Chapter 02. 데이터와 표본분포 - 설유환

CHAPTER 2 데이터와 표본분포

2.1 랜덤표본추출(random sampling)과 표본편향

표본추출의 중요성

- 빅데이터 시대가 되면서 더는 표본추출이 필요 없을 것이라고 생각하는 사람이 증가
    - 왜냐면 엄청나게 많은 데이터의 양을 다룰 수 있어지니까
- 그러나 데이터의 질과 적합성을 일정 수준 이상으로 담보할 수 없으면서 데이터 크기만 늘어나는게 오늘날 상황
    - 데이터 크기만 엄청 커지고 검증이 안되는 데이터들도 끼워짐
- 오히려 다양한 데이터를 효과적으로 다루고 데이터 편향을 최소화하기 위한 방법으로 표본추출의 중요성이 더욱 커짐
- 전통적인 통계학에서는 강력한 가정에 기초한 이론을 통해 왼쪽의 모집단을 밝혀내는 데 초점을 맞춘데 비해, 현대 통계학에서는 이러한 가정이 더 이상 필요하지 않은 오른쪽에 대한 연구로 방향이 옮겨지기 시작

🙌 표본 : 큰 데이터 집합으로 부터 뽑은것들(집합)

모집단 : 데이터 집합을 구성하는 전체 대상

랜덤 샘플링 = 임의표본추출 = 임의 표집 = 랜덤표본 추출 : 무작위 추출

층화표본추출 : 층으로 나눠서 표본추출

계층 Stratum ⇒ 공통된 특징을 가진 모집단의 동종 하위 그룹

단순 임의표본 (단순 랜덤표본) : 모집단 층화 없이 임의표본 추출로 얻은 표본

편향 (bias): 계통상의 오류

표본편향 : 모집단을 **잘못** 대표하는 표본

편향 (탄착군)

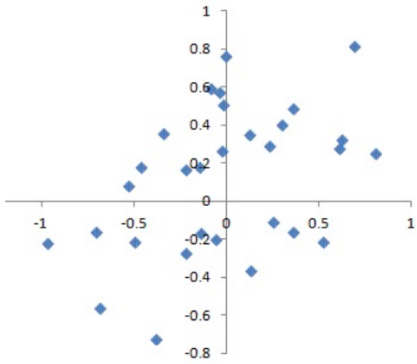


Figure 2-2. Scatterplot of shots from a gun with true aim

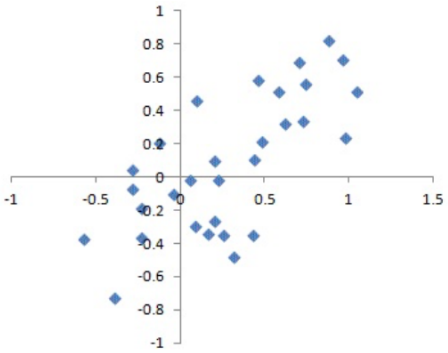


Figure 2-3. Scatterplot of shots from a gun with biased aim

- (위) 랜덤표본추출로 인한 오류: 오차가 랜덤하며 어느 쪽으로 강하게 치우치는 **경향이 없음**
- (아래) 편향에 따른 오류: 랜덤한 오차도 존재하고 특정한 곳으로 치우치는 경향이 있음

  bias란... 한국어로 '**치우침**'이라고 할까요.

  사격에서 총을 여러발 쏘았을 때, 탄착군이 전반적으로 치우치는 형상을 떠올리시면 되겠습니다.

2.1.4 표본평균 모평균

- 예를 들어, **미국 여성들의 평균 신장**을 구하려고 한다. 이를 위해 각 주(state)마다 1000명의 표본을 추출했다. 워싱턴, 뉴욕 등 50개의 주에서 추출한 샘플들로부터 각각 구한 평균값이 표본평균이다.
  이 표본평균은 각 주마다 다르다. 워싱턴의 표본평균은 161일 수 있고, 뉴욕에서의 표본평균은 155일 수 있다. 이런 이유로 아까 위에서 표본평균은 고정된 값이 아니라고 한 것이다.
- 표본평균은 샘플링을 할 때마다 다른 값이 나오므로 당연히 모평균(mean of population)과 같을 수 없다.

(https://m.blog.naver.com/owl6615/221847019934)

2.2 선택 편향

선택 편향 : 관측 데이터를 선택하는 방식 때문에 생기는 편향

데이터 스누핑 : 흥미로운 것을 찾아 광범위하게 데이터를 살피는 것

방대한 검색효과 : 중복 데이터 모델링이나 너무 많은 예측변수를 고려하는 모델링에서 비록되는 편향 혹은 비재현성

- 선택편향을 조심할 필요가 있다.
  - 서울대 IQ 테스트 ⇒ 사실 한국인 평균 지능이 이정도다
  - 아래는 웃자고 넣은것

- 그래서 hold out 사용함 ⇒ 교차검증
- 목푯값 섞기 = 최종적으로는 순열검정

2.21 평균으로의 회귀

> 여기서의 회귀는 돌아간다 라는 의미로서 통계적 모델링 방법중 하나인 '선형회귀'와 구분되어야 한다.

- 평균으로의 회귀는 일종의 선택편향으로 인해 나타나는 결과이다.
    - 성적으로 신인을 뽑을때 ⇒ 운을 고려하기 힘들다
    - 예외적인 변수가 관찰되면 그 다음에는 중간 정도의 경우가 관찰되는 경향이 있다.

⇒ 예시로 신인왕 슬럼프가 있다. 신인왕 수상자들이 2년차에는 죽을쑤는 현상을 의미한다.

> 만일 평균으로의 회귀가 일어나지 않는다면 세대를 거듭하며 키가 큰 사람들이 낳은 자식들은 한없이 키가 커지고, 키가 작은 사람들이 낳은 자식들은 계속 작아지며 결과적으로 세상은 비정상적으로 키가 크거나 작은 사람들로 양분화 되었을 것이다.

- 피하려면 어떻게 해야 하는가 ?
    1. 가설을 구체적으로 명시하고
    2. 임의표본추출 (랜덤샘플링)을 한뒤 데이터를수집하면 편향을 피할 수 있다.
    3. 모든 형태의 데이터 분석은 **수집/분석 단계에서 생기는 편향의 위험성을 항상 가지고 있다.**

추가적으로, **확증편향**도 있다.

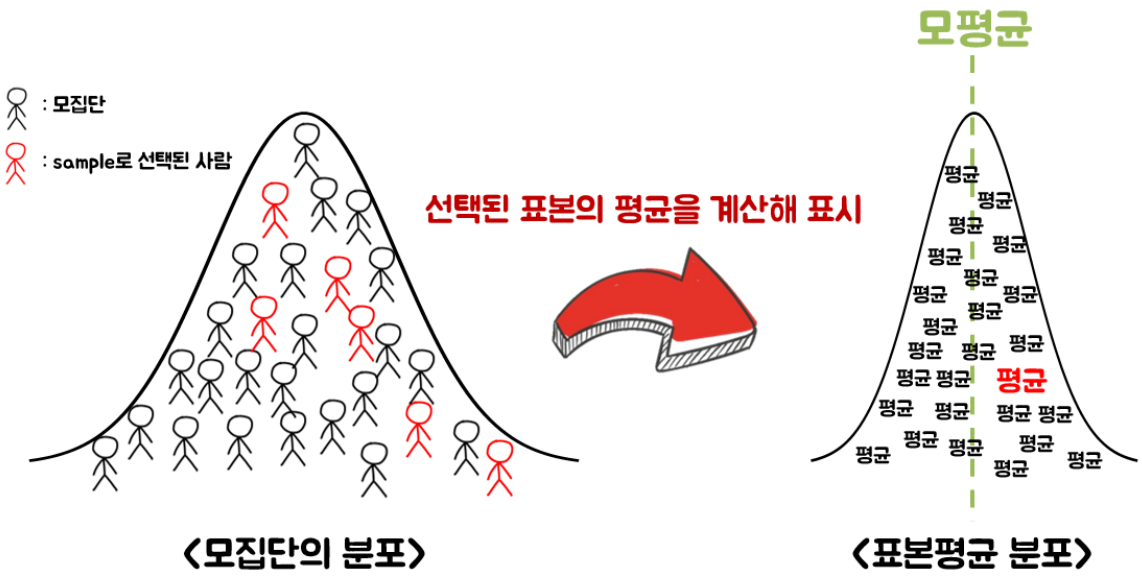
> **확증편향** -
>
> 자기의 의견이나 태도에 유리한 방향으로 편향되게 증거를 생성하고 평가하며 가설을 검증하는 행동 경향성을 자기 측 편향(myside bias)이라 한다.
> 자기 측 편향은 확증 편향의 한 유형이다. **확증 편향(confirmation bias)**이라는 용어는 의견이나 태도에 국한되지 않고 자기의 가설을 확증하는 증거를 선호하는 '모든' 상황에 적용되기 때문에 자기 측 편향보다 광범위하다. 쉽게 말해, **자기가 보고 싶은 것만 보고 믿고 싶은 것만 믿는 현상**을 뜻한다. 이런 편향성은 의사결정 시에만 일어나는 것이 아니라 정보를 수집하는 단계에서부터 나타난다.

> 사람들은 자신의 생각이 사실임을 확인시켜주는 정보를 찾고, 자신의 생각에 반하는 정보는 무시하는 경향이 있다. 동일한 출처, 동일한 내용의 정보를 반복해서 받아들임으로써 기존의 생각을 계속해서 강화하며 자기 합리화를 하는 것이다. 정보의 중복은 자신의 판단에 대한 확신을 더하게는 해주지만 이것이 판단의 정확도를 높여주는 것은 아니다.

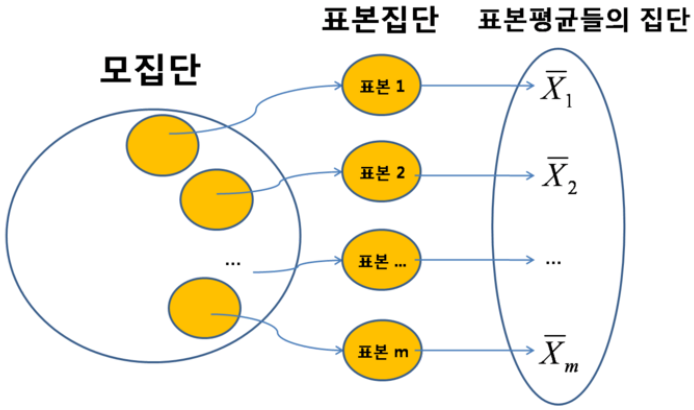
2.3 통계학에서의 표본분포

- 용어정리

👐 - 표본통계량 sample statistics : 더 큰 모집단에서 샘플링된 데이터들로 부터 얻은 측정지표
- 데이터 분포 : 어떤 데이터 집합에서 각 개별 값의 도수 분포
- 표본 분포 : 표본들로부터 얻은 도수분포
- 중심극한정리 : 표본 크기가 커질수록 표본분포가 정규분포를 따르는 경향
- 표준오차 : 여러 표본들로부터 얻은 표본통계량의 변량
- 표준편차 : 개별 데이터 값들의 변량



- **표준오차 vs 표준편차**
  - https://www.youtube.com/watch?v=4BCwUq17Thw
  - 퍼짐 정도(분포)를 안다면 의사결정에 더 유리하다
    - 둘다 분포의 특성을 나타내는 말이긴 함
  - 표준편차는 해당 집단 내 각 표본의 측정값이 평균에서 어느 정도 떨어져 있는지를 나타내는 지표
  - 표준오차는 모평균의 추정치인 표본평균이 가지는 표준편차
    (standard deviation of the sample-mean's estimate of a population mean)



  - 모집단에서 표본집단을 m 개 추출하게 되면 서로 다른 m 개의 표본평균이 계산되기 때문에
    이들 표본평균들 간에도 편차가 존재하게 됩니다.
    이렇게 표본평균 내에 존재하는 편차를 표준오차 라고 합니다.

$$SE = \frac{s}{\sqrt{n}}, \quad s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

X_i : i번째 개체의 측정값

$\bar{X}$: 표본평균

s : 표본표준편차

n : 표본크기

SE : 표준오차

- 평균과 같은 표본 통계량의 분포는 데이터 자체의 분포보다 규칙적이고 종 모양(normal distribution)일 가능성이 높다.

  표본이 클수록 그럴 가능성이 높은 것이 사실이다.

  표본이 클수록 표본 통계량의 분포가 좁아진다.

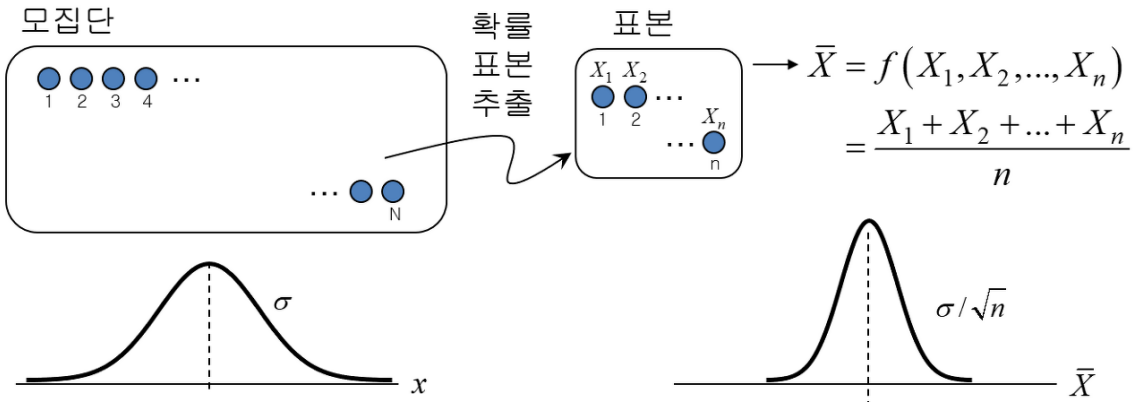
2.3.1 중심극한정리 CLT

> 모집단이 「평균이 μ이고 표준편차가 σ인 임의의 분포」을 이룬다고 할 때, 이 모집단으로부터 추출된 표본의 「표본의 크기 n이 충분히 크다」면 표본 평균들이 이루는 분포는 「평균이 μ 이고 표준편차가σ/√n인 정규분포」에 근접한다.

> 여기서 많은 분들이 헷갈리시는 부분이 있습니다. 생각보다 많은 분들이 중심극한정리를 "**내가 수집한 표본의 크기가 크면, 그 표본의 평균이 모집단의 평균과 같고, 표본의 표준편차가 모집단의 표준편차를 표본수로 나눈 값과 같게 된다.**"라고 이해하곤 합니다.
> 이와 같이 중심극한정리를 이해 했다면, **이건 중심극한 정리를 완전**

> **히 잘못 이해한 것입니다**. 표본은 매번 추출할 때마다 달라지게 되고, 그에 따라 표본의 평균값도 매번 달라지기 때문입니다. 따라서 우리가 연구를 위해 수집한 표본의 평균값이 아무리 크기가 크다고 하더라도 모집단의 평균값과 같다고 말할 수 없습니다.

> 그렇다면 중심극한정리에서 말하는 표본평균분포란 무엇일까요? 중심극한정리에서 말하는 표본평균분포는 내가 수집한 표본을 말하는 것이 아닙니다. **표본평균분포는 영어로 Sampling distribution of sample mean**입니다.
> 즉 **표본평균분포는 "모집단에서 표본크기가 n인 표본(예: 30개)을 여러번 반복해서 추출(예: 200번 추출)했을 때 (즉, X1(nㅈ=30), X2(n=30), X3(n=30), ... X200(n=30), 각각의 표본 평균들이 이루는 분포**"를 말합니다. 그리고 중심극한정리는 그 표본의 크기가 커질 수록 (**보통 30 이상**), 표본 평균들이 이루는 분포가 <모집단의 평균 μ 그리고 표준편차가σ/√n인 정규분포>에 가까워진다는 정리입니다. 이 말을 그림으로 정리하면 아래와 같습니다.



- 그렇다면 왜 중심극한정리가 중요한 것일까요?
  - 그것은 중심극한정리가 표본 수집을 기반으로 한 추리통계에서 **아주 중요한 이론적 근거를 제시하고 있기 때문**입니다.
  - 쉽게 설명드리면 우리는 이 정리를 통해, 모집단이 어떤 분포를 가지고 있던지 간에 (모집단 분포가 모양이던 상관없이) 일단 표본의 크기가 충분히 크다면,

표본평균들의 분포가 모집단의 모수를 기반으로한 정규분포를 이룬다는 점을 이용하여, 특정 사건(내가 수집한 표본의 평균)이 일어날 확률값을 계산할 수 있게 됩니다.
다시 말해 중심극한정리는 **표본 평균들이 이루는 표본 분포와 모집단 간의 관계를 증명함**으로써, 수집한 표본의 통계량(statistics)을 이용해 **모집단의 모수(Parameters)를 추정할 수 있는 수학적(확률적) 근거를 마련해 줍니다.**
이것이 추리통계에서 중심극한정리가 중요한 이유입니다.

2.4 부트스트랩

> 🙌🏻 Bootstrap : (부츠 뒤의) '가죽 손잡이', 혹은 (비유적으로) '혼자의 힘'을 뜻한다고 합니다. 부츠를 신을 때 손잡이를 이용하면 다른 사람의 힘을 빌리지 않고 신을 수 있듯이 Bootstrap sampling이란 비용과 시간이 많이 드는 데이터 수집을 스스로 해결할 수 있는 샘플링 방법이라고 간단히 말할 수 있다고 합니다.

부트스트랩(Bootstrap)이란?

- 현재 있는 표본에서 **추가적으로 표본을 복원-추출**하고 각 표본에 대한 통계량과 모델을 다시 계산하는 것이며, 데이터나 표본통계량이 정규분포를 따라야 한다는 가정은 꼭 필요하지 않다.
- 즉, 원래 표본을 수천, 수백만 번 복제하는 것이라고 할 수 있다.
- 이를 통해 원래 표본으로부터 얻어지는 모든 정보를 포함하는 가상 모집단을 얻을 수 있다.

부트스트랩(Bootstrap)은 표본통계량의 변동성을 평가하는 강력한 도구이다.

부트스트랩 재표본추출 알고리즘

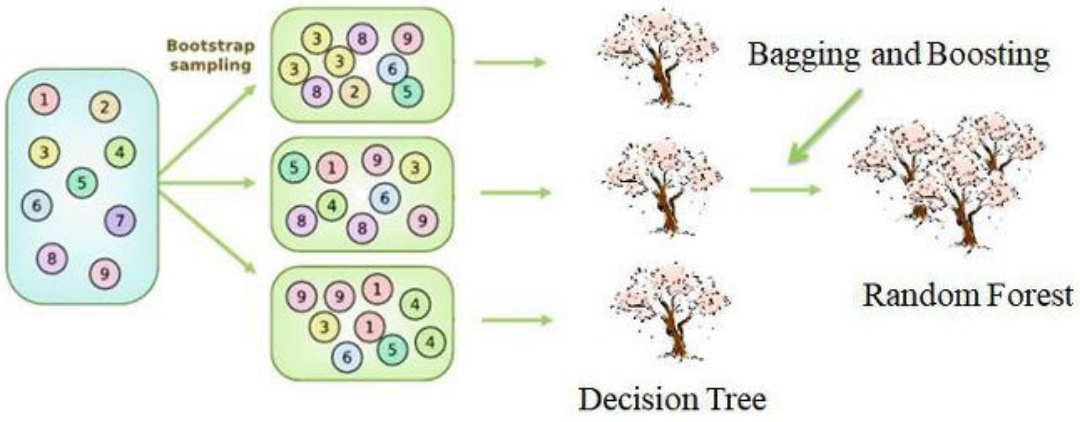
1. 샘플 값을 하나 뽑아서 기록하고 제자리에 놓는다.
2. n번 반복한다.
3. 재표본추출된 값의 평균을 기록한다.
4. 1~3단계를 R번 반복한다.
5. R개의 결과를 사용하여 표준편차, 히스토그램, 신뢰구간 등을 찾는다.

배깅(Bagging)이란?

- 배깅(Bagging)은 앙상블 기법의 종류 중 하나이며,
  부트스트랩(Bootstrap) 데이터를 가지고 모델을 돌려 모델 파라미터의 안정성(변동성)을 추정하거나 예측력을 높일 수 있다.
- 이를 활용하여 분류 및 의사 결정 트리를 사용할 때, **여러 부트스트랩(Bootstrap) 샘플을 가지고 트리를 여러 개 만들어 각 트리에서 나온 예측값을 평균 내는 것이 단일 트리를 사용하는 것보다 효과적**인데, 이 방법을 배깅(Bagging)이라고 한다.

배깅(Bagging) 기법을 활용한 모델이 우리가 흔히 사용하는 랜덤 포레스트 모델(Random Forest Model)이다.

- Developed by Leo Breiman at University of California, Berkeley (1996, 1999)
- Special case of the "model averaging" approach
  - Attempt to reduce bias of single tree

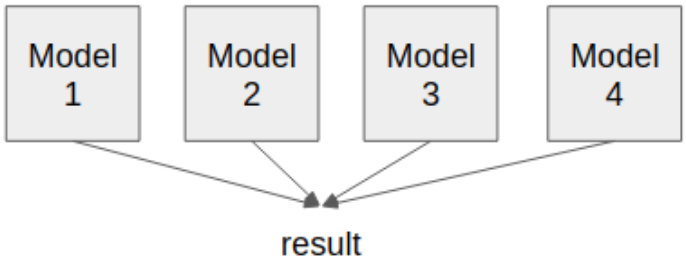


ML에서의 Bootstrap

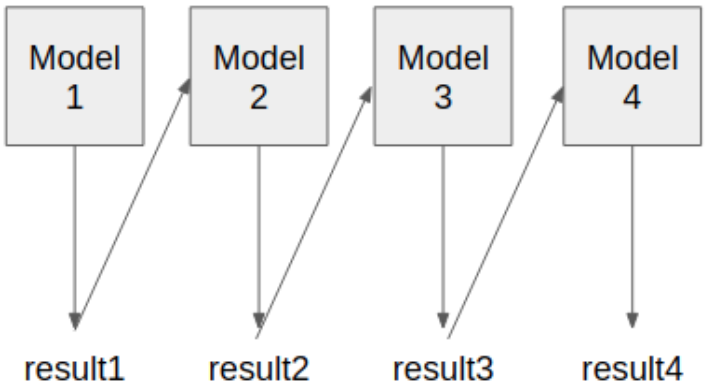
- 머신러닝에는 Bootstrap은 원래의 데이터 셋으로부터 랜덤 샘플링을 통해 학습데이터(Training Data)를 늘리는 방법입니다. 데이터 양을 늘릴 수 있고, 분포를 고르게 만들 수 있는 효과가 있습니다.
- 그리고 Bootstrap을 이용하면 **Ensemble**을 사용할 수 있습니다.

- Ensemble은 Bagging과 Boosting 2가지 방식이 존재합니다.

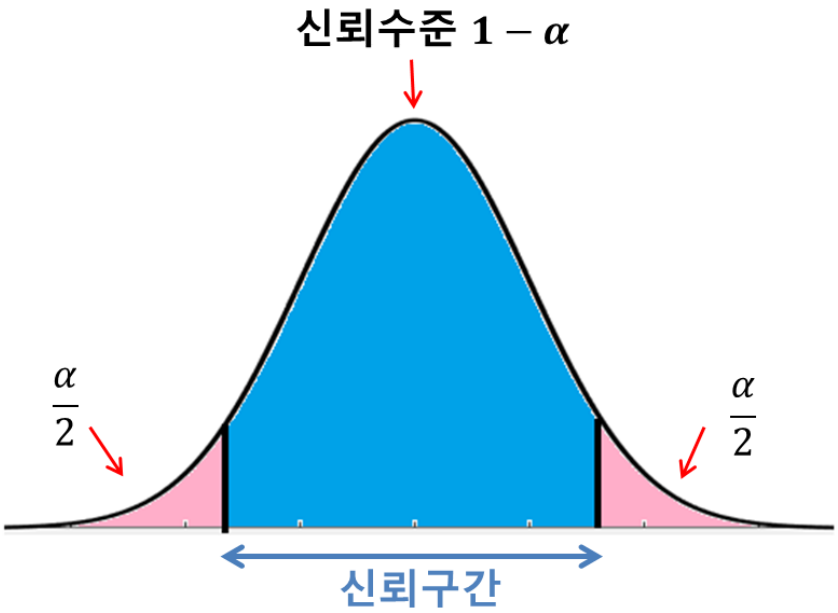
- Bagging은 Bootstrap으로 조금씩 서로 다른 훈련 데이터를 생성하고, 병렬로 처리하여 결과를 결합하는 방법입니다.



- Boosting은 잘못 분류된 객체들에 집중하여 새로운 분류 규칙을 생성하는 단계를 반복하는 순차적 학습 알고리즘입니다.



2.5 신뢰구간



> ‘95% 확률’ 이라는 용어를 썼지만, 어떤 경우에는 확률이라는 말을 ‘95% 신뢰 수준’이라는 말로 바꿔 쓰기도 한다. 즉, 신뢰 수준이라는 말은 확률이라는 말과 궤를 같이 한다고 할 수 있다.

- 신뢰수준 : 같은 모집단으로부터 같은 방식으로 얻은, 관심 통계량을 포함할 것으로 예상되는, 신뢰구간의 백분율
- 구간끝점 : 신뢰구간의 최상위, 최하위 끝점

- 표본추정치 주위의 x% 신뢰구간이란, 평균적으로 유사한 표본추정치 x% 정도가 포함되어야 함(비슷한 샘플링 절차를 따랐을 때)
  - 90% 신뢰구간 : 표본통계량의 부트스트랩 표본분포의 90%를 포함하는 구간

- 부트스트랩 신뢰구간 구하는 법

📌 1. 데이터에서 복원 추출 방식으로 크기 n인 표본을 뽑는다.
2. 재표본추출한 표본에 대해 원하는 통계량을 기록한다.
3. 1~2 단계를 R번 반복한다.
4. x% 신뢰구간을 구하기 위해 R 개의 재표본 결과의 분포 양쪽 끝에서 [(100-x)/2] % 만큼 잘라낸다.
5. 절단한 점들은 x% 부트스트랩 신뢰구간의 양 끝점이다.

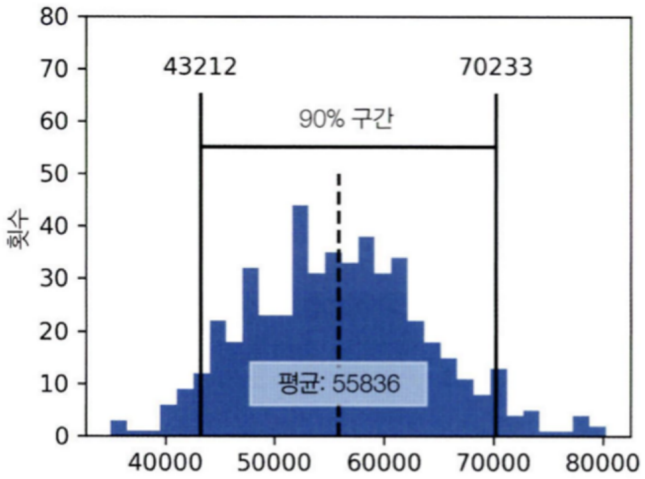


그림 2-9 20명 표본으로 구한 대출 신청자의 연간 소득에 대한 부트스트랩 신뢰구간

- 신뢰구간은 신뢰수준이 높을수록, 표본이 작을수록 구간이 넓어진다.
- 구간이 넓어진다 ⇒ 불확실성이 커진다