LGE - SNU AI Scientist 고급과정

확률통계 및 통계 방법론

홍현경(hyungyeong81@snu.ac.kr)

2026년 01월

Contents

1 확률과 확률분포

1.1 확률

예제 1. 한 모바일 앱의 하루 사용자 행동을 분석하는 상황을 고려하자. 임의로 선택한 한 사용자의 하루 동안의 행동에 대해 다음 두 사건을 정의하자.

- 사건 A: 해당 사용자가 앱 알림을 한 번 이상 클릭한 사건
- 사건 B: 해당 사용자가 앱 내에서 결제를 한 사건

분석 결과 다음과 같은 확률이 주어졌다고 하자.

$$P(A) = \frac{3}{5}, \qquad P(B) = \frac{1}{2}, \qquad P(A \cup B) = \frac{7}{10}$$

다음 물음에 답하여라.

(a) 해당 사용자가 알림을 클릭하고, 동시에 결제까지 했을 확률을 구하여라.

(b) 해당 사용자가 알림을 클릭한 사건과 결제를 한 사건이 서로 독립인지 판단하여라.

Solution.

(a) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 이므로, $P(A \cap B) = P(A) + P(B) - P(A \cup B)$ 이다.
주어진 값을 대입하면, $P(A \cap B) = \frac{3}{5} + \frac{1}{2} - \frac{7}{10} = \frac{6}{10} + \frac{5}{10} - \frac{7}{10} = \frac{4}{10} = \frac{2}{5}$ 이다.
따라서 해당 사용자가 알림을 클릭하고 동시에 결제까지 했을 확률은 $\frac{2}{5}$ 이다.

(b) 두 사건 A와 B가 서로 독립이려면 $P(A \cap B) = P(A)P(B)$가 성립해야 한다.
이 때, $P(A)P(B) = \frac{3}{5} \times \frac{1}{2} = \frac{3}{10}$ 이므로 이는 (a)에서 구한 $P(A \cap B) = \frac{2}{5}$와 같지 않다.
따라서 사건 A와 B는 서로 독립이 아니다.

예제 2. 서버 3대(서버 1, 서버 2, 서버 3)로 구성된 클라우드 서버 시스템이 병렬로 연결되어 있다. 각 서버는 서로 독립적으로 작동하며, 하나의 서버가 고장날 확률은 0.02이다. 이 시스템은 세 대의 서버 중 적어도 두 대의 서버가 정상적으로 작동할 때 정상적으로 작동한다고 한다. 다음의 물음에 답하여라.

(a) 시스템이 정상적으로 작동할 확률을 구하여라.

(b) 시스템이 정상적으로 작동한다는 것이 관측되었을 때, 서버 1이 정상적으로 작동하고 있을 확률을 구하여라.

Solution.

각 서버가 고장날 확률이 0.02이므로, 정상 작동할 확률은 0.98이다.

(a) 시스템이 정상 작동하려면 3대의 서버가 모두 정상이거나 정확히 2대의 서버가 정상이어야 한다.

- 서버 3대 모두 정상: $(0.98)^3$
- 정확히 2대의 서버가 정상: 고장난 서버가 1번일 때, 2번일 때, 3번일 때의 3가지 경우가 있으므로 $3 \times \{(0.98)^2 \times 0.02\}$

따라서 시스템이 정상 작동할 확률은 $(0.98)^3 + 3 \times \{(0.98)^2 \times 0.02\}$이다.

(b) 구하려는 값은 $P(\text{서버 1 정상} \mid \text{시스템 정상}) = \dfrac{P(\text{서버 1 정상} \cap \text{시스템 정상})}{P(\text{시스템 정상})}$.

서버 1이 정상이고 시스템도 정상이라는 것은, 서버 1은 정상이고, 서버 2와 3 중 적어도 한 대는 정상인 경우이다.

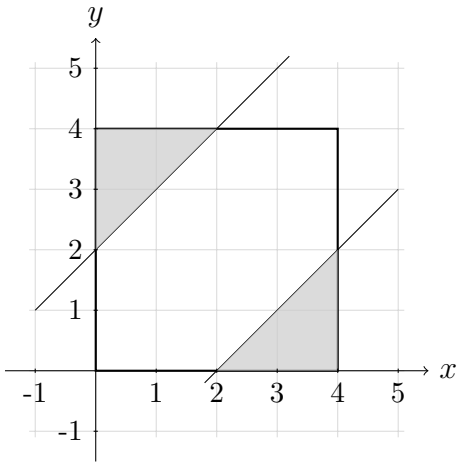
서버 2와 3이 둘 다 고장날 확률은 $(0.02)^2$이므로, 서버 2와 3 중 적어도 한 대가 정상일 확률은 $1 - (0.02)^2$이다.

따라서 $P(\text{서버 1 정상} \cap \text{시스템 정상}) = 0.98 \times \{1 - (0.02)^2\}$이므로,

$P(\text{서버 1 정상} \mid \text{시스템 정상}) = \dfrac{0.98 \times \{1 - (0.02)^2\}}{(0.98)^3 + 3 \times \{(0.98)^2 \times 0.02\}}$.

예제 3. 한 공정 라인에서는 동일한 양을 측정하는 두 개의 센서가 설치되어 있다. 두 센서는 정상 작동 시 측정값이 0부터 4 사이의 값 중에서 서로 독립적으로 균등하게 발생한다고 알려져 있다. 첫 번째 센서의 측정값을 확률변수 X, 두 번째 센서의 측정값을 확률변수 Y 라 하자. 품질 관리 기준에 따라, 두 측정값의 차이가 2보다 크면 측정 시스템에 이상이 있는 것으로 판단한다. 시스템에 이상이 있다고 판단될 확률을 구하여라.

Solution. 우선, 두 확률변수 X, Y 가 독립이며 구간 $[0,4]$ 의 균등분포를 따르므로 표본공간은 $[0,4] \times [0,4]$ 이고, 이 표본공간의 면적은 16이다. 이제, $|X-Y| > 2$ 인 영역을 생각해보자. 확률변수 X, Y 의 실현값을 x, y 라 하면 이 영역은 $x - y > 2$ 와 $x - y < -2$ 의 합집합이므로, 각각의 영역의 면적은 다음 그림과 같다.



이를 통해 회색 부분의 면적을 다음과 같이 구할 수 있다.

$$\begin{aligned} \text{면적} &= \text{면적}(\{(x,y) : x - y > 2\}) + \text{면적}(\{(x,y) : x - y < -2\}) \\ &= \frac{1}{2} \times 2 \times 2 + \frac{1}{2} \times 2 \times 2 = 4 \end{aligned}$$

따라서 $|X-Y| > 2$ 인 영역의 면적은 4이고, 결합확률밀도함수가 $f(x,y) = \frac{1}{16}$ 이므로 $|X-Y| > 2$ 인 사건의 확률은 $\frac{4}{16} = \frac{1}{4}$ 이다.

예제 4. 한 프로야구 구장에서 자동 투구 판정 시스템(Automated Ball-Strike System: ABS)을 운영하고 있다. 이 시스템은 스트라이크 존에 들어오는 투구를 스트라이크(strike), 그 외의 투구를 볼(ball)로 자동으로 판별한다. 이 시스템에 대해 다음과 같은 정보가 주어졌다고 가정하자.

- 전체 투구 중 40%는 실제로 스트라이크이다.
- 실제 스트라이크 투구의 90%를 ABS가 스트라이크라고 올바르게 판정한다.
- 실제 볼 투구의 20%를 ABS가 스트라이크라고 잘못 판정한다.

이 때, 다음의 물음에 답하여라.

(a) ABS가 투구를 스트라이크라고 판단할 확률을 구하여라.

(b) ABS가 스트라이크로 판단한 투구가 실제로 스트라이크일 확률을 구하여라.

Solution.

사건을 다음과 같이 설정하자.

- A: 실제 투구가 스트라이크인 사건
- B: ABS가 투구를 스트라이크라고 판정한 사건

문제에서 주어진 정보에 의해 $P(A) = 0.40$, $P(B \mid A) = 0.90$, $P(B \mid A^c) = 0.20$ 이다.
또한 $P(A^c) = 0.60$이다.

(a) ABS가 투구를 스트라이크라고 판단할 확률은 전확률공식을 이용하여 구할 수 있다. 즉, $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$.

따라서, $P(B) = 0.90 \times 0.40 + 0.20 \times 0.60 = 0.36 + 0.12 = 0.48$이다.

(b) ABS가 스트라이크로 판정한 투구가 실제로 스트라이크일 확률은 베이즈 정리를 이용하여 구할 수 있다. 즉, $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$ 이다.

이 때, $P(A \cap B) = P(B \mid A)P(A)$로 나타낼 수 있다.

(a)에서 구한 $P(B)$를 대입하면,

$P(A \mid B) = \dfrac{0.90 \times 0.40}{0.90 \times 0.40 + 0.20 \times 0.60} = \dfrac{0.36}{0.48} = 0.75$이다.

1.2 확률변수와 확률분포

예제 5. 1부터 5까지의 자연수로 이루어진 집합 $\{1, 2, 3, 4, 5\}$에서 서로 다른 두 원소를 임의로 선택한다. 다음 물음에 답하여라.

(a) 선택된 두 수 중 큰 값을 확률변수 X라고 할 때, $\mathrm{E}(X)$를 구하여라.

(b) 선택된 두 수 중 작은 값을 확률변수 Y라고 할 때, $P(2 \le Y \le 4)$을 구하여라.

Solution.

(a) X의 확률분포는 다음과 같다.

x	2	3	4	5
$p_X(x)$	0.1	0.2	0.3	0.4

$$\therefore E(X) = 1/10 \times 2 + 2/10 \times 3 + 3/10 \times 4 + 4/10 \times 5 = 4$$

(b) Y의 확률분포는 다음과 같다.

y	1	2	3	4
$p_Y(y)$	0.4	0.3	0.2	0.1

이를 이용하면, $P(2 \le Y \le 4) = P(Y = 2) + P(Y = 3) + P(Y = 4) = 0.6$

또는, $P(2 \le Y \le 4) = 1 - P(Y = 1) = 1 - 0.4 = 0.6$.

예제 6. 연속확률변수 X의 확률밀도함수가 다음과 같을 때 물음에 답하여라.

$$f_X(x) = \frac{c(1-x^3)}{3} \quad (0 < x < 1)$$

(a) 상수 c의 값을 구하여라.

(b) $P(\frac{1}{2} < X \leq 1)$을 구하여라.

(c) $\mathrm{E}(X)$, $\mathrm{Var}(X)$을 각각 구하여라.

Solution.

(a) 확률밀도함수의 정의로부터

$$\begin{aligned}\int_0^1 f_X(x)dx &= \int_0^1 \frac{c(1-x^3)}{3}dx \\ &= \frac{c}{3}\Big[x - \frac{1}{4}x^4\Big]_0^1 = \frac{c}{3}\Big(1 - \frac{1}{4}\Big) = \frac{c}{4} = 1\end{aligned}$$

이므로 $c = 4$이다.

(b) (a)에서 $c = 4$이므로, $f_X(x) = \dfrac{4(1-x^3)}{3}$, $\quad (0 < x < 1)$이다.

따라서, $P\Big(\dfrac{1}{2} < X \leq 1\Big) = \displaystyle\int_{1/2}^1 \frac{4(1-x^3)}{3}\,dx = \frac{4}{3}\left[x - \frac{x^4}{4}\right]_{1/2}^1$.

이 때, $\left[x - \dfrac{x^4}{4}\right]_{1/2}^1 = \Big(1 - \dfrac{1}{4}\Big) - \Big(\dfrac{1}{2} - \dfrac{(1/2)^4}{4}\Big) = \dfrac{3}{4} - \Big(\dfrac{1}{2} - \dfrac{1}{64}\Big) = \dfrac{17}{64}$

따라서, $P\Big(\dfrac{1}{2} < X \leq 1\Big) = \dfrac{4}{3} \cdot \dfrac{17}{64} = \dfrac{17}{48}$.

(c) $\mathrm{E}[X]$와 $\mathrm{Var}(X)$는 다음과 같다.

$$\begin{aligned}\mathrm{E}[X] &= \int_0^1 x f_X(x)dx \\ &= \int_0^1 \frac{4x(1-x^3)}{3}dx \\ &= \frac{4}{3}\Big[\frac{1}{2}x^2 - \frac{1}{5}x^5\Big]_0^1 = \frac{4}{3}\Big(\frac{1}{2} - \frac{1}{5}\Big) = \frac{2}{5} \\ \mathrm{E}[X^2] &= \int_0^1 x^2 f_X(x)dx \\ &= \int_0^1 \frac{4x^2(1-x^3)}{3}dx \\ &= \frac{4}{3}\Big[\frac{1}{3}x^3 - \frac{1}{6}x^6\Big]_0^1 = \frac{4}{3}\Big(\frac{1}{3} - \frac{1}{6}\Big) = \frac{2}{9} \\ \mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \frac{14}{225}\end{aligned}$$

2 다양한 확률분포

2.1 확률분포의 예

예제 7. 한 음악 스트리밍 서비스에서는 사용자에게 노래를 추천하고, 사용자의 반응을 바탕으로 추천 시스템을 개선한다. 노래가 추천된 후 일정 시간 이내에 사용자가 아무런 긍정적인 행동(저장, 재생 지속 등)을 보이지 않으면, 해당 추천은 '관심 없음'으로 기록된다.

로그 분석 결과, 임의의 사용자에 대해 관심 없음 반응이 발생할 확률은 0.3이며, 개별 사용자들의 반응은 서로 독립이라고 가정한다. 임의로 선택한 사용자 4명에게 동일한 노래를 추천할 때, 다음 물음에 답하여라.

(a) 한 사용자의 관심 없음 반응 여부를 나타내는 확률변수 X를 다음과 같이 정의할 때, X의 확률분포를 명시하여라.

$$X = \begin{cases} 1, & \text{관심 없음 반응이 발생한 경우} \\ 0, & \text{관심 없음 반응이 발생하지 않은 경우} \end{cases}$$

(b) 4명의 사용자 중 각 사용자의 관심 없음 반응 여부를 나타내는 확률변수를 X_1, X_2, X_3, X_4라고 하자. 관심 없음 반응이 나타난 사용자의 수를 확률변수 Y라 할 때

- X_1, X_2, X_3, X_4와 Y 사이의 관계식을 바탕으로 Y의 확률 분포를 명시하고,
- $P(Y \geq 2)$를 계산하여라.

Solution.

(a) 확률변수 X는 한 사용자의 관심 없음 반응 여부를 나타내며, 관심 없음 반응이 발생할 확률은 0.3이다. 따라서 X의 확률질량함수는

$$P(X = 1) = 0.3, \qquad P(X = 0) = 0.7$$

이고, X는 성공 확률이 0.3인 베르누이분포를 따른다.

즉, $X \sim \text{Bernoulli}(0.3)$.

(b) 각 사용자의 관심 없음 반응 여부를 나타내는 확률변수 X_1, X_2, X_3, X_4는 서로 독립이며 모두 Bernoulli(0.3)을 따른다.

관심 없음 반응이 나타난 사용자의 수 Y는 $Y = X_1 + X_2 + X_3 + X_4$로 표현되며, 이는 서로 독립인 베르누이 확률변수의 합이므로 $Y \sim \text{Binomial}(4, 0.3)$ 이다.

따라서, $P(Y \geq 2) = 1 - \{P(Y = 0) + P(Y = 1)\}$.

$P(Y = 0),\ P(Y = 1)$을 계산하면,

$$P(Y = 0) = \binom{4}{0}(0.3)^0(0.7)^4 = (0.7)^4 = 0.2401,$$

$$P(Y=1) = \binom{4}{1}(0.3)(0.7)^3 = 4 \times 0.3 \times 0.343 = 0.4116.$$

따라서, $P(Y \geq 2) = 1 - (0.2401 + 0.4116) = 0.3483.$

예제 8. 확률변수 X의 확률밀도함수가 다음과 같이 주어져 있다.

$$f_X(x) = \begin{cases} c\,x^2(1-x), & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

다음 물음에 답하여라.

(a) $f_X(x)$가 베타분포의 확률밀도함수 형태임을 이용하여 X의 분포를 $\text{Beta}(\alpha, \beta)$로 가정할 때, (α, β)를 구하여라.
(참고: 베타분포의 확률밀도함수)

$$X \sim \text{Beta}(\alpha, \beta) \iff f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1$$

(b) (a)에서 구한 분포의 형태를 이용하여 상수 c의 값을 구하여라.
(참고: 감마함수 계산공식)

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \quad \Gamma(n) = (n-1)! \ (n\text{은 자연수})$$

Solution.

(a) $f_X(x) = cx^2(1-x)$ $(0 \le x \le 1)$ 이고,

$$x^2(1-x) = x^{3-1}(1-x)^{2-1}$$

이므로 $f_X(x)$는 Beta 분포의 pdf 형태

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1$$

와 비교하여 $\alpha = 3$, $\beta = 2$에 대응한다. 따라서

$$X \sim \text{Beta}(3, 2)$$

로 가정할 수 있다.

(b) 정규화 조건 $\int_0^1 f_X(x)\,dx = 1$을 이용하면

$$1 = \int_0^1 c\,x^2(1-x)\,dx = c\int_0^1 x^{3-1}(1-x)^{2-1}\,dx = c \cdot B(3,2).$$

Beta 함수 $B(\alpha, \beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$이므로

$$B(3,2) = \frac{\Gamma(3)\Gamma(2)}{\Gamma(5)} = \frac{2!\,1!}{4!} = \frac{2}{24} = \frac{1}{12}.$$

따라서

$$c \cdot \frac{1}{12} = 1 \quad \Rightarrow \quad c = 12.$$

예제 9. 확률변수 X가 평균이 10, 분산이 4인 정규분포를 따른다고 하자. 즉,

$$X \sim N(10, 4).$$

다음과 같이 정의된 확률변수 Y를 고려하자.

$$Y = 3X - 5.$$

(a) 확률변수 Y의 분포를 구하여라.

(b) 표준정규분포표를 이용하여 $P(19 \le Y \le 31)$을 계산하여라.

z	0.00	0.01	0.02	0.03
0.8	0.7881	0.7910	0.7939	0.7967
0.9	0.8159	0.8186	0.8212	0.8238
1.0	0.8413	0.8438	0.8461	0.8485

$P(Z \le z)$ 표준정규분포표 일부

Solution.

$X \sim N(\mu, \sigma^2)$일 때, 상수 $a \neq 0$, b에 대해 다음이 성립한다.

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

(a) 주어진 확률변수 $Y = 3X - 5$에 대해

$$Y \sim N(3 \times 10 - 5,\ 3^2 \times 4) = N(25, 36).$$

(b) (a)에서 $Y \sim N(25, 36)$이므로 평균은 25, 표준편차는 6이다.
확률변수 $Z = \dfrac{Y - 25}{6}$로 표준화하면 $Z \sim N(0, 1)$이다. 따라서

$$P(19 \le Y \le 31) = P\Big(\frac{19 - 25}{6} \le Z \le \frac{31 - 25}{6}\Big) = P(-1 \le Z \le 1).$$

주어진 표준정규분포표로부터

$$P(Z \le 1) = 0.8413, \qquad P(Z \le -1) = 1 - 0.8413 = 0.1587$$

이므로

$$P(19 \le Y \le 31) = 0.8413 - 0.1587 = 0.6826.$$