2023/11/22, 확률 기초 세미나

확률 및 통계학

- 1장 통계학과 자료 분석(보충) -

우 승 찬(seungchan@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 통계학과 자료
- 표본추출(Sampling)
- 측도(Measure)
- 자료의 표현
- 통계의 함정

통계학과 자료

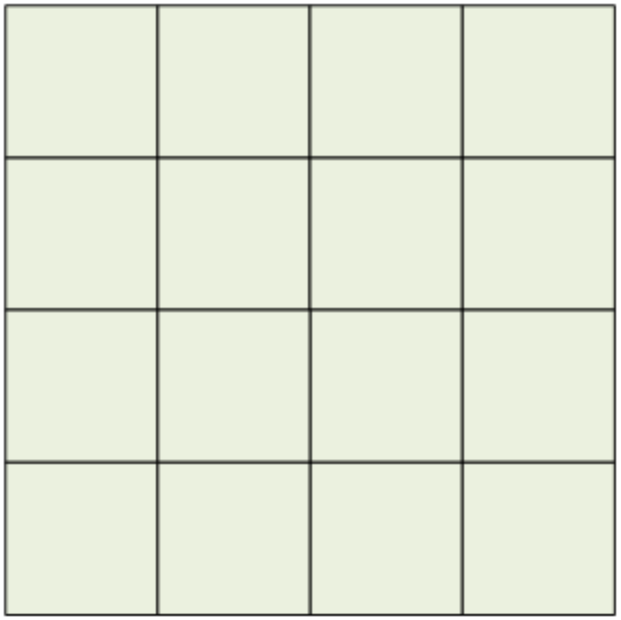
- 통계학(Statistics)
  - 정의
    - 산술적 방법을 기초로 하여, 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 분야
  - 용어
    - 모집단(Populations)
      - 연구 대상이 되는 집단의 전체
    - 표본 집단(Sample)
      - 연구 조건에 맞게 추출한 모집단의 일부 집단

통계학과 자료

- 통계학(Statistics)
  - 종류
    - 추론 통계학(Inferential Statistics)
      - 자료에 내포되어 있는 정보를 분석해서 불확실한 사실을 추론하여 검정, 추정, 예측 등을 수행하는 학문
        - e.g., 세종대학교 학생들의 평균 체중을 추정하기 위해 표본 집단을 선택하여 체중을 측정하고 해당 표본 데이터를 사용하여 전체 학생들의 평균 체중 추정
    - 기술 통계학(Descriptive Statistics)
      - 자료를 수집하고 정리해서 표, 도표 등을 만들거나 요약하여 변동의 크기, 대푯값, 분산, 평균 등을 구하는 학문
        - e.g., 히스토그램, 줄기-잎 그림 등

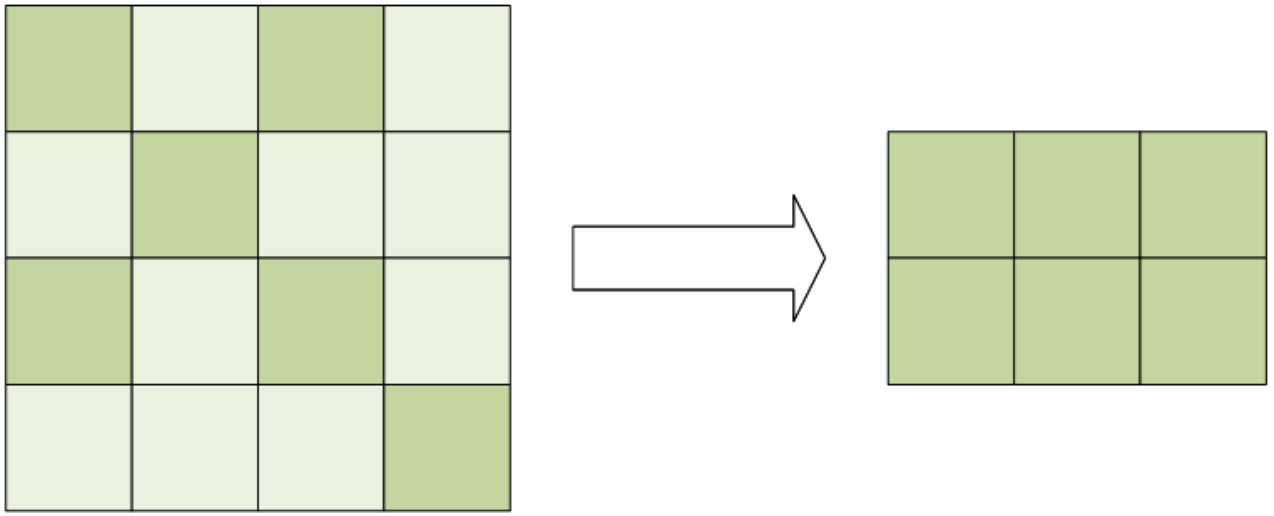
통계학과 자료

- 자료의 수집
  - 전수 조사
    - 모집단을 이루는 모든 개체들을 조사하여 모집단의 특징을 나타내는 방법
      - e.g., 통계청 인구주택총조사 등
    - 한계
      - 막대한 비용과 시간의 소모
      - 모든 대상에게 응답 얻기가 어려움
      - 모집단이 접근하기 어려운 지역에 분포 되었을 경우, 실질적으로 조사하기 어려움



통계학과 자료

- 자료의 수집
  - 표본 조사
    - 연구 대상이 되는 모집단 중 표본을 선택하고 조사를 실시
    - 조사 결과를 통해 모집단의 특성을 추정
  - 특징
    - 모집단에서 추출한 표본이 전체 모집단의 특징을 대표해야 조사의 의미가 존재함



통계학과 자료

- 자료의 사용
  - 수집된 자료를 분석, 해석하여 통계학적 결과에 대한 불확실성의 척도 및 근거 제공
    - e.g., 추론 통계학, 기술 통계학 등
  - 불확실성(Uncertainty)
    - 결과나 사건의 발생 여부가 확실하지 않고 예측하기 어려운 상태
      - 자료의 변동성, 오차, 오류에 의해 불확실한 성질이 나타남

통계학과 자료

- 자료의 분류
  - 이산형(Discrete) 자료
    - 측정되는 값이 연속적이지 않고 특정 값을 가지는 자료
      - e.g., 계수자료(학생 수, 책의 페이지 수 등),
        이진자료(제품 결함 여부, 설문 조사 응답 등)
    - 표본비율(Sample Proportion)
      - 주어진 표본에서 특정한 속성이나 결과가 나타나는 빈도를 백분율 형태로 표현하는 것
        - e.g., 설문 조사 응답에서 100명의 응답자 중 40명이 특정 질문에 '예'라고 답했다면, 표본비율은 $\frac{40}{100} \times 100 = 40\%$로 표현됨
  - 연속형(Continuous) 자료
    - 측정되는 값이 연속적으로 구성되는 자료
      - e.g., 사람의 키, 무게, 온도, 시간 등

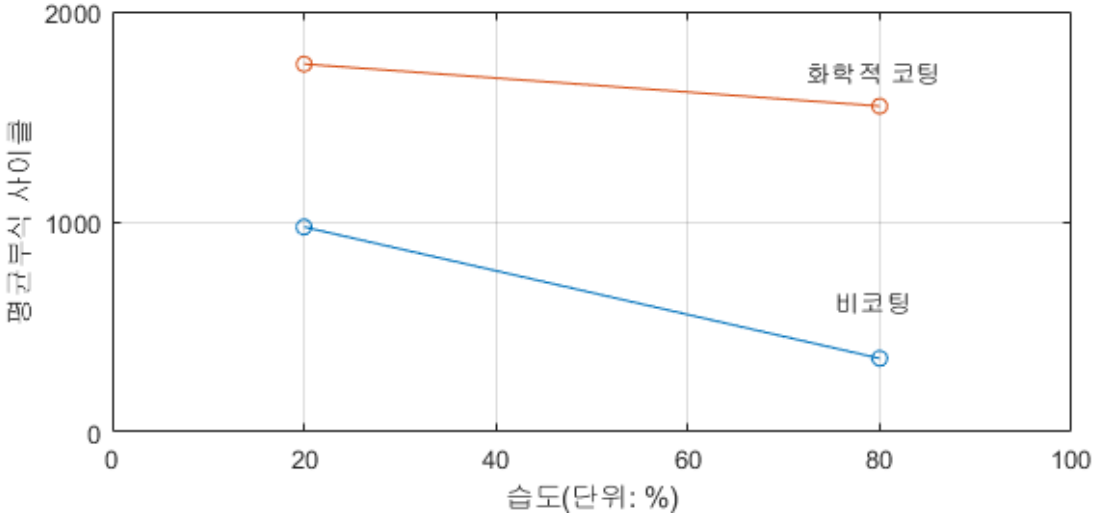
통계학과 자료

- 자료의 변동(1/2)
  - 관찰된 데이터 값들이 여러 요소에 영향을 받아 서로 다르게 나타나는 정도
    - e.g., 식물 생장 연구의 경우, 온도, 물의 양, 공기 성분 등
  - 변동 사용에 따른 통계적 방법
    - 실험 계획(Experimental Design)
      - 자료에 영향을 주는 요소를 측정자가 제어하여 실험을 통해 분석하는 방법
        - e.g., 식물 생장에 물이 끼치는 영향 연구의 경우, 물의 양을 조절하여 표본을 여러 개로 구성하고 식물의 특징을 조사할 수 있음
    - 관측 연구(Observation Study)
      - 자료에 영향을 주는 요소를 제어하지 않고 관측을 통해 분석하는 방법
        - e.g., 식물 생장 연구의 경우, 길이, 크기, 색 등을 측정하여 식물의 특징을 조사할 수 있음

통계학과 자료

- 자료의 변동(2/2)
  - 교호작용(Interaction)
    - 정의
      - 두 개 이상의 변수가 있을 경우, 하나의 변수가 다른 변수의 효과에 영향을 미치는 상황
    - 평행성(Parallelism)
      - 변수 간의 효과가 일정하게 유지되는지를 나타내는 것
        - e.g., 부식방지제를 통한 알루미늄의 부식 정도 측정
          화학적 코팅이 효과적일 뿐 아니라 코팅은 습도의 영향을 무시할 만 하다.

코팅	습도	파손 평균사이클
비코팅	20%	975
	80%	350
화학적 코팅	20%	1750
	80%	1550



통계학과 자료

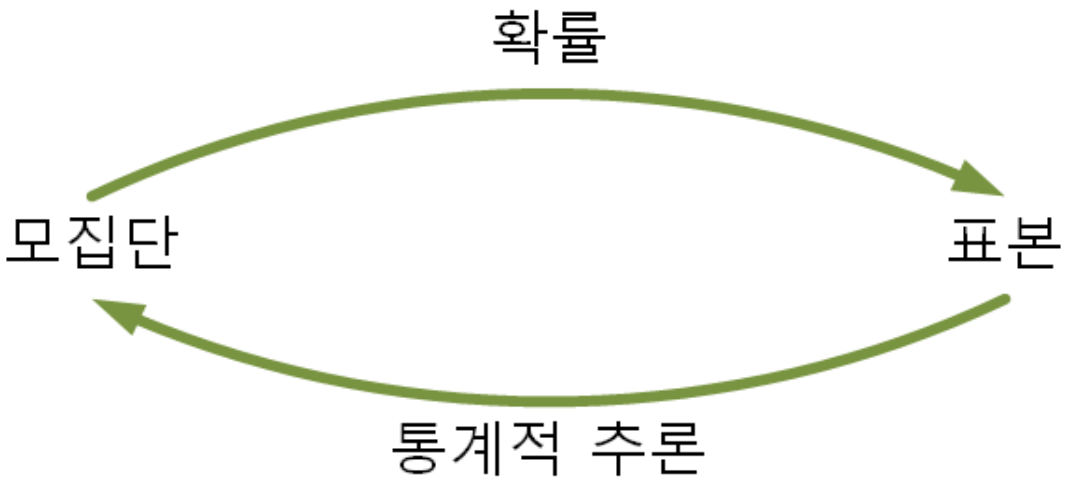
- 확률(Probability)
  - 정의
    - 특정 사건이 발생할 가능성을 수치로 나타낸 것
  - 통계학에서 확률의 역할
    - 자료로부터 얻어지는 추론에 대한 신뢰성 검증
    - P-값(P-Value)
      - 추론의 신뢰성 검증에 대한 지표로 사용되는 값
      - 신뢰성을 입증하기 위해서는 P-값이 0.05 미만이여야 함
        - P-값이 0.05 미만 일 경우, 가설을 만족할 확률이 95% 이상으로 관찰된 효과가 우연이 아닐 가능성이 높다는 것을 나타냄
        - P-값이 0.05 이상 일 경우, 가설을 만족할 확률이 95% 미만으로 관찰된 효과를 입증할 수 없음

통계학과 자료

- 확률과 추론의 관계(1/3)
  - 귀납법(Inductive Reasoning)
    - 특정 사례나 관찰에서 일반적인 결론을 도출하는 방법
      - e.g., 여러 번의 경험을 통해 비가 올 때마다 길이 젖는 것을 관찰할 경우, “비가 오면 길이 젖는다”는 일반적인 결론 도출 가능
  - 연역법(Deductive Reasoning)
    - 일반적인 원칙이나 가정을 바탕으로 특정 상황에 적용되는 결론을 도출하는 방법
      - e.g., 대전제 – 소전제 – 결론의 삼단논법
        대전제: “모든 인간은 죽는다”, 소전제: “소크라테스는 사람이다.”
        그러므로 소크라테스는 죽는다.

통계학과 자료

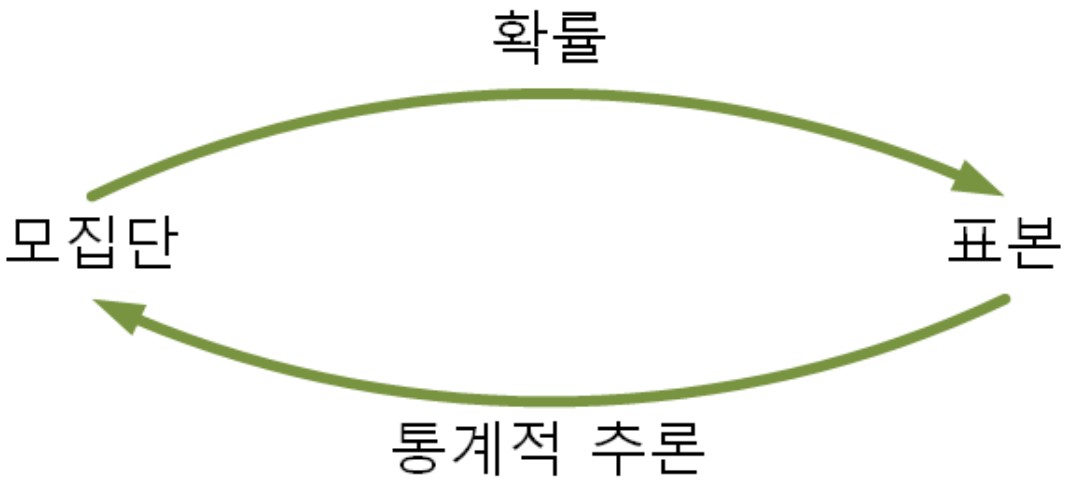
- 확률과 추론의 관계(2/3)



  - 귀납적 방법
    - 표본을 바탕으로 통계적 추론을 통해 모집단에 대한 결론을 이끌어냄
      - e.g., 여러 학교에서 수집한 학생들의 수학 성적 데이터를 기반으로, “해당 지역 학생들의 평균 수학 성적이 X이다.”라는 일반화된 결론 도출

통계학과 자료

- 확률과 추론의 관계(3/3)



- 연역적 방법
  - 확률을 통해 특징들이 알려진 모집단으로부터 가상적으로 추출된 표본 자료의 특성에 대한 결론을 이끌어냄
    - e.g., 정규 분포 모집단으로부터 가상적으로 여러 번 표본을 추출하여, 표본들의 평균과 분산이 모집단의 평균 및 분산과의 유사함을 확인함

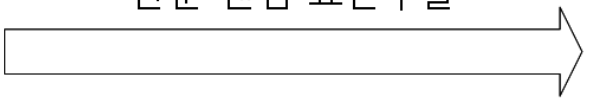
표본추출(Sampling)

- 정의
  - 모집단에서 조건을 만족하는 일부 개체를 추출하는 과정
- 목적
  - 전체 모집단을 대표할 수 있는 표본을 효율적으로 선택하고 분석함으로써 모집단에 대한 결론
    - 통계적 추론(Statistical Inference)
- 종류
  - 단순 랜덤 표본추출(Simple Random Sampling)
  - 층화 랜덤 표본추출(Stratified Random Sampling)
  - 군집 표본추출(Cluster Sampling)

표본추출(Sampling)

- 단순 랜덤 표본추출(Simple Random Sampling)
  - 정의
    - 모집단의 모든 개체가 표본으로 선택될 가능성이 같도록 표본을 추출하는 방법
  - 특징
    - 모집단에 대한 사전 지식 불필요
    - 모집단의 모든 구성원이 추출될 확률이 동등함
      - e.g., 25 개체로 구성된 모집단(N)에서 9개의 표본(n) 추출
        특정 개체가 표본에 포함될 확률: 특정 개체를 포함하는 모든 가능한 표본 조합의 수를 전체 가능한 표본 조합의 수로 나눈 것 $= {}_{N-1}C_{n-1} / {}_{N}C_{n}$

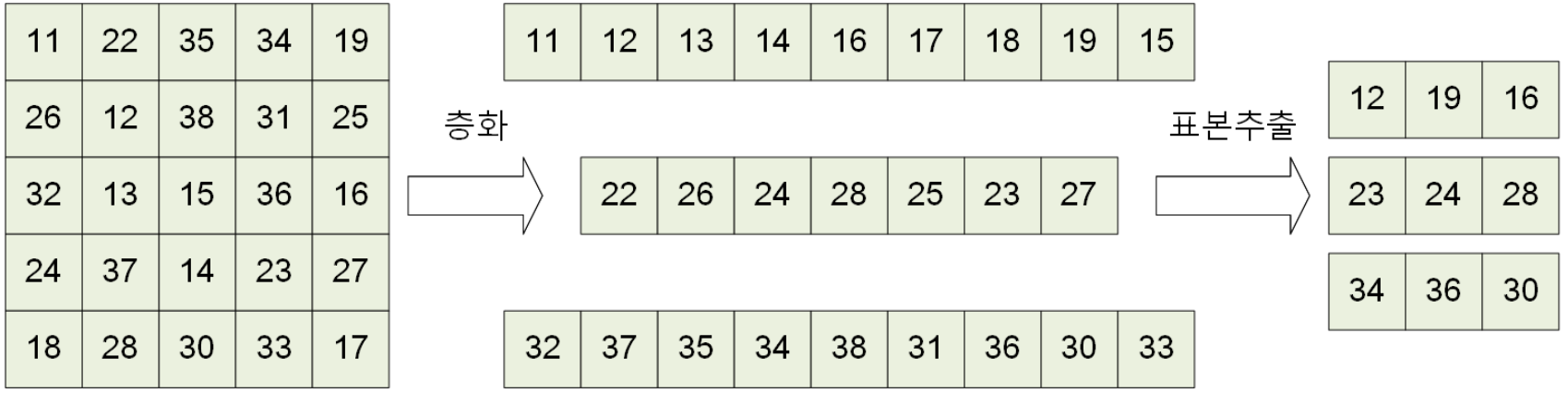
11	22	35	34	19
26	12	38	31	25
32	13	15	36	16
24	37	14	23	27
18	28	30	33	17



11	34	25
23	14	16
24	15	33

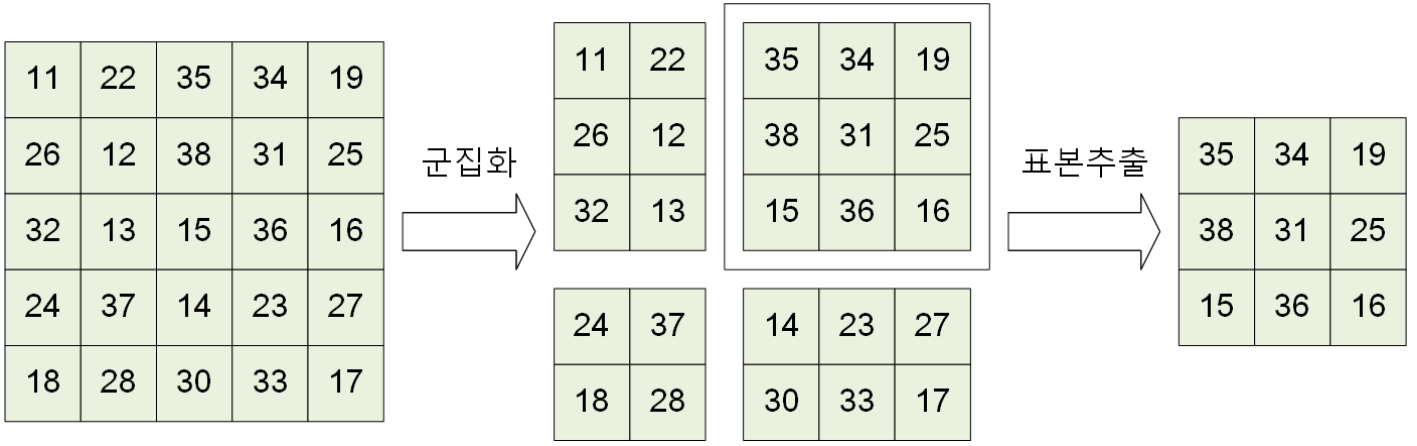
표본추출(Sampling)

- 층화 랜덤 표본추출(Stratified Random Sampling)
  - 정의
    - 모집단을 서로 겹치지 않는 여러 하위 층으로 나누고, 각 층에서 무작위로 표본을 추출하는 방법
  - 특징
    - 다양한 하위 층을 고르게 대표할 수 있어, 모집단의 다양성을 잘 반영함
    - 각 층의 특성에 대한 추정 및 비교 가능
      - e.g., 세종대학교에 재학 중인 모든 학생들을 학년으로 나누어 만족도 조사 수행



표본추출(Sampling)

- 군집 표본추출(Cluster Sampling)
  - 정의
    - 모집단을 여러 군집으로 나누고, 일부 군집을 무작위로 선택하여 군집 내의 모든 개체를 표본으로 추출하는 방법
  - 특징
    - 큰 모집단에서 표본을 추출할 때 비용과 시간 절약 가능
    - 군집 내의 다양성이 모집단 전체의 다양성을 대표하지 못하여 편향 발생 가능성 존재
      - e.g., 전국의 학교들을 지역별로 군집화한 후, 무작위로 선택된 몇 개 지역의 모든 학교를 대상으로 교육 수준을 조사하여 전국적인 교육 수준 평가



측도(Measure)

- 정의
  - 자료의 특정 특성이나 양을 수치적으로 나타내는 방법
- 특징
  - 자료를 해석하는 데에 중요한 특징을 요약
  - 자료 간의 비교를 가능하도록 함
- 종류
  - 위치 측도(Measures of Position)
  - 산포 측도(Measures of Dispersion)

측도(Measure)

- 위치 측도(Measures of Position) (1/2)
  - 정의
    - 자료 내의 특정 값들의 위치의 분포를 나타내는 측도
  - 특징
    - 자료 내에서 특정 관측치가 어디에 위치하는지를 나타내는 값

이름	식	설명
표본 평균 (Sample Mean)	$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$	• 일반적으로 사용하는 산술적 평균
표본 중앙 값 (Sample Median)	$\tilde{x} = \begin{cases} x_{(n+1)/2}, & n\text{이 홀수} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & n\text{이 짝수} \end{cases}$	• 전체 데이터 중 가운데에 있는 수 • 극단 값이나 특이점에 영향 받지 않음
절사 평균 (Trimmed Mean)	$\bar{x}_{tr(P)}$	• 자료의 가장 크거나 작은 일부분을 제외한 평균 • 극단 값이나 특이점에 대해 표본 평균보다 상대적으로 영향을 덜 받음

측도(Measure)

- 위치 측도(Measures of Position) (2/2)
  - e.g., 질소의 사용이 뿌리 성장에 미치는 영향
    두 개의 개별적인 모집단 떡갈나무 묘목 10그루,
    140일 후의 줄기 무게 (단위: g)

무질소	질소
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

표본 평균

- $\bar{x}(\text{무질소}) = \frac{0.32+0.53+0.28+0.37+0.47+0.43+0.36+0.42+0.38+0.43}{10} = 0.399\text{g}$
- $\bar{x}(\text{질소}) = \frac{0.26+0.43+0.47+0.49+0.52+0.75+0.79+0.86+0.62+0.46}{10} = 0.565\text{g}$

표본 중앙값

- $\tilde{x}(\text{무질소}) = \frac{0.38+0.42}{2} = 0.4\text{g}$
- $\tilde{x}(\text{질소}) = \frac{0.49+0.52}{2} = 0.505\text{g}$

절사 평균

- $\bar{x}_{tr(10)}(\text{무질소}) = \frac{0.32+0.37+0.47+0.43+0.36+0.42+0.38+0.43}{8} = 0.3975\text{g}$
- $\bar{x}_{tr(10)}(\text{질소}) = \frac{0.43+0.47+0.49+0.52+0.75+0.79+0.62+0.46}{8} = 0.56625\text{g}$

측도(Measure)

- 산포 측도(Measures of Dispersion) (1/2)
  - 정의
    - 자료들이 중심 위치에서 얼마나 퍼져 있는지를 나타내는 측도
  - 특징
    - 분포의 넓이나 변동성을 나타내는 값

이름	식	설명
표본 범위 (Sample Range)	$x_{max} - x_{min}$	• 가장 단순한 산포도로 표본들의 범위를 나타냄
표본 분산 (Sample Variance)	$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$	• 자료들이 표준 평균으로부터 얼마나 퍼져 있는지 나타내는 측정치
표본 표준 편차 (Sample Standard Deviation)	$s = \sqrt{s^2}$	• 자료 값이 표본 평균 주변에 얼마나 밀집해 있는지 나타내는 측정치 • 제곱 단위인 표본 분산의 단위를 맞춤

측도(Measure)

- 산포 측도(Measures of Dispersion) (2/2)
  - 편향(Bias)
    - 측정값이나 추정치가 실제 값에서 일관되게 벗어나는 경향
  - 자유도(Degrees of Freedom)
    - 자료 추정에서 독립적으로 변할 수 있는 값들의 수
      - 분산의 경우, 평균이 이미 계산되었다면 마지막 자료값은 불변값이므로 자유도가 N-1임
  - e.g., pH 계량기의 편향을 시험하고자 7.0pH값의 물질 10번 측정

pH값				
7.07	7.00	7.10	6.97	7.00
7.03	7.01	7.01	6.98	7.08

표본 범위

- $x_{max} - x_{min} = 7.10 - 6.97 = 0.13$

표본 분산

- $\bar{x} = \frac{7.07+7.00+7.10+6.97+7.00+7.03+7.01+7.01+6.98+7.08}{10} = 7.025\text{pH}$
- $s^2 = \frac{1}{9}\{(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + \cdots + (7.08 - 7.025)^2\} = 0.001939(\text{pH})^2$

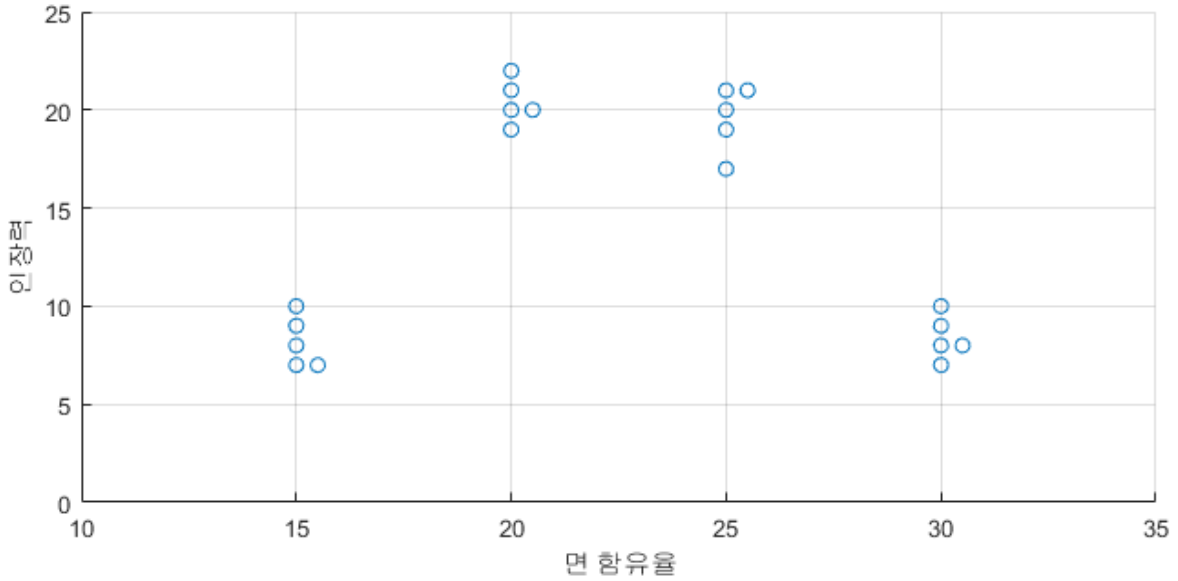
표본 표준편차

- $s = \sqrt{0.001939} = 0.044\text{pH}$

자료의 표현

- 산점도(Scatter Plot)
  - 정의
    - 두 변수 간의 관계를 시각적으로 표현하는 통계적 방법
  - 특징
    - x축, y축에 두 변수의 값들을 표시하여 변수 간의 상관 관계, 패턴, 추세, 이상치 등을 파악함
      - 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 줄기와 잎의 두 부분으로 나누어 시각적으로 표현하는 통계적 방법
  - 특징
    - 데이터의 분포, 중앙값, 이상치 등을 시각적으로 빠르게 파악
    - 중소규모 자료에 적합하며 대규모 자료에는 적합하지 않음
      - e.g., 자료의 값이 34일 경우, 줄기는 3을 가지고 잎은 4를 가짐

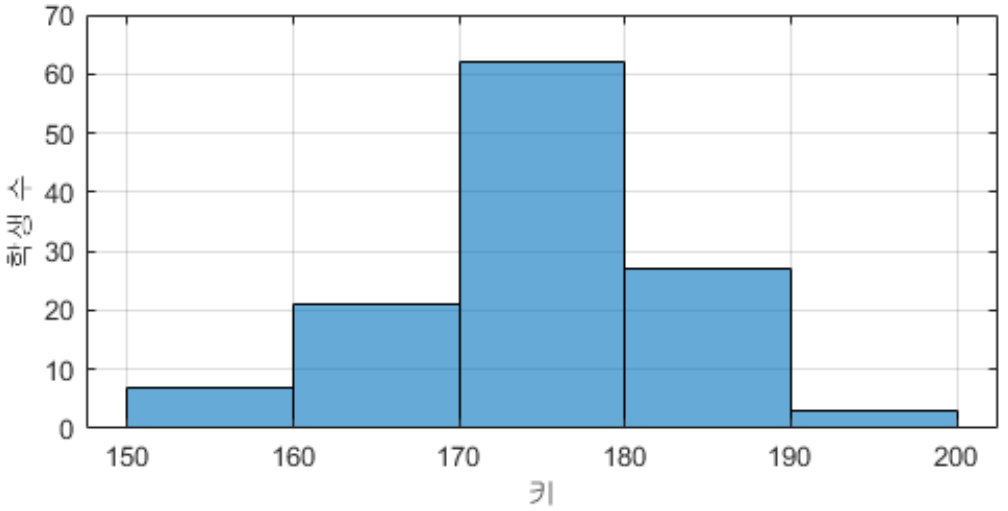
22	41	35	45	32
34	16	31	33	38
25	43	34	36	29
33	31	37	44	32
47	38	32	26	17

줄기	잎	도수
1	67	2
2	2596	4
3	52413846317282	14
4	15347	5

자료의 표현

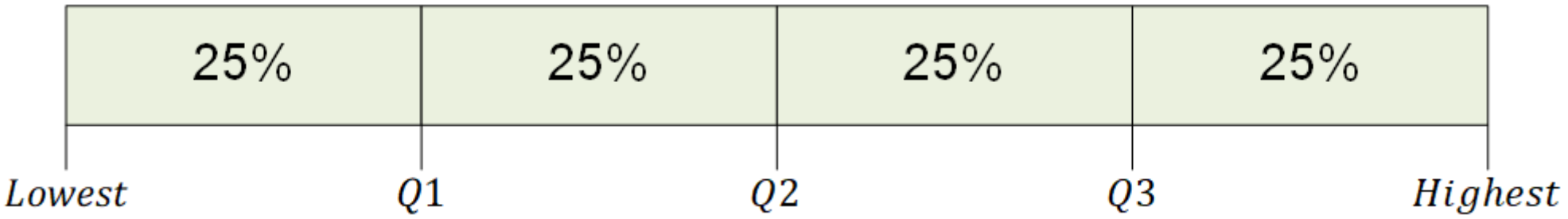
- 히스토그램(Histogram)
  - 정의
    - 연속적인 자료 값의 분포를 막대 그래프 형태로 표현하는 통계적 방법
  - 특징
    - 연속적인 자료 값을 범위로 나누어 어떠한 범위에 집중되어 있는지 시각적으로 나타냄
      - e.g., 정보보호학과 학생들의 키 분포

계급 구간(키)	중간점	학생 수
150 ~ 160	155	7
160 ~ 170	165	21
170 ~ 180	175	62
180 ~ 190	185	27
190 ~ 200	195	3



자료의 표현

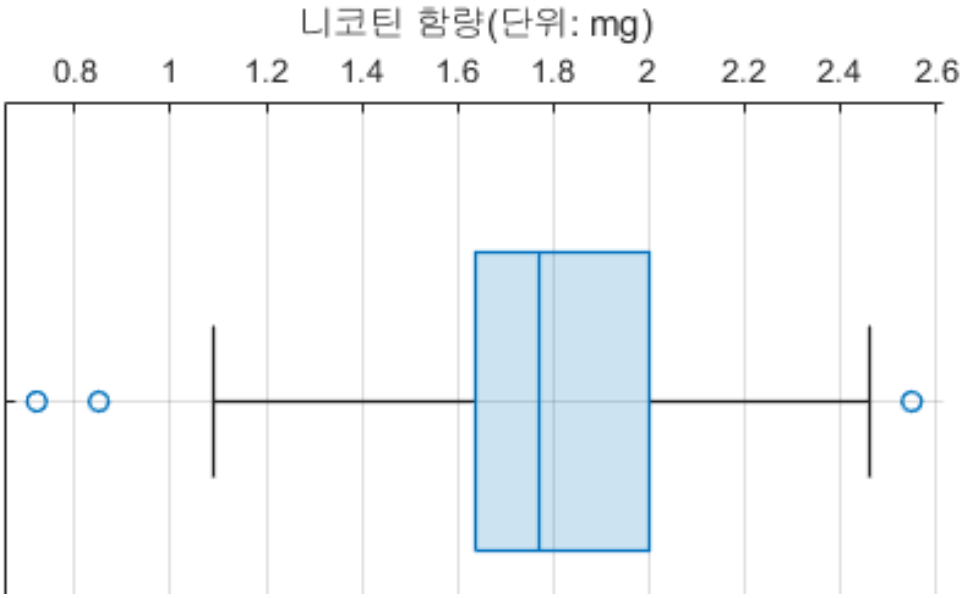
- 상자-수염 그림(Box and Whisker Plot) (1/3)
  - 정의
    - 수치적 자료를 다섯 숫자 요약 값과 사분 범위를 사용하여 그래프 형태로 표현하는 통계적 방법
      - 다섯 숫자 요약
        - 최소값, 제1사분위수(Q1), 중앙값(Q2), 제3사분위수(Q3), 최대값으로 전체 자료를 요약한 것
  - 사분 범위(Interquartile Range)
    - $Q3 - Q1 = IR$
    - 사분 범위의 1.5배 이상 범위에 자료가 있다면 특이점으로 판정할 수 있음



자료의 표현

- 상자-수염 그림(Box and Whisker Plot) (2/3)
  - e.g., 담배 40개비의 니코틴 함량
    - $Q1 = 1.63$
    - $Q2 = 1.75$
    - $Q3 = 1.97$
    - 사분범위 $= 0.34$

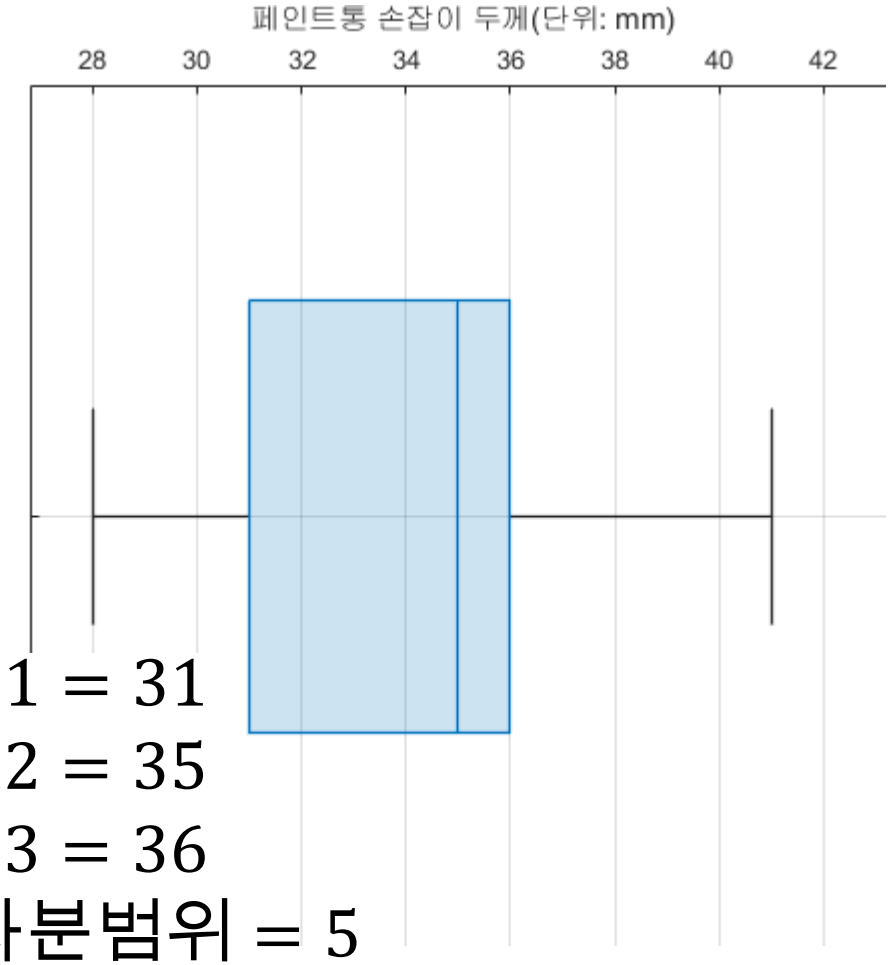
1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69



자료의 표현

- 상자-수염 그림(Box and Whisker Plot) (3/3)
  - e.g., 페인트통 손잡이 두께 측정

표본	측정값	표본	측정값
1	29 36 39 34 34	16	35 30 35 29 37
2	29 29 28 32 31	17	40 31 38 35 31
3	34 34 39 38 37	18	35 36 30 33 32
4	35 37 33 38 41	19	35 34 35 30 36
5	30 29 31 38 29	20	35 35 31 38 36
6	34 31 37 39 36	21	32 36 36 32 36
7	30 35 33 40 36	22	36 37 32 34 34
8	28 28 31 34 30	23	29 34 33 37 35
9	32 36 38 38 35	24	36 36 35 37 37
10	35 30 37 35 31	25	36 30 35 33 31
11	35 30 37 35 31	26	35 30 29 38 35
12	38 34 35 35 31	27	35 36 30 34 36
13	34 35 33 30 34	28	35 30 36 29 35
14	40 35 34 33 35	29	38 36 35 31 31
15	34 35 38 35 30	30	30 34 40 28 30



- $Q1 = 31$
- $Q2 = 35$
- $Q3 = 36$
- 사분범위 $= 5$

통계의 함정

- 정의
  - 통계 데이터를 해석하거나 사용할 경우, 발생할 수 있는 오류나 오해

- 개요
  - 통계를 통해서는 확률적인 결론이 나오지만, 통계의 범위가 커질수록 통제할 수 없는 통제 변수가 늘어남
  - 통계의 함정에 빠져 잘못된 판단을 내릴 수 있음
  - 통계의 논리적 함정 중에서는 거짓이 아닌 것들도 존재하여 통계가 인용되면 오류들을 항상 생각해 두고 면밀히 봐야 함

통계의 함정

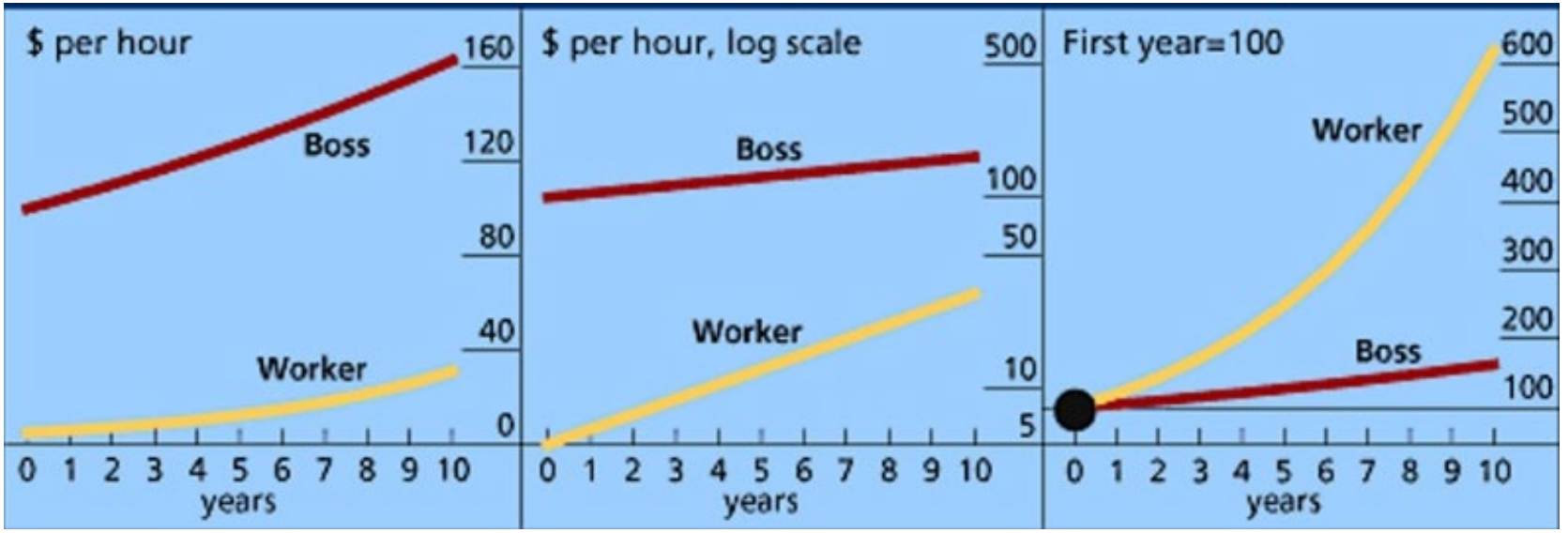
- 잘못된 인과관계 추론
  - 상관관계가 인과관계와 무관할 수 있으며, 직접적인 인과관계를 간과하여 잘못된 결론에 이를 수 있음
    - e.g., 윙슈트 라이더들은 부상 당할 확률이 낮다.
      - 윙슈트 라이더들이 사고가 나면 대부분 죽어버려, 통계에는 부상자가 아닌 사망자로 처리됨
- 의도치 않은 편향
  - 정확하게 수치가 측정되어도, 이를 해석하는 방향이 의도적이지 않게 뒤틀림에 따라 통계가 왜곡됨
    - e.g., 2020년 코로나19 백신 접종자 집단사망 논란
      - 코로나19에 의해 언론의 관심이 극도로 집중된 상황에서 백신접종자가 일정 기간 내에 사망하는 사례가 부각되어 백신 반대 운동이 일어남
      - 백십 접종자의 사망에 관한 실제 수치는 평년 수치와 비슷하였고, 각 사례들은 백신과는 상관없이 평소 기저질환이나 다른 사유로 사망에 이른 경우였음

통계의 함정

- 의도적 편향
  - 특정 집단이 원하는 결과를 도출시키기 위해 의도적으로 편향된 통계를 추론함
    - e.g., 집단이 원하는 답변에는 긍정적인 어휘를, 집단이 원하지 않는 답변에는 부정적인 어휘를 사용하여 답변자가 심리상으로 부정적인 답변을 하지 않게 배치시킴
- 통계적으로 의미 있는 모든 분석에 현실적 의미 부여
  - 통계적으로 의미 있는 결과라 하더라도 현실적으로는 의미가 없을 수 있음
    - e.g., 봄에 태어난 사람이 가을에 태어난 사람보다 키가 0.6cm 크다.
      - Jessica Utts, “What educated citizens should know about statistics and probability”, *The American Statistician*, vol. 57, no.2, pp.74-79, 2023.
      - 현실적으로 키 0.6cm를 위해 출산 시기를 조정할 부모는 거의 없을 것

통계의 함정

- 시각적 자료를 활용한 왜곡
  - 시각적 자료를 통해 편향된 통계를 추론할 수 있음
    - e.g., 세가지 시각적 자료 모두 거짓이 아니며 수학적으로 합리적 그래프
      - 중간 그래프는 노동자들의 임금 증가가 급격하게 이루어져 왔다고 해석될 수 있음
      - 오른쪽 그래프는 노동자들의 임금 증가가 시장의 임금 증가를 초월하고 있다고 해석될 수 있음



<출처> “Logged in”, *The Economist*, pp.76, May 16, 1998.

Thanks!

우 승 찬 (seungchan@pel.sejong.ac.kr)