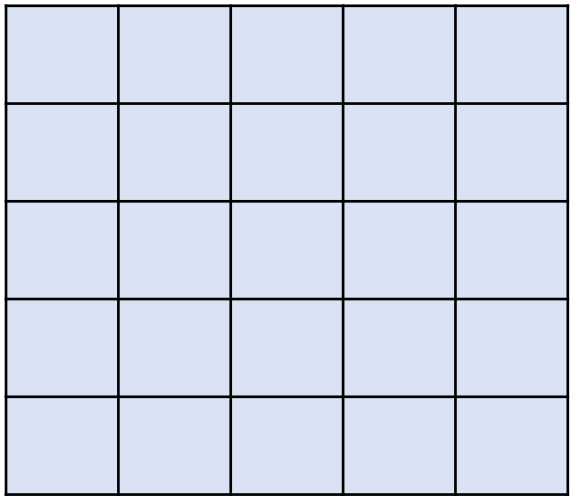
- 통계학의 정의
  - 산술적 방법을 기초로 하여, 데이터를 관찰하고 정리 및 분석하는 방법
- 통계학에서 사용되는 용어
  - 모집단 (Population)
    - 연구 대상이 되는 집단 전체
  - 표본 집단 (sample)
    - 연구 조건에 맞게 추출한 모집단의 일부 집단

통계학과 자료

- 분류
  - 기술 통계학 (Descriptive Statistics)
    - 자료를 수집 및 정리하여 모집단의 특징을 설명하는 방법
    - 자료의 형태를 표현하거나 평균, 중앙값 등의 특성 값 사용
      - e.g., 히스토그램, 줄기-잎 그림 등
  - 추론 통계학 (Inferential Statistics)
    - 표본을 수집하여 모집단의 특징을 추정하고 그 결과가 신뢰성이 있는지 검정하는 방법
      - E.g., 한국인 100명의 수면 시간에 대한 조사에서 평균 6.2시간이 조사 되었다면 한국인의 수면 시간은 6.5시간이 적당할 것으로 추정

통계학과 자료

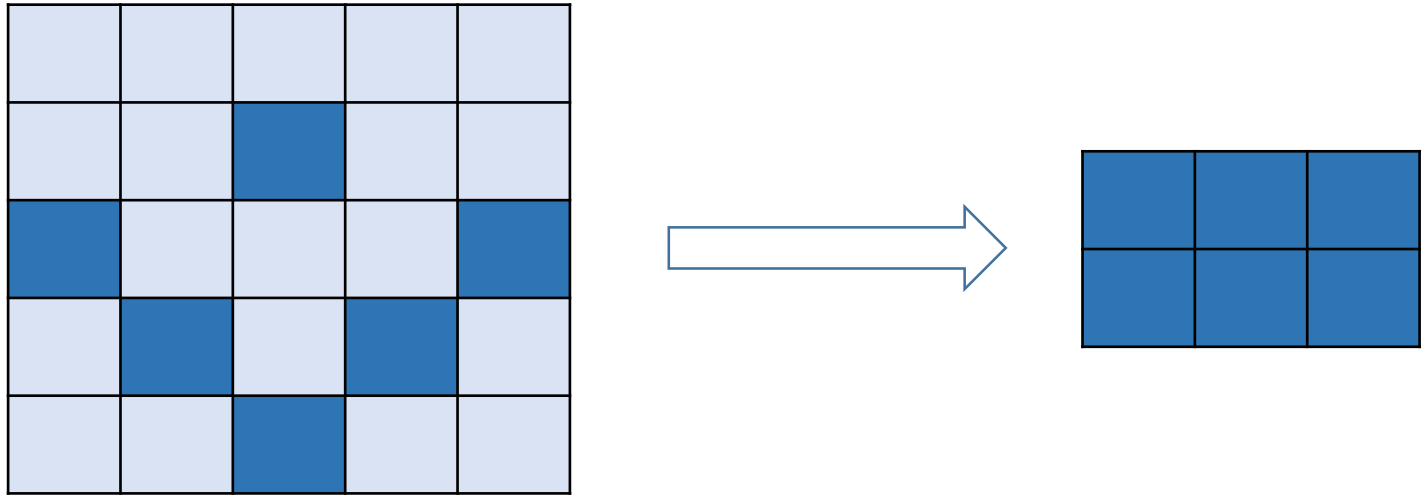
- 자료의 수집
  - 전수 조사
    - 모집단을 이루는 모든 개체들을 조사하여 모집단의 특징을 나타내는 방법
      - e.g., 한국인 전체, 미국인 전체 등



  - 한계
    - 집단 내 모든 개체를 다 조사한다는 것은 현실적으로 불가능

통계학과 자료

- 자료의 수집
  - 표본 조사
    - 연구 대상이 되는 모집단 중 표본을 선택하고 조사를 실시하여 얻은 결과로 모집단의 특성을 추정하는 방법
      - e.g., 광진구 거주자중 세종대학교 학생을 대상으로 게임을 좋아하는 사람에 대한 조사



  - 특징
    - 모집단에서 추출한 표본이 전체 모집단의 특징을 대표해야 함

통계학과 자료

- 자료의 사용
  - 수집된 자료를 사용하여 통계학적 결과에 대한 불확실성의 측정 또는 근거 제공
    - e.g., 추론 통계학, 기술 통계학 등
  - 불확실성 (Uncertainty)
    - 불규칙하여 하나로 단정 지을 수 없는 특성

자료의 표현

- 자료의 분류
  - 이산형 (Discrete)
    - 측정되는 값이 연속적이지 않는 자료 (Count Data)
      - e.g., 기계 반응 횟수, 코드 오류 개수 등
  - 연속형 (Continuous)
    - 측정되는 값이 연속적인 자료
      - 정확히 측정할 수 없기 때문에 자료의 범위를 지정
      - e.g., 사람의 키, 공기의 질량 등

통계학과 자료

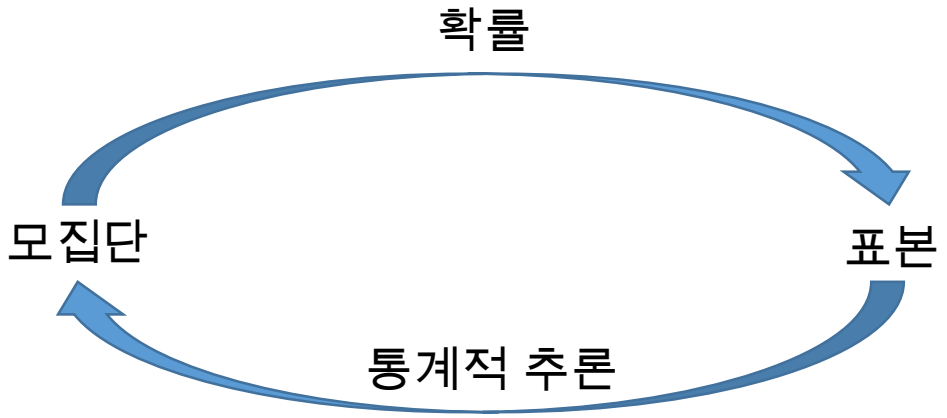
- 자료의 변동
  - 자료는 여러 요소 (Factor)에 영향을 받음
    - e.g., 식물 생장 연구의 경우 온도, 물의 양, 공기 성분 등
  - 변동 사용에 따른 통계적 방법
    - 관측 연구 (Observation Study)
      - 자료에 영향을 주는 요소가 를 제어하 않고 관찰을 통해 분석하는 방법
        - e.g., 토마토의 생장 연구의 경우, 길이나 크기, 색 등을 측정하여 토마토의 특징을 조사 가능
    - 실험 계획 (Experimental Design)
      - 자료에 영향을 주는 요소를 측정자가 제어하여 실험을 통해 분석하는 방법
        - e.g., 토마토의 생장에 물이 끼치는 영향 연구의 경우, 하루에 100ml 물을 주는 토마토와 1L의 물을 주는 토마토 대상으로 토마토의 특징 조사 가능

통계학과 자료

- 확률
  - 정의
    - 어떠한 일이 생길 가능성을 비율로 나타낸 것
  - 통계학에서 확률의 역할
    - 자료로부터 얻어지는 추론에 대한 신뢰성을 검정함
    - P-value (Probability-value)
      - 추론의 신뢰성 검정에 대한 지표로 사용되는 값
      - e.g., P-value가 0.05라면 가설을 만족할 확률이 약 95%라고 의미

통계학과 자료

- 확률과 추론의 관계



  - 귀납적 방법
    - 여러 가지 사실에서 나타난 현상으로 특정한 가설을 증명
    - 표본을 바탕으로 통계적 추론을 통해 모집단에 대한 결론을 이끌어냄
  - 연역적 방법
    - 일반적 원리를 대전제로 두고 특정한 가설을 추론하여 정의
    - 확률을 통해 모집단으로부터 추출된 표본 자료의 특성에 대한 결론을 이끌어냄

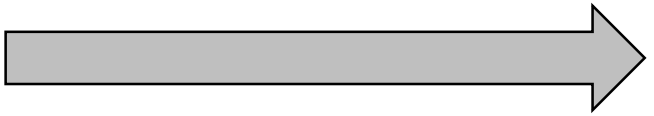
표본 추출 (Sampling)

- 정의
  - 모집단에서 조건을 만족하는 일부 개체를 추출하는 것
  - 종류
    - 단순 랜덤 표본 추출 (Simple Random Sampling)
      - 모집단의 개체가 표본으로 선택될 가능성이 같도록 하는 표본 추출 방법
    - 층화 랜덤 표본 추출 (Stratified Random Sampling)
      - 모집단을 중복되지 않는 층으로 나눈 후 각 층에서 표본을 추출하는 방법

표본 추출 (Sampling)

- 단순 랜덤 표본 추출 (Simple Random Sampling)

11	12	22	33	34
23	13	35	24	36
37	14	38	25	39
26	15	16	30	27
31	28	17	18	19

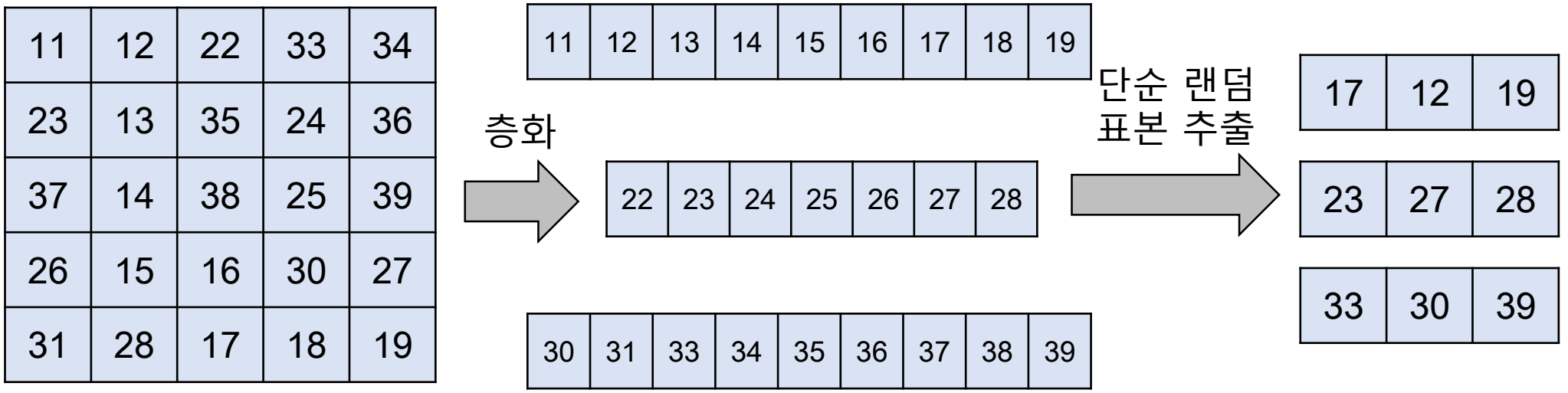


30	23	39
17	38	22
19	11	15

- 특징
  - 모집단에 대한 사전 지식 불필요
  - 랜덤으로 추출하기 때문에 추출될 확률이 동등

표본 추출 (Sampling)

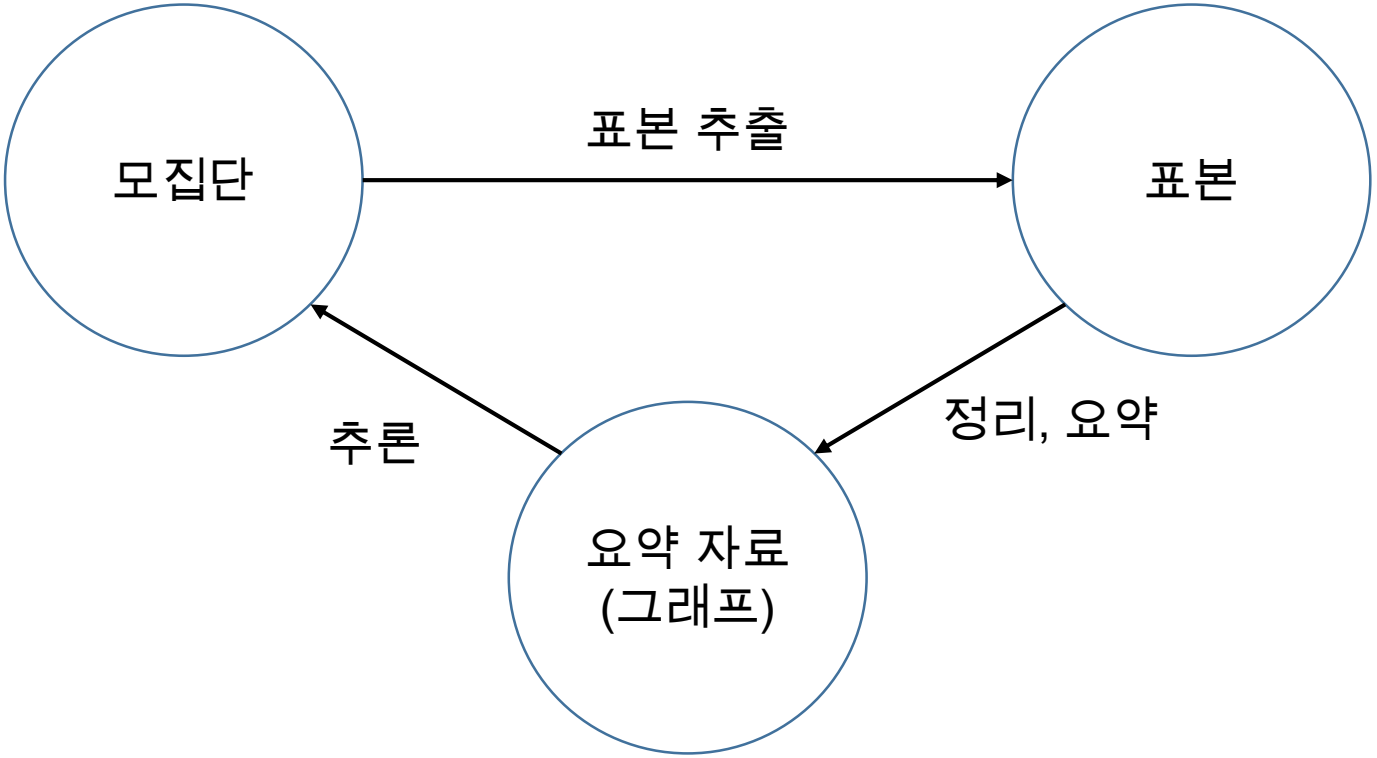
- 층화 랜덤 표본 추출 (Stratified Random Sampling)



- 특징
  - 모집단에 대한 지식이 필요
  - 각 층의 특성에 대한 추정과 비교 가능

표본 추출 (Sampling)

- 목적
  - 연구의 주제가 되는 전체 모집단의 일부를 조건에 맞게 추출하여 이들을 통해 얻은 정보를 바탕으로 모집단을 추정
    - 통계적 추론 (Statistical Inference)



위치 측도와 산포 측도

- 위치 측도
  - 정의
    - 자료들이 어떠한 값을 기준으로 어떤 형태의 분포를 가지는지 나타내는 측도

이름	식	설명
표본 평균 (Sample Mean)	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$	• 일반 적으로 사용하는 산술적인 평균
표본 중앙 값 (Sample Median)	$\tilde{x} = \begin{cases} x_{(n+1)/2}, n = \text{홀수} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2+1}}\right), n = \text{짝수} \end{cases}$	• 극단 값이나 특이점에 영향을 받지 않고 자료를 크기별로 나열 하였을 때 자료의 중심을 측정하기 위한 값
절사 평균 (Trimmed Mean)	$\bar{x}_{tr(K)}$	• 자료의 가장 크거나 작은 일부분을 제외한 평균 • 극단 값 또는 특이점에 대해 영향을 표본 평균과 비교하여 덜 받음

위치 측도와 산포 측도

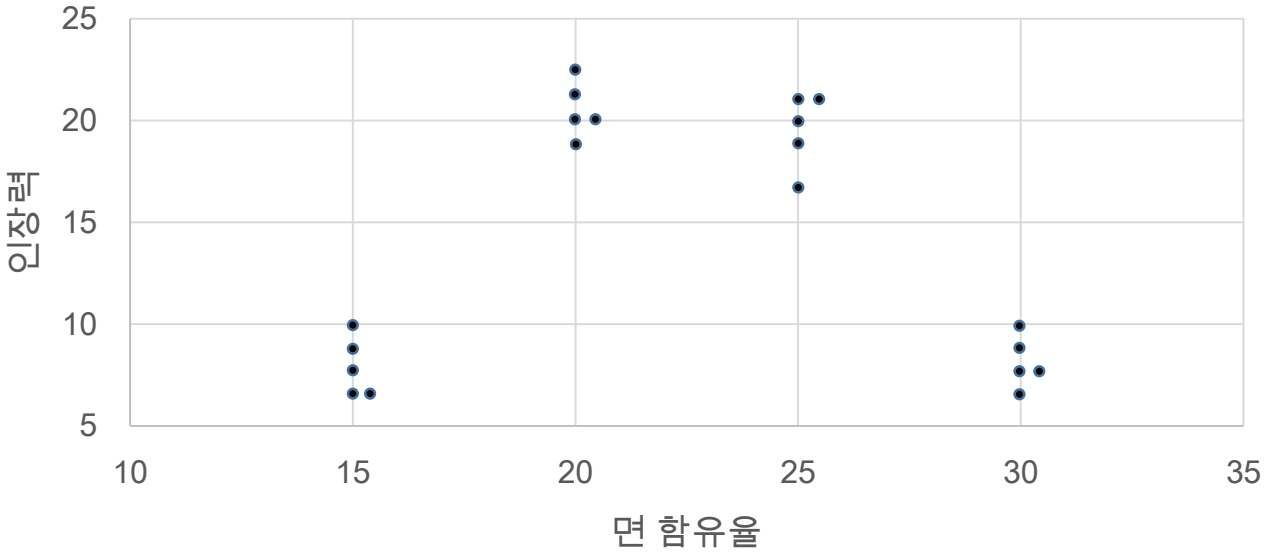
- 산포 측도
  - 정의
    - 자료들이 중심 위치에서 얼마나 떨어져 있는지를 나타내는 측도

이름	식	설명
표본 범위 (Sample Range)	$x_{max} - x_{min}$	• 가장 단순한 산포도, 표본들의 범위
표본 분산 (Sample Variance)	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$	• 표본 평균을 중심으로 자료들이 흩어진 정도를 측정하는 값
표본 표준 편차 (Sample Standard Deviation)	$s = \sqrt{s^2}$	• 제곱 단위인 표본분산의 단위를 맞춤

자료의 표현

- 산점도 (Scatter Diagram)
  - 정의
    - 두 종류 이상의 데이터 사이의 관계를 나타내는 방법
      - e.g., 면 함유율이 인장력에 미치는 영향

면 함유율	인장력
15	7, 7, 9, 8, 10
20	19, 20, 21, 20, 22
25	21, 21, 17, 19, 20
30	8, 7, 8, 9, 10



자료의 표현

- 줄기-잎 그림 (Stem and leaf plot)
  - 정의
    - 자료의 값을 자릿수로 나누어서 앞자리는 줄기, 뒷자리는 잎이라 칭하여 제시하는 방법
      - e.g., 34일 경우 줄기 = 3, 잎 = 4

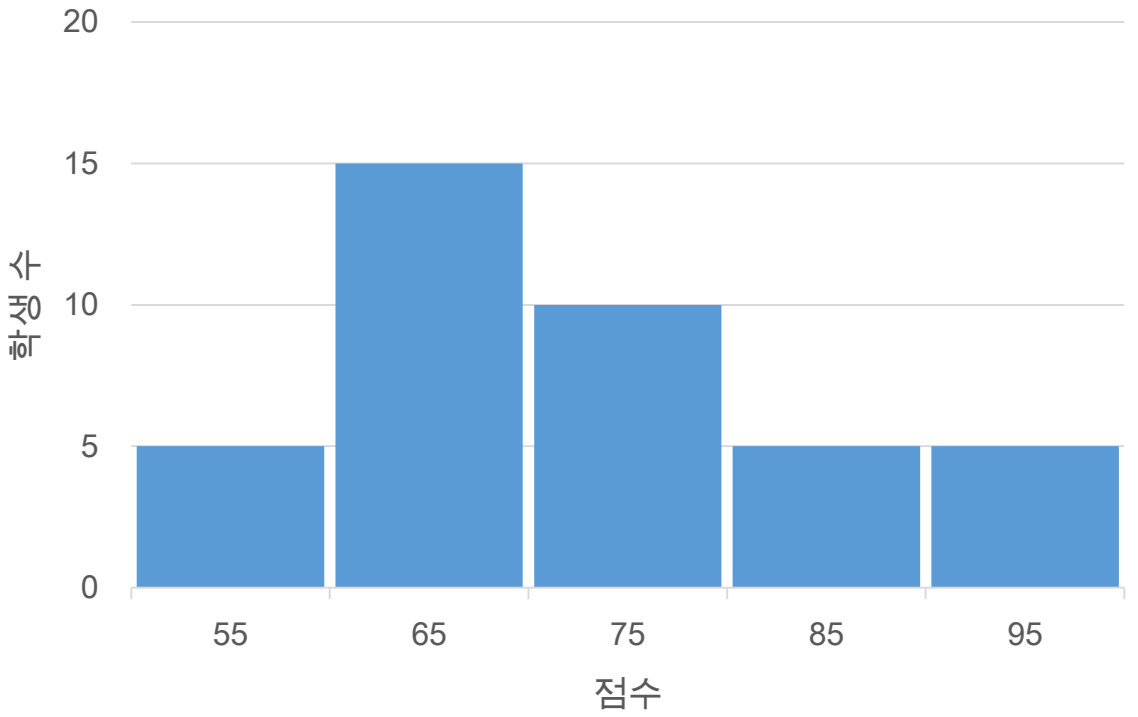
50	64	99	74
87	55	60	77
100	50	50	94
77	70	77	80
98	77	76	77

줄기	잎	빈도
5	0500	4
6	40	2
7	70776477	8
8	70	2
9	894	2
10	0	1

자료의 표현

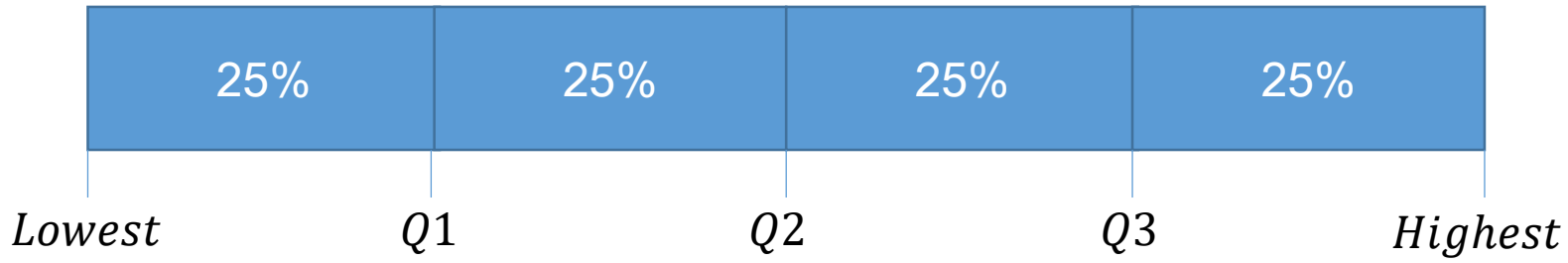
- 히스토그램 (Histogram)
  - 정의
    - 자료의 값이 나타나는 분포를 그래프로 나타낸 것

계급 구간 (점수)	중간점	학생 수
50 ~ 60	55	5
60 ~ 70	65	15
70 ~ 80	75	10
80 ~ 90	85	5
90 ~ 100	95	5



자료의 표현

- 상자-수염 그림 (Box and Whisker Plots)
  - 정의
    - 수치적 자료를 다섯 숫자 요약 값과 사분 범위를 사용하여 그래프로 나타내는 방법
  - 사분 범위 (Interquartile Range)
    - $Q3 - Q1$
    - 사분 범위의 1.5배 넘는 곳에 자료가 있다면 특이점으로 판정 할 수 있음

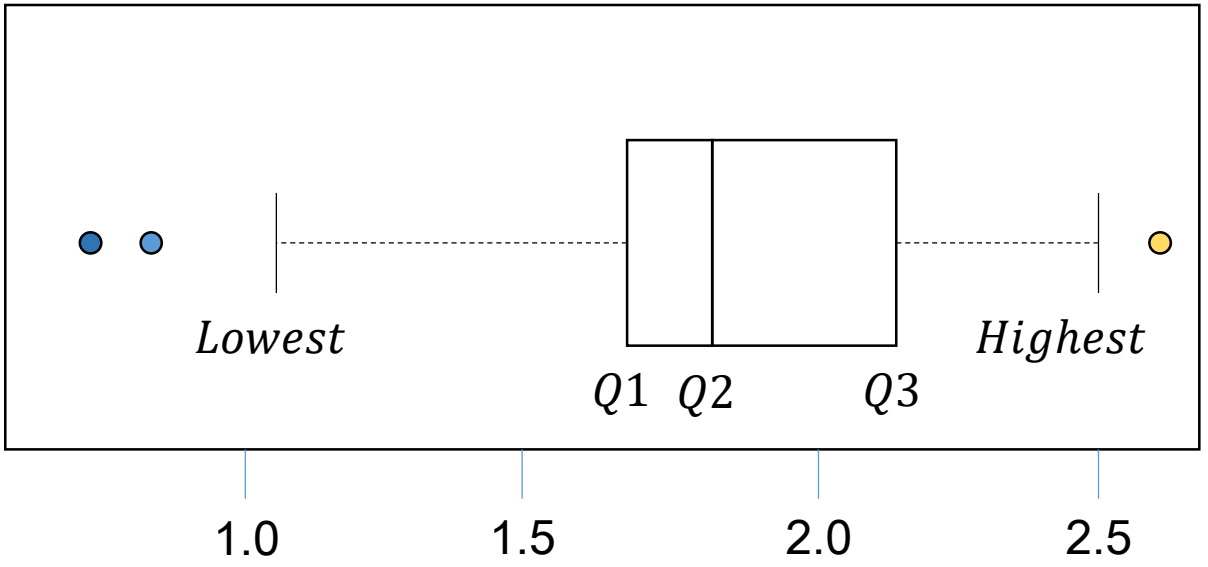


자료의 표현

- 상자-수염 그림 (Box and Whisker Plots)
  - 예시
    - $Q1 = 1.63$
    - $Q2 = 1.75$
    - $Q3 = 1.97$
    - 사분범위 $= 0.34$

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

니코틴 함량(단위: mg) 표



Thanks!

박 재 형 (jaehyoung@pel.sejong.ac.kr)