Biostatistics 602 - Statistical Inference
Lecture 08
Data Reduction - Summary

Hyun Min Kang

February 5th, 2013

Last Lecture

1. What is an exponential family distribution?
2. Does a Bernoulli distribution belongs to an exponential family?
3. What is a curved exponential family?
4. What is an obvious sufficient statistic from an exponential family?
5. When can the sufficient statistic be complete?

Theorem 6.2.25

Suppose $X_1, \cdots, X_n$ is a random sample from pdf or pmf $f_X(x|\theta)$ where

$$f_X(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left[\sum_{j=1}^{k} w_j(\boldsymbol{\theta}) t_j(x)\right]$$

is a member of an exponential family. Then the statistic $T(\mathbf{X})$

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{j=1}^{n} t_1(X_j), \cdots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is complete as long as the parameter space $\boldsymbol{\Theta}$ contains an open set in $\mathbb{R}^k$

Exponential Family Example

Problem

$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$. Determine whether the following statistics are whether (1) sufficient (2) complete, and (3) minimal sufficient.

$$\mathbf{T}_1(\mathbf{X}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right), \mathbf{T}_2(\mathbf{X}) = \left(\overline{X}, s_{\mathbf{X}}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/(n-1)\right)$$

How to solve it

- Decompose $f_X(x|\mu, \sigma)$ in the form of an an exponential family.
- Apply Theorem 6.2.10 to obtain a sufficient statistic and see if it is equivalent to or related to $\mathbf{T}_1(\mathbf{X})$ and $\mathbf{T}_2(\mathbf{X})$.
- Apply Theorem 6.2.25 to show that it is complete.
- Apply Theorem 6.2.28 to show that it is minimal sufficient.

Applying Theorem 6.2.10

$$f_X(x|\mu,\sigma^2) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\exp\left(\frac{\mu}{\sigma^2}x - \frac{x^2}{2\sigma^2}\right)$$

where

$$\begin{cases} h(x) = 1 \\ c(\boldsymbol{\theta}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\ w_1(\boldsymbol{\theta}) = \mu/\sigma^2 \\ w_2(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2} \\ t_1(x) = x \\ t_2(x) = x^2 \end{cases}$$

By Theorem 6.2.10,
$(\sum_{i=1}^n t_1(X_i), \sum_{i=1}^n t_2(X_i)) = (\sum_{i=1}^n X_1, \sum_{i=1}^n X_i^2) = \mathbf{T}_1(\mathbf{X})$ is a sufficient statistic

Applying Theorem 6.2.25. and Theorem 6.2.28

$$\begin{aligned} A &= \{(w_1(\boldsymbol{\theta}), w_2(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \mathbb{R}^2\} \\ &= \left\{\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2} : \mu \in \mathbb{R}, \sigma > 0\right\} \end{aligned}$$

Contains a open subset in $\mathbb{R}^2$, so $\mathbf{T}_1(\mathbf{X})$ is also complete by Theorem 6.2.25. By Theorem 6.2.28, $\mathbf{T}_1(\mathbf{X})$ is also minimal sufficient.

Connecting $\mathbf{T}_2(\mathbf{X})$ to $\mathbf{T}_1(\mathbf{X})$

$$\begin{aligned} \mathbf{T}_1(\mathbf{X}) &= \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) \\ \mathbf{T}_2(\mathbf{X}) &= (\overline{X}, s^2_{\mathbf{X}}) \end{aligned}$$

$$\begin{cases} \overline{X} = \frac{\sum_{i=1}^n X_i}{n} = g_1(\mathbf{T}_1(\mathbf{X})) \\ s^2_{\mathbf{X}} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} = \frac{\sum_{i=1}^n X_i^2 + \sum_{i=1}^n X_i^2/n}{n-1} = g_2(\mathbf{T}_1(\mathbf{X})) \end{cases}$$

$$\begin{cases} \sum_{i=1}^n X_i = n\overline{X} = g_1^{-1}(\mathbf{T}_2(\mathbf{X})) \\ \sum_{i=1}^n X_i^2 = (n-1)s^2_{\mathbf{X}} - n\overline{X}^2 = g_2^{-1}(\mathbf{T}_2(\mathbf{X})) \end{cases}$$

Therefore, $\mathbf{T}_2(\mathbf{X})$ is an one-to-one function of $\mathbf{T}_1(\mathbf{X})$, and also is sufficient, complete, and minimal sufficient.

Example of Curved Exponential Family

Problem

$X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \mu^2)$. Determine whether the following statistic is whether (1) sufficient (2) complete, and (3) minimal sufficient.

$$\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$$

How to solve it

- Decompose $f_X(x|\mu)$ in the form of an an exponential family.
- Apply Theorem 6.2.10 to obtain a sufficient statistic and see if it is equivalent to or related to $\mathbf{T}(\mathbf{X})$
- Apply Theorem 6.2.25 to see if it is complete.
- Apply Theorem 6.2.28 to see if it is minimal sufficient.

Applying Theorem 6.2.10

$$f_X(x|\mu) = \frac{1}{2\pi\mu^2}\exp\left(-\frac{1}{2}\right)\exp\left(\frac{1}{\mu}x - \frac{x^2}{2\mu^2}\right)$$

where

$$\begin{cases} h(x) = 1 \\ c(\mu) = \frac{1}{2\pi\mu^2}\exp\left(-\frac{1}{2}\right) \\ w_1(\mu) = 1/\mu \\ w_2(\mu) = -\frac{1}{2\mu^2} \\ t_1(x) = x \\ t_2(x) = x^2 \end{cases}$$

By Theorem 6.2.10, $\left(\sum_{i=1}^n t_1(X_i), \sum_{i=1}^n t_2(X_i)\right) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) = \mathbf{T}(\mathbf{X})$ is a sufficient statistic for μ

Applying Theorem 6.2.25.

$$\begin{aligned} A &= \{(w_1(\mu), w_2(\mu)) : \mu \in \mathbb{R}\} \\ &= \left\{\frac{1}{\mu^2}, -\frac{1}{2\mu^2} : \mu \in \mathbb{R}\right\} \end{aligned}$$

A does not contains a open subset in $\mathbb{R}^2$, so we cannot apply Theorem 6.2.25. We need to go back to the definition

Is $\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ Complete?

$$\begin{aligned} E\left(\sum_{i=1}^n X_i\right) &= n\mu \\ E\left(\sum_{i=1}^n X_i^2\right) &= nE\left(X_i^2\right) \\ &= n\left[E(X_i)^2 + \mathrm{Var}(X_i)\right] \\ &= n(\mu^2 + \mu^2) = 2n\mu^2 \end{aligned}$$

Note that $\sum_{i=1}^n X_i \sim \mathcal{N}(n\mu, n\mu^2)$.

$$\begin{aligned} E\left[\left(\sum_{i=1}^n X_i\right)^2\right] &= \left[E\left(\sum_{i=1}^n X_i\right)\right]^2 + \mathrm{Var}\left(\sum_{i=1}^n X_i\right) \\ &= (n\mu)^2 + n\mu^2 = n(n+1)\mu^2 \end{aligned}$$

Is $\mathbf{T}(\mathbf{X}) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ Complete? (cont'd)

Define

$$\begin{aligned} g(\mathbf{T}(\mathbf{X})) &= g\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right) \\ &= \frac{\sum_{i=1}^n X_i^2}{2n} - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n(n+1)} \\ E[g(\mathbf{T})|\mu] &= \frac{E\left(\sum_{i=1}^n X_i^2\right)}{2n} - \frac{E\left(\sum_{i=1}^n X_i\right)^2}{n(n+1)} \\ &= \frac{2n\mu^2}{2n} - \frac{n(n+1)\mu^2}{n(n+1)} = 0 \end{aligned}$$

for all $\mu \in \mathbb{R}$. Because there exist $g(\mathbf{T})$ such that $E[\mathbf{T}|\mu] = 0$ and $\Pr(g(\mathbf{T}) = 0) < 1$, $\mathbf{T}(\mathbf{X})$ is NOT complete.

Is $\mathbf{T}(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ Minimal Sufficient?

$$\frac{f_X(\mathbf{x}|\mu)}{f_X(\mathbf{y}|\mu)} = \exp\left[\frac{\sum_{i=1}^n y_i^2 - \sum_{i=1}^n x_i^2}{2\mu^2} + \frac{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}{\mu}\right]$$

The ratio above is a constant to μ if and only if

$$\begin{cases} \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2 \\ \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \end{cases}$$

which is equivalent to $\mathbf{T}(\mathbf{x}) = \mathbf{T}(\mathbf{y})$. Therefore, $\mathbf{T}(\mathbf{X})$ is a minimal sufficient statistic.

Summary of Sufficiency Principle

- Model : $\mathcal{P} = \{f_X(x|\theta), \theta \in \Omega\}$
- Statistic : $T = T(\mathbf{X})$ where $\mathbf{X} = (X_1, \cdots, X_n)$.

Sufficient Statistic

Contains all info about θ

Definition $f_{\mathbf{X}}(\mathbf{x}|T(\mathbf{X}))$ does not depend on θ

Theorem 6.2.2 $f_{\mathbf{X}}(\mathbf{x}|\theta)/q_T(T(\mathbf{X})|\theta)$ does not depend on θ

Factorization Theorem $f_{\mathbf{X}}(\mathbf{x}|\theta) = h(\mathbf{x})g(T(\mathbf{X})|\theta)$

Exponential Family $(\sum_{i=1}^n t_1(X_i), \cdots, \sum_{i=1}^n t_k(X_i))$ is sufficient

Summary of Sufficiency Principle (cont'd)

Minimal Sufficient Statistic

Sufficient statistic that achieves the maximum data reduction

Definition T is sufficient and it is a function of all other sufficient statistics.

Theorem 6.2.13 $f_{\mathbf{X}}(\mathbf{x}|\theta)/f_{\mathbf{X}}(\mathbf{y}|\theta)$ is constant as a function of $\theta \iff T(\mathbf{x}) = T(\mathbf{y})$

Exponential Family (Theorem 6.2.28) Complete and sufficient statistic is minimal sufficient

Summary of Sufficiency Principle (cont'd)

Complete Statistic

This family have to contain "many" distributions in order to be complete. The restriction $E[g(T)|\theta] = 0, \ \forall\theta \in \Omega$ is strong enough to rule out all non-zero functions

Definition $E[g(T)|\theta] = 0$ implies $\Pr(g(T) = 0|\theta) = 1$.

Exponential Family The parameter space Ω is an open subset of $\mathbb{R}^k$.

Example

Problem

The random variable X takes the values 0, 1, 2, according to one of the following distributions:

	$\Pr(X=0)$	$\Pr(X=1)$	$\Pr(X=2)$	
Distribution 1	p	$3p$	$1-4p$	$0<p<\frac{1}{4}$
Distribution 2	p	p^2	$1-p-p^2$	$0<p<\frac{1}{2}$

In each case, determine whether the family of distribution of X is complete.

Solution - Distribution 1

Suppose that there exist $g(\cdot)$ such that $E[g(X)|p]=0$ for all $0<p<\frac{1}{4}$.

$$\begin{aligned}
f_X(x|p) &= p^{I(x=0)}(3p)^{I(x=1)}(1-4p)^{I(x=2)} \\
E[g(X)|p] &= \sum_{x\in\{0,1,2\}} g(x)f_X(x|p) \\
&= g(0)\cdot p + g(1)\cdot(3p) + g(2)\cdot(1-4p) \\
&= p[g(0)+3g(1)-4g(2)] + g(2) = 0
\end{aligned}$$

Therefore, $g(2)=0$, $g(0)+3g(1)=0$ must hold, and it is possible that g is a nonzero function that makes $\Pr[g(X)=0]<1$. For example, $g(0)=1, g(1)=-3, g(2)=0$. Therefore the family of distribution of X is not complete.

Solution - Distribution 2

Suppose that there exist $g(\cdot)$ such that $E[g(X)|p]=0$ for all $0<p<\frac{1}{4}$.

$$\begin{aligned}
f_X(x|p) &= p^{I(x=0)}(p^2)^{I(x=1)}(1-p-p^2)^{I(x=2)} \\
E[g(X)|p] &= \sum_{x\in\{0,1,2\}} g(x)f_X(x|p) \\
&= g(0)\cdot p + g(1)\cdot p^2 + g(2)\cdot(1-p-p^2) \\
&= p^2[g(1)-g(2)] + p[g(0)-g(2)] + g(2) = 0
\end{aligned}$$

$g(0)=g(1)=g(2)=0$ must hold in order to $E[g(X)|p]=0$ for all p. Therefore the family of distribution of X is complete.

Another Example

Problem

Let $X_1,\cdots,X_n$ be iid samples from

$$f_X(x|\mu,\lambda) = \left(\frac{\lambda}{2\pi x^3}\right)^{1/2}\exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right]$$

where $x>0$. Show that $\overline{X}=\frac{1}{n}\sum_{i=1}^n X_i$ and $T=\frac{n}{\sum_{i=1}^n \frac{1}{X}-\frac{1}{\overline{X}}}$ are sufficient and complete.

Solution

$$
\begin{aligned}
f_X(x|\boldsymbol{\theta}) &= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right] \\
&= \left(\frac{\lambda}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\lambda x^2}{2\mu^2 x} + \frac{2\lambda\mu x}{2\mu^2 x} - \frac{\lambda\mu^2}{2\mu^2 x}\right] \\
&= \left(\frac{1}{2\pi x^3}\right)^{1/2} \lambda^{1/2} \exp\left[-\frac{\lambda}{2\mu^2}x + \frac{\lambda}{\mu} - \frac{\lambda}{2}\cdot\frac{1}{x}\right] \\
&= \left(\frac{1}{2\pi x^3}\right)^{1/2} \lambda^{1/2} e^{\lambda/\mu} \exp\left[-\frac{\lambda}{2\mu^2}x - \frac{\lambda}{2}\cdot\frac{1}{x}\right] \\
&= h(x)c(\boldsymbol{\theta})\exp\left[w_1(\boldsymbol{\theta})t_1(x) + w_2(\boldsymbol{\theta})t_2(x)\right]
\end{aligned}
$$

Solution (cont'd)

where

$$
\begin{aligned}
h(x) &= \frac{1}{2\pi x^3} \\
c(\boldsymbol{\theta}) &= \lambda^{1/2} e^{\lambda/\mu} \\
w_1(\boldsymbol{\theta}) &= -\frac{\lambda}{2\mu^2} \\
t_1(x) &= x \\
w_2(\boldsymbol{\theta}) &= -\frac{\lambda}{2} \\
t_2(x) &= \frac{1}{x}
\end{aligned}
$$

Therefore $\mathbf{T}(\mathbf{X}) = (T_1(\mathbf{X}), T_2(\mathbf{X})) = (\sum_{i=1}^n X_i, \sum_{i=1}^n 1/X_i)$ is a complete sufficient statistic because $\boldsymbol{\theta} = (\lambda, \mu)$ contains an open set in $\mathbb{R}^2$.

Solution (cont'd)

Now, we need to show that $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $T = \frac{n}{\sum_{i=1}^n \frac{1}{X} - \frac{1}{\overline{X}}}$ are one-to-one function of $\mathbf{T}(\mathbf{X})$.

$$
\begin{aligned}
\overline{X} &= \frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}T_1(\mathbf{X}) \\
T &= \frac{n}{\sum_{i=1}^n \frac{1}{X} - \frac{1}{\overline{X}}} = \frac{n}{T_2(\mathbf{X}) - \frac{n}{T_1(\mathbf{X})}} \\
T_1(\mathbf{X}) &= n\overline{X} \\
T_2(\mathbf{X}) &= \frac{n}{T} + \frac{1}{\overline{X}}
\end{aligned}
$$

Therefore, $(\overline{X}, T)$ are one-to-one function of $(T_1(\mathbf{X}), T_2(\mathbf{X}))$ and are also a sufficient complete statistic.

Summary

Today

- More Examples of Exponential Family
- Review of Chapter 6

Next Lecture

- Likelihood Function
- Point Estimation