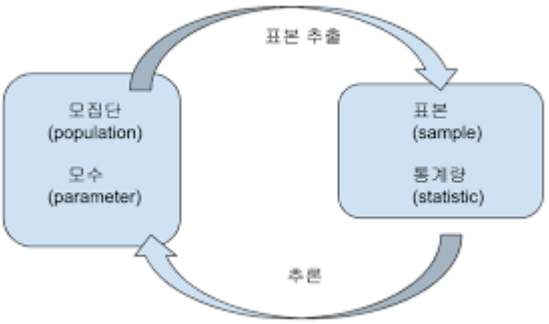
P. 11 ~

DataScience_S01_통계적 사고와 데이터 요약

BDAI 공식홈	https://bdaprogram.oopy.io/
Status	Statistic
문의	official.bdaa@gmail.com
선택	통계학



Outline

- 통계학이란?
- 모집단과 표본
- 자료의 수집
  - 수집방법
  - 표본조사
  - 실험
- 자료의 요약
  - 도표에 의한 요약
  - 수치에 의한 요약

학습목표

이 수업을 통해 아래의 학습목표를 이루고자 한다.

> 표본, 모집단, 변수의 개념을 이해한다.

> 다양한 통계량(평균,분산)등을 해석할 수 있다.

> 도표를 통해 자료를 요약하고, 기술통계로 분포를 요약할 수 있다.

통계학(Statistics)

어원과 역사적 배경

- **어원**: 통계학(Statistics)은 라틴어 *status*(국가, 정치 상태)에서 비롯되었으며, 본래는 *political state* 즉 "국가 운영과 관련된 수치 기록"을 의미했다.
- **고대 국가에서의 활용**
  - **이집트(BC 3000)**: 피라미드 건설에 필요한 인력 동원을 위해 인구조사가 실시되었다.
  - **이스라엘(BC 1500 전후)**: 구약성경 *민수기(Numbers)*에 기록된 인구조사에서 "사람의 수"라는 개념이 등장한다.
  - **인도(16세기)**: 행정 통계 조사 기록이 남아 있으며, 이는 국가 경영을 위한 자료 수집·활용의 초기 형태였다.
- **근대적 발전**: 18세기 말~19세기 초 유럽에서 '통계학'이 체계적인 학문으로 발전. 국가의 행정·재정 관리뿐 아니라 과학적 연구와 산업 혁신에 필수적인 방법론으로 자리 잡았다.

통계적 사고와 과학적 방법론

- **연역법(Deduction)**
  - 전제(공리) → 논리적 추론 → 결론.
  - 수학이나 논리학에서 흔히 사용되며, 불확실성보다는 확실성을 다룬다.
- **귀납법(Induction)**
  - 개별 사례(데이터, 관찰된 현상) → 일반 법칙 추론.
  - 통계학은 귀납적 논리에 기반하며, 불확실성을 다루는 체계적 방법론이다.

> 정의: 통계학은 “데이터와 같은 경험적 사실에 근거하여 불확실성을 수량화하고, 이를 통해 일반화 가능한 결론과 지식을 도출하는 귀납적 학문”이다.

불확실성의 계량화

- **통계 모형**: 확률(Probability)을 사용하여 현실 세계의 불확실성을 수학적으로 표현한다.
  - 질문: “관찰된 현상이 특정 원인 때문인가, 단순한 우연인가?”
  - 답변: 적절한 기준과 검정 절차를 통해 과학적으로 판단할 수 있다.
- **역사적 기여자**
  - **Karl Pearson** (19세기 말): 현대 통계학의 기초를 세운 인물로, 상관계수·χ^2 검정 등 도입.
  - **Ronald A. Fisher** (20세기 초): 최대가능도(MLE) 개념 도입, 실험계획법 (Design of Experiments) 정립, 추정 이론 발전.
  - **Jerzy Neyman**: 신뢰구간(Confidence Interval), 가설검정 이론 정립. “통계학은 모든 과학의 하인이다”라는 말을 남김.

통계학의 정의와 철학적 관점

- **실용적 정의**:
  - 연구 대상 집단으로부터 자료를 수집, 정리, 분석하여 **의사결정에 유용한 정보를 제공하는 학문**.
- **철학적 정의 (C. R. Rao)**:

> “모든 지식은 궁극적으로 역사이고, 모든 과학은 추상적으로 수학이며, 모든 판단은 그 근거에서 통계학이다.”

즉, 통계학은 과거 데이터를 기록하는 동시에, 미래 의사결정을 위한 근거를 제공하는 다리 역할을 한다.

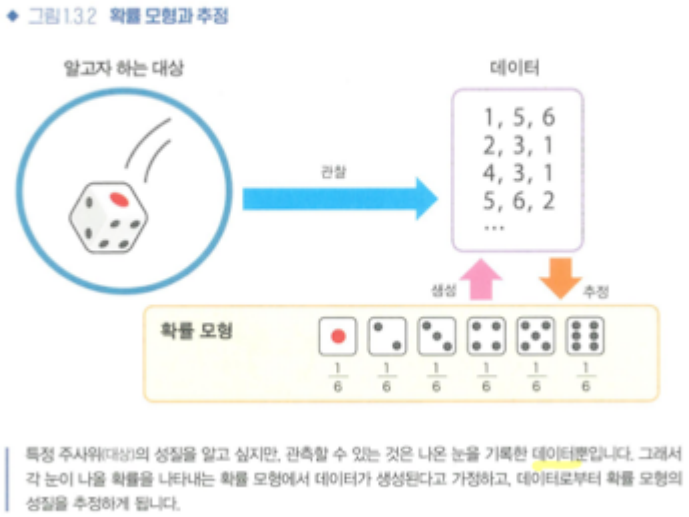
통계학의 두 축

1. **기술통계 (Descriptive Statistics)**
   - 자료의 **정리·요약**이 목적.

- 평균, 분산, 표준편차, 도표·그래프 등을 통해 데이터의 특징을 한눈에 파악.

2. **추론통계 (Inferential Statistics)**
   - 표본 데이터로부터 **모집단 특성을 추론**.
   - 확률분포와 추정, 가설검정 등을 통해 데이터의 발생 원인과 구조를 설명.

데이터가 비교적 단순한 확률 모형에서 생성되었다고 가정
예) 주사위는 각 눈이 확률 1/6로 나타나는 확률 모형을 이용



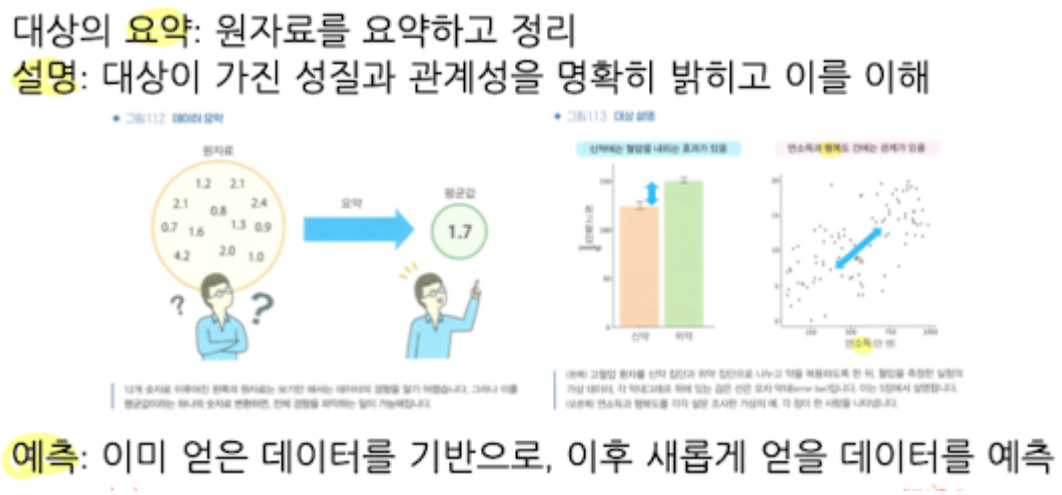
통계적 추론

점추정(point estimation): 데이터로부터 확률모형의 값을 점으로 추정

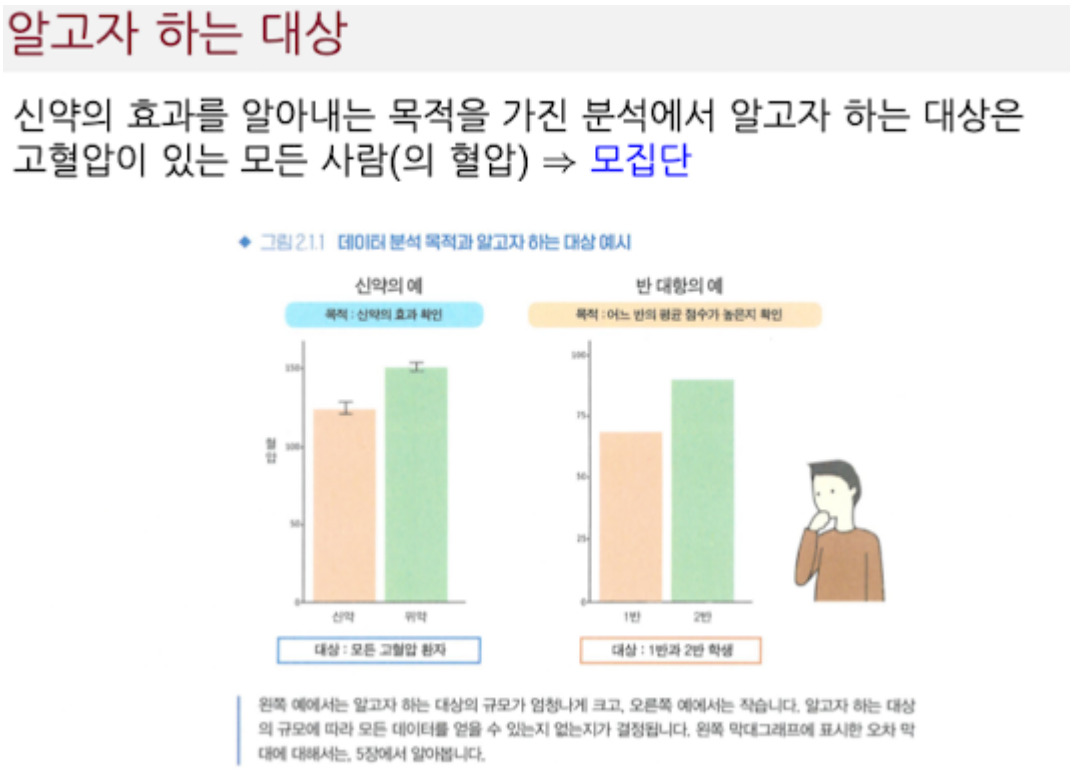
구간추정(interval estimation): 데이터로부터 확률모형의 값을 구간으로 추정

가설검정(statistical test): 세운 가설과 얻은 데이터가 얼마나 맞는지를 평가하여 가설을 채택할 것인가를 판단.

모집단과 표본

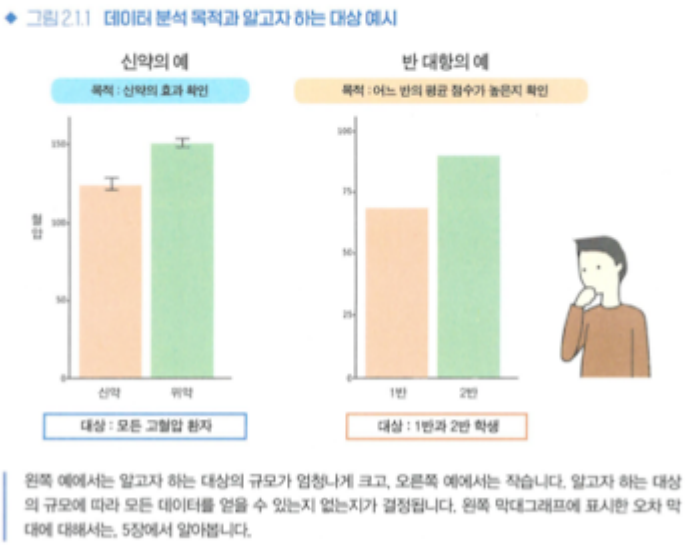


데이터 분석의 목적 설정 사례
신약의 효과 유무와 효과의 크기를 알고 싶다.
소득과 행복도 사이에 어떤 관계가 있는지 알고 싶다.
기온으로부터 올해 농작물 수확량을 예측하고 싶다.



알고자 하는 대상

신약의 효과를 알아내는 목적을 가진 분석에서 알고자 하는 대상은 고혈압이 있는 모든 사람(의 혈압) ⇒ 모집단



모집단의 크기

모집단 크기: 모집단에 포함된 요소(element)의 수

유한모집단: 모집단 중 한정된 요소만 포함한 것

무한모집단: 모집단 중 포함된 요소의 갯수가 무한한 것

자료의 수집

관측(Observation)

- **정의**: 연구자가 **대상에게 아무런 조치나 개입을 하지 않고**, 단순히 관찰을 통해 자료를 얻는 방법.
- **특징**:
  - 현실에서 발생하는 데이터를 있는 그대로 수집한다.
  - 변수 간 연관(association)은 알 수 있으나, 인과관계(causality)를 확정하기 어렵다.
  - 사회조사, 경제 데이터, 인구 센서스 등이 대표적 예시.

실험(Experiment)

- **정의**: 연구자가 **대상에게 특정 조건이나 처치를 가한 뒤**, 그 결과를 관측하는 방법.

- **특징**:
  - 연구 목적에 맞게 **통제(control)**, **무작위화(randomization)**, **반복(replication)** 등을 활용한다.
  - 인과관계(causality)를 파악할 수 있는 가장 강력한 방법이다.
  - 의학 임상시험(Clinical Trial), 교육 효과 검증 등이 대표적 예시.

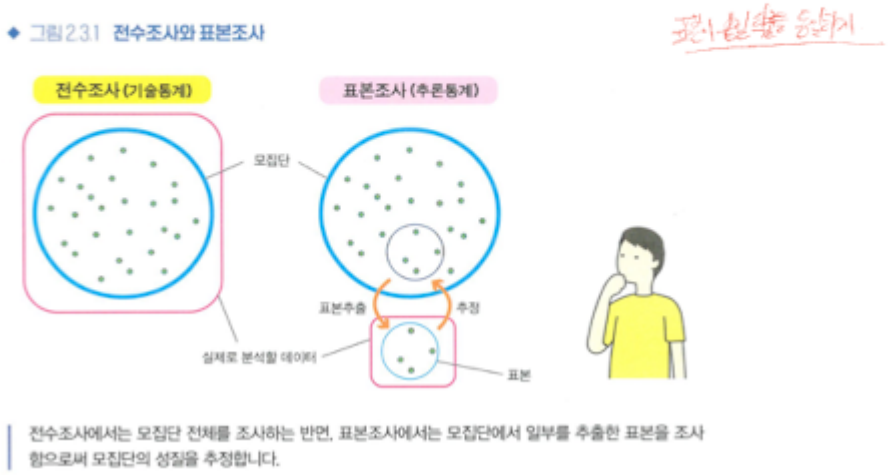
> 연구 질문: 유아기의 영어 조기교육이 중학생 시기의 영어 성취도에 영향을 미치는가?

- **관측연구(Observation Study)**
  - 중학생 100명을 대상으로, 유아기에 영어 조기교육을 받았는지 여부와 현재 영어 성취도를 조사.
  - 연구자는 **개입하지 않고 단순히 자료를 수집**했으므로, 이는 **관측자료**에 해당한다.
  - 결론: “조기교육과 성취도 간에 연관성은 확인 가능하나, 인과관계 확정은 불가능.”
- **실험연구(Experimental Study)**
  - 유아 100명을 무작위로 두 집단으로 나눔.
    - 50명: 영어 조기교육 실시
    - 50명: 영어 조기교육 미실시
  - 이들이 중학생이 되었을 때 영어 성취도를 측정하여 비교.
  - 연구자가 의도적으로 처치(조기교육 여부)를 부여했으므로, 이는 **실험자료**에 해당한다.
  - 결론: 인과관계를 상대적으로 강하게 입증할 수 있다.

변수의 구분

- **설명변수 (Explanatory Variable, 독립변수)**
  - 실험이나 관측에서 **결과에 영향을 줄 수 있는 변수**
  - 예: 영어 조기교육 여부(받음/안 받음)
- **반응변수 (Response Variable, 종속변수)**
  - 설명변수의 변화에 따라 그 값이 결정되는 변수
  - 예: 중학생 시기의 영어 성취도

전수조사 (센서스 census): 모집단 전체를 대상으로 조사
표본조사 (sample survey): 모집단 전체의 일부를 표본(sample)으로 뽑아 여기에 속한 개체의 특성을 나타내 주는 변수를 관측하고 얻어진 자료에 근거하여 모집단 전체에 대한 추론을 진행



표본조사 관련 용어
모집단(population): 우리가 정보를 원하는 대상이 되는 모든 사람, 가구, 사업체 등

개체(unit): 모집단의 개별요소

표본(sample): 모집단에 관한 정보를 얻기 위하여 사용되는 모집단의 일부

표본추출틀(sampling frame): 표본이 추출될 개체들의 목록

변수(variable): 개체의 특성(표본 내 개체들에 대하여 변수가 측정)

모수(parameter): 모집단의 특성을 나타내는 특성치

표본추출기법 (Sampling Methods)

단순랜덤표본추출 (Simple Random Sampling, SRS)

- **정의**: 모집단의 크기를 N, 표본의 크기를 n이라 할 때, N개 중 모든 가능한 n개의 조합이 **동일한 확률**로 선택되는 방식.
- **절차**:
  1. 모집단을 완전한 표본추출틀(sampling frame)에 등록

2. 난수표(random number table)나 난수 생성기를 이용하여 nnn개의 단위를 무작위 추출

- **장점**: 편향이 적고, 통계적 이론이 가장 잘 적용됨
- **단점**: 모집단 리스트가 반드시 필요하며, 비용·시간이 많이 들 수 있음
- **예제**: 베트남전 당시(1970년) 미국의 징병제 → 366일(윤년 포함)을 무작위 추출하여 생일에 따라 징집

층화표본추출 (Stratified Sampling)

- **정의**: 모집단을 동질적인 특성을 가진 층(stratum)으로 나눈 후, 각 층에서 단순랜덤추출 실시
- **조건**: 층 내부는 동질적, 층 간에는 이질적일수록 효율적
- **장점**: 표본의 대표성 ↑, 분산 감소, 소규모 층의 의견도 반영 가능
- **단점**: 층 구분 기준이 명확해야 하고, 사전 정보가 필요
- **예제**: 통계청의 경제활동인구조사 → 7개 도시와 9개 도를 층화, 도는 다시 동부(도시)·읍면부(비도시)로 층화

군집표본추출 (Cluster Sampling)

- **정의**: 모집단을 **집단(cluster)** 단위로 나누고, 일부 집단을 무작위로 뽑아 그 집단에 속한 모든 개체를 조사하는 방식
- **장점**: 비용과 시간이 절약됨, 표본추출틀이 불완전해도 적용 가능
- **단점**: 집단 내부가 이질적일 경우 표본이 대표성을 잃을 수 있음
- **예제**: 서울시 가구조사 → 구 → 동 → 반을 무작위 추출하여 해당 반의 모든 가구를 조사 (다단계추출, multistage sampling)

계통표본추출 (Systematic Sampling)

- **정의**: 모집단을 일정한 간격 k로 나누고, 시작점을 무작위로 정한 후 매 k번째 단위를 표본으로 선택$k = \frac{N}{n}$
- **장점**: 절차가 간단, 현장에서 적용 용이
- **단점**: 모집단에 주기성이 있으면 표본이 왜곡될 수 있음
- **예제**: 총선 출구조사에서 1,000명 중 50명을 조사할 때, 간격 $k = 20$. 무작위 시작점 7 → 7, 27, 47, ...

특수 기법: 워너의 랜덤화 방법 (Warner’s Randomized Response Technique)

- **목적**: 민감한 질문(예: 마약 경험)에 대한 **응답 편향(response bias)** 감소
- **방법**: 두 개 질문 준비 → 동전 던지기 → 앞면 나오면 Q1, 뒷면 나오면 Q2에 '예/아니오' 응답 → 조사자는 어떤 질문에 답했는지 알 수 없음
- **효과**: 개인의 프라이버시를 보호하면서 집단 차원에서 통계적 추정 가능

실험 설계 (Experimental Design)

기본 개념

- **실험단위 (Experimental Unit)**: 실험이 수행되는 개체
- **처리 (Treatment)**: 실험단위에 적용되는 조건
- **설명변수 (Explanatory Variable)**: 처리 조건(독립변수)
- **반응변수 (Response Variable)**: 측정 대상 결과(종속변수)

실험계획의 원리 (Principles of Experimental Design)

1. **비교(Comparison)**: 처리집단과 대조집단을 비교
2. **랜덤화(Randomization)**: 실험단위를 무작위로 배정
3. **블록화(Block)**: 혼란변수(confounding variable)를 통제하기 위해 동질적인 집단으로 나눔
4. **반복(Replication)**: 결과의 신뢰성 확보

실험설계 방법

- **완전임의화설계 (Completely Randomized Design)**
  - 모든 실험단위를 처리집단에 무작위 배정
- **임의화블록설계 (Randomized Block Design)**
  - 블록(동질 집단)을 만든 뒤, 블록 내에서 무작위 배정
- **대응짝설계 (Matched Pair Design)**
  - 유사한 두 단위를 짝지어 처리·대조를 비교 (예: 한 사람의 좌우 발)

자료의 종류와 요약

자료(Data)의 기본 구분

범주형 자료 (Categorical Data)

DB: 통합, 저장, 공유, 운영되는 데이터

기업 → 기업 데이터 / - 실시간 접근성 / 내용기반

- **정의**: 자료가 취할 수 있는 값이 **한정된 범주(category)**로만 표현되는 경우.
- **종류**:
  - **명목형 (Nominal Scale)**
    - 범주 간 **순서 없음**.
    - 예: 성별(남/여), 혈액형(A/B/AB/O), 국적
  - **순서형 (Ordinal Scale)**
    - 범주 간 **순서 존재**, 하지만 범주 간 **간격의 크기를 정량적으로 비교 불가**.
    - 예: 차량 크기(소형/중형/대형), 만족도(매우 불만족 → 매우 만족)

양적 자료 (Quantitative Data, 수치형 자료)

- **정의**: 수치로 측정되고, 산술연산이 가능한 자료.
- **연속형 (Continuous)**
  - 특정 구간 내에서 **이론적으로 무한한 값**을 가질 수 있음.
  - 예: 키, 몸무게, 온도, 시간
- **이산형 (Discrete)**
  - 값이 **셀 수 있는 개수**로만 표현되는 경우.
  - 예: 자녀 수(0,1,2,...), 사고 발생 횟수, 주사위 눈금

+ 사사횟수

척도 수준 (Scale of Measurement)

- **구간형 (Interval Scale)**
  - 값들 사이의 **차이(difference)**는 의미 있지만, **비율(ratio)**은 의미 없음.
  - **절대적 0 없음**.
  - 예: 섭씨 온도(20도와 30도 차이는 의미 있지만, 30도가 20도의 1.5배라는 해석은 불가능), 연도
- **비율형 (Ratio Scale)**
  - **0이 의미 있는 절대적 기준**을 가짐 → 비율 비교 가능.
  - 예: 몸무게, 길이, 소득, 시간

도표에 의한 자료 요약

자료의 종류에 따라 적절한 **시각화 방법**을 선택하는 것이 중요하다.

1. **도수분포표 (Frequency Table)**
    - 범주형 또는 이산형 자료를 요약할 때 사용
    - 각 범주별 빈도(도수)와 상대도수(%) 제시
2. **원형그래프 (Pie Chart)**
    - 전체에 대한 **비율(%)** 표현
    - 범주형 자료 요약에 적합
3. **막대그래프 (Bar Chart)**
    - 범주형 자료의 비교 시 사용
    - 범주의 순서가 있을 수도(ordinal) 없을 수도(nominal) 있음
4. **히스토그램 (Histogram)**
    - 연속형 자료의 분포를 구간(bin)으로 나누어 표현
    - 막대 사이에 간격 없음 (연속형의 특징 반영)
5. **줄기-잎 그림 (Stem-and-Leaf Plot)**
    - 데이터의 원자료(raw data) 형태를 유지하면서 분포 요약
    - 작은 표본에 적합
6. **산점도 (Scatter Plot)**
    - 두 개의 양적 변수 간 관계를 시각화
    - 패턴, 상관관계, 이상치(outlier) 확인 가능

도수분포표(frequency distribution plot)

자료들을 특성에 따라 몇 개의 항목 또는 계급으로 분류하여 빈도수를 정리한 것.

범주형 자료(서열자료), 이산형 수치자료, 연속형 수치자료의 도수분포표

범죄 피해에 대한 두려움

두려움	비율(%)
매우느낌	15.6
약간느낌	41
보통이다	26.1
별로느끼지못함	13.8
전혀느끼지못함	3.5
합계	100

통계청, 한국의 사회지표 12-4

자동차 보유대수 별 가구 수

자동차수	가구 수	비율(%)
0	6,317,367	36.4
1	8,344,406	48.1
2	2,399,445	13.8
3대이상	280,748	1.6
합계	17,341,966	100

통계청, 인구총조사

연령별 실업자의 수

연 령	2014년 실업자 수	비율(%)
15-19세	25	2.7
20-24세	154	16.4
25-29세	206	22.0
30-34세	105	11.2
35-39세	79	8.4
40-44세	75	8.0
45-49세	78	8.3
50-54세	69	7.4
55-59세	65	7.0
60세이상	81	8.6
합 계	937	100

통계청, 경제활동인구조사

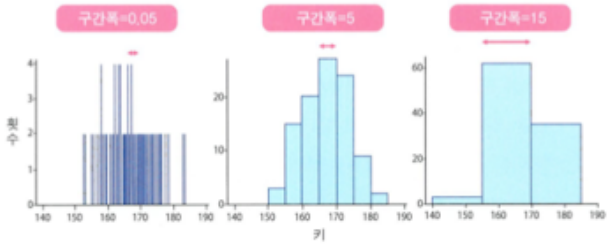
히스토그램 (histogram)

도수분포표를 그래프로 표현한 것으로 가로축은 도수분포표의 계급구간, 세로축은 빈도(도수)

특성

1. 양적 자료에만 적용되며, 각각의 막대는 연속적이어서 빈 간격이 없음.
2. 가로축의 동일한 간격은 동일한 크기를 나타냄.
3. 막대그래프는 높이가 자료의 양을 나타내지만, 히스토그램은 면적이 자료의 양을 나타냄.

◆ 그림 3.22 서로 다른 구간폭으로 나타낸 히스토그램

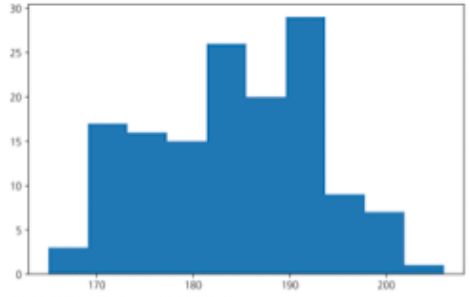


같은 데이터라도 구간폭에 따라 주는 인상이 다름

> 구간폭이 너무 좁으면 그림이 지나치게 세세하게 그려지므로, 어떤 형태의 분포인지 읽기 어렵습니다. 한편, 구간폭이 너무 넓으면 분포 형태 정보가 사라집니다. 그러므로 적절한 구간폭을 설정해야만 합니다. 통계 소프트웨어를 사용하면 통상 자동으로 구간폭을 정해 히스토그램을 그려 줍니다.

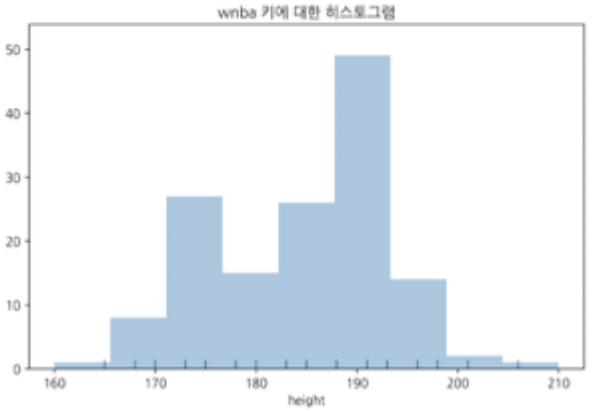
히스토그램

```
plt.hist(wnba['Height'], bins = 10)
plt.savefig("hist_height.png")
plt.show()
```



히스토그램

```
bins = np.linspace(160, 210, 10)
sns.distplot(wnba['Height'], rug = True, kde = False, bins = bins)
plt.title('wnba 키에 대한 히스토그램')
plt.xlabel('height')
plt.show()
```



줄기-잎 그림 (stem and leaf plot)

원자료를 그대로 볼 수 있게 해 주면서 막대그래프와 같은 시각적 효과를 갖는 요약기법

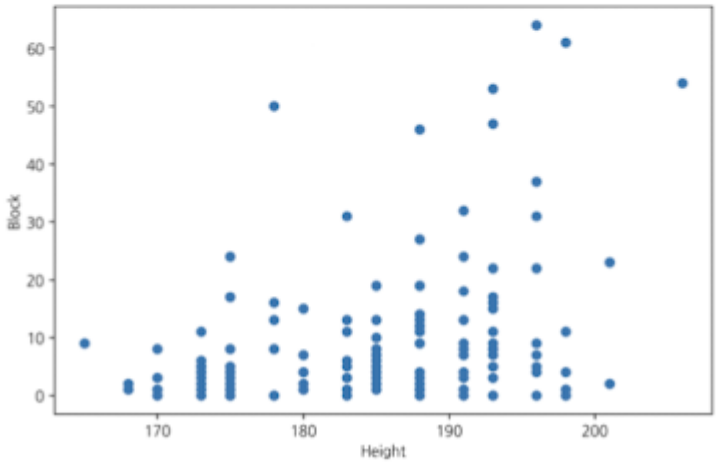
자료를 요약하면서도 정보의 손실이 없음. 그러나 자료의 수가 많거나 흩어진 정도가 클 때는 자료의 파악이 어려움.

```
import stemgraphic
stemgraphic.stem_graphic(wnba['Height'], scale = 10)
plt.show()
```

산점도 (scatter plot)

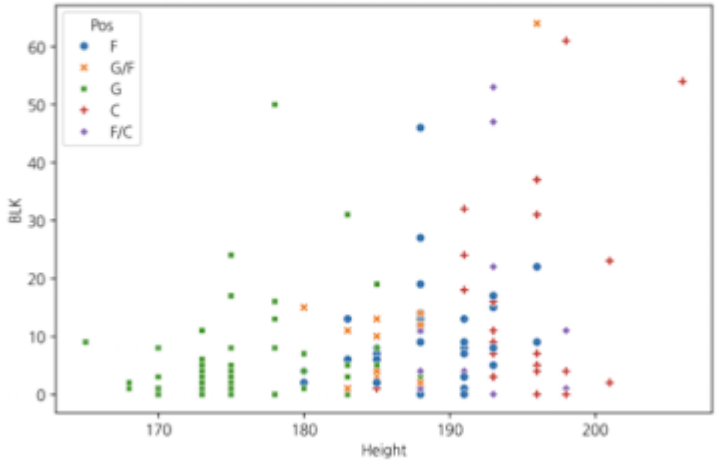
두 수치형 변수간의 관계를 나타내기 위해 사용

```
plt.scatter(wnba['Height'], wnba['BLK'])
plt.xlabel('Height')
plt.ylabel('Block')
plt.show()
```



산점도

```
import seaborn as sns
sns.scatterplot(x = 'Height', y = 'BLK', hue = 'Pos', style = 'Pos', data = wnba)
plt.show()
```



수치에 의한 자료 요약 (Numerical Summaries of Data)

중심 경향치 (Measures of Central Tendency)

(1) 평균 (Mean, Average)

- **정의**: 모든 관측치의 합을 개수로 나눈 값.
- **공식**:
  - 표본평균(sample mean):$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - 모평균(population mean):$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$
- **특징**:
  - 계산이 쉽고 가장 널리 사용됨.
  - *극단값(outlier)**에 크게 영향을 받음.
- **예제**: 시험 점수가 (50, 70, 80, 100)일 때 평균 = (50+70+80+100)/4 = 75.

---

(2) 중위수 (Median)

- **정의**: 모든 관측치를 크기순으로 배열했을 때 **가운데 위치하는 값**.
- **계산법**:
  - n이 홀수일 때: 가운데 위치한 값
  - n이 짝수일 때: 가운데 두 값의 평균
- **특징**:
  - 데이터의 “순서 중심”을 나타냄.
  - 극단값의 영향을 거의 받지 않음.
- **예제**: (10, 20, 90) → 중위수 = 20 / (10, 20, 90, 100) → 중위수 = (20+90)/2 = 55.

---

(3) 최빈값 (Mode)

- **정의**: 가장 자주 나타나는 값.
- **특징**:
  - 범주형 자료(categorical data) 분석에 자주 사용됨.
  - 여러 개의 최빈값이 존재할 수 있음 (이중최빈, 다중최빈).
  - 데이터 변화에 민감.
- **예제**: (2, 3, 3, 5, 7, 7, 7, 8) → 최빈값 = 7.

산포도 (Measures of Dispersion)

(4) 범위 (Range)와 백분위수 (Percentile)

- **범위**: 최대값 – 최소값
  - 간단하지만, 극단값에 민감
- **백분위수**: 데이터의 분포를 100등분한 값
  - p번째 백분위수 P_p: 전체 데이터 중 p%가 그 이하에 위치
  - 예: 25% 백분위수(Q1), 50% 백분위수(중위수), 75% 백분위수(Q3)

---

(5) 사분위수 범위 (Interquartile Range, IQR)

- **정의**: Q3 – Q1
- **특징**: 극단값의 영향을 줄인 변동 척도
- **활용**: 이상치(outlier) 탐지에 자주 사용
  - 통상적으로 $[Q1 - 1.5IQR, Q3 + 1.5IQR]$바깥의 값은 이상치로 간주

---

(6) 다섯 숫자 요약 (Five-number summary)와 상자그림 (Boxplot)

- **다섯 숫자 요약**: 최소값, Q1, Q2(중위수), Q3, 최대값
- **상자그림(Boxplot)**: 다섯 숫자를 시각적으로 표현한 그래프
  - 박스: Q1 ~ Q3
  - 선(수염, whisker): 최소값과 최대값
  - 점: 이상치(outlier)

---

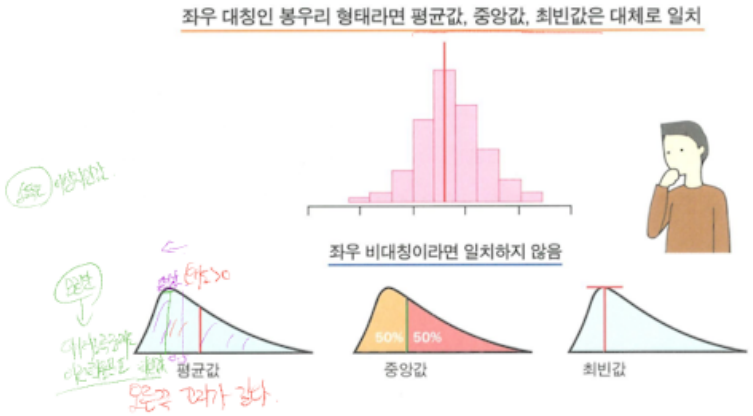
(7) 분산 (Variance)과 표준편차 (Standard Deviation)

- **분산 (Variance)**:$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - 데이터가 평균으로부터 얼마나 퍼져 있는지를 제곱해 평균한 값
- **표준편차 (Standard Deviation)**:$s = \sqrt{s^2}$
  - 분산의 제곱근 → 데이터와 동일한 단위로 표현
- **특징**:
  - 데이터 변동성의 대표적인 척도
  - 극단값에 민감
- **예제**: 데이터 (2, 4, 4, 4, 5, 5, 7, 9)
  - 평균 = 5

- 분산 = 4
- 표준편차 = 2

분포형태와 대푯값

◆ 그림 3.3.3 분포 형태와 대푯값



평균, 중위수, 최빈값

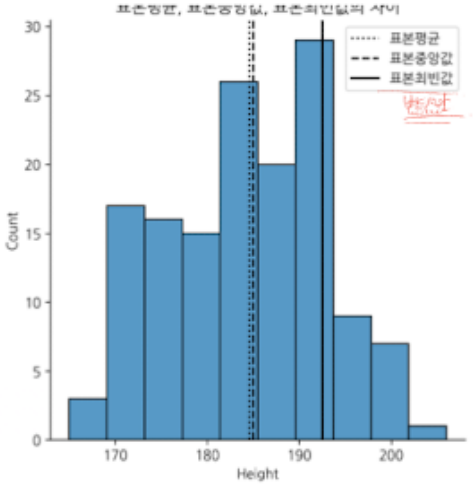
```
bins = np.linspace(160, 210, 11)
ns, _ = np.histogram(wnba['Height'])
mode_index = np.argmax(ns)

sample_mean = np.mean(wnba['Height'])
sample_median = np.median(wnba['Height'])
sample_mode = 0.5*(bins[mode_index]+bins[mode_index+1])
print(sample_mode)
```

```
192.5
```

```
fig = plt.figure(figsize = (3,2))
sns.displot(wnba['Height'], bins = 10)
plt.axvline(sample_mean, c= 'k', ls=':', label='표본평균')
plt.axvline(sample_median, c = 'k', ls = '--', label='표본중앙값')
plt.axvline(sample_mode, c = 'k', ls='-', label='표본최빈값')
plt.title('표본평균, 표본중앙값, 표본최빈값의 차이')
plt.xlabel('Height')
plt.legend()
plt.savefig("기술통계3.png")
plt.show()
```

평균, 중위수, 최빈값



퍼진 정도의 측정 (Measures of Dispersion)

1. 왜 필요한가?

- 자료의 중심(평균, 중위수, 최빈값)만으로는 **분포의 전체적인 특성**을 파악하기 어렵다.
- 예:
  - 자료 1: (6, 7, 8, 9, 10)
  - 자료 2: (1, 3, 8, 12, 16)

    → 두 자료 모두 평균 = 8, 중위수 = 8

    → 하지만 **퍼진 정도(variability)**가 다르므로 분포 형태가 전혀 다름.
- 따라서 분산, 표준편차, 범위, IQR 등을 통해 **데이터의 산포(variability)**를 측정해야 한다.

---

2. 위치 기반 측정치

(1) 범위 (Range)

- **정의**: 자료에서 최대값과 최소값의 차이$R = \max(x_i) - \min(x_i)$
- **특징**: 계산이 간단하나, 극단값(outlier)에 민감

---

(2) 백분위수 (Percentile)

- **정의**: 데이터를 크기순으로 배열했을 때, 전체 데이터 중 $p\%$가 그 값 이하에 위치하는 수치
- **예**: 70번째 백분위수(P70) = 데이터 중 70%가 그 값 이하

- **활용**: 성적, 시험점수, 소득분포 등 상대적 위치를 판단할 때 사용

---

(3) 사분위수 (Quartiles)

- **Q1 (제1사분위수)**: 하위 25% 위치 (25th percentile)
- **Q2 (제2사분위수)**: 중앙값 (50th percentile)
- **Q3 (제3사분위수)**: 상위 25% 위치 (75th percentile)
- **사분위수 범위 (IQR)**:IQR=Q3−Q1

  $IQR = Q3 - Q1$

  → 데이터의 중간 50% 범위, 극단값에 덜 민감

---

3. 평균과의 차이에 의한 측정치

(4) 분산 (Variance)

- **정의**: 각 관측치가 평균으로부터 떨어진 정도를 제곱해 평균한 값
- **공식**:
  - 모집단 분산:$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
  - 표본 분산:$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- **특징**: 단위가 제곱으로 변해 직관성이 떨어짐

---

(5) 표준편차 (Standard Deviation)

- **정의**: 분산의 제곱근$s = \sqrt{s^2}$
- **특징**: 데이터와 동일한 단위 → 직관적 해석 가능
- **의미**: 표본의 대부분 데이터는 $\bar{x} \pm 2s$ 범위 내에 존재 (정규분포 가정 시 약 95%)

---

(6) 변동계수 (Coefficient of Variation, CV)

- **정의**: 표준편차를 평균으로 나눈 값$CV = \frac{s}{\bar{x}}$
- **특징**:
  - 단위(scale)에 의존하지 않음 (무차원 지표)
  - 평균 대비 상대적 산포를 비교할 때 유용
- **예제**: 두 반의 시험 결과 평균이 다를 때, 단순 표준편차 대신 CV를 사용하면 **상대적 변동성** 비교 가능

---

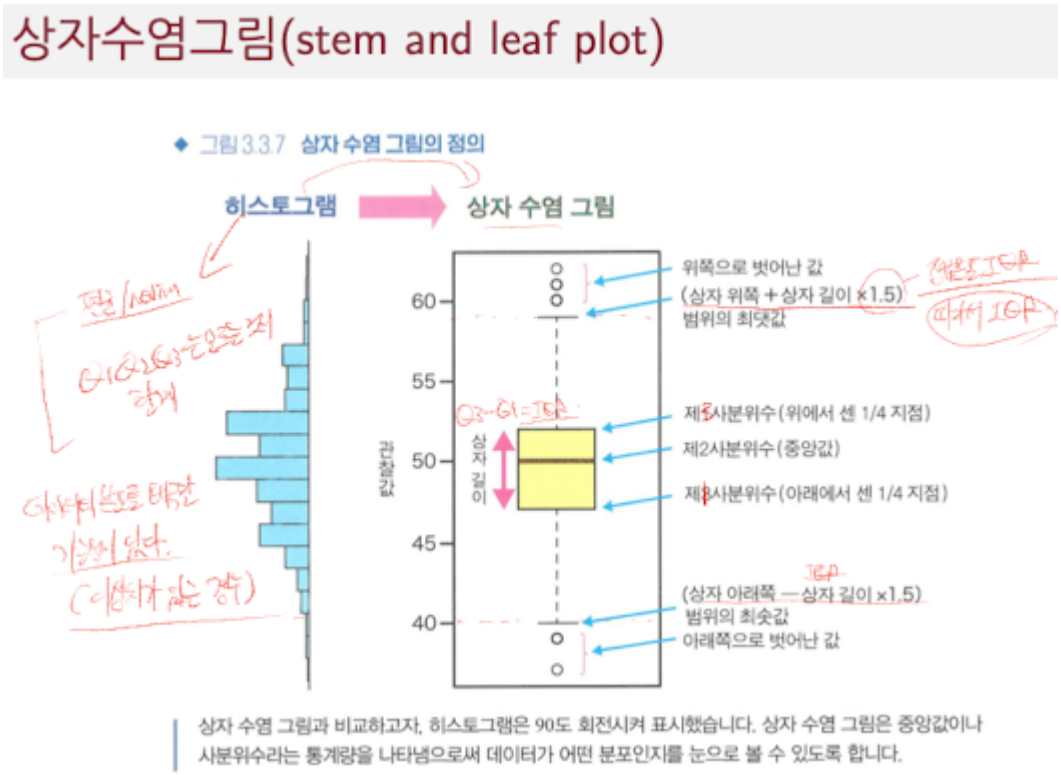
4. 시각화 방법

(1) 다섯 숫자 요약 (Five-number summary)

- 최소값, Q1, Q2(중위수), Q3, 최대값

(2) 상자그림 (Boxplot)

- Q1~Q3 범위를 박스로 표시, Q2는 선으로 표시
- 수염(whisker)은 최소값~최대값 표시
- 이상치는 점으로 별도 표시
- **특징**: 데이터의 분포·중심·산포·이상치를 한눈에 확인 가능



다섯숫자 요약 및 상자그림 예제

```
np.percentile(wnba['Height'], [25,50,75])
```

```
array([176.5, 185. , 191. ])
```

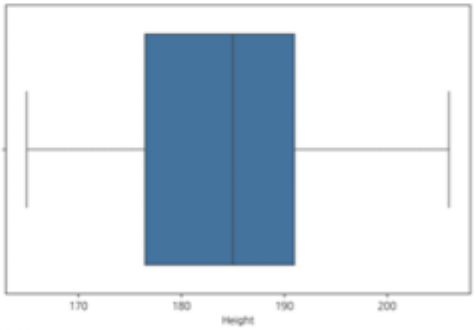
```
np.min(wnba['Height'])
```

```
165
```

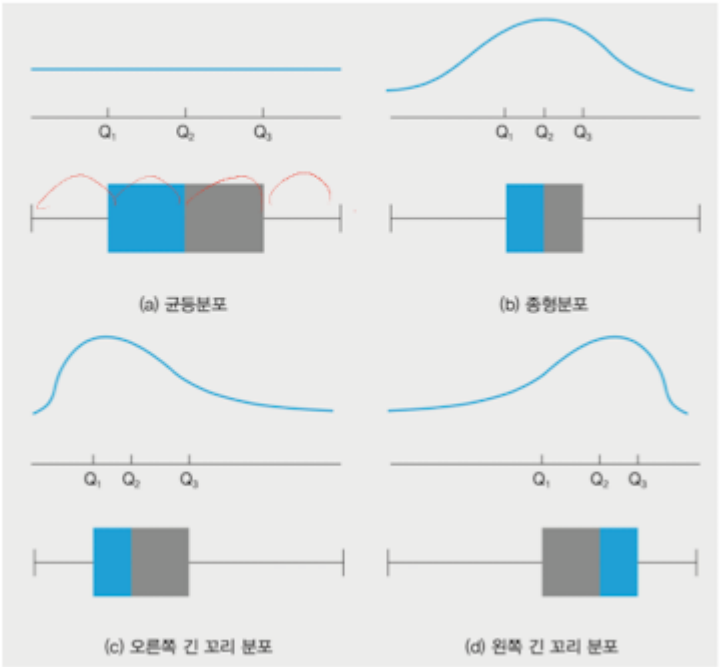
```
np.max(wnba['Height'])
```

```
206
```

```
sns.boxplot(x='Height', data = wnba)
```



분포의 형태와 상자그림



분산과 표준편차, 그 외의 요약치

1. 분산 (Variance)

- **정의**: 각 관측치가 평균으로부터 얼마나 떨어져 있는지를 나타내는 값.

- **편차 (Deviation)**:$d_i = x_i - \bar{x}$

  → 개별 자료값과 평균의 차이

- **표본분산 (Sample Variance)**:$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
  - 분모에 n−1n-1n−1을 사용하는 이유: 불편추정량(unbiased estimator)을 얻기 위함.
- **모분산 (Population Variance)**:$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
- **특징**:
  - 단위가 제곱(squared unit)이라 직관적 해석은 어렵지만, 데이터 흩어짐 정도를 수학적으로 명확히 표현한다.

---

2. 표준편차 (Standard Deviation)

- **정의**: 분산의 제곱근.$s = \sqrt{s^2}, \quad \sigma = \sqrt{\sigma^2}$
- **특징**:
  - 데이터와 동일한 단위로 표현 → 직관적 해석 가능.
  - 값이 작으면 데이터가 평균 주변에 모여 있고, 크면 퍼져 있음.
  - 정규분포 가정 시, 약 68%의 데이터가 $\bar{x} \pm s$, 약 95%가 $\bar{x} \pm 2s$범위 내에 존재.
- **예제**: 데이터 (2, 4, 4, 4, 5, 5, 7, 9)
  - 평균 = 5
  - 분산 = 4
  - 표준편차 = 2

---

3. 왜도 (Skewness)

- **정의**: 분포의 **비대칭 정도**를 측정.$Skewness = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n \cdot s^3}$
- **해석**:
  - >0: 오른쪽 꼬리 길다 (right-skewed, 양의 왜도)
  - =0: 대칭 분포 (정규분포와 유사)
  - <0: 왼쪽 꼬리 길다 (left-skewed, 음의 왜도)
- **예시**:
  - 소득 분포: 대체로 오른쪽 꼬리(고소득자) 길어 양의 왜도

---

4. 첨도 (Kurtosis)

- **정의**: 분포의 **뾰족한 정도(peakness)와 꼬리 두께**를 측정.$Kurtosis = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot s^4}$
- **해석**:
  - 3 기준(정규분포):
    - 3 → 뾰족, 꼬리 두꺼움 (leptokurtic)
    - = 3 → 정규분포와 유사 (mesokurtic)
    - < 3 → 평평, 꼬리 얇음 (platykurtic)
- **의의**: 극단값(outlier) 발생 가능성과 분포 형태를 판단하는 지표

---

5. 표준점수 (Standard Score, Z-score)

- **정의**: 각 관측값이 평균으로부터 몇 표준편차 떨어져 있는지 나타내는 값.$z_i = \frac{x_i - \bar{x}}{s}$
- **특징**:
  - 단위 없음 (dimensionless)
  - 표준화된 데이터: 평균 0, 표준편차 1
  - 다른 분포 간 비교 가능 (예: 국어 점수 80점, 수학 점수 90점 → 어떤 과목 성적이 상대적으로 우수한가?)
- **예제**:
  - 시험 평균 = 70, 표준편차 = 10
  - 학생 점수 = 85 → Z = (85-70)/10 = 1.5 → 평균보다 1.5 표준편차 위