ChatGPT를 활용한 실무 데이터 분석

- 데이터와 기술통계2 -

presented by A.Iglue

오늘의 학습 목표

- 모집단과 표본의 관계를 설명할 수 있다.
- 기술통계량의 종류와 역할을 설명할 수 있다.
- Python과 라이브러리를 이용하여 기술통계량을 계산할 수 있다.

데이터 분석 목적 설정

■ 데이터 분석의 주된 목적은 대상의 요약, 설명, 예측임.

■ 데이터 분석의 시작은 구체적인 데이터 분석 목적을 정하는 것.

■ 데이터 분석의 목적에 따라 어떤 실험이나 관측으로 데이터를 얻어야 할지, 어떻게 데이터를 분석해야 할지 달라지기 때문임.

■ 예시

- ✓ 신약의 효과 유무와 효과의 크기를 알고 싶다.
- ✓ 소득과 행복도 사이에 어떤 관계가 있는지 알고 싶다.
- ✓ 기온으로부터 올해 농작물 수확량을 예측하고 싶다.

알고자 하는 대상 설정

■ 알고자 하는 대상을 명확히 설정해야 자료 수집과 분석의 타당성이 생김.

■ ex1) “혈압을 낮추는 신약”의 효과를 알고 싶다.

✓ 알고자 하는 대상은 “고혈압이 있는 모든 사람의 혈압”임.

✓ 그러나 모든 사람을 대상으로 실험할 수 없으므로, 고혈압 환자 80명을 모집해, 이 중 40명에게는 신약을, 나머지는 위약을 복용시킨 후 조사함.

✓ 여기서 중요한 것은 알고자 하는 대상은 이 실험에 참가한 80명뿐만이 아니라 모든 고혈압 환자가 됨.

■ ex2) “한 고등학교 2학년 1반, 2반의 영어 시험 점수”의 차이를 알고싶다.

✓ 분석의 대상은 한 고등학교 2학년 1반, 2반 학생들의 영어 시험 점수임.

✓ 다른 학년, 다른 반도 아니고, 다른 시험 점수도 아닌 1, 2반 학생들의 영어 시험 점수임.

모집단이란?

■ 모집단(population)이란 알고자 하는 대상 전체를 의미함.

✓ “고혈압을 낮추는 신약의 효능”분석에서는 신약 복용자가 포함된 모집단과 위약 복용자가 포함된 2개의 모집단을 가정함.

✓ “어떤 주사위의 모든 눈이 균등한 확률로 나오는지 검증”분석에서는 그 주사위를 무한하게 던졌을 때 나오는 눈의 전체 집합을 모집단으로 가정함.

■ 모집단을 설정할 때는 데이터 분석 목적과 알고자 하는 대상에 기초해야 함.

알고자 하는 대상이 전체일지라도, 실제로 데이터를 얻을 가능성이 없는 요소를 포함한 모집단은 적절하지 않음.

■ 알고자 하는 대상에서 데이터 획득 조건까지 고려하여 모집단을 설계해야 함.

✓ 모든 고혈압 환자가 모집단 대상이지만, 어떤 이유로 여성 환자의 데이터를 얻을 수 없다면, 모집단은 남성 고혈압 환자로 설정되어야만 함.

모집단의 크기

■ 유한모집단

- ✓ 모집단 중 포함된 요소의 개수가 유한한 집단
- ✓ ex) 1, 2반의 영어 성적 비교
- ✓ ex) 대한민국 국민의 평균 간수치(유한모집단이나 시간과 비용으로 인해 원칙적으로 모든 요소를 조사하기엔 불가능)

■ 무한모집단

- ✓ 모집단 중 포함된 요소의 개수가 무한한 집단
- ✓ ex) 고협알을 낮추는 신약의 경우, 현재의 고혈압 환자 뿐 아니라 미래에 고혈압 약을 복용할 사람도 대상에 포함됨으로 요소 개수에 제한이 없음.
- ✓ ex) 어떤 주사위의 각 눈이 나올 확률의 예에서는 무한하게 주사위를 던져야 함으로 무한모집단이라고 할 수 있음.

모집단의 성질

■ 모집단은 데이터 분석에서 알고자 하는 대상 전체를 가르키기 때문에, 모집단의 성질을 알 수 있다면 대상을 설명, 이해, 예측할 수 있음.

■ 모집단 성질의 예시

✓ 한국인 남성의 평균 키는 173cm이다.

✓ 신약을 복용한 사람의 최고 혈압 평균은 120mmHg이다.

✓ 이 주사위는 모든 눈이 균등하게 나온다.

✓ 이 주사위는 3의 눈이 1/4 확률로 나온다.

전수조사

- 모집단에 포함된 모든 요소를 조사하는 방법
- 전수조사의 경우, ”분석할 데이터 = 모집단“이므로, 획득한 데이터의 특징을 파악하고 기술하기만 해도, 모집단의 성질을 설명할 수 있음.
- 기술통계란, 데이터 그 자체의 특징을 기술하고 요약하는 것.
- 그러나, 전수조사는 비용이나 시간 면에서 부담이 커 실현 불가능한 경우가 대부분임.
- 극단적으로, 무한모집단인 경우는 전수조사가 아예 불가능함.

표본조사

- 모집단 전체가 아닌 모집단으로부터 추출한 표본을 이용해 모집단의 성질을 조사하는 방법
- 이처럼 모집단의 일부를 분석하여 모집단 전체의 성질을 추정하는 것을 추론통계(inferential statistics)라고 함.
  - ✓ ex) 선거 출구조사
- 표본(sample): 추론통계에서 조사하는 모집단의 일부
- 표본추출(sampling): 모집단에서 표본을 뽑는 것.
- 표본수: 추출한 표본의 개수
- 표본크기(sample size): 표본에 포함된 요소의 개수
  - ✓ 표본크기는 모집단의 성질을 추정할 때의 신뢰구간이나 가설검정의 결과에 영향을 미치는 중요한 요소임.

통계량이란?

■ 대상의 성질을 이해하기 위해 수집한 데이터로부터 다양한 계산을 통해 얻은 값

■ 기술통계량(요약통계량): 데이터 그 자체의 성질을 기술하고 요약하는 통계량

✓ 주로 수치형 변수 대상으로 계산 수행

✓ 범주형의 경우, '특정 범주의 값이 몇 개인지' 같은 개수(또는 비율)로만 데이터를 기술•요약 할 수 있음.

■ 결국 통계량으로 요약한다는 것은, 데이터에 있는 정보 중 버려지는 부분이 있다는 것을 의미함.

기술통계량 - 대푯값

■ 대푯값이란, 데이터의 대략적인 분포 위치, 즉 데이터가 어디를 중심으로 분포하는지를 알려주는 통계량

■ 평균값(mean) - 데이터의 중심 경향을 나타내는 통계량

✓ 데이터의 분포 위치를 파악하기 좋으나, 극단적인 값(이상치, outlier)의 영향을 많이 받음.

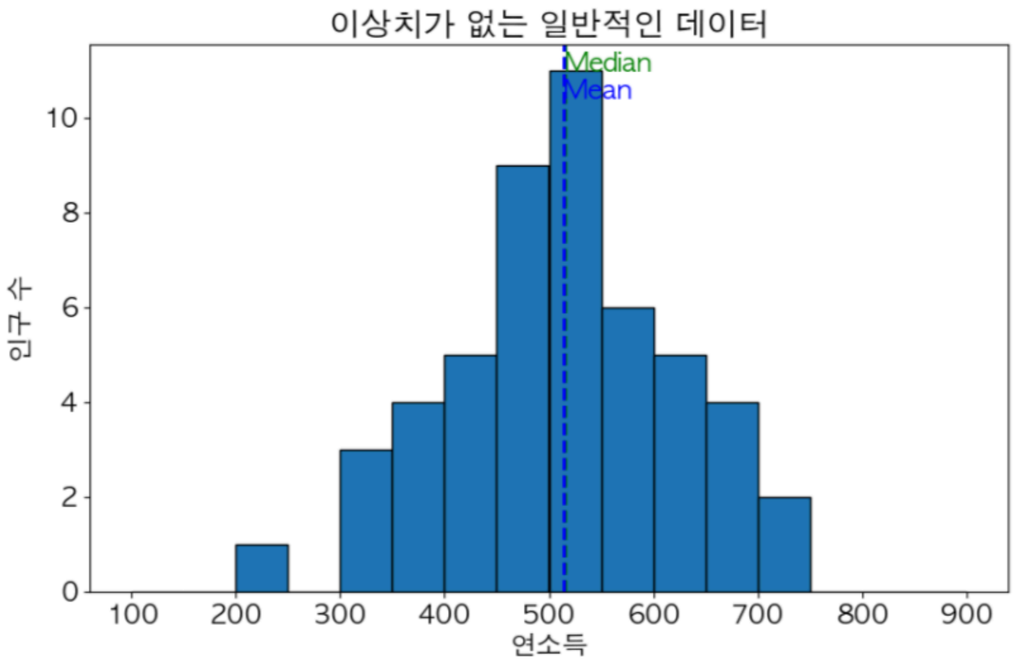
■ 중앙값(median) - 크기 순으로 값을 정렬했을 때 한 가운데 위치한 값

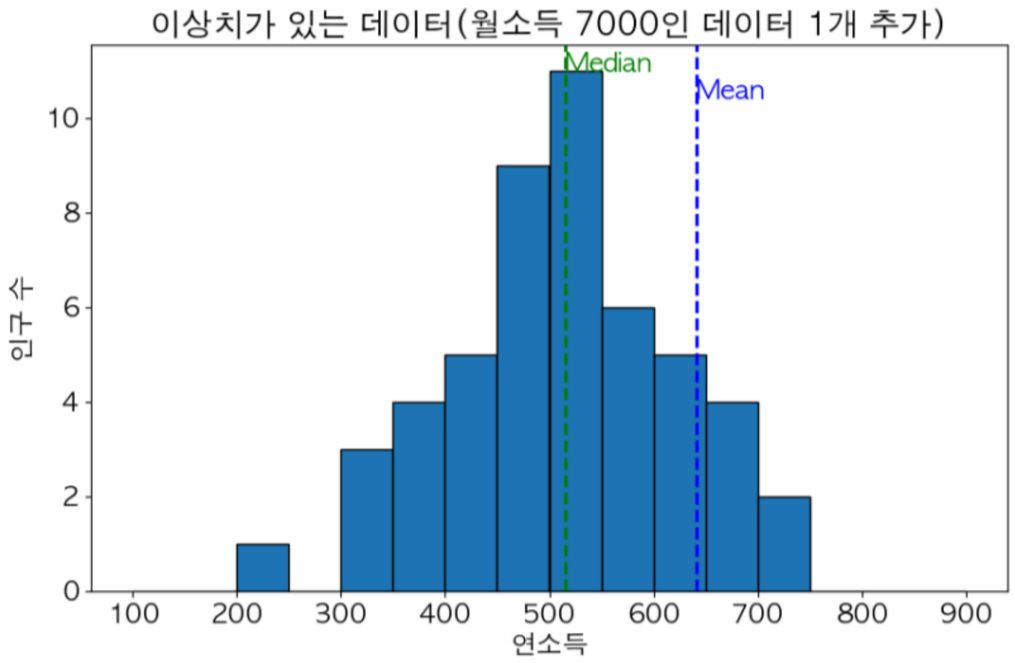
✓ 중앙값은 수치 자체의 정보가 아닌 순서에만 주목하기에, 극단적인 값의 영향을 받지 않음.

■ 최빈값(mode) - 데이터 중 가장 자주 나타나는 값

$$\text{평균} = \frac{a_1 + a_2 + a_3 + \ldots + a_n}{n}$$

이상치와 대푯값


이상치가 없는 일반적인 데이터
Median
Mean
인구 수
연소득
0
2
4
6
8
10
100
200
300
400
500
600
700
800
900


이상치가 있는 데이터(월소득 7000인 데이터 1개 추가)
Median
Mean
인구 수
연소득
0
2
4
6
8
10
100
200
300
400
500
600
700
800
900

기술통계량 - 분산

■ 분산(variance) - 데이터가 평균으로부터 어느 정도 퍼져 있는지(흩어져 있는지) 파악하는 통계량

✓ 모집단분산(population variance): 모집단의 모든 데이터를 사용하여 분산을 계산함.

✓ 표본분산(sample variance): 모집단의 일부인 표본 데이터를 사용하여 분산을 계산함.

■ 불편분산(unbiased variance) - 표본분산의 편향을 보정한 분산

✓ 표본분산은 모집단분산에 비해 과소평가하는 부분이 있기 때문에 자유도를 고려하여 데이터 개수(n)가 아니라 (n–1)로 나누어서 보정해주는데, 이를 불편분산이라고 함.

■ 분산의 특징

✓ 분산의 값은 항상 0보다 큼.

✓ 데이터의 값이 모두 같다면, 분산은 0임.

✓ 데이터의 퍼짐 정도가 크면, 분산의 값은 커짐.

기술통계량 - 분산

모집단분산 계산식	표본분산(불편분산) 계산식
$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

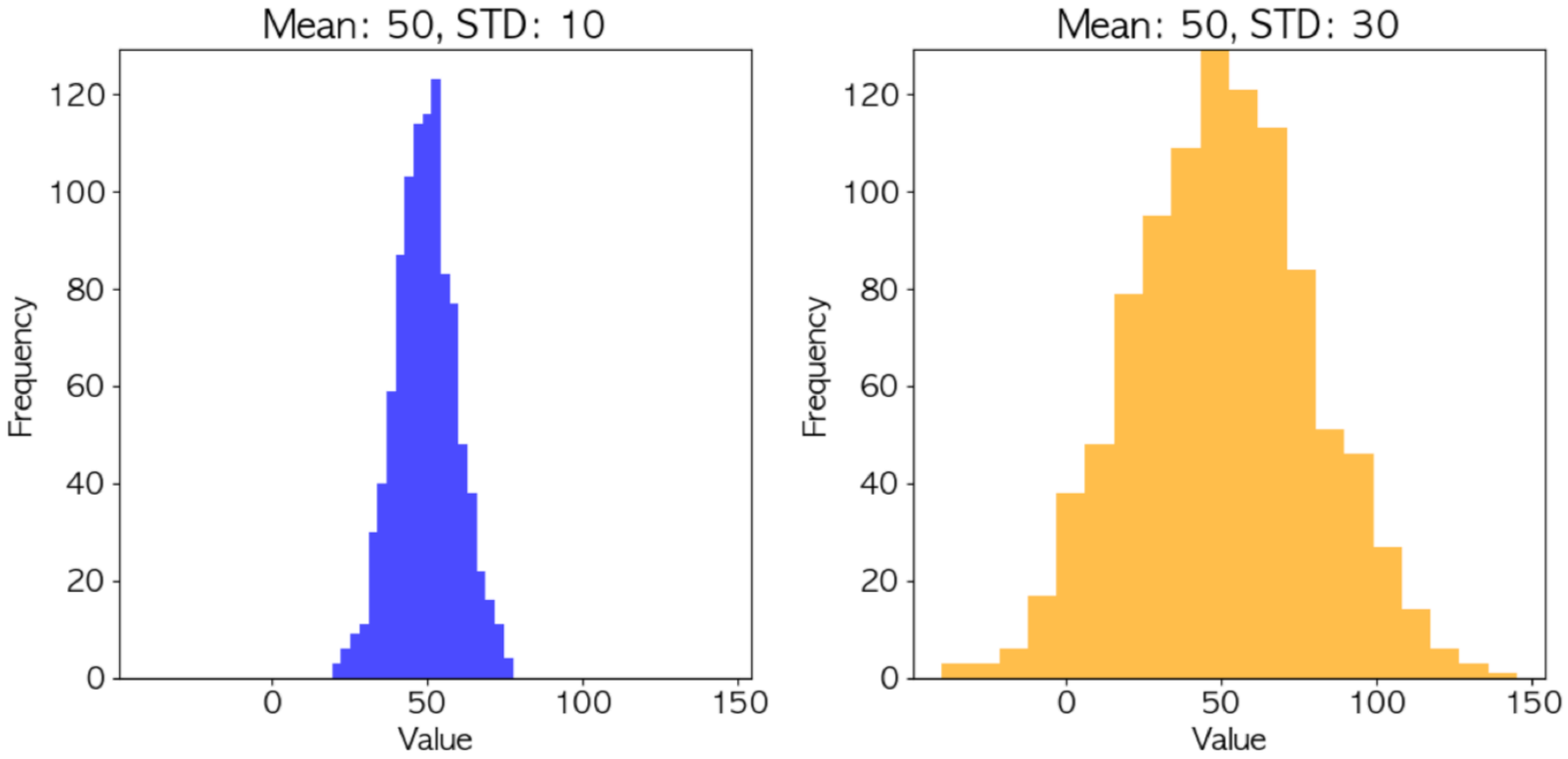
기술통계량 - 표준편차

■ 표준편차(standard deviation):

- ✓ 모집단 표준편차(population standard deviation)와 표본 표준편차(sample standard deviation)가 있음.
- ✓ 분산의 제곱근으로 계산되며, 데이터의 퍼짐 정도를 평균과 동일한 단위로 나타냄으로서 활용도가 높은 통계량

모집단 표준편차 계산식	표본 표준편차(불편분산) 계산식
$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

기술통계량 - 표준편차



▲ 같은 평균, 다른 분산의 분포 비교

오늘의 실습

■ python으로 이해하는 기술통계량

✓ python으로 기술통계량 수식을 구현하여 계산

✓ Numpy library를 활용한 기술통계량 계산