2020/08/13, 2020 확률 세미나

확률 및 통계학

- 1장 통계학과 자료 분석 -

박 재 형(jaehyoung@pel.sejong.ac.kr)

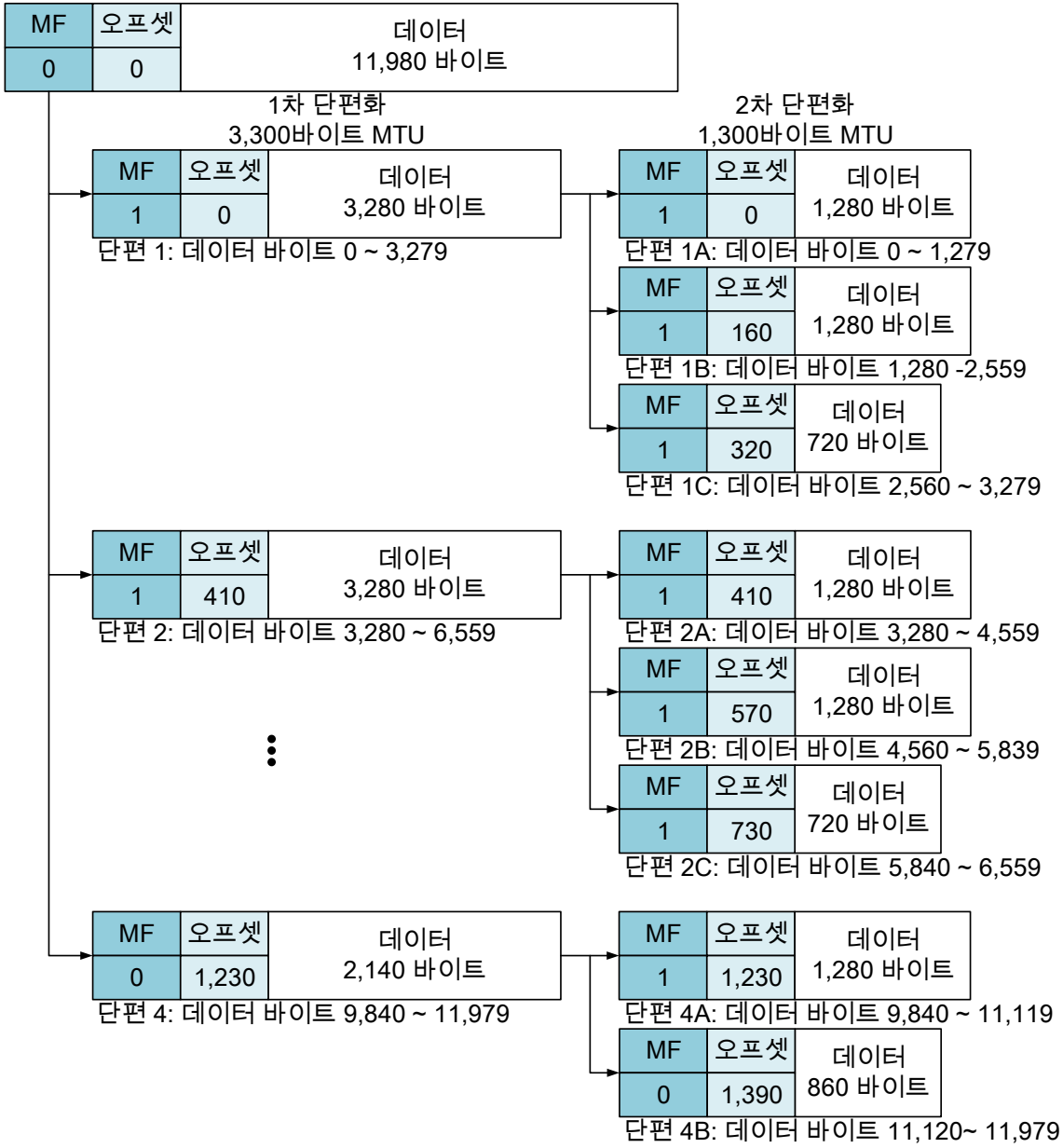
세종대학교 프로토콜공학연구실

목 차

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

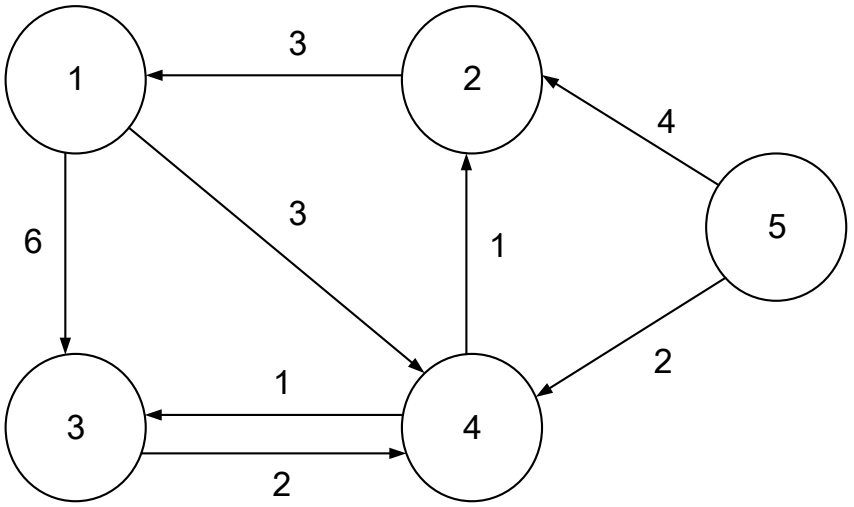
보충

- 단편화 과정
  - MF는 마지막 단편이나 하나의 단편 의미
  - 오프셋은 단편의 위치를 의미
  - DF 값이 1이 아닌 경우 1차 단편화 이후 MTU 크기에 맞지 않는다면 추가 단편화 가능
    - DF 값은 단편화 가능 여부를 나타내는 필드



보충

- IP 멀티캐스팅
  - 주요 기능
    - 멀티캐스트 데이터그램 라우팅
      - 데이터그램을 전송할 특수 알고리즘 사용하여 효율적으로 전송하기 위한 라우팅
        - 최단 경로 우선 알고리즘 (Shortest Path First Algorithm)
          - 목적지 네트워크 까지의 라우터 경로와, 거쳐야할 라우터 수를 가진 라우팅 테이블을 이용하여 경로설정



노드	5 (정점)	4	3	2	1
최단 경로	0				
		2		4	
			3	3	
					6

보충

- 질문
  - 지구 한바퀴를 돌 때 필요한 홉수?
    - 홉은 네트워크내에 출발지 장비와 목적지 장비사이의 경로를 뜻함으로 1개의 홉이 필요하다고 생각합니다.
  - 장비가 멀티캐스트 그룹을 탈퇴하는 경우, 라우터는 어떻게 알아 차리는가?
    - IGMP (Internet Group Management Protocol)를 사용하여 호스트에게 주기적으로 메시지를 전송하여 탈퇴여부를 라우터에게 알려줌
      - IGMP는 장비와 라우터들이 서로 그룹 간 가입 정보를 교환할 수 있도록 하는 프로토콜

통계학과 자료

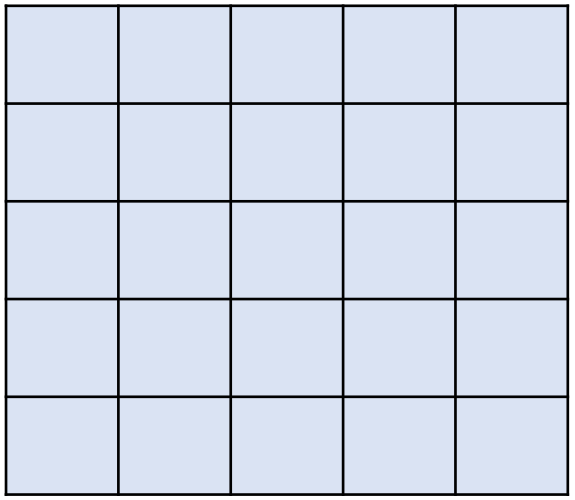
- 정의
  - 산술적 방법을 기초로 하여, 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 분야
- 용어
  - 모집단 (Population)
    - 연구 대상이 되는 집단 전체
  - 표본 집단 (sample)
    - 표본 조사를 통해서 추출한 모집단의 일부
  - 자료 (material)
    - 표본집단을 조사해 얻은 데이터

통계학과 자료

- 분류
  - 기술통계학 (Descriptive Statistics)
    - 표본 값을 요약하여 나타내어 모집단의 특성을 설명하기 위한 방법
    - 그래프가 사용됨
      - e.g., 히스토그램, 줄기-잎 그림 등
    - 특성 설명
      - e.g., 평균, 중앙값, 표준편차 등 사용
  - 추론통계학 (Inferential Statistics)
    - 표본을 수집하여 표본 값이 나타내는 의미를 해석하여 모집단의 특성을 추측하는 방법
      - e.g., 확률을 사용하여 결과 예측

통계학과 자료

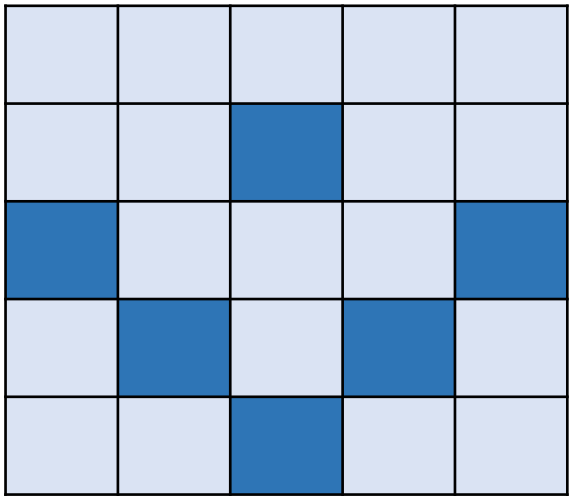
- 자료의 수집
  - 전수 조사
    - 관심 대상이 되는 집단을 이루는 모든 개체들을 조사하여 모집단의 특성을 측정하는 방법
      - e.g., 한국인 전체, 미국인 전체 등



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

통계학과 자료

- 자료의 수집
  - 표본 조사
    - 관심 대상이 되는 모집단 중 일부 표본을 선택하고 조사를 실시하여 얻은 결과로 모집단의 특성을 추정하는 방법
      - e.g., 한국인 남성, 미국인 1000명 등



  - 한계
    - 모집단에서 추출한 표본이 전체 모집단의 특성을 대표해야 함

통계학과 자료

- 자료의 사용
  - 수집된 자료를 사용하여 불확실성의 측정 또는 결과에 대한 논리나 방법을 제공
    - e.g., 추론통계학, 기술통계학 등
  - 불확실성 (Uncertainty)
    - 불규칙하여 하나로 단정 지을 수 없는 특성
      - e.g., 오차 등

통계학과 자료

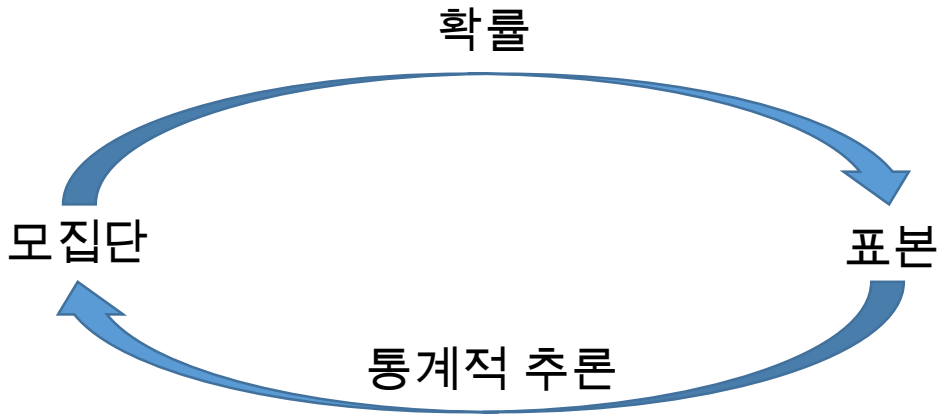
- 자료의 변동
  - 자료는 여러 요소 (Factor)에 영향을 받음
    - e.g., 식물 생장 연구의 경우 온도, 물의 양, 공기 성분 등
  - 변동의 사용
    - 관측연구 (Observation Study)
      - 자료에 영향을 주는 요소가 많고 다양하여 예측할 수 없어 관측자가 제어할 수 없는 경우
        - e.g., 서로 다른 종의 식물에 500ml의 물을 주는 경우, 길이나 크기를 측정하거나 특징을 조사 할 수 있음
    - 실험계획 (Experimental Design)
      - 자료에 영향을 주는 요소를 측정자가 제어할 수 있는 경우
        - e.g., 100ml의 물을 주는 식물과 1L의 물을 주는 식물을 비교하는 경우, 두 집단의 차이점을 알 수 있음

통계학과 자료

- 확률
  - 정의
    - 어떠한 일이 생길 가능성을 비율로 나타낸 것
  - 역할
    - 추출한 표본이 모집단을 얼마나 대표하는가를 나타내는지 계산
  - P-value
    - 가설 검정에 대한 지표로 사용되는 값
    - e.g., P-value가 0.05라면 가설을 만족할 확률이 약 95%라고 의미

통계학과 자료

- 확률과 추론의 관계



- 확률을 이용한 통계적 추론
  - 귀납적 방법
    - 여러 가지 사실에서 나타난 현상으로 특정한 가설을 증명
      - 일반화
  - 연역적 방법
    - 일반적인 원리를 대전제로 특정한 가설을 추론하여 정의
      - 삼단 논법

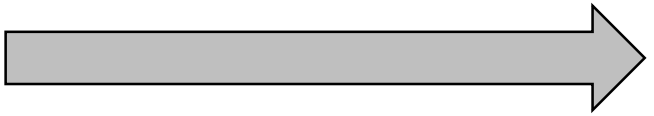
표본 추출 (Sample Sampling)

- 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 (Sample Sampling)

- 단순 랜덤 표본 추출 (Simple Random Sampling)

1	1	2	3	3
2	1	3	2	3
3	1	3	2	3
2	1	1	3	2
3	2	1	1	1

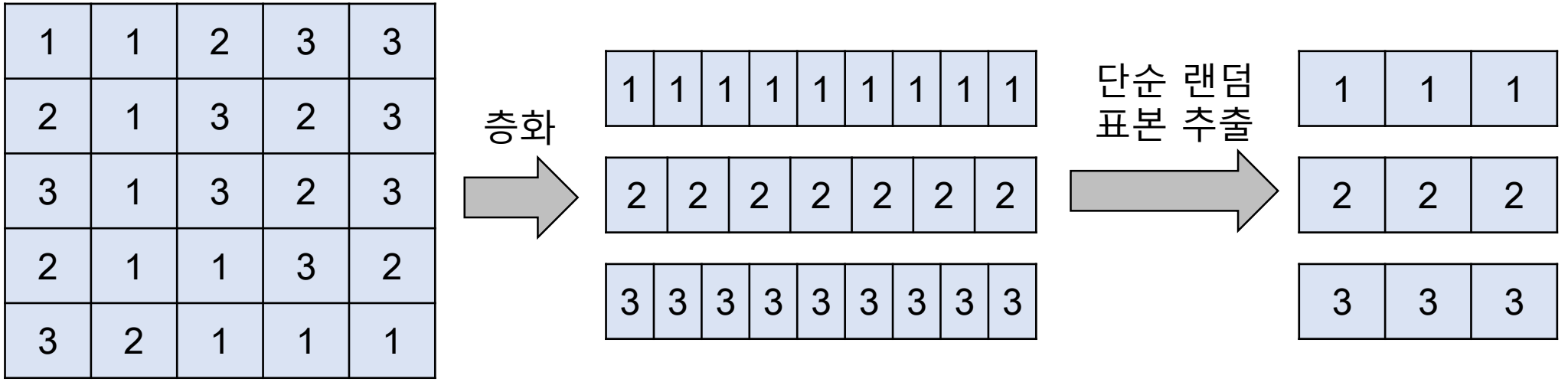


3	2	3
1	3	2
1	1	1

- 특징
  - 모집단에 대한 사전 지식 불필요
  - 추출 기회가 동등하고 독립적이므로 표본의 대표성이 높음

표본 추출 (Sample Sampling)

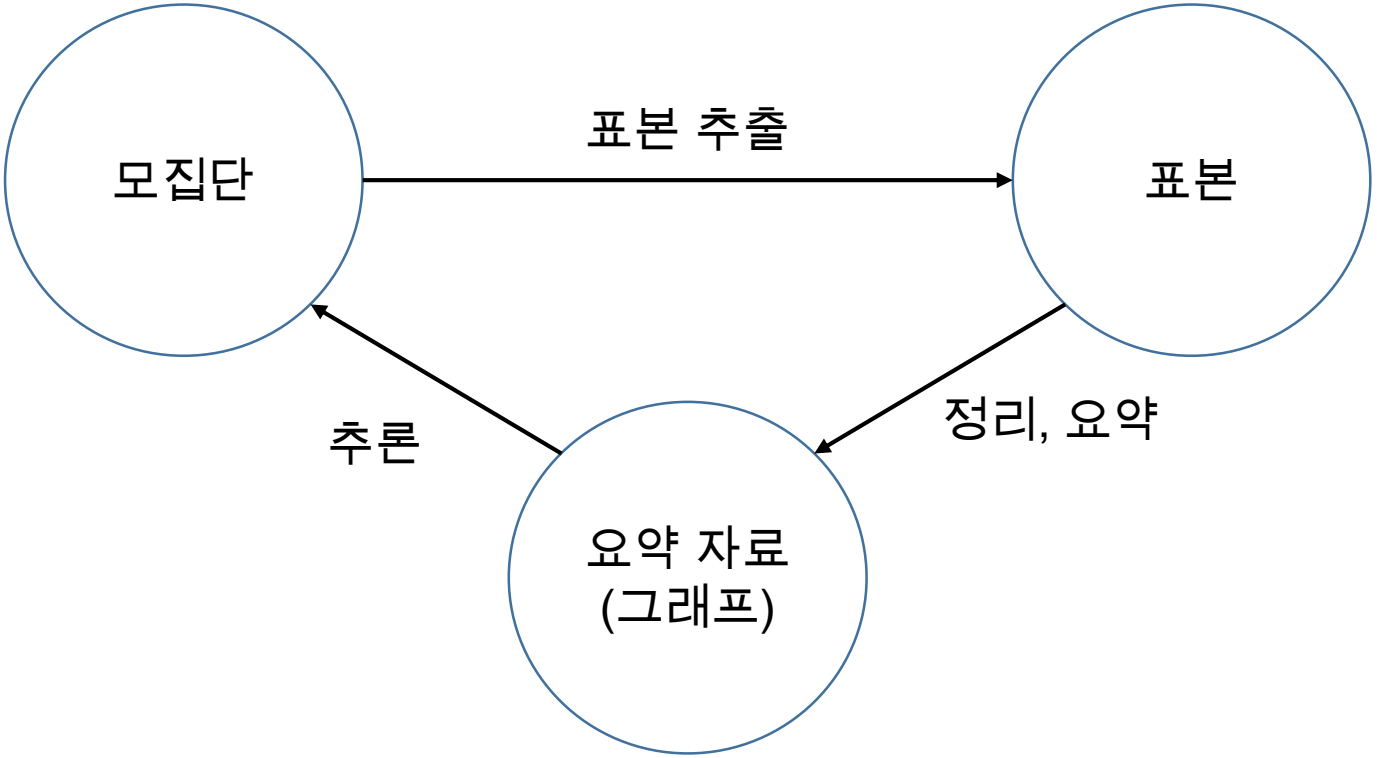
- 층화 랜덤 표본 추출 (Stratified Random Sampling)



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

표본 추출 (Sample Sampling)

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



측도

- 위치 측도
  - 정의
    - 자료들의 어떠한 값을 기준으로 어떤 형태의 분포를 가지는지 나타내는 측도
  - 종류
    - 표본 평균 (Sample Mean)
      - $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
    - 표본 중앙 값 (Sample Median)
      - 자료를 크기 별로 정렬했을 때 극단 값 또는 특이점 (Outline)의 영향을 받지 않는 자료의 중심 값
      - 표본 자료의 무게중심
      - $\tilde{x} = \begin{cases} x_{(n+1)/2}, n = \text{홀수} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2+1}}\right), n = \text{짝수} \end{cases}$

측도

- 위치 측도
- 종류
  - 절사 평균 (Trimmed Mean)
    - 자료의 가장 크거나 작은 일부분을 제외한 평균
    - 극단 값 또는 특이점에 대해 영향을 표본 평균과 비교하여 덜 받음
    - e.g., 1, 2, 2, 3, 3, 4, 5, 5, 7, 100
      10% 절사 평균: $\bar{x}_{tr(10)} = \frac{2+2+3+3+4+5+5+7}{8} = 3.875$
  - 극단 값에 의한 측도의 민감도
    - 표본 평균 > 절사 평균 > 표본 중앙 값

측도

- 산포 측도
  - 자료들이 중심 위치에서 얼마나 떨어져 있는지를 나타내는 측도
  - 종류
    - 표본 범위 (Sample Range)
      - 가장 단순한 산포도, 표본들의 범위
      - $x_{max} - x_{min}$
    - 표본 분산 (Sample Variance)
      - 표본 평균을 중심으로 자료들이 흩어진 정도를 측정하는 값
      - $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$

측도

- 산포 측도
  - 종류
    - 표본 표준 편차 (Sample Standard Deviation)
      - 표본분산의 양의 제곱근
      - 제곱 단위인 표본분산의 단위를 맞춤
      - $s = \sqrt{s^2}$

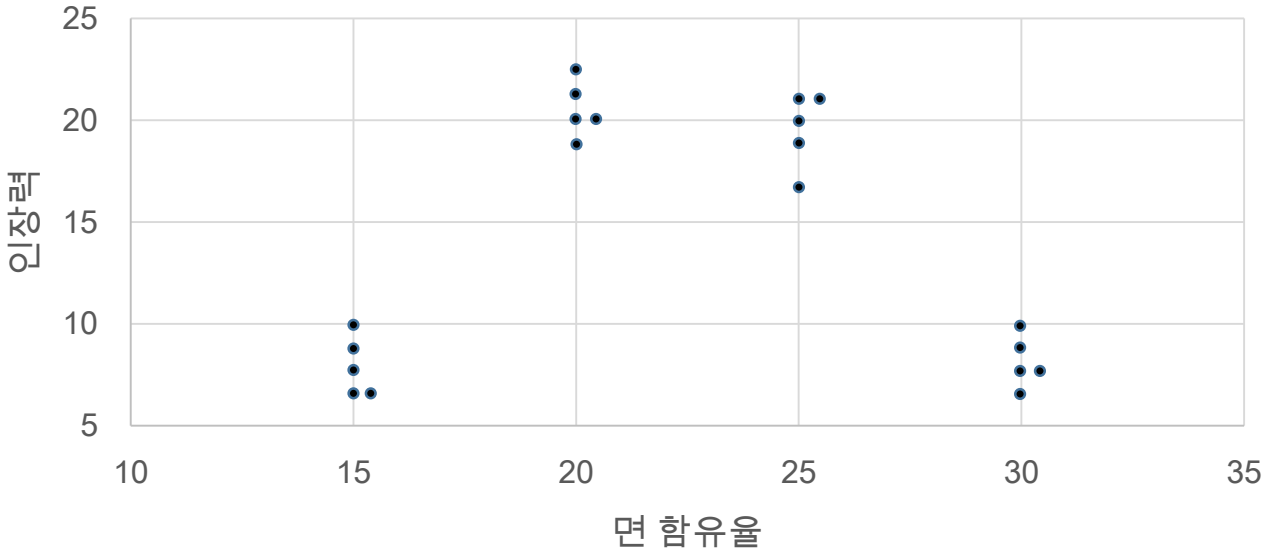
자료의 표현

- 이산형 자료와 연속형 자료
  - 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

자료의 표현

- 산점도 (Scatter Diagram)
  - 정의
    - 두 종류 이상의 데이터 사이의 관계를 고려 할 때 사용하는 분석 방법
  - 특징
    - 두 종류 이상의 자료들의 관계를 볼 수 있음

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
    - 줄기와 잎을 이용하여 자료를 나타내는 방법
      - 줄기 = 높은 자리 값, 잎 = 낮은 자리 값
        - e.g., 34일 경우 줄기 = 3, 잎 = 4;
  - 특징
    - 빈도수를 한눈에 볼 수 있음

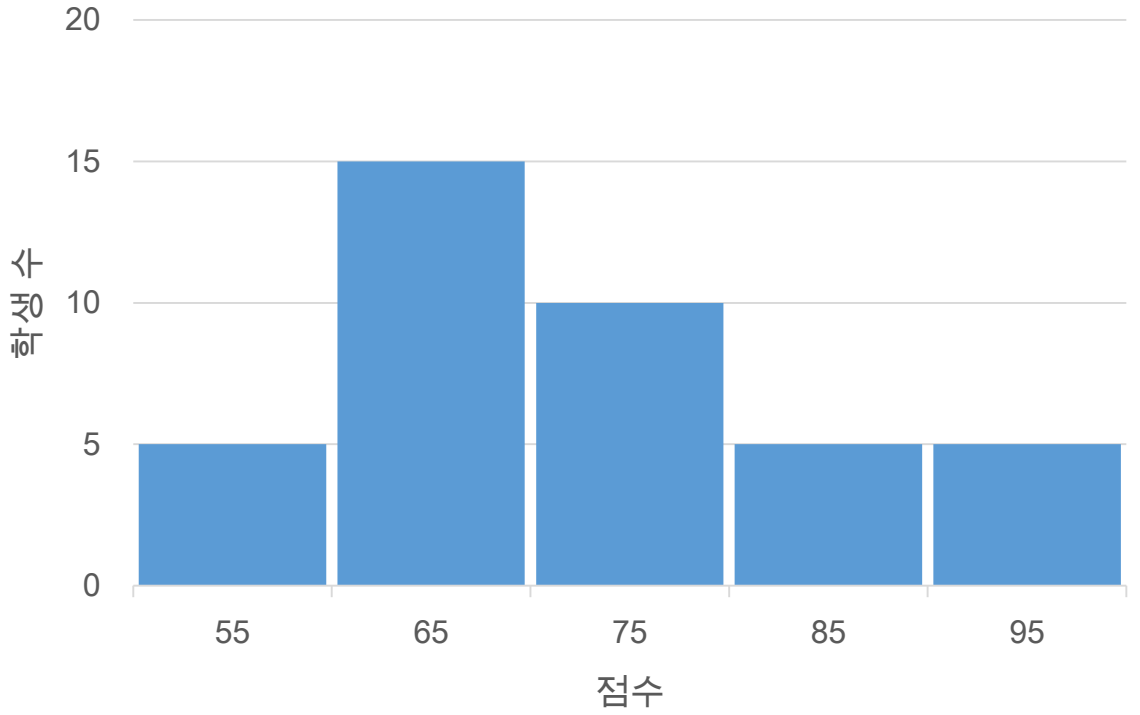
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

줄기	잎	빈도
1	69	2
2	25696	5
3	43185147236282971300971 45	25
4	71354172	8

자료의 표현

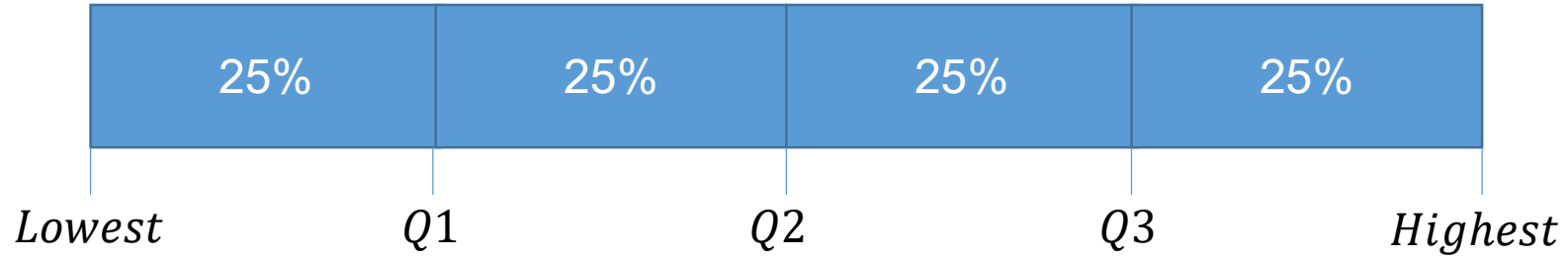
- 히스토그램 (Histogram)
  - 정의
    - 자료의 값이 나타나는 빈도를 그래프로 나타낸 것
  - 특징
    - 구간별 빈도수 비교가능

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
  - 특징
    - 특이점을 비교적 쉽게 찾을 수 있음

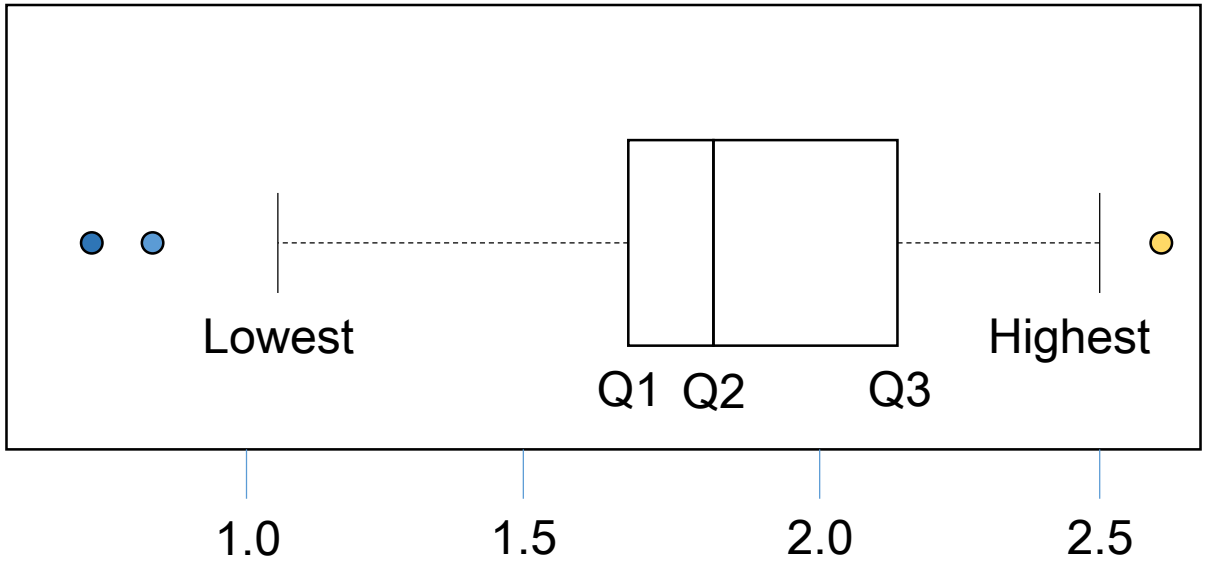


자료의 표현

- 상자-수염 그림 (Box and Whisker Plots)
  - 예시

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)