Sufficient Statistics and the Exponential Family

Statistic: (informal) A summary of the data.

Statistic: (formal) Let $X_1, \dots, X_n$ be a random sample of size n from a population and let $T(x_1, \dots, x_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \dots, X_n)$. Then the random variable or random vector $Y = T(X_1, \dots, X_n)$ is called a statistic.

Sufficient Statistic: (informal). A summary of the data that contains all of the information about any unknown parameters θ in the data

Sufficient Statistic: (formal, not that useful): A statistic $Y = T(\underset{\sim}{X})$ is a sufficient statistic for θ if the conditional distribution of the sample data $\underset{\sim}{X}$ given the value of $T(\underset{\sim}{X})$ does not depend on θ.

Intuition: $T(\underset{\sim}{X})$ follows a distribution that depends on θ.
$\Rightarrow$ we could use $T(\underset{\sim}{X})$ to estimate θ

Once we know $T(\underset{\sim}{X})$ the distribution of $\underset{\sim}{X}$ does not depend on θ
$\Rightarrow$ knowing $\underset{\sim}{X}$ doesn't give you any more information about θ than what we had in $T(\underset{\sim}{X})$.

Factorization Theorem: Let $f(\underset{\sim}{x} \mid \theta)$ denote the joint pdf or pmf of a sample $\underset{\sim}{X}$. A statistic $T(\underset{\sim}{X})$ is a sufficient statistic for θ if and only if there exist functions $g(t \mid \theta)$ and $h(\underset{\sim}{x})$ such that for all sample points $\underset{\sim}{x}$ and all parameter values θ,

$$f(\underset{\sim}{x} \mid \theta) = g(T(\underset{\sim}{x}) \mid \theta) \cdot h(\underset{\sim}{x})$$

Pf: See book

Note 1: Suppose we want to obtain a maximum likelihood estimate, and $t = T(\underset{\sim}{x})$ is a sufficient statistic. Then

$$\mathcal{L}(\theta \mid \underset{\sim}{x}) = f(\underset{\sim}{x} \mid \theta)$$
$$= g(T(\underset{\sim}{x}) \mid \theta) \cdot h(\underset{\sim}{x})$$

$\Rightarrow$ log-likelihood is

$$\ell(\theta \mid \underset{\sim}{x}) = \log\{\mathcal{L}(\theta \mid \underset{\sim}{x})\}$$
$$= \log\{g(T(\underset{\sim}{x}) \mid \theta)\} + \log\{h(\underset{\sim}{x})\}$$

$$\Rightarrow \frac{d}{d\theta}\ell(\theta \mid \underset{\sim}{x}) = \frac{d}{d\theta}\log\{g(T(\underset{\sim}{x}) \mid \theta)\}$$

MLE depends only on $T(\underset{\sim}{x})$, not the full data vector $\underset{\sim}{x}$

Note 2: Suppose we have a prior distribution $\Theta \sim f_\Theta(\theta)$.

The posterior for Θ has pdf/pmf:

$$\cancel{f_{\Theta \mid \underset{\sim}{X}}(\theta \mid \underset{\sim}{x}) = c \cdot f_\Theta(\theta) \cdot f_{\underset{\sim}{X} \mid \Theta}(\underset{\sim}{x} \mid \theta)}$$
$$\cancel{= c \cdot f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta) \cdot h(\underset{\sim}{x})}$$
$$\cancel{= c_2 \cdot f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta) \quad (\text{where } c_2 = c \cdot h(\underset{\sim}{x}))}$$

~~The posterior distribution of~~

$$f_{\Theta \mid \underset{\sim}{X}}(\theta \mid \underset{\sim}{x}) = \frac{f_{\Theta, \underset{\sim}{X}}(\theta, \underset{\sim}{x})}{f_{\underset{\sim}{X}}(\underset{\sim}{x})} = \frac{f_{\Theta, \underset{\sim}{X}}(\theta, \underset{\sim}{x})}{\int f_{\Theta, \underset{\sim}{X}}(\theta, \underset{\sim}{x})\, d\theta} = \frac{f_\Theta(\theta)\, f_{\underset{\sim}{X} \mid \Theta}(\underset{\sim}{x} \mid \theta)}{\int f_\Theta(\theta) \cdot f_{\underset{\sim}{X} \mid \Theta}(\underset{\sim}{x} \mid \theta)\, d\theta}$$

$$= \frac{f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta)\, \cancel{h(\underset{\sim}{x})}}{\int f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta)\, \cancel{h(\underset{\sim}{x})}\, d\theta} = \frac{f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta)}{\int f_\Theta(\theta) \cdot g(T(\underset{\sim}{x}) \mid \theta)\, d\theta}$$

posterior depends only on $T(\underset{\sim}{x})$, not full data vector $\underset{\sim}{x}$

Example: Suppose $X_1, \ldots, X_n \sim$ iid Normal(μ, σ^2)

~~both μ and σ^2 unknown. $\theta = (\mu, \sigma^2)$.~~

The joint pdf of $\underset{\sim}{X}$ is

$$f_{\underset{\sim}{X}}(\underset{\sim}{x} \mid \mu, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ \frac{-(x_i - \mu)^2}{2\sigma^2} \right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x} - \mu)^2 \right\}$$

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

Suppose μ is unknown, σ^2 known.

$$f_{\underset{\sim}{X}}(\underset{\sim}{x} \mid \mu, \sigma^2) = \underbrace{\exp\left\{ -\frac{1}{2} \cdot \frac{n(\bar{x} - \mu)^2}{\sigma^2} \right\}}_{\text{involves } \mu} \cdot \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \right\}}_{\text{no } \mu}$$

$$= g(T(\underset{\sim}{x}) \mid \mu) \cdot h(\underset{\sim}{x})$$

where $T(\underset{\sim}{x}) = \bar{x}$,

$$g(t \mid \mu) = \exp\left\{ \frac{-1}{2} \frac{n(t - \mu)^2}{\sigma^2} \right\},$$

$$h(\underset{\sim}{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ \frac{-1}{2} \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{\sigma^2} \right\}$$

Now suppose both μ, σ^2 are unknown. Set $\theta = (\mu, \sigma^2)$

Our sufficient statistics are $\underset{\sim}{T}(\underset{\sim}{X}) = (T_1(\underset{\sim}{X}), T_2(\underset{\sim}{X}))$, where

$$T_1(\underset{\sim}{X}) = \bar{X}, \quad T_2(\underset{\sim}{X}) = S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Set $h(\underset{\sim}{x}) = 1$ and

$$g(\underset{\sim}{t} \mid \underset{\sim}{\theta}) = g(t_1, t_2 \mid \mu, \sigma^2)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\left\{n(t_1 - \mu)^2 + (n-1)t_2\right\} / 2\sigma^2\right]$$

Then $f(\underset{\sim}{x} \mid \mu, \sigma^2) = g(\underset{\sim}{t} \mid \underset{\sim}{\theta}) \cdot h(\underset{\sim}{x})$

so $\underset{\sim}{T}(\underset{\sim}{X}) = (\bar{X}, S^2)$ is a sufficient statistic for the normal model.

Example: Suppose that $X_1, \ldots, X_n \overset{iid}{\sim}$ Binomial(n, θ), θ unknown, Show that $\sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

Exponential Family (not to be confused with the exponential distribution)

A family of probability distributions is called an exponential family if its pdfs/pmfs can be expressed as

$$f(x|\underset{\sim}{\theta}) = h(x)\cdot c(\underset{\sim}{\theta})\cdot \exp\left\{\sum_{i=1}^{k} w_i(\underset{\sim}{\theta})t_i(x)\right\}$$

Example: Binomial exponential family

Suppose $X \sim \text{Binomial}(n, \theta)$

$$\begin{aligned} f(x|\theta) &= \binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= \binom{n}{x}(1-\theta)^n\left(\frac{\theta}{1-\theta}\right)^x \\ &= \binom{n}{x}(1-\theta)^n \exp\left\{\log\left(\frac{\theta}{1-\theta}\right)x\right) \end{aligned}$$

$\therefore$ this is an exponential family with

$h(x) = \binom{n}{x}$, $c(\theta) = (1-\theta)^n$, $w_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$,

and $t_1(x) = x$.

Other distributions that are exponential families:

Normal, exponential, gamma, χ^2, beta, Dirichlet, Poisson, geometric

Why do we care?

Thm. (Pitman-Koopman-Darmois)

Suppose that $X_1, \ldots, X_n$ are iid r.v.'s with pdf $f_{X_i}(x_i|\theta)$ and the support of $f_{X_i}(x_i|\theta)$ does not depend on θ (Counter-example: Uniform$[0,\theta]$). Only if $f_X(x|\theta)$ is an exponential family is there a sufficient statistic $T(\underset{\sim}{X}) = (T_1(\underset{\sim}{X}), \ldots, T_k(\underset{\sim}{X}))$ whose length k does not increase as n increases.

Thm: Let $X_1, \dots, X_n$ be iid where $f_{X_i|\theta}(x_i|\theta)$ is in an exponential family with pdf is

$$f(x|\underset{\sim}{\theta}) = h(x) \cdot c(\underset{\sim}{\theta}) \cdot \exp\left\{ \sum_{j=1}^{k} w_j(\underset{\sim}{\theta}) t_j(x) \right\},$$

where $\underset{\sim}{\theta} = (\theta_1, \dots, \theta_d)$ for $d \le k$.

Then $T(\underline{X}) = \left(\sum_{j=1}^{n} t_1(X_i), \dots, \sum_{j=1}^{n} t_k(X_i) \right)$

is a sufficient statistic for θ.

Rao-Blackwell Thm:

Let $\hat{\theta}$ be an estimator of a parameter θ, and $T(\underset{\sim}{X})$ a sufficient statistic for θ.

Define $\tilde{\theta} = E[\hat{\theta} \mid T(\underset{\sim}{X})]$.

Then $MSE(\tilde{\theta}) \le MSE(\hat{\theta})$.

We have "Rao-Blackwellized" the original estimator $\hat{\theta}$ to obtain an improved estimator $\tilde{\theta}$.

Note: If $\hat{\theta}$ was unbiased, $\tilde{\theta}$ is still unbiased.

Basic idea: If an estimator $\hat{\theta}$ wasn't based on a sufficient statistic, it can be improved by conditioning on a sufficient statistic.

Suppose $X_1, \ldots, X_n \sim \text{Normal}(\theta, \sigma^2)$ with σ^2 known.

Set $\hat{\theta} = X_1$. Not based on the sufficient statistic $\bar{X}$, clearly not optimal.

It can be shown that $X_1 \mid \bar{X} \sim \text{Normal}(\bar{X}, \frac{n-1}{n}\sigma^2)$

$\tilde{\theta} = E(\hat{\theta} \mid \bar{X}) = E(X_1 \mid \bar{X}) = \bar{X}$

$\tilde{\theta} = \bar{X}$ is the Rao-Blackwellized estimator, and

$MSE(\tilde{\theta}) = \text{Bias}(\tilde{\theta}) + \text{Var}(\tilde{\theta}) = \frac{\sigma^2}{n} < \sigma^2 = \text{Var}(\hat{\theta}) = MSE(\hat{\theta})$,

as given by the Rao-Blackwell theorem.