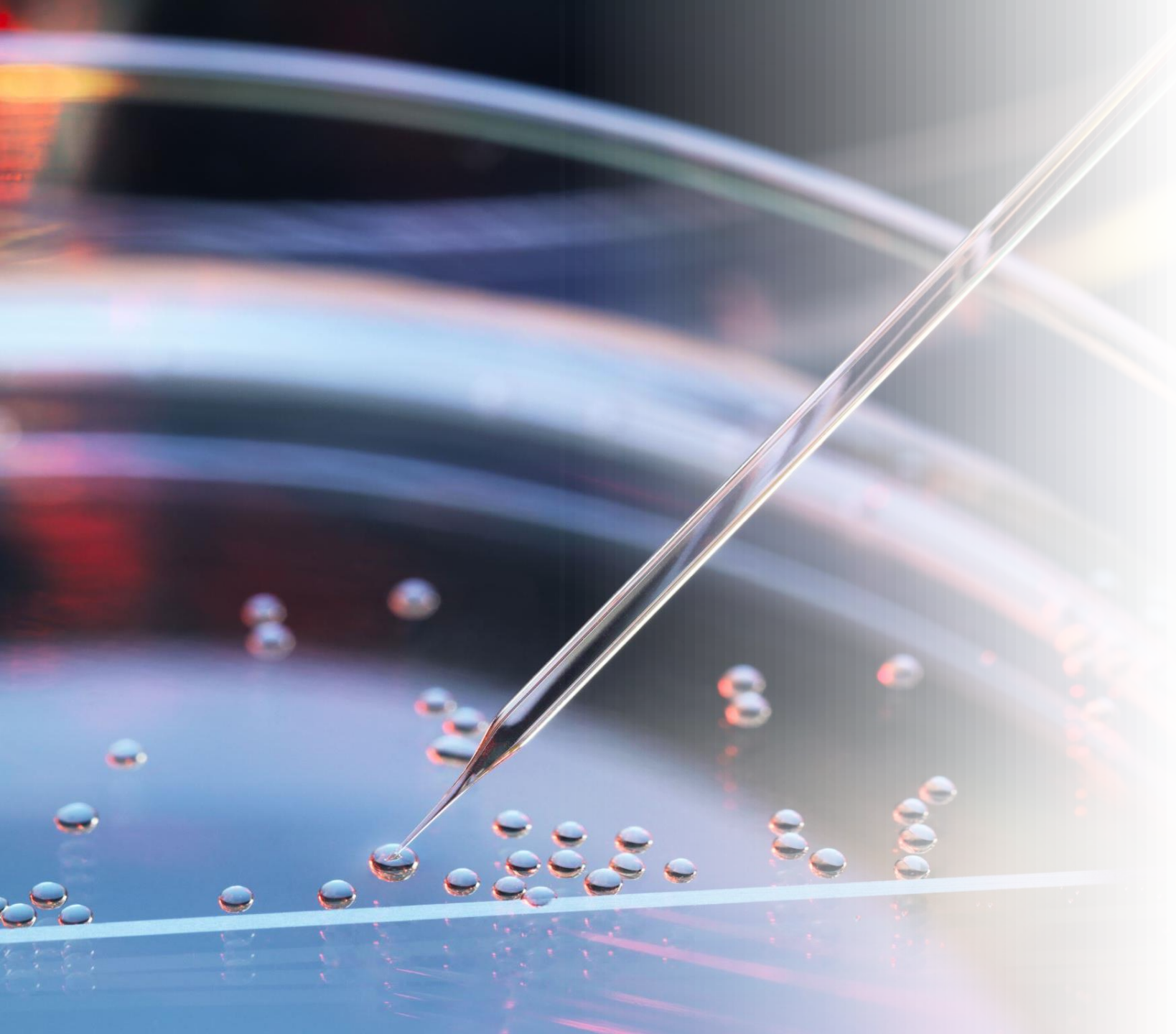
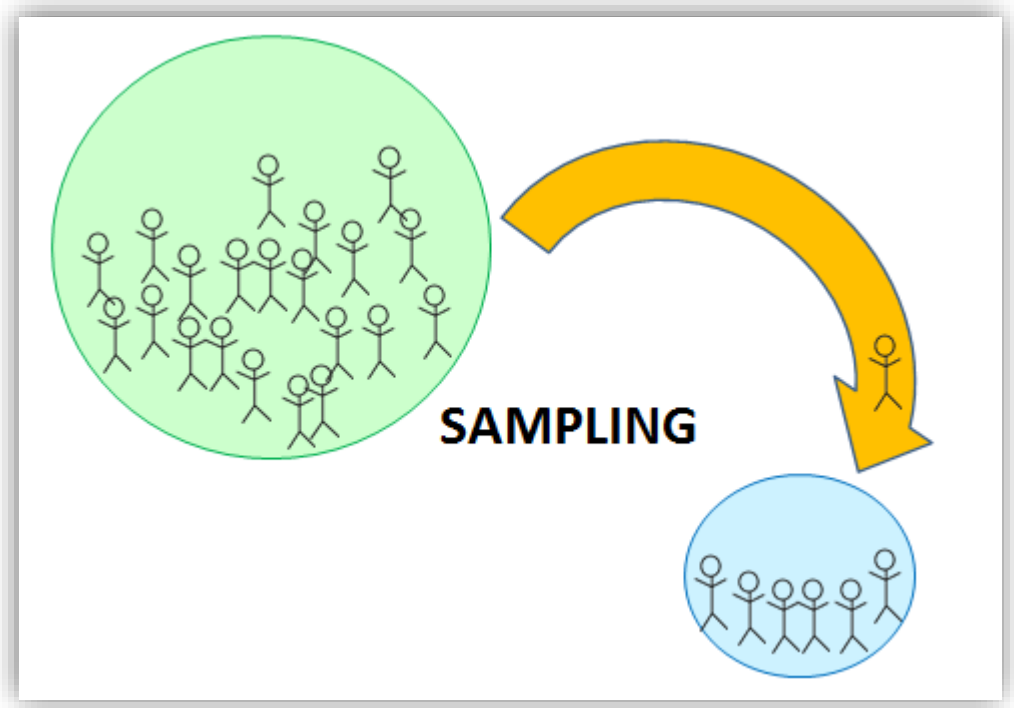


Sampling

Sampling

Sampling?

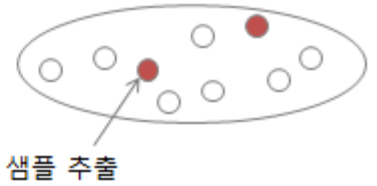
전체 모집단(population)에서
일부 데이터를 선택하여 대표성 있는 표본(sample)을 추출하는 과정

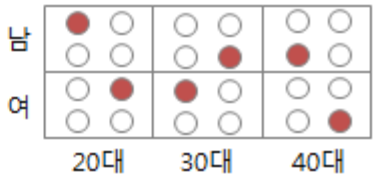


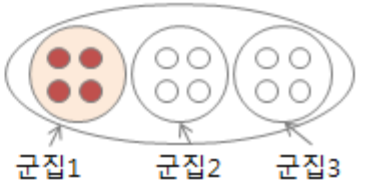
Sampling

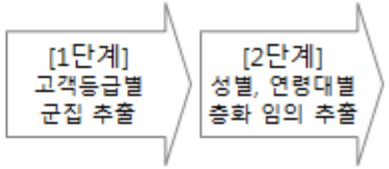
Sampling 기법

단순 임의 추출 (simple random sampling)	모집단으로부터 샘플을 균등하게 임의로 추출하는 방법 (복원추출, 비복원추출)
체계적 추출 (systematic sampling)	무작위로 배열된 표본에서 시간적 혹은 공간적으로 일정한 간격을 두고 표본 추출 (예: 매 5번째 추출 등)
층화 임의 추출 (stratified random sampling)	모집단이 몇 개의 계층(stratum)으로 구성되어 있을 때 각 계층 원소로부터 임의 추출 (예; 성별, 연령대별 등)
군집 추출 (cluster sampling)	모집단이 여러 군집을 이룰 때, 우선 군집을 선택하고 그 집단에서 표본을 추출
다단계 추출 (multi-stage sampling)	여러 단계로 나누어서 표본 추출 수행









https://rfriend.tistory.com/58

Sampling

Sampling 이 필요한 이유

1.시간 및 비용 제약 극복

> 통상 전체 데이터를 모두 수집하고 분석하는 것에는 시간과 비용이 많이 소요

> 데이터 샘플링을 통해 일부 데이터를 추출하여 시간과 비용을 절약

2. 데이터의 신뢰도 향상

> 샘플링을 통해 불필요한 부분을 선택하지 않으므로 Noise 등 데이터 왜곡 요소 제거

> 샘플링을 통해 이 '데이터 정제' 역할 수행

3. 효율적으로 인사이트 도출

> 패턴이나 인사이트 등이 샘플링 한 후에는 조금 더 쉽게 보이는 경우 있음

> 더 작은 규모의 데이터는 핸들링이 수월하여 분석 작업의 효용 제공

Sampling

Sampling 종류

- **단순 무작위 샘플링 (Simple Random Sampling)**
  - ✓ 모집단에서 임의로 샘플을 선택하는 가장 기본적인 방법
  - ✓ 모집단의 크기가 N인 경우 크기가 n인 모든 가능한 샘플을 동일한 확률로 추출 하는 방법

- **통계적 의미**

  통계학에서는 데이터를 모집단에서 추출한 표본, 즉 모집단의 일부로 간주. **표본을 통해 모집단을 추론 (평균 등)**

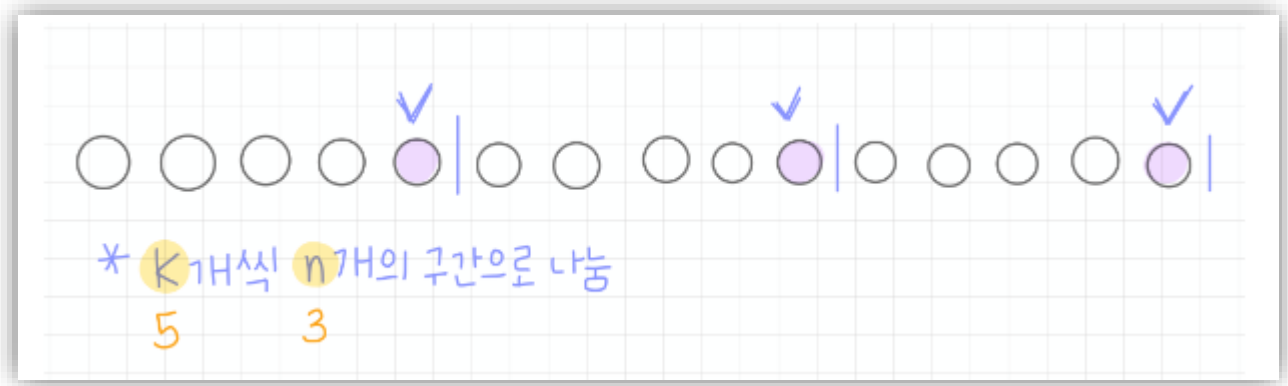
  표본을 모집단에서 추출할 때 **편향되지 않도록** 해야

  '표본이 편향되지 않았다'라는 것은 표본을 추출할 때 모집단의 각 개체가 모두 동일한 확률로 뽑혔다는 의미

Sampling

Sampling 종류

- **계통 샘플링(Systematic Sampling)**
  - ✓ 첫 번째 샘플을 무작위로 선택한 후, 그 다음 샘플은 사전에 정해진 간격으로 선택
  - ✓ 모집단에 대한 데이터 목록이 정렬되어 있거나 데이터에 규칙적인 패턴이 있는 경우에 유용



- Ex)
  : 런칭한 제품에 대한 마케팅 조사, 일정한 간격으로 구매한 고객을 선정
  : 공장 생산 조립 라인에서 매 시간마다 나오는 제품의 품질 분석

Sampling

Sampling 종류

- **층화 샘플링(Stratified Sampling)**
  - ✓ 모집단을 비슷한 특성을 가진 여러 개의 층(계층)으로 나눈 후, 각 층에서 단순 무작위 샘플링
  - ✓ 모집단의 크기가 N인 경우 크기가 n인 모든 가능한 샘플을 동일한 확률로 추출 하는 방법

- Ex)

  : 연령별, 성별

  : 연령대별로 조사하는 경우, 각 연령대를 층으로 나누고, 각 연령대에서 n개의 샘플을

  : 전체 모집단 뿐만 아니라 각 층의 특성에 대한 추정도 할 수 있다는 장점

Sampling

Sampling 종류

- **군집 샘플링(Cluster Sampling)**

  모집단을 여러 군집으로 나눈 후 일부 군집을 무작위로 선택, 선택된 군집의 **모든** 구성원을 조사

  모집단의 크기가 N인 경우 크기가 n인 모든 가능한 샘플을 동일한 확률로 추출 하는 방법

- Why?

  : 샘플링 프레임을 얻기 어려운 경우(ex 서울시 가구 대상 연구, 가구 정보 필요)

  : 각 군집에 속한 개체들은 서로 이질적이고 각 군집들은 유사한 특성을 지녀야

Sampling

비 확률적 Sampling

✓ 특정 표본이 샘플에 포함될 확률을 정확히 알 수 없는 방법

✓ 주로 시간, 비용, 데이터 접근성 등의 제약이 있을 때 사용

- **편의 샘플링(Convenience Sampling)**: 접근하기 쉬운 표본을 샘플로 선택하는 방법
- **판단 샘플링(Judgment Sampling)** 또는 **목적 샘플링(Purposive Sampling)**:
  연구자의 판단에 따라 특정 기준을 충족하는 표본을 선택
- **눈덩이 샘플링(Snowball Sampling)**:
  초기 샘플 구성원들이 다른 구성원들을 추천하는 방식으로 샘플을 확장해 나가는 방법
  (ex 초기 인터뷰 대상이 다음 인터뷰 대상을 소개)

Sampling

샘플링 기법의 선택

> 연구의 **목적**, 모집단의 **특성**, 시간 및 **비용**의 제약, 필요한 데이터의 **정확성**과 **대표성** 등 고려

> 확률 샘플링 방법은 일반적으로 더 높은 대표성과 정확성을 제공하지만, 비용과 노력 필요

> 비확률 샘플링은 더 쉽고 빠르게 데이터를 수집할 수 있지만, 선택된 샘플이 모집단을 정확하게 대표하는지 확신하기 어려움.

> **데이터의 특성과 분석의 목적**에 대한 고민이 선행되어야

Sampling

샘플링 기법의 선택 - 머신러닝

1. 데이터의 특성과 분포

> **데이터 크기와 복잡성**: 대규모 또는 복잡한 데이터 세트의 경우, **계산 비용**을 줄이고 **효율성**을 높이기 위해 샘플링 실행

> **데이터 분포**: 데이터에 편향(bias)이 있는 경우 **층화 샘플링** 같은 방법을 사용, 각 범주의 데이터 샘플이 균등하게 표현되도록 조치

> **클래스 불균형**: 대부분의 데이터가 한 클래스에 속해 있고 다른 클래스는 소수만 존재하는 경우, **오버샘플링**이나 **언더샘플링** 기법 고려필요

2. 연구 목적 및 모델 요구 사항

> **정확도 대 효율성**: 고정밀도가 필수적인 경우, 더 많은 데이터를 포함시키거나 층화 샘플링을 적용하여 대표성을 높여야. 빠른 프로토타이핑이 목적이라면 더 간단한 샘플링 방법 사용 가능

> **모델의 종류**: 사용하는 머신러닝 모델에 따라 샘플링 전략이 달리해야 하는 경우 고려

ex) 클래스 불균형에 더 민감한 모델인 경우

Sampling

샘플링 기법의 선택 - 머신러닝

3. 실행 시간 및 리소스 제약

> **계산 리소스**: 제한된 계산 리소스를 가진 경우, 효율적인 샘플링을 통해 데이터 크기를 줄이고 모델 학습 시간을 단축. 적절한 샘플링을 통해 메모리 문제 방지

4. 샘플링 편향 방지

> **대표성 유지**: 샘플링 과정에서 데이터의 대표성을 유지해야 샘플이 모집단의 특성을 충분히 반영하지 못하면? -> 모델의 **일반화 능력** 저하

5. 재현성 및 검증

> **샘플링 과정의 재현성**: 실험의 재현성을 위해 샘플링 과정을 명확히 문서화하고, 고정된 시드값을 사용하여 샘플링 수행

> **교차 검증**: 샘플링 된 데이터에 대해 교차 검증을 수행하여 모델의 성능을 검증하고, 샘플링이 모델 성능에 미치는 영향 평가

Sampling

Q. 빅데이터 시대에 샘플링?

Sampling

- 실습

  4.1.sampling.ipynb

  4.1.sampling.modeling.ipynb

THANK YOU