Chapter4. 확률과 통계

01 확률과 의사결정

- 통계의 목적 "표본으로부터 모수(parameter) 추정"

  but, 표본으로부터 모수 추정 시, `오차 발생`

  => 통계량이 모수와 일치할 확률 나타냄으로 해결

확률론

수학적 확률

- 통계 자료의 결과를 확률(probability)와 함께 표현
  - 전혀 맞지 않을 확률: 0%, 모두 맞을 확률: 100%
  - 일정한 조건 아래, 동일한 실험 지속적으로 N회 반복 시, 사건 A가 n번 발생할 확률
    - P(A) = n(A)/N
  - 확률은 아래 조건 만족
    1. 확률은 0~1의 값을 가진다.
    2. 모든 사건에 대한 확률의 합은 1이다.

$$\sum_{i=1}^{n} P(E_i) = 1$$

```
(E:사건(Event), i: 시행 횟수, P: 확률)
```

통계적 확률

- 사건 실행 횟수:n , 사건 A가 일어난 횟수: r
  - n을 충분히 크게 한다면, 상대도수로 나타나는 r/n은 일정한 확률값 p로 근사함. (이때, p = 사건 A가 발생할 `통계적 확률` or `경험적 확률`)
    - n을 무한대로 수렴 시,

$$r \div n = P(A)$$

확률의 덧셈법칙

- $$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
  - 확률사건
    - $$합사건(A \cup B)곱사건(A \cap B)배반사건(A \cap B \neq \varnothing)여사건(A^C)$$

조건부 확률과 확률의 곱셈법칙

- 조건부 확률
    - 사건 A가 먼저 발생하고 이어서 B가 발생하는 확률
        - P(B|A)
- 확률의 곱셈법칙
    - $$P(B \mid A) = P(A \cap B) \div P(A)$$
    - $$P(A \cap B) = P(A) \times P(B \mid A)$$

확률변수와 확률함수

- 확률변수(random variable)

  : 실험결과(사건)에 실수값을 대응시키고 그 값에 확률을 부여한 것

  : 결과의 수에 확률이 부여된 것

  : 확률변수 X로 표현
    - 이산 확률변수(discrete random variable)

      : 셀 수 있는 특정한 값들로 구성, 일정한 범위로 나타냄
    - 연속 확률변수(continuous random variable)

      : 연속형, 무한한 경우와 같이 셀 수 없는 경우
- 확률함수

  : 확률 P를 가진 어떤 사건이 n회 시행 중에서 x회 나타날 때, 확률변수 x와 이에 대응되는 P(x)의 관계를 나타낸 함수

  : 확률함수를 논하려면, 표본의 개수가 많아야 함. (최소 30개 표본 필요)

---

02 확률변수의 평균과 분산

표본의 특성 파악 시, 표본으로부터 얻은 자료에서 평균과 분산 구하여 표본분포의 특성 확인

확률변수를 통해 확률을 확인하고 확률분포를 구성하면 그 자료에 대한 판단, 예측 내리기 수월해짐

확률변수의 평균

통계학에서 확률변수의 평균 = 기대값

기대값(expected value) : 어떤 사건에 대해 그 사건이 벌어질 확률을 곱해서 전체 사건에 대해 합한 값

$$E(X) = \sum_{i=1}^{n} x_i P(x_i) = x_1 P(x_1) + x_2 P(x_2) + \ldots + x_n P(x_n)$$

확률변수의 분산과 표준편차

- 확률변수의 분산

  : 기대값의 특성 나타내는 값

  : 기대값과 어느 정도 차이가 있는지

  - 평균(m)과 분산(var^2)
    - 평균(m)

$$m = 1 \div n \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i P(x_i)$$

  - 분산(var^2)

$$\sigma^2 = 1 \div n \sum_{i=1}^{n} (x_i - m)^2 \times (x_i 의 상대도수) = \sum_{i=1}^{n} (x_i - m)^2 \times (x_i 의 상대도수) \div n = \sum_{i=1}^{n} (x_i - m)^2 P(x_i)$$

  - 표준편차(var)

$$\sigma = \sqrt{V(X)} = \sqrt{\sum (x_i - m)^2 P(x_i)}$$

확률변수에서 평균과 분산(표준편차)의 성질

확률변수 평균의 성질

a가 상수, X와 Y가 확률변수일 때 다음이 성립

$$(1) E(a) = a$$

$$(2) E(aX) = aE(X)$$

$$(3) E(aX + -b) = aE(X) + -b$$

$$(4) E(aX + -bY) = aE(X) + -bE(Y)$$

$$(5) E(XY) = E(X)E(Y), X 와 Y 는 확률적으로독립$$

확률변수 분산(표준편차)의 성질

a가 상수, X와 Y가 확률변수일 때 다음이 성립

$$(1) V(a) = 0$$

$$(2) V(aX) = a^2 var(X)$$

$$(3) V(X + Y) = V(X) + V(Y) + 2COV(X, Y) V(X + Y) = V(X) + V(Y), X 와 Y 는 확률적으로독립$$

$$(4) V(X + Y) = V(X) + X(Y) + 2COV(X, Y) V(X + Y) = V(X) + X(Y), X 와 Y 는 확률적으로독립$$

$$(5) sigma(aX + b) = barabarsigma(X)$$

$$(6) V(X) = E(x^2) - [E(X)]^2$$

연습문제

#1

확률분포: 미래 발생할 사건에 대해 확률 나열한 것

ex) 12월에 눈이 온 날 수

#2

평균 = (0.04+0.18+0.19+0.21+0.18+0.13+0.07)/7 = 0.174

분산 =

$$(1//7 - 0.147)^2 * (0.04 + 0.18 + 0.19 + 0.24 + 0.18 + 0.13 + 0.07) = 1.768$$

#3

평균 = (0.89+0.68+0.59+0.42+0.31)/5 = 0.578

분산

$$= 1/5 * sum(x_i - x)^2$$

= 0.040936

: 약 57.8%의 생존율이 평균, 예외 확률이 0.040936%

#4

평균 =

$$0.11 * 1 + 0.32 * 2 + 0.21 * 3 + 0.15 * 4 + 0.11 * 5 + 0.09 * 6 + 0.01 * 7 = 3.14$$

분산

$$(1 - 3.14)^2 * 0.11 + (2 - 3.14)^2 * 0.32 + ... + (7 - 3.14)^2 * 0.01$$

= 2.3004

표준편차 = 1.516707

#5

(a)

$$1 * 0.02 + 2 * 0.05 + 3 * 0.07 + 4 * 0.15 + 5 * 0.22 + 6 * 0.20 + 7 * 0.14 + 8 * 0.08 + 9 * 0.02 = 5.03$$

(b)

$$(1 - 5.03)^2 * 0.02 + (2 - 5.03)^2 * 0.05 + (3 - 5.03)^2 * 0.07 + ... + (9 - 5.03)^2 * 0.02 = 2.984$$

(c)

$$1 - (0.02 + 0.05 + 0.07 + 0.15 + 0.22 + 0.20 + 0.14 + 0.08 + 0.02) = 0.05$$

#6

국내 투자 기대값

$$400(0.7 * 0.5) + 200(0.5 * 0.4) + 100(0.2 * 0.3) = 186$$

해외 투자 기대값

$$800(0.7*0.6)+400(0.5*0.4)+200(0.2*0.2)=424$$

국내 투자 분산

$$(100-186)^2*0.2*0.3+(200-186)^2*0.5*0.4+(400-186)^2*0.7*0.5=16,511.56$$

해외 투자 분산

$$(200-186)^2*0.2*0.2+(400-186)^2*0.5*0.4+(800-186)^2*0.7*0.6=61,500.16$$

: 기대값은 해외 투자가 더 높지만, 해외투자는 동시에 분산도 높다. 그러나 표준편차로 환산해보면, 국내 투자는 약 128, 해외투자는 약 248이므로 해외 투자가 국내 투자의 편차를 감안하더라도 높다. 따라서 해외 투자 선택.