2장 이산형분포

1. 이산형 확률변수
2. 수학적 기대값
3. 평균과 분산
4. 적률생성함수
5. 베르누이 시행과 이항분포
6. 초기하 분포
7. 기하 분포
8. 음이항 분포
9. 포아송 분포

1. 이산형 확률변수

• 이산형(discrete): 대소비교의 의미가 있는 셀 수 있는 정수 자료형
(예) 자녀수, 사고 횟수, 제품의 개수 등.

• 확률변수(random variable): 확률실험을 통해 얻어지는 기본결과 각각에 수치를 대응시킨 것

• 이산형 확률변수(discrete random variable)와 이산형 확률분포
이산형 변수 X의 모든 가능한 실현치 $x_1, x_2, \cdots$에 대해 확률질량(즉, 확률)

$$f(x_1) = P(X = x_1), \quad f(x_2) = P(X = x_2), \quad \cdots$$

가 대응될 때, X를 이산형 확률변수라 하고, $f(x_1), f(x_2), \cdots$를 이산형 확률분포라 한다.

함수 $f(x)$를 확률질량함수(probability mass function: pmf)라 부른다.

(1) $f(x) > 0, \quad x \in R$

(2) $\sum_{x \in R} f(x) = 1$

(3) $P(X \in A) = \sum_{x \in A} f(x), \quad A \subset R.$

$x \notin R$일 때는 $f(x) = 0$으로 정의하며, $f(x)$의 정의구역은 실수전체의 집합이 된다.

예 2.1-1 어제 낚시를 하여 물고기 10마리를 낚았는데 그 중에서 4마리는 30cm 이하였다. 집에 돌아와서 2마리를 비복원 추출의 방법으로 꺼낼 때, X를 30cm이하의 물고기 수라고 하면 X의 pdf는

$$f(x) = \frac{\binom{4}{x}\binom{6}{2-x}}{\binom{10}{2}}, \quad x = 0, 1, 2$$

이다. 이 때 확률변수 X는 **초기하 분포** (Hypergeometric distribution)를 갖는다고 한다.

예 2.1-2 이산확률변수 X의 밀도함수가 $f(x)=\frac{x}{6}$, $x=1, 2, 3$이라고 하자. 그러면 다음의 결과를 쉽게 얻을 수 있다.

$$P(X \le 1) = f(1) = \frac{1}{6},$$

$$P(X \le 3) = f(1) + f(2) + f(3) = \frac{1}{6} + \frac{2}{6} + \frac{3}{6} = 1,$$

$$P(X \le 0) = 0,\ P\left(X \le \frac{3}{2}\right) = \frac{1}{6},\ P\left(X \le \frac{7}{3}\right) = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}.$$

집합 A를 $A=\{t: t \le x,\ x \in R\}$라고 하자. $F(x)=P(X \in x)=\sum_{t \in A} f(t)$라고 정의할 때, $F(x)$를 확률변수 X의 **누적분포함수** (Cumulative distribution function) 또는 간단히 분포함수라 하고 간략하게 cdf 라 표기한다. 또한, $F(\infty)=\lim_{x \to \infty} F(x)$, $F(-\infty)=\lim_{x \to -\infty} F(x)$라 약속한다. 다음은 분포함수의 중요한 성질이다.

(1) $F(x)$는 확률이므로 $0 \le F(x) \le 1$이다.

(2) $F(x)$는 x의 비감소 함수 (Nondecreasing function)이다. 왜냐하면, $x_1 < x_2$일 때, $\{x: x \le x_1\} \subset \{x: x \le x_2\}$이므로 $F(x_1)=P(X \le x_1) \le P(X \le x_2)=F(x_2)$ 이기 때문이다.

(3) $F(\infty)=1$, $F(-\infty)=0$이다.

(4) F는 오른편 연속함수 (Right-continuous function)이다.

(5) X가 이산확률변수일 때, $F(x)$는 계단함수 (Step function)가 된다.

2. 수학적 기댓값

[예 2.2-1] X: 주사위를 던졌을 때 뒷면의 눈금을 나타내는 확률변수

X의 확률밀도함수: $f(x)=\frac{1}{6},\ x=1,2,3,4,5,6$

x의 값에 대하여 의 상금이 주어진다.(단위 만원)

$$u(x)=\begin{cases}1, & x=1,2,3,\\ 5, & x=4,5,\\ 35, & x=6.\end{cases}$$

이때 지불되는 상금의 수학적 기대값은 다음과 같다.

$$\sum_{x=1}^{6}u(x)f(x)=(1)\left(\frac{1}{6}\right)+(1)\left(\frac{1}{6}\right)+(1)\left(\frac{1}{6}\right)+(5)\left(\frac{1}{6}\right)+(5)\left(\frac{1}{6}\right)+(35)\left(\frac{1}{6}\right)$$
$$=8(\text{만원}).$$

[정의 2.2-1]

$f(x)$가 이산확률변수 X의 pdf 일 때,

$$E[u(X)]=\sum_{x\in R}u(x)f(x)$$

의 값이 존재할 때, 이 값을 $u(X)$의 **수학적 기대값** (Mathematical expectation)이라고 한다.

정리 2.2-1 (수학적 기대값의 성질)

수학적 기대값이 존재하면 다음의 성질이 성립한다.

(1) c가 상수이면 $E(c)=c$이다.

(2) $E(cX)=cE(X)$.

(3) $E(cu(X))=cE(u(X))$.

(4) $E[c_1u_1(X)\pm c_2u_2(X)]=c_1E[u_1(X)]\pm c_2E[u_2(X)]$.

(5) $u_1(x)\le u_2(x)$이면 $E[u_1(X)]\le E[u_2(X)]$이다.

[예 2.2-2]

(세인트 피터스버어그의 역설)

뒷면이 나타날 때까지 동전을 던지는 실험을 한다. 처음 시행에 뒷면이 나오면 2원, 두 번째 시행에서 뒷면이 나오면 2^2원, … , x 번째 시행에서 뒷면이 나오면 2^x원을 받는 게임을 할 때, 처음에 얼마를 거는 것이 합당한가?

풀이>

이 경우의 확률밀도함수는 $f(x)=\frac{1}{2^x},\ x=1,2,\cdots$ 이다. 합당한 게임이 되려면 상금의 기대값과 처음에 거는 액수가 같아야 하므로

$$E[2^X]=\sum_{x=1}^{\infty}2^x\cdot\frac{1}{2^x}=1+1+\cdots=\infty$$

가 되는데 이 값은 존재하지 않는다.

[예 2.2-3]

X의 pdf가 다음과 같이 주어진다고 하자.

$$f(x)=\frac{x}{10},\ x=1,2,3,4.$$

그러면 다음의 값을 쉽게 계산할 수 있다.

$$E(X)=\sum_{x=1}^{4}x\left(\frac{x}{10}\right)=(1)\left(\frac{1}{10}\right)+(2)\left(\frac{2}{10}\right)+(3)\left(\frac{3}{10}\right)+(4)\left(\frac{4}{10}\right)=3,$$

$$E(X^2)=\sum_{x=1}^{4}x^2\left(\frac{x}{10}\right)=(1)^2\left(\frac{1}{10}\right)+(2)^2\left(\frac{2}{10}\right)+(3)^2\left(\frac{3}{10}\right)+(4)^2\left(\frac{4}{10}\right)=10,$$

$$E[X(5-X)]=5E(X)-E(X^2)=(5)(3)-10=5.$$

3. 평균과 분산

[정의 2.3-1]

확률변수 X가 이산확률변수 $f(x)$를 갖고 그 공간이 $R=\{b_1, b_2, b_3, \cdots\}$일 때,

$$E(X)=\sum_R xf(x)=b_1f(b_1)+b_2f(b_2)+b_3f(b_3)+\cdots$$

를 X의 **평균**(Mean)이라고 한다.

확률변수 X의 평균 $E(X)$를 μ로 표기하기도 한다.

[예 2.3-1] 확률변수 X의 pdf

$$f(x)=\begin{cases}\frac{1}{8}, & x=0,\ 3,\\ \frac{3}{8}, & x=1,\ 2.\end{cases}$$

이 때 X의 평균은 $\mu=E(X)=0\left(\frac{1}{8}\right)+1\left(\frac{3}{8}\right)+2\left(\frac{3}{8}\right)+3\left(\frac{1}{8}\right)=\frac{3}{2}$

[예2.3-2] 확률변수 X: 주사위를 던져서 나오는 눈금

확률밀도함수: $f(x)=1/6,\ x=1, 2, 3, 4, 5, 6$

평균: $E(X)=\sum_{x=1}^{6}x\left(\frac{1}{6}\right)=\frac{1+2+3+4+5+6}{6}=\frac{7}{2}$

[정의 2.3-2]

이산확률변수 X의 **분산**(Variance)을 다음과 같이 정의한다.

$$\sigma^2=E[(X-\mu)^2]=\sum_R(x-\mu)^2f(x).$$

분산은 다음과 같은 방법으로도 계산할 수 있다.

$$\sigma^2=E[(X-\mu)^2]=E(X^2-2\mu X+\mu^2)=E(X^2)-2\mu E(X)+\mu^2=E(X^2)-\mu^2.$$

[정의 2.3-3]

분산의 양의 제곱근을 확률변수 X의 **표준편차**(Standard deviation)라고 한다.

[예2.3-3] 확률변수 X 분산($Var(X)$)

$$\sigma^2=E[(X-3.5)^2]=\sum_{x=1}^{6}(x-3.5)^2\cdot\frac{1}{6}$$

$$=[(1-3.5)^2+(2-3.5)^2+\cdots+(6-3.5)^2]\left(\frac{1}{6}\right)=\frac{35}{12}$$

표준편차 $\sigma=\sqrt{35/12}=1.708$

$E(X^2)=\sum_{x=1}^{6}x^2\left(\frac{1}{6}\right)=\frac{1^2+2^2+3^2+4^2+5^2+6^2}{6}=\frac{91}{6}$ 이므로

$\sigma^2=Var(X)=\frac{91}{6}-\left(\frac{7}{2}\right)^2=\frac{35}{12}$ 이다.

[예2.3-4] 확률변수 X, 확률밀도함수 $f(x)=1/3,\ x=-1, 0, 1$

평균 $\mu_X = E(X) = \sum_{x=-1}^{1} xf(x) = (-1)\left(\frac{1}{3}\right)+(0)\left(\frac{1}{3}\right)+(1)\left(\frac{1}{3}\right)=0$

분산 $\sigma_X^2 = E(X)^2-\mu_X^2 = \sum_{x=-1}^{1} x^2 f(x)-0^2 = (-1)^2\left(\frac{1}{3}\right)+(0)^2\left(\frac{1}{3}\right)+(1)^2\left(\frac{1}{3}\right)=\frac{2}{3}$

표준편차 $\sigma_X = \sqrt{2/3}$

확률변수 Y, 확률밀도함수 $g(y)=1/3,\ y=-2, 0, 2$

평균 $\mu_Y = E(Y) = \sum_{y=-1}^{1} yf(y) = (-2)\left(\frac{1}{3}\right)+(0)\left(\frac{1}{3}\right)+(2)\left(\frac{1}{3}\right)=0$

분산 $\sigma_Y^2 = E(Y)^2-\mu_Y^2 = \sum_{y=-1}^{1} y^2 f(y)-0^2 = (-2)^2\left(\frac{1}{3}\right)+(0)^2\left(\frac{1}{3}\right)+(2)^2\left(\frac{1}{3}\right)=\frac{8}{3}$

표준편차 $\sigma_X = 2\sqrt{2/3}$

--> 확률변수 Y의 분산은 확률변수 X의 분산의 4배이며 표준편차는 2배임을 알 수 있다. (Y가 X보다 더 많이 퍼져 있는 확률변수)

[정리 2.3-1] (체비세프의 부등식)

확률변수 X가 유한인 평균 μ와 분산 σ^2을 갖는다면 $k\geq 1$인 모든 k에 대하여

$$P(|X-\mu|\geq k\sigma)\leq \frac{1}{k^2}.$$

증명 확률변수 X의 pdf를 $f(x)$라 하고 $A=\{x : |x-\mu|\geq k\sigma\}$로 두면

$$\sigma^2 = E[(X-\mu)^2] = \sum_{x\in R}(x-\mu)^2 f(x) = \sum_{x\in A}(x-\mu)^2 f(x)+\sum_{x\in A^c}(x-\mu)^2 f(x)$$

이며 마지막 등식의 두 번째 항은 0보다 크거나 같고, 집합 A에서 $|x-\mu|\geq k\sigma$이므로

$$\sigma^2 \geq \sum_{x\in A}(x-\mu)^2 f(x) \geq \sum_{x\in A}(k\sigma)^2 f(x) = k^2\sigma^2\sum_{x\in A} f(x) = k^2\sigma^2 P(X\in A)$$
$$= k^2\sigma^2 P(|X-\mu|\geq k\sigma)$$

이다. 즉, $P(|X-\mu|\geq k\sigma)\leq \frac{1}{k^2}$ 이다.

[참고]

① 체비세프의 부등식에서 $\varepsilon = k\sigma$로 두면 $P(|X-\mu|\geq \varepsilon)\leq \frac{\sigma^2}{\varepsilon^2}$

② 체비세프의 부등식은 $P(|X-\mu|<k\sigma)\geq 1-\frac{1}{k^2}$으로도 표현

예 2.3-5 확률변수 X가 평균 25, 분산 16 (즉, $\sigma=4$)를 갖는다면

$$P(17<X<33)=P(|X-25|<8)=P(|X-\mu|<2\sigma)\geq 1-\frac{1}{4}=0.75$$

이다. 이는 X의 값이 구간 (17, 33)에 포함될 확률이 적어도 0.75 이상이라는 뜻이다. 나중에 다루겠지만 X가 정규분포를 갖는다면 이 확률은 0.95가 된다.

예 2.3-6 확률변수 Y의 분산이 $\frac{d}{n}$라고 하자. Y의 분산은 n의 값이 증가할수록 감소하게 된다. 체비세프의 부등식을 적용하면

$$P(|Y-\mu|\geq\varepsilon)\leq\frac{d}{n\varepsilon^2}$$

이므로 고정된 $\varepsilon>0$에 대하여 n이 충분히 크면 $\frac{d}{n\varepsilon^2}$는 0에 가깝게 된다. 따라서 Y의 값들은 대부분 구간 $(\mu-\varepsilon,\ \mu+\varepsilon)$에 포함된다.

4. **적률생성함수;** 확률변수 X의 적률(Moment)은 X의 거듭제곱으로 정의

[정의 2.4-1]

확률변수는 pdf가 $f(x)$인 이산확률변수라고 하자.

(a) $E(X^r)=\sum_R x^r f(x)$를 원점에 대한 X의 r차 적률이라고 한다.

(b) $E[(X-b)^r]=\sum_R (x-b)^r f(x)$를 중심이 b인 X의 r차 적률이라고 한다.

(c) $M(t)=E[e^{tX}]=\sum_R e^{tx}f(x)$, $-h<t<h$가 존재할 때, $M(t)$를 X의 적률생성함수 (Moment generating function)라고 한다.

예 2.4-1 확률변수 X가 주사위의 표면에 나타나는 눈금의 수를 나타낸다고 하면 X의 pdf는 $f(x)=\frac{1}{6}$, $x=1, 2, \cdots, 6$이므로 X의 적률생성함수는

$$M(t)=\sum_x e^{tx}f(x)=f(1)e^{1t}+f(2)e^{2t}+\cdots+f(6)e^{6t}=\frac{1}{6}e^t+\cdots+\frac{1}{6}e^{6t}$$

예 2.4-2 확률변수 X의 적률생성함수가 $M(t)=\frac{1}{15}e^{t}+\frac{2}{15}e^{2t}+\frac{3}{15}e^{3t}+\frac{4}{15}e^{4t}+\frac{5}{15}e^{5t}$로 주어졌다면, X의 pdf는 $f(x)=\frac{x}{15}$, $x=1, 2, 3, 4, 5$로 표현할 수 있다. 이제 새로운 확률변수 Y를 $Y=X^2$으로 정의하면 Y의 적률생성함수는

$$M_Y(t) = E[e^{tY}] = E[e^{tX^2}] = \sum_x e^{tx^2} f(x)$$
$$= \frac{1}{15}e^{t} + \frac{2}{15}e^{4t} + \frac{3}{15}e^{9t} + \frac{4}{15}e^{16t} + \frac{5}{15}e^{25t}$$

이며 Y의 확률분포는 다음과 같이 주어진다.

Y	1	4	9	16	25
$f_Y(y)$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{3}{15}$	$\frac{4}{15}$	$\frac{5}{15}$

이제 적률생성함수에 대하여 좀더 알아보도록 한다. 적률생성함수 $M(t)$를 t에 대해 미분을 하면 다음의 결과를 얻는다. 이 결과를 이용하여 평균과 분산을 쉽게 얻을 수 있으므로 꼭 알아두도록 한다.

$$\frac{d}{dt}M(t) = \frac{d}{dt}\sum_x e^{tx} f(x) = \sum_x \frac{d}{dt} e^{tx} f(x) = \sum_x x e^{tx} f(x)$$
$$\vdots$$
$$M^{(r)}(t) = \sum_x x^r e^{tx} f(x)$$

이므로 $M^{(r)}(0) = E(X^r)$ 임을 알 수 있고, $M'(0)=E(X)$, $M''(0)=E(X^2)$이므로 분산은 $\sigma^2 = M''(0)-[M'(0)]^2$으로 쉽게 구한다.

[참고] 맥로린의 급수 (Maclaurin's series) 이용

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots \text{ 이므로}$$

$$M(t) = E[e^{tX}] = E\left[1 + Xt + X^2\frac{t^2}{2!} + X^3\frac{t^3}{3!} + \cdots\right]$$
$$= E(1) + E(X)t + E(X^2)\frac{t^2}{2!} + E(X^3)\frac{t^3}{3!} + \cdots$$
$$= \sum_{r=0}^{\infty} E(X^r)\left(\frac{t^r}{r!}\right)$$

예 2.4-3 확률변수 Y의 적률생성함수가 $M(t)=(1-t)^{-1}$, $t<1$이라고 하면 맥로린의 급수 전개를 이용하여 다음의 결과를 얻는다.

$$M(t) = 1+t+t^2+t^3+\cdots+t^r+\cdots$$

$$= 1+(1!)\left(\frac{t}{1!}\right)+(2!)\left(\frac{t^2}{2!}\right)+(3!)\left(\frac{t^3}{3!}\right)+\cdots+(r!)\left(\frac{t^r}{r!}\right)+\cdots$$

따라서, $E(X^r)=r!$, $r=0,1,2,3,\cdots$이며 $\mu=E(X)=1!=1$,
$\sigma^2=E(X^2)-[E(X)]^2=2!-1^2=1$임을 알 수 있다.

예 2.4-4 확률변수 X의 적률이 $E(X^r)=0.8$, $r=1,2,3,\cdots$으로 주어졌다고 하면 X의 적률생성함수는

$$M(t) = M(0)+\sum_{r=1}^{\infty}0.8\left(\frac{t^r}{r!}\right)=1+0.8\sum_{r=1}^{\infty}\frac{t^r}{r!}=0.2+0.8\sum_{r=0}^{\infty}\frac{t^r}{r!}$$

$$= 0.2e^{0t}+0.8e^{1t}$$

이므로 $P(X=0)=0.2$, $P(X=1)=0.8$임을 알 수 있다.

※ $M(0)=E(e^{0\cdot x})=E(1)=1$

예 2.4-5 확률변수 X의 pdf가 $f(x)=2\left(\frac{1}{3}\right)^x$, $x=1,2,3,\cdots$이라고 하면 X의 적률생성함수는 $M(t)=\sum_{x=1}^{\infty}e^{tx}2\left(\frac{1}{3}\right)^x=\sum_{x=0}^{\infty}2\left(\frac{e^t}{3}\right)^x=\frac{2(e^t/3)}{1-e^t/3}$

$=\frac{2e^t}{3-e^t}$이며, 급수의 합이 존재하기 위해서는 $\frac{e^t}{3}<1$ 즉, $t<\ln 3$

이어야 한다. X의 적률생성함수를 t에 관하여 미분하면

$$M'(t) = \frac{(3-e^t)2e^t-2e^t(-e^t)}{(3-e^t)^2}=\frac{6e^t}{(3-e^t)^2}$$

이므로 $\mu=M'(0)=\frac{3}{2}$을 얻는다.

[참고] (누가적률 (Cumulants))

확률변수 X의 적률생성함수 $M(t)$가 $-h<t<h$에서 존재한다고 하자. X의 누가적률 (Cumulants)을 $R(t)=\ln M(t)$로 정의한다. $R(t)$를 이용하여 X의 평균과 분산을 구해본다.

$$R'(t)=\frac{M'(t)}{M(t)},\quad R''(t)=\frac{M(t)M''(t)-[M'(t)]^2}{[M(t)]^2}$$

이고, $M(0)=1$이므로 $R'(0)=M'(0)=\mu$, $R''(0)=M''(0)-[M'(0)]^2=\sigma^2$이 된다.

예 2.4-6 ([예 2.4-5]의 계속) [예 2.4-5]에서 확률변수 X의 적률생성함수는 $M(t)=2e^t/(3-e^t)$, $t<\ln 3$ 임을 알았다. $R(t)=\ln M(t)=\ln 2+t-\ln(3-e^t)$이므로

$$R'(t)=1+\frac{e^t}{3-e^t},\quad R''(t)=\frac{(3-e^t)e^t-e^t(-e^t)}{(3-e^t)^2}$$

이고 $\mu=R'(0)=\frac{3}{2}$, $\sigma^2=R''(0)=\frac{3}{4}$이다.

정리 2.4-1 서로 독립인 확률변수 $X_1, X_2, \cdots, X_n$의 적률생성함수가 각각 $M_{X_1}(t), \cdots, M_{X_n}(t)$라고 할 때, $Y=a_0+a_1X_1+\cdots+a_nX_n$ ($a_0, a_1, \cdots, a_n$은 상수)의 적률생성함수는 $M_Y(t)=e^{a_0t}M_{X_1}(a_1t)\cdot\cdots\cdot M_{X_n}(a_nt)$이다.

따름정리 확률변수 X의 적률생성함수를 $M(t)$라 할 때, $Y=a+bX$ (a, b는 상수)의 적률생성함수는 $M_Y(t)=e^{at}M(bt)$이다.

5. 베르누이시행과 이항분포

• 베르누이 시행에서는 두 가지의 서로 배반이고, 완전한 분할이 되는 확률실험(예를 들면, 성공과 실패, 합격품과 불량품, 동전의 앞면과 뒷면 등)의 결과를 얻게 된다. 베르누이 시행에서 성공할 확률을 p라고 하면 실패할 확률은 q=1-p가 된다.

[정의 2.5-1]

확률변수 X의 pdf가 $f(x)=p^x(1-p)^{1-x},\ x=0,1,\ 0\le p\le 1,\ q=1-p$ 로 주어질 때, X는 **베르누이 분포** (Bernoulli distribution)를 갖는다고 한다.

• 베르누이 분포의 평균과 분산

$$\mu=E(X)=\sum_{x=0}^{1}x\cdot p^x(1-p)^{1-x}=0\cdot(1-p)+1\cdot p=p,$$

$$\sigma^2=Var(X)=\sum_{x=0}^{1}(x-p)^2\cdot p^x(1-p)^{1-x}=p^2\cdot(1-p)+(1-p)^2\cdot p=p(1-p).$$

• 베르누이 분포의 적률생성함수 $M(t)=E[e^{tX}]=\sum_{x=0}^{1}e^{tX}\cdot p^x(1-p)^{1-x}=(1-p)+pe^t$

$\mu=M'(0)=E(X)=p,\ \sigma^2=M''(0)-[M'(0)]^2=p(1-p)$

[정의 2.5-2]

확률변수 X의 밀도함수가

$$f(x)=\binom{n}{x}p^x(1-p)^{n-x},\quad x=0,1,2,\cdots,n,\quad q=1-p$$

로 주어질 때, X는 모수가 $(n,\ p)$인 **이항분포** (Binomial distribution)를 갖는다고 한다.

• 모수가 (n,p)인 이항분포를 $B(n,p)$로 표기, 베르누이분포: $B(1,p)$

[예 2.5-1] 주사위를 다섯 번 독립적으로 던질 때, 6이 두 번, 6이 아닌 것이 세 번 나올 확률은

$$f(2)=\binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3$$

확률변수 X: $n=5$회 던져서 6이 나오는 회수, $X\sim B(5,1/6)$

[예 2.5-2] 동전을 독립적으로 10회 던질 때, 앞면이 꼭 6회 나타날 확률

$$\binom{10}{6}\left(\frac{1}{2}\right)^6\left(\frac{1}{2}\right)^4 = P(X \le 6) - P(X \le 5) = 0.8281 - 0.6230 = 0.2051$$

적어도 앞면이 6회 나타날 확률

$$\sum_{x=6}^{10}\binom{10}{x}\left(\frac{1}{2}\right)^x\left(\frac{1}{2}\right)^{10-x} = 1 - P(X \le 5) = 1 - 0.6230 = 0.3770.$$

[참고] (이항분포의 성질)

①이항정리 $(a+b)^n = \sum_{x=0}^{n}\binom{n}{x}b^x a^{n-x}$

$\sum_{x=0}^{n}\binom{n}{x}p^x(1-p)^{n-x} = [(1-p)+p]^n = 1$ --> 확률밀도함수의 요건 충족

②이항분포에서의 적률생성함수

$$M(t) = E[e^{tX}] = \sum_{x=0}^{n} e^{tx}\binom{n}{x}p^x(1-p)^{n-x}$$

$$= \sum_{x=0}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x} = [(1-p)+pe^t]^n$$

$M(t)$의 t에 대한 1차, 2차 도함수를 계산하면

$M'(t) = n[(1-p)+pe^t]^{n-1}pe^t$

$M''(t) = n(n-1)[(1-p)+pe^t]^{n-2}(pe^t)^2 + n[(1-p)+pe^t]^{n-1}(pe^t)$

이항분포 확률변수 X의 평균과 분산

$\mu = M'(0) = np$

$\sigma^2 = M''(0) - \mu^2 = n(n-1)p^2 + np - (np)^2 = np(1-p).$

③확률변수 $X \sim B(1, p)$, 확률변수 $Y \sim B(1, p)$, X와 Y가 서로 독립

각각의 적률생성함수 $M_X(t) = M_Y(t) = pe^t + (1-p)$

$X+Y$의 적률생성함수 $M_{X+Y}(t) = M_X(t)\,M_Y(t) = [pe^t + (1-p)]^2$; $X+Y \sim B(2, p)$

④확률변수 $X \sim B(1, p_1)$, 확률변수 $Y \sim B(1, p_2)$, X와 Y가 서로 독립

$X+Y$의 적률생성함수

$$M_{X+Y}(t) = [p_1e^t + (1-p)][p_2e^t + (1-p_2)]$$

$$= p_1p_2e^{2t} + [p_1(1-p_2) + p_2(1-p_1)]e^t + (1-p_1)(1-p_2)e^{0t}$$

$Z = X+Y$의 pdf는 다음과 같이 주어진다.

$$f_Z(z) = \begin{cases} (1-p_1)(1-p_2), & z = 0 \\ p_1(1-p_2) + p_2(1-p_1), & z = 1 \\ p_1p_2, & z = 2. \end{cases}$$

[예 2.5-3] 어떤 제품의 생산과정에서 평균적으로 10개 중에 한 개가 불량품이라고 한다. 독립적으로 제품을 5개 선택하여 검사를 한다. 확률변수 X를 $n=5$개 가운데 불량품의 수, $X \sim B(5, 0.1)$

X의 평균과 분산

$$E(X) = 5 \cdot (0.1) = 0.5, \quad Var(X) = 5 \cdot (0.1) \cdot (0.9) = 0.45.$$

불량품이 많아야 한 개 있을 확률

$$P(X \le 1) = \binom{5}{0}(0.1)^0(0.9)^5 + \binom{5}{1}(0.1)^1(0.9)^4 = 0.9185$$

[참고] (약한 대수의 법칙)

X가 이항분포 $B(n,p)$를 갖는다면 X/n은 성공에 대한 상대도수이며, p를 모르는 경우에는 p의 추정량으로 사용할 수 있다. (p의 근사값: X/n)

$\varepsilon > 0$일 때,

$$P\left(\left|\frac{X}{n} - p\right| \ge \varepsilon\right) = P(|X - np| \ge n\varepsilon) = P\left(|X - np| \ge \frac{\sqrt{n}\varepsilon}{\sqrt{pq}}\sqrt{npq}\right)$$

이고, X의 평균과 표준편차는 각각 $\mu = np$, $\sigma = \sqrt{npq}$이므로 $k = \sqrt{n}\,\varepsilon / \sqrt{pq}$라 두면 체비세프의 부등식을 이용하여 다음 식을 얻는다.

$$P\left(\left|\frac{X}{n} - p\right| \ge \varepsilon\right) = P(|X - \mu| \ge k\sigma) \le \frac{1}{k^2} = \frac{pq}{n\varepsilon^2}.$$

즉, $P\left(\left|\frac{X}{n} - p\right| < \varepsilon\right) \ge 1 - \frac{p(1-p)}{n\varepsilon^2}$를 얻는다. $p(1-p)$는 $p = 1/2$일 때 최대값 $1/4$를 갖게 되므로

$$P\left[\left|\frac{X}{n} - p\right| < \varepsilon\right] \ge 1 - \frac{p(1-p)}{n\varepsilon^2} \ge 1 - \frac{1}{4n\varepsilon^2}$$

이다. 고정된 $\varepsilon > 0$, $0 < p < 1$에 대하여 다음 식이 성립함을 알 수 있다.

$$1 \ge \lim_{n \to \infty} P\left[\left|\frac{X}{n} - p\right| < \varepsilon\right] \ge \lim_{n \to \infty}\left(1 - \frac{p(1-p)}{n\varepsilon^2}\right) = 1.$$

즉, $\lim_{n \to \infty} P\left[\left|\frac{X}{n} - p\right| < \varepsilon\right] = 1$임을 알 수 있다.

[예 2.5-4] 국회의원 선거에서 김 후보와 박 후보가 경쟁을 하고 있다고 하자. 김 후보의 진영에서는 500명의 유권자를 랜덤하게 뽑아서 그 중 300명이 김 후보를 지지하고 있다는 것을 알았다. 김 후보 진영에서는 선거에서 이길 것을 어느 정도 확신할 수 있는가?(단 허용오차 $\epsilon = 0.1$)

풀이> X: $n = 500$명의 유권자 중에서 김 후보를 지지하는 유권자의 수

p: 김 후보를 지지하는 총유권자의 비율

대수의 법칙을 이용하여

$$P\left(\left|\frac{X}{500} - p\right| < 0.1\right) \geq 1 - \frac{1}{4 \cdot 500 \cdot (0.1)^2} = 0.95$$

임을 알 수 있다. $X/500 = 300/500 = 0.6$이므로 그들은 $0.5 < p < 0.7$임을 95%이상 확신할 수 있다.

6. 초기하분포

[정의 2.6-1] 확률변수 X의 밀도함수

$$f(x) = \frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}, \quad x = 0, 1, \cdots, r, \quad r = \min(n, K)$$

로 주어질 때, X는 모수가 (M, K, n)인 초기하 분포 (Hypergeometric distribution)를 갖는다고 한다.

[예 2.6-1] 위의 정의에서 $M = 100$, $K = 10$(즉, 불량률이 10%)일 때, 비복원으로 8개를 꺼내는 경우 불량품이 2개 이하인 확률은?

풀이> 초기하 분포를 이용하면 $P[X \leq 2] = \sum_{x=0}^{2} \frac{\binom{10}{x}\binom{90}{8-x}}{\binom{100}{8}} = 0.97$

이항분포를 이용하면 $\sum_{x=0}^{2} \binom{8}{x}(0.1)^x(0.9)^{8-x} = 0.96.$

[정리 2.6-1] $K/M=p$를 고정시키고, $M\to\infty$일 때, 초기하 분포는 근사적으로 이항분포와 같다. 즉,

$$\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} \to \binom{n}{x}p^x(1-p)^{n-x}.$$

증명 $(M)_n = M\cdot(M-1)\cdot\cdots\cdot(M-n+1) = \binom{M}{n}n!$이라 두면,

$$\begin{aligned}\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} &= \frac{(K)_x(M-K)_{n-x}n!}{x!(n-x)!(M)_n} = \frac{(K)_x(M-K)_{n-x}}{(M)_n}\frac{n!}{x!(n-x)!}\\ &= \frac{K}{M}\frac{K-1}{M-1}\cdots\frac{K-x+1}{M-x+1}\frac{M-K}{M-x}\frac{M-K-1}{M-x+1}\cdots\frac{M-K-(n-x)+1}{M-x-(n-x)+1}\binom{n}{x}\\ &\to \binom{n}{x}p^x(1-p)^{n-x}\end{aligned}$$

[정리 2.6-2] 확률변수 X가 초기하 분포를 따르면 X의 평균과 분산은 다음과 같다.

$$E(X)=np,\quad Var(X)=np(1-p)\frac{M-n}{M-1},\quad p=\frac{K}{M}.$$

증명 $\binom{K}{x}=\frac{K}{x}\binom{K-1}{x-1}$, $\binom{M}{n}=\frac{M}{n}\binom{M-1}{n-1}$이므로

$$E(X)=\sum_{x=0}^{n}x\cdot\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} = n\cdot\frac{K}{M}\sum_{x=1}^{n}\frac{\binom{K-1}{x-1}\binom{M-K}{n-x}}{\binom{M-1}{n-1}}$$

이며, $y=x-1$로 두면, $\sum_{y=0}^{n-1}\frac{\binom{K-1}{y}\binom{M-1-K+1}{n-1-y}}{\binom{M-1}{n-1}}=1$ (초기하 분포의 pdf는 밀도함수의 성질을 만족하므로, 또는 조합의 성질을 이용하여)이므로 X의 기대값은

$$E(X)=n\cdot\frac{K}{M}\sum_{y=0}^{n-1}\frac{\binom{K-1}{y}\binom{M-1-K+1}{n-1-y}}{\binom{M-1}{n-1}} = np.$$

$E[X(X-1)]$을 구하기 위하여 $y=x-2$로 둔다.

$$\begin{aligned}E[X(X-1)] &= \sum_{x=0}^{n}x(x-1)\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}\\ &= n(n-1)\frac{K(K-1)}{M(M-1)}\sum_{x=2}^{n}\frac{\binom{K-2}{x-2}\binom{M-K}{n-x}}{\binom{M-2}{n-2}}\\ &= n(n-1)\frac{K(K-1)}{M(M-1)}\sum_{y=0}^{n-2}\frac{\binom{K-2}{y}\binom{M-2-K+2}{n-2-y}}{\binom{M-2}{n-2}}\\ &= n(n-1)\frac{K(K-1)}{M(M-1)}\\ &= n(n-1)p\cdot\frac{K-1}{M-1}\end{aligned}$$

이므로 X의 분산은 다음과 같다.

$$\begin{aligned}Var(X) &= E(X^2)-E(X)^2 = E(X(X-1))+E(X)-E(X)^2\\ &= np(1-p)\frac{M-n}{M-1}.\end{aligned}$$

[예 2.6-2] 부산의 어떤 동에 M명이 살고 있는데 그 중 절반이 여성이라고 한다. 이 동에서 랜덤하게 10명을 선발했을 때 여성(x)이 0, 1, 2, 3, 4, 5명이 선발될 확률을 구해본다. M=50, 100, 200, ∞인 경우에 구한 확률 값은 다음표와 같고 $M=\infty$인 경우의 확률은 이항분포에서의 확률과 같고 M의 값이 커짐에 따라 이항분포 확률 값에 가까워짐을 알 수 있다.

x	$M=50$	$M=100$	$M=200$	$M=\infty$
0	0.0003	0.0006	0.0008	0.0010
1	0.0050	0.0072	0.0085	0.0098
2	0.0316	0.0380	0.0410	0.0439
3	0.1076	0.1131	0.1153	0.1172
4	0.2181	0.2114	0.2082	0.2051
5	0.2748	0.2539	0.2525	0.2461

7. 기하분포

• 성공할 확률이 p인 베르누이 시행

X: 처음 성공이 일어날 때까지 계속 실험을 수행하여 처음 성공(S)이 일어날 때까지의 실패(F)회수

실험결과 $F, F, F, S, \cdots$,인 경우 $X=3$이며, 그 확률은

$$P(X=3)=(q)(q)(q)(p)=q^3p=(1-p)^3p$$

[정의 2.7-1] 확률변수 X의 pdf가

$$f(x)=(1-p)^xp,\ x=0,1,2,\cdots$$

로 주어질 때 X의 모수가 p인 기하분포(Geometric distribution)를 갖는다고 하며, $G(p)$로 표현

[기하분포의 성질]

(1) 무한등비급수의 합은 등비 r이 $|r|<1$일 때, $\sum_{k=0}^{\infty} ar^k=\frac{a}{1-r}$이므로

$\sum_{x=0}^{\infty} f(x)=\sum_{x=0}^{\infty}(1-p)^xp=\frac{p}{1-(1-p)}=1$이다. 즉, $f(x)$는 pdf 의 조건을 만족한다.

(2) $P(Y\geq k)=\sum_{x=k}^{\infty}(1-p)^xp=\frac{(1-p)^kp}{1-(1-p)}=(1-p)^k$ 이므로

$P(X<k)=\sum_{x=0}^{k-1}(1-p)^xp=1-(1-p)^k$ 이다.

(3) X의 적률생성함수는 아래와 같이 구해진다.

$$M(t) = E[e^{tX}] = \sum_{x=0}^{\infty} e^{tx}(1-p)^x p = \sum_{x=0}^{\infty}[(1-p)e^t]^x p = \frac{p}{1-(1-p)e^t}.$$

이때, 위의 급수가 수렴할 구간은 $(1-p)e^t < 1$, 즉 $t < -\ln(1-p)$이다.

[정리 2.7-1] 확률변수 X가 모수 p인 기하분포를 따를 때, X의 평균과 분산

$$E(X) = \frac{1-p}{p}, \quad Var(X) = \frac{1-p}{p^2}.$$

[증명] X의 적률생성함수 $M(t)$의 1차, 2차 도함수

$$M'(t) = \frac{p(1-p)e^t}{[1-(1-p)e^t]^2},$$

$$M''(t) = \frac{p(1-p)e^t}{[1-(1-p)e^t]^2} + \frac{2p(1-p)^2e^{2t}}{[1-(1-p)e^t]^3}.$$

$$E(X) = M'(0) = \frac{1-p}{p}, \quad E(X^2) = M''(0) = \frac{1-p}{p} + \frac{2(1-p)^2}{p^2}$$

$$Var(X) = \frac{1-p}{p^2}$$

[예 2.7-1] 주사위를 독립적으로 던질 때, 처음으로 4가 6번째 시행에 나타날 확률은

$$P(X=5) = \left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right) = 0.067$$

4가 나타나는데 적어도 6번 시행해야 할 확률은

$$P(X \geq 5) = \sum_{x=5}^{\infty}\left(\frac{5}{6}\right)^x\left(\frac{1}{6}\right) = \left(\frac{5}{6}\right)^5 = 0.402$$

4가 나타나는데 많아야 5회 시행해야 할 확률은

$$P(X \leq 4) = \sum_{x=0}^{4}\left(\frac{5}{6}\right)^x\left(\frac{1}{6}\right) = 1-\left(\frac{5}{6}\right)^5 = 0.598$$

[참고] 확률변수 X가 모수 p인 기하분포를 따를 때, 새로운 확률변수 $Y = X+1$이라고 정의해 보자. 즉, 확률변수 Y는 처음 성공이 일어날 때까지의 총 시행회수를 뜻한다. 이때, Y의 기댓값은

$$E(Y) = E(X+1) = E(X)+1 = \frac{1-p}{p}+1 = \frac{1}{p}$$

이며, Y의 분산은 X의 분산과 같다.

8. 음이항분포

x: 성공할 확률이 p인 베르누이 시행에서 r회 성공이 나타날 때까지의 실패의 회수
$x+r$: r회 성공이 나타날 때까지의 총 시행회수
r번째 성공은 $x+r$회째 시행에서 있게 된다.

[정의 2.8-1] 확률변수 X의 pdf가

$$f(x)=\binom{x+r-1}{x}p^r q^x,\quad x=0,1,\cdots,\quad q=1-p$$

로 주어질 때, X는 모수가 (r,p)인 음이항분포(Negative binomial distribution)를 갖는다. $(X\sim Nb(r,p))$

※ $\binom{x+r-1}{r-1}=\binom{x+r-1}{x}$

[예 2.8-1] 주사위를 6이 두 번째 나올 때까지 계속해서 던진다. 두 번째 6이 나오기 전에 6이 아닌 것이 꼭 10회 나타날 확률은

$$\binom{10+2-1}{1}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^{10}=0.049$$

[음이항분포의 성질]

(1) pdf의 조건: $\sum_{x=0}^{\infty}f(x)=1$성립

맥로린의 급수 전개 $(1+t)^n=\sum_{x=0}^{\infty}\binom{n}{x}t^x$이므로 $t=p-1$, $n=-r$을 대입하면

$$[1+(p-1)]^{-r}=[1-(1-p)]^{-r}=\sum_{x=0}^{\infty}\binom{-r}{x}(p-1)^x=\sum_{x=0}^{\infty}\binom{-r}{x}(1-p)^x(-1)^x$$

$$\binom{-r}{x}(1-p)^x(-1)^x=\binom{-r}{x}(p-1)^x=\frac{-r(-r-1)\cdots(-r-x+1)}{x!}(p-1)^x$$

$$=\frac{r(r+1)\cdots(r+x-1)}{x!}(1-p)^x$$

이므로 $[1+(p-1)]^{-r}=\sum_{x=0}^{\infty}\binom{r+x-1}{x}(1-p)^x$이다. 따라서,

$$p^r[1-(1-p)]^{-r}=\sum_{x=0}^{\infty}\binom{r+x-1}{x}(1-p)^x p^r=\sum_{x=0}^{\infty}f(x)=1$$

(2) 음이항분포를 따르는 확률변수 X의 적률생성함수

$$\begin{aligned}M(t)&=E[e^{tX}]\\&=\sum_{x=0}^{\infty}e^{tx}f(x)=\sum_{x=0}^{\infty}e^{tx}\binom{r+x-1}{x}(1-p)^x p^r\\&=p^r\sum_{x=0}^{\infty}\binom{r+x-1}{x}[(1-p)e^t]^x=p^r[1-(1-p)e^t]^{-r}\\&=\left(\frac{p}{1-(1-p)e^t}\right)^r.\end{aligned}$$

정리 2.8-1 음이항분포를 따르는 확률변수 X의 평균과 분산은 다음과 같다.

$$E(X)=\frac{rq}{p},\quad Var(X)=\frac{rq}{p^2}.$$

증명 적률생성함수 $M(t)$의 1차, 2차 도함수를 구하면

$$M'(t)=r(1-p)p^r e^t[1-(1-p)e^t]^{-r-1},$$
$$M''(t)=r(1-p)p^r[\ (1-p)(r+1)e^{2t}\{1-(1-p)e^{2t}\}^{-r-2}+e^t\{1-(1-p)e^t\}^{-r-1}].$$

따라서 평균과 분산은 아래와 같이 쉽게 구해진다.

$$E(X)=M'(0)=\frac{r(1-p)}{p},$$
$$Var(X)=M''(0)-[M'(0)]^2=\frac{r(1-p)}{p^2}.$$

[참고] ① 확률변수 $X_1, X_2, \cdots, X_r$이 서로 독립이고 모수 p인 기하분포를 갖는다면 확률변수 $Y=X_1+X_2+\cdots+X_r$은 모수가 (r, p)인 음이항 분포를 갖는다.

② 확률변수 $X_1, X_2, \cdots, X_m$이 서로 독립이고 각각의 분포가 $Nb(r_1, p)$, $Nb(r_2, p)$, $\cdots$, $Nb(r_m, p)$일 때 확률변수 $W=X_1+\cdots+X_m$는 모수가 $(r_1+\cdots+r_m, p)$인 음이항 분포를 갖는다.

③ 확률변수 Y를 r번째 성공이 일어날 때까지의 총 시행회수라고 하면 $Y=X+r$로 둘 수 있다. Y의 평균과 분산은 다음과 같다.

$$E[Y]=E[X+r]=r\cdot\frac{1-p}{p}+r=\frac{r}{p},$$
$$Var[Y]=Var[X+r]=Var(X)=\frac{r(1-p)}{p^2}.$$

9. 포아송분포

주어진 시간이나 물리적인 현상에서 특정한 사건이 일어나는 회수
(일정시간대에 전화 통화의 수, 고객 수, 교통사고의 수, 책의 각 쪽에 오타의 수 등)

[정의 2.9-1]

어떤 실험이 **포아송 과정** (Poisson process)이 되기 위해서는 다음의 4가지 조건을 만족해야 한다.

(a) 겹치지 않는 구간에서 일어나는 사건의 수는 서로 독립이다.
(b) 짧은 구간에 하나의 사건이 일어나는 확률은 구간의 길이에 비례한다.
(c) 충분히 짧은 구간에서 둘 이상의 사건이 일어날 확률은 0이다.
(d) 전 구간을 통하여 어느 부분에서나 사건은 동일하게 일어난다.

- 확률변수X: 길이가 1인 구간에서 일어나는 사건의 수, $P(X=k)$, k; 음이 아닌 정수
- 단위구간(길이:$1/n$, $n>k$)에서 사건이 한번 일어날 확률: 조건(b)에 의해서 $\lambda(1/n)$, 조건 (a)에 의하여 확률 $p=\lambda(1/n)$인 베르누이 실험을 n회 시행하는 경우와 같다.

- 포아송분포를 이용하면 이항분포의 근사적 확률 값을 구하기 쉽다.

모수 (n,p)인 이항분포에서 $np:=\lambda$라 고정시키고 n을 매우 크게 하여 이항분포 pdf

$$\begin{aligned} f(x) &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!}p^x(1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{n^x}\frac{\lambda^x}{x!}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-x} \end{aligned}$$

$\lim_{n\to\infty}\left(1+\frac{c}{n}\right)^n = e^c = 1+c+\frac{c^2}{2!}+\frac{c^3}{3!}+\cdots$, 여기서 자연대수(오일러수) $e=\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$

$\lim_{n\to\infty}\frac{n(n-1)\cdots(n-x+1)}{n^x} = \lim_{n\to\infty}\left[1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)\right]=1$이고

$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$, $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}=1$이므로 다음이 성립한다. $\lim_{n\to\infty}f(x) = \frac{\lambda^x}{x!}e^{-\lambda}$.

[정의 2.9-2] $X \sim Po(\lambda)$; 확률변수 X의 pdf가

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

으로 주어지면 X는 모수가 λ인 포아송 분포(Poisson distribution)를 갖는다.

[포아송 분포의 성질]

(1) $f(x) \geq 0$이고 $\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{-\lambda} e^{\lambda} = 1$; $f(x)$는 pdf의 조건을 만족한다.

(2) 포아송 분포의 적률생성함수는 모든 실수 값 t에 대하여

$$M(t) = \sum_{x=0}^{\infty} e^{tx}\left(\frac{\lambda^x e^{-\lambda}}{x!}\right) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)}$$

[정리 2.9-1] 확률변수 X가 포아송 분포를 갖는다면 X의 평균과 분산

$$E(X) = Var(X) = \lambda.$$

즉, 포아송 분포에서의 평균과 분산은 같고 그 값은 λ이다.

증명> 확률변수 X의 적률생성함수 $M(t)$의 1차, 2차 도함수를 구하면

$$M'(t) = \lambda e^{-\lambda} e^t e^{\lambda e^t},$$

$$M''(t) = \lambda e^{-\lambda} e^t e^{\lambda e^t}(\lambda e^t + 1)$$

$$E(X) = M'(0) = \lambda, \quad Var(X) = M''(0) - [M'(0)]^2 = \lambda(\lambda + 1) - \lambda^2 = \lambda$$

[예 2.9-1] 10,000명이 사는 도시에 100명의 색맹이 있다고 하자. 이 도시에서 200명을 랜덤하게 뽑을 때, 색맹이 3명 이하로 나타날 확률은?

풀이> (1) 초기하분포를 이용하면 다음과 같고 계산이 복잡하다.

$$P[X \leq 3] = \sum_{x=0}^{3} \frac{\binom{100}{x}\binom{9900}{200-x}}{\binom{10000}{200}}$$

(2) 이항분포를 이용하면 100/10000=0.01이므로

$$P[X \leq 3] \approx \sum_{x=0}^{3} \binom{200}{x}(0.01)^x(0.99)^{200-x} \approx 0.858$$

(3) 포아송분포를 이용하면 $\lambda = 200 \times 0.01 = 2$ $P[X \leq 3] \approx \sum_{x=0}^{3} \frac{e^{-2} 2^x}{x!} \approx 0.857$

※일반적으로 $n \geq 20$, $p \leq 0.05$일 때 이항분포와 포아송분포의 확률 값은 거의 같다.

[정의]

확률변수 X의 pdf가

$$f(x_1, x_2, \cdots, x_k) = \frac{n!}{x_1!\ x_2!\ \cdots\ x_k!} p_1^{x_1}\ p_2^{x_2} \cdots\ p_k^{x_k}, \quad \sum_{i=1}^{k} x_i = n, \quad \sum_{i=1}^{k} p_i = 1$$

과 같이 주어질 때, X는 **다항분포** (Multinomial distribution)를 갖는다고 한다.

[정의]

위의 정의에서 $k=3$, $X_3 = n - X_1 - X_2$, $p_3 = 1 - p_1 - p_2$라 두면

$$f(x_1, x_2) = \frac{n!}{x_1!\ x_2!\ (n-x_1-x_2)!} p_1^{x_1}\ p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2}$$

이 되는데 이는 **삼항분포** (Trinomial distribution)의 pdf이다.

예 어떤 기업체에서 생산하는 제품의 95%는 합격품, 4%는 준합격품, 1%는 불합격품이라고 한다. 이 기업체에서는 하루에 생산하는 제품 중에서 20개를 랜덤하게 뽑아서 준합격품과 불합격품을 선별하는데, 준합격품이 3개 이상이거나 불합격품이 2개 이상이면 생산부장이 책임 추궁을 당한다고 한다. 생산부장이 책임 추궁을 당할 확률은 얼마인가?

풀이 준합격품의 개수를 X, 불합격품의 개수를 Y라고 두면 구하고자 하는 확률은

$$\begin{aligned} P[(X \geq 3) \cup (Y \geq 2)] &= 1 - P[(X \leq 2) \cap (Y \leq 1)] \\ &= 1 - \{P[X=0, Y=0] + P[X=0, Y=1] \\ &\quad + P[X=1, Y=0] + P[X=1, Y=1] \\ &\quad + P[X=2, Y=0] + P[X=2, Y=1]\} \end{aligned}$$

이며, (X, Y)의 밀도함수 $f(x, y) = P[X=x, Y=y]$는 다음과 같은 삼항분포의 형태이다.

$$f(x, y) = \frac{20!}{x!\ y!\ (20-x-y)!} (0.04)^x (0.01)^y (0.95)^{20-x-y}.$$

따라서,

$$f(0, 0) = 0.3585, \quad f(0, 1) = 0.3019, \quad f(1, 0) = 0.0755,$$
$$f(1, 1) = 0.0604, \quad f(2, 0) = 0.1208, \quad f(2, 1) = 0.0229$$

이므로 구하는 확률은 $P[(X \geq 3) \cup (Y \geq 2)] = 0.06$이다.

[정의]

(a) $P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n) = P(X_1=x_1)P(X_2=x_2)\cdots P(X_n=x_n)$ 일 때, $X_1, X_2, \cdots, X_n$을 서로 독립인 확률변수라고 한다.

(b) $f_i(x_i)$를 확률변수 X_i, $i=1, 2, \cdots, n$의 pdf라 할 때,

$$P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n)$$

을 확률변수 $X_1, X_2, \cdots, X_n$의 **결합확률밀도함수** (Joint probability density function)라 한다.

(c) $P(X_1=x_1, X_2=x_2, \cdots, X_n=x_n) = f(x_1)f(x_2)\cdots f(x_n)$이면 확률변수 $X_1, X_2, \cdots, X_n$은 서로 독립이고 동일한(혹은 항등인) 분포 (Independent and identical distribution, 간략하게 iid로 표현함)를 갖는다고 말한다.

[정의]

iid인 확률변수들의 모임을 pdf가 $f(x)$인 분포로부터의 **확률표본** (Random sample)이라 하고 n을 **표본의 크기** (Sample size)라 한다.

[정의]

확률변수의 공간이 구간(혹은 구간의 합집합)으로 주어질 때, 연속형 확률변수라고 한다.

[정의]

$x_1, x_2, \cdots, x_n$을 확률표본 $X_1, X_2, \cdots, X_n$의 측정값이라 하고, $N(\{x_i : x_i \le x\})$를 x보다 작거나 같은 측정값들의 개수라 할 때,

$$F_n(x) = \frac{1}{n} N(\{x_i : x_i \le x\})$$

를 **실험분포함수** (Empirical distribution function)라 한다.

[참고] (실험분포함수 $F_n(x)$의 성질)

① 대수의 법칙에 의하여 $n \to \infty$일 때, $F_n(x) \to F(x)$이다.

② $0 \le F_n(x) \le 1$.

③ $x \nearrow \infty$일 때, $F_n(x) \nearrow 1$이다.

④ $F_n(x) = \begin{cases} 0, & x < \min\{x_i\} \\ 1, & x \ge \max\{x_i\}. \end{cases}$

⑤ $F_n(x)$는 계단함수(Step function)이다.

[정의]

$h(x) = \dfrac{f_i}{n(c_i - c_{i-1})}$, $c_{i-1} < x \le c_i$, $i=1, 2, \cdots, k$를 **상대도수 히스토그램** (Relative frequency histogram)이라고 한다.

[정의]

(a) 표본평균 : $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) 표본분산 : $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$.

(c) 표본 중위수 (Median) : $m = \begin{cases} X_{((n+1)/2)}, & n\text{이 홀수일 때} \\ \frac{1}{2}(X_{(n/2)} + X_{(n/2+1)}), & n\text{이 짝수일 때} \end{cases}$

(d) 표본 중간범위수 (Midrange) : $M = \frac{1}{2}(X_{(1)} + X_{(n)})$.

(e) 범위 (Range) : $R = X_{(n)} - X_{(1)}$.

(f) 사분위수 범위 (Interquartile range) = 제 3사분위수 − 제 1사분위수.

(g) 절사평균 (Trimmed mean) : 관측 값의 양쪽에서 일정 비율 α의 극단 값을 버린 후 나머지의 평균.

(h) 윈저화 평균 (Winzorized mean) : 관측 값의 양쪽에서 각각 극단 값을 100α분위수와 $100(1-\alpha)$분위수의 값으로 대체하여 구한 평균.