Information in data — captured by the likelihood

→ but in what format?

Data $\underset{\sim}{x}$ reduced to summary statistics $T(\underset{\sim}{x})$

sufficient for estimating parameter θ : $\hat{\theta} = f(T(\underset{\sim}{x}))$

Definition A statistic $T(\underset{\sim}{x})$ is sufficient for θ if the distribution $\underset{\sim}{X} | T = t$ does not depend on θ

↓

That is, knowing $T(\underset{\sim}{x})$ captures all of the information in $\underset{\sim}{x}$ about θ.

Factorization Theorem

If T is sufficient for θ we can write the model as

$$f_\theta(\underset{\sim}{x}) = g(T(\underset{\sim}{x}), \theta)\, h(x)$$

g: depends on $\underset{\sim}{x}$ only through $T(\underset{\sim}{x})$

h: does not depend on θ.

Example from normal-mean model

$X_i \sim^{iid} N(\mu, \sigma^2)$

$\theta = (\mu, \sigma^2)$

$$\Rightarrow f_\theta(\underset{\sim}{x}) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right\}$$

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

Sufficient statistics $\sum x_i^2$ and $\sum x_i$

But ... $\sum_{i=1}^{m} x_i$ and $\sum_{i=m+1}^{n} x_i$ (partial sums) are also sufficient statistics

$\Rightarrow$ (except of) minimal sufficiency : maximum data reduction that still retains all the information about θ.

Definition $T(\underset{\sim}{x})$ is minimally sufficient if it is a function of any other sufficient statistic.

May not be unique : any one-to-one map of minimal sufficient T is minimal sufficient

If $T(\underset{\sim}{x})$ is sufficient for $\theta \Rightarrow$ likelihood based on $\underset{\sim}{x} \equiv$ likelihood based on $T(\underset{\sim}{x})$ (follows from Factorization Thm)

$\Rightarrow L(\theta)$ is a function of any T

$\Rightarrow$ $L(\theta)$ is minimal sufficient

Exponential Families

Important class of distributions for development of theory of statistics.

Has nice properties (but) also covers many important models used in practice.

One-parameter model

$$p_\theta(x) = \exp\{\eta(\theta)T(x) - A(\theta) + c(x)\}$$

- $A(\theta)$: mean-function
- $\eta(\theta)$: function of θ
- $T(x)$: sufficient statistic
- $c(x)$: normalizing constant

n iid samples

$$p_\theta(\underset{\sim}{x}) = \exp\{\eta(\theta)\underbrace{\sum_{i=1}^{n} T(x_i)}_{T(\underset{\sim}{x})} - \underbrace{nA(\theta)}_{\tilde{A}(\theta)} + c(\underset{\sim}{x})\}$$

$\sum_{i=1}^{n} T(x_i)$: nice property of exponential family

When $\eta(\theta) = \theta$ we say we have the distribution in its <u>canonical form</u> and it looks especially simple

$$p_\theta(\underset{\sim}{x}) = \exp\{\theta T(\underset{\sim}{x}) - \tilde{A}(\theta) + c(\underset{\sim}{x})\}$$

Some properties $\Rightarrow$ $\boxed{E(T(\underset{\sim}{x})) = \tilde{A}'(\theta)}$

$\boxed{V(T(\underset{\sim}{x})) = \tilde{A}''(\theta)}$

(proof via moment generating functions)
(google)

This mean-variance relationship important for selecting appropriate model to use.

Example normal-mean model (σ^2 known)

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

$E(\bar{x}) = \mu$, $V(\bar{x}) = \frac{\sigma^2}{n}$ does not depend on μ

Poisson

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

$E(\bar{x}) = \mu$ $V(\bar{x}) = \frac{\mu}{n}$ does depend on mean μ.

Multiparameter case

$\underset{\sim}{\theta} = \{\theta_1, \ldots, \theta_p\}$

$$f_{\underset{\sim}{\theta}}(\underset{\sim}{x}) = \exp\left\{ \sum_{j=1}^{p} \eta_j(\underset{\sim}{\theta})\, T_j(\underset{\sim}{x}) - A(\underset{\sim}{\theta}) + C(\underset{\sim}{x}) \right\}$$

vector of sufficient statistics $\{T_1(\underset{\sim}{x}) \ldots T_p(\underset{\sim}{x})\}$

vector of functions of $\underset{\sim}{\theta}$ - $\eta_j(\underset{\sim}{\theta})$ $j = 1, \ldots, p$ where each component may involve more than one component of θ.

Example normal mean model

$$p_\theta(x) = \exp\left\{ -\frac{x^2}{2\sigma^2} + \frac{\mu}{\sigma^2}x - \frac{1}{2}\left\{ \frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right\} \right\}$$

$\theta = (\mu, \sigma^2)$

$A(\theta)$

$$\eta(\theta) = \left\{ \frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2} \right\} \qquad T(x) = \{ x, x^2 \}$$

Likelihood & exponential families

$$L(\theta|x) = \exp\{ \eta(\theta)T(x) - A(\theta) + C(x) \}$$

$$\log L(\theta|x) = \eta(\theta)T(x) - A(\theta) + C(x)$$

$\Downarrow$

score $S(\theta) = \eta'(\theta)T(x) - A'(\theta)$

If $\eta(\theta) = \theta$, $S(\theta) = T(x) - A'(\theta)$

MLE $\Rightarrow \{\theta : S(\theta) = 0\} \Rightarrow$ solution to $\eta'(\theta)T(x) = A'(\theta)$

or if canonical form

$T(x) = A'(\theta)$

Curvature, Fisher information

$$I(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = -\eta''(\theta)T(x) + A''(\theta)$$

$\Rightarrow$ canonical form $I(\theta) = A''(\theta)$

Remember for sample size n

$\Rightarrow T(\underset{\sim}{X}) = \sum T(X_i)$ and $\tilde{A}(\theta) = n A(\theta)$

$\Rightarrow I_n(\theta) = n A''(\theta)$

Also remember we have

$$\mathrm{Var}(\hat{\theta}) = I_n(\theta)^{-1} = n^{-1} \left(I_1(\theta)^{-1} \right)$$

n^{-1}: sample size dependency of variance

$I_1(\theta)^{-1}$: constant in variance expression that depends on shape of $L(\theta)$ (specific family model)

Looking at the score in more detail

$S(\theta) = \frac{\partial}{\partial \theta} \log L(\theta)$

↓

Take expectation under true model p_θ

$$E_\theta S(\theta) = \int \left(\frac{\partial}{\partial \theta} \log L(\theta|x) \right) p_\theta(x)\, dx$$

$$\overset{\text{chain rule}}{=} \int \frac{\frac{\partial}{\partial \theta} L(\theta|x)}{L(\theta|x)} p_\theta(x)\, dx$$

$$= \int \frac{\partial}{\partial \theta} L(\theta|x)\, dx$$

for nice distributions (can move $\frac{\partial}{\partial \theta}$ outside $\int$)

$$\frac{\partial}{\partial \theta} \underbrace{\int L(\theta|x)\, dx}_{=1 \text{ since } L(\theta|x) = p_\theta(x)} = 0$$

true for exponential families

So, the expected value of $S(\theta)$ is 0

⇒ logical to use $\hat{\theta} = \{\theta : S(\theta) = 0\}$ as an estimate since we hope $S(\theta) = \frac{\partial}{\partial\theta} \log L(\theta|x)$ observed (for a particular x) not too far from expected

How much can $S(\theta)$ vary?

$$\mathrm{Var}_\theta S(\theta) = E_\theta(S(\theta)^2) = \int \left(\frac{\partial \log L(\theta)}{\partial \theta}\right)^2 p_\theta(x)\,dx$$

since $E_\theta(S(\theta)) = 0$

$$= \int \left(\frac{\partial \log L(\theta)}{\partial \theta}\right)\left(\frac{\frac{\partial}{\partial\theta} L(\theta)}{L(\theta)}\right) p_\theta(x)\,dx$$

$$= \int \left(\frac{\partial \log L(\theta)}{\partial \theta}\right)\left(\frac{\partial}{\partial\theta} L(\theta|x)\right) dx$$

integration by parts $=$ $\frac{\partial \log L(\theta)}{\partial\theta}\left(\frac{\partial}{\partial\theta}\int p_\theta(x)\,dx\right) - \int \frac{\partial^2 \log L(\theta|x)}{\partial\theta^2} p_\theta(x)\,dx$

$= 0$

$$\Rightarrow \mathrm{Var}_\theta S(\theta) = -\int \frac{\partial^2 \log L(\theta|x)}{\partial\theta^2} p_\theta(x)\,dx \overset{\text{definition}}{\equiv} I_x(\theta)$$

the expected Fisher information.

So, estimate $\hat{\theta} = \{\theta : S(\theta) = 0\}$

$\Rightarrow$ $\hat{\theta}$ inherits properties of $S(\theta)$ — mean 0, variance $FI(\theta)$

$\Downarrow$ — mean, variance for θ

(more later)

$\hat{\theta} \sim N(\theta, FI(\theta)^{-1})$

Example normal mean model

$S(\theta) = \frac{n}{\sigma^2}(\bar{x} - \theta)$ (σ^2 known)

$\Downarrow$

$\Rightarrow \begin{cases} E(S(\theta)) = 0 \\ V(S(\theta)) = \frac{n}{\sigma^2} \end{cases}$

and $I(\theta) = -\frac{\partial}{\partial\theta} S(\theta) = \frac{n}{\sigma^2} \Rightarrow FI(\theta) = \frac{n}{\sigma^2}$

Here, $I(\theta) = FI(\theta)$ happens for exponential families on canonical form.

Exponential families

$S(\theta) = \eta'(\theta) T(x) - A'(\theta)$

$\Rightarrow \hat{\theta} = \{\theta : \eta'(\theta) T(x) = A'(\theta)\}$

and we have

$E(S(\theta)) = \eta'(\theta) E(T(x)) - A'(\theta) = 0$

Now, $I(\theta) = -\frac{\partial^2}{\partial\theta^2} \log L(\theta) = A''(\theta) - \eta''(\theta) T(x)$

and $\mathcal{I}(\theta) = E(\quad) = A''(\theta) - \eta''(\theta) E(T(x))$

At solution to score-equation $\hat{\theta}$ (function of data $\hat{\theta} = \vec{\theta}(T(x))$)

$I(\hat{\theta}) = A''(\hat{\theta}) - \eta''(\hat{\theta}) T(x)$ where we know $\eta'(\hat{\theta}) T(x) = A'(\hat{\theta})$

and

$\mathcal{I}(\hat{\theta}) = A''(\hat{\theta}) - \eta''(\hat{\theta}) E(T(x))$ where we know $\eta'(\theta) E_\theta(T(x)) = A'(\theta)$ all θ.

function $\mathcal{I}$ evaluated at $\hat{\theta}$ value

$\Rightarrow$ $\underline{\underline{I(\hat{\theta}) = \mathcal{I}(\hat{\theta})}}$

so observed Fisher information coincides with the expected evaluated at the value $\hat{\theta}$

If we use canonical form we get an even stronger result

$I(\theta) = A'(\theta)$ so $I(\theta) = \mathcal{I}(\theta)$ functions coincide

$\mathcal{I}(\theta) = A''(\theta)$

Interesting result. $I(\theta)$ - curvature obtained from actual sample $\underline{x}$

→ uncertainty about estimating θ in this sample

$\mathcal{I}(\theta)$ is a long-run frequency property about how hard estimation of θ is.

Turns out $FI(\theta)$ has another important meaning.

(Cramér-Rao lower bound

- Let's say θ estimator $T(x)$ is unbiased such that $E_\theta T(x) = \theta$
- Then $V_\theta(T(x)) \geq \frac{1}{FI(\theta)}$

That is, we can never do better than $FI(\theta)^{-1}$ in terms of estimation efficiency

We can meet this bound for exponential family models!