Seoul National University

- 일반통계학 -
제3장
확률과 확률분포

1 확률의 정의

1.1 Why?

모집단에서 일부만 관측하고 이를 바탕으로 모집단 전체에 대한 결론을 이끌어 내는데 논리적 근거가 된다.

1.2 용어정리

(1) 표본공간(sample space)

- 통계적 조사에서 얻을 수 있는 모든 가능한 결과들의 집합

예) 공정한 주사위를 던지는 실험 S={ }

(2) 사건(event)

- 표본공간의 부분집합 (관심이 있는 실험 결과의 집합)

예) a. 주사위의 1과 2의 눈금이 나오는 경우

b. 전구의 수명시간이 10이상인 경우

(3) 근원사건(elementary event), 단순사건(simple event)

-한 개의 원소로 이루어진 사건

1.3 사건의 연산

사건 A,B에 대하여

- 합사건 : A∪ B
- 곱사건 : A∩ B
- 여사건 : A^c
- A∩ B $= \emptyset$ 이면 A와B 는 서로 배반사건

※ 공집합은 모든 집합과 배반임.

1.4 분할

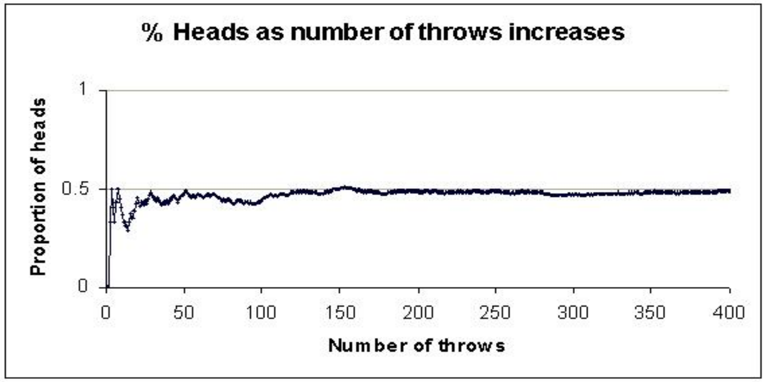
1.5 확률의 정의

<u>고전적 확률(Laplace 1749~1827) (또는 사전확률)</u>

- 각각의 근원사상이 일어날 가능성이 같다는 가정
- 표본공간 S가 n 개의 근원사건으로 이루어져 있고, 각 근원사건이 일어날 가능성이 같다면, m 개의 원소로 구성된 사건 A의 확률 $P(A) = m/n$
- 예:
- 단점 :

<u>상대도수확률 (또는 사후확률)</u>

- 무수히 많이 시행하였을 때 그 사건이 일어난 비율(relative frequency)이 수렴해 가는 값
- 동전을 던져서 나오는 앞면이 나올 확률 =0.5
- 단점:



1 확률의 정의

주관적 확률

•각자 생각하고 있는 어떤 사건이 일어날 가능성에 대한 믿음의 정도(degree of belief)-상대도수에 의존할 수 없는 사건에 적용

•예)

•단점 :

공리적 확률 - Kolmogorov(1903-1987)

• 현대수학자들이 보통 생각하는 확률개념

• 사전,사후,주관적 확률이 공통으로 갖는 성질/ 세 공리를 만족하면서도 한 사건에 대응되는 확률이 여러 개 존재할 수 있음.

(공리 1)

(공리 2)

(공리 3)

1 확률의 정의

(예제) 표본공간 $S = \{\mathrm{H}, \mathrm{T}\}$

사건의 집합={ }

- 사전 확률

- 사후확률

1.6 확률의 성질

(1) P(∅)=0

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

(3) $P(A^c)=1-P(A)$

(4) $A \subset B$ 이면 $P(A) \leq P(B)$

(5) 서로 배반인 사건 $A_i(i = 1,2, \cdots n)$에 대하여 $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$

2 조건부 확률과 독립사건

2.1 조건부 확률

- 사건 A가 주어졌을 때 사건B가 일어날 확률 (단, P(A)>0)

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

- 예) 두개의 주사위를 던지는 실험에서 첫 번째 던진 주사위 눈이 두 번째 던진 주사위 의 눈보다 클 때 두 주사위 눈의 합이 10일 확률은?

- 곱셈법칙 : $P(A) > 0, P(B) > 0$ 이면 $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

예) 불량품 20개와 양호품 80개로 구성된 로트에서 2개의 제품을 단순랜덤추출할 때, 2개 모두 불량품일 확률을 구하여라.

2.3 전확률 공식

- 사건 $A_1, A_2, \cdots, A_n$에 대하여 $A_i \cap A_j = \emptyset (i \neq j), A_1 \cup \cdots \cup A_n = S, P(A_i)>0$ 이면

$P(B) = P(B|A_1)P(A_1) + \cdots P(B|A_n)P(A_n) = \sum_{k=1}^{n} P(B|A_k)P(A_k)$

예) 통계학과 : 1학년(30%), 2학년(25%), 3학년(25%), 4학년(20%)
수학과목 수강 : 1학년의 50%, 2학년의 30%, 3학년의 10%, 4학년의 2%
통계학과 학생 중 한 학생을 단순랜덤추출하였을 때 그 학생이 수학 과목의 수강생일 확률을 구하여라.

2.4 베이즈 정리

표본공간 S의 분할 $A_1, A_2, \cdots, A_n$ 과 $P(A_i)>0$, $P(B)>0$ 에 대하여

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{k=1}^{n} P(B|A_k)P(A_k)}$$

예) 어떤 지역의 결핵환자의 비율이 0.001로 알려져 있다. 결핵에 결려있는지를 알아보는 검사에서 결핵에 걸렸을 때 양성 반응이 나타날 확률은 0.95이고 그렇지 않을 때 양성 반응이 나타날 확률은 0.011이라고 한다. 양성 반응이 나타났을 때 결핵에 걸렸을 확률을 구하여라.

2.5 서로 독립 (mutually independent)

- $P(B|A) = \mathrm{P}(B)$ 이면 사건 A와 사건 B는 서로 독립이라 한다.
- 사건 A, B가 독립이면 $P(A \cap B) = P(A)P(B)$, $P(A|B) = \mathrm{P}(A)$, $P(B|A) = \mathrm{P}(B)$

예) "불량품 20개와 양호품 80개로 구성된 로트에서 2개의 제품을 단순랜덤추출할 때, 2개 모두 불량품일 확률을 구하여라. " 이 문제에 제시된 두 사건은 독립인가? 만약 단순랜덤복원추출을 한다면 독립인가?

- $A_1, \cdots, A_n$에 대하여 다음이 성립할 때, $A_1, A_2, \cdots, A_n$은 서로 독립이라고 한다.

P($A_i \cap A_j$)=P(A_i)P(A_j) $(1 \leq i < j \leq n)$

P($A_i \cap A_j \cap A_k$)=P(A_i)P(A_j)P(A_k) $(1 \leq i < j < k \leq n)$

$\cdots$

P($A_1 \cap A_2 \cap \cdots \cap A_n$)=P($A_1$)P($A_2$) $\cdots$P(A_n)

3.1 동전 1개를 던져 나오는 결과를 관측하는 실험

- 표본공간 :S ,S={H,T}, H ='표면',T='이면',
- 확률 : P ({H})=0.5 P({T})=0.5
- 위와 같이 표본공간의 원소, 즉 가능한 결과가 숫자가 아닐 경우 표본공간과 확률을 나타내는 것이 복잡해진다. 따라서, 다음과 같이 표본공간 S에서 정의되어 실수값을 갖는 함수 $X : S \to \boldsymbol{R}$ 를 정의하고 X를 확률변수라고 부른다.

예) 함수 X=앞면의 수라 할 때, X(H)=1, X(T)=0이 되고 이 때 새로운 표본공간 S_X와 확률 P_X 를 얻을 수 있다.

(S_X, P_X) : 표본공간 $S_X = \{0,1\}$

확률 P_X : $P_X(\{0\})=P(\{T\})=0.5$, $P_X(\{1\})=P(\{H\})=0.5$

우리는 확률변수 X를 이용해서 P({H})=P(X=1)이라고 간단히 표기한다.

3.2 확률변수

- 표본공간 위에 정의 된 실수값 함수 ($X : S \longrightarrow \mathbf{R}$)

(1) 이산형 확률변수

- 확률변수 X 가 취할 수 있는 모든 값이 셀 수 있을 경우

예) 동전을 2번 던지는 실험을 한다. X=앞면의 수

(2) 연속형 확률변수

-확률변수 X 가 어떤 구간 내의 모든 값을 취할 수 있는 경우

예) 표본공간 $S=\{x : x \in [0,1]\}$ 일 때, 확률변수X 를 바늘이 가리키는 눈금 이라고 하면, 확률변수X 가 취할 수 있는 값은 $\{x: x \in [0,1]\}$ 이다.

3.3 확률분포

-확률변수의 값에 따라 확률이 어떻게 흩어져 있는지를 합이 1인 양수로서 나타낸 것

예) S={H,T} , X=동전의 앞면의 수

x	0	1	합계
P(X=x)	1/2	1/2	1

-확률밀도함수(probability density function)에 의해 확률분포를 편리하게 표현

- 이산형 확률변수 X의 확률밀도함수

(일반적으로 확률질량함수라고 부름, **p**robability **m**ass **f**unction , **pmf**)

a. 확률변수 X가 취할 수 있는 값이 $x_1, x_2, x_3,$...일 때, 확률질량함수 $p(x)$를 다음과 같이 정의한다.

$$p(x) = \begin{cases} P(X = x_i) & x = x_i, i = 1,2,\dots \\ 0 & o.w \end{cases}$$

b. 확률질량함수의 성질

$(1)\, 0 \le p(x) \le 1,$

$(2)\, \sum_{all\ x} p(x) = 1$

$(3)\, P(a < X \le b) = \sum_{a<x\le b} p(x)$

(예) 동전을 두번 던지는 실험에서 확률변수 X=앞면의 수라고 할 때, 확률분포 및 확률질량함수를 구하여라.

- 연속형 확률변수의 확률밀도함수(**p**robability **d**ensity **f**unction, **pdf**)

a. 연속확률변수 X에 대하여 함수 $p(x)$가

(1) $p(x) \geq 0, \int_{-\infty}^{\infty} p(x)dx = 1$

(2) $P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$ (단, $-\infty \leq a < b \leq \infty$)

를 만족시킬 때, $p(x)$를 X의 확률밀도함수라고 한다.

b. 연속형 확률변수의 성질

(1) P(c)=0 (c는 상수)

(2) $P(a \leq X \leq b)$=$P(a < X \leq b)$=$P(a \leq X < b)$=$P(a < X < b)$

(예) 바늘이 구간 [a,b] (a,b ∈[0,1]) 사이의 눈금에 멈출 확률

X=바늘이 저절로 멈추면서 가리키는 눈금

$P(a \leq X \leq b)$=(b-a)/(1-0)=b-a (단, $0 \leq a < b \leq 1$)

$$p(x) = \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{o.w} \end{cases}$$

3 확률변수와 확률분포

- 누적분포함수 (**c**umulative **d**istribution **f**unction, **cdf** 또는 분포함수 **df**)

$$F(x) = P(X \le x) = \begin{cases} \displaystyle\sum_{x_i \le x} p(x_i)\,, & X\text{: 이산형} \\ \displaystyle\int_{-\infty}^{x} p(t)\,dt, & X\text{: 연속형} \end{cases}$$

\- 누적분포함수는 확률밀도함수로부터 얻어진다. 그 역도 성립한다.

(예) 동전을 두번 던지는 실험에서 확률변수 X=앞면의 수라고 할 때, 누적분포함수를 구하여라.

4 기대값과 그 성질

4.1 확률변수의 기대값(expected value) 또는 평균(mean)

- $p(x)$ 를 X의 확률밀도함수라고 할 때, 확률변수 X의 기대값

$$E(X) = \begin{cases} \sum_{\text{모든 } x} xp(x) & X: \text{이산확률변수} \\ \int_{-\infty}^{\infty} xp(x)dx & X: \text{연속확률변수} \end{cases}$$

예) 동전을 2회 던질 때 앞면의 개수 X 의 확률밀도함수와 기대값

예) 한 사람이 오후 12시부터 1시 사이에 우연히 정거장에 오는 시간 X의 확률밀도함수와 기대값

4 기대값과 그 성질

- 함수 g (**R**→**R**)에 대하여 확률변수g(X)의 기대값

$$E(g(X)) = \begin{cases} \displaystyle\sum_{\text{모든 } x} g(x)p(x) & X\text{: 이산확률변수} \\ \displaystyle\int_{-\infty}^{\infty} g(x)p(x)dx & X\text{: 연속확률변수} \end{cases}$$

예) Y=(X-1)^{4}의 확률분포와 기대값 (X는 앞의 동전 2회 던지는 실험의 확률변수)

- **기대값의 성질**

(1) 임의의 상수 a, b에 대하여 $E(aX+b) = aE(X) + b$
(2) 함수 g_1, g_2와 임의의 상수 a, b에 대하여,

$$E(ag_1(X) + bg_2(X)) = aE(g_1(X)) + b\mathrm{E}(g_2(X))$$

(3) $X \geq 0$이면 $E(X) \geq 0$

4.2 분산 (variance)과 표준편차 (standard deviation)

X의 평균이 μ이고 X의 확률밀도함수가 $p(x)$일 때,

(1) $\mathrm{Var}(X) = E[(X-\mu)^2] = \begin{cases} \sum_{\text{모든 } x}(x-\mu)^2 p(x) & X\text{: 이산확률변수} \\ \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx & X\text{: 연속확률변수} \end{cases}$

(2) sd(X)=$\sqrt{\mathrm{Var}(X)}$
(3) 분산의 간편계산 : $\mathrm{Var}(X) = \mathrm{E}(X^2) - [E(X)]^2$
(4) 분산의 성질 : $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$ $(a, b$는 상수)

4.3 표준화(standardization)

$$Z = \frac{X - E(X)}{sd(X)}$$

E(Z)=0

Var(Z)=1

5.1 두 확률변수의 결합분포

: 두 개의 확률변수가 취할 수 있는 값들의 모든 순서 짝에 확률이 흩어진 정도를 합이 1인 양수로 나타낸 것

5.2 결합확률밀도함수 (joint pdf)

- 이산형 확률변수(X,Y) 의 확률밀도함수

$$p(x,y) = \begin{cases} P(X = x_i, Y = y_j) & (x,y) = (x_i, y_j), \quad i,j = 1,2,\dots \\ 0 & o.w \end{cases}$$

- 이산형 확률밀도함수의 성질

(1) $0 \le p(x,y) \le 1,\ \sum_{\text{모든}(x,y)} p(x,y) = 1$

(2) $P(a < X \le b, c < X \le d) = \sum_{a<x\le b} \sum_{c<y\le d} p(x,y)$

- 연속확률변수 (X,Y) 에 대하여 함수 $p(x,y)$가 다음을 만족할 때,

$(1) p(x,y) \geq 0, \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y)dxdy = 1$

$(2) P(a \leq X \leq b, c \leq X \leq d) = \int_{c}^{d} \int_{a}^{b} p(x,y)dxdy$ (단, $-\infty \leq a < b \leq \infty, -\infty \leq c < d \leq \infty$)

$p(x,y)$를 (X,Y) 의 확률밀도함수라고 한다.

5.3 주변확률밀도함수 (marginal pdf)

결합확률밀도함수 $p(x,y)$ 에서 유도된 각 확률변수 X,Y의 확률밀도함수

(1) 이산형 : $p_X(x) = \sum_{모든\ y} p(x,y)$, $p_Y(y) = \sum_{모든\ x} p(x,y)$

(2) 연속형 : $p_X(x) = \int_{-\infty}^{\infty} p(x,y)dy$, $p_Y(x) = \int_{-\infty}^{\infty} p(x,y)dx$

예) X,Y의 결합분포가 다음과 같을 때, $p_X(x)$, $p_Y(y)$를 구하여라.

x \ y	0	1	2	3	행의 합
0	0.05	0.05	0.10	0.00	0.20
1	0.05	0.10	0.25	0.10	0.50
2	0.00	0.15	0.10	0.05	0.30
열의 합	0.10	0.30	0.45	0.15	1.00

5.4 두 확률변수의 함수의 기대값

X,Y의 결합확률밀도함수 $p(x,y)$ 및 함수 g 에 대하여 확률변수 g(X,Y)의 기대값

(1) $E(g(X,Y)) = \begin{cases} \sum_{\text{모든 } x} \sum_{\text{모든 } y} g(x,y)p(x,y) & (X,Y)\text{: 이산확률변수} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)p(x,y)\,dx\,dy & (X,Y)\text{: 연속확률변수} \end{cases}$

(2) $E(ag_1(X,Y) + bg_2(X,Y)) = aE(g_1(X,Y)) + b\mathrm{E}(g_2(X,Y))$, a, b는 상수

5.5 두 확률변수의 독립성

X,Y의 결합확률밀도함수 $p(x,y)$와 각각의 주변밀도함수 $p_X(x)$, $p_Y(y)$에 대하여 $p(x,y)$=$p_X(x)p_Y(y)$ 가 모든 (x,y)에 대하여 성립할 때, X와 Y는 서로 독립이다.

예) X,Y의 결합분포가 다음과 같을 때, E(XY)를 구하시오. X와 Y는 서로 독립인가?

x \ y	0	1	행의 합
0	1/8	1/8	1/4
1	1/4	1/4	1/2
2	1/8	1/8	1/4
열의 합	1/2	1/2	1.00

6 공분산과 상관계수

6.1 공분산 (covariance)

$Cov(X,Y) = \mathrm{E}(\mathrm{X} - \mu_X)(Y - \mu_Y)$, μ_X=E(X), μ_Y=E(Y)

- 확률변수 X,Y의 선형관계의 유무 및 방향성을 나타내는 척도
- 각 확률변수가 취하는 값의 단위에 의존
- 간편계산법 $Cov(X,Y)$=E(XY)-E(X)E(Y)

6.2 상관계수 (correlation coefficient)

$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{sd(X)sd(Y)}$

- 확률변수 X,Y의 선형관계의 유무, 방향성 및 강도를 나타내는 척도
- 공분산의 단위에 대한 의존도를 없애준 것

6.3 성질

- $Cov(aX+b, cY+d) = acCov(X,Y)$
- $Corr(aX+b, cY+d) = \begin{cases} Corr(X,Y), & ac>0 \\ -Corr(X,Y), & ac<0 \end{cases}$
- $\mathrm{Var}(X \pm Y) = \mathrm{Var(X)} + \mathrm{Var(Y)} \pm 2\ Cov(X,Y)$
- $-1 \le \rho \le 1$
- $Y = a + bX$의 관계에 있으면 ρ=1(b>0) 또는 ρ=-1(b<0)

6.4 독립인 두 확률변수의 기대값과 성질

(1) E(XY)=E(X)E(Y)

(2) Cov(X,Y)=0, Corr(X,Y)=0

(3) Var(X±Y)=Var(X) + Var(Y)