

확률 통계 기초

⇒ 표본공간 (Sample Space)

• 발생 가능한 모든 결과의 집합 (Ω, S)

ex) 동전 던졌을 때 나올 수 있는 결과의 표본공간 $\Omega = \{\text{앞}, \text{뒤}\}$

주사위를 던졌을 때 나올 수 있는 결과의 표본공간 $\Omega = \{1, 2, 3, 4, 5, 6\}$

⇒ 사건 (Event)

• 표본공간의 부분집합

• 표본공간에서 특정한 결과 A가 나오면, "사건 A가 발생했다"라고 함

ex) 주사위를 던졌을 때 홀수가 나오는 사건 $B = \{1, 3, 5\}$

⇒ 확률 (Probability)

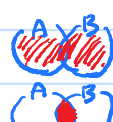
• 어떤 사건이 발생하는 가능성을 0과 1사이의 숫자로 나타냄

• 어떤 사건 A의 확률은 $P(A)$ 로 표기

$$P(A) = \frac{\text{사건 A의 원소의 개수}}{\text{발생 가능한 모든 결과의 개수}}$$

⇒ 여러 사건에 대한 확률

• 사건 A 또는 사건 B가 나오는 사건은 $A \cup B$ 라고 표기



• 사건 A와 사건 B가 동시에 나오는 사건은 $A \cap B$ 라고 표기



$$\rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad P(A \cap B) = P(A, B)$$

⇒ 확률이 독립 (Independent) (두 사건의 관계)

• 두 사건이 발생한 확률을 곱한 것과, 두 사건이 동시에 나올 확률이 같은 경우

$$\rightarrow P(A \cap B) = P(A) \cdot P(B)$$

⇒ 확률이 배반 (Disjoint) (두 사건의 관계)

• 두 사건이 동시에 발생한 확률이 0인 경우

$$\rightarrow P(A \cap B) = 0 \rightarrow A \text{가 발생하면 } B \text{는 절대 발생하지 않음}$$

• A, B가 서로 배반인 경우, 다음이 성립

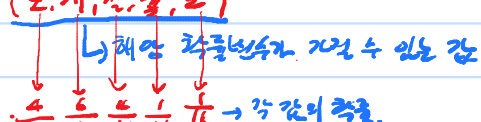
$$\rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B)^0 = P(A) + P(B)$$

⇒ 확률 변수 (Random variable)

• 어떤 시험의 점수에 따라 값이 순서적 수를 갖는 변수

ex) 주사위를 던졌을 때 나오는 눈의 값 (1, 2, 3, 4, 5, 6)

눈의 값에서 나올 수 있는 개 (도, 개, 닭, 물, 무)



• 확률 변수 X가 특정한 값 x를 가질 확률: $P(X=x) = P_X(x)$

⇒ 확률 분포 (Probability distribution) ⇒ 인공적으로 만든 확률 분포를 인공적으로 조작

• 어떤 확률 변수가 특정한 값을 가질 확률을 나타내는 함수

• 확률 변수와 특정한 값을 연결해 주는 개념

• 이산 확률 분포

→ 확률 변수가 가질 수 있는 값이 여러 개가 아닌, 정수인 경우

→ 이산 확률 분포를 나타내는 함수: 확률 질량 함수

ex) 이산 확률 분포, 베르누이 분포, 이항 분포...

주사위를 던졌을 때 나오는 눈의 값에 대한 분포는 이산 분포인 값 1~6에 대해

이산 확률 분포를 갖음

• 연속 확률 분포

→ 연속 확률 분포를 나타내는 함수: 확률 밀도 함수

ex) 정규 분포, 카이제르 분포, 감마 분포

• 확률의 합은 항상 1

• 확률은 0보다 크거나 같음

• 확률 밀도 함수 / 확률 질량 함수의 범위는 1

⇒ 평균과 기댓값

• 평균 (Mean): 데이터가 실제로 주어졌을 때 사용

• 기댓값 (Expectation): 확률 분포가 주어졌을 때 사용

⇒ 이산형 확률 변수의 기댓값

• 이산형 확률 변수 X의 확률 질량 함수가 $P(X=x_i)$ 인 경우,

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

⇒ 연속형 확률 변수의 기댓값

• 연속형 확률 변수 X가 구간 $[a, b]$ 에서 모든 값을 취할 수 있고,

확률 밀도 함수가 $f_X(x)$ 인 경우,

$$E(X) = \int_a^b x f_X(x) dx$$

⇒ 기댓값의 성질

• 확률 변수 X, Y와 상수 a, b에 대해 다음이 성립

1. $E(a) = a$
2. $E(aX) = aE(X)$
3. $E(X+Y) = E(X) + E(Y)$
4. $E(X-Y) = E(X) - E(Y)$
5. $E(aX+bY) = aE(X) + bE(Y)$

⇒ 분산 (Variance)

$$\hookrightarrow Var(X) = E[(X-\mu)^2]$$

• 표준편차 (Standard deviation): 분산의 양의 제곱근

$$\sqrt{E[(X-\mu)^2]}$$

⇒ 이산형 확률 변수의 분산

$$\hookrightarrow \sigma^2 = Var(X) = \sum_{i=1}^n (x_i - \mu)^2 \cdot P(X=x_i)$$

⇒ 연속형 확률 변수의 분산

$$\hookrightarrow \sigma^2 = Var(X) = \int_a^b (x - \mu)^2 f_X(x) dx$$

⇒ 표본의 분산

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

→ Sample의 개수
→ mean
→ 표본의 분산을 이용하는 예제를 찾아서, n개의 데이터가 아닌 n-1개의 데이터로 계산

⇒ 분산의 성질

1. $Var(X) = E[(X-\mu)^2] = E(X)^2 - \mu^2$
2. $Var(aX+b) = a^2 Var(X)$
3. $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$
4. $Var(X-Y) = Var(X) + Var(Y) - 2Cov(X, Y)$
5. $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$

⇒ 공분산 (Covariance)

• 두 확률 변수의 상관관계를 나타내는 값

$$Cov(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

• $Cov(X, Y) > 0$: X, Y는 양의 상관관계가 있음

• $Cov(X, Y) < 0$: X, Y는 음의 상관관계가 있음

• $Cov(X, Y) = 0$: X, Y는 상관관계가 없음 (독립사건으로 상관관계가 없음)

→ 독립사건: $E(XY) = E(X)E(Y)$

$$Cov(X, X) = Var(X)$$

$$Cov(X, Y) = Cov(Y, X)$$

$$Cov(aX, bY) = abCov(X, Y)$$

$$Cov(X+Y, Z) = Cov(X, Z) + Cov(Y, Z)$$