**1주차 강의 노트: 기초 통계 개념과 데이터 유형**

**강의 목표**

* 기초 통계 개념을 이해하고, 이를 활용하여 데이터를 요약하고 해석하는 방법을 학습합니다.
* 다양한 데이터 유형을 이해하고, 각 유형에 맞는 적절한 분석 방법을 선택하는 능력을 배양합니다.

1. **통계학의 기본 개념**
2. **통계학의 정의와 역할**
   * **통계학:** 데이터를 수집, 분석, 해석하여 유용한 정보를 도출하는 학문
   * **역할:** 데이터에 기반한 의사결정, 패턴 발견, 미래 예측 등에 사용
3. **기술 통계와 추론 통계**
   * **기술 통계:** 데이터를 요약하고 설명하는 기법. 예: 평균, 분산, 히스토그램
   * **추론 통계:** 표본 데이터를 사용하여 모집단에 대해 결론을 도출하는 기법. 예: 가설검정, 신뢰구간
4. **데이터 유형**
5. **질적 데이터 (Qualitative Data)**

* **명목형 데이터 (Nominal Data):** 카테고리 또는 이름으로 표현되는 데이터로 순서가 없음.
  + - 예제: 성별 (남성, 여성), 혈액형 (A, B, AB, O)
  + **순서형 데이터 (Ordinal Data):** 순서가 있는 카테고리 데이터로 순서는 있으나 간격은 일정하지 않음.
    - 예제: 학년, 고객 만족도 (매우 불만족, 불만족, 보통, 만족, 매우 만족)

1. **양적 데이터 (Quantitative Data)**

* **이산형 데이터 (Discrete Data):** 셀 수 있는 데이터로 주로 정수로 표현됨.
  + - 예제: 학생 수, 판매량, 사고 발생 건수
  + **연속형 데이터 (Continuous Data):** 측정 가능한 데이터로 주로 실수로 표현됨.
    - 예제: 키, 몸무게, 시간

1. **기초 통계 지표**
2. **중심 경향성 측정**

* **평균 (Mean):** 데이터 값의 산술 평균. 모든 값을 더한 후 값의 개수로 나눈 값.

평균 =

* + 예제: 10, 20, 30의 평균 = (10 + 20 + 30) / 3 = 20
* **중앙값 (Median):** 데이터를 크기 순서대로 정렬했을 때 중앙에 위치한 값.
  + 홀수 개의 데이터: 중앙에 위치한 값
  + 예제: 10, 20, 30의 중앙값 = 20
* **최빈값 (Mode):** 데이터에서 가장 빈번하게 나타나는 값.
  + 예제: 10, 20, 20, 30의 최빈값 = 20

1. **변동성 측정**

* **분산 (Variance):** 데이터 값이 평균에서 얼마나 떨어져 있는지를 나타내는 값. 각 데이터 값에서 평균을 뺀 값을 제곱한 후 평균을 냄.

분산 =

* **표준편차 (Standard Deviation):** 분산의 제곱근. 데이터 값이 평균을 기준으로 얼마나 퍼져 있는지 나타냄

표준편차 =

* + 예제: 10, 20, 30의 분산과 표준편차
    - 평균 = 20
    - 편차 = (-10, 0, 10)
    - 분산 = (100 + 0 + 100) / 3 = 66.67
    - 표준편차 = ≈ 8.16
* **IQR(Interquartile Range, 사분위 범위):** 데이터의 중간 50%가 얼마나 퍼져 있는지를 나타내는 값으로 데이터의 분산 정도를 파악하는 데 사용

IQR = Q3(세 번째 사분위수)-Q1(첫 번째 사분위수)

1. **데이터 요약 방법**

* **도수 분포표 (Frequency Distribution Table):** 데이터를 카테고리별로 나누어 각 카테고리의 빈도를 나타낸 표.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

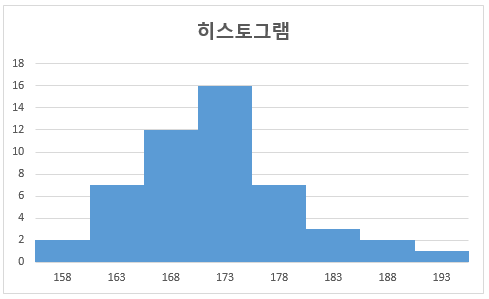
[전체 데이터]

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

[도수분포표 작성 예시]

* **히스토그램 (Histogram):** 도수 분포표를 그래프로 나타낸 것. 연속형 데이터의 분포를 시각적으로 표현.



[히스토그램]

* **상자 그림 (Box Plot):** 데이터의 중앙값, 사분위수, 최소/최대값 등을 시각적으로 나타낸 그래프.
  + 예제: 데이터 10, 20, 30, 40, 50의 상자 그림
    - 중앙값 = 30
    - 1사분위수 (Q1) = 20
    - 3사분위수 (Q3) = 40
    - 최소값 = 10
    - 최대값 = 50

텍스트, 스크린샷, 폰트, 도표이(가) 표시된 사진

자동 생성된 설명

**예제 데이터**

데이터셋: [12, 15, 14, 10, 18, 20, 12, 17, 16, 19]

1. 평균: 15.3
2. 중앙값: 15.5
3. 최빈값: 12
4. 분산 및 표준편차: 9.81, 3.132091952673165
5. Q3, Q1, IQR : 17.75, 12.5, 5.25