ChatGPT를 활용한 실무 데이터 분석

- 4. 정규분포와 중심극한정리 -

presented by A.Iglue

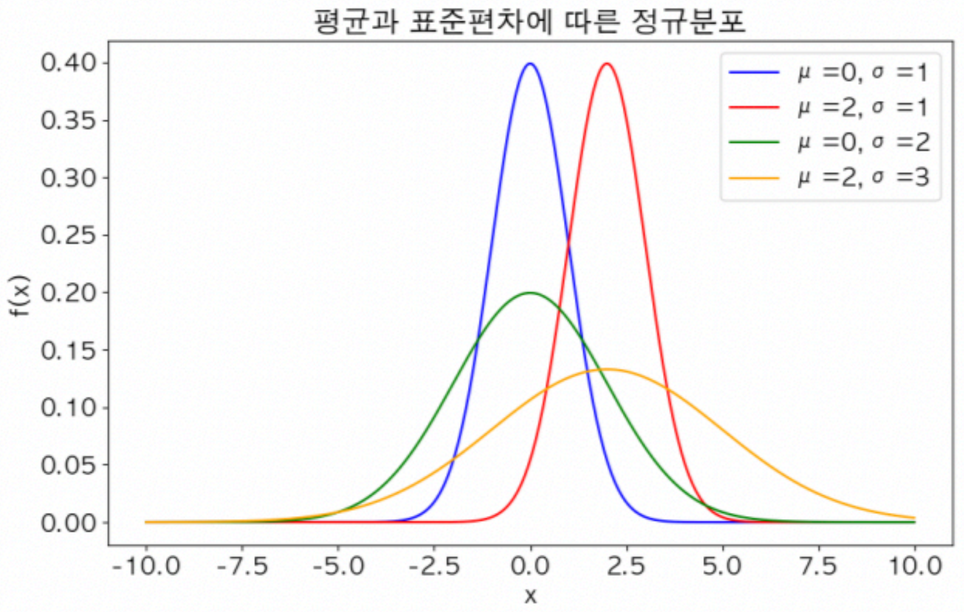
오늘의 학습 목표

- 정규분포와 그 의미를 설명할 수 있다.
- 큰수의 법칙을 이해하고, 시뮬레이션으로 구현할 수 있다.
- 중심극한정리를 이해하고, 시뮬레이션으로 구현할 수 있다.

정규분포

- 정규분포(normal distribution) 또는 가우스분포(gaussian distribution)는 통계학에서 가장 자주 등장하고 중요한 확률분포임.
- 정규분포는 연속형 확률변수를 대상으로 정의되고, 확률밀도함수는 평균(μ), 표준편차(σ)라는 2개의 파라미터로 결정됨.
- 정규분포는 $N(\mu, \sigma^2)$으로 표기함.
- 특히, 평균(μ=0), 표준편차(σ=1)인 정규분포 N(0, 1)인 경우를 표준정규분포라고 함.

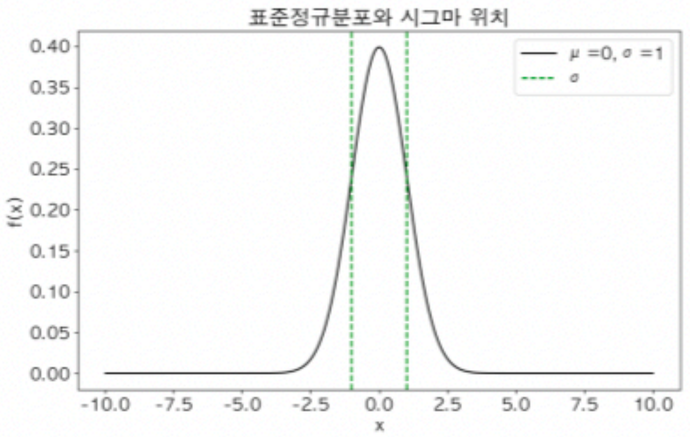
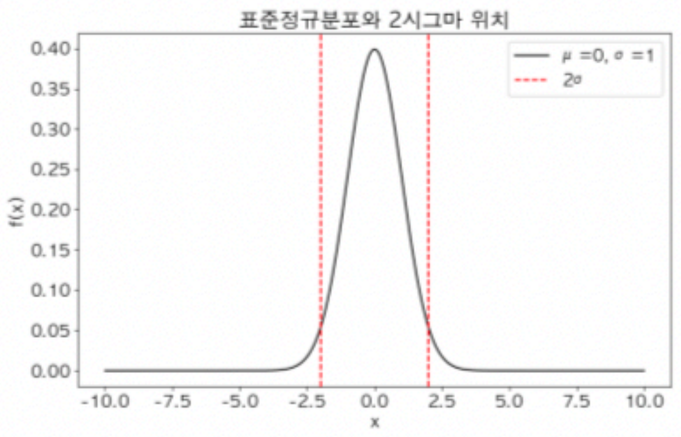
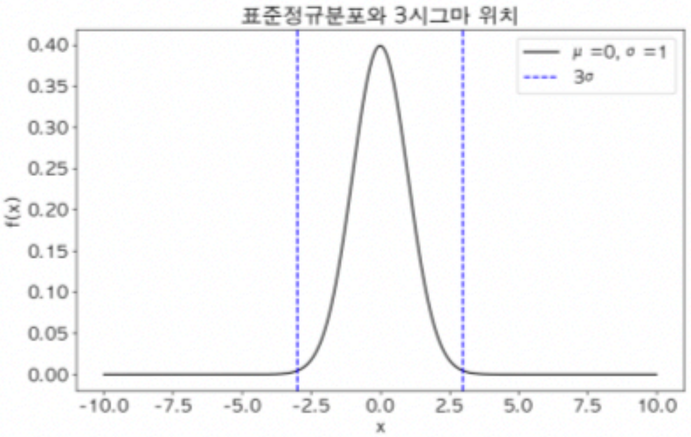
정규분포의 확률밀도함수

정규분포 확률밀도함수	평균과 표준편차에 따른 정규분포
$f(x\|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	

정규분포의 특징

- 평균을 중심으로 한 종형으로, 좌우대칭 분포임.
- 평균 근처에 값이 가장 많고, 평균에서 멀어질수록 적어짐.
- 키, 몸무게, 시험 성적 등 정규분포로 근사할 수 있는 현상이 많음.
- $\mu - \sigma$부터 $\mu + \sigma$까지의 범위에 값이 있을 확률은 약 68%임.
- $\mu - 2\sigma$부터 $\mu + 2\sigma$까지의 범위에 값이 있을 확률은 약 95%임.
- $\mu - 3\sigma$부터 $\mu + 3\sigma$까지의 범위에 값이 있을 확률은 약 99.7%임.

정규분포의 시그마 위치

1시그마 위치(약 68%)	2시그마 위치(약 95%)	3시그마 위치(약 99.7%)
		

표준화

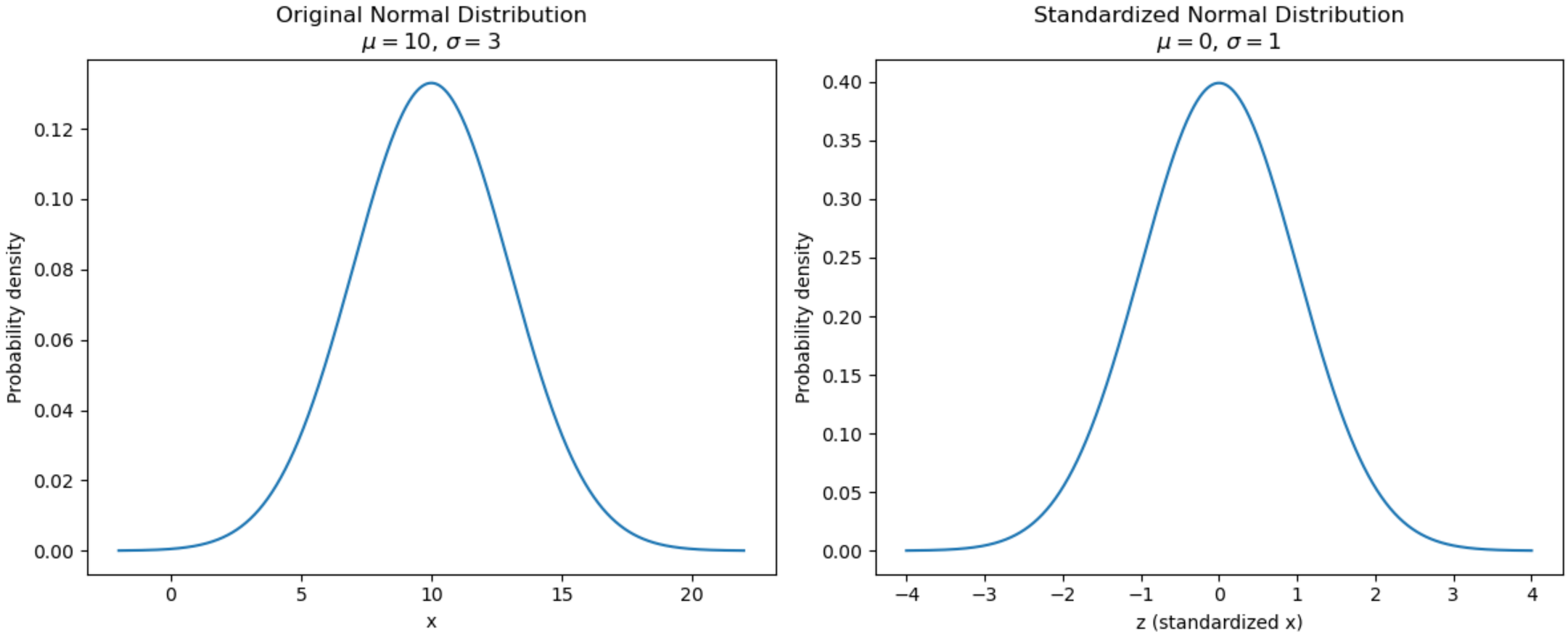
■ 일반적으로 확률변수 x는 데이터의 평균 μ와 표준편차 σ를 이용하여 평균이 0, 표준편차가 1인 확률변수 z로 변환할 수 있음.

■ 이를 표준화(standardizing, normalizing)라 하며, 새로운 값을 z값이라고 부름.

■ ex) 대학수학능력시험에 등장하는 표준점수는 평균을 50, 표준편차를 10으로 변환한 값(10z + 50)임. 예를 들어 평균 50, 표준편차 15이 점수 분포에서 80점을 받았다면, z=2이고, 표준점수는 70이 됨. 즉, 평균과 표준편차가 다른 과목별 분포에서 각 과목 당 점수의 위치를 계산하고 비교할 수 있음.

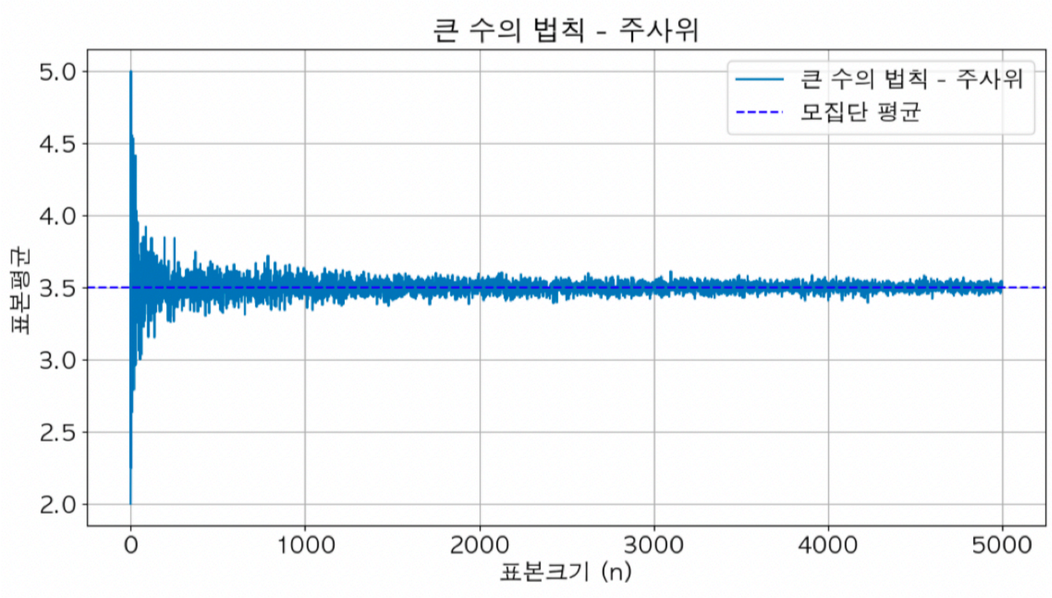
표준화 공식

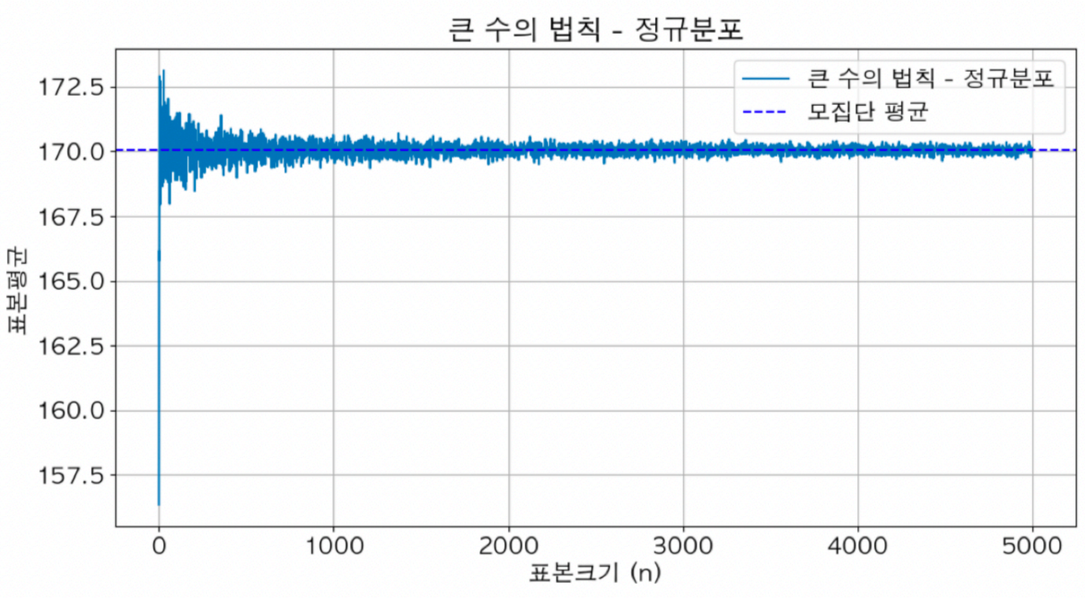
$$z = \frac{(x - \mu)}{\sigma}$$



큰수의 법칙

- 표본평균과 모집단평균의 관계는 큰 수의 법칙을 따름.
- 큰 수의 법칙(law of large numbers)이란, 표본크기 n이 커질수록 표본평균이 모집단평균에 한없이 가까워진다는 법칙임.
- 이 말은 곧, 표본오차(표본평균 - 모평균)가 0에 한없이 가까워진다는 말과 동치임.





중심극한정리

■ 독립적이고 동일한 분포를 가진 임의의 변수들의 합(또는 평균)이 충분히 큰 표본 크기에서 정규분포에 근사한다는 통계학의 원리

■ 보통 표본 크기는 30 이상이고, 모집단에서 여러 번 무작위추출을 각 표본의 평균을 구하고 그의 분포를 보면, 정규분포에 근사한다는 의미임.

■ 중심극한정리가 중요한 이유는, 모집단의 분포를 고려하지 않고도, 적당히 큰 표본크기를 갖는 표본평균분포를 알 수 있다면, 정규분포를 이용해 모집단의 평균과 표준편차를 추론할 수 있다는 이론적 근거가 되기 때문임.

■ 모집단의 평균과 표준편차 구하기

✓ 정규분포의 평균 = 모집단의 평균(μ)

✓ 정규분포의 표준편차 = $\frac{\sigma}{\sqrt{n}}$

오늘의 실습

- python으로 무작위추출 구현하기
- python으로 큰수의 법칙 구현하기
- python으로 중심극한정리 구현하기