ChatGPT를 활용한 실무 데이터 분석

- 3. 확률과 추론통계 연결하기 -

presented by A.Iglue

오늘의 학습 목표

- 확률변수, 확률질량함수, 확률밀도함수를 설명할 수 있다.
- 추정량의 의미를 설명할 수 있다.
- 비편향 추정량에 대해 설명할 수 있다.

확률이란?

- 확률이란, '발생 여부가 불확실한 사건의 발생 가능성을 숫자로 표현한 것'
- 사건 전체 경우의 수에 대해 해당 사건이 발생하는 경우의 수의 비율
- 사건 A의 확률은 P(A)로 나타냄.
- ex) 주머니에 붉은 구슬 4개와 흰 구슬 1개가 들어 있을 때, 안을 보지 않고 구슬 하나를 꺼냈을 때, 흰 구슬일 확률은?
- 실제 경우의 수로부터 확률을 정의하려면 어떤 구슬이든 동등한 확률로 꺼내진다는 가정이 있어야 함.

확률변수란?

- 확률변수(probability variable)란, 임의의 실험 결과에 수치를 대응시키는 변수
- 즉, 무작위로 발생하는 어떤 사건의 결과를 수치적으로 표현함.
- ex1) 변수 X={붉은 구슬, 흰 구슬}, P(X=붉은 구슬)=4/5, P(X=흰 구슬)=1/5
- ex2) 변수 X=주사위의 눈의 수, P(X=1)=1/6, P(X=6)=1/6
- 위 예에서 X와 같이 확률이 달라지는 변수를 확률변수라고 부르고, 확률변수가 실제로 취하는 값을 실현값이라고 함.

확률변수의 종류

■ 이산확률변수(discrete probability variable)

✓ 확률변수가 셀 수 있는 값을 취할 수 있음.

✓ 일반적으로 유한하거나 셀 수 있는 무한집합의 값을 가짐.

✓ ex) 동전 던지기의 앞면, 뒷면을 0과 1로 표현하는 경우

■ 연속확률변수(continuous probability variable)

✓ 확률변수가 연속적인 값을 취할 수 있으며, 일반적으로 모든 실수를 가질 수 있음.

✓ ex) 특정 지역에서 하루 동안의 강우량을 측정하는 경우

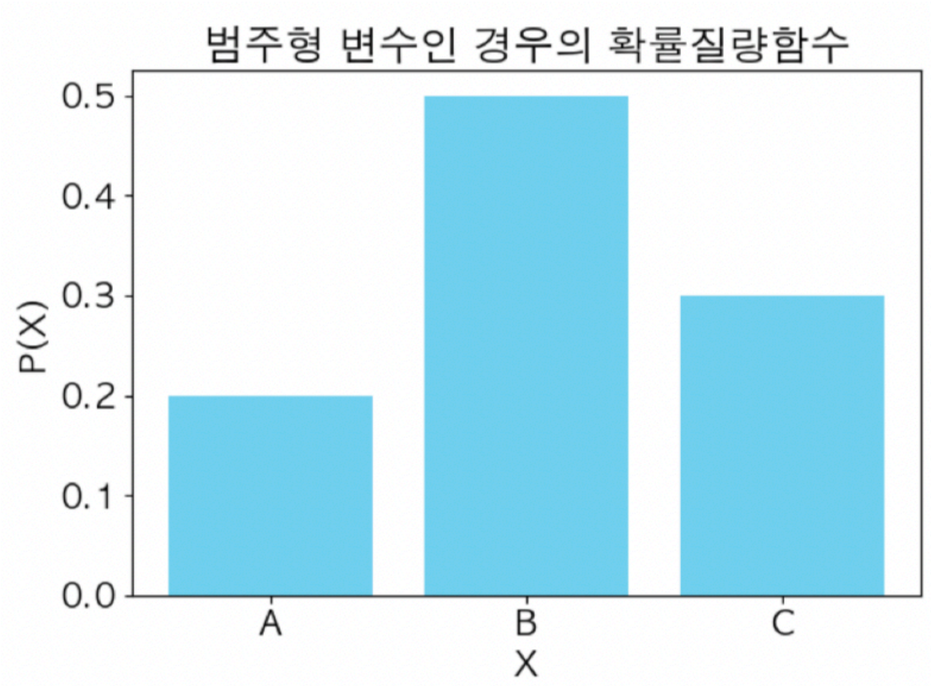
확률변수의 확률분포란?

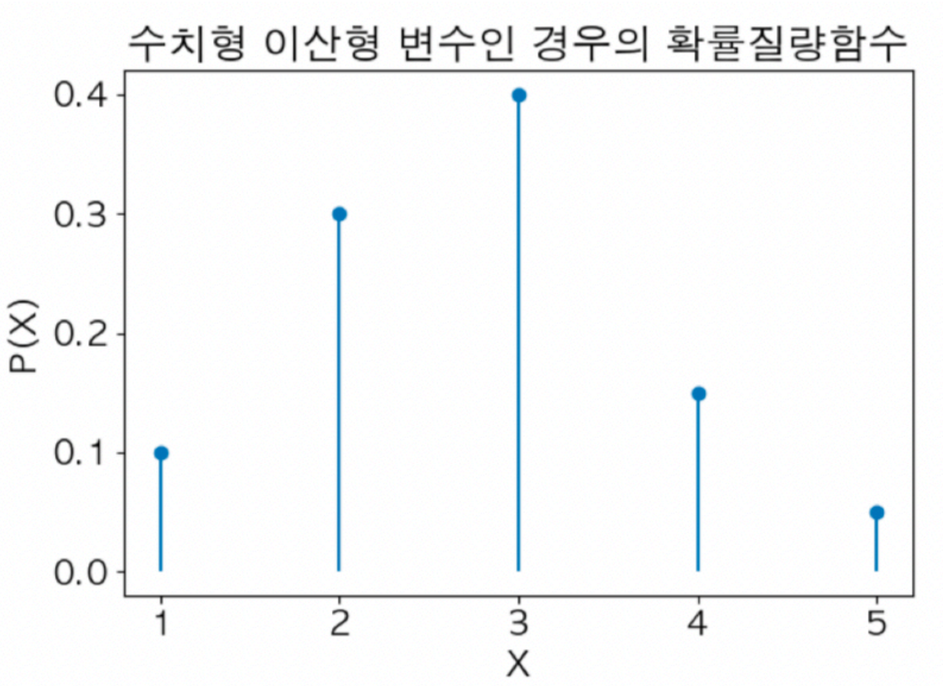
- 확률변수가 특정 값을 취할 확률을 표현함.
- 가로축에 확률변수, 세로축에 그 확률변수의 발생 가능성을 표시함.
- P(X=x)의 의미: ”어떤 확률변수(X)가 어떤 실현값(x)이 될 확률”
- 이산확률변수는 확률질량함수(probablility mass function, PMF), 연속확률변수는 확률밀도함수(probability density function, PDF)로 나타냄.

확률질량함수

■ 확률질량함수란, X가 이산형 확률변수이고, X의 실현값을 지정하면 그 값에 대한 확률을 즉시 계산할 수 있는 함수

■ 확률은 0 이상이고, 모두 더했을 때 1이 됨.

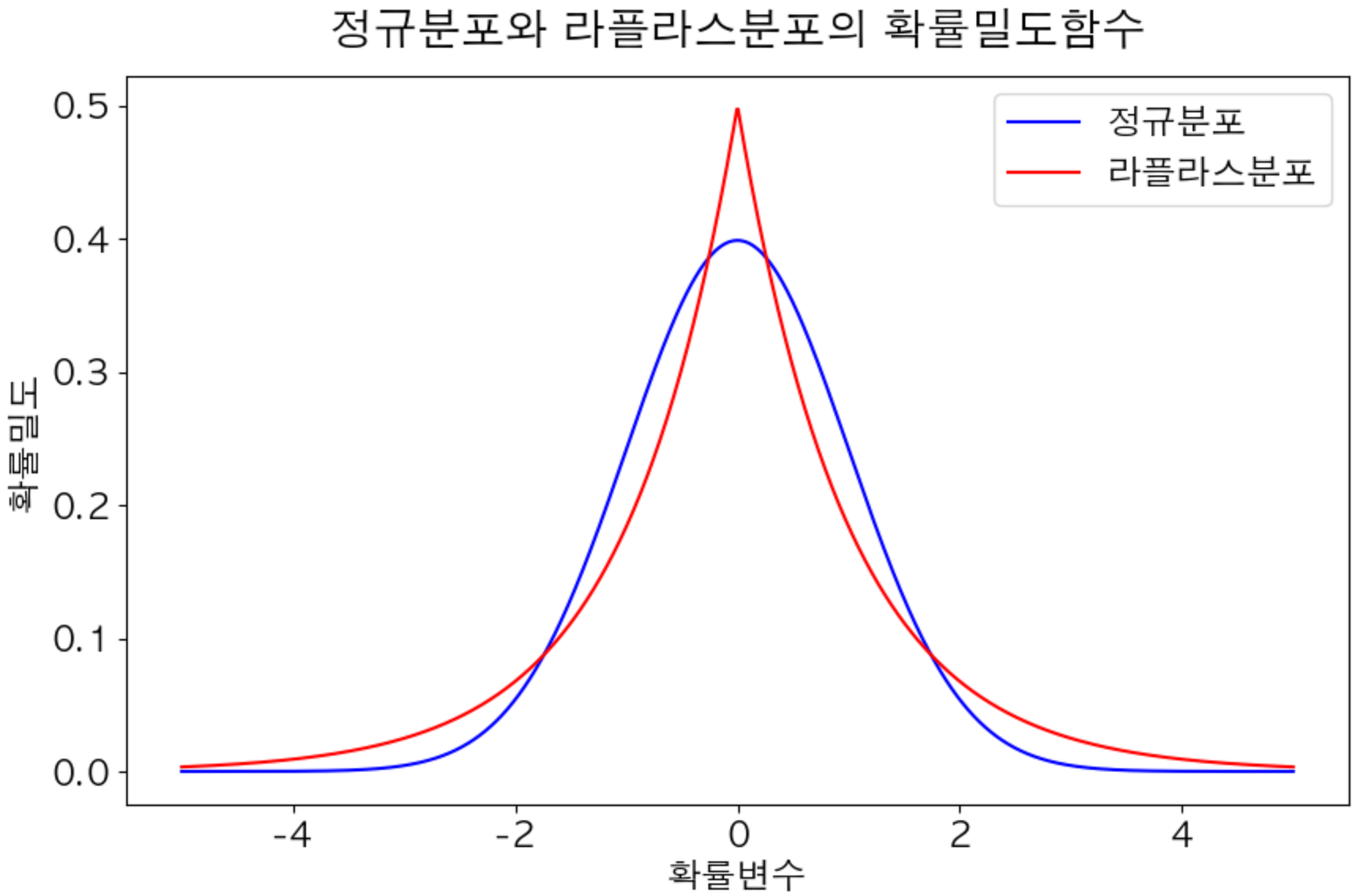




확률밀도함수

- X가 연속형 확률변수이고, 실수값을 갖는 변수 X가 $x < X < x+\Delta x$를 취할 확률이 $f(x)\cdot\Delta x$ 일 때, f(x)를 확률밀도함수라고 함. 여기서 Δx는 0에 가까운 극한값이라 봄.
- 이렇게 정의된 확률을 구하는 이유는 연속형 확률변수가 X=특정값 일때의 확률은 0이기 때문임(연속된 값이라 특정값을 가질 수 없음).
- 결국, '특정 값이 범위에 들어갈 확률'은 확률밀도함수 f(x)를 해당 구간에 대해 적분하여 구해야 함.
- 확률밀도함수의 세로축은 확률 그 자체의 값이 아니라, 상대적인 발생 가능성을 표현한 값임.
- 확률변수가 어떤 값에서 어떤 값까지의 범위에 들어갈 확률을 알고 싶다면, 확률밀도함수를 적분하여 x축과 확률밀도함수로 둘러싸인 부분의 넓이를 구함.
- 확률변수의 정의역 전체를 적분하면 1이 됨.

확률밀도함수 예시


정규분포와 라플라스분포의 확률밀도함수
정규분포
라플라스분포
확률밀도
0.5
0.4
0.3
0.2
0.1
0.0
-4
-2
0
2
4
확률변수

확률 구하기

확률질량함수의 확률 구하기	확률밀도함수의 확률 구하기
$P(X = x) = p(x)$	$P(a \leq X \leq b) = \int_a^b f(x)\,dx$

확률분포와 실현값

- 확률분포와 실현값은 모집단과 표본의 관계와 매우 비슷함.
- 즉, 모집단=확률분포, 표본=확률분포를 따르는 실현값
- 얻은 표본으로 모집단을 추정한다는 말은, 얻은 실현값으로 이 값을 발생시킨 확률분포를 추정한다는 말과 동일함.
- 모집단을 수학적으로 다룰 수 있는 확률분포(모형)에 근사하여 작업을 진행할 수 있게 되어, 모집단의 추정이 용이해짐.
- 수학적인 확률분포로 모집단분포를 근사하는 것을 모형화(modeling)이라고 함.

추론통계와 확률분포

- 추론통계: 모집단의 일부인 표본에서 모집단의 성질을 추정하는 통계
- 그러나 실제 모집단은 직접 관측할 수 없고 이해하기도 어려운 대상이기에, 표본으로 추정하는 일 역시 어려움.
- 따라서 현실 세계의 모집단을 수학 세계의 확률분포로 가정하고, 표본 데이터는 그 확률분포에서 생성된 실현값인 것으로 가정하여 추론통계를 진행함.
- 즉, 다루기 어려운 “모집단과 표본 데이터” 대신 “확률분포와 그 실현값”으로 분석 대상을 치환함.
- 표본 데이터(실현값)을 가지고 확률분포를 특징 짓는 모수(파라미터)를 추정함으로써, 모집단을 이해할 수 있음.
- 즉, 모수를 찾으면 확률분포를 규정할 수 있고, 모집단이 확률분포를 띈다는 가정 아래, 모집단을 추정하게 됨.

추정량과 추론통계량

■ 추정량(Estimator)

- ✓ 모집단의 성질(파라미터 - 평균, 분산 등)을 추정하기 위해 사용하는 표본 통계량
- ✓ 추정량은 확률변수의 확률분포를 계산하여 계산할 수 있음.
- ✓ 일치추정량: 표본크기 n을 무한대로 생각했을 때, 모집단의 성질과 일치하는 추정량
- ✓ 비편향추정량: 추정량의 평균값(기댓값)이 모집단의 성질과 일치할 때의 추정량
- ✓ 추정량 하나하나는 모집단의 성질(모평균 μ)에서 벗어나지만, 이를 모아 구한 평균값이 μ와 일치하는 경우 이를 비편향추정량이라고 부름.

■ 추론통계량(Inferential Statistics)

- ✓ 모집단에 대한 추론을 가능하게 하는 표본으로부터의 다양한 통계량을 포함하는 추정량보다 넓은 개념
- ✓ 추정량, 신뢰구간, 검정통계량 등 다양한 통계적 추론 과정에 사용되는 통계량을 의미함.

추정량 – 기댓값

■ 기댓값(expectation value)

- ✓ "변수가 확률적으로 얼마나 발생하기 쉬운가"를 평균적으로 나타낸 값
- ✓ 평균과 거의 동일함. 단지 확률을 통해 평균을 구한다는 의미로 기댓값을 사용함.
- ✓ 확률변수 X 전체에 대한 기댓값이므로 적분 구간은 전체가 됨.
- ✓ 확률변수 X의 각 실현값과 해당 확률의 곱을 한 후 더함.

■ 기댓값 계산하기

이산형 변수의 기댓값	연속형 변수의 기댓값
$E[X] = \sum_i x_i p(x_i)$	$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$

추정량 - 분산

■ 분산(variance)

✓ 확률분포가 기댓값 주변에 어느 정도 퍼졌는지를 나타내는 값.

✓ 기댓값과의 차이를 제곱한 숫자를 이용해 데이터가 기댓값에서 어느 정도 떨어져 있는지를 평가함.

■ 분산 계산하기

이산형 변수의 분산	연속형 변수의 분산
$\mathrm{Var}(X) = E[(X - E[X])^2] = \sum_i (x_i - E[X])^2 p(x_i)$	$\mathrm{Var}(X) = E[(X - E[X])^2] = \int_{-\infty}^{\infty} (x - E[X])^2 f(x)\, dx$

추정량 - 표준편차

■ 표준편차(standard deviation)

✓ 분산의 제곱근

✓ 분산에 비해 데이터와 동일한 단위를 사용하기 때문에 전체적으로 데이터의 분포가 평균으로부터 얼마나 떨어져 있는지 해석하기 용이함.

■ 표준편차 계산하기

이산형 변수의 표준편차	연속형 변수의 표준편차
$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\sum_i (x_i - E[X])^2 p(x_i)}$	$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)} = \sqrt{\int_{-\infty}^{\infty} (x - E[X])^2 f(x)\,dx}$

추정량 - 왜도와 첨도

■ 왜도(skewness)

✓ 분포가 좌우대칭에서 어느 정도 벗어났는가를 표현하는 수치.

✓ 0에 가까우며, 양수면 오른쪽으로 꼬리가 길고, 음수면 왼쪽으로 꼬리가 길게 됨.

■ 첨도(curtios)

✓ 분포가 얼마나 뾰족한지와 그래프의 꼬리가 차지하는 비율(분포의 양쪽 끝 꼬리의 확률 크기)이 얼마인지로 평가함.

✓ 첨도가 높으면 분포의 중심이 더 뾰족하고, 꼬리가 더 두꺼워짐.

■ 왜도•첨도 계산하기

왜도	첨도
$\text{Skewness}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$	$\text{Kurtosis}(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4 - 3$

비편향추정량 - 평균

■ 중심극한정리에서 표본평균 분포의 평균은 모집단 평균과 일치하므로, 비편향추정량이라고 할 수 있음.

■ 즉, 표본평균 분포의 평균값은 모집단의 평균값을 잘 대변한다고 인정함.

비편향추정량 - 비편향분산

■ 그러나, 표본 분산은 모집단 분산을 과소평가하는 문제가 있어, 계산식에서 n(표본크기)이 아닌 n-1로 나누어서 보정을 해줌.

✓ 표본분산이 모분산을 과소평가하는 이유는, 원래 편차(데이터와 평균의 차이)를 구할 때, 모평균을 사용해야 하지만, 모평균이 미지수이므로 표본평균(추정량)을 사용하면서, 모평균을 사용했을 때보다 편차가 더 적어지게 됨. 따라서 모분산 및 모표준편차에 비해 과소평가 됨.

■ 이 보정을 통해 계산된 분산을 비편향분산(불편분산)이라고 함.

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

비편향추정량 - 비편향표준편차

■ 비편향분산의 제곱근으로 보정된 표준편차 추정량

확률변수가 2개인 경우

■ 동시확률분포

✓ 확률변수 2개를 동시에 생각할 때의 확률분포

✓ ex) 2개의 주사위 A, B가 있을 때 주사위A에서 나온 눈을 X, 주사위B에서 나온 눈을 Y라 하면, 주사위A가 1의 눈이 나오는 동시에 주사위B가 2의 눈이 나올 확률은 P(X=1, Y=2)로 표현함.

■ 독립확률분포

✓ X, Y 2개 확률변수가 독립이라는 말은 한쪽이 어떤 값을 취하든지, 다른 한쪽의 발생 확률은 변하지 않음.

✓ 독립인 경우의 확률분포는 P(X)•P(Y)로 표현할 수 있음.

✓ ex) 주머니에 든 공을 뽑을 때, 먼저 뽑은 공을 다시 넣고 다음 사람이 공을 뽑을 때

✓ ex) 만약 공을 다시 집어 넣지 않더라도, 뽑는 횟수에 비해 주머니에 공이 엄청 많은 경우라면, 독립이라고 할 수 있음.

확률분포함수와 파라미터

■ 이론적인 확률분포는 수식으로 표현되며, 분포의 형태를 정하는 숫자인 파라미터(parameter, 모수)를 가짐.

■ 데이터 분석의 목적은 모집단의 성질을 파악하는 것이고, 이는 곧 확률분포함수의 모수를 파악하는 것과 같은 의미임.

■ 즉, 모집단을 “○○이라는 파라미터를 가진 □□이라는 확률분포”로 나타낼(근사할) 수 있다면, 이는 곧 모집단의 성질을 아는 것이 되고, 이것이 바로 데이터 분석의 목적이 되는 것임.