확률 및 통계학

- 1장 통계학과 자료 분석-

명 세인(sein@pel.smuc.ac.kr)

상명대학교 프로토콜공학연구실

목 차

- 확률 및 통계학
- 표본추출
- 측도
- 논의

확률 및 통계학

- 개요
  - 경영자들의 통계적 방법을 통한 품질개선으로 사용됨
  - 예측하기위해 자료를 수집하여 수학적근거를 이용한 추정

  - 추론통계학(Inferential Statistics)
    - 단순히 수집한 자료를 보여주는것이아닌 자료가 나타내는 값을 해석하여 의미를 보여줄 수 있는것
    - **표본을통해 모집단을 추측**

확률 및 통계학

- 자료의 사용
  - 통계적 방법을 사용하기위해 자료 또는 정보(Scientific Data)의 수집이 필요
  - 수집된자료는 일반적으로 불확실성(Uncertainty)을 가지며, 변동에 대응하는 과학적 판단을할 수 있도록 추론통계학이 발전
    - 실제값, 즉 참(True)값과 **관측값의 차이가 의미가 있거나 없음을 따질** 수 있어야 함
    - 관측시의 오류가 없는지를 의미

확률 및 통계학

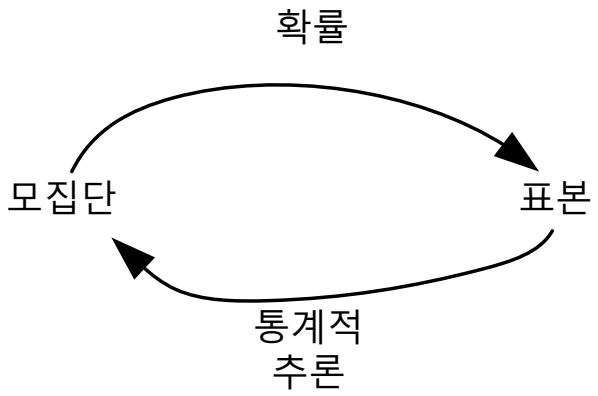
- 자료의 변동
  - 자료는 여러 요소(Factor)에 영향을 받으며, 구제적으로 정의된 방법으로 표본을 추출하고 해석 하여야 함
    - 수집된 자료가 항상 같고, 목표값과 동일하면 통계적 방법은 불필요
      - 불확실성(Uncertainty)에 의해 불가능
  - 실험계획법(Experimental Design)
    - 관측대상에 영향을 주는 Factor를 측정자가 제어할 수 있는 경우
  - 관측연구(Observation Study)
    - 관측대상에 영향을 주는 Factor가 많고 다양하여 예측할 수 없으므로, 관측자가 제어할 수 없는 경우

확률 및 통계학

- 자료의 변동
  - 기술통계학(Descriptive Statistics)
    - 수집된 자료를 요약적으로 표현하는 방법으로 평균(Mean), 중앙값(Median), 표준편차(Standard Deviation)등으로 표현
    - 히스토그램(Histogram), 줄기-잎 그림(Stem & Leaf Plot), 점그림(Dot Plot), 상자 그림(Box Plot)등의 **그래프로 요약하여 표현**
  - 통계적 추론(Statistical Inference)
    - 수집된 자료에서 알려지지 않은 대상(모집단)을 통계학적 근거를 이용하여 추측

확률 및 통계학

- 확률의 역할
  - 추출한 표본이 모집단을 얼마나 대표하는가
    - P-value문제
      - 특정 가설을 전제로 가설이 성립하는지에 대한 근거로 p값을 제시
  - 확률과 추론의 관계



확률 및 통계학

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

목 차

- 확률 및 통계학
- 표본추출
- 측도
- 논의

표본 추출

- 표본추출(Sampling)
  - 정의된 집합(모집단)에서 조건을 만족하는 일부 개체를 선정
  - 표본을 구성하고 추론하여 모집단을 추정

- 단순랜덤표본추출(Simple Random Sampling)
  - 특정 표본크기(Sample Size)내의 표본들이 선택될 확률이 동일
    - 모집단에 대해 표본추출시 집단분포가 고르지 않다면 편향표본(Biased Sample)

표본 추출

- 층화랜덤표본추출(Stratified Random Sampling)
  - 표본추출 단위(Unit)들이 균질하지 않고 자연스럽게 균질한 군(Group)들로 겹침 없이 나뉘는 경우
    - 각 층 내에서 표본을 랜덤하게 추출하여 특정층의 강조를 없앰
    - 시골과 도시에 대한 정치적 성향 조사
      - 시골과 도시에대해 표본추출의 균일함에대한 오류
      - 인구 비율, 시골과 도시의 크기등 전체에대해 일반적인 단순랜덤표본추출은 편향표본이 될 가능성

표본 추출

- 실험계획법(Experimental Design)
  - 임의성(Randomness), 랜덤할당(Random Assignment)개념은 실험계획법분야에서 중요한 역할
  - 처리, 처리조합(Treatment Combination)들이 연구 및 비교 대상 모집단이 됨
    - 투약 효과, 위약 <-> 투약
    - 금속피 부식, 도장 <-> 비도장 / 저습 <-> 다습
  - 각 처리조합에 대해 표본은 변동(Variation)이 발생
    - 각 변동은 실험단위에서 발생(Experimental unit)
    - 실험단위가 균일하지 않아 변동이 커지면 두 집단의 차이가 검출되지 못하고 감춰짐

표본 추출

- 실험단위의 임의성
  - 완전확률설계(Completely Randomized Design)
    - 표본추출시 편향되지 않는 추출을 구성
  - 변동이 편향되는 표본은 과학적 발견에 방해되거나 근거자료로써의 가치가 부족

목 차

- 확률 및 통계학
- 표본추출
- 측도
- 논의

측도

- 위치측도
  - 자료의 중심에 대한 계량 측도
    - 표본평균(Sample Mean)
      - 산술평균(Numerical Average)이며 자료의 무게중심
        - $\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$
    - 표본 중앙값(Sample Median)
      - 극단값 또는 특이점(Outline)의 영향을 덜받는 자료의 중심 측도
        - $\overline{x} = \begin{cases} x_{(n+1)/2}, & n = \text{홀수} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & n = \text{짝수} \end{cases}$

측도

- 위치측도
  - 자료의 중심에 대한 계량 측도
    - 절사평균(Trimmed Mean)
      - 자료의 가장 크거나 작은 일부분을 제외한 평균
      - 극단값 또는 특이점에대해 영향으 덜 받음
  - 극단값에 의한 측도의 민감도
    - 중앙값 > 절사평균 > 평균

측도

- 산포 측도
  - 표본의 산포(Variability)
    - 자료의 분포도, 중심으로 부터 분포도등을 의미
  - 표본범위(Sample Range)
    - 가장 단순한 산포도, 표본들의 범위
    - $X_{max} - X_{min}$

측도

- 산포 측도
  - 표본의 산포(Variability)
    - 표본분산
      - 평균을 기준으로 자료의 분포도
      - 평균과의 차는 0이므로 제곱으로 분산을 구함
      - $s^2 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$
    - 표본표준편차(Sample Standard Deviation)
      - 표본분산의 양의 제곱근, 제곱단위인 표본분산의 단의를 제곱근을통해 표본단위와 맞추어줌
      - 사실살 변동의 측도가 됨
      - $s = \sqrt{s^2}$

목 차

- 확률 및 통계학
- 표본추출
- 측도
- 논의

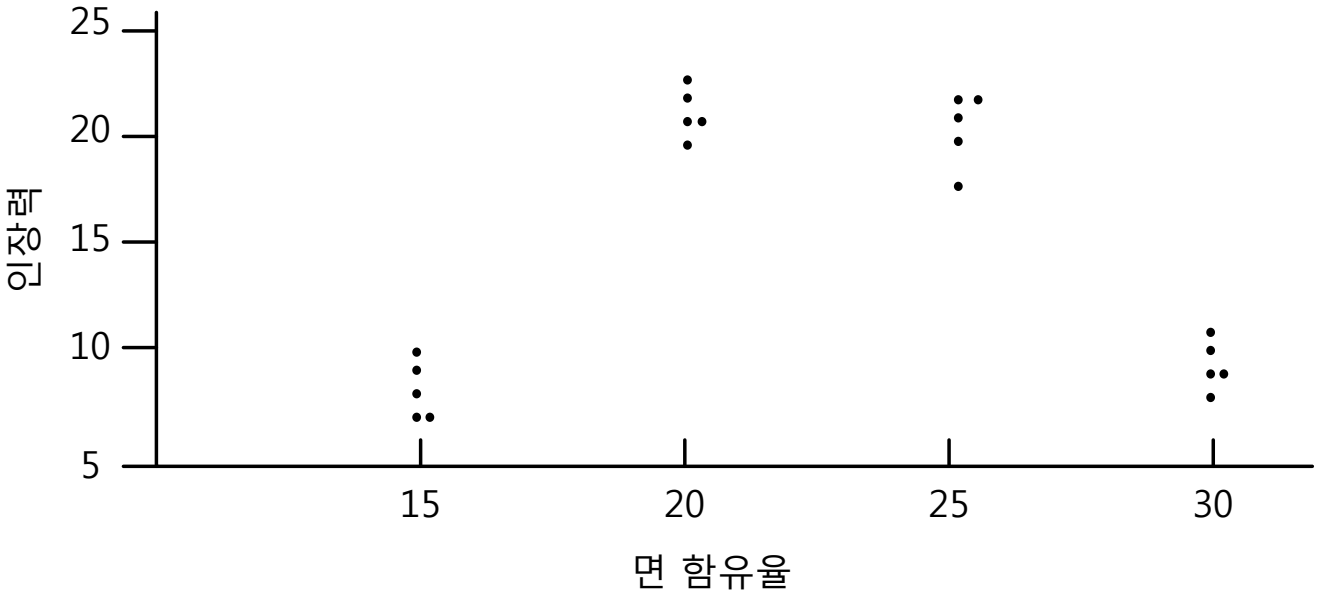
논의

- 이산형 자료와 연속형 자료
  - 이산형(Discrete)
    - 연속적이지 않는 계수자료(Count Data)
      - 반응 횟수, 결함이 있는 기계의 개수 등
  - 연속형(Continuous)
    - 측정되는 값이 연속적이기때문에 정확히 측정할 수 없는경우 범위를 지정
      - 사람의 키, 무체의 질량 등

논의

- 자료의 표현
  - 산점도
    - 두종류 이상의 자료들의 관계를 볼 수 있음
      - 양/음/원형 분포 등
      - 강/약의 선형 관계 등

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



논의

- 자료의 표현
  - 줄기-잎 그림(Stem & Leaf Plots)
    - 줄기와 잎 표현으로 자료를 간략히 표현
    - 비교적 적은 자료에대해 간략히 표현됨

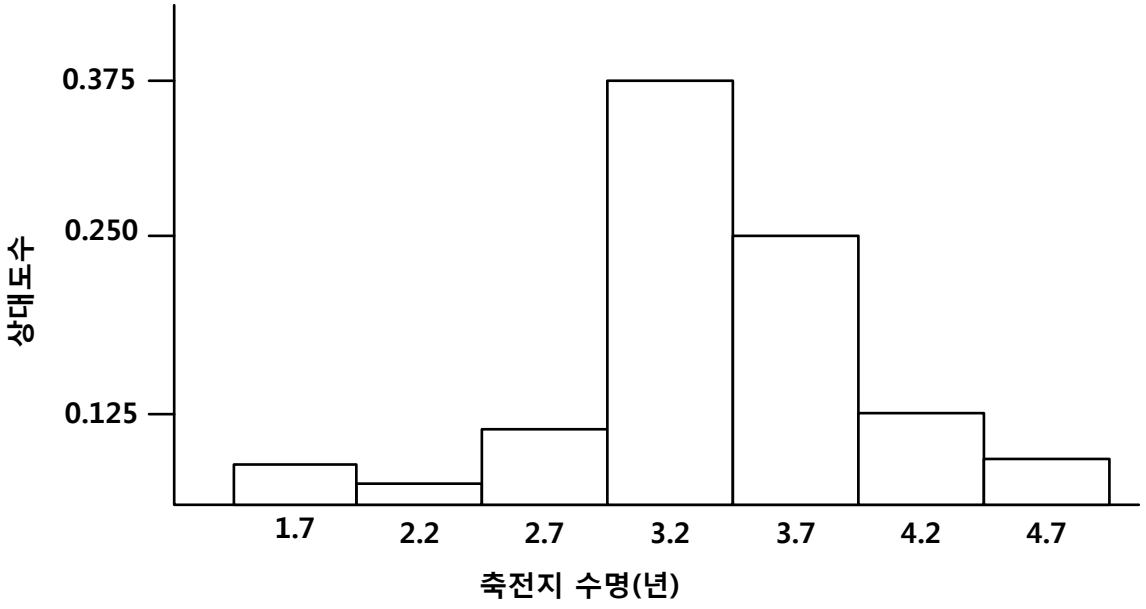
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

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

논의

- 자료의 표현
  - 히스토그램(Histogram)
    - 측정된 도수의 분포도를 표현

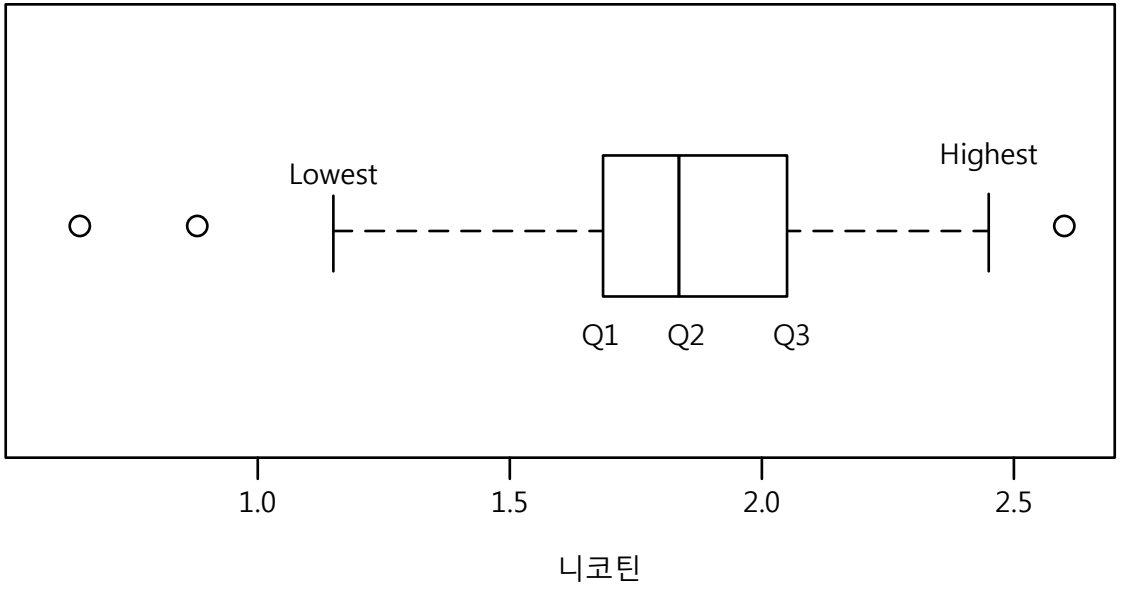
계급구간	중간점	도수	상대도수
1.5-1.9	1.7	2	0.050
2.0-2.4	2.2	1	0.025
2.5-2.9	2.7	4	0.100
3.0-3.4	3.2	15	0.375
3.5-3.9	3.7	10	0.250
4.0-4.4	4.2	5	0.125
4.5-4.9	4.7	5	0.075



논의

- 자료의 표현
  - 상자-수염 그림(Box & Whisker Plots)
    - 최고값, 최저값, 자료의 사분위수 Q1, Q2, Q3를 이용하여 그래프로 표현
    - 최저극단값 2개와 최고 극단값 하나를 제외하고 그래프로 표현

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



논의

- 통계학의 연구 및 활용
  - 모든 통계적 실험에 대해 적절한 통계방법을 적용하여야 함
  - 교호작용(Interaction)에 의하여 표본수집과 통계적 해석은 여러가지 요인에 의해 상호 변동이 주어짐
  - 실험계획
    - 표분추출의 요인을 관측자가 통제할 수 있는경우
  - 관측연구
    - 모든요인을 통제할 수 없어 특정기간동안 관측된 표본을 추출하는 경우
  - 후향연구
    - 역사적으로 수집된 데이터를 활용하는 경우

감사합니다!