Big Data

데이터 분석 기획

류영표 강사
youngpyoryu@dongguk.edu

강사 소개

류영표

Youngpyo Ryu

現 동국대학교 수학과/응용수학 석사수료

現 SD아카데미 국비과정 강사

現 Upstage AI X 네이버 부스트캠프 멘토

前 메가 IT아카데미(파이썬,빅데이터) 강사

한국파스퇴르연구소 Image Mining 인턴(Deep learing)

前 ㈜셈웨어(수학 컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- 딥러닝 집중 교육과정 강사
- (재)윌튼블록체인 6일 과정 (파이썬기초, 크롤링,머신러닝)
- 서울특별시 X AI 양재허브 X 모두의연구소 (중급 NLP과정) 보조강사
- SK아카데미_HLP(임원) 1차/2차 보조강사
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- LG전자 / LG 인화원 보조강사
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의

주요 프로젝트 및 기타사항

- 제1회 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 최우수상 수상(Q&A 챗봇)
- 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

통계학

- 통계학 이란?

  - 자료로부터 유용한 정보를 이끌어 내는 학문(자료의 수집, 정리, 해석하는 방법 등을 포함)

  - ex) 일기예보, 경제통계, 사회조사 분석통계, 실험결과 분석통계 등 다양한 형태

- 통계 분석이란?

  - 특정한 집단이나 불확실한 현상을 대상으로 자료를 수집 -〉 대상 집단에 대한 정보를 구함 -〉 적절한 통계 분석 방법을 이용한 의사결정(통계적 추론) 과정을 말함.

  - 통계적 추론에는 대상 집단의 특정값을 추측하는 추정 / 가설 설정 후 채택 여부를 결정하는 가설검정 / 미래에 대한 예측이 있다.

모집단

- 알고자 하는 대상 / 모집단을 구성하는 개체를 추출단위 혹은 원소라고 함.
- 모집단에 대해 조사하는 방법

  1 .총조사 : 모든 개체를 조사하는 방법, 많은 비용과 시간이 소요되므로 특별한 경우를 제외하고 실시 하지 않음.

  2. 표본조사

  – 모집단의 일부분(표본)만 조사하여 모집단에 대해 추론

  – 표본추출 방법에 따라 분석결과의 해석이 큰 차이가 발생할 수 있으므로, 모집단의 정의/표본의 크기/조사방법/조사기간/표본 추출 방법을 명확하게 밝히거나 확인해야 한다.

표본추출 방법


표본추출방법
비확률표본추출방법
편의표본추출법
판단표본추출법
할당표본추출법
확률표본추출방법
단순무작위 표본추출법
층화표본추출법
군집표본추출법

표본추출 방법

1) 비확률표본추출방법 (Non-Probability Sampling)

→ 모집단 내의 각 구성요소가 표본으로 선택될 확률을 알 수가 없음

- 편의추출법 (Convenience Sampling): 임의로 선정한 지역과 시간대에 조사자가 원하는 사람들을 표본으로 선택하는 방법

  → 표본추출비용이 거의 들지 않고 절차가 간단, 대표성 부족

- 판단표본추출 (Purposive Sampling): 조사문제를 잘 알고 있거나 모집단의 의견을 반영할 수 있을 것으로 판단되는 특정집단을 표본으로 선택하는 방법

  → 대표성 부족, 사전조사에 활용

- 할당표본추출방법 (Quota Sampling): 미리 정해진 분류기준에 의해 전체표본을 여러 집단으로 구분하고 각 집단별로 필요한 대상을 추출하는 방법

  → 성별, 연령, 지역이 할당 기준으로 많이 활용되어짐.

  → 모집단의 특성이 반영, 상업적 마케팅조사에 가장 널리 활용됨

표본추출 방법

2) 확률표본추출방법

→ 특정 조사대상이 뽑힐 확률, 발생될 오류의 정도에 대한 추정이 가능한 방법

→ 표본프레임 확보가능, 모집단의 수가 적은 경우 효과적

- 단순무작위 표본추출방법 (Random Sampling): 표본프레임내의 각 표본들에 대해 일련번호를 부여하고, 이를 이용해 일정수의 표본을 무작위(random)로 추출하는 방법으로 확률표본추출방법 중 가장 기본적인 방법

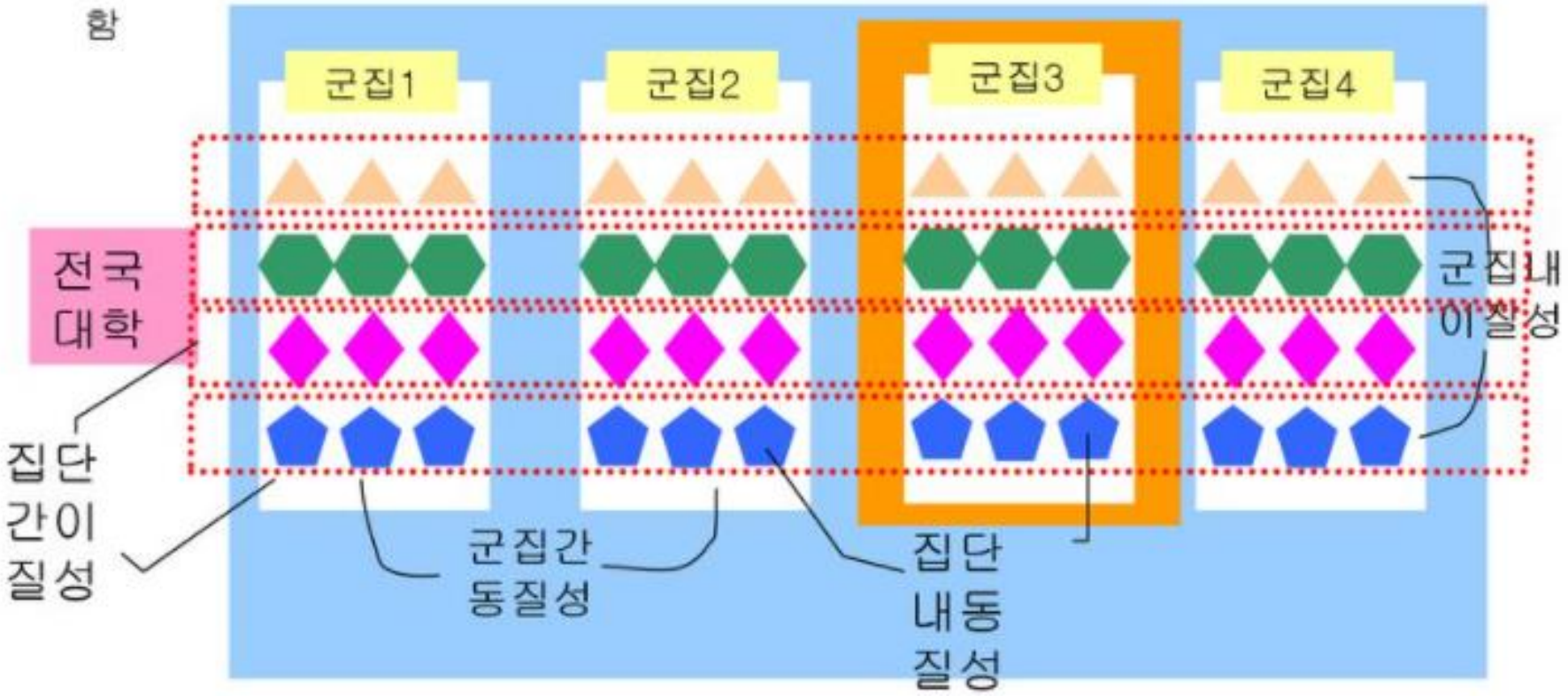
  ex) 복권 추첨

- 층화표본추출법(Stratified Sampling): 모집단을 어떤 기준에 따라 서로 상이한 소집단들로 나누고 이들 각 소집단들로부터 빈도에 따라 적절한 수의 표본을 무작위로 추출하는 방법 ex) 대학생연구에서 학년별로 몇 명씩 표본을 선정하는 방법

표본추출 방법

- 군집표본추출법(Cluster Sampling): 모집단을 소집단(군집)들로 나누고 일정수의 소집단을 무작위적으로 표본 추출한 다음, 추출된 소집단내의 구성원들을 모두 조사하는 방법

ex) 모집단을 총 4개의 군집으로 나누고 이중 1개의 군집을 선택하되 그 군집내의 모든 구성원 표본함



자료의 종류 및 확률

- 자료의 종류

  1) 질적자료 : 대상에 속하는 집단을 분류하는 명목척도 / 서열관계나 선호도를 관측하는 순서척도

  2) 양적자료 : 온도, 지수 등 속성의 양을 측정하는 구간 척도 / 무게, 나이 등 숫자로 관측되는 일반적인 비율

- 확률

  – 확률 : 특정 사건이 일어날 가능성의 척도

  – 표본공간 : 나타낼 수 있는 모든 결과들의 집합

  – 근원사건 : 한 개의 원소로만 이루어진 사건

  – 배반사건 : 교집합이 공집합인 사건들

  – 조건부 확률 : 특정 사건 A가 일어났다는 가정하의 사건 B의 확률

  – 독립 : 사건 A가 일어났는지 여부와 상관없이 사건 B의 확률이 동일하면 서로 독립이라고 함.

확률변수와 확률분포

- 확률 변수 : 정의역이 표본공간이고 치역이 실수값인 함수

- 1) 이산형 확률변수

  – 사건의 확률을 각 이산점에서의 확률의 크기로 나타내는 확률질량함수로 표현

  – ex) 베르누이 확률분포, 이항분포, 기하분포, 다항분포, 포아송 분포 등

- 2) 연속형 확률변수

  – 사건의 확률을 함수의 면적으로 표현하는 확률밀도함수로 표현

  – 면적으로 표현되므로 한점에서의 확률은 0이다.

  Ex) 균일분포, 정규분포, 지수분포 등

추정

- 통계적 추론은 추정과 가설검정으로 나눌 수 있는데, 그 중에서도 추정은 점추정과 구간추정으로 나뉜다.
- 1) 점추정
  - 가장 참값이라고 여겨지는 하나의 모수의 값을 택하는 것(모수는 특정한 값일 것이라고 추정)
  - 사실상 추정이 얼마나 정확한가를 판단하기가 불가능
  - 대표적인 예로 모평분산과 모 분산을 추정하기 위한 추정량인 표본평균과 표본분산이 있다.
- 2) 구간 추정
  - 점 추정의 정확성을 보완하는 방법
  - 일정한 크기의 신뢰수준(90%, 95%, 99% 등)으로 모수가 특정한 구간(신뢰구간)에 있을 것이라고 선언하는 것.

점추정

- 불편성(Unbiasedness): 모든 가능한 표본에서 얻은 추정량의 기댓값은 모집단의 모수와 편의(차이가 없다)
- 효율성(efficiency) : 추정량의 분산이 작을수록 좋다.
- 일치성(consistency) : 표본의 크기가 아주 커지면, 추정량이 모수가 거의 같아진다.
- 충족성(sufficient): 추정량은 모수에 대하여 모든 정보를 제공한다.
- 표본평균(sample mean) : 모집단의 평균(모평균)을 추정하기 위한 추정량, 확률표본의 평균값
- 표본분산(sample variance) : 모집단의 분산(모집단)을 추정하기 위한 추정량

가설 검정

- 모집단에 대한 귀무가설(H0)과 대립가설(H1)을 설정한 뒤, 표본관찰 또는 실험을 통해 하나를 선택하는 과정

  1) 귀무가설(H0) : 대립가설과 반대의 증거를 찾기 위해 정한 가설

  2) 대립가설(H1) : 증명하고 싶은 가설

- 귀무가설이 옳다는 전제하에서 관측된 검정통계량의 값보다 더 대립가설을 지지하는 값이 나타날 확률을 구하여 가설의 채택여부를 결정한다.