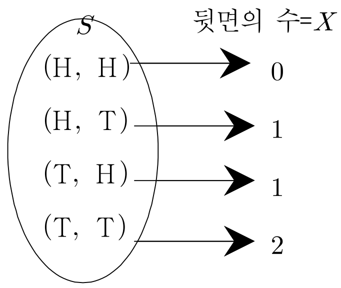
2. 이산확률분포

2-1. 확률변수

① 확률변수 : 표본공간의 각 표본값에 하나의 실숫값을 대응시켜 주는 함수

〚예〛 한 개의 동전을 두 번 던지는 시행



뒷면의 수(X)	0	1	2	합계
확률	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	1

② 확률변수의 종류

이산확률변수	연속확률변수
확률변수 X가 취할 수 있는 값이 유한개이거나 셀 수 있는 무한개(countably infinite)인 경우	확률변수 X가 셀 수 없는 무한개 (uncountably infinite)의 실숫값을 갖는 경우
〚예〛	〚예〛

③ 확률변수 X의 평균 $\mathrm{E}(X)$, 분산 $\mathrm{V}(X)$, 표준편차 $\sigma(X)$

> 확률변수 $Y=aX+b$ (a, b는 상수)의 평균, 분산, 표준편차는 각각
> ㉠ $\mathrm{E}(Y)=\mathrm{E}(aX+b)=a\mathrm{E}(X)+b$
> ㉡ $\mathrm{V}(Y)=\mathrm{V}(aX+b)=a^2\mathrm{V}(X)$
> ㉢ $\sigma(Y)=\sigma(aX+b)=|a|\sigma(X)$

〚예〛 확률변수 X에 대하여 $\mathrm{E}(X)=3$, $\mathrm{V}(X)=2$일 때, 확률변수 $Y=-2X+5$의 평균, 분산, 표준편차

㉠ $\mathrm{E}(-2X+5)=-2\mathrm{E}(X)+5=-2\times 3+5=-1$

㉡ $\mathrm{V}(-2X+5)=(-2)^2\mathrm{V}(X)=4\times 2=8$

㉢ $\sigma(-2X+5)=|-2|\sigma(X)=2\times\sqrt{2}=2\sqrt{2}$

④ 확률함수 : 확률변수에 대하여 정의된 실수를 구간 [0, 1] 사이의 확률에 대응시키는 함수

이산확률변수	연속확률변수
확률질량함수 $\mathrm{P}(X=x_i)$	확률밀도함수 $f(x)$
1. $0 \le \mathrm{P}(X=x_i) \le 1$ (단, $i=1,\ 2,\ \cdots\cdots,\ n$)	1. $f(x) \ge 0$
2. $\sum_{i=1}^{n}\mathrm{P}(X=x_i)=1$	2. $\int_{-\infty}^{\infty} f(x)dx=1$
3. $\mathrm{P}(x_i \le X \le x_j) = \sum_{k=i}^{j}\mathrm{P}(X=x_k)$	3. $\mathrm{P}(a \le X \le b) = \int_a^b f(x)\,dx$

⑤ 확률분포 ; 불확실한 현상을 관찰하는 경우 실현치와 확률간의 함수 관계로 표현되는 것. 모든 확률변수는 나름대로 고유한 확률분포를 가지며 통계 분석은 이러한 확률분포를 근거로 행해짐

이산확률분포	연속확률분포
이항분포, 포와송분포, 초기하분포, 기하분포, 음이항분포, 다항분포 등	정규분포, 표준정규분포, 지수분포, t분포, F분포, 카이제곱분포, 감마분포, 일양분포, 균등분포 등

㉠ 이산형 균등분포(discrete uniform distribution)

이산형 확률변수 X가 n개의 가능한 값 $x_1,\ x_2,\ x_3,\ \cdots,\ x_n$을 가지며 이들의 확률이 모두 동일할 때, X는 이산형 균등분포에 따른다고 한다.

$$\mathrm{P}(X=x_i)=\frac{1}{n}\ (i=1,2,3,\ \cdots,n)$$

㉡ 이항분포(binomial distribution)

표본공간이 오직 두 가지의 배타적인 원소로 구성된 실험의 시행을 베르누이 시행이라고 하는데, 이 베르누이 시행을 독립적으로 n번 시행한 결과로 얻는 성공 횟수 X에 대한 확률분포이다.

$$\mathrm{P}(X=x)={}_n\mathrm{C}_x p^x q^{n-x}\ \ (\text{단, } x=0,\ 1,\ 2,\ ....,\ p+q=1)$$

㉢ 음이항분포(negative binomial distribution)

성공률 p인 베르누이 시행을 r회의 성공이 나타날 때까지 시행하는 경우 r회 성공할 때까지의 시행하는 횟수 X에 대한 확률분포이다.

$$\mathrm{P}(X=x)={}_{x-1}C_{r-1}p^r q^{x-r}\ \ (\text{단, } x=r,\ r+1,\ldots,\ p+q=1)$$

㉣ 기하분포(geometric distribution)

음이항분포의 특별한 경우로, $r=1$인 경우를 기하분포라고 한다. 즉, 표본의 크기 n을 정하지 않고 표본을 계속하여 택하면서 요구하는 수 많큼의 성공이 일어나면 표본추출을 멈추는 경우로, 처음으로 성공이 일어날 때까지의 실패하는 횟수 X에 대한 확률분포이다.

$$\mathrm{P}(X=x)=pq^x\ \ (\text{단, } x=0,\ 1,\ 2,\ ....,\ p+q=1)$$

ⓜ 초기하분포(hypergeometric distribution)

N개의 공(M개는 정상품, $N-M$개는 불량품)이 들어 있는 상자에서 n개의 공을 비복원으로 임의 추출하는 실험에서 공을 하나 추출했을 때 정상품이면 성공, 불량품이면 실패라고 정한 실험에서 성공 횟수 X에 대한 확률분포이다.

$$\mathrm{P}(X=x)=\frac{{}_{M}\mathrm{C}_{x}\times{}_{N-M}\mathrm{C}_{n-x}}{{}_{N}\mathrm{C}_{n}}\ (\text{단, } x=0,\ 1,\ 2,\ \dots,\ n)$$

ⓑ 포아송분포(Poisson distribution)

단위 구간 또는 단위 시간 내에 어떤 사건이 평균 λ회 발생한다고 할 때, 그 구간(시간) 내에서 발생되는 사건의 횟수 X에 대한 확률분포이다.

$$\mathrm{P}(X=x)=\frac{e^{-\lambda}\lambda^{x}}{x!}\ (\text{단, } x=0,\ 1,\ 2,\ \dots)$$

ⓢ 균등분포(uniform distribution)

연속확률변수 X가 실수 구간 [a, b]에서 나타날 가능성이 균등할 때 X는 균등분포를 따른다고 한다.

$$f(x)=\frac{1}{b-a}\ (\text{단}, a\le x\le b)$$

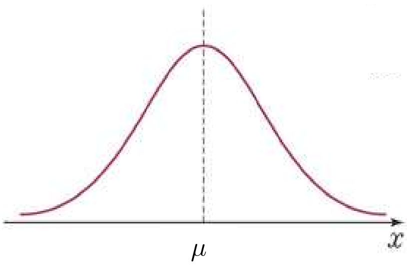
ⓞ 정규분포(normal distribution)

i) 정규분포의 확률밀도함수는 드무아브르(de Moivre, A.; 1667~1754)가 처음 발표함

ii) 비슷한 시기에 물리학 실험에 수반되는 측정오차의 분포에 정규분포를 적용한 독일의 수학자 가우스(Gauss, K.F.;1777~1855)의 이름을 따서 가우스 분포라고도 함

iii) 정규분포는 가능한 값이 구간 $(-\infty,\ \infty)$ 사이의 모든 실숫값이며, 분포의 형태가 중심이 가장 높게 솟아오른 좌우 대칭인 종 모양을 함.

iv) 평균치 μ, 분산 σ^2의 정규분포에 따르는 연속형 확률변수 X의 확률밀도함수



$$f(x)=\frac{1}{\sqrt{2\pi}\,\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ (-\infty<x<\infty)$$

ⓩ 표준정규분포(standard normal distribution)

평균치 $\mu=0$, 분산 $\sigma^2=1$인 정규분포를 표준정규분포라고 한다. 표준정규분포에 따르는 확률변수를 Z로 표시하며, 간략히 $Z\sim N(0,1)$로 표기한다.

$$f(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x)^2}{2}}\ (-\infty<x<\infty)$$

㉧ 지수분포(exponential distribution)
어떤 사건이 발생할 때까지 걸리는 시간 혹은 사건과 사건 사이의 경과시간을 설명해주는 확률분포가 지수분포이다. 어떤 정해진 시간 간격 동안 발생하는 사건의 개수가 평균치 λ의 포아송 분포에 따를 때, 첫 번째 사건이 발생할 때까지의 시간 혹은 사건과 사건 사이의 경과시간은 평균치 $\frac{1}{\lambda}$의 지수분포에 따른다.

$$f(x) = \lambda e^{-\lambda x} \ (\text{단, } x \geq 0)$$

㉨ 얼랑분포(erlang distribution)
지수분포를 일반화시킨 확률분포로, 정해진 시간간격 동안 발생되는 사건의 개수가 포아송분포에 따르는 경우 임의의 한 시점에서부터 r개의 사건이 발생될 때까지의 경과 시간에 대한 확률분포는 얼랑분포를 따른다. 즉, 단위시간당 발생되는 사건의 개수가 평균 λ의 포아송분포를 따를 때, r개의 사건이 발생될 때까지의 경과시간을 확률변수 X라 하면, X는 모수 λ, r을 갖는 얼랑확률변수이다.

$$f(x) = \frac{x^{r-1}\lambda^r e^{-\lambda x}}{(r-1)!} \ (x > 0, r = 1, 2, 3, \cdots)$$

㉩ 감마분포(gamma distribution)
모수 λ, r의 얼랑분포에서 r이 취할 수 있는 값은 양의 정수 값으로 제한되어 있다. 얼랑분포에서 r이 취할 수 있는 값이 양의 실숫값으로 확대되면 이를 감마분포라고 한다. 따라서 얼랑분포는 감마분포의 특수한 경우라고 볼 수 있다.

$$f(x) = \frac{x^{r-1}\lambda^r e^{-\lambda x}}{\Gamma(r)} \ (x > 0)$$

여기서 감마함수 $\Gamma(x) = \int_0^\infty x^{r-1}e^{-x}dx \ (r > 0)$와 같이 정의된다.

㉪ Weibull 분포(weibull distribution)
여러 가지 기계 및 전자 시스템의 수명의 확률분포로 가장 많이 이용되는 확률분포가 와이블 분포이다. 척도모수(scale parameter) $\delta > 0$와 형상모수(shape parameter) $\beta > 0$에 따른 확률밀도함수는 다음과 같다.

$$f(x) = \frac{\beta}{\delta}\left(\frac{x}{\delta}\right)^{\beta-1} e^{-\left(\frac{x}{\delta}\right)^{\beta}} \ (x > 0)$$

㉫ 대수정규분포(lognormal distribution)
확률분포 X의 자연대수 $\ln X$가 정규분포를 따를 때, X를 대수정규분포에 따른다고 한다. 즉, 두 확률변수 W와 X가 $X = e^W$ $(W = \ln X)$의 관계를 가질 때 W가 정규분포에 따른다면 X는 대수정규분포에 따르게 된다. W가 평균 θ, 분산 ω^2을 갖는 정규분포의 확률변수일 때, X의 확률밀도함수는 다음과 같다.

$$f(x) = \frac{1}{\sqrt{2\pi x\omega}} e^{-\frac{(\ln x - \theta)^2}{2\omega^2}} \ (0 < x < \infty)$$

대수정규분포는 신뢰성 공학에서 부품의 수명에 관한 확률분포, 혹은 재고관리에서 일정기간 동안의 제품의 수요에 대한 확률분포 등에 널이 이용된다.

2-2. 이산확률분포

① 이산확률변수 X의 각 값 x_1, x_2, ⋯⋯, x_n과 X가 그 값을 취할 확률 p_1, p_2, ⋯⋯, p_n 사이의 대응 관계

② 이산확률변수 X의 확률질량함수(p.m.f)

; 이산확률분포를 나타내는 함수 $\mathrm{P}(X=x_i)$ $(i=1,\ 2,\ \cdots,\ n)$

③ 확률질량함수 $\mathrm{P}(X=x_i)$의 성질

㉠ $0 \le \mathrm{P}(X=x_i) \le 1$(단, $i=1,\ 2,\ \cdots\cdots,\ n$)

㉡ $\sum_{i=1}^{n} \mathrm{P}(X=x_i)=1$

④ 이산확률변수 X의 기댓값(평균) $\mathrm{E}(X)$

$$\mathrm{E}(X)=x_1p_1+x_2p_2+\cdots\cdots+x_np_n=\sum_{i=1}^{n}x_ip_i$$

⑤ 이산확률변수 X의 평균 $\mathrm{E}(X)=m$일 때,

㉠ X의 분산 $\mathrm{V}(X)=\mathrm{E}\{(X-m)^2\}=\sum_{i=1}^{n}(x_i-m)^2p_i=\sum_{i=1}^{n}x_i^2p_i-m^2$

$=\mathrm{E}(X^2)-\{\mathrm{E}(X)\}^2$

㉡ X의 표준편차 $\sigma(X)=\sqrt{\mathrm{V}(X)}$

〚예〛 한 개의 동전을 세 번 던지는 시행, 앞면의 개수를 확률변수 X라고 할 때의 확률분포

X	0	1	2	3	합계
$\mathrm{P}(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	1

[풀이]

㉠ $\mathrm{P}(0 \le X \le 1)=\mathrm{P}(X=0)+\mathrm{P}(X=1)=\frac{1}{8}+\frac{3}{8}=\frac{1}{2}$

㉡ $\mathrm{E}(X)=0\left(\frac{1}{8}\right)+1\left(\frac{3}{8}\right)+2\left(\frac{3}{8}\right)+3\left(\frac{1}{8}\right)=1.5$

㉢ $\mathrm{V}(X)=0\left(\frac{1}{8}\right)+1\left(\frac{3}{8}\right)+4\left(\frac{3}{8}\right)+9\left(\frac{1}{8}\right)-\frac{9}{4}=\frac{3}{4}$

㉣ $\sigma(X)=\sqrt{\mathrm{V}(X)}=\sqrt{\frac{3}{4}}$

㉤ 확률밀도함수 $p(x)={}_3C_x\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{3-x}$ (이항분포)

[풀이추가] 이항분포 $B\left(3,\frac{1}{2}\right)$이므로 평균$=3\times\frac{1}{2}=\frac{3}{2}$, 분산$=3\times\frac{1}{2}\times\frac{1}{2}=\frac{3}{4}$이다.

〖예〗 불량률이 0.01인 A 제품을 하나씩 차례로 임의 추출하여 품질검사 하는 시행. 확률변수 X를 불량품(F)이 발견되기까지 필요한 검사개수라 할 때,

㉠ 시행의 표본공간 $S=\{F,\ SF,\ SSF,\ SSSF,\ \ldots\}$

㉡ $p(x)=(0.99)^{x-1}(0.01),\ x=1,\ 2,\ldots$

㉢ $P(X\le 3)=P(X=1)+P(X=2)+P(X=3)$

$$=0.01+0.0099+0.0098=0.0297$$

㉣ $P(X=125)=(0.99)^{124}(0.01)=0.0029$

㉤ 이 시행은 기하분포(Geometric Distribution)를 따르며, 첫 번째 불량품이 발견되기까지 평균적으로 $\dfrac{1}{p}=\dfrac{1}{0.01}=100$ 즉, 100개마다 하나씩 불량품이 발견된다.

2-3. 이항분포

① 1회의 시행에서 A가 일어날 확률 p, 일어나지 않을 확률 $q(q=1-p)$, n번의 독립시행에서 사건 A가 일어나는 횟수를 X

② 확률변수 X의 확률질량함수

$$\mathrm{P}(X=k)={}_n\mathrm{C}_k p^k q^{n-k}\ (\text{단, } k=0,\ 1,\ 2,\ \cdots\cdots,\ n,\ p+q=1)$$

X	0	1	2	$\cdots\cdots$	n	합계
$\mathrm{P}(X=x)$	${}_n\mathrm{C}_0 q^n$	${}_n\mathrm{C}_1 p^1 q^{n-1}$	${}_n\mathrm{C}_2 p^2 q^{n-2}$	$\cdots\cdots$	${}_n\mathrm{C}_n p^n$	1

: 이항분포, $\mathrm{B}(n,\ p)$

③ 확률변수 X가 이항분포 $\mathrm{B}(n,\ p)$를 따를 때,

㉠ 평균 $\mathrm{E}(X)=np$, $\mathrm{V}(X)=npq$

㉡ 표준편차 $\sigma(X)=\sqrt{npq}$ (단, $q=1-p$)

〖예〗 어떤 병을 앓고 있는 6명의 환자에게 완치율이 70 %인 신약을 투여할 때, 이 중에서 5명 이상이 치유될 확률?

[풀이]

㉠ 치유되는 환자의 수 X, 이항분포 $\mathrm{B}\left(6,\ \dfrac{7}{10}\right)$

㉡ 5명 이상이 치유될 확률 $\mathrm{P}(X\ge 5)$

$$\mathrm{P}(X\ge 5)=\mathrm{P}(X=5)+\mathrm{P}(X=6)$$

$$={}_6\mathrm{C}_5\left(\frac{7}{10}\right)^5\left(\frac{3}{10}\right)^1+{}_6\mathrm{C}_6\left(\frac{7}{10}\right)^6\left(\frac{3}{10}\right)^0$$

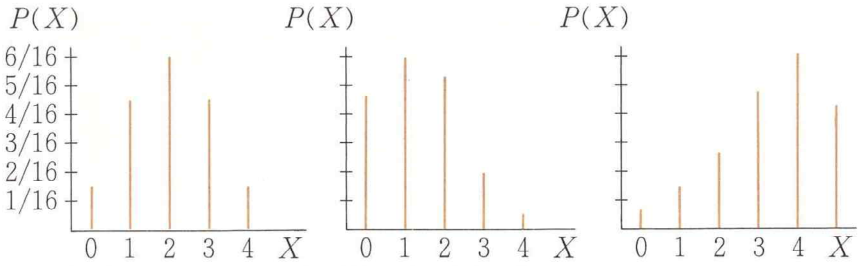
$$=0.420175$$

㉢ $\mathrm{E}(X)=6\times\dfrac{7}{10}=4.2$

㉣ $\mathrm{V}(X)=6\times\dfrac{7}{10}\times\dfrac{3}{10}=1.26$ ㉤ $\sigma(X)=\sqrt{1.26}$

④ 이항분포 ~ 정규분포

㉠ 이항분포 : 시행횟수 n과 성공 확률 p에 영향을 받는다.



(a) $p=1/2$인 이항분포 (b) $p=1/3$인 이항분포 (c) $p=2/3$인 이항분포

㉡ 이항분포는 시행횟수 n에 영향을 받는다.

; $np > 5$ 그리고 $nq > 5$의 조건을 만족할 때