[Data Analysis] 통계의 이해

BY 호쌤(이광호) ON OCTOBER 30, 2022

R, Python 등을 기반으로 데이터 분석 수업을 진행할 때 첫 시간에 소개하던 데이터 분석 기법 개요에 관한 요약글 입니다. 이 글에서 통계학의 정의와 역할에 대하여 알아보고, 데이터의 기본요소를 살펴봅니다. 통계학에서 가장 중심이 되는 개념인 모집단과 표본, 모수와 통계량을 소개합니다.

학습목표

- 통계학의 정의와 역할을 설명할 수 있다.
- 데이터에서 변수와 단위를 찾아낼 수 있다.
- 모집단, 표본, 모수, 통계량의 개념을 설명할 수 있다.
- 데이터 분석 기법의 종류를 구분하고 설명할 수 있다.

#01. 통계학이란

1. 데이터

> *세상을 이해하는 창*

어떤 현상을 이해하기 위해 그 현상을 관찰하여 데이터를 수집

전통적인 데이터 수집 방법 -> 관찰, 설문조사, 실험등

2. 데이터 폭발(Data explosion)

컴퓨터와 정보통신 기술의 발달로 매일 방대한 양의 데이터가 생산됨

> *뉴욕타임즈가 하루에 싣는 정보의 양은 17세기 영국의 평범한 한 사람이 평생 소비 하는 정보의 양과 비슷하다 (Wurman, S.A. (1987) "Information Anxiety" New York: Doubleday , p.32)*

> *페이스북에서는 하루에 4페타 바이트의 정보가 생성된다*
> *(https://kinsta.com/blog/facebook-statistics/, Jan 3, 2021)*

1페타바이트 = 1024테라바이트

3. 통계학

불확실한 현상을 이해하기 위해 데이터를 **수집**하고, 데이터 패턴을 **요약**, **분석**하여 불확실한 현상에 대한 결론을 찾는 학문

> *데이터에서 쓸모 있는 정보를 얻기 위한 별도의 과정*

- 특정집단을 대상으로 수행한 조사나 실험을 통해 나온 결과에 대한 요약된 형태의 표현
- 일기예보, 물가/ 실업률, 정당 지지도，의식조사와 사회조사 분석 통계，임상실험 등의 실험 결과 분석 통계
- 조사 또는 실험을 통해 데이터를 확보
- 조사대상에 따라 총조사 (census)와 표본조사로 구분

4. 통계학의 역할

통계학의 역할에는 데이터 수집, 데이터 요약, 데이터 추론이 있다.

1) 데이터 수집

알고 싶은 현상을 왜곡되지 않게, 잘 반영하는 데이터를 수집하기 위해 통계적 원리를 사용

총조사/전수조사(cencus)

- 대상 집단 모두를 조사하는데 많은 비용과 시간이 소요되므로 특별한 경우를 제외하고는 사용 되지 않는다.

예시: 선거 여론조사

> *대통령 선거를 앞두고 유권자의 지지성향을 조사하여 선거전략을 세우고자 한다. 전체 유권자의 연령별, 성별 분포를 고려하여 전체를 대표할 수 있는 일부 유권자를 뽑아 조사한다.*

표본조사

모집단 내에서 그 집단의 특성을 잘 나타낼 수 있는 일부를 추출하여 이들로부터 자료를 수집하고 수집된 자료를 토대로 전체의 특성을 추정 (대부분의 설문조사)

예시: 임상시험

> *특정 감염병 예방을 위해 개발된 백신의 효과를 알아 보기 위해, 3만명의 자원자를 모집한 후 랜덤으로 두 그룹으로나누고, 한그룹은백신, 다른 그룹은 플라시보를 투여한다. 3개월 동안 추적 관찰하여 백신의 효과를 증명할 수 있는 데이터를 얻는다.*

2) 데이터 요약

데이터가 가진 특징과 패턴을 정확하고 효과적으로 드러내기 위한 통계적 방법을 사용: 기술통계

예시: 소아의 몸무게

> *소아의 몸무게를 조사하여 나이별로 몸무게의 평균, 중간값, 사분위수 등 요약통계량을 구한다. 나이에 따른 몸무게의 변화를 보여주기 위해 그래프를 작성한다.*

예제: 미세먼지

> *지역별 미세먼지 농도를 수집하여 지도 위에 미세먼지 농도를 색깔로 표현한다.*

3) 추론

데이터를 이용하여 우리의 관심 대상에 대해 추측하고 그 추측의 신뢰성을 계량화: 추측통계(추론통계)

예시: 평균연봉

> *대한민국 임금노동자의 평균 연봉을 알아내기 위해서 랜덤 표집한 300명의 연봉을 조사하여 평균 연봉 추정치와95% 신뢰구간을구한다.*

예시: 항암제 효과

> *새로 개발된 항암제의 효과를 알아보기 위하여 무작위 배정 임상시험에서 관측한 치료군과 대조군의 암재발률을 비교한다.*

5. 데이터

하나 이상의 변수에 대한 관찰값의 모음

데이터의 기본 요소

이름	설명
단위(Unit)	관측되는 개별 대상

이름	설명
변수(variable)	각 단위에 대해 관측되는 특성
관찰값(observation)	각 단위로부터 관측한 특성의 값

예시: 4명의 데이터

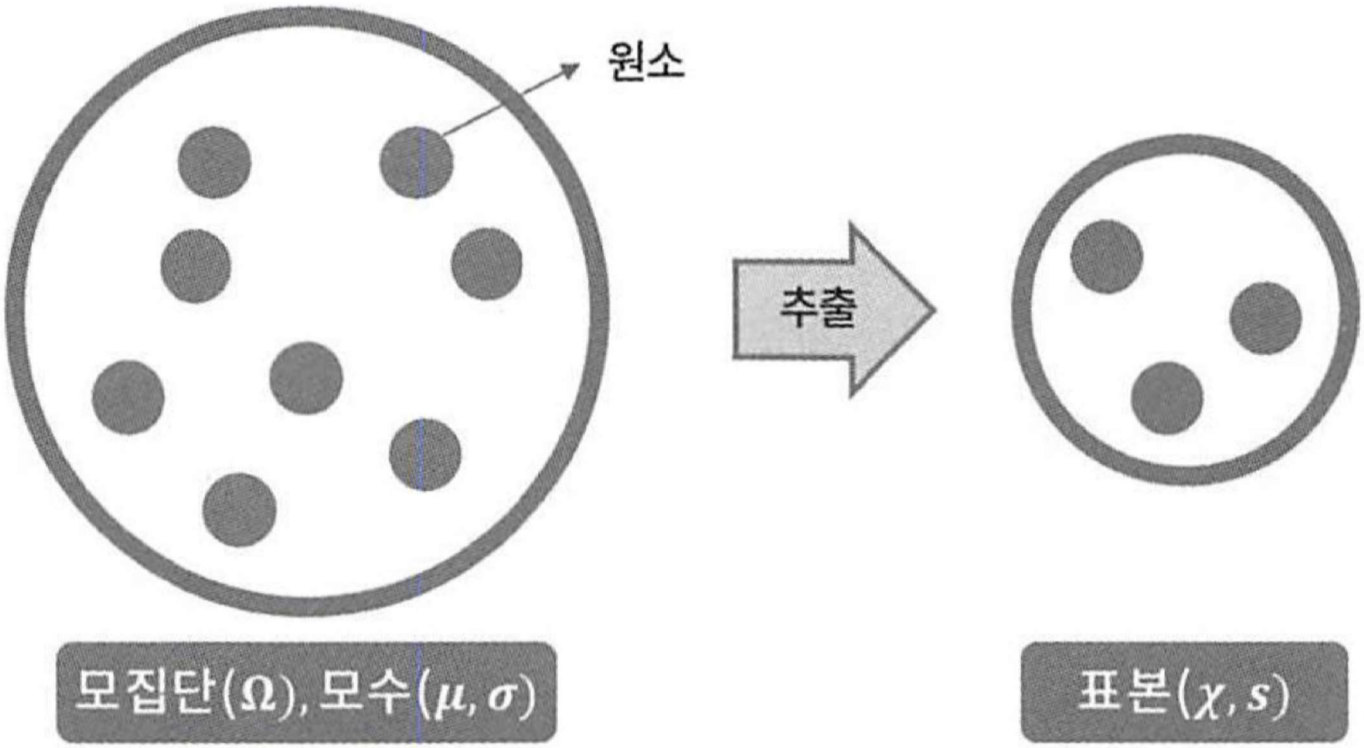
> *시연이는 여성이고 키161cm, 몸무게50kg이다. 이안이는 남성이고 키175cm, 몸무게73kg이다. 연하는 여성이고 키163cm, 몸무게55kg이다. 가현이는 여성이고 키171cm, 몸무게60kg이다.*

이름	성별	키(cm)	몸무게(kg)
시연	여	161	50
이안	남	175	50
연하	여	163	55
가현	여	171	60

- 단위 : 시연이, 이안이, 연하, 가현이
- 변수 : 성별, 키, 몸무게

#02. 통계학의 주요 개념

1. 모집단(population)과 모수(paramter)



1) 모집단

조사하고자 하는 대상 집단 전체로서 관심 대상이 되는 모든 개체의 모임을 의미한다.

대부분의 경우 모집단은 너무 커서 모든 개체를 조사할 수 없다.

- 원소(element) : 모집단을 구성하는 개체

모집단의 종류

종류	설명
유한모집단	개체 수가 유한개
무한모집단	개체 수가 무한개

2) 모수

모집단의 특성을 나타내는 대표값(평균)

대부분의 경우 값을 알 수 없다.

예외) 개체수가 적은 유한모집단인 경우 모든 개체를 조사하면 모수를 알아낼 수 있다.

2. 표본(sample)

모집단을 알기 위해 실제로 관측한 모집단의 일부로서 모집단을 잘 반영하는 표본을 뽑는 것은 매우 중요하다.

1) 확률화

모집단으로부터 편의되지 않은 표본을 추출하는 절차

2) 확률표본

확률화 절차에 의해 추출된 표본

3) 표본조사의 오차

표본오차: 모집단을 대표할 수 있는 표본 단위들이 조사대상으로 추출되지 못함으로서 발생하는 오차

표본편의

- 모수를 작게 또는 크게 할 때 추정하는 것과 같이 표본추출방법에서 기인하는 오차.
- 표본 추출 과정에서 특정 대상이 다른 대상에 비해 우선적으로 추출될 때 생기는 오차를 의미.
- 표본편의(Sampling Bias)는 확률화(Randomization)에 의해 **최소화하거나 없앨 수 있다.**

비표본오차

- 표본오차를 제외한 모든 오차로 모든 부주의나 실수, 알 수 없는 원인 등 모든 오차를 의미한다.
- 조사대상이 증가하면 비표본오차도 커진다.

4) 표본 추출 방법

단순랜덤표집(simple random sampling)

유한모집단에서 n개의 개체로 이루어진 가능한 모든 부분집합이 표본으로 선택될 확률이 같도록 설계된 표본 표집 방법

- 각 샘플에 번호를 부여하여 임의의 n 개를 추출하는 방법
- 각 샘플은 선택될 확률이 동일
- 비복원，복원(추출한 element 를 다시 집어넣어 추출하는 경우) 추출

계통추출법 (systematic sampling)

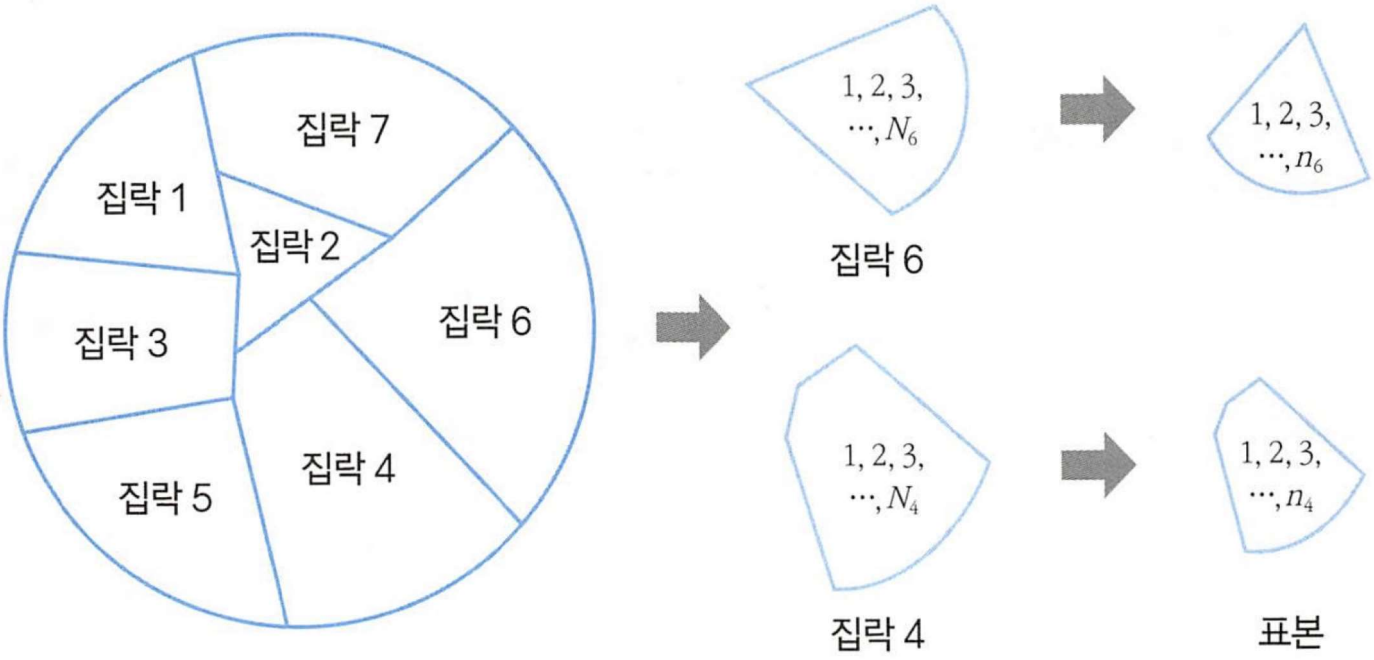
- 단순랜덤추출법의 변형
- 변형된 방식으로 번호를 부여한 샘플을 나열하여 K 개씩 (K=N/n) n 개의 구간 으 로 나누고 첫 구간 (1, 2, ⋯ , K) 에서 하 나 를 임의로 선택한 후에 K 개씩 띄 어서 n 개의 표본을 선택
- 임의 위치에서 매 k 번째 항 목 을 추 출

"5개 마다 조사"

예 15개 중 3개의 샘플 추출

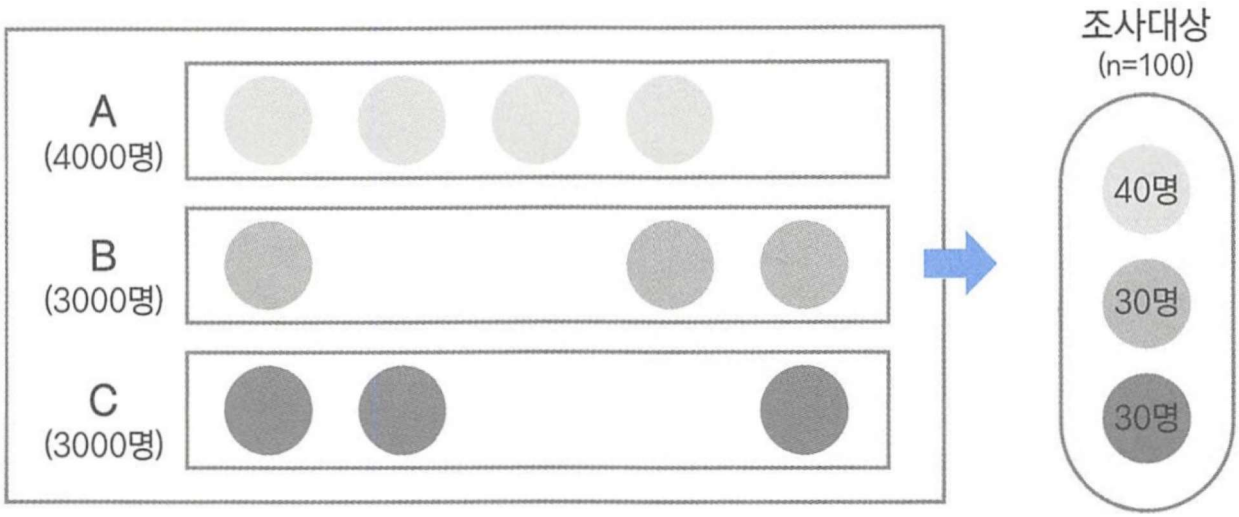
집락추출법 (cluster random sampling)

- 군집을 구분하고 군집별로 단순랜덤 추출법을 수행한 후, 모든 자료를 활용하거나 샘플링
- 지역표본추출, 다단계표본추출



층화추출법 (stratified random sampling)

- 이질적인 원소들로 구성된 모집단에서 각 계층을 고루 대표할 수 있도록 표본을 추출하는 방법
- 유사한 원소끼리 몇 개의 층 (stratum)으로 나누어 각 층에서 랜덤 추출



※ 실험 : 특정 목적 하에서 실험 대상에게 처리를 가한 후에 그 결과를 관측해 자료를 수집하는 방법이다.

4. 통계량(statistic)

표본의 특성을 나타내는 대표값

모수를 추정하기 위해 표본에서 얻은값

표본을 새로 뽑으면 통계량의 값이 달라진다

예시: 주거비

대한민국의 1가구당 평균 주거비를 알아보려고 한다. 전국의 모든 가구의 주거비를 설문하는 것은 너무 많은 시간과 비용이 필요하므로, 랜덤으로 뽑은 1,000가구에 방문하여 주거비를 조사한다.

구분	내용
모집단	대한민국의 모든가구
표본	랜덤으로 뽑은 1000가구
모수	대한민국의 가구당 평균 주거비
통계량	표본 1000가구의 평균 주거비

학습정리

- 통계학이란 불확실한 현상을 이해하기 위해 데이터를 수집하고, 데이터 패턴을 요약, 분석하여 불확실한 현상에 대한 결론을 찾는 학문이다.
- 통계학의 역할에는 데이터의 수집, 데이터의 요약, 추론이있다.
- 데이터는 하나 이상의 변수에 대한 관찰값의 모음이다.
- 데이터에서 관측되는 개별 대상을 단위라 하고, 각 단위에 대해 관측되는 특성은 변수라고 한다.
- 관심 대상이 되는 모든 개체의 모임을 모집단이라 하고, 모집단을 알기위해 실제로 관측한 모집단의 일부를 표본이라고 한다.
- 모집단을 잘 대표하는 표본을 표집하는 방법 중 가장 기본이 되는 방법은 단순랜덤표집이다.
- 모수는 우리가 알고싶은 모집단의 특성을 나타내는 대푯값이고, 모수를 알기 위해 표집한 표본의 특성을 나타내는 대푯값을 통계량이라고 한다

— ABOUT 호쌤(이광호)

메가스터디IT아카데미에서 Java, Spring, Python, Frontend 등을 강의하는 IT 전문 강사이자 프리렌서 개발자 입니다.
https://www.youtube.com/@hossam-codingclub

SEOUL, KOREA HTTP://WWW.HOSSAM.KR