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
- 좋은 추정량을 판단하는 기준: (1) 불편성 (2) 효율성 (3) 최소분산성 (4) 일치성(컨시스턴시)

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: $X_1, \ldots, X_n \overset{iid}{\sim} N(0, \sigma^2)$ 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.

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

note: 불편성이 만족된다고 하여 항상 일치성이 만족되는것은 아니다. $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$이라고 할때 μ를 X_1으로만 추정한다고 하자. 이 추정량은 불편추정량이지만 일치추정량은 아니다.

- **효율성**에 대하여 알아보자.

- 이제부터 특정 추정법이 얼마나 분산이 작은지를 따져볼 것이다.

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

example: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ 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 (1) 불평성도 만족하면 좋겠고 (2) 추정량의 분산도 작으면 좋겠다.

• 크래머-라오는 불편성을 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.

• 이러한 추정량은 불편성을 가지는 추정량중에 최소분산을 가진다. 따라서 이 추정량을 최소분산불편추정량이라고 부른다. 약어로 MVUE라고 부른다.

• 교재에 따라서 (김우철 교수님 교재 포함) MVUE를 전역최소분산불편추정량 즉 UMVUE라고 부르기도 한다.

• 이러한 이름을 붙이게 된 배경은 아래와 같다.

- 불편성을 만족하는 추정량중에 최소분산성을 갖추는 추정량을 찾는다고 하자. 하지만 이 룰은 지나치게 모호해서 실제로는 비상식적인 추정량이 좋은 추정량처럼 보이는 경우도 있다. 즉 치팅이 존재한다.

example: $X_1, \dots, X_n \overset{iid}{\sim} N(19.5, \sigma^2)$ 이라고 하자. 여기에서 μ를 추정한다고 하자. 사람3이 $\hat{\mu} = 19.5$ 이라고 주장한다고 치자. (이분은 그러니까 데이터를 안보고 추론하시는 분이다.)

- 사람3의 추정은 불편성을 만족하며 분산이 0이다. 따라서 완벽한 추정량이다.

- 이러한 추정량을 제외해야함이 옳다. 그래서 데이터를 안보고 추론하는 것은 제외한다고 공지를 내렸다.

- 사람4의 경우 사람3의 치팅을 응용하였다. 아래와 같이 추정한다고 하자.

$$19.5\frac{99}{100} + \frac{1}{100}\bar{X}$$

- 결국 사람4는 데이터를 보는척만 하는 것이다. 사람4의 추정법은 실제 모수가 19.5일 경우만 꽤 정확한 추정처럼 보인다.

- 사람3이나 사람4와 같은 치팅종류는 공통점이 있는데 특정 조건에서만 불편성을 만족한다는 것이다. 즉 $\theta = 19.5$인 경우에서만 불편성을 만족한다. $\theta = 19.4$만 되어도 불편성을 만족하지 않는다.

- 결국 크라머라오의 부등식에서 "불편성을 만족하는 추정량"이라는 표현은 "θ의 참값에 상관없이 항상 불편성을 만족하는 추정량"으로 바뀌어야 한다.

- 이러한 이유로 MVUE를 UMVUE로 부르기도 한다.

note: UMVUE라고 표현하지 않으면 사람3과 추정량이 "내가 모수가 19.5일때 한정 MVUE" 라고 주장해도 딱히 반박할 말이 없어보인다.

UMVUE, 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.

- 아래와 같은 스토리를 가진 예제 혹은 문제가 많은 교재에 실려있다.

(1) MME이든 MLE이든 어떤 방법으로 모수 θ를 추론하였다고 치자. 예를들어 MME로 추정했다고 치자.

(2) $\hat{\theta}^{MME}$의 평균을 계산해봤더니 모수와 일치하여 불편성을 만족함을 알 수 있었다.

(3) 혹시나하고 $\hat{\theta}^{MME}$의 분산을 계산해봤더니 그 분산이 크래머-라오의 하한이 나왔다.

(4) 그렇다면 우리가 구했던 MME가 UMVUE라는 인증을 받게 된다.

note: 쉽게 생각하면 MLE, MME는 추정을 하는 방법? 혹은 원리? 를 나타낸다. 이것이 좋은 추정량이라는 증거는 없다.

note: UMVUE는 좋은 추정량임을 인증하는 마크라고 생각하면 된다.

UMVUE를 구하는 방법?

- 좋은 추정량 즉 UMVUE를 만들기 위해서 어떻게 하면 좋을까?

- 좋은 추정량을 만들기 위해서는 MME, MLE와 같은 좋은 추정법을 개발하는 방법도 있다. 하지만 맛있는 음식은 좋은 재료에서 나오듯이 좋은 추정량을 만들기 위해서 좋은 통계량을 먼저 골라보자고 생각할 수 있다.

note: 추정량과 통계량의 차이는 무엇인가? 추정량은 모수의 추론을 위해 만들어진 특별한 통계량이라고 해석하면 된다. 따라서 통계량은 추정량을 만드는 재료라고 볼 수 있다.

- 그렇다면 어떠한 통계량이 θ의 UMVUE를 만들기에 적절할까?

- 라오-블랙웰은 $\hat{\theta}^{UMVUE}$를 얻기 위해서는 우선 θ에 대한 충분통계량을 만들어야 함을 증명하였다. (정리 8.2.5.)

- 레만-셰퍼는 θ에 대한 완비충분통계량으로 만든 θ의 불편추정량이 결국 θ의 UMVUE가 됨을 증명하였다. (정리 8.3.1.)

- 따라서 우리의 전략은 아래와 같다.

(1) 데이터를 손질하여 완비충분통계량을 만든다.

(2) 손질된 완비충분통계량으로 불편추정량을 만들어 UMVUE를 만든다.

Step1. (최소)충분통계량을 구한다.

example: $X_1, \dots, X_4 \overset{iid}{\sim} Bernoulli(p)$ 라고 하자. 만약에 $X_1, \dots, X_4$를 관측한 결과가 아래와 같다고 하자.

$$(x_1, x_2, x_3, x_4) = (1, 1, 0, 0)$$

이 샘플을 가지고 불량률을 $\hat{p}$를 추론한다면 아래와 같이 추론하는것이 합리적으로 보인다.

$$\hat{p} = 2/4$$

이는 좋은 추정방법이다.

- 아래는 $X_1, \dots, X_4$를 관측한 결과에 따라서 추정한 결과이다.

(1) $(1, 1, 0, 0) \quad \Rightarrow \quad \hat{p} = 2/4 = 0.5.$

(2) $(1, 0, 1, 0) \quad \Rightarrow \quad \hat{p} = 2/4 = 0.5.$

(3) $(0, 0, 1, 1) \quad \Rightarrow \quad \hat{p} = 2/4 = 0.5.$

- p를 추론하는데 총 불량품수만 기억하면 되지 불량품들이 나온 순서 즉 개개의 x_i의 값은 알 필요가 없다. 즉 각 실험결과에 대하여 모든 관측치를 일일히 기록할 필요가 없고 아래만 기록해도 **충분**하다.

$$Y = \sum_{i=1}^{4} X_i$$

- 이때 Y는 확률표본의 함수이므로 분명히 통계량이다. 또한 직관적으로 Y만 알고있으면 우리는 $\hat{p}$를 추론하기에 충분하다는 것을 알 수 있다. 이때 $Y = u(X_1, \dots, X_n)$를 θ에 대한 충분통계량이라고 한다.

- 어떻게 Y를 찾을 수 있을까? 매번 직관적으로 찾아야 하나? 그리고 어떻게 $\hat{p}$를 추론함에 있어서 Y만 고려해도 **충분**하다고 주장할 수 있을까? **충분**하다는 것의 수학적 정의는 무엇일까?

- $Y = y$가 주어졌을때 $f(X_1, \ldots, X_n | Y = y)$가 모수 p의 함수가 아니라면 Y는 p의 충분통계량이다.

---

example: 위의 예를 다시 풀어보자.

$$\begin{aligned}
&P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0 \mid Y = 2) \\
&= \frac{P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, Y = 2)}{P(Y = 2)} \\
&= \frac{P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0)}{P(Y = 2)} \\
&= \frac{P(X_1 = 1)P(X_2 = 1)P(X_3 = 0)P(X_4 = 0)}{P(Y = 2)} \\
&= \frac{p \cdot p \cdot (1-p) \cdot (1-p)}{P(Y = 2)} \\
&= \frac{p \cdot p \cdot (1-p) \cdot (1-p)}{\frac{4!}{2!2!} p^2 (1-p)^2} = \frac{1}{6}
\end{aligned}$$

note: 이때 $P(Y = 2) = \frac{4!}{2!2!} p^2 (1-p)^2$의 계산은 아래의 논리에 기초한다.

(1) $X_1, X_2, X_3, X_4 \overset{iid}{\sim} Bernoulli(p)$ 이므로 $Y \sim b(4, p)$ 이다.

(2) 따라서

$$P(Y = 2) = \frac{4!}{2!2!} p^2 (1-p)^2.$$

- 아무튼

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0 \mid Y = 2) = \frac{1}{6}$$

인데 이 확률은 p와 무관하다.

- 관측한 샘플이

$$X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 0$$

의 경우에도 동일하게 구할 수 있다. 일반적으로는 아래와 같이 확장할 수 있다.

(1) $X_1, \ldots, X_n \overset{iid}{\sim} Bernoulli(p)$.

(2) $Y = \sum_{i=1}^{n} X_i \sim b(n, p)$.

$$P(X_1 = x_1, \ldots, X_n = x_n | Y = y) = \frac{y!(n-y)!}{n!}$$

• $Y = y$가 주어지면 $X_1, \ldots, X_n$의 조건부 분포는 p와 무관하다. $\Longrightarrow$ $Y = y$가 주어졌을때 $X_1, \ldots, X_n$의 조건부 분포에는 p의 정보가 없다. $\Longrightarrow$ Y는 p의 충분통계량이다.

• 따라서 아래의 순서로 모수 p를 추론하는 것이 합리적이다.

(1) 자료 $X_1, \ldots, X_n$를 잘 요약하여 p의 충분통계량 Y를 만든다.

(2) Y를 이용하여 p의 추정량 $\bar{X}$를 만든다.

note: 물론 $y = \sum_{i=1}^{4} x_i$ 대신에 $\bar{x} = \sum_{i=1}^{4} x_i/4$를 기록해도 된다. $\bar{x}$ 역시 p 를 추론하기에 충분하다. 즉 $\bar{X}$ 역시 그 자체로 충분통계량이다.

note: 위의 경우 직관적으로 $\bar{X}$역시 충분통계량임을 예측할 수 있으므로 굳이 (1) Y를 만들고 (2) $\bar{X}$를 만드는 것이 바보같이 보일수 있다. 하지만 문제는 대부분의 경우 θ의 정보를 요약하기 위해서 무엇을 기록하면 충분한지 감을 못잡는 경우가 많다는 것이다. 즉 충분통계량을 항상 직관적으로 얻을 수 있지는 않다.

example: $X_1, \ldots, X_n \overset{iid}{\sim} Gamma(\alpha, \beta)$ 이라고 하자. (α, β)를 추론하기에 적절한 충분통계량은 무엇인가?

• 정답은 $(\sum_{i=1}^{n} X_i, \prod_{i=1}^{n} X_i)$이다. 즉 (α, β)를 추론하기위해서 관측치들의 합과 관측치들의 곱만 기억해도 충분하다. (예제 8.2.5.)

• 이런 경우는 직관적으로 찾기 힘든 경우이다. 따라서 충분통계량을 직관적으로 찾기 힘든경우라도 찾게 만들어주는 수학적 도구가 있으면 좋겠다. 이것이 바로 분해정리이다.

(정리) (분해정리) 교재의 정리 8.2.1.

최소충분통계량

• $(\sum_{i=1}^{n} X_i, \prod_{i=1}^{n} X_i)$가 (α, β)의 충분통계량이라는 사실에서 보듯이 충분통계량이 벡터도 될 수 있다. 따라서 벡터 $(X_1, \ldots, X_n)$, 즉 데이터 그 자체는 모든 모수의 충분통계량이라는 논리가 성립한다.

• 하지만 데이터 자체는 좋은 충분통계량이 되지 않는데 그 이유는 압축을 하지 않기 때문이다. 압축의 개념을 좀 더 수학적으로 표현할 수 있을까? 만약에 A라는 자료가 있고 B라는 자료가 있다고 하자. A를 통해서는 B를 만들수 있는데 B를 통해서 A를 못만든다면 B가 A보다 압축된 자료라고 볼 수 있다. 즉 A를 B로 바꾸는 변환은 존재하는데 B를 A로 바꾸는 변환이 존재하지 않는다면 B는 A보다 압축된 자료라고 볼 수 있다.

- 다시 감마분포의 예제로 돌아가보자.

(1) $(X_1, \ldots, X_n)$로는 $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$를 만들 수 있다. 즉 아래를 만족하는 변환이 존재한다.

$$g : (X_1, \ldots, X_n) \to \left(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i \right)$$

(2) $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$로는 $(X_1, \ldots, X_n)$를 만들 수 없다. 즉 아래를 만족하는 변환은 없다.

$$g : \left(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i \right) \to (X_1, \ldots, X_n)$$

- 이 경우 $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$ 와 $(X_1, \ldots, X_n)$은 모두 충분통계량이 되지만

$$\left(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i \right)$$

이 $(X_1, \ldots X_n)$보다 작은 자료, 즉 더욱 압축된 자료라고 볼 수 있다.

- 이 논의를 확장하여 보자. 특정한 충분통계량 $T^*(X_1, \ldots, X_n)$를 상상하자. 그리고 어떠한 충분통계량 $T(X_1, \ldots, X_n)$ 에 대하여서도 $T^*(X_1, \ldots, X_n)$으로 가는 함수 g를 찾을 수 있다고 하자. 즉

$$\text{for all sufficient statistics } T \text{ there exists } g \;\; st. \quad g : T \to T^*$$

이라고 하자. 그러면 T^*는 충분통계량중에서 가장 작은 충분통계량이 된다. 이러한 충분통계량을 최소충분통계량이라고 한다.

- 일반적으로 충분통계량은 당연히 최소충분통계량을 의미한다고 보면 된다. (최소충분통계량이 아니면 충분성의 개념을 만들 이유가 없음.)

note: 당연한 소리지만 최소충분통계량에 1:1변환을 태운것도 최소충분통계량이다. 베르누이 분포에서 p를 추정한다고 할때 $\bar{X}$와 $\sum_{i=1}^n X_i$는 모두 최소충분통계량이다.

- 충분통계량을 구하는 것도 매우 힘든데 (직관적 추측 혹은 분해정리 이용) 그것이 최소충분통계량인지도 따져야 한다면 매우 힘들것이다.

- 다행이 아래의 정리들이 있어서 최소충분통계량을 얻는것은 매우 쉽다.

(정리) 지수족에서 θ에 대한 MLE를 구했는데 (1) 그 MLE가 유일한 MLE이고 (2) 그 MLE가 θ 충분통계량이라면

$\hat{\theta}^{MLE}$ 가 최소충분통계량임!

이라는 정리가 존재한다. (정리 8.2.4.)

(정리) 지수족에서 $T(x)$는 (최소)충분통계량이면서 완비통계량임.

note: 지수족은 많은 분포를 포함한다. (1) 베르누이/바이노미알/멀티노미알 (2) 기하/음이항 (3) 포아송/지수/감마 (4) 정규/카이제곱 등.

step 2. 완비성을 따져보자.

- θ를 잘 추정하는 좋은 추정량을 얻기 위해서 우선 좋은 통계량을 얻을 필요가 있음을 밝혔다. 좋은 통계량의 성질 중 하나가 (최소)충분성이었는데 다른하나는 완비성이다. 우리는 θ에 대한 충분성을 만족하면서 완비성도 만족하는 통계량을 찾아야하는데 이러한 통계량을 완비충분통계량이라고 하고 기호로 CSS라고 표현한다. 즉

$$\begin{aligned} & T(x) \text{ is CSS of } \theta \\ \iff & T(x) \text{ is SS of } \theta \text{ and } T(x) \text{ is CS of } \theta \end{aligned}$$

- 다행이 지수족의 경우 $T(x)$가 CSS가 됨이 증명되어있다.