2017/07/13, 2017 확률 세미나

확률 및 통계학

- 통계학과 자료분석 -

송 영 준(youngjun@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

목 차

- 개요
- 표본추출
- 위치 측도 와 산포 측도
- 이산형 자료와 연속형 자료
- 통계적 모형화, 과학적 조사, 그래프 진단
- 통계학 연구의 일반적인 유형: 실험계획, 관측연구, 후향연구

개요

- 통계학
  - 사회 현상을 통계에 의하여 관찰-연구하는 학문
  - 관심의 대상에 대한 자료를 수집하여 정리, 요약하고 이들 자료에 포함된 정보를 토대로 불확실한 사실에 대해 과학적 판단을 내릴 수 있도록 그 방법을 제시해 주는 학문
  - 관측 자료를 바탕으로 추론(Inference)를 하는 과학의 한 분야로서 불확실성(Uncertainty)하에서 보다 합리적인 의사 결정을 하는 방법을 제시해 주는 학문

개요

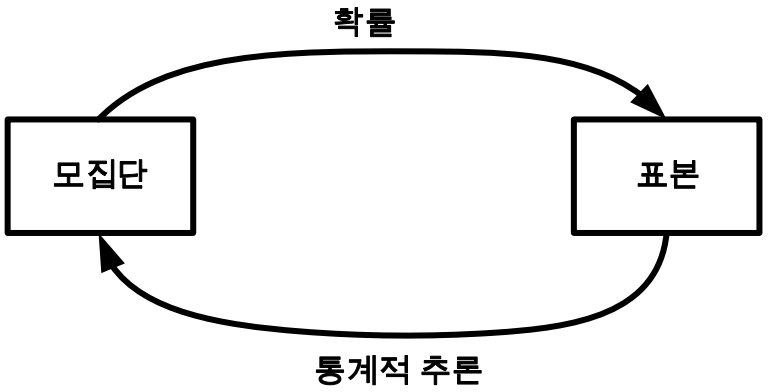
- 통계학
  - 분류
    - 기술 통계학
      - 수집된 자료를 정리 및 요약하는 방법을 다루는 통계학
      - 자료를 표,그래프 등으로 나타내고, 대표 값(평균,중간 값 등)과 산포도(분산, 표준 편차)로 자료의 전반적인 특성을 표현
    - 추측 통계학
      - 주어진 자료의 정보를 분석해서 미래에 일어날 상황을 예측하는 통계학
      - 통계적 추론을 통해 얻어진 추측이나 결론은 항상 옳은 것은 아님
      - 어느 정도의 불확실성을 가지고 있는데 이 불확실성의 정도를 확률로 표현
      - 확률론을 바탕으로 발전

개요

- 자료의 변동
  - 자료는 여러 요소(Factor)에 영향을 받으며 , 구체적으로 정의된 방법으로 표본을 추출하고 해석 하여야 함
    - 수집된 자료가 항상 같고, 목표 값과 동일하면 통계적 방법은 불필요
  - 실험 계획법(Experimental Design)
    - 관측대상에 영향을 주는 Factor를 측정자가 제어할 수 있는 경우
  - 관측 연구(Observation Study)
    - 관측대상에 영향을 주는 Factor가 많고 다양하여 예측할 수 없으므로, 관측자가 제어할 수 없는 경우

개요

- 확률의 역할
  - 추출한 표본이 모집단을 얼마나 대표하는가를 나타내는지를 계산
    - P-value 문제
      - 특정 가설을 전제로 가설이 성립하는지에 대한 근거로 P값을 제시
  - 확률과 추론의 관계



개요

- 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

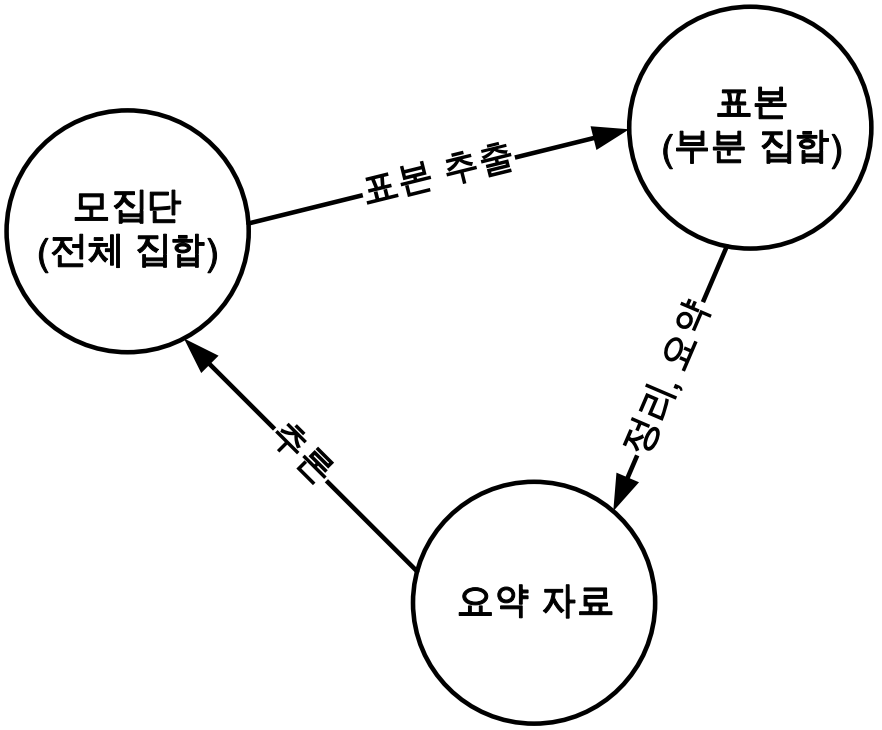
표본 추출(Sample Sampling)

- 정의
  - 모집단(Population)에서 표본을 추출하는 일
  - 표집(Sampling)이라고도 부름

- 용어
  - 모집단(Population): 어떤 정보를 얻고자 하는 전체 대상 또는 전체 집합을 의미
  - 표본(Sample): 모집단으로부터 뽑은 부분 집합, 조사 대상을 의미
  - 모수(Parameter): 모집단의 특성을 나타내는 양적인 값으로 고유한 상수로 나타냄

표본 추출(Sample Sampling)

- 목적
  - 연구의 주제가 되는 전체 모집단(Population)의 일부를 일정한 방법에 따라 추출하여 이들을 통해 얻은 정보를 바탕으로 모집단을 추정
    - 통계적 추론(Statistical Inference)



표본 추출(Sample Sampling)

- 종류
  - 단순랜덤표본추출(Simple Random Sampling)
    - 표본 추출법 중 가장 기본적인 방법
    - 특정 표본크기(Sample Size)내의 표본들이 선택될 확률이 동일
    - 제한된 모집단을 대표하는 표본을 추출하게 되는 문제점을 제거할 수 있음
      - 모집단에 대해 표본 추출 시 집단 분포가 고르지 않다면 편향표본
      - Biased Sample
  - 층화랜덤표본추출(Stratified Random Sampling)
    - 모집단이 중복되지 않도록 층으로 나눔
    - 각 층 내에서 표본을 랜덤하게 추출하여 특정 층의 강조를 없앰

표본 추출(Sample Sampling)

- 실험 계획법(Experiment Design)
  - 관측 대상에 영향을 주는 Factor를 측정자가 제어할 수 있는 경우
  - 임의성 (Randomness), 랜덤할당(Random Assignment)의 개념은 중요
  - 처리,처리조합(Treatment Combination)들이 연구 및 비교 대상 모집단이 됨
    - 금속피로 부식연구 : 도장 vs 비 도장, 저습도 노출 vs 고습도 노출

표본 추출(Sample Sampling)

- 실험 계획법(Experiment Design)
  - 처리조합에 대해 표본은 변동(Variation)이 존재
    - 변동은 실험단위(Experimental Unit)에서 발생
    - 실험 단위가 지나치게 균질 하지 않아서 변동이 너무 커지면, 변동으로 인해 두 모집단의 차이가 검출되지 않음
    - 변동이 편향되는 표본은 과학적 발견에 방해되거나 근거자료로써의 가치가 부족

위치 측도 와 산포 측도

- 위치 측도
  - 자료의 중심이 어디인가를 알 수 있는 계량적인 측도

  - 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)
    - 극단 값 또는 특이점(Outline)의 영향을 덜 받는 자료의 중심 측도
    - $\tilde{x} = \begin{cases} x_{(n+1)/2}, n = \text{홀수} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), n = \text{짝수} \end{cases}$

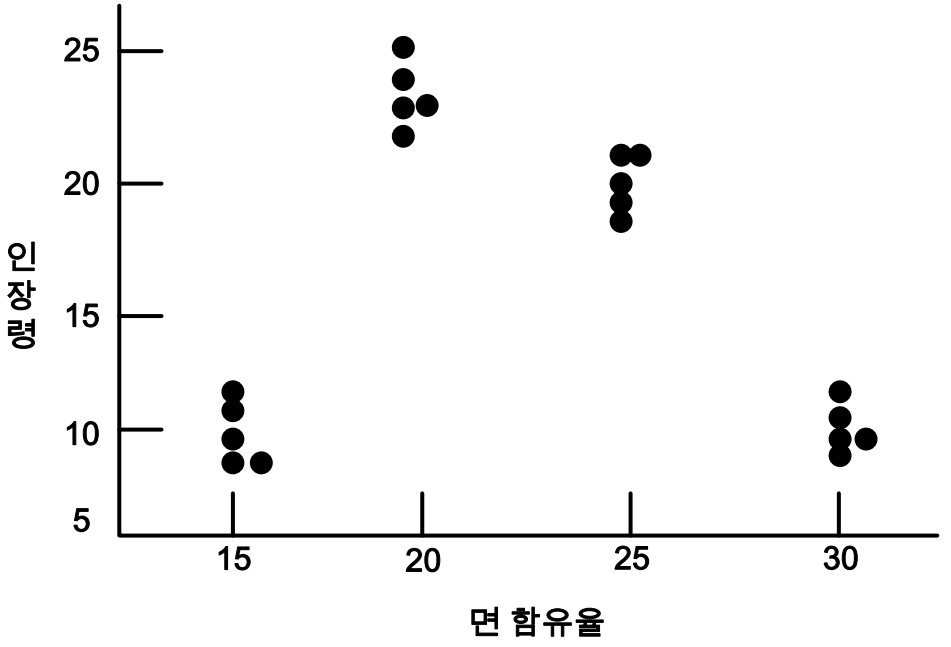
위치 측도 와 산포 측도

- 위치측도
  - 절사 평균(Trimmed Mean)
    - 자료의 가장 크거나 작은 일부분을 제외한 평균
    - 극단 값 또는 특이점에 대해 영향을 덜 받음
  - 극단 값에 의한 측도의 민감도
    - 중앙값 > 절사 평균 > 평균

통계적 모형화, 과학적 조사, 그래프 진단

- 산점도
  - 변수 간에 있어서 관계를 규명하기 위해서 변수가 다른 변수에 끼치는 영향을 나타내는 기법

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



통계적 모형화, 과학적 조사, 그래프 진단

- 줄기-잎 그림
  - 통계적 자료를 표 형태와 그래프가 혼합된 방법
  - Stem and leaf plot
  - 줄기와 잎 표현으로 자료를 간략히 표현
  - 비교적 적은 자료에 대해 간략히 표현

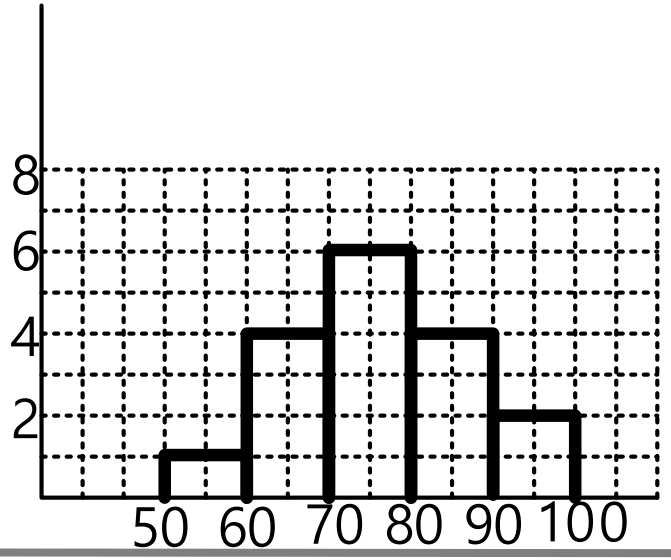
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

줄기	잎	도수
1	69	2
2	25696	5
3	43185147236282913009714 5	25
4	71354172	8

통계적 모형화, 과학적 조사, 그래프 진단

- 히스토그램
  - 표로 되어있는 도수 분포를 막대그래프로 나타낸 것
    - 도수분포표를 그래프로 나타낸 것을 의미
      - 도수분포 표: 자료를 일정한 수의 범위로 나누어 분류하고, 각 범위별로 수량을 정리한 표
      - 변량 : 점수, 시간과 같은 여러 자료를 수량으로 나타낸 것(바뀌는 값)
      - 도수 : 각 계급에 속하는 변량의 개수
      - 계급 : 변량을 일정한 구간으로 나눈 것

점수(점)	학생 수(명)
50~60	1
60~70	4
70~80	6
80~90	4
90~100	2
합계	17명

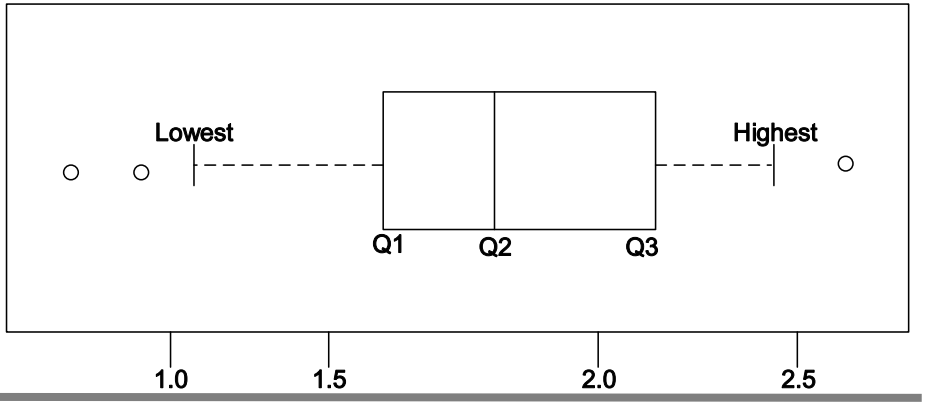


어떤 학급의 학생들의 수학 점수 : 92점, 84점, 88점, 87점, 81점, 76점, 96점, 71점, 72점, 79점, 75점, 73점, 60점, 68점, 66점, 63점, 52점

통계적 모형화, 과학적 조사, 그래프 진단

- 상자-수염 그림(그림상자)
  - 최고값, 최저값, 자료의 사분위수 Q1,Q2,Q3를 이용하여 그래프로 표현(5가지 요약 수치)
  - 집단이 여러 개인 경우에 한 공간에 표현하여 비교
    - 사분위수 : 데이터의 일부를 거의 균등한 관찰값 수를 포함하는 3개의 그룹으로 나눈 값, 합계 100%를 3개의 균등한 부분으로 분할
      - 20%, 50%, 75%, 100%
      - 사분위수 범위(IQR) : Q3-Q1

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



통계학 연구의 일반적인 유형

- 통계학 연구 및 활용
  - 통계적 방법에 의해 알아낸 정보들은 많은 과학적, 공학적 분야의 의사결정과 문제해결에 영향력을 발휘
  - 모든 통계적 실험에 대해 적절한 통계방법을 적용해야 함

  - 교호작용(Interaction)에 의하여 표본수집과 통계적 해석은 여러 가지 요인(Factor)에 의해 상호 변동이 주어짐
    - 단일요인, 통제 불가능한 요인
      - 통제 불가능한 요인의 연구
        - 관측 연구

감사합니다!