이산확률분포

著 : 雀
sukita1729@gmail.com

Ⅰ. 이산확률분포

표본공간 S에 대하여,

$$X : S \to \mathbb{R}, \quad X(e_i) = x_i \in \mathbb{R}$$

의 함수 X를 생각하자. S가 유한집합이거나 가산집합인 경우(자연수 집합 $\mathbb{N}$과 동등한 경우) X를 이산확률변수라 한다. 이때 $X(e_i) = x_i$를 편의상 $X = x_i$로 쓴다.

또한

$$f : X \to P(X) \subset [0, 1]$$

의 함수에서 x_i는 $P(X = x_i)$에 대응되고, 이를 확률질량함수 p_i라 한다. 이때 p_i는 다음의 조건을 만족한다.

(i) $0 \le p_i \le 1$
(ii) $\sum_i p_i = 1$

이산확률분포를 시각화하기 위하여 주로 이산확률분포표와 이산확률분포 그래프를 그린다. (본문에서는 생략한다.)

또한 평균 m, 분산 σ^2, 표준편차 σ는 각각 다음과 같이 정의되는 값이다.

① $m = E(X) = \sum_i x_i p_i$
② $\sigma^2 = V(X) = \sum_i (x_i - m)^2 p_i = E(X^2) - E(X)^2$
③ $\sigma = S(X) = \sqrt{V(X)}$

위 정의를 이용하면 $a, b, c \in \mathbb{R}$에 대하여 다음 성질들을 쉽게 증명할 수 있다.

① $E(aX + b) = aE(X) + b$
② $E(aX^2 + bX + c) = aE(X^2) + bE(X) + c$

③ $V(aX+b)=a^2V(X)$
④ $S(aX+b)=|a|S(X)$

이중 ①과 ②는 $\sum$의 선형성에 의해 자명한 결과이다. 다음 절부터는 대표적인 이산확률분포들의 정의를 살펴보고 그 확률질량함수와 기댓값, 그리고 분산을 구해보도록 하겠다.

Ⅱ. 베르누이 분포

베르누이 분포는 가장 간단한 형태의 이산확률분포이다. 베르누이 시행이란 성공할 확률이 p $(0 \le p \le 1)$, 실패할 확률이 $q=1-p$로 일정한 시행을 의미하며, 베르누이 분포의 확률변수 X는 이 베르누이 시행을 1회 실시했을 때의 성공 횟수이다. 따라서 다음과 같은 값들을 얻을 수 있다.

$$E(X)=p, \quad E(X^2)=p, \quad V(X)=p-p^2=pq$$

Ⅲ. 이항분포

이항분포는 고등학교 교육과정 확률과 통계에서 유일하게 배우는 이산확률분포이다. 이항분포의 확률변수 X는 앞서 정의한 베르누이 시행을 n회 실시하였을 때의 성공 횟수이며, 확률변수 X가 이와 같은 이항분포를 따른다는 것을 $X \sim \mathrm{B}(n, p)$와 같이 표기한다. (p는 베르누이 시행의 성공 확률이다.)

이때 확률질량함수는 n회 중 i번 성공한 횟수를 고르는 경우의 수 ${}_n\mathrm{C}_i$에 확률을 곱하면

$$P(X=i)=p_i={}_n\mathrm{C}_i p^i q^{n-i}$$

이다. 또한 기댓값 m은

$$m=E(X)=\sum_{i=0}^{n} i \cdot {}_n\mathrm{C}_i p^i q^{n-i}=\sum_{i=1}^{n} n \cdot {}_{n-1}\mathrm{C}_{i-1} p^i q^{n-i}$$

$$=np\sum_{i=1}^{n} {}_{n-1}\mathrm{C}_{i-1} p^{i-1} q^{n-i}=npR$$

한편 R은 $\mathrm{B}(n-1, p)$를 따르는 확률분포의 확률질량함수의 총합이므로 확률질량함수의 성질에 의해 $R=1$이다. 따라서

$$m=E(X)=np$$

이다. (이는 미분을 통해서도 증명할 수 있다.)

분산을 구하기 위해 $E(X(X-1))$을 구하면,

$$E(X(X-1))=\sum_{i=0}^{n} i(i-1)\cdot {}_nC_i p^i q^{n-i} = \sum_{i=2}^{n} n(n-1)\cdot {}_{n-2}C_{i-2} p^i q^{n-i}$$

$$= n(n-1)p^2 \sum_{i=2}^{n} {}_{n-2}C_{i-2} p^{i-2} q^{n-i} = n(n-1)p^2 R$$

이다. 한편 R은 $\mathrm{B}(n-2, p)$을 따르는 확률분포의 확률질량함수의 총합이므로 확률질량함수의 성질에 의해 $R=1$이다. 따라서

$$E(X(X-1)) = (n^2-n)p^2, \quad E(X^2) = E(X(X-1)) + E(X) = (n^2-n)p^2 + np$$

이고, 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = np(1-p) = npq$$

이다.

Ⅳ. 초기하분포

초기하분포의 확률변수 X는 흰 공 M개, 검은 공 $N-M$개가 들어 있는 주머니에서 비복원 추출로 n개를 뽑았을 때 나온 흰 공의 개수이다. 확률변수 X가 이와 같은 초기하분포를 따른다는 것을 $X \sim \mathrm{HYP}(N, M, n)$과 같이 표기하며, 그 확률질량함수는

$$p_i = \frac{{}_MC_i \cdot {}_{N-M}C_{n-i}}{{}_NC_n}$$

와 같이 구해진다. 이때 기댓값 m은

$$m = E(X) = \sum_{i=0}^{n} \frac{i \cdot {}_MC_i \cdot {}_{N-M}C_{n-i}}{{}_NC_n} = \sum_{i=1}^{n} \frac{M \cdot {}_{M-1}C_{i-1} \cdot {}_{N-M}C_{n-i}}{{}_NC_n}$$

$$= \frac{{}_{N-1}C_{n-1}}{{}_NC_n} M \times \sum_{i=1}^{n} \frac{{}_{M-1}C_{i-1} \cdot {}_{N-M}C_{n-i}}{{}_{N-1}C_{n-1}} = \frac{{}_{N-1}C_{n-1}}{{}_NC_n} M \times R$$

이다. 한편 R은 $\mathrm{HYP}(N-1, M-1, n-1)$을 따르는 확률분포의 확률질량함수의 총합이므

로 확률질량함수의 성질에 의해 $R=1$이다. 따라서

$$m=E(X)=\frac{{}_{N-1}\mathrm{C}_{n-1}}{{}_{N}\mathrm{C}_{n}}M=\frac{Mn}{N}=np$$

이다. $(p=\frac{M}{N})$

분산을 구하기 위해 $E(X(X-1))$을 구하면,

$$E(X(X-1))=\sum_{i=0}^{n}\frac{i(i-1)\cdot {}_{M}\mathrm{C}_{i}\cdot {}_{N-M}\mathrm{C}_{n-i}}{{}_{N}\mathrm{C}_{n}}=\sum_{i=2}^{n}\frac{M(M-1)\cdot {}_{M-2}\mathrm{C}_{i-2}\cdot {}_{N-M}\mathrm{C}_{n-i}}{{}_{N}\mathrm{C}_{n}}$$

$$=\frac{{}_{N-2}\mathrm{C}_{n-2}}{{}_{N}\mathrm{C}_{n}}M(M-1)\times\sum_{i=2}^{n}\frac{{}_{M-2}\mathrm{C}_{i-2}\cdot {}_{N-M}\mathrm{C}_{n-i}}{{}_{N-2}\mathrm{C}_{n-2}}=\frac{{}_{N-2}\mathrm{C}_{n-2}}{{}_{N}\mathrm{C}_{n}}M(M-1)\times R$$

이다. 한편 R은 $\mathrm{HYP}\,(N-2, M-2, n-2)$를 따르는 확률분포의 확률질량함수의 총합이므로 확률질량함수의 성질에 의해 $R=1$이다. 따라서

$$E(X(X-1))=\frac{{}_{N-2}\mathrm{C}_{n-2}}{{}_{N}\mathrm{C}_{n}}M(M-1)=\frac{n(n-1)}{N(N-1)}M(M-1),$$

$$E(X^2)=E(X(X-1))+E(X)=\frac{Mn}{N}\left(1+\frac{(n-1)(M-1)}{N-1}\right)$$

이고, 분산 $V(X)$는

$$V(X)=E(X^2)-E(X)^2=\frac{Mn}{N}\cdot\frac{N(N-1)+N(n-1)(M-1)-Mn(N-1)}{N(N-1)}$$

$$=\frac{nM(N-M)(N-n)}{N^2(N-1)}=n\cdot\frac{M}{N}\cdot\frac{N-M}{N}\cdot\frac{N-n}{N-1}=npq\frac{N-n}{N-1}$$

이다. $(p=\frac{M}{N})$

Ⅴ. 기하분포

기하분포의 확률변수 X는 성공할 확률이 p $(0\le p\le 1)$인 베르누이 시행이 처음 성공할 때까지 시행한 횟수이다. 확률변수 X가 이와 같은 초기하분포를 따른다는 것을

$X \sim \mathrm{GEO}(p)$과 같이 표기하며, 그 확률질량함수는

$$p_i = pq^{i-1} \ (i \in \mathbb{N})$$

이다. 기댓값과 분산을 구하기 위해서는 다음과 같이 $\frac{1}{1-x}$의 전개식을 이용해야 한다.

$\frac{1}{1-x}$의 급수전개는

$$\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i \quad (|x| < 1)$$

이고, 이는 $|x| < 1$일 때 우변의 무한 등비급수가 좌변의 값으로 수렴하는 것을 통해 확인할 수 있다. 이제 양변을 x로 미분하면

$$\frac{1}{(1-x)^2} = \sum_{i=0}^{\infty} i \cdot x^{i-1} \quad \cdots \quad [1]$$

이고, x를 곱한 후 다시 양변을 x로 미분하면

$$\frac{x}{(1-x)^2} = \sum_{i=0}^{\infty} i \cdot x^i, \quad \frac{1+x}{(1-x)^3} = \sum_{i=0}^{\infty} i^2 \cdot x^{i-1} \quad \cdots \quad [2]$$

이다. 이제 기댓값 m을 구해보면

$$m = E(X) = \sum_{i=1}^{\infty} i \cdot pq^{i-1} = p\sum_{i=1}^{\infty} i \cdot q^{i-1} = \frac{p}{(1-q)^2} = \frac{1}{p} \quad (\because [1])$$

이고, 제곱의 평균 $E(X^2)$은

$$E(X^2) = \sum_{i=1}^{\infty} i^2 \cdot pq^{i-1} = p \times \frac{1+q}{(1-q)^3} = \frac{1+q}{p^2} \quad (\because [2])$$

이다. 따라서 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = \frac{q}{p^2}$$

이다.

Ⅵ. 음이항분포 (시행횟수 관점)

음이항분포의 확률변수 X는 성공할 확률이 p $(0 \le p \le 1)$인 베르누이 시행이 n번 성공할 때까지 시행한 횟수이다. 확률변수 X가 이와 같은 음이항분포를 따른다는 것을 $X \sim \mathrm{NB}(n, p)$과 같이 표기하며, 그 확률질량함수는

$$p_i = {}_{i-1}\mathrm{C}_{n-1} p^n q^{i-n}$$

과 같이 구해진다. (이항분포는 시행 횟수가 고정되어 있고 성공 횟수가 확률변수인 반면, 음이항분포는 시행 횟수가 확률변수이고 성공 횟수는 고정되어 있다. 즉 음이항분포라는 이름은 이와 같이 이항분포의 반대라는 의미에서 유래하였다.)

이때 기댓값 m은

$$m = E(X) = \sum_{i=n}^{\infty} i \cdot {}_{i-1}\mathrm{C}_{n-1} p^n q^{i-n} = \sum_{i=n}^{\infty} n \cdot {}_{i}\mathrm{C}_{n} p^n q^{i-n}$$

$$= \frac{n}{p} \sum_{i=n}^{\infty} {}_{i}\mathrm{C}_{n} p^{n+1} q^{i-n} = \frac{n}{p} R$$

이다. 한편 R은 $\mathrm{NB}(n+1, p)$를 따르는 확률분포의 확률질량함수의 총합이므로 확률질량함수의 성질에 의해 $R = 1$이다. 따라서

$$m = E(X) = \frac{n}{p}$$

이다. 분산을 구하기 위해 $E(X(X-1))$을 구하면,

$$E(X(X-1)) = \sum_{i=n}^{\infty} i(i-1) \cdot {}_{i-1}\mathrm{C}_{n-1} p^n q^{i-n} = \sum_{i=n}^{\infty} n(i-1) \cdot {}_{i}\mathrm{C}_{n} p^n q^{i-n}$$

$$= \frac{n}{p} \sum_{i=n}^{\infty} (i+1-2) \cdot {}_{i}\mathrm{C}_{n} p^{n+1} q^{i-n}$$

$$= \frac{n}{p} \sum_{i=n}^{\infty} (i+1) \cdot {}_{i}\mathrm{C}_{n} p^{n+1} q^{i-n} - \frac{2n}{p} \sum_{i=n}^{\infty} {}_{i}\mathrm{C}_{n} p^{n+1} q^{i-n}$$

$$= \frac{n}{p} R - \frac{2n}{p} S$$

이다. 한편 R과 S는 각각 $\mathrm{NB}(n+1, p)$를 따르는 확률분포의 기댓값과 확률질량함수의 총

합이므로 확률질량함수의 성질에 의해 $R=\frac{n+1}{p}$, $S=1$이다. 따라서

$$E(X(X-1))=\frac{n(n+1)}{p^2}-\frac{2n}{p}=\frac{n(n+1-2p)}{p^2},$$

$$E(X^2)=E(X(X-1))+E(X)=\frac{n(n+1-p)}{p^2}$$

이고, 분산 $V(X)$는

$$V(X)=E(X^2)-E(X)^2=\frac{n(1-p)}{p^2}=\frac{nq}{p^2}$$

이다.

Ⅶ. 푸아송분포

푸아송분포는 단위 시간 동안 어떤 사건의 발생 횟수를 나타내며, 그 확률질량함수는 이항분포에서 n이 충분히 클 때($n\to\infty$), p가 충분히 작을 때($p\to 0$), 그리고 $np=\lambda$로 일정할 때이다. 이때 λ는 평균 발생 횟수이고, 이를 푸아송 분포의 모수라고 한다. 즉,

$$p'_i={}_n\mathrm{C}_i p^i q^{n-i}=\frac{n(n-1)(n-2)\cdots(n-i+1)}{i!}\times\left(\frac{\lambda}{n}\right)^i\times\left(1-\frac{\lambda}{n}\right)^{n-i}$$

$$=\frac{\lambda^i}{i!}\times\left(1-\frac{\lambda}{n}\right)^n\times\frac{n(n-1)(n-2)\cdots(n-i+1)}{n^i}\times\left(1-\frac{\lambda}{n}\right)^{-i}$$

$$=\frac{\lambda^i}{i!}\times\left(1-\frac{\lambda}{n}\right)^n\times\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{i-1}{n}\right)\times\left(1-\frac{\lambda}{n}\right)^{-i}$$

에서

$$p_i=\lim_{n\to\infty}p_i'=\frac{\lambda^i}{i!}\times e^{-\lambda}\times 1\times 1=\frac{\lambda^i e^{-\lambda}}{i!}$$

이다. (확률변수 X가 이와 같은 푸아송분포를 따른다는 것을 $X\sim\mathrm{POI}(\lambda)$과 같이 표기한다.)

이때 기댓값 m은

$$m = E(X) = \sum_{i=1}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-1)!} = \lambda e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda$$

이다. (e^x의 테일러 전개식을 고려하면 $\sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = e^{\lambda}$이다.)

분산을 구하기 위해 $E(X(X-1))$을 구하면,

$$E(X(X-1)) = \sum_{i=2}^{\infty} \frac{\lambda^i e^{-\lambda}}{(i-2)!} = \lambda^2 e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} = \lambda^2,$$

$$E(X^2) = E(X(X-1)) + E(X) = \lambda + \lambda^2$$

이고, 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = \lambda$$

Ⅷ. 이항분포의 근사

이항분포의 근사가 성립함을 증명하기 위해서는 이항분포의 적률생성함수가 각 조건에서 해당 확률분포의 적률생성함수로 수렴한다는 것을 증명해야 하지만, 본문에서 적률생성함수에 관한 논의는 제외하였으므로 증명 없이 근사를 적용하는 방법만 기술한다.

$\mathrm{B}(n, p)$를 따르는 확률변수 X에 대하여, 만약 np와 nq의 값이 모두 5보다 작다면 X를 푸아송분포 $\mathrm{POI}(\lambda)$ ($\lambda = np$)로 근사한다. 만약 np와 nq 중 5 이상의 값이 있다면 X를 정규분포 $\mathrm{N}(n, p)$로 근사한다. (정규분포에 관한 내용은 연속확률분포 칼럼에서 다룰 것이다.)

예를 들어, $X \sim \mathrm{B}(100, 0.02)$인 확률변수 X에 대하여 $P(X=1) = 100 \times (0.02)^1 \times (0.98)^{99}$로 계산이 아주 복잡하지만, 이를 푸아송분포 $\mathrm{POI}(2)$로 근사하면 이는 $\frac{2}{e^2}$가 된다. 실제로 각각의 값을 계산해보면 오차율이 낮아, 나름대로 타당한 근사임을 알 수 있다.