1회차

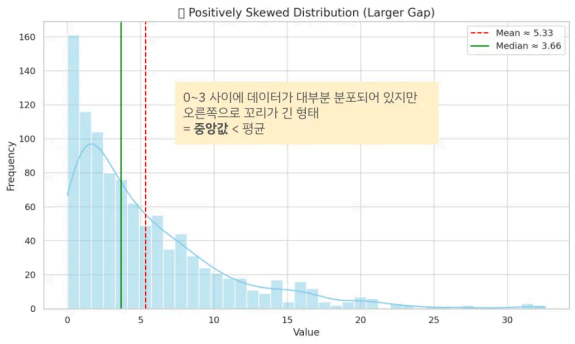
기술통계와 추론통계

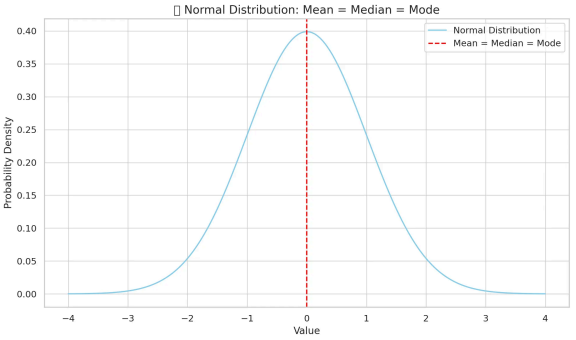
🖋 기술 통계 (Descriptive Statistics)

> **현재의 데이터를 요약하고 설명(기술)하는 통계" - 관찰된 데이터에 집중**

- 중심 경향치 : 평균, 중앙값, 최빈값

개념	설명	기호
평균(Mean)	모든 값을 더한 뒤 개수로 나눈 값	μ
중앙값(Median)	나열했을 때, 가운데 위치한 값	
최빈값(Mode)	가장 많이 나타나는 값	

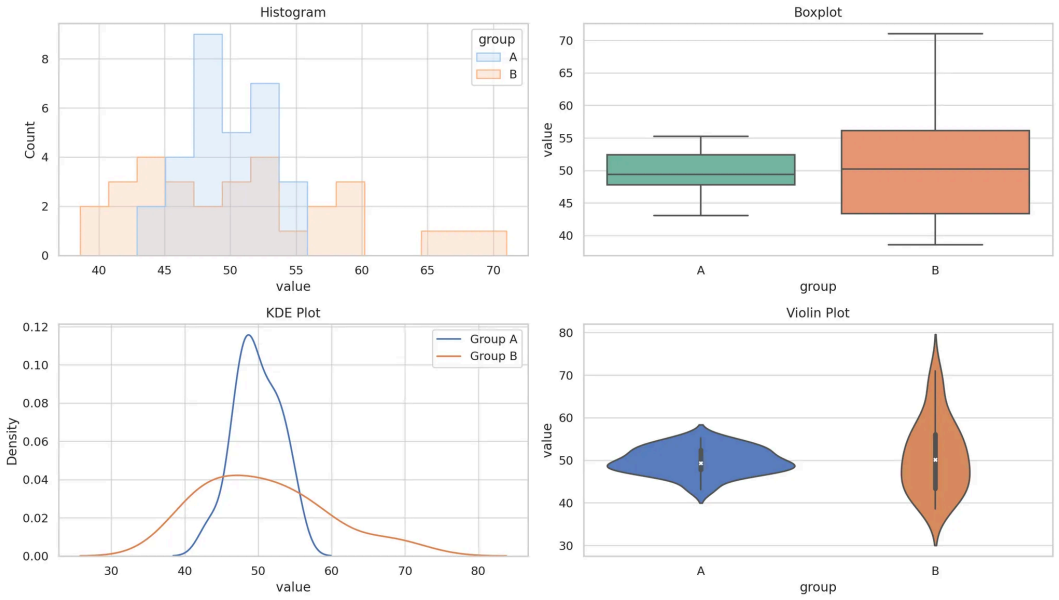




- 흩어진 정도 : 분산, 표준편차 (데이터가 중심에서 얼마나 흩어지고 퍼져 있는지 정도)

개념	설명	기호
편차(Deviation)	각 데이터가 평균에서 얼마나 떨어져 있는지	
분산(Variance)	편차를 제곱해서 평균낸 값	σ^2
표준편차(Standard Deviation)	분산에 루트를 씌운 값 (원래 단위로 복원)	σ

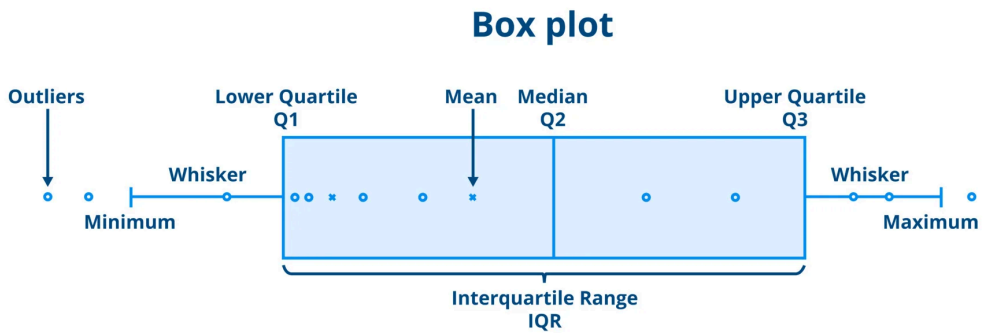
  - 예시
    - 데이터 A : 10, 10, 10, 10 → 평균 10, 표준편차 0
    - 데이터 B : 7, 8, 10, 15 → 평균 10, 표준편차 약 3.08
  - 분산을 확인할 수 있는 시각화
    - 히스토그램, 박스플랏, 밀도곡선, 바이올릿플랏 ...



- 분위수 (Quantile) : 데이터를 크기순으로 정렬했을 때, **특정 비율에 해당하는 위치의 값**
  - 일반적으로 p-분위수 는 아래에서부터 $p \times 100\%$ 위치에 있는 값

개념	설명
사분위수 (Quartiles)	데이터를 4등분 • 1사분위수(Q1, 25%) : 하위 25% 지점 값 • 2사분위수(Q2, 50%) : 중앙값 (Median) • 3사분위수(Q3, 75%) : 상위 25% 지점 값
백분위수 (Percentiles)	데이터를 100등분 • 90번째 백분위수(P90) = 전체 데이터 중 90%가 이 값 이하
십분위수 (Deciles)	데이터를 10등분

  - **IQR : Q3-Q1**



    - IQR 방식의 이상치 처리

☝🏻 하단 whisker보다 작거나 상단 whisker보다 큰 값을 제거

- whisker(하단) : $Q1 - 1.5 * IQR$
- whisker(상단) : $Q3 + 1.5 * IQR$

▼ 여기서 잠깐 Quiz~!

❓ Q1. 아래 리스트의 평균과 최빈값을 구해보세요.

[4,6,6,7,8]

Q2. 다음 표를 보고, IQR을 구해보세요.

	bmi
count	4909.00
mean	28.89
std	7.85
min	10.30
25%	23.50
50%	28.10
75%	33.10
max	97.60

기술통계 실습

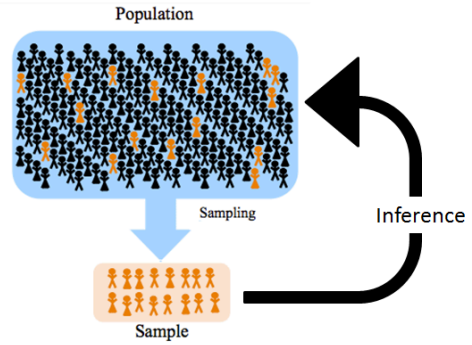
✅

- describe() 로 데이터 요약하기
- 히스토그램으로 데이터 분포 시각화
- IQR 방식으로 이상치 제거
- boxplot 그리기

🔍 추론 통계 (Inferential Statistics)

> **일부 데이터(표본)를 바탕으로 전체 모집단을 추정(예측)하거나, 어떤 주장이 맞는지 검정하는 통계**

- 기술 통계는 '있는 데이터를 요약'하고, **추론 통계는 '없는 모집단을 예측'한다**는 차이!! (ex. 여론조사)
- 모집단은 전부 관측할 수 없어서 표본을 추출하지만 이 **표본이 얼마나 신뢰할 수 있는 정보인지 추정**해야 함



추정과 가설검정

- 추정 (Estimation)
  - 모집단의 평균, 비율 등은 알 수 없기에 '표본'을 통해 '추정'
  - 값을 추정하거나 구간을 추정할 수 있음

개념	설명	예시
점추정	하나의 숫자로 모수 추정	모집단 평균은 약 65다
구간추정	신뢰 가능한 범위를 제시	모집단의 평균은 62~68 사이일 것이다

- 가설검정 (Hypothesis Testing)
  - 어떤 주장이 **우연**인지, 아니면 우연이 아닌지 확인하는 과정
    - 우연이 아니다 = 통계적으로 유의미하다

모집단과 표본

✅ **[Key point!]**
모집단을 추정한다 = 표본을 통해 모집단의 특성(평균/분산)을 추정한다.

- **표본평균과 표본분산**

모집단과 표본이 차이가 날텐데?

- **표본오차(sampling error)**

어떻게 추정이 가능할 수 있는거야?

- **중심극한정리!**

모집단의 특성을 알아보자!

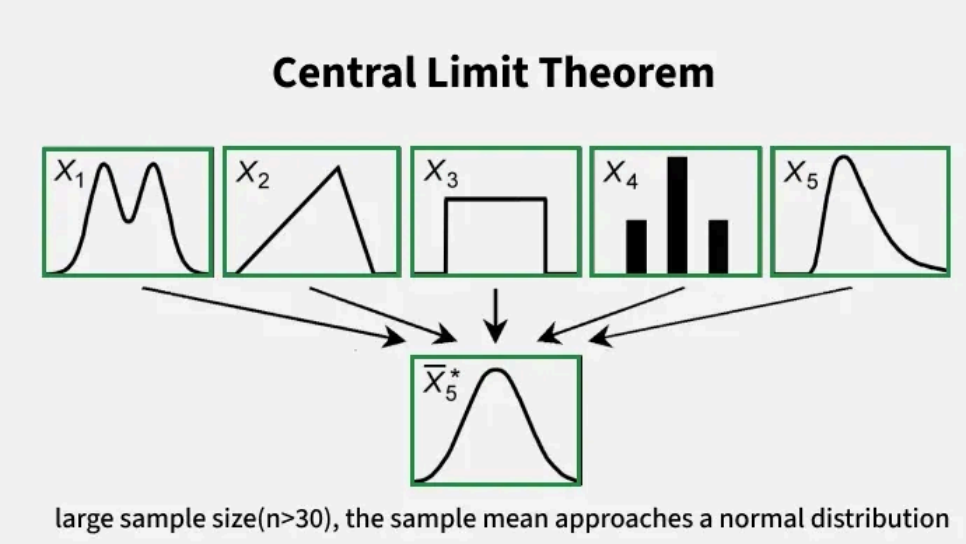
📌표본의 평균과 분산

- 우리는 표본의 평균($\bar{X}$)과 분산(s^2)으로 모집단의 평균과 분산을 추정할 수 있다.

  ▼ **why? 중심극한정리**

❓ 중심극한정리란 쉽게 말해 **표본평균은 정규분포를 따른다**는 정리입니다.

→ 그말은 즉슨 표본평균으로 모평균을 추정할 수 있다는 것!

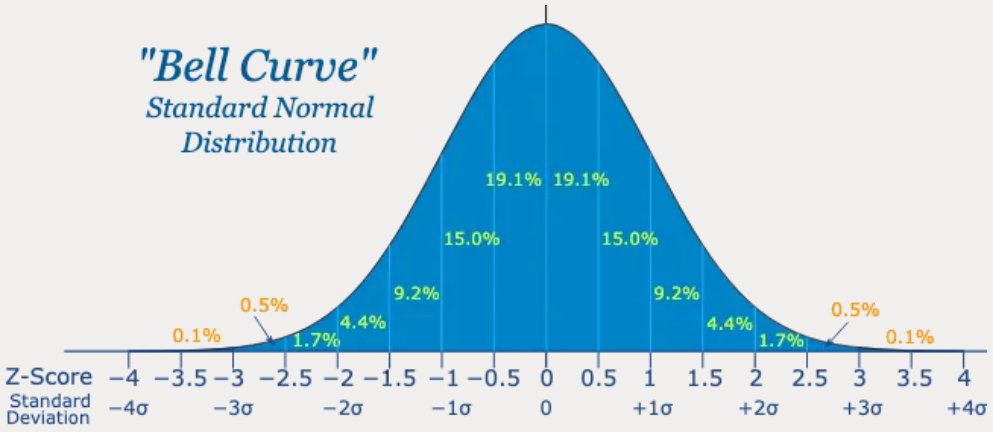


중심극한정리에 따르면, 표본평균은 평균이 모집단의 평균(μ) 이고 분산이 모집단의 분산/표본의 크기 ($\frac{\sigma^2}{n}$)인 정규분포를 따릅니다.

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

여기서 잠깐, 정규분포란?

: 아래와 같이 종모양을 갖는 분포를 말해요 (오늘은 여기까지만 알기!)

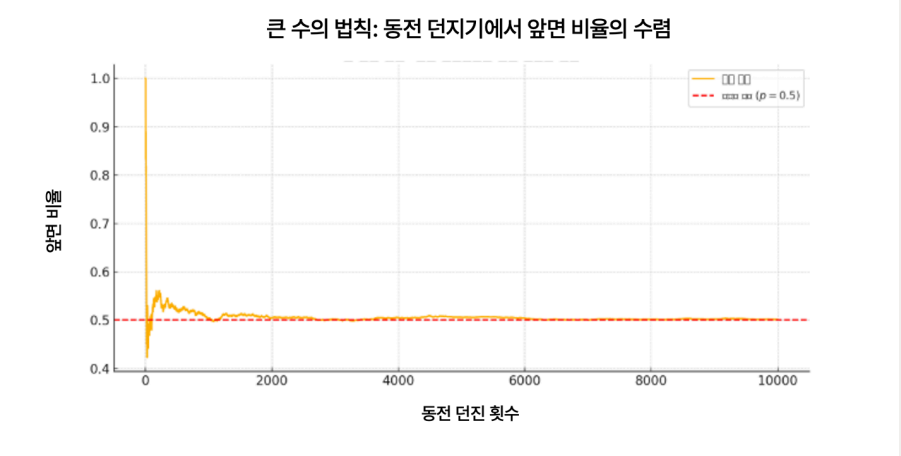


📌 표본오차(Sampling Error)

- **: 표본평균($\bar{x}$) - 모집단평균(μ)**
  - → 우리가 측정한 값과 실제 값(모집단)의 차이
- 표본을 무작위로 뽑는 과정에서 표본마다 평균이 다를 것이다. = 표본오차도 달라진다.
- 그렇다면 표본으로 모집단의 평균 (μ) 을 어떻게 알 수 있지?

▼ why? 큰 수의 법칙(Law of large numbers)

❓ 표본크기가 n이 커질수록 표본평균 $\bar{x}$가 모집단평균 μ에 한없이 가까워질 것이다.



⇒ 표본크기 n이 커질수록 표본평균은 모집단 평균에 가까워지고, 표본오차는 작아지는 경향을 갖는다.

📌 표준오차(Standard Error $\frac{s}{\sqrt{n}}$)

- **: sampling error의 표준편차= 표본평균의 표준편차**
- 중심극한 정리에 따르면, sampling error의 분산은 $\frac{\sigma^2}{n}$이다.
- 하지만 우리는 모집단의 분산(σ^2)을 모르니까 우리가 아는 표본의 분산(s^2)으로 대체한다 → $\frac{s^2}{n}$

뭐라는 건지 도통 모르겠다면, 바로 실습으로 넘어가봅시다 😇

추론통계 실습(1)

✅
- df.sample() 로 데이터 샘플링하기
- 샘플의 평균과 분산 구해서 전체 데이터와 비교하기
- 중심극한 정리
    - 샘플 여러개를 구해서 표본평균의 히스토그램 그려보기

다음 수업 안내

- 시간 : 10/24(수) 10시
- 내용
    - 가설검정
    - 신뢰구간과 p-value
    - 분포: 정규분포, t분포, F분포