Data Science

Basic Statistics

노기섭 교수

(kafa46@hongik.ac.kr)

Lecture Goals

- 데이터 요약을 위한 기초 통계 개념 이해
- 평균, 분산 등 기초 통계량의 정의와 계산법 숙지
- 다양한 확률분포의 특징 파악
- 표본 추정 및 신뢰구간 개념 이해

주요 개념

- 평균(Mean), 중앙값(Median), 최빈값(Mode)
- 분산(Variance), 표준편차(Standard Deviation)
- 확률분포 (이항분포, 포아송분포, 정규분포 등)
- 추정(Estimation): 점추정과 구간추정

기술 통계

기술통계의 의미와 역할

- **정의**
  - 데이터를 요약해 대표적인 특성과 분포를 파악하는 방법
- **목적**
  - 수집된 데이터를 전체적으로 이해하고 이후 분석 기법 선택의 기초 제공
- **주요 질문**
  - 평균적으로 어떤 경향을 보이는가? → 평균
  - 데이터의 흩어짐은 어느 정도인가? → 표준편차
  - 가장 자주 나타나는 값은 무엇인가? → 최빈값
  - 극단값은 존재하는가? → 이상치(Outlier)
- **기술통계는 데이터를 간단히 요약해 데이터의 전반적 성격을 한눈에 이해하도록 돕는다.**

간단한 예제

- 예를 들어, 10명의 학생 시험 점수

  [72, 77, 77, 83, 88, 91, 95]

- **이 데이터를 요약하면:**
  - 평균: 약 83.3
  - 중앙값: 83
  - 표준편차: 약 9.5
  - 최빈값: 최빈값은 77 (두 번 등장 → 가장 많이 등장)
- **성적이 80점대 초반에 몰려 있고, 흩어짐은 약 ±10점 수준이라는 것을 직관적으로 파악**

중심 경향 (Measures of Central Tendency)

- **데이터의 전반적인 위치나 대푯값을 나타내는 지표를 중심 경향이라 한다.**
- **사람들이 중심 경향을 사용하는 이유**
  - 복잡한 데이터 요약
    - 수많은 데이터 값을 하나의 대푯값으로 단순화하여 전체 경향을 빠르게 이해
  - 비교 가능성 제공
    - 집단 간 차이를 평균이나 중앙값으로 비교함으로써 직관적 해석 가능
  - 의사결정 지원: 기업의 매출, 학생들의 성적, 사회 조사 결과 등에서

    중심값을 활용하면 정책 결정이나 전략 수립에 근거를 제공
  - 데이터 해석 기준: 분산, 표준편차 같은 다른 통계 지표들을 해석할 때 기준점 역할
    - 예를 들어, 평균을 중심으로 데이터가 얼마나 흩어져 있는지 파악할 수 있다.

평균(Mean)의 개념과 종류

산술평균은 데이터를 대표하는 가장 기본적이고 직관적인 통계량

- 모든 값을 더해 개수로 나누며, 전체 자료의 중심 경향을 나타냄.
- 장점: 계산이 쉽고 이해하기 쉽다.
- 단점: 극단값(outlier)에 민감해 평균이 왜곡될 수 있다.

다른 평균

- 데이터 특성과 목적에 따라 기하평균, 조화평균 등 다른 형태의 평균을 사용할 수 있다.

산술평균 (Arithmetic Mean)

- 가장 기본적인 평균으로, 모든 값을 동일한 비중으로 더한 뒤 개수로 나눈 값

- 대부분의 데이터에서 기본적으로 사용되지만, 극단값에 민감함

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

가중평균 (Weighted Mean)

■ **각 값에 중요도(가중치)를 곱해 평균을 계산**

- 예: 시험 점수를 계산할 때 중간고사 40%, 기말고사 60%와 같은 가중치 적용

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

- 한 학생의 점수가 다음과 같을 경우
  - 중간고사: 80점 (가중치 40%)
  - 기말고사: 90점 (가중치 60%)

$$\bar{x} = \frac{0.4 \times 80 + 0.6 \times 90}{0.4 + 0.6} = 86$$

기하평균 (Geometric Mean)

■ **n 개의 값을 모두 곱한 뒤 n 제곱근을 취해 계산**

- 주로 성장률이나 비율 데이터 분석에 사용
- 예: 투자 수익률이 매년 10%, 20%, -5%일 때, 실제 투자 성과를 설명하는 데 적합

$$G = \left(\prod_{i=1}^{n} x_i \right)^{\frac{1}{n}}$$

매년 일정한 성장률 r로 3년 동안 불어나서 1.254배가 되었다고 하면, 그 r 은 기하평균으로 구할 수 있다.

$$r = (1.254)^{\frac{1}{3}} - 1 \approx 7.8\%$$

- 투자금이 3년 동안 다음과 같은 수익률을 기록했다고 하자.
  - 1년차: +10% → 1.1배
  - 2년차: +20% → 1.2배
  - 3년차: −5% → 0.95배
  - 최종 (3년간) 성장 배율은 다음과 같이 계산할 수 있다.
  - 즉, 3년 뒤에는 원금이 약 1.254배가 된다.

$$1.1 \times 1.2 \times 0.95 = 1.254$$

산술평균으로 계산할 경우

$$r = \frac{10\% + 20\% - 5\%}{3} \approx 8.3\%$$

조화평균 (Harmonic Mean)

■ **데이터 값들의 역수 $\frac{1}{n}$ 의 평균을 구한 뒤, 다시 역수를 취한 값이다.**

- 주로 속도, 비율, 단위당 데이터 분석에 적합

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

- 산술평균과의 차이점
  - 산술평균은 "그냥 더해서 나눈 값" → 큰 값에 민감
  - 조화평균은 "역수의 평균" → 작은 값에 민감
  - 따라서 속도, 비율, 단위당 성능처럼 분모가 중요한 상황에서 필요

조화평균 (Harmonic Mean) – 예시 1

- 평균 속도
  - 서울 → 부산(200km): 100 km/h
  - 부산 → 서울(200km): 50 km/h
  - 산술평균으로 구하면?

$$\frac{100 + 50}{2} = 75\ km/h$$

- 하지만 실제 평균 속도는 전체 거리 ÷ 전체 시간으로 계산해야 한다.
  - 총 거리 = 400 km
  - 총 시간 = 200/100 + 200/50 = 2 + 4 = 6 시간
  - 평균 속도 = 400/6 = 66.67 km/h

$$H = \frac{2}{\frac{1}{100} + \frac{1}{50}} = 66.67\ km/h$$

  - 위와 같은 계산법은 조화평균 공식을 적용하면 정확하게 같은 값을 얻을 수 있다

조화평균 (Harmonic Mean) – 예시 2

- CPU의 평균 성능
  - CPU A: 초당 100회 연산 (100 ops/s)
  - CPU B: 초당 300회 연산 (300 ops/s)
  - 산술평균으로 구하면?

$$\frac{100 + 200}{2} = 200\ ops/s$$

  - 하지만 조화평균 개념을 적용하면 다음과 같다.

$$H = \frac{2}{\frac{1}{100} + \frac{1}{300}} = 150\ ops/s$$

  - 두 CPU가 같은 양의 작업을 순차적으로 처리할 경우, 산술평균은 으로 과대 평가 ➔ 다음 슬라이드

조화평균 (Harmonic Mean) – 예시 2 (계속)

- 순차적으로 같은 작업량을 처리할 경우 속도는 총 시간은 구간별 시간이 합이 되어야 한다.
- 같은 작업량 1 단위를 각각 처리하면 걸리는 시간은 각각 $\frac{1}{100}$초, $\frac{1}{300}$초가 된다.
- 2개의 CPU가 작업하므로 총 작업량은 2 단위이다. 그리고 총 시간은 $\frac{1}{100}+\frac{1}{300}$초 이다.
- 두 속도로 같은 작업량을 순차적으로 처리한 경우의 평균은 전체작업량/전체시간 이므로, 평균 속도는 다음과 같다.

$$H = \frac{2}{\frac{1}{100}+\frac{1}{300}} = 150\ ops/s$$

조화평균은 속도, 처리율, 가동율,
인구밀도, 자동차 연비, 생산성과 같이
역수 단위 1/어떤값 (역수 단위) 로
표현되는 값들의 평균을 낼 때 사용

- 즉, "작업 시간"을 고려하지 않은 단순 산술평균은 빠른 CPU의 성능을 지나치게 반영하고, 느린 CPU의 영향을 과소평가한다.

절사평균 (Trimmed Mean)

■ **데이터에서 상위와 하위 일부 값을 제외하고 산술평균을 계산하는 방법**

- 극단값의 영향을 줄여 안정적인 대표값을 얻는 데 효과적

$$T_p = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

$x_{(i)}$: 오름차순으로 정렬된 데이터

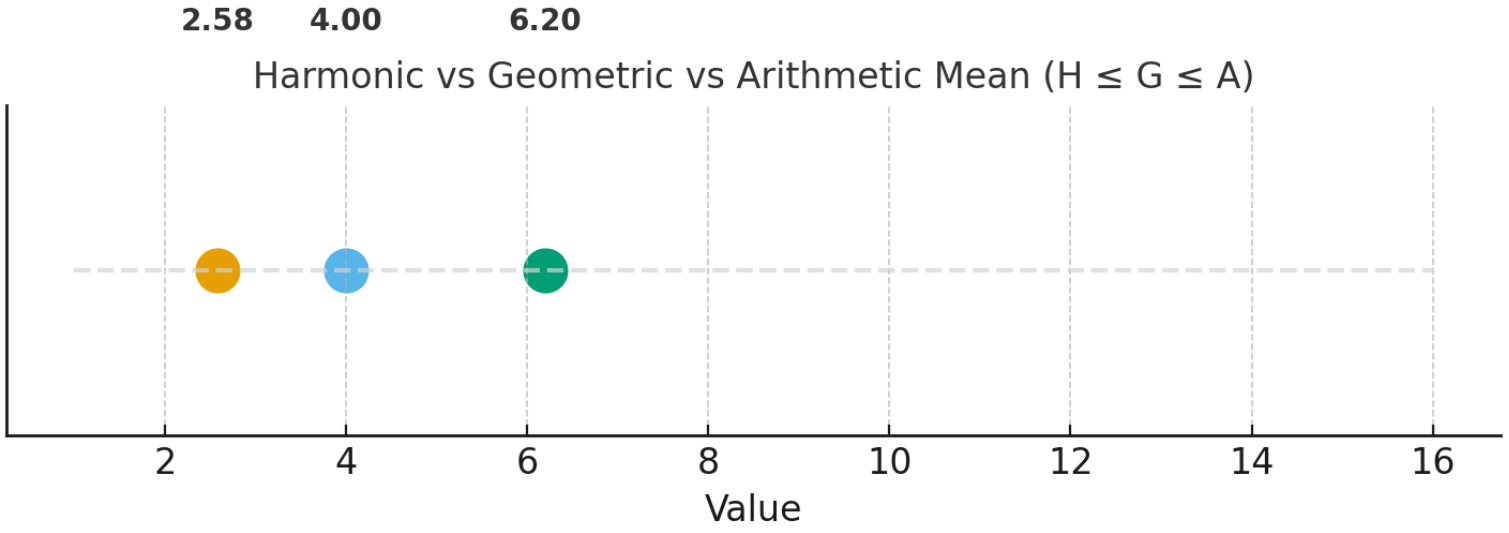
$k = \lfloor p \cdot n \rfloor$: 제거할 데이터 개수

언제 어떤 대표값(평균)을 사용할 것인가?

평균 종류	정의	언제 쓰는가	대표 사례
산술평균 (Arithmetic Mean)	값들의 합 ÷ 개수	값이 단순히 더해지는 상황	시험 점수 평균, 키·몸무게, 소득
기하평균 (Geometric Mean)	값들의 곱의 n 제곱근	곱셈·성장률·비율의 누적 효과가 중요한 상황	투자 수익률, 인구 증가율, 성장률 비교
조화평균 (Harmonic Mean)	값 개수 ÷ (역수들의 합)	속도·효율·비율처럼 분모가 중요한 상황	평균 속도, 연비, CPU 성능, 처리율

평균들의 특성 비교 – 직관적 이해

■ [1, 2, 4, 8, 16] 이용해 평균을 구했을 경우 위치



- 평균은 데이터를 대표하는 중요한 통계량이지만, 단순히 산술평균만으로는 충분하지 않다.
- 데이터의 특성과 목적에 따라 가중평균, 기하평균, 조화평균, 절사평균 등을 적절히 활용해야 한다.

다양한 평균을 구하는 실습

https://www.deepshark.org/courses/data_science/w/05_basic_statistics#mean_practice

산포도 (Measures of Dispersion)

산포도 (Measures of Dispersion)

산포도 (Measures of Dispersion)

- 산포도는 데이터가 평균을 중심으로 얼마나 퍼져 있는지를 나타내는 지표

주요 지표:

- **분산(Variance)**: 데이터가 평균에서 떨어진 정도를 제곱해 평균낸 값. 값이 클수록 데이터가 넓게 퍼짐.
- **표준편차(Standard Deviation)**: 분산의 제곱근으로, 데이터 단위와 동일해 직관적 해석이 가능.
- **사분위수(Quartiles)**: 데이터를 4등분해 분포 범위와 이상치를 파악하는 데 유용(Q1, Q2, Q3).

활용 예시

- 평균만 제시하는 것보다 중앙값, 표준편차, 최빈값 등을 함께 제시하면 데이터의 분포와 특성을 더 잘 설명
- 기술통계에서 산포도는 데이터의 흩어짐과 분포 특성을 요약하며, 확률분포·추정·가설검정 등 통계적 추론의 기초 됨.

확률변수(Random Variable)

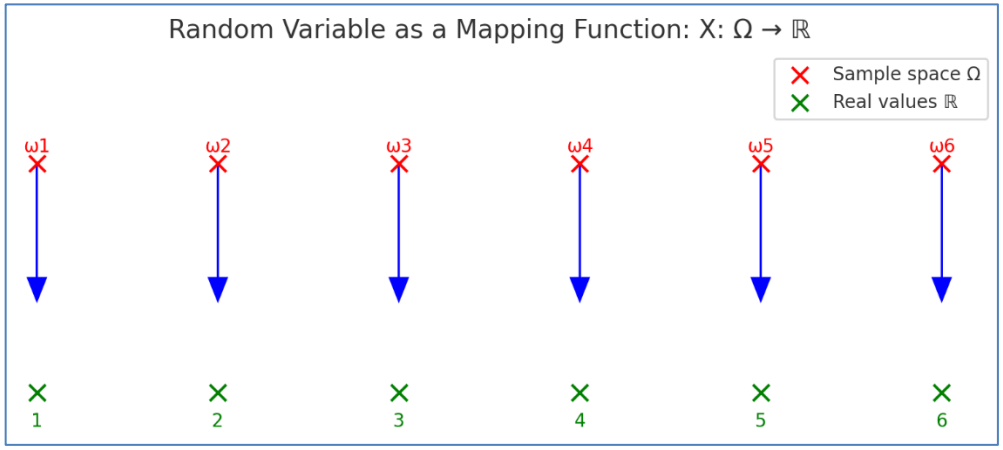
확률변수(Random Variable) - 1/3

확률변수(Random Variable)

- 우연적인 사건의 결과를 수치로 표현한 것, 예: 주사위 눈금(1 ~ 6), 시험 점수(0 ~ 100)

확률변수는 단순히 결과 값이 아니라, 확률적 실험의 표본공간(Sample Space)에서 실제 수치 집합으로의 사상(mapping function) 이다.

$$X: \Omega \rightarrow \mathbb{R}$$



- Ω : 가능한 모든 사건들의 집합(표본공간)
- $\mathbb{R}$: 실수 집합
- 예: 주사위 실험에서

$\Omega = \{w_1, w_2, \cdots, w_6\}$ 이라 하면, 확률변수 X는 각 사건 w_i를 눈금 값 $\{1, 2, 3, 4, 5, 6\}$으로 매핑한다.

확률변수(Random Variable) - 2/3

왜 함수로 이해해야 할까?

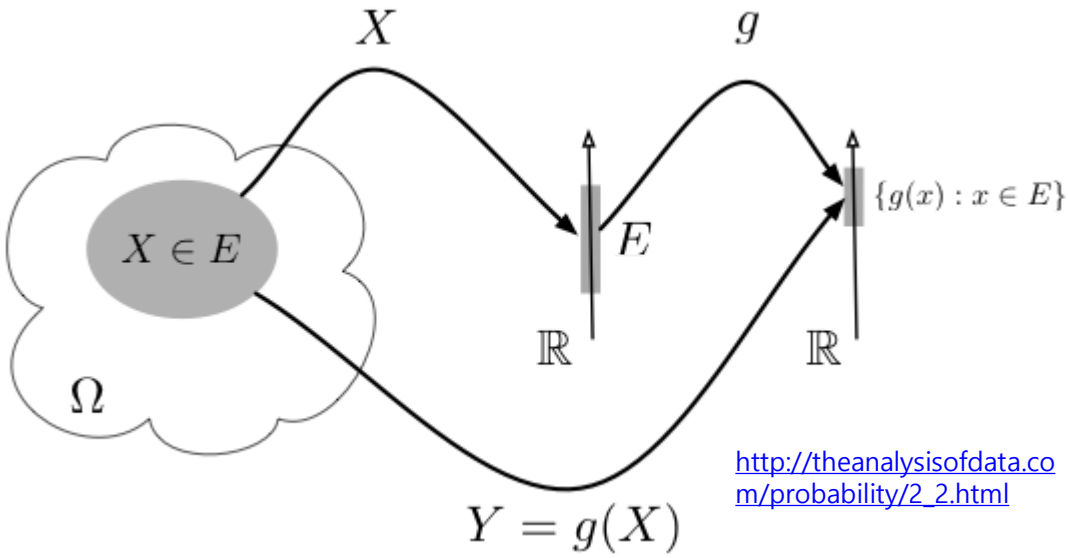
- 수학적 일관성 사건에 확률을 부여하고 이를 계산하기 위해 사건을 수치로 변환하는 규칙이 필요
- 확률분포 정의의 기반 확률분포는 "확률변수가 특정 수치를 가질 확률" 을 나타내므로, 확률변수의 함수적 성질이 전제되어야 함.
- 응용의 편리성 실제 데이터 분석에서는 사건 자체보다는 수치화 된 결과를 다루는 것이 효율적

Given an RV X,

the RV $Y = g(X)$ is a mapping

from Ω to $\mathbb{R}$ realized by

$$\omega \xrightarrow{\text{yields}} g\big(X(\omega)\big)$$



http://theanalysisofdata.com/probability/2_2.html

확률변수(Random Variable) - 3/3

■ **수학에서의 변수**

- 보통 미지수나 값을 임의로 대입할 수 있는 기호로, 그 자체가 값(value)을 직접 나타낸다.

■ **확률변수**

- 표본공간의 사건을 실수로 변환하는 함수이며, 사건에 확률이 결합되어 있다는 점이 본질적인 차이다.

■ **예제**

- $\boldsymbol{x = 3}$ 같은 단순한 대입은 수학적 변수의 개념이고,
- 주사위를 던져서 나온 사건 $\boldsymbol{w}$를 $\boldsymbol{X(w) = 3}$으로 매핑하는 것은 확률변수의 개념

■ **따라서 확률변수는 단순한 "숫자 기호"가 아니라, 사건을 수치로 연결시켜 확률적 계산을 가능하게 하는 함수적 개념**

이산형 vs. 연속형 확률변수 – 1/3

- 확률변수의 구분
  - 이산형(discrete) 확률변수
  - 연속형(continuous) 확률변수
  - 확률변수가 가질 수 있는 값의 형태와 확률을 표현하는 방식의 차이
- **이산형 확률변수 (Discrete Random Variable)**
  - 특징: 셀 수 있는 값들만 취한다. 값의 집합이 유한하거나 가산무한(1, 2, 3, ...)일 수 있다.
  - 확률 표현: 각 가능한 값마다 확률을 직접 부여하며, 모든 확률의 합은 1이다.
- **예시**
  - 주사위 눈금 (1, 2, 3, 4, 5, 6)
  - 동전 던지기의 결과 (앞면=1, 뒷면=0)
  - 하루에 도착하는 손님 수 (0명, 1명, 2명, ...)

이산형 vs. 연속형 확률변수 – 2/3

연속형 확률변수 (Continuous Random Variable)

- 특정 구간 안에서 무한히 많은 값을 가질 수 있다. 값은 연속적인 실수 범위에 속한다.
- 확률 표현: 특정 값 하나의 확률은 0이며, 구간에 대한 확률로 정의
  - 확률밀도함수(pdf)를 사용

예제

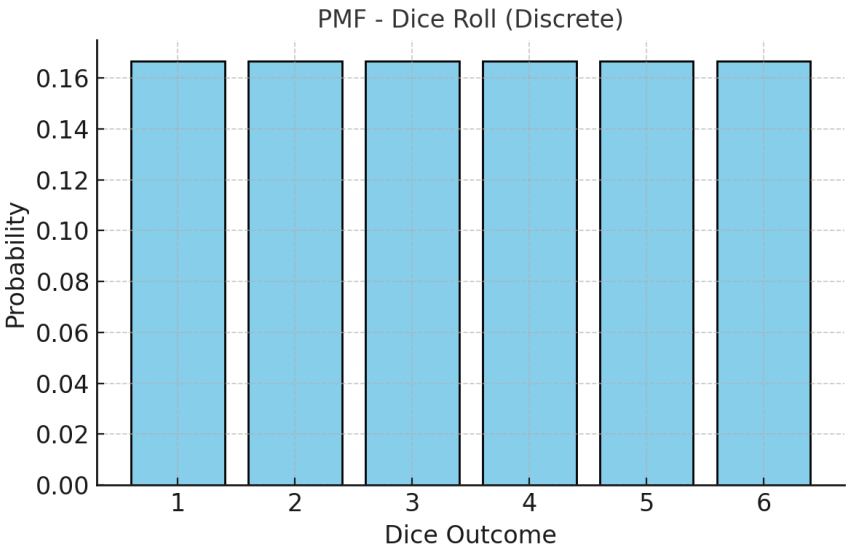
- 사람의 키, 몸무게
- 특정 지역의 기온, 강수량
- 기계가 고장 나기까지 걸리는 시간

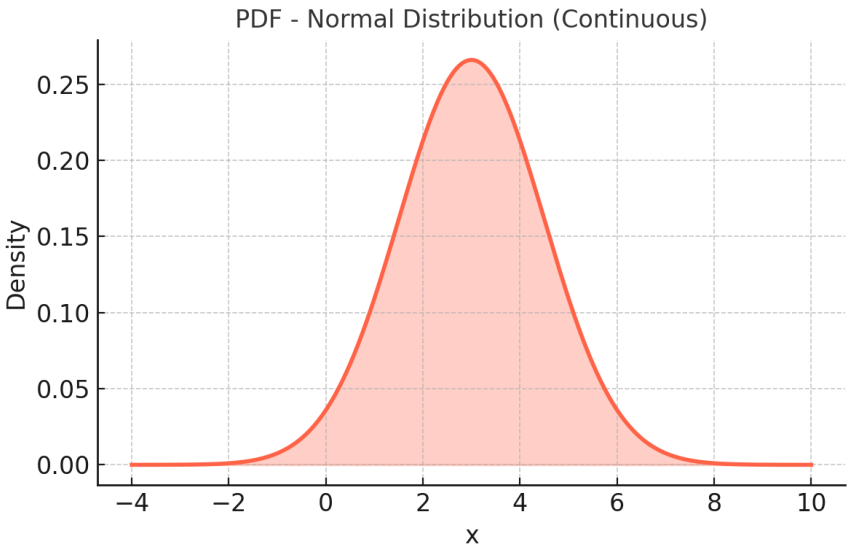
이산형 vs. 연속형 확률변수 – 3/3

구분	이산형 확률변수	연속형 확률변수
값의 형태	셀 수 있는 값 (유한 또는 가산무한)	연속적인 실수 구간
확률 표현	확률질량함수 (PMF)	확률밀도함수 (PDF)
특정 값의 확률	가능	구간 확률로 정의
예시	주사위, 동전, 손님 수	키, 몸무게, 시간, 온도

확률변수 시각화

https://www.deepshark.org/courses/data_science/w/05_basic_statistics#random_variable_practice





확률분포(Probability Distribution)

확률분포(Probability Distribution)

- 확률변수(Random Variable)가 가질 수 있는 값들과 그 값들이 나타날 확률을 체계적으로 정리한 것
- 즉, "어떤 값이 얼마나 자주 나타나는가"를 수학적으로 표현한 것

이산형 확률분포

이산형 확률분포 (Discrete Probability Distribution)

이산형 확률분포란

- 확률변수가 가질 수 있는 값이 정수처럼 하나하나 세어질 수 있는 경우에 해당하는 분포
- 확률변수가 취할 수 있는 값들은 뚝뚝 끊어져 있으며, 각각의 값마다 확률이 주어진다.
- 확률변수의 가능한 값들은 유한 개일 수도 있고, 무한히 많더라도 셀 수 있는 경우라면 모두 이산형에 속함
  - 동전을 여러 번 던졌을 때 앞면이 나오는 횟수
  - 주사위를 던졌을 때 나오는 눈금
  - 편의점에 1시간 동안 들어오는 손님 수

이산형 확률분포에서는...

- 각 값에 대한 확률을 직접 더할 수 있음
- 모든 확률의 합은 항상 1이 된다.
- 이산형 확률분포는 결과가 셀 수 있는 값들로만 이루어져 있고, 각각의 값에 확률이 배분

이항분포, 포아송분포, 기하분포가 대표적임

확률질량함수 (Probability Mass Function, PMF)

확률질량함수 (Probability Mass Function, PMF)

- 이산형 확률변수의 분포를 정의하는 함수
- 확률변수 X가 특정 값 x를 가질 확률을 다음과 같이 표현

$$P(X = x) = f(x)$$

- 특징
  - 가능한 모든 값에 대한 확률의 합은 1
  - PMF를 알면 평균(기대값)과 분산 같은 통계적 특성을 계산 가능

$$E[X] = \sum_{x} x \cdot f(x)$$

$$Var(X) = \sum_{x} (x - E[X])^2 \cdot f(x)$$

이항분포 (Binomial Distribution) - 1/2

이항분포

- 어떤 사건이 두 가지 결과(성공/실패, 참/거짓, 앞/뒤 등)로만 나뉘는 상황에서, 동일한 확률을 가진 독립적인 시행을 여러 번 반복했을 때 성공이 발생하는 횟수를 나타내는 분포
- 예를 들어 동전을 10번 던져서 앞면이 몇 번 나오는지를 모델링

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

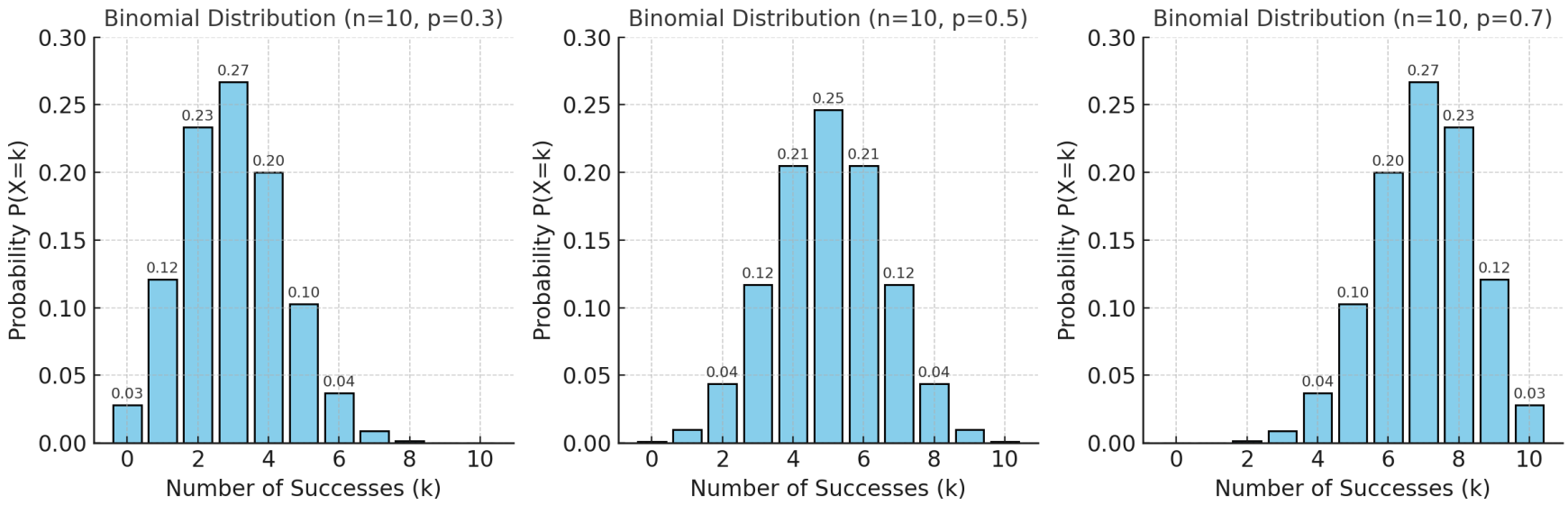
- n: 독립 시행의 횟수
- p: 한번의 시행에서 성공할 확률
- k: 성공 횟수

특징

- 모든 시행은 서로 독립
- 각 시행의 성공 확률은 동일
- 평균: $E[X] = np$
- 분산: $Var[X] = np(1-p)$

이항분포 (Binomial Distribution) - 2/2

$n = 10$ 일때 성공확률 p에 따른 이항분포 확률질량함수(PMF)의 변화



적용 분야

- 통계학: 설문조사 응답(성공/실패), 품질관리(불량품 개수), 임상시험 성공률 분석
- 산업: 생산 라인에서 불량품 비율 추정, 마케팅 캠페인의 성공 횟수 예측
- 자연과학: 유전자 발현 확률, 생물학적 성공·실패 실험

- 시행 횟수 n이 커지고 p가 적당할 때, 정규분포로 근사할 수 있다.
- n 이 크고 p가 매우 작을 때는 포아송분포로 근사할 수 있다.

포아송분포 (Poisson Distribution) - 1/2

포아송분포

- 일정한 시간이나 공간에서 드물게 일어나는 사건의 발생 횟수를 모델링할 때 사용되는 확률분포
- 사건이 발생하는 평균 횟수(λ)가 주어졌을 때, 실제로 관측되는 사건 횟수가 어떤 확률로 일어나는 지를 설명 (예: 한 시간 동안 콜센터에 걸려오는 전화의 수)

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \qquad k = 0, 1, 2, \cdots$$

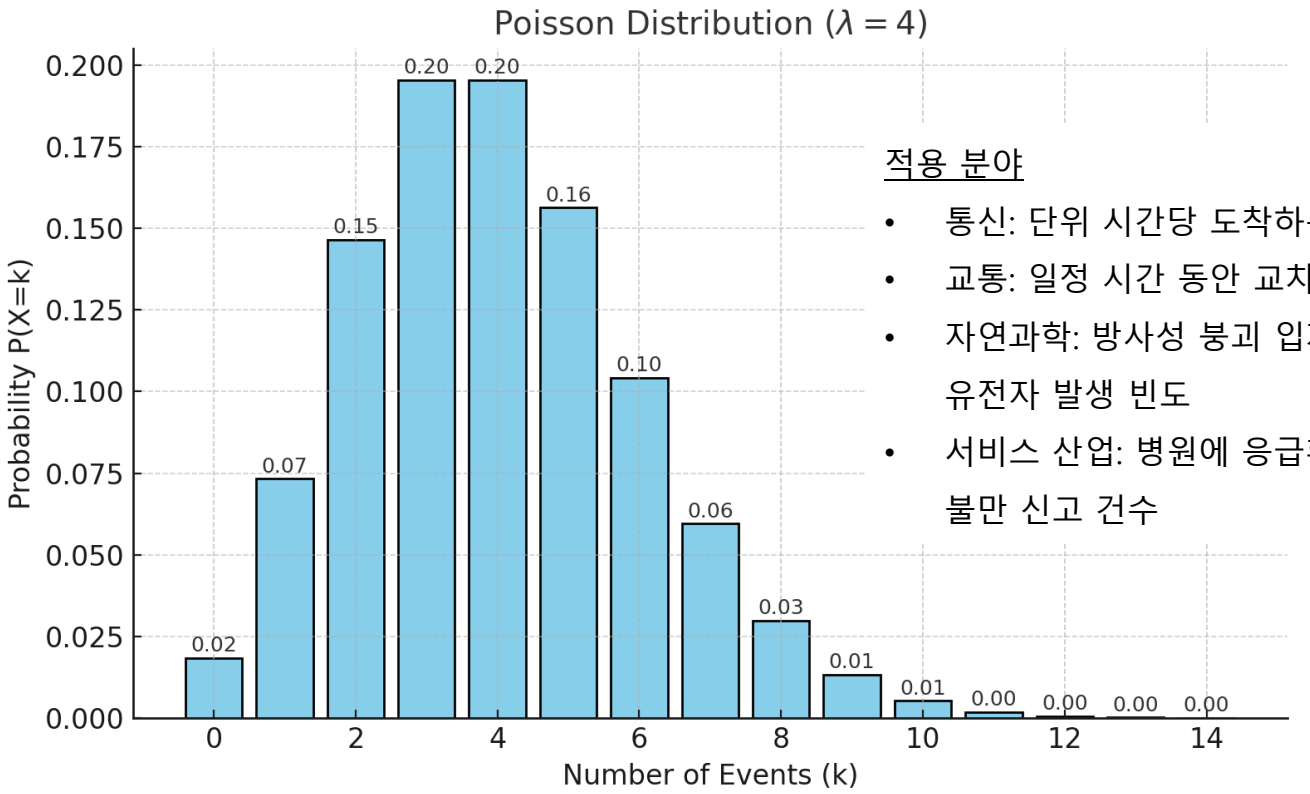
- λ: 단위 시간(또는 단위 공간)당 평균 발생 횟수
- k: 실제로 관측된 사건의 횟수

특징

- 사건은 서로 독립적으로 발생
- 사건발생은 짧은 시간 간격에서 비례적으로 발생
- 동시에 여러 사건 발생 확률은 매우 낮음
- 평균: $E[X] = \lambda$
- 분산: $Var[X] = \lambda$

포아송분포 (Poisson Distribution) - 2/2

■ **평균 발생 횟수 $\lambda = 4$ 일 때의 포아송분포 확률질량함수(PMF)**



적용 분야

- 통신: 단위 시간당 도착하는 전화 호출 수
- 교통: 일정 시간 동안 교차로에 도착하는 차량의 수
- 자연과학: 방사성 붕괴 입자의 발생 횟수, 돌연변이 유전자 발생 빈도
- 서비스 산업: 병원에 응급환자가 도착하는 횟수, 고객 불만 신고 건수

기하분포 (Geometric Distribution)

기하분포

- 어떤 사건이 두 가지 결과(성공/실패)로만 나뉘는 베르누이 시행에서, 첫 번째 성공이 나올 때까지의 시행 횟수를 확률변수로 하는 분포
- 예를 들어, 동전을 계속 던졌을 때 처음으로 앞면이 나올 때까지 걸린 횟수를 모델링하고 싶은 경우

$$P(X = k) = (1 - p)^{k-1}p,$$

$$k = 1, 2, 3, \cdots$$

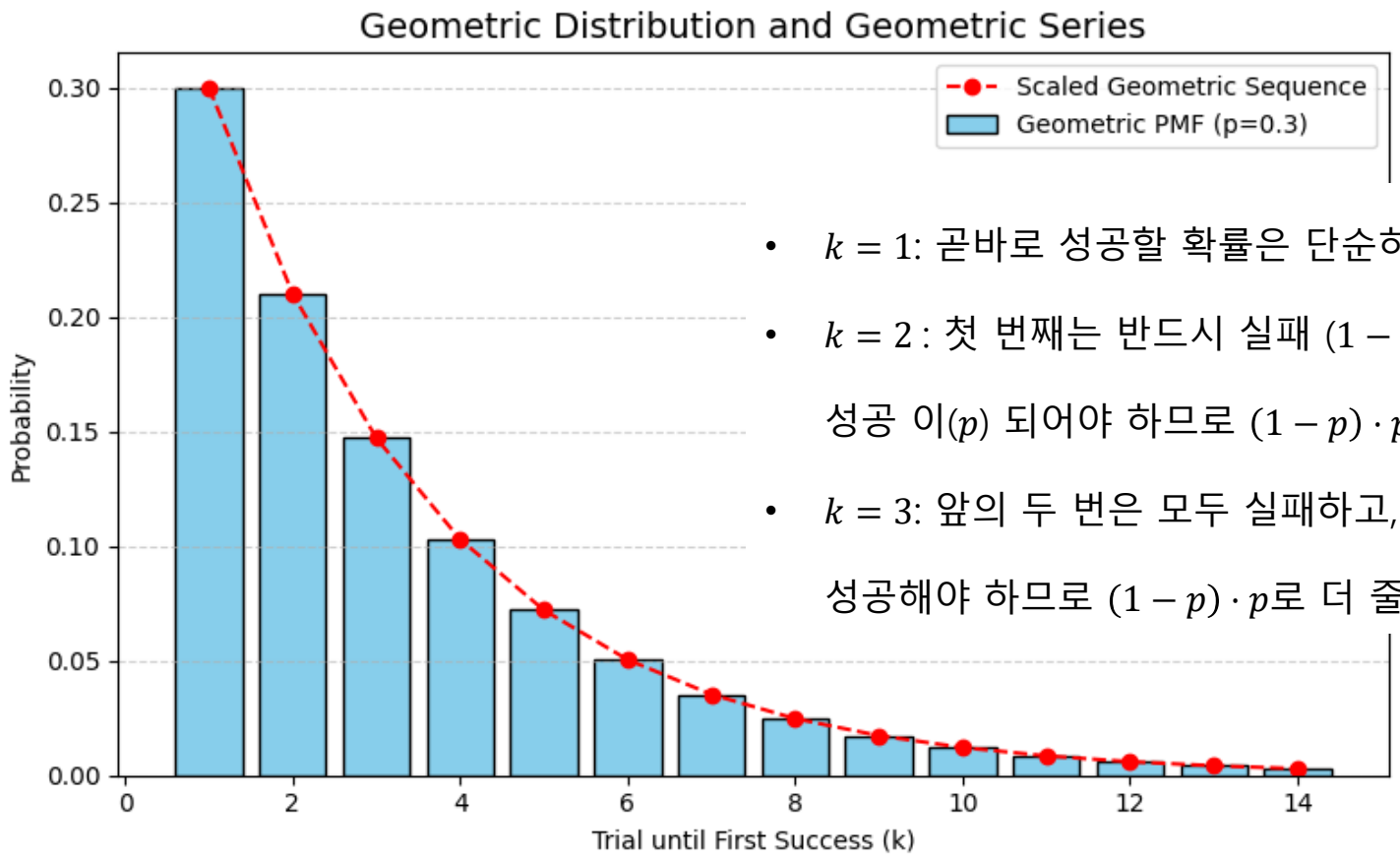
- p: 성공 확률
- k: 첫 성공이 일어나기까지의 시행 횟수

특징

- 모든 시행은 서로 독립
- 각 시행의 성공 확률은 동일
- 평균: $E[X] = \frac{1}{p}$
- 분산: $Var[X] = \frac{1-p}{p^2}$
- Memoryless (이미 s번 실패해도 앞으로 성공 확률은 동일)

기하분포 (Geometric Distribution)

■ 성공 확률 $p = 0.3$으로 나타낸 기하분포의 확률질량 함수



- $k = 1$: 곧바로 성공할 확률은 단순히 p이다.
- $k = 2$: 첫 번째는 반드시 실패 $(1 - p)$, 두 번째는 성공 이(p) 되어야 하므로 $(1 - p) \cdot p$가 된다.
- $k = 3$: 앞의 두 번은 모두 실패하고, 세 번째에 성공해야 하므로 $(1 - p) \cdot p$로 더 줄어든다.

연속형 확률분포

연속형 확률분포 (Continuous Probability Distribution)

- **확률변수가 특정 구간 내에서 무수히 많은 실수 값을 가질 수 있는 경우**
- **개별 값에 확률을 부여하지 않고, 구간에 대해 확률을 정의**
- **특징**
  - 특정 값 하나의 확률은 항상 0
  - 확률밀도함수(PDF) 적분으로 구간 확률 계산
  - 곡선 아래 면적이 전체 확률이며, 전체 면적 = 1
- **예시**
  - 온도, 키, 몸무게, 시간 등 → 연속형 확률변수로 다룸

확률밀도함수 (Probability Density Function, PDF)

확률밀도함수(Probability Density Function, PDF)

- 연속형 확률변수가 특정 구간에 속할 확률을 나타내는 함수
- 연속형 확률변수는 특정 값 하나에 대한 확률이 0이므로, $P(X = x)$ 형태로는 의미가 없음.
- 대신, PDF $f(x)$는 값의 "밀도(density)"를 표현하며, 구간 단위의 확률을 적분을 통해 계산

특징

- 모든 값에 대해 음수가 아니다 (Non-negative)

$$f(x) \geq 0$$

- 전체 영역에 대한 확률은 1이다. $\int_{-\infty}^{\infty} f(x)dx = 1$
- 임의의 구간 확률은 구간 적분으로 계산한다. $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = 1$

정규분포 (Normal Distribution)

정규분포

- 연속형 확률분포의 대표로, 평균을 중심으로 좌우가 대칭인 종(bell) 모양의 밀도 곡선
- 자연·사회에서 관측되는 많은 측정값(키, 시험 점수, 센서 노이즈 등)이

  여러 요인의 작은 영향이 합쳐져 나타날 때 정규분포로 잘 근사됨

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \Phi(z) = P(Z < z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

특징

- 대칭성: 평균 μ를 중심으로 좌우 대칭, 왜도는 0.
- 표준정규분포: 평균 $\mu = 0$ 이고, 표준편차 $\sigma = 1$ 인 경우, $Z \sim N(0,1)$로 표현
- 선형결합의 폐포성: 가 독립이면 도 정규분포
  - 폐포성(closure)이란 수학에서 어떤 연산을 해도 그 결과가 다시 같은 집합 안에 머문다는 성질
- 모집단에서 데이터를 여러 개를 뽑아 평균을 계산하면 그 평균값 자체도 확률변수가 되며 정규분포
- 최대엔트로피: 주어진 평균·분산 제약 하에서 엔트로피가 최대인 분포

정규분포와 가우시안분포의 비교

정규분포 (Normal Distribution)

- 통계학에서 주로 사용되는 용어
- 19세기 Karl Pearson이 "자연 현상에서 정상(normal)적인 분포"라는 의미로 명명
- 심리학, 사회과학, 통계 교재 등에서 흔히 사용

가우시안분포 (Gaussian Distribution)

- 수학·공학·물리학 분야에서 자주 사용되는 용어
- Carl Friedrich Gauss가 천문학 오차 분석에서 공식화
- 신호처리, 영상처리, 머신러닝 등 공학적 맥락에서 널리 사용

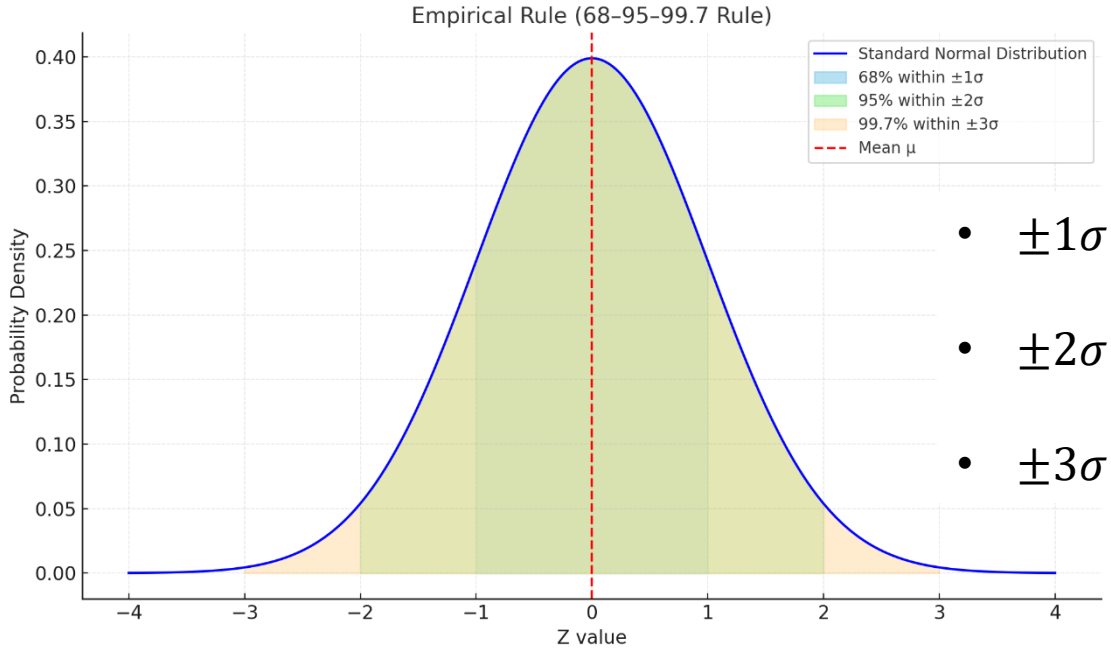
결론

- 차이는 사용되는 분야와 전통에 있음
- 정규분포 = 가우시안분포
- 맥락과 분야에 따라 명칭만 다를 뿐, 수식·성질은 완전히 동일함

Z 점수

표준화(Z-점수)

- 정규분포 $X \sim N(x; \mu, \sigma)$를 표준정규분포 $Z \sim N(0,1)$ 로 변환
- 임계값, 구간확률 계산 등에 사용
- 경험 규칙: 68–95–99.7 법칙



- $\pm 1\sigma$ 구간 (약 68%)
- $\pm 2\sigma$ 구간 (약 95%)
- $\pm 3\sigma$ 구간 (약 99.7%)

정규분포와 머신러닝

■ 회귀의 손실함수 해석

- 선형회귀에서는 예측값 $\hat{y}$ 와 실제값 y의 차이(오차)가 정규분포를 따른다고 가정

$$y = \hat{y} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

실제값은 예측값 주변에 정규분포 형태로 퍼져있다는 가정이 적용됨

- 예측 손실에 대한 확률 밀도 함수

$$p(y|\hat{y}) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right)$$

- 로그를 취하고 부호를 바꾸면

$$-\log p(y|\hat{y}) = \frac{(y-\hat{y})^2}{2\sigma^2} + constant$$

$(y-\hat{y})^2$ 항이 제곱오차를 의미 (Squared Error)

정규분포를 활용한 머신러닝 모델

정규분포의 활용

- 가우시안 나이브 베이즈: 연속 특성의 조건부분포를 정규로 가정해 사후확률 계산
- 가우시안 혼합모형(GMM): 여러 정규의 혼합으로 복잡한 분포 근사

딥러닝

- 가중 초기화·배치정규화에서 활성 분포를 근사 정규로 가정하는 논리.
- 확산모형은 학습·샘플링 과정에 가우시안 노이즈를 점진적으로 주입/제거.

모델	발표 시기	기관/회사	특징
Stable Diffusion XL (SDXL)	2023	Stability AI	- 오픈소스 텍스트-투-이미지 확산모델 - 뛰어난 해상도(1024×1024), 세밀한 컨트롤 - ControlNet, LoRA 등 커뮤니티 생태계 활성화
Imagen / Imagen 2	2022~2024	Google DeepMind	- 대규모 텍스트-이미지 모델 - 고해상도, 사실적 표현력 우수 - Imagen 2는 Parti와 결합해 향상된 세밀 묘사
DALL·E 3	2023	OpenAI	- GPT-4와 결합된 고품질 이미지 생성 - 프롬프트 해석 능력 강화, 복잡한 장면 표현

t-분포 (Student's t-Distribution)

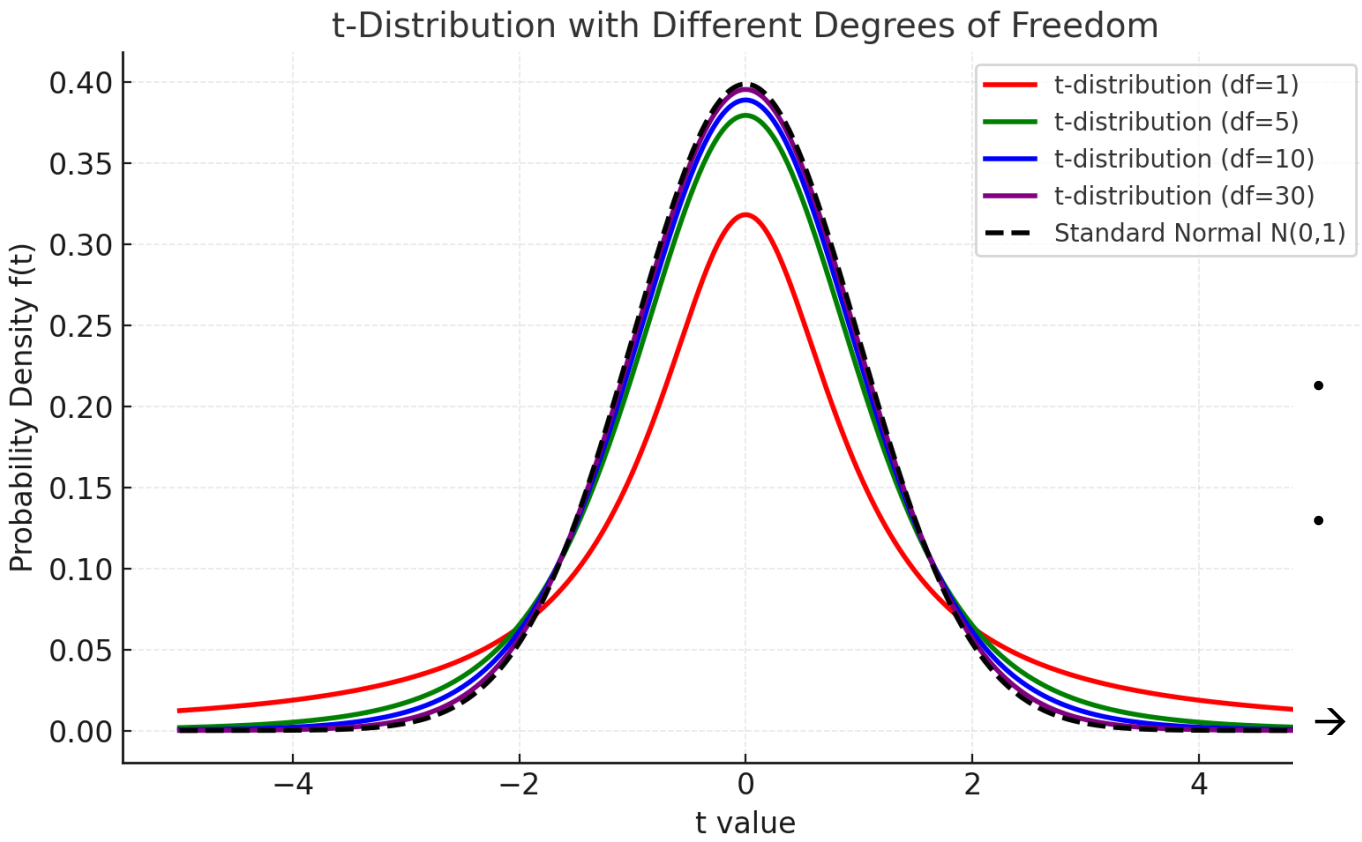
t-분포

- 표본 크기가 작거나 모분산을 알 수 없는 상황에서 평균에 대한 추론을 할 때 사용되는 연속형 확률분포
- 정규분포와 비슷한 종 모양을 가지고 있지만, 꼬리가 더 두꺼워 극단값(이상치)에 대한 확률을 더 크게 부여
- 자유도(degree of freedom)가 커질수록 정규분포에 점점 가까워지고, 자유도가 무한대에 수렴하면 정규분포와 동일

$$f(t;v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\cdot\Gamma\left(\frac{v}{2}\right)} \times \left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}}, \qquad -\infty < t < \infty$$

- v: 자유도 (일반적으로 $n-1$, 표본 크기 n일 때)
- $\Gamma(\cdot)$: 감마 함수 (https://en.wikipedia.org/wiki/Gamma_function)

t-분포 (Student's t-Distribution)



- 자유도가 작을수록 분포의 꼬리가 두껍게 나타남
- 표본 크기가 커져 자유도가 증가할수록 t-분포는 점점 정규분포에 가까워짐

→ 표본이 많아질수록 추정이 안정된다는 의미

자유도 (Degree of Freedom)

자유도 (Degree of Freedom)

$$DoF = n - (\text{제약조건 수})$$

- 자유도는 통계학에서 데이터가 가질 수 있는 독립적인 정보의 개수를 의미
- 즉, 전체 데이터 중에서 통계량을 계산할 때 제약 조건에 의해 자유롭게 변할 수 있는 값의 수

자유도에 대한 직관적 이해

- 표본의 크기가 n인 데이터가 $x_1,\ x_2,\ x_3$ 있다고 하자.
- 이 데이터들의 평균 $\bar{x}$를 이미 알고 있다면, 2개의 값이 정해지면 마지막 값은 자동으로 결정된다.
- 따라서 평균을 계산할 때 자유롭게 선택할 수 있는 값은 2개이며, 이를 자유도라고 한다.

중심극한정리 (Central Limit Theorem)

중심극한정리(Central Limit Theorem, CLT)

- 통계학의 핵심 정리로 표본의 크기가 충분히 크다면 모집단의 분포가 어떤 형태이든지

  표본평균의 분포는 정규분포에 가까워진다는 내용

- 즉, 원래 데이터가 정규분포가 아니어도,

  표본평균을 반복해서 구하면 그 값들의 분포가 점점 정규분포로 수렴

  · 독립이고 동일한 분포(i.i.d.)를 따르는 확률변수 $X_1,\ X_2, \cdots,\ X_n$ 이

  기댓값 $E[X_i] = \mu$, 분산 $Var(X_i) = \sigma^2$ 에 대하여 다음이 성립한다.

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ 에 대하여 다음이 성립 } \quad \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

  · 즉, n이 커질수록 표본평균의 표준화된 분포는 **표준정규분포**에 수렴

중심극한정리 (Central Limit Theorem)

■ 중심극한정리 직관적인 이해

- 주사위를 던지는 실험에서 한 번 던질 때의 결과는 균등분포(1~6)이다.
- 하지만 주사위를 30번 던져 평균을 구하는 실험을 수천 번 반복하면, 이 평균들의 분포는 정규분포 형태로 근사
- 즉, 원래 데이터의 분포가 어떤 모양이든 간에, 평균값의 분포는 정규분포로 수렴

■ 인공지능(AI)에서의 적용 사례

- 모델 성능 추정 및 불확실성 추정
  - 모델의 예측값에 대한 평균 손실(loss)을 여러 배치에서 측정하면,

    개별 손실의 분포가 어떻든 간에 배치 평균 손실은 정규분포로 근사
- 배치 정규화(Batch Normalization)
  - 미니배치(batch)마다 평균과 분산을 계산해 정규화할 때, 각 배치의 평균은 표본평균
  - 대규모 표본에서의 평균 추정
  - 수집된 로그 데이터나 센서 데이터 표본 평균을 이용한 추정에서는 CLT를 적용해 정규 근사

추정(Estimation)

추정(Estimation)

- 데이터 분석의 중요한 목표 중 하나는 표본(sample)으로부터 모집단(population)의 특성을 추론하는 것
- 추정은 관측된 표본 데이터를 이용해 모집단의 모수(parameter)를 예측하는 절차

점추정 (Point Estimation)

- 점추정은 모수(parameter)의 값을 하나의 구체적인 수치(점) 로 추정하는 방법
- 예를 들어, 모집단의 평균이나 분산 같은 모수를 모를 때, 표본 데이터를 이용해 이를 하나의 값으로 추정
- 점추정 예시
  - 어떤 학교 학생들의 평균 키를 알고 싶을 때, 전체 학생을 조사할 수 없다면 표본 50명을 뽑아 평균 키를 계산한다.
  - 이때 표본평균 $\bar{x}$ 는 모평균 μ의 점추정량이며, 계산된 값이 점추정치가 된다.
- 점추정치 하나만으로는 추정의 불확실성을 알 수 없음
- 실무 통계적 추론에서는 구간추정(Interval Estimation)과 함께 사용되어 신뢰도를 보완

구간추정 (Interval Estimation)

구간추정 (Interval Estimation)

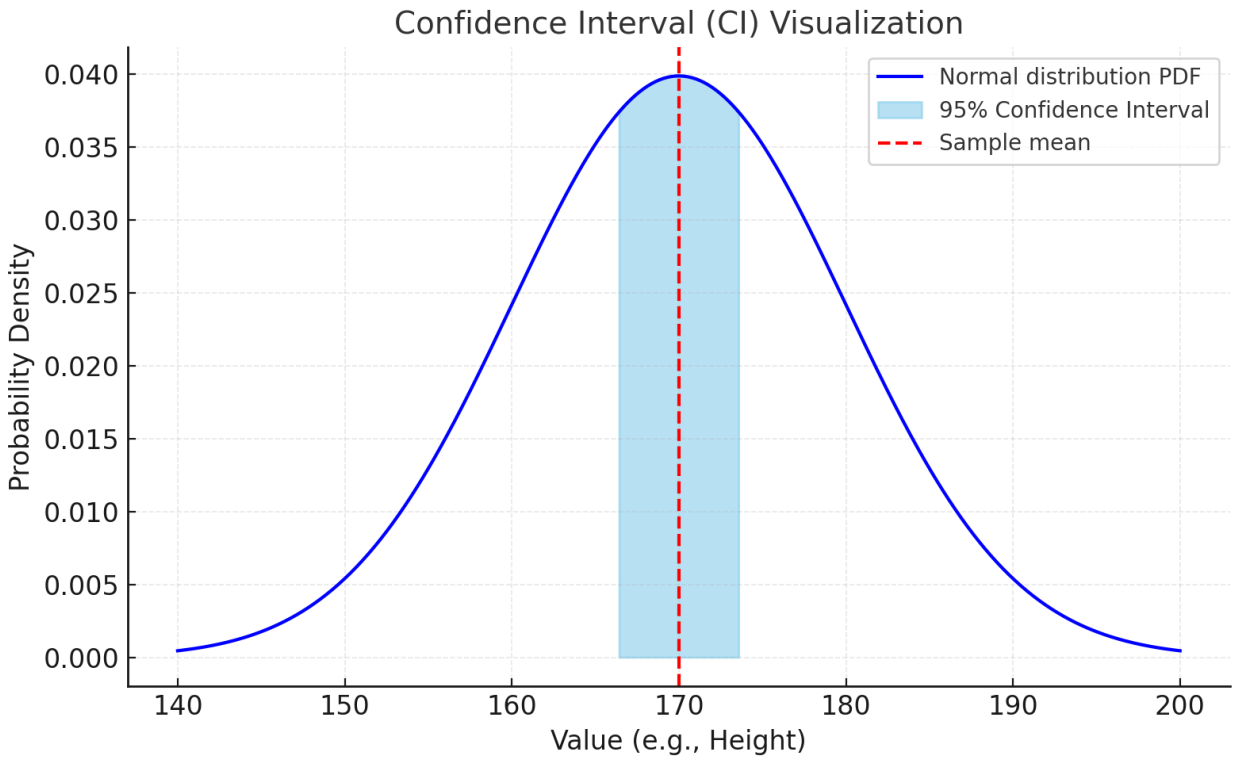
- 모수(parameter)의 값을 하나의 수치로 추정하는 점추정과 달리,
- 모수가 포함될 가능성이 높은 구간(interval)을 제시하는 방법
- 즉, 추정값에 대한 불확실성을 반영하여 "모수가 이 구간 안에 있을 것이다"라는 형태로 표현
- 보통 **신뢰구간(confidence interval, CI)** 을 사용

신뢰구간 (Confidence Interval)

- 모집단의 모수가 포함될 가능성이 높은 구간을 제시
- "모수가 이 구간 안에 있을 것이다"라는 정보를 제공

신뢰구간의 해석

- **모평균 μ의 95% 신뢰구간이 [166.08, 173.92]라고 한다면,**
  - "모평균 μ가 이 구간 안에 있을 확률이 95%다"라고 해석
  - "같은 방식으로 표본을 무수히 반복 추출하여 신뢰구간을 만들면, 그 중 약 95%가 실제 모평균 μ을 포함한다" 라고 해석



신뢰수준을 산출하는 방법

■ **표본평균 $\bar{x}$, 표본 표준편차 s, 표본크기 n이 주어졌을 때 모평균 μ의 신뢰구간은?**

- 모분산을 모르는 경우 (t-분포 사용)

$$\bar{x} \pm t_{\alpha/2,n-1} \times \frac{s}{\sqrt{n}}$$

- 모분산을 아는 경우 (정규분포 사용)

$$\bar{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$$

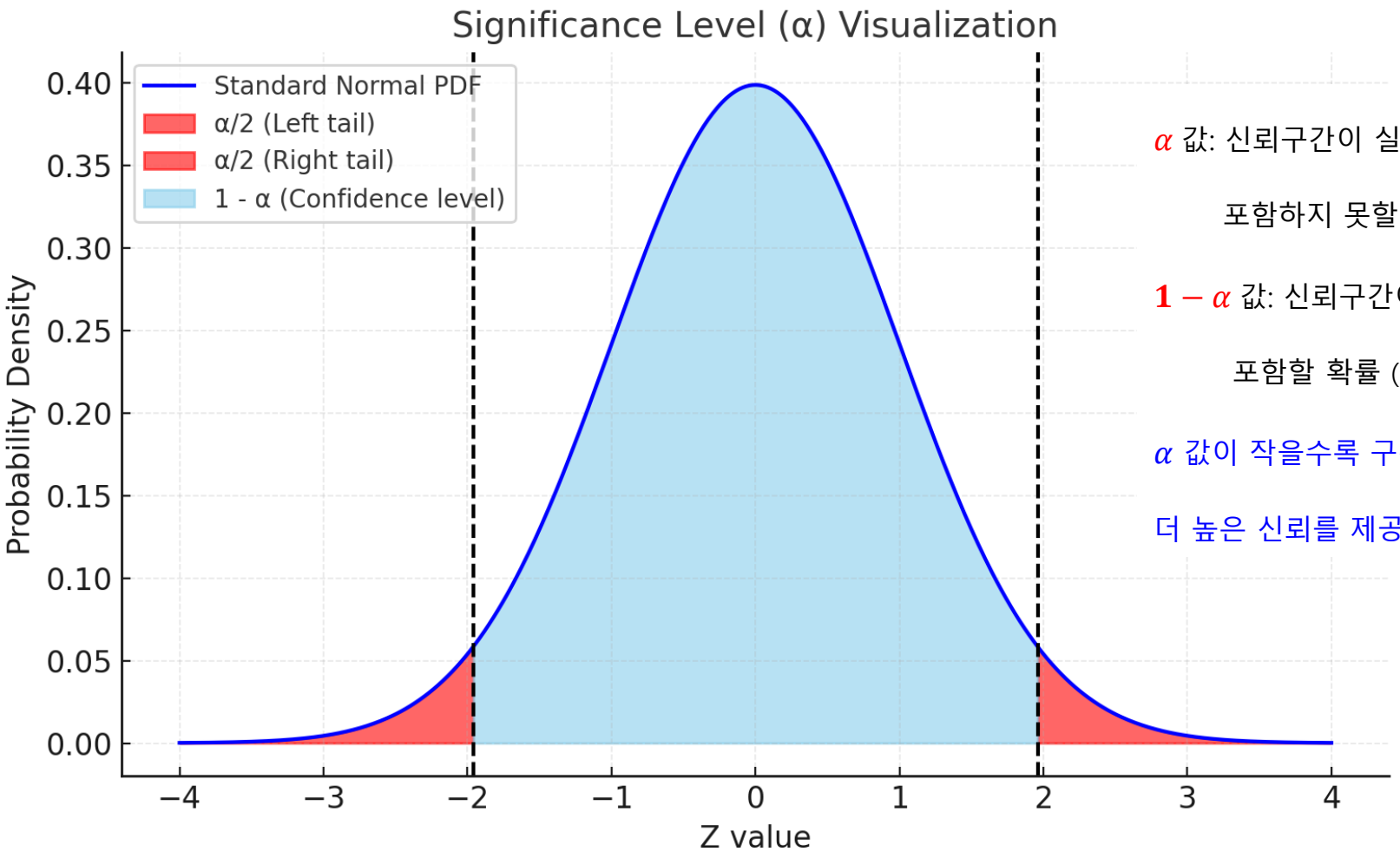
- 표본이 커질수록 t 값은 z 값에 가까워져, 결국 두 값은 동일하게 수렴한다.

- 유의수준 α의 의미

  · 신뢰구간에서 α 는 유의수준(significance level) 을 나타낸다.

  · 이는 곧 "실제 모수가 신뢰구간 밖에 있을 확률"을 의미한다.

유의수준 (Significance Level)



α 값: 신뢰구간이 실제 모수를 포함하지 못할 확률

$1-\alpha$ 값: 신뢰구간이 실제 모수를 포함할 확률 (신뢰수준)

α 값이 작을수록 구간은 넓어지고, 더 높은 신뢰를 제공

신뢰구간 예제 1/2

- **대한민국 군인 모집단 6만 명의 평균 키 μ를 추정할 때, 육군 10,000명의 표본을 구성하여 측정한 평균($\bar{x}$)은 170cm 이고, 표준편차가 2cm인 경우라고 하자.**

- **95% 신뢰구간(유의수준 $\alpha = 0.05$)은 무엇인가?**

  - 실제로는 모집단을 표준편차를 모르기 때문에 t 분포를 사용
  - 실제로 **샘플 수가 30개 이상**이면, 복잡하게 t 값을 찾지 않고 정규분포(z 값)을 써도 충분히 정확한 신뢰구간 추정 가능

자유도(df)	t 분포의 형태	정규분포와의 차이
5	꼬리가 매우 두꺼움	정규분포와 차이 큼
30	**조금 두꺼움**	**어느 정도 근사 가능**
100 이상	거의 정규분포와 동일	값과 값 차이 미미
10,000	완전히 정규에 수렴	$t_{\alpha/2,0000} \approx 1.96$

신뢰구간 예제 2/2

- **주어진 문제에서 자유도$(n-1)$가 매우 크기 $(30 \ll 30 \ll n-1)$ 때문에**

  $t_{\alpha/2,9999} \approx z_{\alpha/2} = 1.96$ **과 거의 동일하다.**

- **따라서 정규분포를 사용해도 된다.**

- **신뢰구간은**

$$170 \pm 1.96 \times 2 = [166.08, 173.92]$$

- "대한민국 군인의 평균키(모평균)는 95% 확률로 [166.08cm, 173.92cm] 구간 안에 있다"라고 해석

수고하셨습니다 ..^^..