충분통계량 개요

정의

- 개요

1. 어떤 통계량 Y=$u(x_1, \dots, x_n)$이 모수 θ에 대해 모든 정보를 가지고 있다면,

1) $x_1, \dots, x_n$에 대한 Y의 조건부 분포는 모수 θ에 대해 그 어떤 정보도 포함하고 있지 않으며

2) Y_2=k$(x_1, \dots, x_n)$ 에 대하여

(1) Y_2가 f$(x_1, \dots, x_n|Y)$를 pdf로 갖는 분포에서 정의한 통계량이라면,

(2) 이제 Y_2를 토대로 θ를 추정하는 것은 불가능 할 것이다.

3) 이처럼, Y가 θ에 대한 충분한 정보를 가지고 있을 때

(1) $\frac{f(x_1;\theta)\cdot f(x_2;\theta)\dots f(x_n;\theta)}{f(u(x_1,\dots,x_n);\theta)} = h(x_1, \dots, x_n)$,

(2) 즉 θ에 종속되지 않을 때 이를 충분통계량 이라고 한다.

정의

- *네이만의 인수분해 정리*

1. $X_1, \dots, X_n$을 $\theta \in \Omega$에 대해 pdf 혹은 pmf를 갖는 분포에서 추출한 확률표본이라고 하자.

1) 이 때, 통계량 $Y_1 = u(x_1, \dots, x_n)$ 에 대하여

2) $\prod_{i=1}^{n} f(x_i;\theta) = K_1(u(x_1, \dots, x_n);\theta) \cdot K_2(x_1, \dots, x_n)$
즉, θ 의존하는 Y_1 에 대한 함수 K_1과 θ 에 독립적인 K_2로 인수분해가 가능할 경우
$Y_1 = u(x_1, \dots, x_n)$ 는 충분통계량이다.

2. 증명

1) $Y_1 = u_1(x_1, \dots, x_n)$ 으로 재정의하고, $Y_2 = u_2(x_1, \dots, x_n)$ … $Y_n = u_n(x_1, \dots, x_n)$ 으로 정의하면

(1) 그 역함수는 $x_1 = w_1(y_1, \dots, y_n)$, … $x_n = w_n(y_1, \dots, y_n)$이고

(2) 그 때의 야코비안 |J|는 변환에 따라 정의 가능하다.

정의

- *네이만의 인수분해 정리*

2. 증명

2) $\prod_{i=1}^{n} f(x_i;\theta) = K_1(u_1(x_1,\dots,x_n);\theta)\cdot K_2(x_1,\dots,x_n)$*를 다시 쓰면*

(1) $g(y_i;\theta) = K_1(y_1;\theta)\cdot K_2(w_1,\dots,w_n)|J|$

3) y_1만의 주변 pdf를 구하면

(1) $g(y_1;\theta) = \int\cdots\int K_1(y_1;\theta)\cdot K_2(w_1,\dots,w_n)|J|\,dy_2\cdots dy_n$
$= K_1(y_1;\theta)\int\cdots\int K_2(w_1,\dots,w_n)|J|\,dy_2\cdots dy_n$
$= K_1(y_1;\theta)\cdot \mathrm{m}(y_1)$

4) 한편, 가정 $\prod_{i=1}^{n} f(x_i;\theta) = K_1(u_1(x_1,\dots,x_n);\theta)\cdot K_2(x_1,\dots,x_n)$ *에서*

(1) $K_1(u_1(x_1,\dots,x_n);\theta) = \frac{\prod_{i=1}^{n} f(x_i;\theta)}{K_2(x_1,\dots,x_n)}$ 이므로, 이를 이용하면

(2) $g(y_1;\theta) = \prod_{i=1}^{n} f(x_i;\theta)\,\frac{\mathrm{m}(y_1)}{K_2(x_1,\dots,x_n)}$ 에서 $\frac{\prod_{i=1}^{n} f(x_i;\theta)}{g(y_1;\theta)} = \frac{K_2(x_1,\dots,x_n)}{\mathrm{m}(y_1)}$

5) 마찬가지로 $\frac{K_2(x_1,\dots,x_n)}{\mathrm{m}(y_1)}$ 는 θ에 의존하지 않으므로 $g(y_1;\theta)$는 충분통계량이다.

예제

- $X_1, \dots, X_n$을 $f(x;\theta)$ = θ^{x}(1- $\theta)^{(1-x)}$를 가지는 분포에서 추출한 확률표본이라고 하자.

1. Y_1= $X_1 + \cdots + X_n$ 일 때, 통계량 Y_1의 pmf는

1) $f_{Y_1}(Y_1;\theta)$ = $\binom{n}{y}\theta^{\sum x_i}(1- \ \theta)^{(n-\sum x_i)}$ 이다.

2) 이 때, 우도함수와 Y_1의 비율은

(1) $\dfrac{\prod_{i=1}^{n} f(x_i;\theta)}{\theta^{\sum x_i}(1- \ \theta)^{(n-\sum x_i)}} = \dfrac{\theta^{x_1}(1- \ \theta)^{(1-x_1)}.\theta^{x_2}(1- \ \theta)^{(1-x_2)}\dots.}{\binom{n}{y}\theta^{\sum x_i}(1- \ \theta)^{(n-\sum x_i)}} = \dfrac{\theta^{\sum x_i}(1- \ \theta)^{(n-\sum x_i)}}{\binom{n}{y}\theta^{\sum x_i}(1- \ \theta)^{(n-\sum x_i)}} = \dfrac{1}{\binom{n}{y}}$

(2) 결과 h($X_1, \dots, X_n$) = $\frac{1}{\binom{n}{y}}$ 로 , 이는 θ에 의존하지 않으므로 Y_1은 충분통계량이다.

예제

- $X_1, \ldots, X_n$을 $N(\theta, \sigma^2)$를 가지는 분포에서 추출한 확률표본이라고 하자.

1. 이 때, 결합 pdf는

1) $\prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i-\theta)^2}{2\sigma^2}\right)$
$= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \theta)^2\right)$

2) 한편, $\sum (x_i - \theta)^2 = \sum [(x_i - \overline{x}) - (\overline{x} - \theta)]^2$
$= \sum (x_i - x)^2 + 2\,(x_i - \theta) \sum (x_i - x) + (\overline{x} - \theta)^2$

(1) 이 때, $\sum (x_i - \overline{x}) = 0$ 이므로, 이는 결국

(2) $\sum (x_i - \theta)^2 = \sum (x_i - \overline{x})^2 + n(\overline{x} - \theta)^2$

3) $\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \overline{x})^2 + n(\overline{x} - \theta)^2\right)$ 에서, 이를 분리하면

(1) $\prod_{i=1}^{n} f(x_i;\theta) = \exp\left(-\frac{1}{2\sigma^2} n(\overline{x} - \theta)^2\right) \cdot \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \overline{x})^2\right)$

(2) 이는 명백히

- θ에 의존하는 통계량 $\overline{x}$ 에 대한 함수 $\exp\left(-\frac{1}{2\sigma^2} n(\overline{x} - \theta)^2\right)$ 와
- θ에 의존하지 않는 $\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \overline{x})^2\right)$ 로 인수분해 되므로
- $\overline{x}$는 충분통계량이다.