Take-away from "Intuition about Fisher Information":

$$\frac{d^2}{d\theta} \ell(\theta|x_1,\ldots,x_n)\Big|_{\theta=\hat{\theta}^{MLE}}$$ large in magnitude

$\iff \ell''(\theta|x_1,\ldots,x_n)$ very curved at $\hat{\theta}^{MLE}$

$\iff \ell(\theta|x_1,\ldots,x_n)$ changes quickly as we move away from $\hat{\theta}^{MLE}$

$\iff$ ~~a~~ the data provide a lot of "information" that $\hat{\theta}^{MLE}$ is better than other possibilities

---

Consider the form of $\frac{d^2}{d\theta} \ell(\theta|x_1,\ldots,x_n)$:

$$\frac{d^2}{d\theta^2} \ell(\theta|x_1,\ldots,x_n) = \frac{d^2}{d\theta^2} \log\left[\mathcal{L}(\theta|x_1,\ldots,x_n)\right]$$

$$= \frac{d^2}{d\theta^2} \log\left[\prod_{i=1}^{n} f(x_i|\theta)\right]$$

$$= \frac{d^2}{d\theta^2} \sum_{i=1}^{n} \log\{f(x_i|\theta)\}$$

$$= \underbrace{\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log\{f(x_i|\theta)\}}$$

overall magnitude of $\frac{d^2}{d\theta^2} \ell(\theta|x_1,\ldots,x_n)$ grows with the sample size.

- Subset 1 had a sample size of 56, more information about λ.
- Subset 2 had $n = 4$, less information about λ.

Note: $\frac{d^2}{d\theta^2} \ell(\hat{\theta}^{MLE} | x_1, \ldots, x_n) < 0$ ($\hat{\theta}^{MLE}$ maximizes $\ell(\theta)$)

$-\frac{d^2}{d\theta^2} \ell(\hat{\theta}^{MLE} | x_1, \ldots, x_n) > 0$ has a more intuitive sign: larger value <=> more information.

Def. (to be refined): The observed Fisher information about a parameter θ is

$$J(\theta^*) = -\frac{d^2}{d\theta^2} \ell(\theta | x_1, \ldots, x_n) \Big|_{\theta = \theta^*}$$

Note: Most often evaluated at the MLE:

$$J(\hat{\theta}^{MLE}) = -\frac{d^2}{d\theta^2} \ell(\theta | x_1, \ldots, x_n) \Big|_{\theta = \hat{\theta}^{MLE}}$$

The observed Fisher information from one observation $X_i = x_i$ is

$$J_i(\theta^*) = -\frac{d^2}{d\theta^2} \ell(\theta | x_i) \Big|_{\theta = \theta^*} = \frac{d^2}{d\theta^2} \log\{f_{X_i}(x_i | \theta)\} \Big|_{\theta = \hat{\theta}^{MLE}}$$

If observations independent, $J(\theta^*) = \sum_{i=1}^{n} J_i(\theta^*)$

Ex.: Suppose $X_1, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda)$.

$E(X_i) = \lambda$, MLE is $\hat{\lambda}^{MLE} = \bar{X}$, $f_{X_i}(x_i | \lambda) = e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}$

Find the observed Fisher information about λ if we observe $x_1, \ldots, x_n$.

$$J_i(\lambda^*) = -\frac{d^2}{d\lambda^2} \log\{f_{X_i}(x_i | \lambda)\} \Big|_{\lambda = \lambda^*} = -\frac{d^2}{d\lambda^2} \log\left[e^{-\lambda} \cdot \frac{\lambda^{x_i}}{x_i!}\right] \Big|_{\lambda = \lambda^*}$$

$$= -\frac{d^2}{d\lambda^2} \left[-\lambda + x_i \log(\lambda) - \log(x_i!)\right] \Big|_{\lambda = \lambda^*}$$

$$= -\frac{d}{d\lambda} \left[-1 + \frac{x_i}{\lambda}\right] \Big|_{\lambda = \lambda^*}$$

$$= + \frac{x_i}{\lambda^2}$$

$$J(\lambda) = \sum_{i=1}^{n} J_i(\lambda) = \sum_{i=1}^{n} \frac{x_i}{\lambda^2} = \frac{1}{\lambda^2} \cdot \sum_{i=1}^{n} x_i = \frac{1}{\lambda^2} \cdot n \cdot \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{n}{\lambda^2} \bar{x}$$

Evaluate at $\hat{\lambda}^{MLE} = \bar{x}$:

$$J(\hat{\lambda}^{MLE}) = J(\bar{x}) = \frac{n}{(\bar{x})^2} \cdot \bar{x} = \frac{n}{\bar{x}}$$

Def.: The Fisher information is the expected value of the observed Fisher information.

$$I(\theta^*) = -E\left[\frac{d^2}{d\theta^2} \ell(\theta \mid X_1, \ldots, X_n)\Big|_{\theta=\theta^*}\right]$$

capital X's

- on average across all samples of size n, what is the curvature of the log-likelihood function at the parameter value θ^*?

Def.: The Fisher information from one observation is

$$I_i(\theta^*) = -E\left[\frac{d^2}{d\theta^2} \ell(\theta \mid X_i)\Big|_{\theta=\theta^*}\right]$$

↑ this is the closest our book comes to defining Fisher information, p. 276

Note: $I(\theta^*) = \sum_{i=1}^{n} I_i(\theta^*)$

Actual Definition of Fisher information

$$I(\theta^*) = E\left[\left\{\frac{d}{d\theta} \ell(\theta \mid X_1, \ldots X_n)\right\}^2\right] = E\left[\left\{\frac{d}{d\theta} \sum_{i=1}^{n} \log(f_X(x_i \mid \theta))\right\}^2\right]$$

Thm.: If $f_X(x \mid \theta)$ satisfies "regularity conditions"
↳ twice continuously differentiable, support of f_X doesn't depend on θ, other stuff, ...

then the two definitions are equivalent:

$$E\left[\left\{\frac{d}{d\theta} \ell(\theta \mid X_1, \ldots, X_n)\right\}^2\right] = -E\left[\frac{d^2}{d\theta^2} \ell(\theta \mid X_i)\Big|_{\theta=\theta^*}\right]$$

Pf: Enough to show for $n=1$.

Two preliminary claims:

1) $\frac{d}{d\theta}\int_{-\infty}^{\infty} f_X(x|\theta)\,dx = \frac{d}{d\theta} 1 = 0$

2) $\frac{d}{d\theta}\log\{f_X(x|\theta)\} = \frac{1}{f_X(x|\theta)}\cdot\frac{d}{d\theta} f_X(x|\theta)$

$$\Rightarrow \frac{d}{d\theta} f_X(x|\theta) = \left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]\cdot f_X(x|\theta)$$

We get:

$0 = \frac{d}{d\theta}\int_{-\infty}^{\infty} f_X(x|\theta)\,dx$ (claim 1)

$= \int \frac{d}{d\theta} f_X(x|\theta)\,dx$ (if regularity conditions satisfied)

$= \int \left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]\cdot f_X(x|\theta)\,dx$ (claim 2)

Now take second derivative wrt θ:

$0 = \frac{d}{d\theta}\int \left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]\cdot f_X(x|\theta)\,dx$

$= \int \frac{d}{d\theta}\left(\left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]\cdot f_X(x|\theta)\right)dx$ (regularity conditions)

$= \int\left[\frac{d^2}{d\theta^2}\log\{f_X(x|\theta)\}\right] f_X(x|\theta)\,dx + \int\left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]\cdot\left(\frac{d}{d\theta} f_X(x|\theta)\right)dx$ (product rule)

$= \int\left[\frac{d^2}{d\theta^2}\log\{f_X(x|\theta)\}\right] f_X(x|\theta)\,dx + \int\left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]^2\cdot f_X(x|\theta)\,dx$ (claim 2)

Rearrange:

$$\int\left[\frac{d}{d\theta}\log\{f_X(x|\theta)\}\right]^2\cdot f_X(x|\theta)\,dx = -\int\left[\frac{d^2}{d\theta^2}\log\{f_X(x|\theta)\}\right] f_X(x|\theta)\,dx$$

$$\Leftrightarrow E\left[\left\{\frac{d}{d\theta}\ell(\theta|X)\right\}^2\right] = -E\left[\frac{d^2}{d\theta^2}\ell(\theta|X)\right]$$

:)