빅데이터 시대에는 표본추출이 필요없다? 오히려 데이터 질이 떨어지고 크기만 늘어나 표본추출이 중요해졌다.

* 빅데이터 프로젝트라도, 먼저 작은 샘플로 모델링과 테스트를 거친다.
* 모집단보다는 표본추출과정과 추출한 데이터에 집중한다.

# 2.1 랜덤표본추출과 표본편향

* 표본은 더 큰 데이터 집합(모집단)에서 얻은 데이터의 부분집합이다.
  + 모집단은 유한한 전체집합이라 볼 수도 있고, 이론적인 가상의 집합을 뜻하기도 한다.
* **랜덤표본추출**(random sampling): 모집단에서 무작위로 데이터를 추출하는 과정. 모집단 내 모든 원소들은 동일한 추출 확률을 가진다.
  + 랜덤표본추출로 얻어진 표본을 단순랜덤표본이라 한다.
  + 추출된 샘플을 다시 모집단에 포함시키면 복원추출, 추출된 샘플을 사용하지 않는다면 비복원추출.
  + 자기 선택 편향: 리뷰 사이트 같은 경우 표본 자체가 특수하다. (긍정적이든 부정적이든) 특정 사건을 겪은 사람일 가능성이 높다. 전체 모집단을 파악하기에는 무리가 있지만, 단순 비교의 목적으로서는 신뢰할만 하다.
* 샘플 기반 추정이나 모델링에선 데이터 양보다 질이 더 중요하다. 표본편향이 발생하지 않도록 해야한다.
  + 랜덤표본추출을 사용한다 하더라도, 모집단 자체에 오류가 있다면 비임의추출이다.
  + 자기선택 표본편향: 리뷰 사이트같은 경우, 모든 모집단을 대표하는 것이 아니라, (좋든 싫든) 특정한 사건을 겪은 사람일 가능성이 높다.

편향

* 통계적 편향은 측정 과정 혹은 표본추출 과정에서 생겨난 계통적 오차.
  + 랜덤표본추출으로 인한 오류인지, 편향에 따른 오류인지 살펴봐야 한다.

랜덤 선택

* 층화표본추출은 모집단을 여러 층으로 나누어 무작위로 샘플을 추출하는 방법.
  + 특정 계층의 데이터 수가 적을 땐 해당 층에 가중치를 두면 계층마다 표본크기를 얻을 수 있다.

크기와 품질: 크기는 언제 중요해질까?

* 데이터의 개수가 적을수록 더 유리한 경우가 있다.
  + 랜덤표본추출에 시간과 노력을 기울일수록 편향이 줄고 데이터 탐색과 데이터 품질에 집중할 수 있다.
    - 수천 개의 샘플에서는 결측값이나 특잇값 찾기에 수월하겠지만, 몇백만 개의 데이터라면 일일히 조사는 불가능하다.
* 빅데이터가 가치있을 때는 **데이터가 큰 동시에 희박할 때.**
* 모집단과 표본의 평균을 구분하는 이유?
  + 표본에 대한 정보는 관찰을 통해 얻어지고, 모집단에 대한 정보는 주로 작은 표본들로부터 추론한다.