

LGE - SNU AI Scientist 고급과정

확률통계 및 통계 방법론

홍현경(hyungyeong81@snu.ac.kr)

2026년 01월

Contents

| | |
|----------------------|-----------|
| 1 확률과 확률분포 | 2 |
| 1.1 확률 | 2 |
| 1.2 확률변수와 확률분포 | 6 |
| 2 다양한 확률분포 | 8 |
| 2.1 확률분포의 예 | 8 |
| 2.2 다변량 확률변수와 확률분포 | 13 |
| 3 표본분포 | 17 |
| 3.1 표본분포의 성질 | 17 |
| 4 데이터 기반 추론 1 | 21 |
| 4.1 통계적 추정 | 21 |
| 4.2 통계적 추정 방법과 알고리즘 | 26 |
| 5 데이터 기반 의사결정 | 30 |
| 5.1 유의성 검정의 개념 | 30 |

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.

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

다음 물음에 답하여라.

- 해당 사용자가 알림을 클릭하고, 동시에 결제까지 했을 확률을 구하여라.
- 해당 사용자가 알림을 클릭한 사건과 결제를 한 사건이 서로 독립인지 판단하여라.

Solution.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ 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.

- 두 사건 A와 B가 서로 독립이라면 $P(A \cap B) = P(A)P(B)$ 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.

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

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

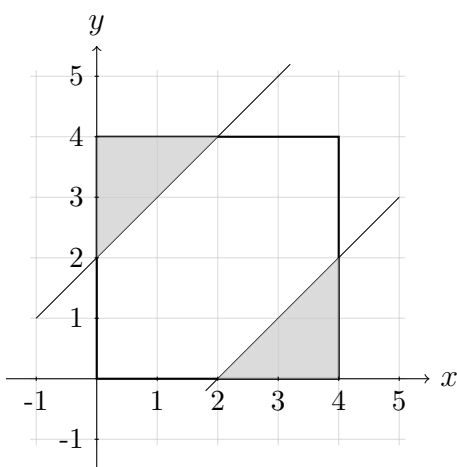
서버 2와 3이 둘 다 고장날 확률은 $(0.02)^2$ 이므로, 서버 2와 3 중 적어도 한 대가 정상일 확률은 $1 - (0.02)^2$ 이다.

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

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

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 생각해 보자. 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 $P(A) = 0.40$, $P(B | A) = 0.90$, $P(B | A^c) = 0.20$ 이다.
또한 $P(A^c) = 0.60$ 이다.

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

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, $P(A | B) = \frac{P(A \cap B)}{P(B)}$ 이다.

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

(a)에서 구한 $P(B)$ 를 대입하면,

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

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 X 라고 할 때, $E(X)$ 를 구하여라.
- (b) 선택된 두 수 중 작은 값을 확률변수 Y 라고 할 때, $P(2 \leq Y \leq 4)$ 을 구하여라.

Solution.

- (a) X 의 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) Y 의 확률분포는 다음과 같다.

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

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

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

예제 6. 연속확률변수 X 의 확률밀도함수가 다음과 같을 때 물음에 답하여라.

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

- (a) 상수 c 의 값을 구하여라.
 (b) $P(\frac{1}{2} < X \leq 1)$ 을 구하여라.
 (c) $E(X)$, $\text{Var}(X)$ 을 각각 구하여라.

Solution.

- (a) 확률밀도함수의 정의로부터

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

이므로 $c = 4$ 이다.

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

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

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

$$\text{따라서, } P\left(\frac{1}{2} < X \leq 1\right) = \frac{4}{3} \cdot \frac{17}{64} = \frac{17}{48}.$$

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

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

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 사용자에게 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 X 를 다음과 같이 정의할 때, X 의 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_4 라고 하자. 관심 없음 반응이 나타난 사용자의 수를 확률변수 Y 라 할 때
- X_1, X_2, X_3, X_4 와 Y 사이의 관계식을 바탕으로 Y 의 확률 분포를 명시하고,
 - $P(Y \geq 2)$ 를 계산하여라.

Solution.

- (a) 확률변수 X 는 한 사용자의 관심 없음 반응 여부를 나타내며, 관심 없음 반응이 발생할 확률은 0.3이다. 따라서 X 의 확률질량함수는

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

이고, X 는 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_4 는 서로 독립이며 모두 $\text{Bernoulli}(0.3)$ 을 따른다.

관심 없음 반응이 나타난 사용자의 수 Y 는 $Y = X_1 + X_2 + X_3 + X_4$ 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 X 의 확률밀도함수가 다음과 같이 주어져 있다.

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

다음 물음에 답하여라.

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

(참고: 베타분포의 확률밀도함수)

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

- (b) (a)에서 구한 분포의 형태를 이용하여 상수 c 의 값을 구하여라.

(참고: 감마함수 계산공식)

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

Solution.

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

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

이므로 $f_X(x)$ 는 Beta 분포의 pdf 형태

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

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

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

Beta 함수 $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ 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 X 가 평균이 10, 분산이 4인 정규분포를 따른다고 하자. 즉,

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

다음과 같이 정의된 확률변수 Y 를 고려하자.

$$Y = 3X - 5.$$

(a) 확률변수 Y 의 분포를 구하여라.

(b) 표준정규분포표를 이용하여 $P(19 \leq Y \leq 31)$ 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 |

$P(Z \leq z)$ 표준정규분포표 일부

Solution.

$X \sim N(\mu, \sigma^2)$ 일 때, 상수 $a \neq 0$, b 에 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 $Z = \frac{Y - 25}{6}$ 로 표준화하면 $Z \sim N(0, 1)$ 이다. 따라서

$$P(19 \leq Y \leq 31) = P\left(\frac{19 - 25}{6} \leq Z \leq \frac{31 - 25}{6}\right) = P(-1 \leq Z \leq 1).$$

주어진 표준정규분포표로부터

$$P(Z \leq 1) = 0.8413, \quad P(Z \leq -1) = 1 - 0.8413 = 0.1587$$

이므로

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

2.2 다변량 확률변수와 확률분포

예제 10. 한 온라인 플랫폼에서 사용자의 유료 구독 여부를 X , 모바일 앱 사용 여부를 Y 라고 하자.

$$X = \begin{cases} 1, & \text{유료 구독자} \\ 0, & \text{비구독자} \end{cases}, \quad Y = \begin{cases} 1, & \text{모바일 앱 사용} \\ 0, & \text{사용하지 않음} \end{cases}$$

조사를 통해 얻은 (X, Y) 의 결합확률질량함수는 다음과 같다.

| | $Y = 0$ | $Y = 1$ |
|---------|---------|---------|
| $X = 0$ | 0.30 | 0.10 |
| $X = 1$ | 0.20 | 0.40 |

- (a) 모바일 앱을 사용하는 사용자가 유료 구독자가 아닐 확률을 구하여라.
 (b) $E(X)$, $E(Y)$, $\text{Cov}(X, Y)$ 를 구하여라.
 (c) $U = 2X + 1$, $V = 3Y - 2$ 라 할 때, $\text{Cov}(U, V)$ 를 구하여라.

Solution.

- (a) 제시된 문장은 사건 $\{Y = 1\}$ 이면서 $\{X = 0\}$ 인 확률을 의미하므로

$$P(\text{모바일 앱 사용, 비구독자}) = P(X = 0, Y = 1) = 0.10.$$

- (b) 먼저 X, Y 의 주변확률을 표에서 직접 구하면

$$P(X = 0) = 0.30 + 0.10 = 0.40, \quad P(X = 1) = 0.20 + 0.40 = 0.60,$$

$$P(Y = 0) = 0.30 + 0.20 = 0.50, \quad P(Y = 1) = 0.10 + 0.40 = 0.50.$$

따라서

$$E(X) = 0 \cdot 0.40 + 1 \cdot 0.60 = 0.60, \quad E(Y) = 0 \cdot 0.50 + 1 \cdot 0.50 = 0.50.$$

또한

$$E(XY) = \sum_{x=0}^1 \sum_{y=0}^1 xy P(X = x, Y = y) = 1 \cdot 1 \cdot P(X = 1, Y = 1) = 0.40.$$

따라서

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0.40 - (0.60)(0.50) = 0.10.$$

- (c) $U = 2X + 1$, $V = 3Y - 2$ 이므로 공분산의 성질에 의해

$$\text{Cov}(U, V) = \text{Cov}(2X + 1, 3Y - 2) = 2 \cdot 3 \text{Cov}(X, Y) = 6 \cdot 0.10 = 0.60.$$

예제 11. 두 개의 센서 A, B가 동일한 물체의 온도($^{\circ}\text{C}$)를 측정한다고 하자. 센서 A의 측정값을 확률변수 X_1 , 센서 B의 측정값을 확률변수 X_2 라고 하자. $X_1 \sim N(25, 4^2)$, $X_2 \sim N(25, 3^2)$ 이고 X_1 과 X_2 는 서로 독립이라고 할 때, 두 센서의 측정값 차이가 4°C 보다 클 확률을 다음 표를 이용하여 구하여라. (단, $Z \sim N(0, 1)$)

| z | $P(0 \leq Z \leq z)$ |
|-----|----------------------|
| 0.2 | 0.079 |
| 0.4 | 0.155 |
| 0.6 | 0.226 |
| 0.8 | 0.288 |
| 1.0 | 0.341 |
| 1.2 | 0.385 |

Solution.

X_1 과 X_2 가 독립이므로

$$X_1 - X_2 \sim N(25 - 25, 4^2 + 3^2) = N(0, 25).$$

따라서 $D = X_1 - X_2$ 라 두면

$$D \sim N(0, 5^2).$$

구하고자 하는 확률은

$$P(|X_1 - X_2| > 4) = P(|D| > 4) = P(D < -4) + P(D > 4).$$

표준화하면

$$P(|D| > 4) = P\left(Z < -\frac{4}{5}\right) + P\left(Z > \frac{4}{5}\right) = 2 \times P(Z > 0.8).$$

주어진 표가 $P(0 \leq Z \leq z)$ 형태이므로

$$P(Z > 0.8) = 1 - P(Z \leq 0.8) = 1 - (0.5 + P(0 \leq Z \leq 0.8)) = 0.5 - P(0 \leq Z \leq 0.8).$$

표에서 $P(0 \leq Z \leq 0.8) = 0.288$ 이므로

$$P(|D| > 4) = 2 \times (0.5 - 0.288) = 0.424.$$

예제 12. KBO리그에서 투구 구속은 측정 방식에 따라 차이가 발생할 수 있다. 구속을 측정하기 위해서는 보통 **카메라 기반의 투구 추적 시스템(PTS)**과 **레이더 기반의 트랙맨**이 사용되며, 트랙맨은 도플러 레이더를 이용해 공의 궤적을 비행 전 구간에서 걸쳐 추적하여 투수 손을 떠난 지점에 가까운 구속을 측정한다. 반면, PTS는 여러 대의 카메라로 공의 이동 구간 일부를 추적하여 구속을 산출하므로, 트랙맨에 비해 상대적으로 낮은 구속이 기록되는 경우가 있다. 이러한 측정 결과는 경기 중 문자 중계에 표출되는 공식 구속 기록으로 사용된다.

한 투구의 실제 구속을 $v(\text{km/h})$ 라고 하자. 여기서 v 는 오차가 없는 값이라고 가정한다. 문자 중계에 표출되는 구속은 측정 과정에서 발생하는 차이를 포함하여 다음과 같이 나타난다고 가정한다.

$$\text{표출 구속} = v + X$$

여기서 X 는 실제 구속과 표출 구속 사이의 **속력 차이(편차)**를 나타내는 **확률변수**로, $X < 0$ 은 표출 구속이 실제 구속보다 느리게 기록된 경우를 의미한다.

구장 및 방송 환경의 차이로 인해, 문자 중계에 사용되는 구속 측정 방식은 다음 두 유형으로 구분된다고 가정한다. 한 경기에서는 하나의 측정 방식이 일관되게 사용되며, 경기 단위로 볼 때 **트랙맨 기준 방식이 사용될 확률은 0.7, 카메라 기준 방식이 사용될 확률은 0.3**이다.

- (유형 A: 카메라 기반 측정) 카메라 영상을 이용한 추정 방식으로, 표출 구속이 실제 구속보다 더 낮게 기록되는 경향이 있다. 이 경우 측정 오차 X 는

$$X \sim N(-4, 1^2).$$

- (유형 B: 트랙맨 기준 측정) 보정 절차가 적용된 측정 방식으로, 트랙맨을 이용한 측정 방식으로, 표출 구속과 실제 구속의 차이가 상대적으로 작다. 이 경우 측정 오차 X 는

$$X \sim N(-1, 1^2).$$

이 때, 문자 중계 구속이 실제보다 3km/h 이상 느리게 표출될 확률, 즉 $P(X \leq -3)$ 을 구하여라. (단, $\alpha = P(Z \leq z_\alpha)$, $Z \sim N(0, 1)$ 에 대해 $z_{0.999} = 3$, $z_{0.977} = 2$, $z_{0.933} = 1.5$, $z_{0.841} = 1$, $z_{0.691} = 0.5$ 임을 이용하여라.)

Solution.

전확률공식을 이용하면

$$P(X \leq -3) = 0.3 P(X \leq -3 | A) + 0.7 P(X \leq -3 | B).$$

유형 A에서는 $X \sim N(-4, 1^2)$ 이므로

$$P(X \leq -3 | A) = P\left(\frac{X + 4}{1} \leq \frac{-3 + 4}{1}\right) = P(Z \leq 1).$$

주어진 값 $z_{0.841} = 1$ 을 이용하면 $P(Z \leq 1) = 0.841$ 이므로

$$P(X \leq -3 \mid A) = 0.841.$$

유형 B에서는 $X \sim N(-1, 1^2)$ 이므로

$$P(X \leq -3 \mid B) = P\left(\frac{X+1}{1} \leq \frac{-3+1}{1}\right) = P(Z \leq -2).$$

주어진 값 $z_{0.977} = 2$ 에서 $P(Z \leq 2) = 0.977$ 이므로 대칭성을 이용하면

$$P(Z \leq -2) = 1 - P(Z \leq 2) = 1 - 0.977 = 0.023.$$

따라서

$$P(X \leq -3) = 0.3 \times 0.841 + 0.7 \times 0.023 = 0.2523 + 0.0161 = 0.2684.$$

(추가 코멘트)

위 계산은 조건부 확률분포와 전확률 공식을 이용한 풀이이며, 이러한 상황은 혼합 분포에서 확률을 구한 것으로 생각할 수도 있다. 유형 A(카메라 기준)일 확률이 0.3, 유형 B(트랙맨 기준)일 확률이 0.7이므로, X 의 확률밀도함수는 가우시안 혼합분포 형태로 다음과 같이 나타낼 수 있다.

$$f_X(x) = 0.3 \cdot \frac{1}{\sqrt{2\pi} \cdot 1^2} \exp\left(-\frac{(x+4)^2}{2 \cdot 1^2}\right) + 0.7 \cdot \frac{1}{\sqrt{2\pi} \cdot 1^2} \exp\left(-\frac{(x+1)^2}{2 \cdot 1^2}\right), \quad -\infty < x < \infty$$

3 표본분포

3.1 표본분포의 성질

예제 13. 어떤 공장에서 생산되는 부품의 길이는 평균 50mm, 표준편차 0.05mm의 분포를 따르고, 각 부품의 생산은 서로 독립적으로 이루어진다고 한다. 생산된 제품 중 무작위로 100개의 부품을 추출했을 때, 추출된 부품의 평균 길이가 $[50\text{mm}-0.01\text{mm}, 50\text{mm}+0.01\text{mm}]$ 범위에 들어오면 해당 공정 상태를 ‘우수’로 평가한다. 공정 상태가 ‘우수’로 평가될 확률을 아래의 표를 참고하여 구하여라. (단, Z 는 표준정규분포를 따르는 확률변수이다.)

| z | $P(0 \leq Z \leq z)$ |
|-----|----------------------|
| 1.0 | 0.341 |
| 1.5 | 0.433 |
| 2.0 | 0.477 |
| 2.5 | 0.494 |
| 3.0 | 0.499 |

Solution.

추출된 부품의 길이를 X_1, \dots, X_{100} 라고 하면,

$$E(X_i) = 50, \quad \text{Var}(X_i) = 0.05^2 \quad (i = 1, \dots, 100).$$

표본의 크기가 크므로 중심극한정리에 의해 $\frac{\bar{X}-50}{0.005} \sim N(0, 1)$.

따라서, 표준정규분포를 따르는 확률변수 Z 에 대하여

$$P(50 - 0.01 \leq \bar{X} \leq 50 + 0.01) \approx P(-2 \leq Z \leq 2) = 2 \times 0.477 = 0.954$$

예제 14. 한 카페의 에스프레소 한 샷 추출량(ml)을 확률변수 X 라 하자. 추출량은 평균 30ml, 표준편차 3ml인 분포를 따른다고 가정하며, 각 샷의 추출은 서로 독립적으로 이루어진다. 무작위로 100샷을 추출하여 추출량을 측정했을 때, 추출량의 표본평균 \bar{X} 가 $[30\text{ml} - 0.3\text{ml}, 30\text{ml} + 0.3\text{ml}]$ 범위에 들어오면 해당 날의 추출 상태를 정상으로 평가한다. 추출 상태가 정상으로 평가될 확률을 아래의 표준정규분포표를 참고하여 구하여라.

| z | $P(0 \leq Z \leq z)$ |
|-----|----------------------|
| 1.0 | 0.341 |
| 1.5 | 0.433 |
| 2.0 | 0.477 |
| 2.5 | 0.494 |
| 3.0 | 0.499 |

Solution.

에스프레소 한 샷의 추출량을 X_1, \dots, X_{100} 이라고 하면,

$$E(X_i) = 30, \quad \text{Var}(X_i) = 3^2 \quad (i = 1, \dots, 100).$$

표본의 크기가 크므로 중심극한정리에 의해

$$\frac{\bar{X} - 30}{0.3} \sim N(0, 1).$$

따라서, 표준정규분포를 따르는 확률변수 Z 에 대하여

$$P(30 - 0.3 \leq \bar{X} \leq 30 + 0.3) \approx P(-1 \leq Z \leq 1) = 2 \times 0.341 = 0.682.$$

예제 15. 음악 스트리밍 서비스에서 임의로 한 명의 사용자를 선택했을 때, 그 사용자가 시스템이 추천한 곡을 끝까지 재생하면 1, 중간에 종료하면 0으로 기록한다고 하자. 이를 확률변수 X 로 두면, 확률변수 X 는 다음과 같은 분포를 가진다.

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

서로 다른 $n = 100$ 명의 사용자를 무작위로 선택하여 얻은 결과를 X_1, \dots, X_{100} 이라 하자. 다음 물음에 답하여라.

- (a) 새로운 확률변수 Y 를 추천한 곡을 끝까지 재생한 사용자의 수로 정의하려고 한다. 이 때, Y 를 X_1, \dots, X_{100} 을 이용하여 나타내고, 이를 바탕으로 Y 의 분포를 명시하여라.
- (b) 추천한 곡을 끝까지 재생한 사용자의 수가 75명 이상일 확률을 근사적으로 구하여라. (단, $Z \sim N(0, 1)$ 에서, $P(Z \leq 1.091) = 0.862$ 임을 이용하여라.)

Solution.

- (a) 추천한 곡을 끝까지 재생한 사용자의 수를 $Y = \sum_{i=1}^{100} X_i$ 로 정의한다.

X_i 들이 서로 독립인 베르누이 확률변수이므로, Y 는 성공확률이 0.7인 이항분포를 따른다. 즉,

$$Y \sim \text{Bin}(100, 0.7),$$

이며 확률질량함수는

$$P(Y = y) = \binom{100}{y} (0.7)^y (0.3)^{100-y}, \quad y = 0, 1, \dots, 100.$$

- (b) 표본평균 $\bar{X} = \frac{1}{100} \sum_{i=1}^{100} X_i$ 를 생각하면

$$\bar{X} = \frac{Y}{100} \iff Y \geq 75 \iff \bar{X} \geq 0.75.$$

따라서

$$P(Y \geq 75) = P(\bar{X} \geq 0.75).$$

이때,

$$\mu = E(X) = 0.7, \quad \sigma^2 = \text{Var}(X) = 0.7 \cdot 0.3 = 0.21$$

이므로

$$E(\bar{X}) = \mu = 0.7, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{100} = \frac{0.21}{100} = 0.0021.$$

중심극한정리에 의해

$$\frac{\bar{X} - 0.7}{\sqrt{0.0021}} \sim N(0, 1).$$

따라서

$$\begin{aligned} P(Y \geq 75) &= P(\bar{X} \geq 0.75) \\ &\approx P\left(Z \geq \frac{0.75 - 0.7}{\sqrt{0.0021}}\right) \\ &= P\left(Z \geq \frac{0.05}{\sqrt{0.0021}}\right). \end{aligned}$$

여기서, $\frac{0.05}{\sqrt{0.0021}} \approx \frac{0.05}{0.04583} \approx 1.091$. 따라서

$$P(Y \geq 75) \approx 1 - P(Z \leq 1.091) = 0.138.$$

4 데이터 기반 추론 1

4.1 통계적 추정

예제 16. 확률통계 및 통계방법론 과목을 수강한 학생들의 시험 성적을 분석하고자 한다. 이 시험은 20점 만점이며, 수강생 34명의 성적을 하나의 모집단으로 간주한다. 이 모집단의 모평균 성적은 17점이라고 알려져 있다. 이 모집단에서 비복원추출을 이용하여 크기가 5인 랜덤 표본을 선택할 때, 모평균에 대한 두 가지 추정량인 표본평균 \bar{X} 와 표본중앙값 X_{med} 의 확률분포가 다음과 같이 주어져 있다. 아래의 물음에 답하시오.

| | | | | | |
|--------------|------|------|------|------|------|
| \bar{X} | 15 | 16 | 17 | 18 | 19 |
| $P(\bar{X})$ | 0.10 | 0.20 | 0.40 | 0.20 | 0.10 |

| | | | | | |
|---------------------|------|------|------|------|------|
| X_{med} | 15 | 16 | 17 | 18 | 19 |
| $P(X_{\text{med}})$ | 0.25 | 0.15 | 0.20 | 0.15 | 0.25 |

- 표본평균과 표본중앙값의 기댓값을 구하고, 두 추정량이 비편향추정량(편향 = 0)인지 여부를 확인하여라.
- 표본평균과 표본중앙값의 분산을 구하여라.
- 앞의 결과를 바탕으로, 표본평균과 표본중앙값 중 어떤 추정량이 더 좋은 추정량인지에 대해 논하여라.

Solution.

- (a) 기댓값 계산:

$$E(\bar{X}) = 15 \cdot 0.10 + 16 \cdot 0.20 + 17 \cdot 0.40 + 18 \cdot 0.20 + 19 \cdot 0.10 = 17 = \mu,$$

$$E(X_{\text{med}}) = 15 \cdot 0.25 + 16 \cdot 0.15 + 17 \cdot 0.20 + 18 \cdot 0.15 + 19 \cdot 0.25 = 17 = \mu.$$

따라서, $E(\bar{X}) = \mu$ 이고 $E(X_{\text{med}}) = \mu$ 이므로 \bar{X} 와 X_{med} 모두 비편향 추정량이다.

- (b) 분산 계산:

$$E(\bar{X}^2) = 15^2 \cdot 0.10 + 16^2 \cdot 0.20 + 17^2 \cdot 0.40 + 18^2 \cdot 0.20 + 19^2 \cdot 0.10 = 290.2,$$

$$\text{Var}(\bar{X}) = E(\bar{X}^2) - [E(\bar{X})]^2 = 290.2 - 17^2 = 290.2 - 289 = 1.2.$$

$$\begin{aligned}E(X_{\text{med}}^2) &= 15^2 \cdot 0.25 + 16^2 \cdot 0.15 + 17^2 \cdot 0.20 + 18^2 \cdot 0.15 + 19^2 \cdot 0.25 = 291.3, \\ \text{Var}(X_{\text{med}}) &= E(X_{\text{med}}^2) - [E(X_{\text{med}})]^2 = 291.3 - 17^2 = 291.3 - 289 = 2.3.\end{aligned}$$

(c) 결론:

\bar{X} 와 X_{med} 모두 비편향 추정량이지만, $\text{Var}(\bar{X}) = 1.2$ 이고 $\text{Var}(X_{\text{med}}) = 2.3$ 이므로 표본평균 \bar{X} 의 분산이 더 작다. 따라서 \bar{X} 가 X_{med} 보다 더 좋은 추정량이다.

예제 17. 어떤 전자부품의 수명(시간)은 정규분포를 따르며, 모표준편차는 $\sigma = 100$ 시간으로 알려져 있다고 하자. 이 전자부품의 평균 수명을 μ (시간)라 하자. 무작위로 $n = 30$ 개의 전자부품을 선택하여 수명을 측정한 결과, 수명에 대한 표본평균이 2000시간으로 나타났다. 다음 물음에 답하여라. (단, 표준정규분포를 따르는 확률변수 Z 에 대하여 $z_{1-\alpha}$ 는 $P(Z \geq z_{1-\alpha}) = \alpha$ 를 만족한다고 할 때, $z_{0.995} = 2.58$, $z_{0.99} = 2.33$, $z_{0.975} = 1.96$, $z_{0.95} = 1.65$ 임을 이용하시오.)

(a) 모평균 μ 에 대한 95% 신뢰구간을 구하여라.

(b) 모평균 μ 에 대한 99% 신뢰구간의 길이가 (a)에서 구한 95% 신뢰구간의 길이보다 작아지기 위해 필요한 표본의 크기 n 의 최소값을 구하여라.

Solution.

X_1, \dots, X_n 이 정규모집단 $N(\mu, 100^2)$ 에서 추출된 랜덤포본이고, $n = 30$, $\bar{X} = 2000$ 이라 하자. 모표준편차 $\sigma = 100$ 이 알려져 있으므로

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mu}{100/\sqrt{30}} \sim N(0, 1).$$

(a) $P(-z_{0.975} \leq Z \leq z_{0.975}) = 0.95$ 를 이용하면, 95% 신뢰구간은

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{100/\sqrt{30}} \leq 1.96\right) = 0.95$$

이므로, 이를 μ 에 대해 정리하면

$$\mu \in \left(\bar{X} - 1.96 \cdot \frac{100}{\sqrt{30}}, \bar{X} + 1.96 \cdot \frac{100}{\sqrt{30}}\right).$$

$\bar{X} = 2000$ 을 대입하면

$$\mu \in \left(2000 - 1.96 \cdot \frac{100}{\sqrt{30}}, 2000 + 1.96 \cdot \frac{100}{\sqrt{30}}\right) \approx (1964.2, 2035.8).$$

(b) 표본크기가 n 일 때, 모평균 μ 에 대한 양측 $100(1 - \alpha)\%$ 신뢰구간의 길이는

$$2 \times z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

이다. (a)에서의 95% 신뢰구간 길이는

$$2 \times z_{0.975} \frac{100}{\sqrt{30}} = 2 \times 1.96 \times \frac{100}{\sqrt{30}}.$$

99% 신뢰구간의 길이는

$$2 \times z_{0.995} \frac{100}{\sqrt{n}} = 2 \times 2.58 \times \frac{100}{\sqrt{n}}.$$

이 때,

$$\begin{aligned}2 \times 2.58 \times \frac{100}{\sqrt{n}} &< 2 \times 1.96 \times \frac{100}{\sqrt{30}} \\ \frac{2.58}{\sqrt{n}} &< \frac{1.96}{\sqrt{30}} \\ \sqrt{n} &> \frac{2.58}{1.96} \times \sqrt{30}.\end{aligned}$$

따라서,

$$n > 30 \times \left(\frac{2.58}{1.96}\right)^2 \approx 51.981,$$

필요한 표본 크기의 최소값은 52이다.

예제 18. 한 식품회사는 새로 출시한 과자의 평균 중량(g) μ 가 규격을 만족하는지 확인하려고 한다. 생산 공정이 안정적이라 과자 중량은 정규분포를 따른다고 가정하며, 공정 데이터로부터 모표준편차는 $\sigma = 3(\text{g})$ 으로 알려져 있다. 회사는 μ 에 대한 95% 신뢰구간을 구성할 때, 신뢰구간의 오차한계를 0.8g 이하로 하고자 한다. 이를 위해서는 표본이 최소 몇 개 있어야 하는지 구하여라. (단, 표준정규분포를 따르는 확률변수 Z 에 대하여 $z_{1-\alpha}$ 는 $P(Z \geq z_{1-\alpha}) = \alpha$ 를 만족한다고 할 때, $z_{0.995} = 2.58$, $z_{0.99} = 2.33$, $z_{0.975} = 1.96$, $z_{0.95} = 1.65$ 임을 이용하시오.)

Solution.

과자 중량이 정규분포를 따르고 모표준편차가 $\sigma = 3\text{g}$ 로 알려져 있으므로, 모평균 μ 에 대한 95% 신뢰구간은

$$\bar{X} \pm z_{0.975} \frac{\sigma}{\sqrt{n}}$$

이다. 이때 오차한계는

$$d = z_{0.975} \frac{\sigma}{\sqrt{n}}$$

이므로, $d \leq 0.8$ 이 되도록 하는 n 은

$$z_{0.975} \frac{\sigma}{\sqrt{n}} \leq 0.8 \iff \sqrt{n} \geq \frac{z_{0.975}\sigma}{0.8} \iff n \geq \left(\frac{z_{0.975}\sigma}{0.8} \right)^2$$

를 만족해야 한다. 주어진 값 $z_{0.975} = 1.96$, $\sigma = 3$ 을 대입하면

$$n \geq \left(\frac{1.96 \times 3}{0.8} \right)^2 = \left(\frac{5.88}{0.8} \right)^2 = (7.35)^2 = 54.0225.$$

표본의 크기는 자연수이므로 최소 표본크기는 $n = 55$ 이다.

4.2 통계적 추정 방법과 알고리즘

예제 19. 어떤 온라인 쇼핑 플랫폼에서는 하나의 제품에 대해 사용자가 남기는 리뷰 별점 유형을 세 가지로 분류한다. 각 반응은 다음과 같이 숫자 1, 2, 3으로 기록된다.

- 1: 낮은 수준의 만족도
- 2: 보통 수준의 만족도
- 3: 높은 수준의 만족도

한 사용자가 남기는 리뷰 별점을 확률변수 X 로 나타낼 때, 이 반응은 다음과 같은 분포를 따른다고 알려져 있다.

$$P(X = 1) = \theta^2, \quad P(X = 2) = 2\theta(1 - \theta), \quad P(X = 3) = (1 - \theta)^2, \quad (0 < \theta < 1).$$

이제 서로 독립인 사용자의 리뷰 별점 X_1, \dots, X_n 을 관측하였다. 관측 결과, 리뷰 별점이 1 또는 2인 경우의 총 횟수가 m 번이었고, 반응 값이 3인 경우는 $n - m$ 번 관측되었다고 하자. 이때 모수 θ 의 최대가능도추정량 $\hat{\theta}_{MLE}$ 을 m 과 n 을 이용하여 구하여라(2차 미분은 보이지 않아도 무방하다).

Hint (미분 공식 및 근의 공식)

- (로그 미분 공식)
양의 함수 $f(\theta) > 0$ 에 대하여

$$\frac{d}{d\theta} \log f(\theta) = \frac{f'(\theta)}{f(\theta)}.$$

- (유리함수 미분 공식)
미분 가능한 함수 $u(\theta), v(\theta)$ 에 대하여

$$\frac{d}{d\theta} \left(\frac{u(\theta)}{v(\theta)} \right) = \frac{u'(\theta)v(\theta) - u(\theta)v'(\theta)}{v(\theta)^2}, \quad v(\theta) \neq 0.$$

- (근의공식)
이차방정식 $a\theta^2 + b\theta + c = 0$ ($a \neq 0$)의 해는 $\theta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ 이다.

Solution.

값 1 또는 2가 관측될 확률은

$$P(X = 1 \text{ or } 2) = \theta^2 + 2\theta(1 - \theta) = 2\theta - \theta^2,$$

값 3이 관측될 확률은

$$P(X = 3) = (1 - \theta)^2$$

이다. 따라서 가능도함수는

$$L(\theta) = \{2\theta - \theta^2\}^m \{(1 - \theta)^2\}^{n-m}.$$

로그가능도함수는

$$\ell(\theta) = m \log(2\theta - \theta^2) + 2(n - m) \log(1 - \theta).$$

이를 θ 에 대해 미분하면

$$\ell'(\theta) = m \frac{2 - 2\theta}{2\theta - \theta^2} - \frac{2(n - m)}{1 - \theta}.$$

$\ell'(\theta) = 0$ 을 정리하면

$$m(1 - \theta)^2 - (n - m)\theta(2 - \theta) = 0 \iff n\theta^2 - 2n\theta + m = 0.$$

따라서

$$\theta^2 - 2\theta + \frac{m}{n} = 0$$

이고, $0 < \theta < 1$ 을 만족하는 해는

$$\hat{\theta}_{\text{MLE}} = 1 - \sqrt{1 - \frac{m}{n}}.$$

이제 2차 미분을 계산하면

$$\ell''(\theta) = m \left(-\frac{2}{2\theta - \theta^2} - \frac{(2 - 2\theta)^2}{(2\theta - \theta^2)^2} \right) - \frac{2(n - m)}{(1 - \theta)^2}.$$

$0 < \theta < 1$ 에서 각 항은 모두 음수이므로

$$\ell''(\theta) < 0 \quad (0 < \theta < 1).$$

따라서 위에서 구한 $\hat{\theta}_{\text{MLE}}$ 은 로그가능도함수를 최대화하는 최대가능도추정량이다.

예제 20. 엘리베이터 두 대가 설치된 어떤 건물에서 연구자는 엘리베이터를 이용하는 사람들의 버튼 사용 습관을 관찰한다. 무작위로 서로 다른 n 명을 표본으로 선택하여, 각 사람이 엘리베이터를 기다릴 때 버튼을 양쪽 모두 누르는지 여부를 기록하였다. i 번째 사람의 행동을 다음과 같이 정의한다.

- $X_i = 1$: 버튼을 양쪽 모두 누름
- $X_i = 0$: 버튼을 한쪽만 누르거나 누르지 않음

각 사람은 서로 독립적으로 표본추출되었으며, 한 사람이 두 버튼을 모두 누르는 행동은 다른 사람의 행동에 직접적인 영향을 주지 않는다고 가정한다. 한 사람이 버튼을 양쪽 모두 누를 확률을 p 라고 하면,

$$P(X_i = 1) = p, \quad P(X_i = 0) = 1 - p, \quad (0 < p < 1),$$

즉, $X_i \sim \text{Bernoulli}(p)$ 이다. n 명의 관측 결과가 x_1, \dots, x_n 으로 주어졌을 때, 모수 p 의 최대가능도추정량 \hat{p}_{MLE} 을 구하여라.

Hint (미분 공식)

- (로그 미분 공식)
양의 함수 $f(\theta) > 0$ 에 대하여

$$\frac{d}{d\theta} \log f(\theta) = \frac{f'(\theta)}{f(\theta)}.$$

- (유리함수 미분 공식)
미분 가능한 함수 $u(\theta), v(\theta)$ 에 대하여

$$\frac{d}{d\theta} \left(\frac{u(\theta)}{v(\theta)} \right) = \frac{u'(\theta)v(\theta) - u(\theta)v'(\theta)}{v(\theta)^2}, \quad v(\theta) \neq 0.$$

Solution.

관측값이 x_1, \dots, x_n 일 때 가능도함수는

$$L(p) = \prod_{i=1}^n P(X_i = x_i) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum_{i=1}^n x_i} (1-p)^{n - \sum_{i=1}^n x_i}.$$

로그가능도함수는

$$\ell(p) = \log L(p) = \left(\sum_{i=1}^n x_i \right) \log p + \left(n - \sum_{i=1}^n x_i \right) \log(1-p).$$

이를 p 에 대해 미분하면

$$\ell'(p) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p}.$$

최대가능도추정량은 $\ell'(p) = 0$ 을 만족하는 해이므로

$$\frac{\sum_{i=1}^n x_i}{p} = \frac{n - \sum_{i=1}^n x_i}{1 - p} \iff \left(\sum_{i=1}^n x_i\right)(1 - p) = \left(n - \sum_{i=1}^n x_i\right)p \iff \sum_{i=1}^n x_i = np.$$

따라서

$$\hat{p}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}.$$

또한

$$\ell''(p) = -\frac{\sum_{i=1}^n x_i}{p^2} - \frac{n - \sum_{i=1}^n x_i}{(1 - p)^2} < 0 \quad (0 < p < 1)$$

이므로 위의 해는 로그가능도함수를 최대화한다.

5 데이터 기반 의사결정

5.1 유의성 검정의 개념

예제 21. 한 온라인 구독 서비스에서는 하루 동안 접속한 사용자들 중 서비스 이용을 중단하는 사용자 수를 기록한다. 각 사용자는 서로 독립적으로 행동하며, 정상적인 서비스 상태에서는 한 사용자가 서비스 이용을 중단할 확률이 $p = 0.03$ 으로 알려져 있다.

하루에 총 200명의 사용자를 무작위로 관측하여, i 번째 사용자가 서비스를 중단하면 $X_i = 1$, 그렇지 않으면 $X_i = 0$ 이라 하자. 각 X_i ($i = 1, \dots, 200$)는 서로 독립이라고 가정한다. 이제 다음 가설을 검정하고자 한다.

$$H_0 : p = 0.03 \quad \text{vs.} \quad H_1 : p > 0.03.$$

기각역을 자연수 M 에 대하여

$$X_1 + \dots + X_{200} \geq M$$

의 형태로 정하고, 제1종 오류를 범할 확률을 0.05 이하로 하고자 한다. 다음 이항분포 누적확률표를 이용하여 자연수 M 의 최소값을 구하여라.

| k | 9 | 10 | 11 | 12 | 13 |
|---|-------|-------|-------|-------|-------|
| $P(S \leq k), S \sim \text{Bin}(200, 0.03)$ | 0.919 | 0.960 | 0.982 | 0.992 | 0.997 |

Solution.

$X_i \sim \text{Bernoulli}(p)$ ($i = 1, \dots, 200$) 이고 서로 독립이므로

$$S := X_1 + \dots + X_{200} \sim \text{Bin}(200, p).$$

귀무가설 H_0 하에서

$$S \sim \text{Bin}(200, 0.03).$$

제1종 오류 확률은

$$P(H_0 \text{ 기각} \mid H_0 \text{ 참}) = P(S \geq M \mid H_0).$$

문제의 조건에 따라 $P(S \geq M \mid H_0) \leq 0.05$ 를 만족해야 한다. 이는

$$P(S \geq M \mid H_0) = 1 - P(S \leq M - 1 \mid H_0) \leq 0.05 \iff P(S \leq M - 1 \mid H_0) \geq 0.95$$

와 동치이다. 주어진 표에서 $P(S \leq 9 \mid H_0) = 0.919 < 0.95$, $P(S \leq 10 \mid H_0) = 0.960 \geq 0.95$ 이므로 $P(S \leq M - 1 \mid H_0) \geq 0.95$ 를 만족하는 최소의 M 은

$$M - 1 \geq 10 \implies M = 11.$$

예제 22. 연속형 확률변수 X 를 한 병에 담긴 소주의 실제 알코올 도수라고 하자. 해당 소주가 빨간뚜껑 소주인지, 혹은 초록뚜껑 소주인지를 판별하고자 한다. 제조 과정상의 변동과 측정 오차로 인해 알코올 도수는 정규분포를 따른다고 가정한다. 귀무가설과 대립가설을 다음과 같이 설정하였다.

$$H_0 : X \sim N(16, 4), \quad H_1 : X \sim N(20, 4).$$

측정된 도수가 18도를 초과하면 H_0 를 기각한다고 하자. 주어진 표준정규분포표를 이용하여 다음을 구하여라.

| z | $P(0 \leq Z \leq z)$ |
|-----|----------------------|
| 1.0 | 0.341 |
| 1.5 | 0.433 |
| 2.0 | 0.477 |
| 2.5 | 0.494 |
| 3.0 | 0.499 |

(a) 제1종 오류를 범할 확률을 구하여라.

(b) 제2종 오류를 범할 확률을 구하여라.

Solution.

(a)

$$\begin{aligned}
 (\text{제1종 오류의 확률}) &= P(X > 18 \mid H_0 \text{이 참}) \\
 &= P\left(\frac{X - 16}{2} > 1\right) \\
 &= P(Z > 1) \\
 &= 0.5 - P(0 \leq Z \leq 1) \\
 &= 0.5 - 0.341 \\
 &= 0.159
 \end{aligned}$$

(b)

$$\begin{aligned}
 (\text{제2종 오류의 확률}) &= P(X \leq 18 \mid H_1 \text{이 참}) \\
 &= P\left(\frac{X - 20}{2} \leq -1\right) \\
 &= P(Z \leq -1) \\
 &= P(Z \geq 1) \\
 &= 0.5 - P(0 \leq Z \leq 1) \\
 &= 0.5 - 0.341 \\
 &= 0.159
 \end{aligned}$$

예제 23. 한 커피 프랜차이즈의 라지 사이즈 커피에 대한 실제 평균 용량(ml)을 μ 라고 하자. 실제 커피의 평균 용량이 표기된 용량인 355ml와 일치하는지 확인하기 위해 아래의 가설을 검정하고자 한다.

$$H_0 : \mu = 355 \quad \text{vs} \quad H_1 : \mu \neq 355$$

해당 프랜차이즈의 라지 사이즈 커피 $n = 36$ 잔을 랜덤추출하여 용량을 측정한 결과, 표본평균과 표본표준편차가 다음과 같이 나타났다.

$$\bar{x} = 351, \quad s = 12$$

유의수준 5%에서 가설을 검정하고자 할 때, 아래의 물음에 답하시오.

- (a) 검정통계량과 임계치(critical value)를 이용하여 (즉, 기각역을 설정하여) 가설을 검정하여라.
- (b) 유의확률(p-value)를 이용하여 가설을 검정하여라.

(단, $P(T \geq t_{1-\alpha}) = \alpha$, $T \sim t(35)$ 에 대해 $t_{0.975} = 2.03$, $t_{0.95} = 1.69$ 이며, $P(T \leq 2) = 0.974$ 임을 이용하여라.)

Solution.

- (a) 가설: $H_0 : \mu = 355$ vs. $H_1 : \mu \neq 355$

유의수준: $\alpha = 0.05$

검정통계량:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{351 - 355}{12/\sqrt{36}} = \frac{-4}{2} = -2$$

기각역:

$$|t| > t_{0.975}(35), \quad t_{0.975}(35) = 2.03$$

이때 $|t| = 2 < 2.030$ 이므로 H_0 를 기각하지 못한다.

즉, 유의수준 5%에서 평균 용량이 355ml와 다르다고 할 충분한 통계적 근거가 없다.

- (b) 가설, 유의수준, 검정통계량은 (a)와 동일하며 검정통계량은 $t = -2$ 이다.

유의확률:

$$\text{p-value} = P(|T| \geq |t|) = 2P(T \geq 2), \quad T \sim t(35).$$

문제에서 $P(T \leq 2) = 0.974$ 이므로 $P(T \geq 2) = 0.026$ 이고,

$$\text{p-value} = 2 \times 0.026 = 0.052.$$

따라서 $\text{p-value} = 0.052 > 0.05$ 이므로 H_0 를 기각하지 못한다.

즉, 유의수준 5%에서 평균 용량이 355ml와 다르다고 할 충분한 통계적 근거가 없다.

예제 24. 한 F1 팀이 새 피트 전략 A 를 도입할지 판단하기 위해, 기존 전략 B 대비 랩타임 (초) 개선량을 시뮬레이션으로 평가한다고 하자. 독립적인 시뮬레이션을 $n = 100$ 번 수행하여, 각 반복에서 얻는 개선량을 X_1, \dots, X_n 이라 하자. X_i 는 서로 독립이고 정규분포를 따른다고 가정하며, 랩타임 개선량의 모분산은 알려져 있지 않다고 하자. 시뮬레이션 결과, 표본평균과 표본표준편차는 각각 $\bar{x} = 0.06$, $S = 0.3$ 로 관측되었다.

새 피트 전략이 통계적으로 유의하다고 판단되기 위해서는 평균 랩타임 개선량이 양수여야 한다. 따라서 다음 가설을 검정한다.

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0$$

유의수준 5%에서 주어진 가설을 검정하고자 할 때, 아래의 물음에 답하시오.

(a) 검정통계량과 임계치(critical value)를 이용하여 (즉, 기각역을 설정하여) 가설을 검정하여라.

(b) 유의확률(p-value)를 이용하여 가설을 검정하여라.

(단, $P(T \geq t_{1-\alpha}) = \alpha$, $T \sim t(99)$ 에 대해 $t_{0.975} = 1.98$, $t_{0.95} = 1.66$ 임을 이용하여라.)

Solution.

(a) 주어진 가설은

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu > 0$$

이며, 유의수준은 $\alpha = 0.05$ 이다.

표본의 크기는 $n = 100$, 표본평균과 표본표준편차는 각각 $\bar{X} = 0.06$, $S = 0.3$ 으로 주어졌다. 모분산이 알려져 있지 않으므로 검정통계량은

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

이며, 귀무가설 하에서 $T \sim t(99)$ 를 따른다.

검정통계량의 값을 계산하면

$$t = \frac{0.06 - 0}{0.30/\sqrt{100}} = \frac{0.06}{0.03} = 2$$

이다.

단측검정에서 유의수준 $\alpha = 0.05$ 이므로 기각역은

$$t > t_{0.95}(99)$$

이며, 문제에서 주어진 값에 따라

$$t_{0.95}(99) = 1.66$$

이다. 따라서

$$t = 2 > 1.66$$

이므로 귀무가설 H_0 를 기각한다.

즉, 유의수준 5%에서 평균 랩타임 개선량이 양수라고 할 충분한 통계적 근거가 있다.

- (b) 주어진 가설, 유의수준, 검정통계량은 (a)와 동일하며, 계산된 검정통계량의 값은 $t = 2$ 이다. 유의확률은

$$\text{p-value} = P(T \geq 2), \quad T \sim t(99)$$

로 정의된다. 문제에서 $t_{99,0.95} = 1.66$ 이고 이는

$$P(T \leq t_{0.95}(99)) = 0.95 \iff P(T \geq 1.66) = 0.05$$

를 의미한다. 또한 계산된 검정통계량의 값이 $2 > 1.66$ 이므로

$$\text{p-value} = P(T \geq 2) < P(T \geq 1.66) = 0.05.$$

따라서 $\text{p-value} < 0.05$ 이므로 유의수준 5%에서 H_0 를 기각한다.

즉, 평균 랩타임 개선량이 양수라고 판단할 충분한 통계적 근거가 있다.