연속확률분포

著 : 雀
sukita1729@gmail.com

Ⅰ. 연속확률분포

표본공간 S에 대하여,

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

의 함수 X를 생각하자. S가 비가산집합인 경우(실수 집합 $\mathbb{R}$과 동등한 경우) X를 연속확률변수라 한다. 이때 $X(e_i) = x_i$를 편의상 $X = x_i$로 쓴다.

또한

$$f : X \to \mathbb{R}_0^+$$

의 함수에서 x_i는 $f(x_i)$에 대응되고, 이를 확률밀도함수 f라 한다. 이때 f는 다음의 조건을 만족한다.

(i) $f(x_i) \geq 0$

(ii) $\displaystyle\int_{-\infty}^{\infty} f(x)dx = 1$

연속확률분포는 변수가 무한히 많고 셀 수 없으므로 일반적으로 확률밀도함수의 그래프를 그려 표현한다.

또한 확률 $\mathrm{P}(a \leq X \leq b)$와 평균 m, 분산 σ^2, 표준편차 σ는 각각 다음과 같이 정의된다.

① $\mathrm{P}(a \leq X \leq b) = \displaystyle\int_a^b f(x)dx$

② $E(X) = \displaystyle\int_{-\infty}^{\infty} xf(x)dx$

③ $\sigma^2 = V(X) = \displaystyle\int_{-\infty}^{\infty} (x-m)^2 f(x)dx = E(X^2) - E(X)^2$

④ $\sigma = S(X) = \sqrt{V(X)}$

①의 양변을 x로 미분하면 미적분학의 기본 정리에 의해 다음을 얻는다.

$$\frac{d}{dx}\mathrm{P}\,(a \le X \le b) = f(x)$$

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 $\int$의 선형성에 의해 자명한 결과이다. 다음 절부터는 대표적인 연속확률분포들의 정의를 살펴보고 그 확률질량함수와 기댓값, 그리고 분산을 구해보도록 하겠다.

Ⅱ. 균등분포

균등분포는 가장 간단한 형태의 연속확률분포이다. 균등분포는 모든 지점에서의 확률이 동일한 분포이며, 확률변수 X가 a와 b 사이에서의 균등분포를 따른다는 것을 $X \sim \mathrm{U}\,(a, b)$와 같이 표기한다. 따라서 균등분포의 확률밀도함수는

$$f(x) = \begin{cases} \dfrac{1}{b-a} & (a \le x \le b) \\ 0 & otherwise \end{cases}$$

이다. 이때 기댓값 m은

$$m = E(X) = \int_a^b \frac{x}{b-a}dx = \frac{a+b}{2}$$

이고,

$$E(X^2) = \int_a^b \frac{x^2}{b-a}dx = \frac{a^2+ab+b^2}{3}$$

이므로 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = \frac{(a-b)^2}{2}$$

이다.

Ⅲ. 지수분포

시간당 발생률이 μ인 사건 A를 생각하자. 이때 시간 $(0, t)$에서 사건 A의 발생 횟수를 확률변수 X_t라 하면 $X_t \sim \mathrm{POI}(\mu t)$임이 알려져 있다. 즉, X_t의 확률질량함수는

$$\mathrm{P}(X_t = i) = \frac{(\mu t)^i e^{-\mu t}}{i!}$$

이다. 이때 지수분포의 확률변수 X는 사건 A가 처음으로 일어날 때까지 걸린 시간이며, 확률변수 X가 이와 같은 지수분포를 따른다는 것을 $X \sim \mathrm{EXP}(\mu)$와 같이 표기한다. 한편

$$\mathrm{P}(X \le t) = \mathrm{P}(X_t \ge 1) = 1 - \mathrm{P}(X_t = 0) = 1 - e^{-\mu t} \quad (t \ge 0)$$

이므로 지수분포의 확률밀도함수는 양변을 t로 미분하여

$$f(t) = \mu e^{-\mu t} \quad (t \ge 0)$$

이다. 이때 기댓값 m은

$$m = E(X) = \int_0^\infty \mu x \cdot e^{-\mu x} dx = \left[x \cdot (-e^{-\mu x})\right]_0^\infty + \int_0^\infty e^{-\mu x} dx$$

$$= 0 + \left[-\frac{1}{\mu} e^{-\mu x}\right]_0^\infty = \frac{1}{\mu}$$

이고,

$$E(X^2) = \int_0^\infty \mu x^2 \cdot e^{-\mu x} dx = \left[x^2 \cdot (-e^{-\mu x})\right]_0^\infty + \int_0^\infty 2x \cdot e^{-\mu x} dx$$

$$= 0 + \frac{1}{\mu} \cdot \frac{2}{\mu} = \frac{2}{\mu^2}$$

이므로 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = \frac{1}{\mu^2}$$

이다.

Ⅳ. 감마분포

감마분포는 지수분포의 확장의 일종으로, 감마분포의 확률변수 X는 시간당 발생률이 μ인 사건 A가 n번 일어날 때까지 걸린 시간이다. 확률변수 X가 이와 같은 감마분포를 따른다는 것을 $X \sim \mathrm{GAM}(\mu, n)$과 같이 표기하며, 지수분포에서와 같은 방법으로 $\mathrm{P}(X \le t)$를 구하면

$$\mathrm{P}(X \le t) = \mathrm{P}(X_t \ge n) = 1 - \sum_{i=0}^{n-1} \mathrm{P}(X_t = i)$$

$\mathrm{POI}(\mu t)$를 따르는 확률변수 X_t의 확률질량함수를 대입하면

$$\mathrm{P}(X \le t) = 1 - \sum_{i=0}^{n-1} \frac{(\mu t)^i e^{-\mu t}}{i!}$$

이고, 확률밀도함수를 구하기 위해 양변을 t로 미분하면

$$f_n(t) = \frac{d}{dt}\mathrm{P}(X \le t) = \mu e^{-\mu t} - \sum_{i=1}^{n-1}\left[\mu^i e^{-\mu t}\frac{t^{i-1}}{(i-1)!} - \mu^{i+1}e^{-\mu t}\frac{t^i}{i!}\right]$$

$$= \mu e^{-\mu t} + e^{-\mu t}\sum_{i=1}^{n-1}\left[\mu^{i+1}\frac{t^i}{i!} - \mu^i\frac{t^{i-1}}{(i-1)!}\right]$$

$$= \mu e^{-\mu t} + e^{-\mu t}\left[\mu^2 t - \mu + \mu^3\frac{t^2}{2!} - \mu^2 t + \cdots + \mu^n\frac{t^{n-1}}{(n-1)!} - \mu^{n-1}\frac{t^{n-2}}{(n-2)!}\right]$$

$$= \frac{\mu^n t^{n-1} e^{-\mu t}}{(n-1)!} \quad (t \ge 0)$$

이다. 한편 확률밀도함수의 성질에 의해

$$\int_0^\infty \frac{\mu^n t^{n-1} e^{-\mu t}}{(n-1)!} dt = 1$$

이고, 이 형태가 감마함수의 적분 형태와 동일하다는 것에서 감마분포의 이름이 유래하였다.

이때 기댓값 m은

$$m = E(X) = \int_0^\infty \frac{\mu^n t^n e^{-\mu t}}{(n-1)!} dt$$

$$= \frac{\mu^n}{(n-1)!}\left[-\frac{1}{\mu}t^n e^{-\mu t}\right]_0^\infty - \frac{\mu^n}{(n-1)!}\int_0^\infty nt^{n-1}\cdot\left(-\frac{1}{\mu}\right)e^{-\mu t}dt$$

$$= \frac{n\mu^{n-1}}{(n-1)!}\int_0^\infty t^{n-1}e^{-\mu t}dt = \frac{n}{\mu}\int_0^\infty \frac{\mu^n t^{n-1}e^{-\mu t}}{(n-1)!}dt = \frac{n}{\mu}R$$

이다. 한편 R은 $\mathrm{GAM}(\mu, n)$을 따르는 확률분포의 확률밀도함수의 전구간에 대한 정적분 값이므로 확률밀도함수의 성질에 의해 $R=1$이다. 따라서

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

이다. 또한 제곱의 평균 $E(X^2)$은

$$E(X^2) = \int_0^\infty \frac{\mu^n t^{n+1}e^{-\mu t}}{(n-1)!}dt$$

$$= \frac{\mu^n}{(n-1)!}\left[-\frac{1}{\mu}t^{n+1}e^{-\mu t}\right]_0^\infty - \frac{\mu^n}{(n-1)!}\int_0^\infty (n+1)t^n\cdot\left(-\frac{1}{\mu}\right)e^{-\mu t}dt$$

$$= \frac{(n+1)\mu^{n-1}}{(n-1)!}\int_0^\infty t^n e^{-\mu t}dt = \frac{n(n+1)}{\mu^2}\int_0^\infty \frac{\mu^{n+1}t^n e^{-\mu t}}{n!}dt = \frac{n(n+1)}{\mu^2}R$$

이다. 한편 R은 $\mathrm{GAM}(\mu, n+1)$을 따르는 확률분포의 확률밀도함수의 전구간에 대한 정적분 값이므로 확률밀도함수의 성질에 의해 $R=1$이다. 따라서

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

이고, 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = \frac{n}{\mu^2}$$

이다.

Ⅴ. 정규분포

정규분포는 이항분포를 근사하면서 등장한 분포로, 가장 일반적으로 적용 가능한 연속확률분포이다. 이항분포 $\mathrm{B}(n, p)$를 따르는 확률변수 Y의 확률질량함수는 드무아브르 - 라플라

스 정리에 의해 다음과 같이 근사할 수 있으며, 이를 정규분포의 확률밀도함수라고 한다.

$$\mathrm{P}\,(Y=i)={}_n\mathrm{C}_i p^i q^{n-i} \approx \frac{1}{\sqrt{2\pi}\,\sqrt{npq}}\,e^{-\frac{(i-np)^2}{2npq}}$$

$m=np$, $\sigma=\sqrt{npq}$ 를 이용하여 이를 다시 쓰면

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

이며, 확률변수 X가 이와 같은 감마분포를 따른다는 것을 $X\sim \mathrm{N}\,(m,\sigma^2)$과 같이 표기한다. 정규분포의 그래프는 다음과 같은 치환을 통해 평균이 0이고 분산이 1인 표준정규분포로 변환할 수 있으며, 이 과정을 정규분포의 표준화라고 한다.

$$Z=\frac{X-m}{\sigma},\quad f(z)=\frac{1}{\sqrt{2\pi}}\,e^{-\frac{z^2}{2}}\quad(-\infty<z<\infty)$$

정규분포의 확률밀도함수인 e^{-x^2}은 그 부정적분이 초등함수들의 유한 연산으로 표현 불가능하므로 그 정적분의 값을 정확하게 구할 수 없고, 정규분포에서의 확률을 구할 때는 수치적으로 계산한 값들을 취해야 한다. 무한히 많은 정규분포에 대한 정적분 값을 모두 계산해놓을 수는 없으므로, 평균이 0이고 분산이 1인 표준정규분포에 대하여 계산을 해 놓고 다른 정규분포에서는 이를 표준정규분포로 바꾸어 확률을 계산한다. 0부터 양수 a까지의 정적분 값을 모아놓은 표를 정규분포표라고 한다.

한편 e^{-x^2}은 일반적으로 정적분의 정확한 값을 구할 수 없지만 0부터 무한대까지의 정적분 값은 정확하게 구할 수 있는데, 이를 가우스 적분이라 한다. 앞서 제시한 표준정규분포의 확률밀도함수를 고려하면 다음이 성립함을 유추해볼 수 있다.

$$\int_0^\infty e^{-x^2}dx=\frac{\sqrt{\pi}}{2},\quad \int_{-\infty}^\infty e^{-x^2}dx=\sqrt{\pi}$$

증명 방법으로는 극좌표 변환을 통한 방법이 가장 널리 알려져 있으며, 복소 경로적분이나 파인만 적분 테크닉을 통해서도 증명 가능하다. (본문에서는 극좌표 증명만을 제시한다.)

$$I=\int_{-\infty}^\infty e^{-x^2}dx=2\int_0^\infty e^{-x^2}dx$$

$$I^2=4\int_0^\infty e^{-x^2}dx\cdot\int_0^\infty e^{-y^2}dy=4\int_0^\infty\int_0^\infty e^{-(x^2+y^2)}dxdy$$

$$= 4\int_0^{\frac{\pi}{2}}\int_0^{\infty} e^{-r^2} r dr d\theta \ \ (r = x^2 + y^2,\ \theta = \tan^{-1}\left(\frac{y}{x}\right),\ dxdy = rdrd\theta)$$

$$= 4\int_0^{\frac{\pi}{2}} d\theta \cdot \int_0^{\infty} re^{-r^2} dr = 2\pi\left[-\frac{1}{2}e^{-r^2}\right]_0^{\infty} = \pi$$

$$\therefore I = \int_{-\infty}^{\infty} e^{-x^2} dx = \sqrt{\pi}$$

표준정규분포의 기댓값 m을 구해보면, $zf(z)$는 기함수이므로

$$m = E(X) = \int_{-\infty}^{\infty} \frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = 0$$

이고, 제곱의 평균 $E(X^2)$은

$$E(X^2) = \int_{-\infty}^{\infty} \frac{z^2}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \left[-\frac{z}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\right]_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{z^2}{2}} dz$$

$$= 0 + \frac{1}{\sqrt{\pi}}\int_{-\infty}^{\infty} e^{-u^2} du = 1 \quad (z = \sqrt{2}\,u,\ dz = \sqrt{2}\,du)$$

이므로 분산 $V(X)$는

$$V(X) = E(X^2) - E(X)^2 = 1$$

이다.

Ⅵ. 연속확률분포의 활용

현대의 양자역학 모델과 보른 해석에 의하면 슈뢰딩거 파동방정식의 근인 파동함수 ψ에 대하여 $|\psi|^2$은 해당 물질의 존재 확률에 대한 확률밀도함수가 된다. 원자 내의 전자 역시 확률로 존재하므로 각 오비탈 별 파동함수가 존재하며, 이때 $|\psi|^2 dV$는 미소부피 dV에서 전자가 발견될 확률을 뜻한다. 여기서 확률밀도함수의 전구간에 대한 정적분 값이 1이라는 성질이 파동함수의 계수를 결정하는 데 유용하게 작용하며, 이를 규격화 조건이라고 한다.

한편 수소 원자에는 원자핵(양성자)이 있으므로 퍼텐셜이 0이 아니고, 따라서 이를 푸는 것은 매우 까다롭다. 이를 간단하게 만든 것이 3차원 상자 속 입자이며, 이를 더 간단하게 한

2차원과 1차원 상자 속 입자도 있다. 2차원 상자 속 입자부터는 이변수함수와 중적분이 필요하므로, 본문에서는 1차원 상자 속 입자만 다루도록 하겠다.

1차원 상자, 즉 수직선에서 움직이는 입자의 퍼텐셜 V가 위치 x에 대하여 다음과 같이 주어졌다고 가정하자.

$$V(x)=\begin{cases}0 & (0<x<L)\\ \infty & otherwise\end{cases}$$

즉, 입자는 오직 구간 $(0, L)$에서만 존재할 수 있으며 그때의 퍼텐셜 에너지는 0이다. 그 밖의 구간에서는 퍼텐셜 에너지가 무한대로 발산하므로 절대로 도달할 수 없다.

이때 $V(x)=0$이므로 이를 적용하면 다음과 같은 시간에 무관한 슈뢰딩거 방정식을 얻는다.

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2}=E\psi(x)$$

이를 풀고 표준화를 하면 양의 상수 A에 대하여 아래와 같은 파동함수를 얻는다.

$$\psi_n=A\sin\left(\frac{n\pi x}{L}\right)$$

이때 확률밀도함수는 $|\psi_n|^2$이므로 규격화 조건을 이용하면

$$1=\int_{-\infty}^{\infty}|\psi_n|^2dx=A^2\int_0^L\sin^2\left(\frac{n\pi x}{L}\right)dx$$

$$=\frac{A^2}{2}\int_0^L\left[1-\cos\left(\frac{2n\pi x}{L}\right)\right]dx=\frac{A^2}{2}\left[x-\frac{L}{2n\pi}\sin\left(\frac{2n\pi x}{L}\right)\right]_0^L=\frac{L}{2}A^2$$

이므로 $A=\sqrt{\frac{2}{L}}$을 얻는다.

한편 확률밀도함수를 이용하면 특정 구간에서 전자가 존재할 확률을 계산해낼 수 있다.

예제) 수소 원자의 바닥상태의 파동함수 $\psi(r)=\frac{1}{\sqrt{\pi a_0^3}}e^{-\frac{r}{a_0}}$으로 주어진다. 다음 물음에 답하시오. (단, a_0는 보어 반지름이다.)

(1) 주어진 함수 ψ에 대하여 파동함수의 규격화 조건이 성립함을 확인하시오.

(2) 원자핵으로부터 거리가 $\frac{a_0}{2}$인 영역 바깥에서 전자가 발견될 확률을 구하시오.

(이 문제의 $\psi(r)$ 역시 수소 원자에 대한 슈뢰딩거 방정식을 풀어서 얻은 파동함수이다.)

sol) (1) 문제에서 전자는 3차원 공간에 존재하므로 확률밀도함수에 $4\pi r^2$을 곱한 후 정적분을 해야 한다.

$$\int_0^\infty 4\pi r^2 |\psi(r)|^2 dr = \frac{4}{a_0^3}\int_0^\infty r^2 e^{-\frac{2r}{a_0}} dr = \frac{4}{a_0^3}\left\{\left[-\frac{a_0}{2}r^2 e^{-\frac{2r}{a_0}}\right]_0^\infty + \frac{a_0}{2}\int_0^\infty 2re^{-\frac{2r}{a_0}} dr\right\}$$

$$= \frac{4}{a_0^3}\left\{0 + a_0\left[-\frac{a_0}{2}re^{-\frac{2r}{a_0}}\right]_0^\infty + \frac{a_0^2}{2}\int_0^\infty e^{-\frac{2r}{a_0}} dr\right\}$$

$$= \frac{4}{a_0^3}\left\{\frac{a_0^2}{2}\left[-\frac{a_0}{2}e^{-\frac{2r}{a_0}}\right]_0^\infty\right\} = \frac{4}{a_0^3}\left(\frac{a_0^3}{4}\right) = 1$$

원자 내에 전자가 존재하는 것은 확실하므로 전구간에서 전자를 발견할 확률은 자명하게 1이다.

(2) $$\int_{\frac{a_0}{2}}^\infty 4\pi r^2 |\psi(r)|^2 dr = \frac{4}{a_0^3}\int_{\frac{a_0}{2}}^\infty r^2 e^{-\frac{2r}{a_0}} dr = \frac{4}{a_0^3}\left\{\left[-\frac{a_0}{2}r^2 e^{-\frac{2r}{a_0}}\right]_{\frac{a_0}{2}}^\infty + \frac{a_0}{2}\int_{\frac{a_0}{2}}^\infty 2re^{-\frac{2r}{a_0}} dr\right\}$$

$$= \frac{4}{a_0^3}\left\{\frac{a_0^3}{8e} + a_0\left[-\frac{a_0}{2}re^{-\frac{2r}{a_0}}\right]_{\frac{a_0}{2}}^\infty + \frac{a_0^2}{2}\int_{\frac{a_0}{2}}^\infty e^{-\frac{2r}{a_0}} dr\right\}$$

$$= \frac{4}{a_0^3}\left\{\frac{3a_0^3}{8e} + \frac{a_0^2}{2}\left[-\frac{a_0}{2}e^{-\frac{2r}{a_0}}\right]_{\frac{a_0}{2}}^\infty\right\} = \frac{4}{a_0^3}\left(\frac{5a_0^3}{8e}\right) = \frac{5}{2e} \simeq 91.97\%$$

보어 반지름 a_0는 보어 모델에서 전자가 존재할 것으로 예상되는 거리이므로 전자가 핵으로부터 $\frac{a_0}{2}$보다 더 멀리 떨어져 있을 확률이 약 0.9197이라는 것은 타당한 수치이다.

Ⅶ. 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.

예를 들어, 주사위를 100번 던져 짝수의 눈이 나온 경우의 수를 확률변수로 가지는 확률분포는 이항분포 $\mathrm{B}(100, 0.5)$를 따르고, 이때 짝수의 눈이 45번 이상 60번 이하 나온 확률을 계산하라고 한다면 이항분포의 확률질량함수를 16번 계산하여 모두 더해야 한다. 계산과정을 간단히 하게 위해 이를 정규분포 $\mathrm{N}(50, 5^2)$으로 근사할 수 있으며, 이때 확률을 구해보면 표준정규분포를 따르는 확률변수 Z에 대하여

$$\mathrm{P}(45 \le X \le 60) = \mathrm{P}(-1 \le Z \le 2) = \mathrm{P}(0 \le Z \le 1) + \mathrm{P}(0 \le Z \le 2)$$

$$= 0.3413 + 0.4772 = 0.8185$$

로 아주 간단하게 계산이 가능하다. 실제로 각각의 값을 계산해보면 오차율이 낮아, 타당한 근사임을 알 수 있다.