Definition 6.11 (score equation, score function)

- The log likelihood of an i.i.d. sample of size n from a pdf/pmf $f(x|\theta)$ is
$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$
score function of n obs. = sum of score function of ith obs

(Exercise) Check the previous MLE examples and identify the score equation & score function.

- The MLE maximizes $l(\theta)$ and is usually obtained by solving the **score equation**
$$0 = \frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta).$$
sum of scores = 0 at $\hat{\theta}_{MLE}$

Score of ith observation X_i

- $\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$ is called **score function**.

a function of observations X_i's and parameter θ

3/18

Definition 6.12 (Fisher information for one-dimensional parameter, TBp.263)

Let $X_1, \ldots, X_n$ be a sample of size n with a joint pdf/pmf f. Define

This is a function of unknown parameter θ

can allow not to be i.i.d.

average (w.r.t. the parameter θ)

$$I_{X_1,\ldots,X_n}(\theta) = \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1,\cdots,X_n|\theta)\right]^2$$

$\frac{\partial}{\partial\theta}\log f(X|\theta)$ is a r.v., but not statistic

treated as random variable

weighted average of score^2 → always ≥ 0

which is called the **(Fisher) information** of θ contained in $X_1, \ldots, X_n$.

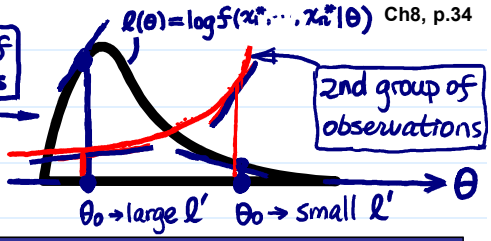
Question: What *information* does $I_{X_1,\ldots,X_n}(\theta)$ offer?

$$I_{X_1,\ldots,X_n}(\theta) = \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1,\cdots,X_n|\theta)\right]^2$$

score function

log-likelihood

observations: fixed → random



Theorem 6.4 (TBp. 276)

Let $X_1, \ldots, X_n$ be an i.i.d. sample of size n from a pdf/pmf $f(x|\theta)$.

$Y_1, \ldots, Y_n$ indep.

$$I_{X_1,\cdots,X_n}(\theta) = \mathrm{E}_\theta\left(\frac{\partial}{\partial\theta}\log\left[\prod_{i=1}^{n} f(X_i|\theta)\right]\right)^2 = \mathrm{E}_\theta\left[\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]^2$$

score function of ith observation

Y_i

$\because$ independent

$$= \sum_{i=1}^{n}\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]^2 + 2\sum_{i<j}\underbrace{\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta)\right]\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_j|\theta)\right]}_{=0,\ \text{by } (\Delta) \text{ in } E(Y_iY_j)=E(Y_i)E(Y_j)\ \text{LNp.35}}$$

$\because$ identical

$$= n\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 \equiv n\cdot I_{X_1}(\theta)$$

- $I_{X_1}(\theta)$ is the Fisher information contained in a sample of size one.
- $nI_{X_1}(\theta)$: interpreted as the information of θ contained in a sample of size n from $f(\cdot|\theta)$.

larger sample size, more information.

more information

X, Y indep $\Rightarrow I_{X,Y}(\theta) = I_X(\theta) + I_Y(\theta)$

$\mathrm{cor}(X,Y)=0$ ⟷ $\mathrm{cor}(X,Y)=\pm 1$ ($Y=aX+b$)

less information

- The Fisher informations of independent samples are additive.

Theorem 6.4 (TBp. 276)

Under appropriate smoothness conditions on f, (pdf/pmf)

$$I_{X_1}(\theta) \equiv \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 = Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right]$$

(cf. same r.v. Y_1)

Proof (for pdf case): Since $\int f(x|\theta)\,dx = 1$ for all θ,

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x|\theta)\,dx = \int \left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] f(x|\theta)\,dx \\
&= \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]. \quad \cdots(\Delta) \\
&\Rightarrow Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right] \quad (Y_1) \\
&= \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 - \left\{E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]\right\}^2 \quad (Y_1) \\
&= \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2.
\end{aligned}$$

(score function of one observation ($\Rightarrow$ average of scores $=0$))

$$\begin{aligned}
0 &= \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)\,dx = \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\
&= \int \left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] f(x|\theta)\,dx + \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)\,dx \\
&= E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right] + \underbrace{E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2}_{I_{X_1}(\theta)}.
\end{aligned}$$

(need smoothness of f for interchanging integration and differentiation.)

Note (TBp. 278, 279).

For i.i.d. case, (l: log-likelihood of n observations)

$$E_\theta[l'(\theta)^2] = I_{X_1,\ldots,X_n}(\theta) = n\cdot I_{X_1}(\theta) = -n\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right]$$

(same θ; use joint distribution of $X_1,\cdots,X_n$)

$$\overset{\because \text{identical}}{=} -\sum_{i=1}^n \mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^n \log f(X_i|\theta)\right] = -E_\theta[l''(\theta)]$$

(log-likelihood of $X_1,\cdots,X_n$)

- **interpretation**: when $|E_\theta[l''(\theta)]|$ is large at $\theta=\theta_0$, $l(\theta)$ is, on average, changing rapidly in a vicnity of θ_0

Example 6.18 (Fisher information of i.i.d. Bernoulli $B(\theta)$)

Let $X_1,\ldots,X_n$ be i.i.d. from Bernoulli distribution $B(\theta)$ (i.e., the pmf of X_i is, $\theta^{x_i}(1-\theta)^{1-x_i}$, for $x_i\in\{0,1\}$), then $E(X_i)=\theta$ and $Var(X_i)=\theta(1-\theta)$.

(statistical modeling (e.g., toss a coin); joint pmf $\propto \theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}$)

- For a single observatoin X_i, the first and second derivatives of its log likelihood are:

($\log(\theta)$: $\log(\theta)$, when $x=1$; $\log(1-\theta)$, when $x=0$)

$$\begin{aligned}
\log f(x|\theta) &= x\log\theta + (1-x)\log(1-\theta), \\
\partial/\partial\theta \log f(x|\theta) &= x/\theta - (1-x)/(1-\theta) = (x-\theta)/(\theta(1-\theta)), \\
\partial^2/\partial\theta^2 \log f(x|\theta) &= -x/\theta^2 - (1-x)/(1-\theta)^2.
\end{aligned}$$

(x: fixed; score; $1/\theta$, when $x=1$; $-1/(1-\theta)$, when $x=0$)

- The Fisher information of a single observation, say X_1, is

X: random

$$I_{X_1}(\theta) \overset{\text{by definition 6.12 LNp.33}}{=} E_\theta\left[\frac{X_1-\theta}{\theta(1-\theta)}\right]^2 = \frac{E_\theta[(X_1-\theta)^2]}{\theta^2(1-\theta)^2} = \frac{Var_\theta(X_1)}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}.$$

$I_{X_1}(\theta) = Var_\theta\left[\frac{X_1-\theta}{\theta(1-\theta)}\right]$ (Ec) $\to \frac{1}{\theta(1-\theta)}$

$$I_{X_1}(\theta) \overset{\text{by Thm. 6.4 LNp.35}}{=} -E_\theta\left[-\frac{X_1}{\theta^2} - \frac{1-X_1}{(1-\theta)^2}\right] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

- The Fisher information of observations $X_1, \ldots, X_n$ is

(Exercise) Use joint pmf to obtain

$$I_{X_1,\cdots,X_n}(\theta) = nI_{X_1}(\theta) = \frac{n}{\theta(1-\theta)}.$$

$(X_1, \cdots, X_n) \to Y = \sum X_i$ ✗

Notice that $I_{X_1,\cdots,X_n}(\theta)$

$\bar{X}$ estimate θ; $Var(\bar{X}) = \frac{\theta(1-\theta)}{n}$; • large when $\theta \approx 1/2$; • small when $\theta \approx 0$ or 1

- increases when n increases, ← larger sample size, more information
- increases when $\theta \downarrow 0$ or $\theta \uparrow 1$, ← more information when θ is close to 0 or 1
- reaches a minimum $4n$ at $\theta = 0.5$. ← minimum information when $\theta = 1/2$

- Consider a single observation $Y \sim$ Binomial(n, θ). The pmf of Y is

$X_1, \cdots, X_n \to$ non-invertible transformation $\to \sum X_i$

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \text{for } y \in \{0, 1, \ldots, n\}.$$

y: fixed

The second derative of log likelihood is

$$\partial^2 \log f(y|\theta)/\partial^2\theta = -y/\theta^2 - (n-y)/(1-\theta)^2.$$

check sufficiet statistic in Ex.b.21 LNp.48

Y: random

$(X_1, \cdots, X_n)$ & $Y = \sum X_i$ carry same amount of information

The Fisher information of Y, is

$$I_Y(\theta) = -E_\theta\left[-\frac{Y}{\theta^2} - \frac{n-Y}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

Note that $I_Y(\theta)$ is the same as $I_{X_1,\cdots,X_n}(\theta)$. ← reasonable?

3/20

Theorem 6.5 (consistency of MLE, TBp. 275)

Under appropriate smoothness conditions of f, the MLE from an i.i.d. sample is consistent. → $\hat{\theta}_{MLE} \xrightarrow{P} \theta_0$ ← more general than LLN

$X_1, \cdots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$

Proof (sketch, for pdf case): Denote the true value of θ by θ_0. The MLE maximizes $\frac{l(\theta)}{n} \overset{(*)}{=} \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$.

Y_i

$Y_1, \cdots, Y_n$ are i.i.d.

What is the difference between $\int [\log f(x|\theta)] f(x|\theta_0)dx$ and $\int [\log f(x|\theta)] f(x|\theta)dx$?

The weak law of large numbers implies

function of $\underset{\sim}{X}$ & θ → $\forall\, \theta$, $\frac{l(\theta)}{n} \overset{(*)}{\xrightarrow{P}} E_{\theta_0}[\log f(X|\theta)] = \int [\log f(x|\theta)] f(x|\theta_0)\, dx$ as $n \to \infty$.

$E_{\theta_0}[\log f(X|\theta)]$: a function of θ; $\log f(x|\theta)$: variable; $f(x|\theta_0)$: fixed