Probability & Statistics

확률 기초

강사 양석환

확률과 확률 변수, 확률 밀도 함수

모집단과 표본

- **모집단 (Population)**
  - **관심있는 대상이 되는 모든 측정 값의 집합**
- **표본 (Sample)**
  - **모집단에서 추출한 측정 값의 집합**
- **모수 (Population Parameter)**
  - **모집단의 분포 특성을 규정짓는 척도**
  - **관심의 대상이 되는 모집단의 대푯값**
- **표본 통계량 (Sample Statistic)**
  - **표본의 대푯값**

경기도에 살고 있는 모든 고등학생의
키의 평균에 관심이 있다고 했을 때

전체 고등학생의 키를 측정하기 어려워
임의의 고등학생 30명을 추출해서 키를 측정하였다면...

- 모집단: 경기도에 살고 있는 모든 고등학생의 키의 집합
- 모수: 경기도에 살고 있는 모든 고등학생의 평균 키
- 표본: 임의로 선택한 고등학생 30명의 키
- 표본 통계량: 헤딩 30명의 키의 평균

확률의 개념

- **확률 (Probability)이란?**
  - 어떤 사건이 발생할 가능성을 0과 1 사이의 숫자로 수치화 시킨 것
  - $P(A) = \dfrac{\text{사건 } A\text{의 결과의 개수}}{\text{발생 가능한 모든 경우의 개수}}$
  - 관련 용어
    - 표본 공간(S 또는 Ω): 발생 가능한 모든 결과의 집합
    - 사건(Event): 표본 공간의 부분 집합
    - 단순 사건: 발생 가능한 결과들 중 하나만 발생한 사건

동전던지기를 생각해 보면...

Ω = {앞면, 뒷면}, 사건 A={앞면}

확률 $P(A) = \frac{3}{6} = \frac{1}{2} = 0.5$

집합으로 표현해 보면...

$P(\Omega \cap A)$ = {앞면과 뒷면이 동시에 나오는 경우} = 0

$P(\Omega \cup A)$ = {앞면 또는 뒷면이 나오는 경우}

$= P(\Omega) + P(A) - P(\Omega \cap A)$

$= \frac{1}{2} + \frac{1}{2} - 0 = 1$

확률의 종류

- **이론적 확률 (Theoretical Probability)**
  - 이론에 기반한 확률 → 아직 발생하지 않은 미래에 초점을 맞추는 방법
  - 예: 동전 던지기의 경우
    - 동전을 던질 때 앞면이 나올 확률은 50%이다

      → 아직 동전을 던지지는 않았으며 미래에 동전을 던진다면 앞면이 나올 확률은 50%이다.

- **경험적 확률 (Empirical Probability)**
  - 과거의 경험에 기반한 확률 → 과거의 경험에서 얻은 데이터를 기반으로 확률을 계산하는 방법
  - 예: 동전 던지기의 경우
    - 과거에 동전을 100번 던졌을 때 앞면이 53번 나왔다면 앞면이 나올 경험적 확률은 53%이다.

독립과 배반

- **독립 (Independent)**
  - **두 사건 A와 B가 독립이다**

    → 각각의 두 사건이 발생한 확률을 곱한 결과 = 두 사건이 동시에 발생할 확률

    → $P(A \cap B) = P(A)P(B)$

흔히 하는 오해
"두 사건이 독립이다 → 두 사건이 동시에 발생할 확률이 0 이다"
위의 내용은 독립(Independent)이 아닌 배반(Disjoint)의 개념

독립과 배반

- **배반 (Disjoint)**
  - 두 사건 A와 B가 상호 배반 관계이다

    → 두 사건 A와 B가 동시에 발생할 확률 = 0

    → $P(A \cap B) = 0$
  - 두 사건 A와 B가 상호 배반 관계라면 두 사건 중 적어도 한 사건 발생할 확률은 다음과 같음

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$= P(A) + P(B) - 0$$
$$= P(A) + P(B)$$
$$\therefore P(A \cup B) = P(A) + P(B)$$

  - 동전 던지기의 경우 앞면이 나오는 사건을 A, 뒷면이 나오는 사건을 B라고 하면 두 사건이 동시에 발생할 확률 $P(A \cap B)$는 0 이므로 이는 배반 사건에 해당함

독립과 배반

- Disjoint 식을 일반화 하면,
  - n개의 사건 $A_1, A_1, \cdots, A_n$이 배반일 경우
  - $P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = P\left(\sum_{i=1}^{\infty} A_i\right)$$

- 배반이 성립하지 않는 일반적인 경우에는
  - $P(A \cup B)$를 구하기 위하여 $P(A \cap B)$까지 고려해야 함 → 아래의 식이 성립함

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) \leq P\left(\sum_{i=1}^{\infty} A_i\right)$$

확률의 성질

- **확률의 기본적인 성질(공리적 정의)**
  - $0 \leq P(A) \leq 1$
  - $P(\Omega) = 1$
  - 만약 $A_1, A_2, \cdots$가 상호 배반 사건이라면

$$\mathrm{P} = \left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

확률의 활용

- 확률을 배우는 이유
  - 무작위로 일어나는 사건들을 이해하고 해석하기 위함
  - 확률을 공부함으로써 사건이 발생할 수 있는 확률을 기반으로 어떤 일을 계획하거나 결정할 수 있게 됨

- AI에서 확률은 어떻게 활용되나?
  - 현실 세계의 모든 현상들은 우연성을 가지고 있기 때문에

    → 이런 현상을 표현하려면 확률을 빼고는 설명하기 어려움
  - AI 분야에서는 상황을 판단하는 하나의 방법으로

    → 정답이 될 확률이 가장 높은 것을 정답으로 채택하는 방법을 자주 사용함

연습문제

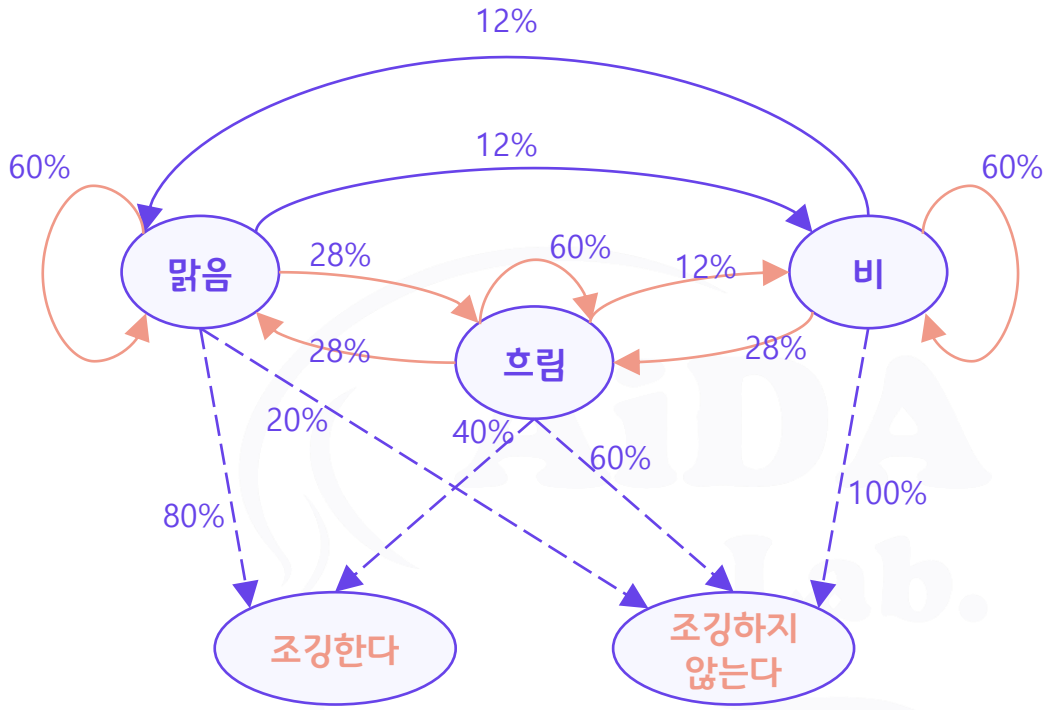
- 날씨를 '맑음'과 '비', 그리고 '흐림의 세 가지로 분류한다고 가정한다.
- 같은 날씨가 연속될 확률을 60%라고 하면, 날씨가 연속되지 않고 변할 확률은 40%가 된다.
- 단, 날씨가 변할 때 '맑음' → '흐림', '비' → '흐림', '흐림' → '맑음'으로 변할 확률은 각각 70%라고 하자.
- 오늘 날씨가 '맑음'일 때, 다음 질문에 대한 답을 구하라.

  - 모레의 날씨가 '비'가 될 확률을 구하라.
  - 조깅을 즐기는 A씨는 '맑음'인 날에 80%의 확률로, '흐림'인 날에 40%의 확률로 조깅을 한다. 만약 날씨가 '비'라면 그 날은 조깅을 하지 않는다. A씨가 오늘과 내일, 이틀 연속으로 조깅할 확률을 구하라.

연습문제

- 풀이
  - 주어진 조건에 의하여
    - 맑음 → 맑음 : 60%
    - 맑음 → 흐림 : 40% * 70% = 28%
    - 맑음 → 비 : 40% * 30% = 12%
  - 비가 올 때의 확률이나 그 날의 날씨에 따라 조깅을 할지 안 할지에 대한 확률을 구해보면...
    - 60% * 12% + 28% * 12% + 12% * 60% = 17.76%

      맑음 → 맑음 → 비 | 맑음 → 흐림 → 비 | 맑음 → 비 → 비
    - 상태 천이도에 따라 80% * (60% * 80% + 28% * 40%) = 47.36%



상태 천이도

확률 변수의 개념

- **확률 변수 (Random Variable)란?**
  - **확률 현상을 통해 값이 확률적으로 정해지는 변수**
    - 표본공간(Ω)이 주어져 있을 때, 하나의 함수 X가 모든 $c \in \Omega$에 대하여 딱 한 개의 숫자만을 할당하는 경우, $X(c) = x$ 를 말함
    - 확률 현상: 발생 가능한 결과들은 알지만 가능한 결과들 중 어떤 결과가 나올지는 모르는 현상
  - **확률 변수와 관련된 표기법(Notation)**
    - 확률변수는 대문자로 표기함. 소문자는 특정 값을 의미함
    - "확률변수 X가 특정 값 x일 확률"을 나타내는 표기는 → $P(X = x) \quad or \quad P_X(x)$
      - P는 확률을 의미 → 즉 P 다음에 나오는 괄호 안의 사건(event)이 발생할 확률
      - X는 확률 변수를 의미, x는 특정 값을 의미
      - $P(X = 3)$ → 확률 변수 X가 특정 값 3일 확률을 의미함

확률 변수의 종류

- 이산 확률 변수 (Discrete Random Variable)
  - 확률 변수의 값이 연속되지 않고, 셀 수 있을 만큼 뿔뿔이 흩어져 있는 것
  - 예시: 주사위의 면에 쓰인 숫자나 어떤 사건이 일어나는 시행 횟수 등

- 연속 확률 변수 (Continuous Random Variable)
  - 확률 변수의 값이 특정 범위 내에서 실수 형태로 존재하며, 소수점 이하까지 내려가는 것(연속 되는 것)
  - 예시: 키, 몸무게, 경과 시간과 같이 끊김이 없이 연속적으로 이어지는 값들

확률 분포의 개념

- **확률 분포(Probability Distribution)란?**
  - **확률 변수가 특정 값을 가질 확률을 나타내는 함수**
    - 확률 분포는 함수(Function)이며 함수는 대응(Mapping)을 의미

      → 확률 분포는 확률 변수와 특정 값을 Mapping 시켜주는 함수를 의미함

확률 분포의 종류

- 이산형 확률 분포 (Discrete Probability Distribution)
  - 확률 변수 X의 표본 공간이 셀 수 있는 경우

    → 확률 변수가 가질 수 있는 값의 종류를 셀 수 있는 경우
  - 확률 질량 함수를 가짐

- 연속형 확률 분포 (Continuous Probability Distribution)
  - 확률 변수 X의 표본 공간이 셀 수 없는 경우

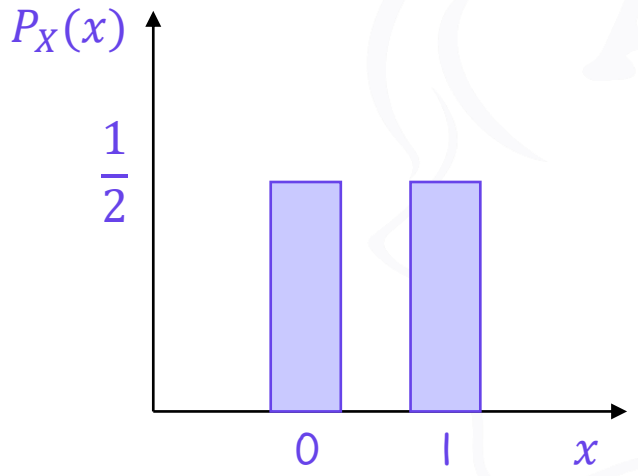
    → 확률 변수가 가질 수 있는 값의 종류를 셀 수 없는 경우
  - 확률 밀도 함수를 가짐

이산형 확률 분포

참고:
확률 질량함수와 함께 많이 사용되는 분포로 누적 분포 함수가 있음.
누적 분포 함수는 확률 변수 X가 취할 수 있는 값들을 누적해서 구하며
확률 변수가 특정 값보다 작거나 같을 확률을 의미함

- **확률 질량 함수 (Probability Mass Function, PMF)**
  - 이산형 확률 변수가 특정 값을 가질 확률
  - 확률 질량 함수의 성질
    - $P_X(x) \geq 0$
    - $\sum_x P_X(x) = 1$
  - 예시
    - 동전을 던졌을 때 앞면을 0, 뒷면을 1이라고 하고
    - 발생하는 결과를 확률 변수 X라고 했을 때
    - 확률 질량 변수는 다음과 같음 →

x	$P(X = x)$
0	1
1	2



동전을 던졌을 때 확률 질량 함수

연속형 확률 분포

- 확률 밀도 함수 (Probability Density Function, PDF)
  - 확률 변수가 특정 값을 가질 확률이 0
  - 확률 밀도 함수의 성질
    - $f_X(x) \geq 0 \quad for \;\; all \;\; x$
    - $\int_{-\infty}^{\infty} f_X(x) = 1$
  - 확률 질량 함수에서는 특정 점이 확률의 의미를 가지지만, 확률 밀도 함수에서는 특정 범위(Range)가 확률의 의미를 가짐
    - $P(a < X < b) = \int_b^a f_X(t)dt$
  - 연속형 확률 변수의 경우, 다음의 확률은 모두 동일함
    - $P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b)$

확률 질량과 확률 밀도의 차이

- 질량: 어떤 물질에 존재하는 물질의 양을 의미

  → 확률 질량: 각 이산형 확률 변수가 가지는 확률의 양을 의미

  - $P(X = x) = 0.5, \quad x = 0, 1$ → 이산형 확률 변수가 0, 1을 각각 취할 확률이 0.5임을 알 수 있음

- 밀도: 단위 부피에 존재하는 질량의 양을 의미

  → 확률 밀도 함수의 표현식: $P(a < X < b) = \int_b^a f_X(x)dx$

  - 연속형 확률 변수에 대하여 a와 b 사이의 확률은 적분을 이용 → 확률 밀도= $f_X(x)$
  - 밀도는 단위 부피에 존재하는 질량의 양 → 위의 식에서 단위 부피는 dx → dx 는 길이이므로

    → 확률밀도란 단위 길이 dx 당 질량 $f_X(x)$를 의미함

    → $f_X(x)$가 크다는 것은 해당 단위 길이에서 밀도가 높다는 것을 의미, 해당 영역을 적분한 것이 확률이 됨

확률 분포의 활용

- AI에서 확률 분포는 어떻게 활용되나?
  - 어떤 현상에 대한 관측 결과들을 이산 확률 변수로 취급 →이에 대한 이산 확률 분포를 구할 수 있다면,
  - 다음에 일어날 사건에 대한 확률을 과거의 데이터로부터 추측할 수 있음

  - 적절한 연속 확률 분포를 선택한다면
  - 적은 수의 시행만으로도 앞으로 일어날 사건의 확률을 상당히 높은 정확도로 추측할 수 있음

결합 확률과 조건부 확률

- **결합 확률**
  - 서로 독립적인 사건 A 와 B 가 동시에 일어날 확률
  - 사건 A 와 사건 B 가 서로 독립된 사건일 때, 두 사건의 결합 확률은 다음과 같음

$$P(A \cap B) = P(A, B) = P(A)P(B)$$

- **조건부 확률**
  - 사건 B 가 일어난다는 것을 전제로 한 사건 A 의 확률

    → 사건 B 가 일어난 후에 사건 A 가 일어날 확률
  - 사건 B가 일어났을 때, 사건 A가 일어날 조건부 확률은 다음과 같음

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

결합 확률과 조건부 확률

- 예시
  - 병 A를 앓고 있는지 판정하는 양성판정 정확도가 90%인 검사기가 있고
  - 어떤 사람이 이 검사기로 검사를 시행해서 양성 판정이 나왔다면
  - 이 사람은 90%의 확률로 병에 걸려 있다고 이야기 할 수 있을까? → 말할 수 없음

  - 검사가 알려주는 확률과 우리가 알고 싶은 확률은 조건부 확률의 의미에서 정반대이므로.
  - 검사의 양성판정 정확도 '90%'는 검사가 병을 가진 사람을 정확하게 포착할 확률, 즉 병을 가지고 있다는 전제 하에 검사 결과가 양성일 확률이 90%임을 의미
  - 우리가 알고 싶은 것은 검사 결과가 양성이라는 전제 하에 병을 앓고 있을 확률

결합 확률과 조건부 확률

	전제	관심 사건	수학적 표현
검사의 정확도	병을 앓고 있다	검사 결과: 양성	P(검사 결과: 양성|병을 앓고 있다) = 0.90
우리의 관심사	검사 결과: 양성	병을 앓고 있다	P(병을 앓고 있다|검사 결과: 양성) = ...?

- 검사의 정확도만을 가지고 우리의 관심사인 '(양성인 사람이) 병을 앓고 있을 확률'을 알 수 없음
- 그러나 검사 대상인 질병의 유병률을 알고 있다면,
- 예를 들어 전세계 인구 중 1%가 병 A를 앓는다고 알려져 있다고 가정한다면...
- 그리고 음성판정 정확도(병 A가 걸리지 않은 사람이 실제로 테스트 결과 음성으로 나올 확률)도 양성판정 정확도와 마찬가지로 90%라고 가정한다면...

  → 검사 결과가 양성으로 나온 사람이 실제로 병 A를 앓고 있을 확률은 약 8.3%

결합 확률과 조건부 확률

$$
\begin{aligned}
P(\text{병}|\text{양성}) &= \frac{P(\text{양성}|\text{병})P(\text{병})}{P(\text{양성})} \\
&= \frac{P(\text{양성}|\text{병})P(\text{병})}{P(\text{양성}|\text{병})P(\text{병}) + P(\text{양성}|\text{무병})P(\text{무병})} \\
&= \frac{0.9 \times 0.01}{0.9 \times 0.01 + (1-0.9) \times 0.99} \approx 8.3\%
\end{aligned}
$$

병을 앓고 있지 않는 99% 인구 중에서
병이 있다고 오진을 받은 경우가
검사기가 병을 앓는 사람을
제대로 진단한 경우를 압도하게 됨

결합 확률과 조건부 확률의 활용

- AI에서 결합 확률과 조건부 확률은 어떻게 활용되나?
  - 예측 모델의 정밀도나 정확성을 표현할 때에는 다양한 방법으로 평가할 수 있음
  - 사용하려는 목적에 따라 지표를 잘 선택해서 사용할 필요가 있음
  - 인공지능 모델의 정확성을 표현할 때 사용하는 대표적인 지표로는 정밀도(Precision), 재현율(Recall), F값(F-Score) 등이 있음
    - 정밀도 등의 지표의 계산은 결합 확률과 조건부 확률을 활용하여 계산함

평균과 기댓값

- 평균 (Mean)
  - 데이터의 중심을 나타내기 위하여 사용하는 값
  - 어떤 값의 집합의 적절한 특징을 나타내거나 요약하기 위하여 사용하는 값
  - 중앙값과 최빈수를 포괄하는 값
  - 평균을 알면 데이터 전체를 알지 않아도 데이터가 평균을 중심으로 분포되어 있다는 사실을 추측할 수 있음
  - 평균을 통해 분포의 중심을 알 수 있다 → 분포의 위치를 알 수 있다 → 평균을 알면 위치를 알 수 있다 (이런 이유로 평균을 Location Parameter라고 함)
  - n개의 확률 변수가 각각 $x_1, x_2, \cdots, x_n$ 이라는 값을 가질 때 평균 값 $\overline{x}$ 는..

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

평균과 기댓값

- 기댓값 (Expected Value)
  - 확률 분포의 평균을 계산하는 것을 의미
  - 평균은 표본 데이터가 실제로 주어졌을 때 데이터의 평균을 직접 계산하는 것을 의미하지만
  - 기댓값은 데이터가 주어지기 전에 계산하는 것이며 앞으로 데이터가 주어졌을 때 기대되는 평균을 의미함

    → 데이터가 주어지면 평균, 주어지지 않은 상태라면 기댓값

    → 평균은 과거를 바라보는 단어, 기댓값은 미래를 바라보는 단어

평균과 기댓값

- **평균의 분류**
  - **모집단의 평균 (=모평균, Population Mean)**
    - 확률변수 X의 모평균: μ 또는 μ_X로 표기
    - 주로 확률 변수 하나를 다룰 때는 μ
    - 확률 변수가 2개 이상일 경우, 예를 들어 확률 변수 X와 Y를 다룰 때는 모평균을 구분하기 위해서 μ_X, μ_Y로 표기
  - **표본 평균 (Sample Mean)**
    - 표본의 평균
    - 확률 변수 X이 표본 평균: $\bar{x}$로 표기

이산형 확률 변수의 기댓값

- 이산형 확률 변수 X 의 확률 질량 함수가 $P(X = x_i)$라고 할 때,

$$E(X) = \sum_{i=1}^{\infty} x_i P(X = x_i) = \sum_{i=1}^{\infty} x_i P_X(x_i)$$

  를 이산형 확률 변수 X의 기댓값이라고 함

- 기댓값과 지시 함수의 개념을 이용하여 앞에서 정의했던 확률을 표현할 수 있음

  - 지시 함수 : $I_A(X) = \begin{cases} 1, & if\ X \in A \\ 0, & if\ X \in A^c \end{cases}$

지시 함수 (Indicator Function)
1 또는 0 값을 가질 수 있으며
X가 A에 포함되면 1, 포함되지 않으면 0 값을 가지는 함수

  - 지시 함수의 기댓값 : $E(I_A) = P(A)$
    - 지시함수의 기댓값은 해당 사건이 발생할 확률과 같음
    - $E(I_A) = \sum_{x \in A} x \cdot P(I_A) = 1 \cdot P_{I_A}(X) + 0 \cdot P_{I_{A^c}}(X) = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A)$

연속형 확률 변수의 기댓값

- 연속형 확률 변수 X가 구간 $[a, b]$에서 모든 값을 취할 수 있고, 확률 밀도 함수가 $f_X(x)$일 때,

$$E(X) = \int_b^a x f_X(x) dx$$

를 연속형 확률 변수 X의 기댓값이라고 함

기댓값의 성질

- $E(a) = a$
- $E(aX) = aE(X)$
- $E(X + Y) = E(X) + E(Y)$
- $E(X - Y) = E(X) - E(Y)$
- $E(aX + bY) = aE(X) + bE(Y)$
- $E(aX - bY) = aE(X) - bE(Y)$

X, Y : 확률 변수
a, b : 상수

표본 평균

- 표본 평균은 모평균의 추정치로 사용할 수 있음
- 표본 평균을 구하기 위해서는 모집단에서 표본을 추출한 후 평균을 구함
- 표본 평균은 확률 변수임
  - 예시:
    - 임의의 모집단에서 표본으로 5명을 추출해 표본 평균 키를 구한다고 할 때,
    - 표본 평균 값은 표본으로 누구를 추출하느냐에 따라 달라짐 → 확률에 따른 변수
  - 표본 평균이 확률 변수라는 의미는 고정된 값이 아니라는 의미

    → 평균과 분산을 가진다는 의미
- 표본 평균의 기댓값은 모평균

분산과 표준 편차

- 분산 (Variance)이란 데이터가 흩어져 있는 정도를 의미함
  - 평균 값 만으로는 데이터의 분포 상태를 알 수 없음

    → 각 데이터가 평균 값으로부터 얼마나 떨어져 있는지 알 수 없음
- 평균 값으로부터의 차이 값인 편차(Deviation)를 확인
  - 편차의 총 합은 0 → 애초에 평균 값을 중심으로 계산되었기 때문. (+)와 (-) 값이 상쇄됨
  - 단순히 편차를 구해서 합치는 것으로는 데이터의 흩어진 정도를 확인할 수 없음
- 편차의 (+)와 (-) 값 상쇄를 피하기 위하여 부호를 제거 → 제곱을 이용
  - 편차를 제곱한 다음 합계를 구하고 이것을 다시 평균 값으로 만듦 → 분산 σ^2
- 분산은 제곱 값이므로 단위 표현이 어려움 → 제곱근 계산 → $\sqrt{\sigma^2} = \sigma$ → 표준편차

평균, 분산, 표준편차의 활용

- AI에서 평균, 분산, 표준편차는 어떻게 활용되나?
  - 평균과 분산, 그리고 표준편차는 과거의 데이터로부터 어떤 특징이나 경향을 밝혀낼 수 있는 가장 기본적인 방법임
  - 인공지능 모델을 만들기 전에 데이터의 특징을 파악할 때 사용함

공분산과 상관계수

- 공분산(Covariance)이란?
  - 분산과 표준편차를 사용하면 데이터가 얼마나 흩어져 있고 얼마나 차이가 심한지 알 수 있음
  - 기본적으로 평균과 분산, 표준 편차는 데이터의 경향을 표현할 때 사용됨
  - 그러나 데이터 각각이 어느 정도의 관계성을 가지는지를 파악하기는 부족함
  - 따라서 분산이라는 개념을 확장하여 두 개의 확률 변수의 흩어진 정도를 계산함 → 공분산
  - 공분산은
    - 두 개의 확률 변수의 선형 관계를 나타내는 값
    - 한 확률 변수의 증강에 따른 다른 확률 변수의 증감의 경향에 대한 측도가 됨

공분산과 상관계수

- **공분산**
  - 두 가지 데이터에 대한 n조의 확률변수 $(X, Y) = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$가 있다고 가정하면
  - X 의 평균이 μ_x이고 Y의 평균이 μ_y라고 할 때, 공분산 $Cov(X, Y)$는 다음과 같다.

$$Cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

  - 공분산의 계산에서는 단위에 대하여 신경을 쓸 필요가 없음
    - 처음부터 서로 다른 두 데이터 간의 관계를 표현하는 지표이므로 단위의 차이에는 의미가 없음

공분산과 상관계수

- 공분산의 결과
  - 양의 값을 가질 때
    - 두 가지의 데이터는 양의 관계가 있음 → 두 데이터 중 어느 한 쪽이 증가할 때, 다른 한 쪽도 증가함
  - 음의 값을 가질 때
    - 두 가지의 데이터는 음의 관계가 있음 → 두 데이터 중 어느 한 쪽이 증가할 때, 다른 한 쪽은 감소함
  - 그러나 공분산의 절대값이 크다고 해서 양의 관계나 음의 관계의 강도가 더 세다고 말할 수는 없음

    → 표준편차와 공분산을 통해서 계산할 수 있는 상관계수를 통해서 비교할 수 있음

공분산과 상관계수

- 데이터 간의 상관 관계
  - 두 가지의 데이터 간의 상관 관계는 공분산을 통해서 확인할 수 있음(양/음의 상관관계)
  - 서로 다른 단위 사이의 계산이므로
    - 단위 A와 단위 B 사이의 관계와 단위 A와 단위 C와의 관계를 직접 비교할 수 없음(의미가 없음)
    - 따라서 상관계수(Correlation Coefficient)를 도입함

공분산과 상관계수

- 상관계수(Correlation Coefficient)
  - 공분산을 각각의 표준편차로 나누어 단위를 없애 버린 값

    → 단위가 없는 무차원 수(Dimensionless Number)
  - 확률변수 X 와 Y의 분산이 양수이고 각각의 표준편차가 σ_X, σ_Y, 공분산이 σ_{XY} 라고 할 때

    상관계수는 다음과 같음

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \qquad (-1 \le \rho \le 1)$$

  - 상관계수 계산 시 공분산을 표준편차의 곱으로 나누면서 ρ는 -1에서 +1 사이의 값을 가지게 됨

    → 정규화
  - 값이 제 각각이었던 공분산을 상관계수로 변환함에 따라 상관관계의 강약을 비교할 수 있게 됨

공분산과 상관계수의 활용

- AI에서 상관관계는 어떻게 활용되나?
  - 사람이 직관적으로 분석하기 어려울 만큼의 대규모의 데이터가 있다면
  - 컴퓨터를 이용하여 무수히 많은 파라미터를 조합하고 그들의 상관계수를 계산함으로써
  - 상관관계가 강한 조합을 찾아내도록 만들 수 있음
  - 이런 과정을 거치면 사람이 미처 발견하지 못했던 숨은 관계나 데이터의 특징을 찾을 수 있으므로
  - 데이터를 보다 유용하게 활용할 수 있음