

이공학도를 위한 확률 및 통계학

-1장 통계학과 자료분석-

송 민 희(minhee@pel.sejong.ac.kr)

세종대학교 프로토콜공학연구실

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

개요

- 통계학(Statistics) (1/3)

- 정의

- 산술적 방법을 기초로 하여, 다량의 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 분야

- 기본 용어

- 모집단(Population)
 - 정보를 얻고자 하는 전체 대상 집합
- 표본(Sample)
 - 모집단에서 추출된 일부 자료의 집합
- 모수(Parameter)
 - 모집단의 특성을 나타내는 기술적 통계치
- 통계량(Statistic)
 - 표본의 특성을 나타내는 기술적 통계치

개요

- 통계학 (2/3)

- 분류 (1/2)

- 기술 통계학

- 정의

- 수집된 자료를 요약·정리하여 데이터의 특성과 분포를 파악하는 학문

- 특징

- 표본 자료에 한정하여 분석을 수행
 - 자료의 수치적 요약 사용하여 자료의 특성과 분포를 파악
e.g., 평균(Mean), 중앙값(Median), 표준편차(Standard Deviation)
 - 그래프를 통해 자료의 분포와 형태를 시각적으로 표현
e.g., 히스토그램(Histogram), 줄기-잎 그림(Stem-and-leaf Plot)

개요

- 통계학 (3/3)

- 분류 (2/2)

- 추론 통계학

- 정의

- 표본집단의 데이터를 분석하여 모집단의 특성을 추정하고, 이를 바탕으로 모집단 전체에 대한 결론이나 가설을 도출하는 학문

- 특징

- 통계적 추론을 통해 표본을 모집단의 특성으로 확장하여 설명
 - 확률 이론을 기반으로 결론의 신뢰도를 평가
 - 표본의 변동성과 불확실성을 고려하여 결론을 도출

*통계적 추론(Statistical Inference)
표본자료를 토대로 하여 모집단 자료에
대해 추정하거나 결과를 도출하는 것

개요

- 자료의 수집

- 정의

- 통계 분석을 위해 필요한 데이터를 획득하는 과정

- 분류 (1/2)

- 직접 관찰

- 대상의 행동이나 현상을 직접 관측하여 데이터를 수집하는 방법
- 실제 상황을 그대로 반영한 자료 획득 가능
- 시간이 많이 들며, 관측자의 주관 개입 가능성 존재

- 실험

- 특정 조건을 인위적으로 조작하고 결과를 관측하여 데이터를 수집하는 방법
- 원인과 결과의 관계를 명확히 한 자료 획득 가능
- 시간과 비용이 많이 들며, 현실 상황을 완전히 반영하기 어려움

개요

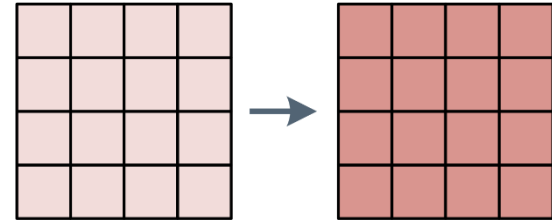
- 자료의 수집

- 분류 (2/2)

- 조사

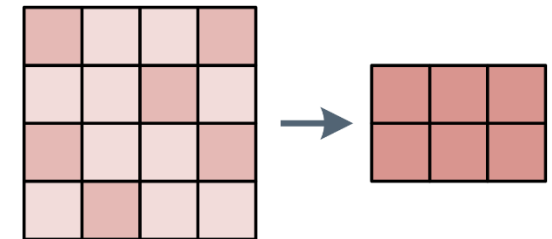
- 전수조사

- 모집단 전체를 대상으로 자료를 수집하는 조사 방법
 - 표본 오차 없이 정확한 모집단 정보 획득 가능
 - 시간과 비용이 많이 소요



- 표본조사

- 모집단의 일부인 표본을 대상으로 자료를 수집하는 조사 방법
 - 시간과 비용을 절감하며 정보 획득 가능
 - 표본 대표성이 부족할 경우 결과 왜곡 가능



개요

- 자료의 분류 (1/2)

- 종류 (1/2)

- 이산형자료(Discrete Data)

- 정의

- 연속적이지 않는 이산적인 값을 갖는 데이터

- 특징

- 관측 결과가 정수 형태로 나타남

- 유형

- 계수자료(Count Data): 정수 값을 가진 데이터

- e.g., 학생 수, 제품 결함 개수

- 이진자료(Binary Data): 두 가지 범주만을 가진 데이터

- e.g., 제품 결함 여부, 시험 합격 여부

- 표본비율(Sample Proportion)

- 이진자료 표본에서 특정 범주가 차지하는 비율

- e.g., 10개의 제품 중 6개에서 결함이 발생할 경우, 표본비율은

- $\frac{6}{10} \times 100 = 60\%$ 로 표현됨

- 표본평균과 동일한 역할

개요

- 자료의 분류 (2/2)

- 종류 (2/2)

- 연속형자료(Continuous Data)

- 정의

- 일정 구간 내에서 연속적인 값을 갖는 데이터
e.g., 길이, 시간, 무게 등

- 특징

- 소수점 단위까지 표현 가능

- 통계적 추론에서의 역할

- 이산형 자료

- 모비율 추정
 - 사건 발생 가능성 판단

- 연속형 자료

- 모평균, 모분산 추정

*모비율(Population Proportion)

모집단의 특정 범주에 속하는 대상의 비율

*모평균(Population Mean)

모집단에 속한 모든 관측값의 중심 위치를 나타내는 상수

*모분산(Population Variance)

모집단에서 관측값들이 모평균을 기준으로 얼마나 퍼져 있는지를 나타내는 상수

개요

- 자료의 사용

- 정의

- 수집된 자료를 분석·해석하여 정보가 불확실한 상황에서 판단의 근거를 제공하는 과정

- 판단의 근거를 요구하는 원인

- 불확실성(Uncertainty)

- 자료의 변동성, 오차, 오류 등에 의해 결과가 일정하지 않아 결과를 확정적으로 알 수 없는 상태

개요

- 자료의 변동 (1/2)

- 변동(Variability)

- 여러 요인에 의해 관측 값이 일정하지 않게 나타나는 성질
e.g., 온도, 습도, 생산량 등에 따라 제품 불량률이 다르게 나타남
- 변동을 다루기 위해 통계학이 필요
 - 시스템을 이해하고 결론을 내리기 위함

- 자료 변동에 따른 수집 방법(1/2)

- 필요성

- 변동이 존재하므로 자료 수집 방법이 결과에 영향을 미칠 수 있음

개요

- 자료의 변동 (2/2)

- 자료 변동에 따른 수집 방법(2/2)

- 실험계획법(Experimental Design)

- 연구자가 요인(factor)과 요인수준을 의도적으로 조절하여 자료를 수집하는 방법

- e.g., 온도와 습도의 정도가 제품의 불량률에 미치는 영향 파악

- 관측연구(Observational Study)

- 요인을 인위적으로 조절하지 않고 자연적으로 발생한 자료를 관측하여 수집하는 방법

- e.g., 기존 생산 공정에서 생산 기록을 수집하여 제품 불량률에 영향을 미치는 요인 파악

개요

- 확률
- 정의
 - 어떠한 사건이 발생할 가능성을 수치로 나타낸 것
- 통계학에서 확률의 역할
 - 통계적 추론에서 도출된 결론의 확신 정도를 계량화
 - 기술통계학과 추론통계학을 연결하여 결과 해석을 도움
 - P-값(P-value)
 - 관측된 결과가 우연일 가능성을 수치로 나타낸 값
 - 통계적 추론에서 결론의 타당성을 판단하기 위해 사용
 - $P\text{-값} < \text{유의수준}(\text{보통 } 0.05)$
 - 해당 결과가 우연이 아닐 가능성이 높음
 - $P\text{-값} \geq \text{유의수준}$
 - 해당 결과가 우연일 가능성이 높음

개요

• 확률의 역할 (1/4)

*확률질량함수(PMF, Probability Mass Function)
이산확률변수가 특정 값 하나를 가질 확률을 구하는 함수

• 예제 1.1 (1/2)

어느 제조공정에서는 가끔 불량품이 제조되는 것이 예상되지만, 이 회사는 공정 불량률을 최대 5%까지 허용할 수 있다고 한다. 임의로 제품 100개를 추출하여 검사한 결과, 불량품 10개가 발견되었다.

불량률이 5%라면, 임의로 추출된 100개 중 불량품이 10개 이상 포함될 확률을 구하고 해당 확률값을 이용하여 공정의 상태에 대해 판단하라.

- 확률모형 설정
 - 한 제품이 불량일 확률 $p = 0.05$
 - 표본 크기 $n = 100$
 - 확률 변수 X : 100개 중 불량품의 개수 (이항분포)
- 확률 계산
 - 구하고자 하는 확률: $P(X \geq 10)$
 - 이항 분포의 확률질량함수:

$$P(X = x) = \binom{100}{x} (0.05)^x (0.95)^{100-x} \Rightarrow P(X \geq 10) = \sum_{x=10}^{100} \binom{100}{x} (0.05)^x (0.95)^{100-x}$$
$$\Rightarrow P(X \geq 10) = 0.0282 = 2.82\%$$

개요

- 확률의 역할 (2/4)

- 예제 1.1 (2/2)

어느 제조과정에서는 가끔 불량품이 제조되는 것이 예상되지만, 이 회사는 공정 불량률을 최대 5%까지 허용할 수 있다고 한다. 임의로 제품 100개를 추출하여 검사한 결과, 불량품 10개가 발견되었다.

불량률이 5%라면, 임의로 추출된 100개 중 불량품이 10개 이상 포함될 확률을 구하고 해당 확률값을 이용하여 공정의 상태에 대해 판단하라.

- 공정 상태 판단

- 실제 불량률이 5%라면, 임의로 추출한 100개 중 10개 이상의 불량품이 나오는 경우가 우연일 확률은 2.82%
- 확률이 매우 작아 해당 표본과 같은 결과가 우연이라고 보기 어려움
- ∴ 공정의 실제 불량률이 5%를 초과했을 가능성이 있으므로 공정이 수용할만한 상태가 아님.

개요

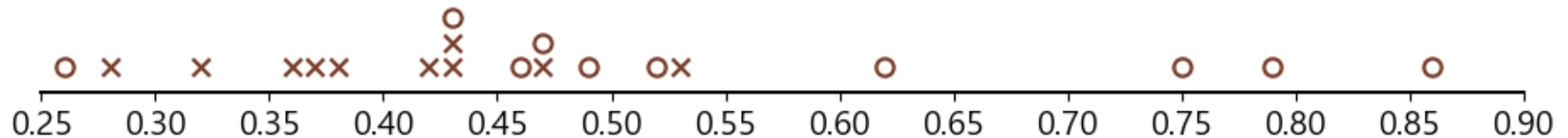
• 확률의 역할 (3/4)

• 예제 1.2 (1/2)

버지니아 대학교에서 나무뿌리와 균류의 작용 사이의 관계를 연구하였다. 북부의 붉은 떡갈나무 묘목을 각각 10그루씩 두 집단으로 나누어, 한 집단에는 질소를 함유시키고 다른 집단에는 질소 없이 온실에 심었다. 그 밖의 모든 조건은 동일하게 하였다. 140일 후 측정된 줄기의 무게(단위 g)는 우측의 표와 같다. 줄기 무게에 대한 점도표를 그리고 분포를 분석하라.

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

▲ 예제 1.2의 자료



▲ 줄기 무게의 점도표

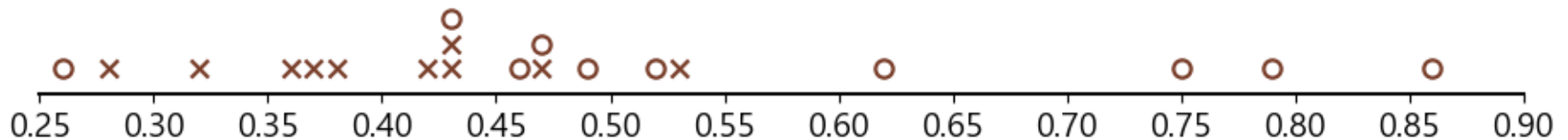
- ○로 표시된 질소 처리 집단의 관측값들은 전반적으로 큰 값 쪽에 분포
- ×로 표시된 무질소 집단의 관측값들은 자료 중심의 아래쪽에 모여 있음
- 질소 처리로 얻어진 4개의 관측값은 무질소 집단의 어떤 관측값보다 큼

개요

- 확률의 역할 (4/4)

- 예제 1.2 (2/2)

점도표에서 질소 처리 집단과 무질소 집단 사이에 뚜렷한 분리가 관찰된다. 질소의 효과가 전혀 없을 경우, 이와 같은 분리 정도가 나타날 확률이 0.03이라고 가정할 때, 이를 바탕으로 질소의 효과 유무를 판단하라.



▲ 줄기 무게의 점도표

- 점도표에서 질소 처리 집단의 관측값들이 무질소 집단보다 전반적으로 큰 값 쪽에 분포
- 질소 효과가 없다는 가정하에 동일한 분리 정도가 나타날 확률이 0.03이라고 가정
 - 해당 값은 매우 작은 확률
 - 위와 같은 결과가 우연히 발생할 가능성이 낮다고 판단됨
 - ∴ 질소의 사용이 줄기 무게(성장)에 영향을 끼쳤다고 판단됨

개요

- 확률과 통계적 추론의 관계 (1/2)

- 귀납법(Induction)

- 정의

- 개별적인 관측 결과로부터 일반적인 결론을 도출하는 추론 방식

- 예시

- ‘A는 죽었다.’, ‘B는 죽었다.’와 같이 개별 사례를 모아 ‘모든 사람은 죽는다.’라는 일반적인 결론 도출

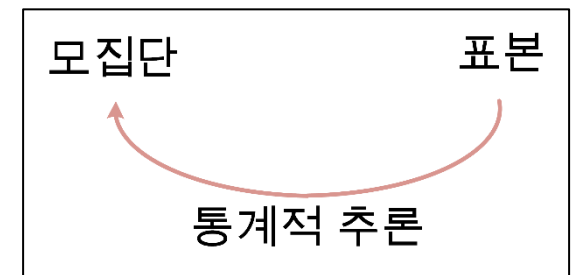
- 귀납적 방법(통계적 추론)

- 정의

- 관측된 표본 결과를 바탕으로 모집단에 대한 결론을 내리는 방법

- 예시

- 일부 표본 제품의 불량 여부 관측을 통해 전체적인 공정 상태를 판단



개요

- 확률과 통계적 추론의 관계 (2/2)

- 연역법(Deduction)

- 정의

- 일반적인 가정이나 법칙(대전제)으로부터 특정 사례(소전제)에 적용하여 특정 결과를 도출하는 추론 방식

- 예시

- ‘모든 사람은 죽는다.’라는 대전제에서, ‘A는 사람이다.’라는 소전제를 통해 ‘A는 죽는다.’라는 필연적 결론 도출

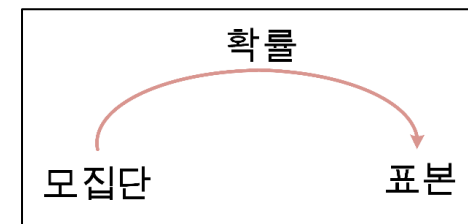
- 연역적 방법(확률 계산)

- 정의

- 모집단에 대한 가정을 바탕으로 표본 결과의 가능성을 계산하는 방법

- 예시

- 어떤 제품의 불량률이 5%라고 가정했을 때, 특정 표본 결과가 나올 확률 계산



목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

표본추출

- 정의

- 모집단에서 일부 개체를 선택하여 자료를 수집하는 방법

- 특징

- 표본의 선택 방법에 따라 결과 신뢰도 변화
- 표본의 대표성이 분석 결과에 직접적 영향
- 표본추출 방법에 따른 편향 발생 가능성 존재

- 분류 (1/4)

- 단순 랜덤 표본 추출(Simple Random Sampling)
- 층화 랜덤 표본 추출(Stratified Random Sampling)
- 군집 표본 추출(Cluster Sampling)

표본추출

- 분류 (2/4)

- 단순 랜덤 표본 추출

- 개요

- 모집단의 모든 개체가 동일한 확률로 선택되도록 표본을 추출하는 방법
 - 무작위 선택에 기반한 표본추출 방식

53	57	32	11	50
39	13	36	47	19
44	31	61	30	55
38	22	41	16	27
14	34	33	12	25

단순랜덤
표본추출

53	32	11
50	47	61
22	41	27

- 장점

- 모집단에 대한 사전 지식 불필요
 - 통계적 분석과 해석이 단순

- 단점

- 표본 크기가 작을 경우 대표성 부족
 - 편향표본(Biased Sample)
 - 제한된 모집단에 한정하여 유효하게 추출된 표본
e.g., 정치적 성향을 추출하는데, 한 지역에서만 추출하는 경우

표본추출

- 분류 (3/4)

*층(strata)

모집단을 특정 기준에 따라 겹침 없이 나눈 하위 집단

- 층화 랜덤 표본 추출

- 개요

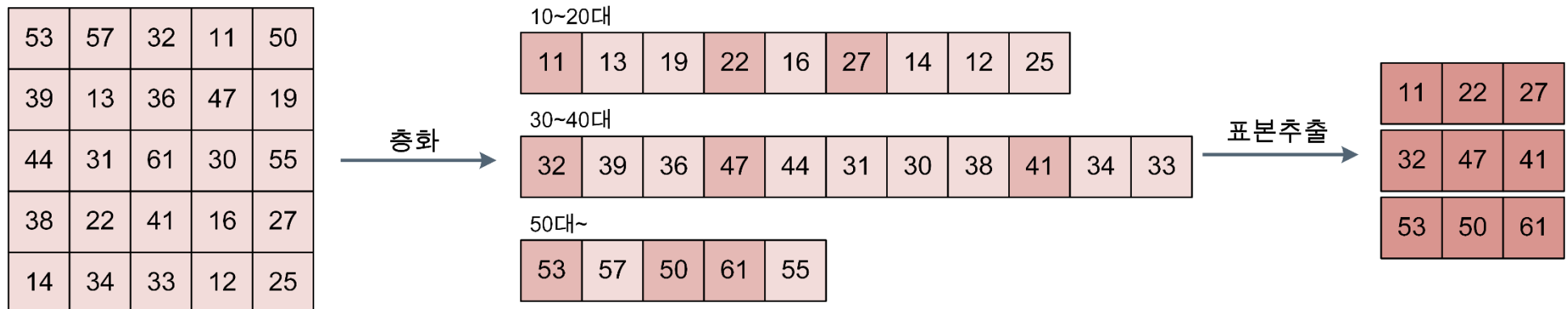
- 모집단을 여러 층으로 구분하고 각 층에서 단순 랜덤 표본 추출을 수행하여 표본을 추출하는 방법

- 장점

- 모든 층이 표본에 포함되도록 구성
 - 표본의 대표성 향상

- 단점

- 사전 정보 및 층 구분 기준 설정 필요



표본추출

- 분류 (4/4)

*군집(Cluster)

모집단을 자연적으로 형성된 집단 단위로 나눈 하위 집단

- 군집 표본 추출

- 개요

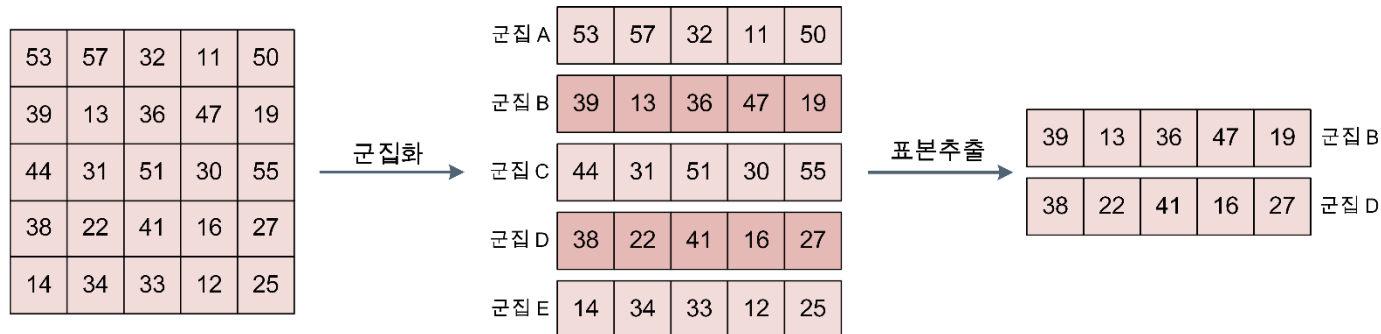
- 모집단을 여러 군집으로 나누고 일부 군집을 무작위로 선택하여 선택된 군집 전체 또는 일부에 대해 표본을 추출하는 방법

- 장점

- 표본 추출 과정이 간단하여 조사 비용 및 시간 절감 가능
 - 모집단 전체 목록이 없을 때도 적용 가능

- 단점

- 선택된 군집이 모집단을 잘 대표하지 못할 가능성 존재
 - 군집 간 이질성이 크면 추정 정확도 감소



표본추출

- 실험계획법 (1/3)

- 정의

- 연구자가 요인과 요인수준을 의도적으로 조절하여 자료를 수집하는 방법

- 실험계획법 핵심 요소

- 개요

- 실험 단위를 처리 또는 처리 조합에 무작위로 배정하기 위함
- 처리 간 비교의 공정성 확보를 위한 필수 조건

- 요소

- 임의성(Randomness)
- 랜덤할당(Random Assignment)

표본추출

- 실험계획법 (2/3)

- 랜덤포본추출의 관점 (1/2)

- 관련 용어

- 처리

- 실험에서 비교하고자 하는 조건 또는 요인 수준

- 처리조합

- 두 개 이상의 요인이 결합된 실험조건
 - 실험계획에서는 각 처리조합을 하나의 비교 대상 모집단으로 간주

- 실험 단위

- 처리가 실제로 적용되는 최소단위
 - 실험 단위간 자연적 차이로 인해 측정 값의 변동이 발생

- 변동의 역할

- 변동이 과도할 경우 처리 효과가 가려져 검출이 어려워짐
 - 실험계획법의 목적
 - 변동을 통제하거나 분리하는 것

표본추출

- 실험계획법 (3/3)

- 랜덤포본추출의 관점 (2/2)

- 실험 단위의 랜덤 할당의 필요성

- 실험 단위의 특성이 특정 처리에 치우치는 것을 방지
 - 처리 간 차이를 공정하게 비교할 수 있는 기반 제공

- 완전확률화설계(Completely Randomized Design)

- 정의

- 모든 실험단위를 처리 또는 처리조합에 완전히 무작위로 할당하는 설계

- 특징

- 보편적이고 표준적인 접근법
 - 실험단위 수가 충분하고 조건이 단순한 경우에 적합
e.g., 환자 100명에게 위약을, 나머지 100명에게 처방약을 랜덤하게 할당

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

위치의 척도

- 위치척도 (1/7)

- 정의

- 자료의 중심이 어디에 위치하는지를 나타내는 수치적 척도

- 특징

- 표본 자료의 대표값을 요약하는데 사용
- 기술통계학의 핵심 구성 요소

- 종류

- 표본평균(Sample Mean)
- 표본중앙값(Sample Median)
- 절사평균(Trimmed Mean)

위치의 측도

- 위치측도 (2/7)

- 종류 (1/3)

- 표본평균

- 정의

- 표본 관측값들의 산술적 평균

- 특징

- 모든 관측값들을 동일한 가중치로 반영
 - 자료 전체 정보를 활용
 - 극단값 또는 특이점(outlier)에 민감

- 수식

표본의 관측값을 $x_1, x_2, \dots, x_n$ 이라고 하면, 표본 평균은

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

위치의 측도

- 위치측도 (3/7)

- 종류 (2/3)

- 표본중앙값

- 정의

- 표본 관측값을 크기순으로 정렬했을 때 가운데 위치한 값

- 특징

- 관측값 개수에 따라 단일값 또는 두 값의 평균을 가짐
 - 자료의 중앙 경향을 표현함
 - 극단값 또는 특이점(outlier)의 영향에 강건함

- 수식

표본의 관측값이 크기순으로 $x_1, x_2, \dots, x_n$ 이라고 하면, 표본중앙값은

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

위치의 측도

- 위치측도 (4/7)

- 종류 (3/3)

- 절사평균

- 정의

- 자료의 양 극단 일부를 제거한 후 계산한 평균

- 특징

- 평균과 중앙값의 절충적 위치
 - 극단값의 영향을 완화하면서 정보 손실 최소화

위치의 측도

- 위치측도 (5/7)

- 예제 1.2 (2) (1/3)

표본자료의 중심위치를 계량화하기 위한 여러 가지 측도가 있다. 다음은 예제 1.2에서 제시된 줄기 무게 자료이다.

- (1) 각 표본에 대해 표본평균과 중앙값을 구하라.
- (2) 각 표본에 대해 10% 절사평균을 구하라.
- (3) 세 위치의 측도를 비교하여 특징을 간략히 설명하라.

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

▲ 예제 1.2의 자료

- 무질소

- 표본평균 $\bar{x} = \frac{0.32+0.53+0.28+0.37+0.47+0.43+0.36+0.42+0.38+0.43}{10} = 0.399$

- 중앙값 $\tilde{x} = \frac{0.38+0.42}{2} = 0.4$

- 질소

- 표본평균 $\bar{x} = \frac{0.26+0.43+0.47+0.49+0.52+0.75+0.79+0.86+0.62+0.46}{10} = 0.565$

- 중앙값 $\tilde{x} = \frac{0.49+0.52}{2} = 0.505$

위치의 측도

- 위치측도 (6/7)

- 예제 1.2 (2) (2/3)

표본자료의 중심위치를 계량화하기 위한 여러 가지 측도가 있다. 다음은 예제 1.2에서 제시된 줄기 무게 자료이다.

- (1) 각 표본에 대해 표본평균과 중앙값을 구하라.
- (2) 각 표본에 대해 10% 절사평균을 구하라.
- (3) 세 위치의 측도를 비교하여 특징을 간략히 설명하라.

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

▲ 예제 1.2의 자료

- 무질소

- 10% 절사평균 $\bar{x}_{tr(10)} = \frac{0.32+0.37+0.47+0.43+0.36+0.42+0.38+0.43}{8} = 0.39750$

- 질소

- 10% 절사평균 $\bar{x}_{tr(10)} = \frac{0.43+0.47+0.49+0.52+0.75+0.79+0.62+0.46}{8} = 0.56625$

위치의 측도

• 위치측도 (7/7)

• 예제 1.2 (2) (3/3)

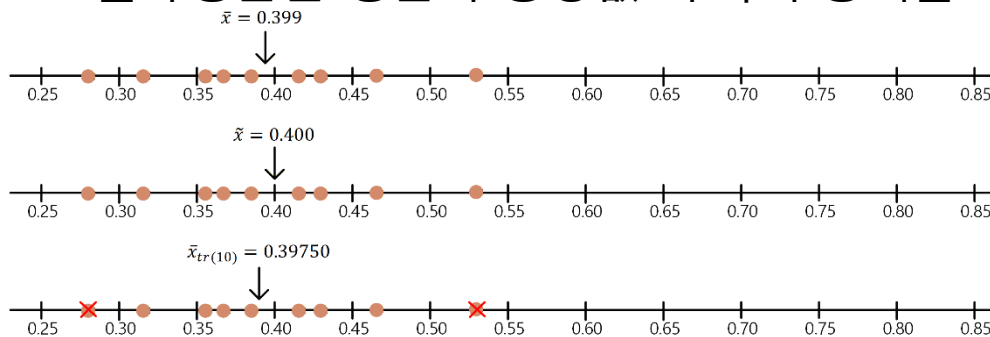
표본자료의 중심위치를 계량화하기 위한 여러 가지 측도가 있다. 다음은 예제 1.2에서 제시된 줄기 무게 자료이다.

- (1) 각 표본에 대해 표본평균과 중앙값을 구하라.
- (2) 각 표본에 대해 10% 절사평균을 구하라.
- (3) 세 위치의 측도를 비교하여 특징을 간략히 설명하라.

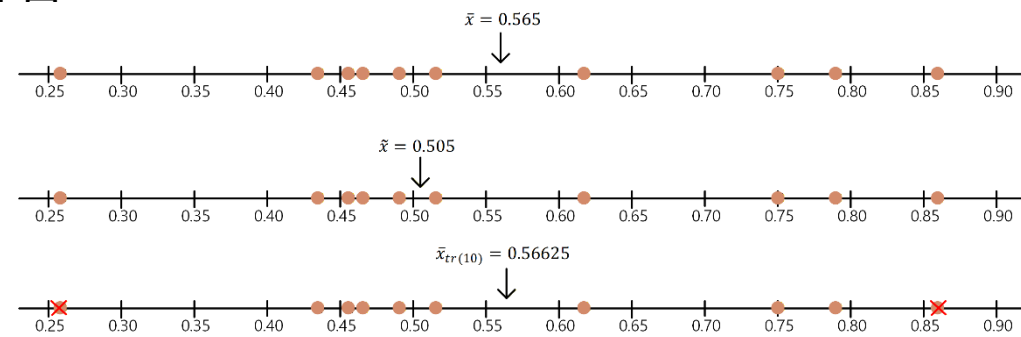
- 표본평균은 극단값의 영향을 받음
- 중앙값은 극단값에 민감하지 않음
- 절사평균은 평균보다 극단값에 덜 민감함
- 절사평균은 평균과 중앙값 사이의 성격을 가짐

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

▲ 예제 1.2의 자료



▲ 무질소 표본에 대한 위치측도



▲ 질소 표본에 대한 위치측도

위치의 측도

• 위치측도의 한계 (1/4)

• 예제 1.3 (1/4)

부식방지제를 알루미늄에 도장했을 때 부식의 정도가 감소하는지를 알아보기 위한 실험을 수행하였다. 도장의 종류는 비코팅과 화학적 코팅, 습도 수준은 20%와 80%로 하였으며, 각 처리 조합마다 알루미늄 시편 2개씩을 무작위로 할당하였다. 부식 정도는 파손될 때까지의 사이클 수(단위: 1,000)로 측정하였고, 각 처리조합의 평균값은 아래의 표와 같다.

- (1) 표 1.2의 평균값을 이용하여, 습도와 도장 종류에 따른 부식 정도를 그래프로 나타내라.
- (2) (1)에서 작성한 그래프를 이용하여 도장과 습도의 영향을 분석하라.
- (3) 평균값만을 사용한 분석의 한계를 설명하라.

(단위: 1,000)

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

▲ 예제 1.3의 자료

위치의 측도

• 위치측도의 한계 (2/4)

• 예제 1.3 (2/4)

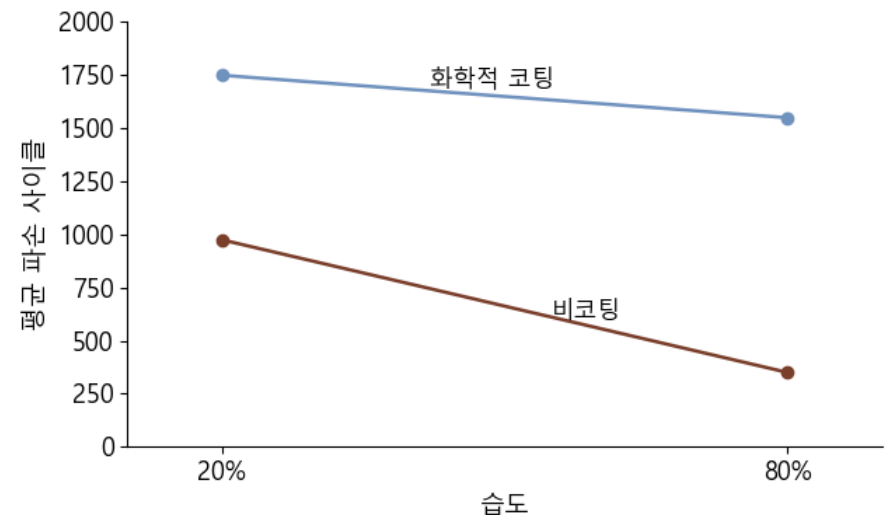
부식방지제를 알루미늄에 도장했을 때 부식의 정도가 감소하는지를 알아보기 위한 실험을 수행하였다. 도장의 종류는 비코팅과 화학적 코팅, 습도 수준은 20%와 80%로 하였으며, 각 처리 조합마다 알루미늄 시편 2개씩을 무작위로 할당하였다. 부식 정도는 파손될 때까지의 사이클 수(단위: 1,000)로 측정하였고, 각 처리조합의 평균값은 아래의 표와 같다.

(1) 표 1.2의 평균값을 이용하여, 습도와 도장 종류에 따른 부식 정도를 그래프로 나타내라.

(단위: 1,000)

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

▲ 예제 1.3의 자료



▲ 예제 1.3의 부식 결과 그래프

위치의 측도

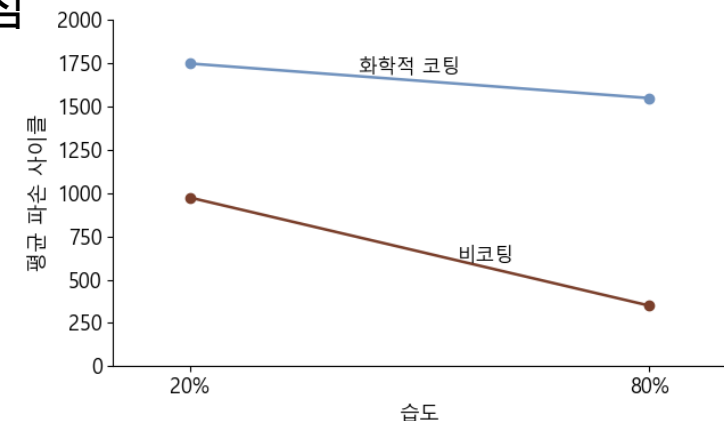
• 위치측도의 한계 (3/4)

• 예제 1.3 (3/4)

부식방지제를 알루미늄에 도장했을 때 부식의 정도가 감소하는지를 알아보기 위한 실험을 수행하였다. 도장의 종류는 비코팅과 화학적 코팅, 습도 수준은 20%와 80%로 하였으며, 각 처리 조합마다 알루미늄 시편 2개씩을 무작위로 할당하였다. 부식 정도는 파손될 때까지의 사이클 수(단위: 1,000)로 측정하였고, 각 처리조합의 평균값은 아래의 표와 같다.

(2) (1)에서 작성한 그래프를 이용하여 도장과 습도의 영향을 분석하라.

- 습도 증가에 따라 두 도장 모두 평균 사이클 수 감소 경향 확인
- 화학적 코팅이 비코팅보다 전반적으로 높은 평균값을 가짐
- 화학적 코팅이 부식을 감소시키는 효과를 가짐을 시사함



▲ 예제 1.3의 부식 결과 그래프

위치의 측도

• 위치측도의 한계 (4/4)

• 예제 1.3 (4/4)

부식방지제를 알루미늄에 도장했을 때 부식의 정도가 감소하는지를 알아보기 위한 실험을 수행하였다. 도장의 종류는 비코팅과 화학적 코팅, 습도 수준은 20%와 80%로 하였으며, 각 처리 조합마다 알루미늄 시편 2개씩을 무작위로 할당하였다. 부식 정도는 파손될 때까지의 사이클 수(단위: 1,000)로 측정하였고, 각 처리조합의 평균값은 아래의 표와 같다.

(3) 평균값만을 사용한 분석의 한계를 설명하라.

- 각 평균값은 시편 2개의 평균에 불과함
- 처리조합 내 변동에 대한 정보가 포함되지 않음
- 평균 차이가 우연에 의한 것인지 판단 불가능함
- 변동성을 고려하지 않으면 분석 결과가 과장될 수 있음
- 추가적인 변동 측정(분산, 표준편차 등)의 필요성 제기됨

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

산포의 척도

- 산포(Variability) (1/6)

- 정의

- 자료가 위치척도를 기준으로 흩어져 있는 정도를 나타내는 척도

- 중요성

- 동일한 평균을 가지는 자료라도, 분포 형태가 다른 경우 존재
 - 산포가 작은 경우 통계적 비교 및 추론의 신뢰도 증가
 - 위치척도만으로는 자료 분석에 한계 존재
 - 평균은 자료의 중심만 요약

- 종류 (1/4)

- 표본범위(Sample Range)
 - 표본분산(Sample Variance)
 - 표본표준편차(Sample Standard Deviation)

산포의 측도

- 산포 (2/6)
 - 종류 (2/4)
 - 표본범위
 - 정의
 - 표본 자료값 중 최대값과 최소값의 차이
 - 특징
 - 계산이 매우 간단함
 - 극단값에 매우 민감함
 - 전체 분포 특성을 충분히 반영하지 못함
 - 수식

표본의 자료값을 $x_1, x_2, \dots, x_n$ 이라고 하면, 표본범위값은

$$X_{max} - X_{min}$$

산포의 측도

- 산포 (3/6)

*자유도(Degrees of Freedom)

표본분산을 계산할 때 사용 가능한 독립적인 정보의 개수

- 종류 (3/4)

- 표본분산

- 정의

- 모집단에서 추출한 표본 데이터들의 평균으로부터 흩어져 있는 정도

- 특징

- 변동의 크기를 정량적으로 측정

- 모든 관측값이 계산에 반영

- 단위가 원래 자료의 제곱

- 수식

표본의 자료값을 $x_1, x_2, \dots, x_n$ 이라고 하면, 표본분산(s^2)은

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

산포의 측도

- 산포 (4/6)

- 종류 (4/4)

- 표본표준편차

- 정의

- 모집단에서 추출한 표본 데이터들의 평균으로부터 흩어져 있는 정도

- 특징

- 단위가 원래 자료와 동일하게 사용됨
 - 자료의 평균적 변동 크기 표현

- 수식

표본의 자료값을 $x_1, x_2, \dots, x_n$ 이라고 하면, 표본표준편차(s)는

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

산포의 측도

- 산포 (5/6)

*모집단 모수(Population Parameter)
모집단의 특성을 나타내는 고정된 상수 값

- 변동측도 활용

- 표본범위
 - 통계적 품질관리에서 활용
- 표본분산
 - 변동의 수학적 측정
- 표본표준편차
 - 원래 단위에서의 변동 크기 표현

- 모집단 모수와의 관계

- 모평균(Population Mean)과 모분산(Population Variance)에 대한 추론을 위한 핵심 역할
 - 모분산 추정
 - 표본 분산
 - 모평균 추정 및 불확실성 표현
 - 표본평균 + 표본표준편차

산포의 측도

- 산포 (6/6)

- 예제 1.4

어느 기사가 pH 계량기의 '편향(bias)'을 시험하고자 7.0pH 값을 갖는 물질을 10번 측정하여 다음과 같은 자료를 얻었다.

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

이에 대하여 표본평균, 표본 분산, 표본표준편차, 자유도를 구하라

- 표본평균 $\bar{x} = \frac{7.07+7.00+7.10+\cdots+7.08}{10} = 7.025$
- 표본분산 $s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939$
- 표본표준편차 $s = \sqrt{0.00193} = 0.044$
- 자유도 $= 10 - 1 = 9$

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

통계적 모형화, 과학적 조사, 그래프 진단

- 산점도(Scatter Plot) (1/2)

- 정의

- 두 변수 간의 관계를 점으로 나타낸 그래프

- 특징

- x축, y축에 각각 하나의 변수 배치
 - 변수 간의 상관 관계, 추세, 패턴, 이상치를 시각적으로 확인

- 활용

- 두 변수 사이의 관계 탐색
 - 회귀분석 및 상관분석의 기초 자료

통계적 모형화, 과학적 조사, 그래프 진단

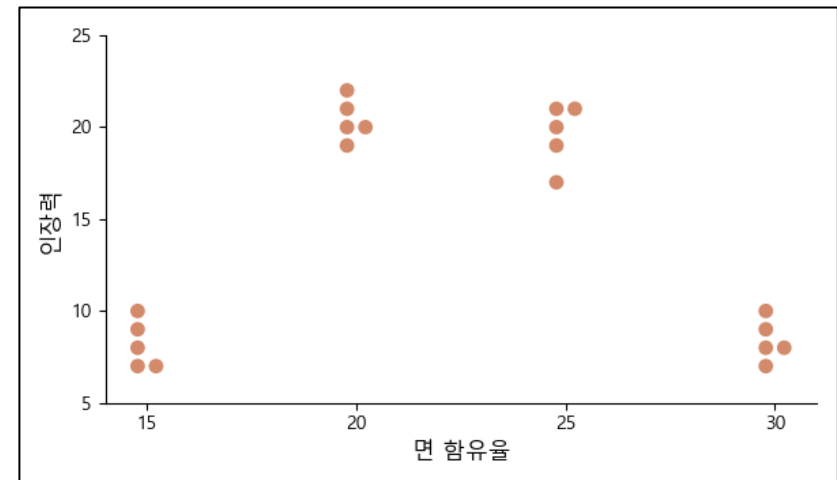
• 산점도 (2/2)

• 예제

한 섬유 제조업자는 여러가지의 면 함유율로 생산된 원사에 대한 실험을 실시하였다. 이에 대하여 산점도 그래프로 표현하여라.

면 함유율	원사의 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

▲ 인장강도



▲ 면 함유율과 인장력의 산점도 그래프

통계적 모형화, 과학적 조사, 그래프 진단

- 줄기-잎 그림(Stem-and-Leaf Plot) (1/2)

- 정의

- 관측값을 줄기와 잎으로 나누어 분포를 나타내는 그래프

- 특징

- 원자료가 그대로 보존됨
- 소규모 자료에 적합

- 활용

- 자료의 분포와 개별 값 동시 확인
- 탐색적 자료 분석(Exploratory Data Analysis)

통계적 모형화, 과학적 조사, 그래프 진단

• 줄기-잎 그림 (2/2)

• 예제

40개의 자동차를 대상으로 축전지의 수명을 조사하였다.
해당 축전지의 수명은 적어도 3년은 지속된다는 것이 보증 되어있다.
각각의 관측값에 대해 줄기-잎 그림으로 표현하여라.

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

▲ 자동차 축전지 수명

줄기	잎	도수
1	69	2
2	25696	5
3	4318514723628297130097145	25
4	71354172	8

▲ 축전지 수명의 줄기-잎 그림

줄기	잎	도수
1	69	2
2*	2	1
2	5696	4
3*	431142622130014	15
3	8576897975	10
4*	13412	5
4	757	3

▲ 축전지 수명의 이중 줄기-잎 그림

통계적 모형화, 과학적 조사, 그래프 진단

- 히스토그램(Histogram) (1/2)

- 정의

- 연속형 자료를 계급으로 나누어 도수 또는 상대도수(Relative Frequency)를 막대로 나타낸 그래프

- 특징

- 막대 사이에 공백이 없음
- 자료의 분포 형태, 중심, 산포 파악 가능
- 계급 폭에 따라 그래프 모양이 달라질 수 있음

- 활용

- 연속형 자료의 전체 분포 파악
- 확률분포 가정의 시각적 확인

통계적 모형화, 과학적 조사, 그래프 진단

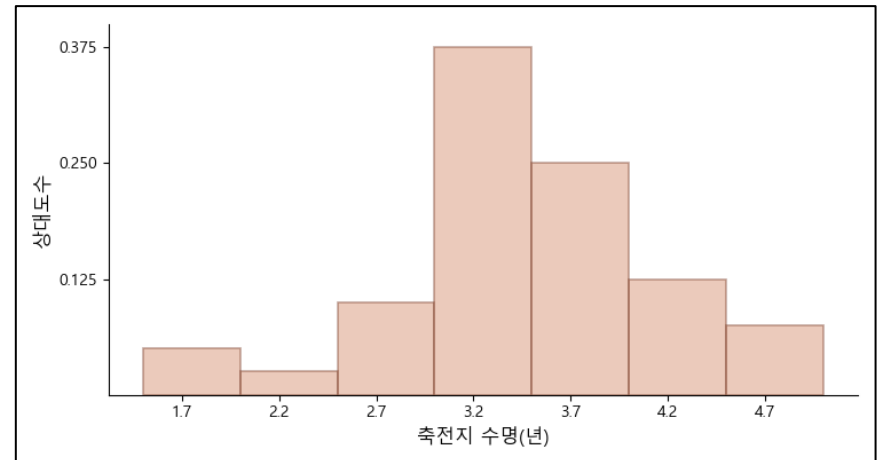
• 히스토그램 (2/2)

• 예제

40개의 자동차를 대상으로 축전지의 수명을 조사하였다.
해당 축전지의 수명은 적어도 3년은 지속된다는 것이 보증 되어있다.
각각의 관측값에 대해 상대도수분포표를 보고 상대도수 히스토그램으로 표현하여라.

계급구간	중간점	도수	상대도수
1.5-1.9	1.7	2	0.050
2.0-2.4	2.2	1	0.025
2.5-2.9	2.7	4	0.100
3.0-3.4	3.2	15	0.375
3.5-3.9	3.7	10	0.250
4.0-4.4	4.2	5	0.125
4.5-4.9	4.7	3	0.075

▲ 축전기 수명의 상대도수분포표



▲ 축전기 수명의 상대도수히스토그램

통계적 모형화, 과학적 조사, 그래프 진단

- 연속형 분포 (1/2)

- 특징

- 표본의 크기가 커질수록 분포의 형태가 매끄럽게 나타남

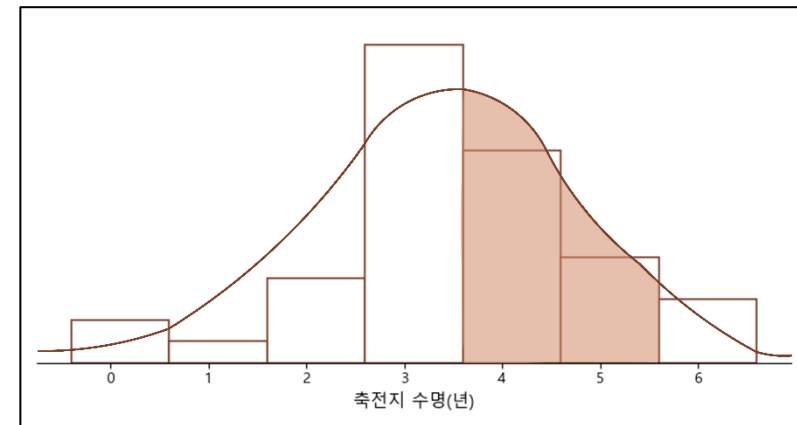
- 확률밀도함수

- 정의

- 연속확률변수의 분포를 나타내는 함수
 - 관측값이 가질 가능성의 상대적 크기를 함수 형태로 표현

- 히스토그램과의 관계

- 히스토그램에서 표현한 구간별 상대도수를 연속적인 곡선으로 연결



▲ 확률밀도함수의 추정

통계적 모형화, 과학적 조사, 그래프 진단

- 연속형 분포 (2/2)

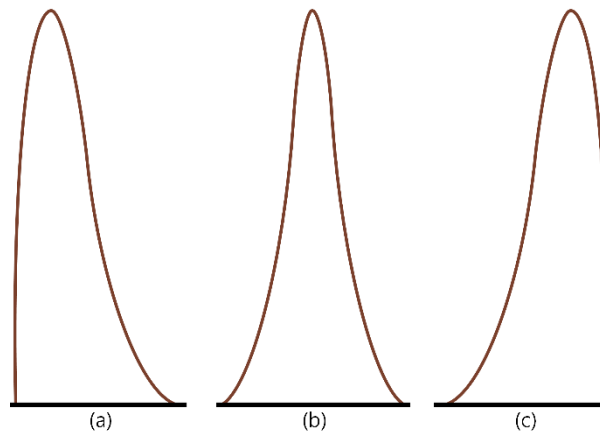
- 분포의 대칭성(Symmetry)

- 대칭(Symmetric)

- 어떤 분포가 하나의 수직선을 따라 좌우의 두 편이 일치하는 경우
 - e.g., 하단 그래프의 (b)

- 비대칭(Skewed)

- 분포의 대칭성이 결여되어 한쪽으로 기울어지는 경우
 - e.g., 하단 그래프의 (a), (c)



▲ 자료의 기움(왜도)

통계적 모형화, 과학적 조사, 그래프 진단

- 상자-수염 그림(Box-and-Whisker Plot) (1/4)
 - 사분위수(Quartiles)와 백분위수(Percentiles)
 - 정의
 - 자료를 일정 비율 기준으로 나누는 위치 측도
 - 사분위수: 자료를 4등분
 - 백분위수: 자료를 100등분
 - 특징
 - 분포의 꼬리 특성 파악 가능
 - 활용
 - 상자-수염 그림 구성 및 해석
 - 상대적 위치 비교(상위/하위 비율)

통계적 모형화, 과학적 조사, 그래프 진단

- 상자-수염 그림 (2/4)

- 정의

- 사분위수를 기준으로 자료 분포를 요약하여 나타낸 그래프

- 특징

- 중앙값, 사분위범위, 극단값 표현
- 이상치 탐지에 효과적
- 분포의 비대칭성 확인 가능

- 활용

- 여러 집단 간 분포 비교
- 자료의 중심과 산포 요약

통계적 모형화, 과학적 조사, 그래프 진단

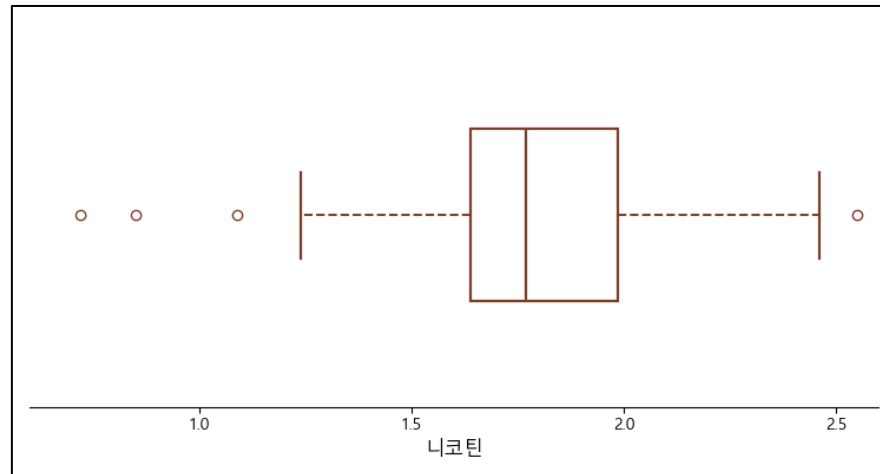
• 상자-수염 그림 (3/4)

• 예제 1.5

40개비의 담배로 구성된 확률표본에 대하여 니코틴 함량(단위: gm)이 측정되었다. 해당 자료를 보고 상자-수염 그림과 줄기-잎 그림으로 표현하여라.

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 측정 자료



▲ 니코틴 자료에 대한 상자-수염 그림

7		2
8		5
9		
10		9
11		
12		4
13		7
14		07
15		18
16		3447899
17		045599
18		2568
19		0237
20		389
21		17
22		8
23		17
24		6
25		5

▲ 니코틴 자료에 대한 줄기-잎 그림

통계적 모형화, 과학적 조사, 그래프 진단

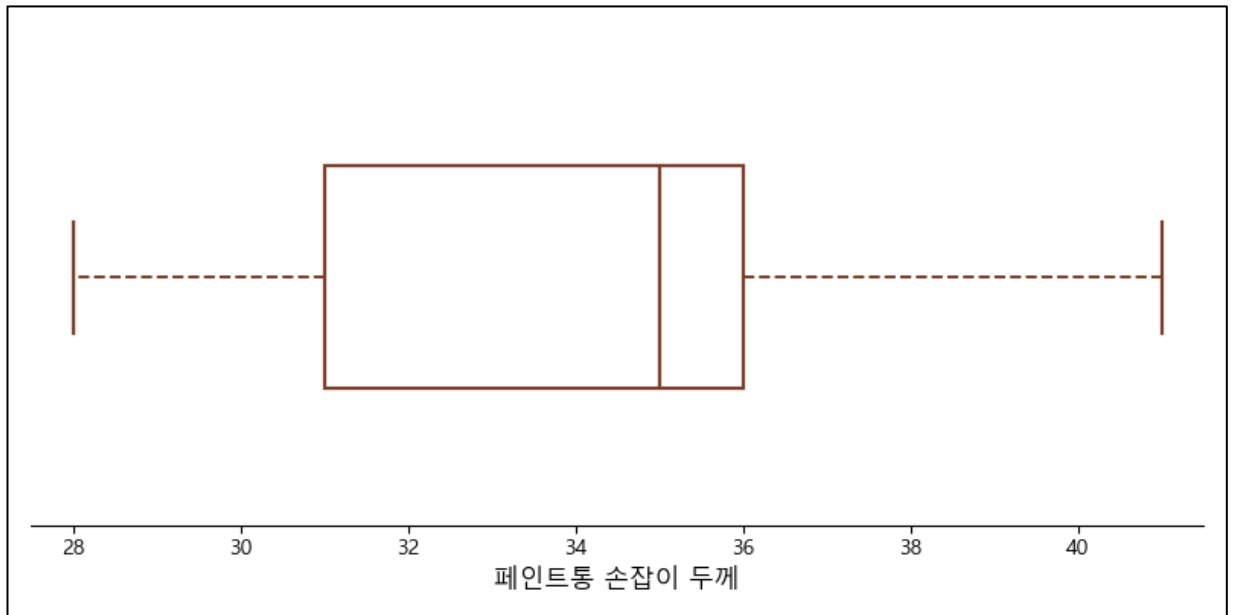
• 상자-수염 그림 (4/4)

• 예제 1.6

30개의 페인트통 손잡이의 두께를 측정하였다.
해당 자료를 보고 상자-수염 그림으로 표현하여라.

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 35 38 35	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

▲ 손잡이 두께 측정 자료



▲ 손잡이 두께에 대한 상자-수염 그림

목 차

- 개요
- 표본추출
- 위치의 측도
- 산포의 측도
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형

통계학 연구의 일반적인 유형

- 연구 설계와 통계적 해석 (1/2)
- 통계적 분석 결과의 의미는 연구 설계 유형에 따라 달라짐
 - 실험계획, 관측연구, 후향연구 각각 해석 가능 범위가 다름
 - 실험계획의 특성
 - 강점
 - 요인을 의도적으로 통제하고 실험 수행
 - 요인을 통제하지 않으면 결과 변화의 원인 불분명
 - 인과관계에 대한 가장 강한 근거 제공
 - 교호작용(Interaction)
 - 한 요인의 효과가 다른 요인의 수준에 따라 달라지는 현상
 - 단순 비교가 아닌 조건 조합에 따른 결과 차이 설명 가능

통계학 연구의 일반적인 유형

- 연구 설계와 통계적 해석 (2/2)
 - 통계적 분석 결과의 의미는 연구 설계 유형에 따라 달라짐
 - 실험계획, 관측연구, 후향연구 각각 해석 가능 범위가 다름
 - 관측연구의 특성
 - 한계
 - 현실적으로 불가피한 경우 다수
 - 요인 통제 및 무작위화 불가능
 - 관계를 파악할 수 있으나 인과관계로 단정짓기 어려움
 - 후향연구의 특성
 - 장점
 - 기존 자료를 활용하여 비용과 시간 절약
 - 한계
 - 통계적 결론의 신뢰도 가장 낮음
 - 자료의 타당성 및 신뢰성 문제 발생
 - 연구 목적과 불일치 가능성 존재

Thanks!

송 민 희(minhee@pel.sejong.ac.kr)