7장. 충분통계량 (Sufficient Ststistics; S.S.)

Data : $X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(x|\theta)$

Based on the observed data, we can make an inference about θ.

Statistic(통계량) : any function of $X_1, X_2, \cdots, X_n$.

Summarize data :

$$\overline{X},\quad S^2,\quad X_{(1)} < X_{(2)} < \cdots < X_{(n)}$$

i.e., any statistic is a data summarization

7.1 충분통계량 (Sufficient Statistic)

Statistical inference(통계적 추론) : Derivation of information about the parameter θ from the given data $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$.

즉, 주어진 자료 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$로부터 모수 θ에 대한 정보를 이끌어 내는 것

Data reduction(자료 축약) : 주어진 자료에서 모수에 대한 정보를 가진 부분만을 간추려 전체 자료 대신에 축약된 정보만을 이용하여 모수에 대한 추론을 함.

Sufficient Statistic(충분통계량)의 의미 이해를 위한 예문 설명 :

Let X, Y be random sample from $N(\theta, 1)$. We are going to do statistical inference for θ based on sample.

$$(X, Y) \xleftrightarrow{1-1} (X+Y, X-Y) \xrightarrow{\textit{data reduction}} (X+Y)$$

Sufficient Statistic(충분통계량)의 의미 : 어떤 통계량 $T(\boldsymbol{X})$가 모수 θ에 대해 가지고 있는 정보가 원 자료 $\boldsymbol{X}$가 모수에 대해 가지고 있는 정보와 같음.

Def 7.1 (*Sufficient Statistic* : 충분통계량)

A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of X given the values of $T(X)$ does not depend on θ.

즉, 통계량 $T(X)$가 θ에 관한 모든 정보를 담고 있어서 $T(X)$가 주어졌을 때 X의 조건부 분포가 θ에 의존하지 않으면 $T(X)$를 θ의 충분통계량이라 함.

● What is the condition dist^n $P_\theta(X=x|T(X)=t(x))$??

$$P_\theta(X=x|T(X)=t(x))=\frac{P_\theta(X=x,T(X)=t(x))}{P_\theta(T(X)=t(x))}$$

$$=\frac{P_\theta(X=x)}{P_\theta(T(X)=t(x))}$$

$$=\frac{p(x|\theta)}{q(t(x)|\theta)}$$

Where $p(x|\theta)$ is the p.m.f of X, $q(T(x)|\theta)$ is the p.m.f of $T(X)$.

Ex 7.1 $X_1,X_2,\cdots,X_n$: random sample from the Bernoulli(p).

What is a sufficient statistic for p?

Sol.

$$X_1,X_2,\cdots,X_n \overset{iid}{\sim} f(x:p)=p^x(1-p)^{1-x}\ ,\quad x=0,1$$

Let $Y=\sum_{i=1}^{n}X_i$. Then, $Y\sim Bin(n,p)$

Joint p.m.f of $X_1,X_2,\cdots,X_n$: $p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\cdots p^{x_n}(1-p)^{1-x_n}$

$$=p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$$

$$P(X_1=x_1,\cdots,X_n=x_n\,|\,Y=y)=\frac{P(X_1=x_1,\cdots,X_n=x_n)}{P(Y=y)}$$

$$=\frac{p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}}{\binom{n}{y}p^y(1-p)^{n-y}}=\frac{1}{\binom{n}{y}}$$

$\Rightarrow$ Dist^n of $X_1,\cdots,X_n$ condition $Y=y$ does not depend on p.

$\Rightarrow$ $Y=\Sigma X_i$ is a sufficient statistic.

Question. When we're finding a S.S., do we need to use the above approach? Are there any other simple methods to do that?

Answer. Using the above approach is not so simple. Luckily, there is a simple device proposed by Neyman.
: Factorization Theorem.

Thm 7.1 (Neyman's factorization theorem : 인수분해정리)

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(\cdot, \theta)$$

$T = t(X_1, \cdots, X_n)$ is a S.S for θ

$$\Leftrightarrow f(x_1, x_2, \cdots, x_n, \theta) = g[\, T(x_1, x_2, \cdots, x_n), \theta\,] \cdot h(x_1, x_2, \cdots, x_n),$$

where $h(x_1, x_2, \cdots, x_n)$ does not depend on θ.

Ex. 7.2

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} Poisson(\lambda)$. Show that $\sum_{i=1}^{n} X_i$ is a S.S. for λ.

Sol.

$$f(x_1, x_2, \cdots, x_n, \lambda) = \frac{\lambda^{\Sigma x_i} e^{-n\lambda}}{x_1!\, x_2! \cdots x_n!} = \left(\lambda^{\Sigma x_i} e^{-n\lambda}\right) \cdot \left(\frac{1}{x_1!\, x_2! \cdots x_n!}\right)$$

$\therefore$ By factorization thm, $\sum_{i=1}^{n} X_i$ is a S.S. for λ.

참고. Any one-to-one function of a S.S. is also S.S.
Therefore, $\overline{X}$ is also a S.S. for λ in Ex 7.2.

<u>즉, 충분통계량의 일대일 함수는 모두 충분통계량이다.</u>

Minimal Sufficient Statistic (최소충분통계량)의 개념 등장

Example.

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, 1), \quad -\infty < \mu < \infty. \text{ Find a S.S for } \mu.$$

Sol.

$$f(x_1, x_2, \cdots, x_n, \theta) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum (x_i - \mu)^2\right]$$

$$= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum \left\{(x_i - \bar{x}) + (\bar{x} - \mu)\right\}^2\right]$$

$$= \left\{\exp\left[-\frac{n}{2}(\bar{x} - \mu)^2\right]\right\} \cdot \left\{\frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum (x_i - \bar{x})^2\right]\right\}$$

$\therefore$ By factorization thm, $\overline{X}$ a is S.S. for μ.

Therefore, $\sum_{i=1}^{n} X_i$ is also a S.S. for μ.

Ex. 7.4 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} U(0, \theta)$. Find a S.S. for θ.

Sol.

Ex. 7.5 $X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(x|\theta) = e^{-(x-\theta)}, \ x > \theta$

Sol.

$$\prod_{i=1}^{n} f(x_i|\theta) = e^{-(\sum x_i - n\theta)}\, \mathrm{I}(\min x_i > \theta)$$

$$= e^{-\sum x_i} e^{n\theta}\, \mathrm{I}(\min x_i > \theta)$$

$\Rightarrow$ By factorization Thm, $\min X_i$: S.S. for θ.

Some other examples

- $f(x|\theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, $0 < x < \infty$

$$\prod_{i=1}^{n} f(x_i|\theta) = \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}\sum_{i=1}^{n} x_i}$$

By factorization Thm, $\sum_{i=1}^{n} X_i$ is a S.S. for θ.

- $f(x|\theta) = \theta x^{\theta-1}$, $0 < x < 1$

$$\prod_{i=1}^{n} f(x_i|\theta) = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}$$

By factorization Thm, $\prod_{i=1}^{n} X_i$ is a S.S. for θ.

7.2 지수족 (Exponential family)

Examples : Binomial, Poisson, Negative binomial, Normal, Gamma, Beta, etc.

Def 7.2 (*Exponential family* of probability distribution : 지수족)

A family of pdfs or pmfs is called an exponential family if the pdf or pmf has the following form :

$$f(x|\theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\},$$

$$\text{or, } f(x|\theta) = g(\theta)h(x)\exp\{c(\theta)T(x)\}.$$

Example 1 : $B(n, p)$

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x},\ x = 0, 1, 2, \cdots, n$$

$$= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n$$

$$= \exp\left(x\log\left(\frac{p}{1-p}\right) + n\log(1-p) + \log\binom{n}{x}\right)$$

$$c(p) = \log\left(\frac{p}{1-p}\right),\ T(x) = x,\ d(p) = n\log(1-p),\ S(x) = \log\binom{n}{x}$$

: Exponential family.

Example 2 : $f(x|\theta) = \frac{1}{\theta} e^{1-\frac{x}{\theta}},\ x > \theta > 0$

==> It is not an exponential family

since $f(x|\theta) = \frac{1}{\theta} e^{1-\frac{x}{\theta}} [I(x > \theta)]$ and $I(x > \theta)$ is a function of x and θ. That is, we can not separate this.

Note. If the support of a distribution family depends on parameters, it can not be an exponential family.

k-th exponential family

A family of pdfs or pmfs is called an exponential family if the pdf or pmf has the following form :

$$f(x|\theta) = \exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(x) + d(\theta) + S(x)\right\},$$

$$\text{or, } f(x|\theta) = g(\theta)h(x)\exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(x)\right\}.$$

and support is not dependent on θ.

Theorem (S.S. for exponential family of dist^n)

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(x|\theta) = \exp\left\{\sum_{j=1}^{k} c_j(\theta)T_j(x) + d(\theta) + S(x)\right\}$$

$$\Rightarrow \left(\sum_{i=1}^{n} T_1(X_i), \sum_{i=1}^{n} T_2(X_i), \cdots, \sum_{i=1}^{n} T_k(X_i)\right) : \text{S.S. for } \theta.$$

Ex. 7.3

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. Show that $\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ is a S.S. for (μ, σ^2).

Sol.

$$f(x_i, x_2, \cdots, x_n|\mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right]$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[\left(-\frac{1}{2\sigma^2}\right)\sum_{i=1}^{n} x_i^{\,2} + \left(\frac{\mu}{\sigma^2}\right)\sum x_i - \frac{n\mu^2}{2\sigma^2}\right]$$

: 2-nd Exponential family

$\therefore$ By factorization thm, $\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ is a joint S.S. for (μ, σ^2).

Therefore, if we let $Y_1 = \sum_{i=1}^{n} X_i$ and $Y_2 = \sum_{i=1}^{n} X_i^2$,

$$\overline{X} = \frac{Y_1}{n} \ \& \ S^2 = \frac{Y_2 - Y_1^{\,2}/n}{n-1} \text{ is also a S.S. for } (\mu, \sigma^2).$$

7.3 완비통계량 (Complete Statistics)

Def 7.3 (*Complete Statistics* ; 완비통계량)

A statistic $T(X)$ is called a complete statistic. if

$$E_\theta[g(T(X))]=0 \text{ for all } \theta \text{ implies } P_\theta\{g(T(X))=0\}=1 \text{ for all } \theta.$$

Ex. 7.6 (Poisson)

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} Poisson(\theta)$. Show that $T(X)=\sum_{i=1}^{n} X_i$ is a complete statistic for θ.

Sol.

Since $T(X)=\sum_{i=1}^{n} X_i \sim Poisson(n\theta)$, $E[g(T(X))]=0$ implies

$$E_\theta[g(T(X))]=\sum_{k=0}^{\infty} g(k)\frac{(n\theta)^k e^{-n\theta}}{k!}=0$$

$$\Rightarrow \sum_{k=0}^{\infty} g(k)\frac{(n\theta)^k}{k!}=0.$$

That is, $g(0)+ng(1)\theta+\frac{n^2 g(2)}{2!}\theta^2+\frac{n^3 g(3)}{3!}\theta^3+\cdots=0$ for all θ.

Therefore, $g(\cdot)\equiv 0$.

$$\therefore \quad T(X)=\sum_{i=1}^{n} X_i \text{ is a complete statistic for } \theta.$$

Ex 7.7 (Binomial)

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} Bernoulli(p)$. Show that $T(X)=\sum_{i=1}^{n} X_i$ is a complete statistic for p.

Sol.

$T(X) \sim Bin(n,p)$. Let g be a ftn such that $E_p(g(T))=0$. Then,

$$0=E_p(g(T))=\sum_{t=0}^{n} g(t)\binom{n}{t}p^t(1-p)^{n-t}=(1-p)^n\sum_{t=0}^{n} g(t)\binom{n}{t}(\frac{p}{1-p})^t$$

$$=(1-p)^n\sum_{t=0}^{n} g(t)\binom{n}{t}r^t,\quad 0<r=\frac{p}{1-p}<\infty$$

$$\Rightarrow g(t)=0 \text{ for } t=0,1,\ ...,n$$

$$\Rightarrow P_\theta(g(T)=0)=1$$

$$\therefore\ T(X)=\sum_{i=1}^{n} X_i \text{ is a complete statistic for } p.$$

Ex 7.8 (Uniform)

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} U(0,\theta)$. Show that $T(X)=X_{(n)}$ is a complete statistic for θ.

Sol.

Recall that the pdf of $T(X)=X_{(n)}$ is

$$f_T(t)=n\frac{t^{n-1}}{\theta^n},\ 0<t<\theta.$$

$E[g(T(X))]=0$ implies

$$E[g(T(X))]=\int_0^\theta g(t)n\frac{t^{n-1}}{\theta^n}dt=0 \text{ for all } \theta>0.$$

$$\Rightarrow \int_0^\theta g(t)t^{n-1}dt=0$$

differentiation with θ gives (양변을 θ에 관해 미분)

$$\Rightarrow g(\theta)\theta^{n-1}=0$$

$\therefore g(\theta)=0$ for all $\theta>0$. Hence, complete statistic.

Thm 7.2 (Complete Statistic for Exp'l family)

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} f(x|\theta)$ with the following exponential family :

$$f(x|\theta) = \exp\{c(\theta)T(x) + d(\theta) + S(x)\}.$$

Then, $\sum_{i=1}^{n} T(X_i)$ is a C.S. for θ.

Def. (*Complete Sufficient Statistic* ; C.S.S. 완비충분통계량)

A sufficient statistic which has completeness is called
Complete Sufficient Statistic. In exponential family, SS is also CSS.

Theorem (C.S.S. for exponential family of dist^n)

$$X_1, X_2, \cdots, X_n \overset{iid}{\sim} f(x|\theta) = \exp\left\{\sum_{j=1}^{k} c_k(\theta) T_j(x) + d(\theta) + S(x)\right\}$$

$$\Rightarrow \left(\sum_{i=1}^{n} T_1(X_i), \sum_{i=1}^{n} T_2(X_i), \cdots, \sum_{i=1}^{n} T_k(X_i)\right) : \text{C.S.S. for } \theta.$$

Ex. 7.9 (Poisson)

$X_1, X_2, \cdots, X_n \overset{iid}{\sim} Poisson(\theta)$. Find a CSS for θ.

Fill out by yourself....