
추정

- 목표: 미지의 모수 θ를 추정하고 싶음.
- 추정량 $\hat{\theta}$을 통하여 미지의 모수 θ를 추정.
- 추정량은 미지의 모수 θ를 추정하는 "rule" 혹은 "법칙" 혹은 "공식"으로 해석가능하다.

example: 평균이 μ인 정규분포에서 10개의 샘플을 뽑았다고 하자.

$$x_1 = 26.3, x_2 = 29.2, \dots$$

우리는 μ를 알고싶다.

- 사람1: 아래와 같이 평균을 추정하자.

$$\hat{\mu} = \frac{26.3 + \cdots + 29.2}{10} = 28.5$$

- 사람2: 아래와 같이 평균을 추정하자.

$$\hat{\mu} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{28.0 + 28.4}{2} = 28.2$$

- 하나의 모수를 추정하는데 여러가지 방법이 있을 수 있다.
- 어떤것이 좋은 추정량인지 판단하는 기준이 필요하다.
- 또한 좋은 추정량들을 구하는 방법도 연구할 필요가 있다.
- 좋은 추정량 구하는 방법: (1) 적률추정법 (2) 최대가능도추정법 (3) 최소분산불편추정법
- 좋은 추정량을 판단하는 기준: (1) 불편성 (2) 효율성 (3) 최소분산성 (4) 일치성(컨시스턴시) (5) 충분성

적률추정법

- 적률추정법은 말그대로 적률을 활용하여 모수를 추정하는 방법을 말함. 적률은 아래와 같은것을 말함.

$$m_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

example: $X_1 \dots X_n \overset{iid}{\sim} Gamma(\alpha, \beta)$ 라고 하자. α, β를 추론해보자. 아래의 관계를 관찰하자.

$$E(X) = \alpha\beta$$
$$V(X) = \alpha\beta^2 + (\alpha\beta)^2$$

그런데 $E(X)$는 m_1으로 추론할 수 있고 $V(X)$는 $m_2 - m_1^2$으로 추론할 수 있다. 따라서 아래를 연립하여 풀면 α,β를 추론할 수 있다.

$$\begin{cases} m_1 = \alpha\beta \\ m_2 - m_1^2 = \alpha\beta^2 + (\alpha\beta)^2 \end{cases}$$

- 장점: WLLN에 의해서 일치성을 보이기 쉽다.

note: 일치성: $\hat{\theta} \xrightarrow{p} \theta \; as \; n \to \infty$.

최대가능도 추정법

example: $X_1, X_2, \dots, X_6 \overset{iid}{\sim} Bernoulli(p)$ 이라고 하자. 아래의 샘플을 관찰했다고 하자.

$$x_1 = 0, \;\; x_2 = 1, \;\; x_3 = 0, \;\; x_4 = 0, \;\; x_5 = 0, \;\; x_6 = 0.$$

p를 추정하고 싶다고 하자.

- 사람1: $p = 0.5$ 라고 하자.

- 사람2: $p = 0.5$ 보다 $p = 0.3$ 이라고 추정하는것이 더 합리적일것 같다. 왜냐하면 $p = 0.5$라고 가정하였을때

$$x_1 = 0, \;\; x_2 = 1, \;\; x_3 = 0, \;\; x_4 = 0, \;\; x_5 = 0, \;\; x_6 = 0.$$

와 같은 샘플을 얻을 확률은

$$\frac{5}{10} \times \frac{5}{10} \times \cdots \times \frac{5}{10} = 0.015625$$

이지만 $p = 0.3$이라고 가정하였을 경우는

$$\frac{7}{10} \times \frac{3}{10} \times \cdots \times \frac{7}{10} = 0.050421$$

가 된다. 따라서 $p = 0.3$ 이라고 추정하는 것이 더 합리적이다.

- 이것이 최대가능도 추정의 모티브이다. 위의 상황을 수식으로 표현하여 보자. 결국 사람2는 아래의 함수를 더 크게 만드는 p가 좋은 추정량임을 주장하는 것이다.

$$L(p) = \prod_{i=1}^{6} pdf(x_i; p)$$

여기에서 $pdf(x_i; p)$는 모수가 p라고 가정하였을 경우 X_i의 pdf이다.

---

note: 위의 예제의 경우 사람1은 $p = 0.5$라고 믿고 있으므로

$$pdf(x_i; p = 0.5) = p^{x_i}(1-p)^{1-x_i}$$

이다. 따라서 $pdf(x_1; p = 0.5) = \cdots = pdf(x_1; p = 0.5) = 0.5$이다. 따라서

$$L(0.5) = (0.5)^6 = 0.015625$$

라고 쓸 수 있다.

note: 사람2는 $p = 0.3$이라고 믿고 있으므로

$$\begin{aligned} pdf(x_1; p = 0.3) =& 0.3^0 \times 0.7^1 = 0.7 \\ pdf(x_2; p = 0.3) =& 0.3^1 \times 0.7^0 = 0.3 \\ & \cdots \\ pdf(x_6; p = 0.3) =& 0.3^0 \times 0.7^1 = 0.7 \end{aligned}$$

와 같이 된다. 따라서

$$L(0.3) = (0.7)^5 \times (0.3)^1 = 0.050421$$

이 된다.

- 그런데 사람3이 $p = 0.2$라고 주장하였다. 이 주장이 더 합리적인지 판단하기 위해서 $L(p)$를 조사해보자. 조사결과

$$L(0.2) = 0.8^5 \times (0.2)^1 = 0.065536$$

따라서 사람3의 주장이 더 합리적이다.

- 전략: 모든 $p \in (0, 1)$에 대하여 아래의 값을 조사하고 이것을 최소화하는 p를 구하자.

$$L(p) = \prod_{i=1}^{n} pdf(x_i; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum_{i=1}^n x_i}$$

로그를 취하면

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

미분을 하면

$$\frac{\partial}{\partial p} \log L(p) = \frac{\sum_{i=1}^n x_i}{p} - \frac{n - \sum_{i=1}^n x_i}{1-p}.$$

따라서 $\frac{\partial}{\partial p} \log L(p) = 0$를 풀면

$$\frac{\sum_{i=1}^n x_i}{p} = \frac{n - \sum_{i=1}^n x_i}{1-p}$$

정리하면

$$p = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

이다.

- 따라서 $p = \bar{x}$로 추정하는게 좋겠다. 즉

$$\hat{p} = \bar{x}$$

따라서 문제의 경우

$$\bar{x} = \frac{0+1+0+0+0+0}{6} = \frac{1}{6}$$

로 p를 추정하는 것이 가장 합리적이다. 이 결과는 놀라울정도로 직관적이고 당연하다. 6번던져서 한번 성공했다면, 성공확률은 대충 1/6로 보는것이 타당할테니까.

- 여기에서 $L(p)$를 p에 대한 likelihood function이라고 한다.

note: 한글로는 우도함수라고 변역하기도 하고 가능도함수라고 번역하기도 한다.

- $\hat{p}$는 likelihood function을 최대화하여 얻은 p의 추정량인데 이러한 이유로 maximum likelihood estimator라고 부르고 줄여서 MLE라고 부른다.

note: 한글로는 최대가능도추정량, 혹은 최대우도추정량이라고 말한다.

- 따라서 $\hat{p}$이 어떠한 방식으로 얻은 추정량인지 더 명확하게 하기위해서 아래와 같이 표기하기도 한다.

$$\hat{p}^{MLE}$$

- 참고로 p를 적률추정법 (method of moments estimator) 으로 추정한다고 하자. 베르누이 분포의 경우

$$EX = p$$

이고 EX는 m_1 즉 $\bar{x}$로 추정할 수 있으므로

$$\hat{p}^{MME} = \bar{x}$$

가 된다. 따라서 베르누이 분포에서 모수 p를 추정하는 경우 적률추정법과 최대가능도 추정법은 같다.

- 모든 경우에서 적률 추정법과 최대가능도 추정법이 같지는 않다. 아래의 예를 살펴보자.

example: $X_1, \ldots, X_n \overset{iid}{\sim} U(0, \theta)$. θ의 MLE를 구해보자. 우도함수는

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta) = \left(\frac{1}{\theta}\right)^n, \quad 0 < x_i < \theta.$$

우도함수에 로그를 취하면 (이것을 로그우도함수라고 부름)

$$\log L(\theta) = \ell(\theta) = -n \log \theta$$

미분을 하면

$$\frac{\partial}{\partial \theta} \ell(\theta) = -n/\theta$$

따라서 $L(\theta)$는 θ의 감소함수이다. 따라서 θ를 작게 고를수록 $L(\theta)$의 값은 커진다. 그런데 아래가 성립하므로

$$0 < x_{(1)} < \cdots < x_{(n)} < \theta$$

θ는 $x_{(n)}$보다 작을 수는 없다. 따라서

$$\hat{\theta}^{MLE} = X_{(n)}$$

- θ를 적률추정법으로 추정하고 싶다면 어떻게 할까?

$$EX = \theta/2$$

이므로

$$\hat{\theta}^{MME} = 2m_1 = 2\bar{x}_1$$

이 된다.

- 따라서 이 경우는 MLE와 MME가 다르다.

note: 최대가능도추정량을 얻는과정은 그렇게 쉽지 않다. (우도함수를 알아야하니까)

note: 그러나 최대가능도추정량의 결과는 매우 직관적인 편이다.

note: 최대가능도추정량은 현재 매우 널리 활용되고 있다. (많은 장점이 있음.)

추정량의 비교

- 추정량이 가져야할 바람직한 성질은 (1) 불편성(언바이어스드니스) (2) 효율성 (3) 최소분산성 (4) 일치성 (5) 충분성이 있다.

- **불편성:** $E(\hat{\theta}) = \theta$.

- $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$라고 하자. p를 아래와 같은 법칙으로 추정한다고 하자.

$$\hat{p} = \frac{\sum X_i}{n}$$

X_i가 확률변수이므로 $\hat{p}$도 확률변수이다.

$$E(\hat{p}) = \frac{\sum EX_i}{n} = p$$

- 이런추정량을 언바이어스드 에스티메이터라고 한다.

note: 적률추정량이나 최대가능도추정량이 항상 언바이어스드 에스티메이터가 되는것은 아니다.

example: $X_1, \dots, X_n \overset{iid}{\sim} N(0, \sigma^2)$ 이라고 하자. 적률추정법으로 σ^2을 추정한다고 하자.

$$\sigma^2 = EX^2 - (EX)^2$$

이므로

$$\hat{\sigma}^2 = m_2 - m_1^2$$

이다. 그런데

$$n(m_2 - m_1^2) = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

이므로

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2$$

이다. 그런데 $E(S^2) = \sigma^2$ 이므로

$$E(\hat{\sigma}^2) \neq \sigma^2$$

이다. 따라서 $\hat{\sigma}^2$은 언바이어스드 에스티메이터가 아니다.

- **일치성**에 대하여 알아보자.

- 위의 예제는 불편성을 만족하지는 않는다. 하지만 n이 충분히 크다면

$$E(\hat{\sigma}) \approx \sigma^2$$

이라고 주장할수 있다. 구체적으로는 아래와 같이 주장할 수 있다.

$$\hat{\sigma}^2 \xrightarrow{p} \sigma^2$$

왜냐하면 $S^2 - \sigma^2 = o_p(1)$이고 $(n-1)/n = O(1)$이기 때문.

- 이 추정량은 잘못된 추정량이라고 보기 아까운데, $E(\sigma^2) \neq \sigma^2$ 이지만 점근적으로는 두 값이 비슷해지기 때문이다.

- 이런성질을 가진 추정량을 컨시스턴시 에스티메이터라고 한다. 구체적으로 $\hat{\theta}$가 θ에 대한 컨시스턴시 에스티메이라고 함은 $\hat{\theta}$가 아래를 만족한다는 의미이다.

$$\hat{\theta} \xrightarrow{p} \theta$$

- **효율성**에 대하여 알아보자.

- 이제부터 특정 추정법이 얼마나 분산이 작은지를 따져볼 것이다.

- 지금까지는 추정량의 평균에 관심이 있었지만 이제는 분산에도 관심을 가질 것이다.

example: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ 이라고 하자. 우리는 여기에서 μ를 추정하고 싶다고 하자. 사람1과 사람2가 있다고 하자.

- 사람1은 아래와 같이 μ를 추정한다.

$$\hat{\mu} = \frac{X_1 + \cdots + X_n}{n}$$

- 사람2는 아래와 같이 μ를 추정한다.

$$\tilde{\mu} = \frac{X_1 + \cdots + X_m}{m}$$

여기에서 $m = n/2$ 이다. (n이 홀수인 경우에는 $m = (n-1)/2$ 라고 하자.)

- 한마디로 말해서 사람2는 사람1과 같은 방법으로 추론하는데 데이터를 절반정도 이유없이 버리고 추론하는 사람이다.

- 직관적으로 생각해도 사람2처럼 추론하는것은 비합리적이다.

- 하지만 어떻게 사람2를 비난할 수 있을까? 어떻게 사람2의 추정량이 나쁜추정량이라고 주장할 수 있을까?

- 불편성의 기준으로 보자.

$$E\hat{\mu} = \frac{n\mu}{n} = \mu$$

이고

$$E\tilde{\mu} = \frac{m\mu}{m} = \mu$$

이다. 따라서 두 추정량 모두 언바이어스드 에스티메이터이다.

- 일치성을 기준으로 보자. 두 추정량 모두 $n \to \infty$ 일때 μ로 수렴한다. 따라서 두 추정량 모두 컨시스턴트 에스티메이터이다.

- 따라서 불편성과 일치성으로는 사람2의 추정량 $\tilde{\mu}$가 나쁜추정량이라고 주장할 근거가 없다.

- 사람2의 추정량을 비난하기 위해 생긴 개념이 효율성이다. 효율성은 추정량의 분산을 판단하는데 분산이 작은 추정량일수록 좋은 추정량이라고 생각한다.

$$V(\hat{\mu}) = \frac{\sigma^2}{n}$$

$$V(\tilde{\mu}) = \frac{\sigma^2}{m}$$

따라서

$$V(\hat{\mu}) < V(\tilde{\mu})$$

가 된다.

- 이때 분산이 작은 추정량을 효율이 좋은 추정량이라 표현한다. 따라서 위의 예제의 경우 $\hat{\mu}$가 $\tilde{\mu}$보다 효율이 좋다.

- 분산이 작은 추정량이라는게 무엇을 의미하는 것일까? 모수를 과녁이라고 하고 추정을 하는 행위를 과녁에 총을 쏘는 행위로 비유해보자. 과녁을 중심으로 총알이 뚫리는것도 중요하지만 총알이 뚫린자리가 밀집해있는것도 중요하다. 분산이 작은 추정량이란 총알이 뚫린자리가 밀집해 있는 추정량이란 의미이다.

크래머-라오 부등식

- 내가 구한 추정량이 불평성도 만족하면 좋겠고 추정량의 분산도 매우 작으면 좋겠다.

- 하지만 (1) 불편성을 만족하면서 (2) 분산도 세상 모든 추정량보다 가장 작은 그런 추정법은 없다. (왜냐하면 분산이 0인 추정법이 존재하기 때문이다.)

note: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ 이라고 하자. 여기에서 μ를 추정한다고 하자. 사람3이 $\hat{\mu} = 21.5$이라고 주장한다고 치자. (이분은 그러니까 데이터를 안보고 추론하시는 분이다.)

note: 사람3이 제시한 추정법은 분산이 0이다.

- 결국 세상에서 가장 작은 분산을 가지는 추정법을 만들려면 사람3처럼 하는수밖에 없기 때문에 이 기준은 포기한다. 대신에 (1) 을 만족하면서 분산이 되도록이면 작은 추정량을 고르는 것이 현실적일것이다.

- 크래머-라오는 (1)을 만족하는 추정량은 분산을 아무리 줄여봤자 특정수준 이하로 줄일수는 없다는 사실을 발견하였다. 이를 크라머-라오 하한이라고 부른다. 그리고 이 내용을 정리하여 아래의 이론을 만들었다.

(정리) 교재 8.4.1.

최소분산불편추정량

- 크라머-라오의 이론은 아래의 상황에서 유용하게 쓸 수 있다.

(1) 내가 θ를 추정하는 어떤 추정법을 만들었다. 이것을 $\hat{\theta}$ 라고 하자.

(2) 그런데 $\hat{\theta}$ 의 평균을 구해서 조사해봤더니 불편추정량임을 알게 되었다.

(3) 또한 $\hat{\theta}$ 의 분산을 구해서 조사해봤더니 이 분산이 크라머-라오 하한이 나왔다.

(4) 그렇다면 크라머-라오의 정리에 따라 내가 구한 추정량이 불편추정량중에 최소분산을 가진다고 주장할 수 있다.

- 이런 추정량을 최소분산불편추정량이라고 한다.

최소분산불편추정량을 구하는 방법

- 크라머-라오의 이론은 아래의 상황에서 유용하게 쓸 수 있다.

(1) 내가 θ를 추정하는 어떤 추정법을 만들었다. 이것을 $\hat{\theta}$ 라고 하자.

(2) 그런데 $\hat{\theta}$ 의 평균을 구해서 조사해봤더니 불편추정량임을 알게 되었다.

(3) 또한 $\hat{\theta}$ 의 분산을 구해서 조사해봤더니 이 분산이 크라머-라오 하한이 나왔다.

(4) 그렇다면 크라머-라오의 정리에 따라 내가 구한 추정량이 불편추정량중에 최소분산을 가진다고 주장할 수 있다.

- 이런 추정량을 최소분산불편추정량이라고 한다.