April 6, 2016 Debdeep Pati

1 Fisher Information

Assume $X \sim f(x \mid \theta)$ (pdf or pmf) with $\theta \in \Theta \subset \mathbb{R}$. Define

$$I_X(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right]$$

where $\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)$ is the derivative of the log-likelihood function evaluated at the true value θ. Fisher information is meaningful for families of distribution which are regular:

1. Fixed support: $\{x : f(x \mid \theta) > 0\}$ is the same for all θ.
2. $\frac{\partial}{\partial\theta}\log f(x \mid \theta)$ must exist and be finite for all x and θ.
3. If $E_\theta|W(X)| < \infty$ for all θ, then

$$\left(\frac{\partial}{\partial\theta}\right)^k E_\theta W(X) = \left(\frac{\partial}{\partial\theta}\right)^k \int W(x) f(x \mid \theta)dx = \int W(x)\left(\frac{\partial}{\partial\theta}\right)^k f(x \mid \theta)dx$$

1.1 Regular families

One parameter exponential families: Cauchy location or scale family:

$$f(x \mid \theta) = \frac{1}{\pi(1 + (x-\theta)^2)}$$
$$f(x \mid \theta) = \frac{1}{\pi\theta(1 + (x/\theta)^2)}$$

and lots more. (Most families of distributions used in applications are regular).

1.2 Non-regular families

$$\text{Uniform}(0, \theta)$$
$$\text{Uniform}(\theta, \theta + 1).$$

1.3 Facts about Fisher Information

Assume a regular family.

1. $$E_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) = 0.$$

   Here $\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)$ is called the "score" function $S(\theta)$.

   Proof.

   $$\begin{aligned} E_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) &= \int\left(\frac{\partial}{\partial\theta}\log f(x \mid \theta)\right) f(x \mid \theta)dx \\ &= \int \frac{\frac{\partial}{\partial\theta} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)dx \\ &= \int \frac{\partial}{\partial\theta} f(x \mid \theta)dx \\ &= \frac{\partial}{\partial\theta}\int f(x \mid \theta)dx = 0 \end{aligned}$$

   since $\int f(x \mid \theta)dx = 1$ for all θ. □

2. $I_X(\theta) = \mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)$.

   Proof. Since $E_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) = 0$

   $$\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right) = E_\theta\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2 = I_X(\theta).$$

   □

3. If $X = (X_1, X_2, \ldots, X_n)$ and $X_1, X_2, \ldots, X_n$ are independent random variables, then $I_X(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta) + \cdots I_{X_n}(\theta)$.

   Proof. Note that

   $$f(x \mid \theta) = \prod_{i=1}^n f_i(x_i \mid \theta)$$

where $f_i(\cdot \mid \theta)$ is the pdf (pmf) of X_i. Observe that

$$\frac{\partial}{\partial\theta} \log f(X \mid \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_i(X_i \mid \theta)$$

and the random variables in the sum are independent. This

$$\mathrm{Var}\left[\frac{\partial}{\partial\theta} \log f(X \mid \theta)\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\frac{\partial}{\partial\theta} \log f_i(X_i \mid \theta)\right]$$

so that $I_X(\theta) = \sum_{i=1}^{n} I_{X_i}(\theta)$ by 2. $\square$

4. If $X_1, X_2, \ldots, X_n$ are i.i.d and $X = (X_1, X_2, \ldots, X_n)$, then $I_{X_i}(\theta) = I_{X_1}(\theta)$ for all i so that $I_X(\theta) = nI_{X_1}(\theta)$.

5. An alternate formula for Fisher information is

$$I_X(\theta) = E_\theta\left(- \frac{\partial^2}{\partial\theta^2} \log f(X \mid \theta)\right)$$

Proof. Abbreviate $\int f(x \mid \theta)dx$ as $\int f$, etc. Since $1 = \int f$, applying $\frac{\partial}{\partial\theta}$ to both sides,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int f = \int \frac{\partial f}{\partial\theta} = \int \frac{\frac{\partial}{\partial\theta}}{f} \cdot f \\ &= \int \left(\frac{\partial}{\partial\theta} \log f\right) f. \end{aligned}$$

Applying $\frac{\partial}{\partial\theta}$ again,

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \int \left(\frac{\partial}{\partial\theta} \log f\right) f \\ &= \int \frac{\partial}{\partial\theta}\left[\left(\frac{\partial}{\partial\theta} \log f\right) f\right] \\ &= \int \left(\frac{\partial^2}{\partial\theta^2} \log f\right) \cdot f + \int \left(\frac{\partial}{\partial\theta} \log f\right) \frac{\partial f}{\partial\theta} \end{aligned}$$

Noting that

$$\begin{aligned} \frac{\partial f}{\partial\theta} &= \frac{\frac{\partial f}{\partial\theta}}{f} \cdot f, \\ &= \left(\frac{\partial}{\partial\theta} \log f\right) f, \end{aligned}$$

this becomes

$$0 = \int \left(\frac{\partial^2}{\partial \theta^2} \log f \right) \cdot f + \int \left(\frac{\partial}{\partial \theta} \log f \right)^2 \cdot f$$

or

$$0 = E\left(\frac{\partial^2}{\partial \theta^2} \log f(X \mid \theta) \right) + I_X(\theta).$$

$\square$

Example: Fisher Information for a Poisson sample. Observe $\underset{\sim}{\mathbf{X}} = (X_1, \ldots, X_n)$ iid Poisson(λ). Find $I_{\underset{\sim}{\mathbf{X}}}(\lambda)$. We know $I_{\underset{\sim}{\mathbf{X}}}(\lambda) = nI_{X_1}(\lambda)$. We shall calculate $I_{X_1}(\lambda)$ in three ways. Let $X = X_1$. Preliminaries:

$$\begin{aligned}
f(x \mid \lambda) &= \frac{\lambda^x e^{-\lambda}}{x!} \\
\log f(x \mid \lambda) &= x \log \lambda - \lambda - \log x! \\
\frac{\partial}{\partial \lambda} \log f(x \mid \lambda) &= \frac{x}{\lambda} - 1 \\
-\frac{\partial^2}{\partial \lambda^2} \log f(x \mid \lambda) &= \frac{x}{\lambda^2}
\end{aligned}$$

Method #1: Observe that

$$\begin{aligned}
I_X(\lambda) &= E_\lambda\left[\left(\frac{\partial}{\partial \lambda} \log f(X \mid \lambda) \right)^2 \right] = E_\lambda\left[\left(\frac{X}{\lambda} - 1 \right)^2 \right] \\
&= \mathrm{Var}_\lambda\left(\frac{X}{\lambda} \right) (\text{since} E\left(\frac{X}{\lambda} \right) = \frac{EX}{\lambda} = 1) \\
&= \frac{\mathrm{Var}(X)}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}
\end{aligned}$$

Method #2: Observe that

$$\begin{aligned}
I_X(\lambda) &= \mathrm{Var}_\lambda\left(\frac{\partial}{\partial \lambda} \log f(X \mid \lambda) \right) = \mathrm{Var}\left(\frac{X}{\lambda} - 1 \right) \\
&= \mathrm{Var}\left(\frac{X}{\lambda} \right) = \frac{1}{\lambda} (\text{as in Method\#1}).
\end{aligned}$$

Method #3: Observe that

$$I_X(\lambda) = E_\lambda\left(-\frac{\partial^2}{\partial \lambda^2} \log f(X \mid \lambda) \right) = E_\lambda\left(\frac{X}{\lambda^2} \right) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}.$$

Thus $I_{\underset{\sim}{\mathbf{X}}}(\lambda) = nI_{X_1}(\lambda) = \frac{n}{\lambda}$.

<u>Example:</u> Fisher information for Cauchy location family. Suppose $X_1, X_2, \ldots, X_n$ iid with pdf

$$f(x \mid \theta) = \frac{1}{\pi(1 + (x-\theta)^2)}.$$

Let $\underset{\sim}{X} = (X_1, \ldots, X_n), X \sim f(x \mid \theta)$. Find $I_{\underset{\sim}{X}}(\theta)$.

Note that $I_{\underset{\sim}{X}}(\theta) = nI_{X_1}(\theta) = nI_X(\theta)$. Now

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log f(x \mid \theta) &= \frac{\frac{\partial f}{\partial \theta}}{f} \\
&= \frac{\frac{-1}{\pi(1+(x-\theta)^2)^2} \cdot 2(x-\theta)(-1)}{\frac{1}{\pi(1+(x-\theta)^2)}} \\
&= \frac{2(x-\theta)}{(1+(x-\theta)^2)}
\end{aligned}$$

Now

$$\begin{aligned}
I_X(\theta) &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right)^2\right] \\
&= E\left(\frac{2(X-\theta)}{1+(X-\theta)^2}\right)^2 \\
&= \int_{-\infty}^{\infty}\left(\frac{2(x-\theta)}{1+(x-\theta)^2}\right)^2 \frac{1}{\pi(1+(x-\theta)^2)}dx \\
&= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{(x-\theta)^2}{(1+(x-\theta)^2)^3}dx.
\end{aligned}$$

Letting $u = x - \theta, du = dx$,

$$\begin{aligned}
I_X(\theta) &= \frac{4}{\pi}\int_{-\infty}^{\infty}\frac{u^2}{(1+u^2)^3}du \\
&= \frac{8}{\pi}\int_{0}^{\infty}\frac{u^2}{(1+u^2)^3}du.
\end{aligned}$$

Substituting $x = 1/(1+u^2)$, $u = (1/x-1)^{1/2}, du = 0.5(1/x-1)^{-1/2}(-1/x^2)dx$,

$$\begin{aligned}
I_X(\theta) &= \frac{8}{\pi}\int_0^\infty \frac{u^2}{(1+u^2)^3}du \\
&= \frac{8}{\pi}\int_0^\infty \frac{u^2}{(1+u^2)}\left(\frac{1}{1+u^2}\right)^2 du \\
&= \frac{8}{\pi}\int_0^1 (1-x)x^2 \cdot (1/2)(1/x-1)^{-1/2}(1/x^2)dx \\
&= \frac{4}{\pi}\int_0^1 x^{1/2}(1-x)^{1/2}dx \\
&= \frac{4}{\pi}\int_0^1 x^{3/2-1}(1-x)^{3/2-1}dx \quad \text{(Beta integral)} \\
&= \frac{4}{\pi}\frac{\Gamma(3/2)\Gamma(3/2)}{\Gamma(3/2+3/2)} = \frac{4}{\pi}\frac{(0.5\sqrt{\pi})^2}{2!} \\
&= \frac{1}{2}.
\end{aligned}$$

Hence $I_{\underset{\sim}{X}}(\theta) = n/2$.

2 Uses of Fisher Information

- Asymptotic distribution of MLE's
- Cramér-Rao Inequality (Information inequality)

2.1 Asymptotic distribution of MLE's

- **i.i.d case:**
  If $f(x \mid \theta)$ is a regular one-parameter family of pdf's (or pmf's) and $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ is the MLE based on $\mathbf{X}_n = (X_1, \ldots, X_n)$ where n is large and $X_1, \ldots, X_n$ are iid from $f(x \mid \theta)$, then approximately,

  $$\hat{\theta}_n \sim \mathrm{N}\left(\theta, \frac{1}{nI(\theta)}\right)$$

  where $I(\theta) \equiv I_{X_1}(\theta)$ and θ is the true value. Note that $nI(\theta) = I_{\mathbf{X}_n}(\theta)$. More formally,

  $$\frac{\hat{\theta}_n - \theta}{\sqrt{\frac{1}{nI(\theta)}}} = \sqrt{nI(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathrm{N}(0,1)$$

as $n \to \infty$.

- **More general case:** (Assuming various regularity conditions) If $f(\underset{\sim}{x} \mid \theta)$ is a one-parameter family of joint pdf's (or joint pmf's) for data $\mathbf{X}_n = (X_1, \ldots, X_n)$ where n is large (think of a large dataset arising from regression or time series model) and $\hat{\theta}_n = \hat{\theta}_n(\mathbf{X}_n)$ is the MLE, then

$$\hat{\theta}_n \sim \mathrm{N}\bigg(\theta, \frac{1}{I_{\mathbf{X}_n}(\theta)}\bigg)$$

where θ is the true value.

2.2 Estimation of the Fisher Information

If θ is unknown, then so is $I_{\mathbf{X}}(\theta)$. Two estimates $\hat{I}$ of the Fisher information $I_{\mathbf{X}}(\theta)$ are

$$\hat{I}_1 = I_{\mathbf{X}}(\hat{\theta}), \quad \hat{I}_2 = -\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{X} \mid \theta)|_{\theta=\hat{\theta}}$$

where $\hat{\theta}$ is the MLE of θ based on the data $\mathbf{X}$. $\hat{I}_1$ is the obvious plug-in estimator. It can be difficult to compute $I_X(\theta)$ does not have a known closed form. The estimator $\hat{I}_2$ is suggested by the formula

$$I_{\mathbf{X}}(\theta) = E\bigg(-\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{X} \mid \theta)\bigg)$$

It is often easy to compute, and is required in many Newton- Raphson style algorithms for finding the MLE (so that it is already available without extra computation). The two estimates $\hat{I}_1$ and $\hat{I}_2$ are often referred to as the "expected" and "observed" Fisher information, respectively.
As $n \to 1$, both estimators are consistent (after normalization) for $I_{\mathbf{X}_n}(\theta)$ under various regularity conditions.
For example: in the iid case: $\hat{I}_1/n, \hat{I}_2/n$, and $I_{\mathbf{X}_n}(\theta)/n$ all converge to $I(\theta) \equiv I_{X_1}(\theta)$.

2.3 Approximate Confidence Intervals for θ

Choose $0 < \alpha < 1$ (say, $\alpha = 0.05$). Let z^* be such that

$$P(-z^* < Z < z^*) = 1 - \alpha$$

where $Z \sim \mathrm{N}(0, 1)$. When n is large, we have approximately

$$\sqrt{I_{\mathbf{X}}(\theta)}(\hat{\theta} - \theta) \sim \mathrm{N}(0, 1)$$

so that

$$P\left\{-z^* < \sqrt{I_{\mathbf{X}}(\theta)}(\hat{\theta}-\theta) < z^*\right\} \approx 1-\alpha$$

or equivalently,

$$P\left\{\hat{\theta} - z^*\sqrt{\frac{1}{I_{\mathbf{X}}(\theta)}} < \theta < \hat{\theta} + z^*\sqrt{\frac{1}{I_{\mathbf{X}}(\theta)}}\right\} \approx 1-\alpha.$$

This approximation continues to hold when $I_{\mathbf{X}}(\theta)$ is replaced by an estimate $\hat{I}$ (either $\hat{I}_1$ or $\hat{I}_2$):

$$P\left\{\hat{\theta} - z^*\sqrt{\frac{1}{\hat{I}}} < \theta < \hat{\theta} + z^*\sqrt{\frac{1}{\hat{I}}}\right\} \approx 1-\alpha.$$

Thus

$$\left(\hat{\theta} - z^*\sqrt{\frac{1}{\hat{I}}}, \hat{\theta} + z^*\sqrt{\frac{1}{\hat{I}}}\right)$$

is an approximate $1-\alpha$ confidence interval for θ. (Here $\hat{\theta}$ is the MLE and $\hat{I}$ is an estimate of the Fisher information.)

3 Cramer-Rao Inequality

Let $\underset{\sim}{X} \sim P_\theta, \theta \in \Theta \subset \mathbb{R}$.

Theorem 1. *If $f(\underset{\sim}{x} \mid \theta)$ is a regular one-parameter family, $E_\theta W(\underset{\sim}{X}) = \tau(\theta)$ for all θ, and $\tau(\theta)$ is differentiable, then*

$$Var_\theta(W(\underset{\sim}{X})) \geq \frac{\{\tau'(\theta)\}^2}{I_{\underset{\sim}{X}}(\theta)}.$$

Proof. Preliminary Facts:

A. $[Cov(X,Y)]^2 \leq (\text{Var}X)(\text{Var}Y)$. This is a special case of the Cauchy-Schwarz inequality. It is better known to statisticians as $\rho^2 \leq 1$ where

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\cdot\text{Var}(Y)}}$$

is the correlation between X and Y.

B. $\text{Cov}(X, Y) = EXY$ if wither $EX = 0$ or $EY = 0$. This follows from the well-known formula.

$$\text{Cov}(X, Y) = EXY - (EX)(EY).$$

Since $E_\theta \frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta) = 0$, from **B**, we have

$$\begin{aligned}
[\text{Cov}_\theta(W(\underset{\sim}{X}), \frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta)] &= E\left[W(\underset{\sim}{X})\frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta)\right] \\
&= \int W(\underset{\sim}{x})\left(\frac{\partial}{\partial\theta} \log f(\underset{\sim}{x} \mid \theta)\right) f(\underset{\sim}{x} \mid \theta) d\underset{\sim}{x} \\
&= \int W(\underset{\sim}{x}) \frac{\partial f(\underset{\sim}{x} \mid \theta)}{\partial\theta} d\underset{\sim}{x} \\
&= \frac{\partial}{\partial\theta} \int W(\underset{\sim}{x}) f(\underset{\sim}{x} \mid \theta) d\underset{\sim}{x} \quad (\text{since } f(\underset{\sim}{x} \mid \theta) \text{ is a regular family}) \\
&= \frac{\partial}{\partial\theta} E_\theta W(\underset{\sim}{X}) = \tau'(\theta).
\end{aligned}$$

Since from **A.**, we have

$$[\text{Cov}_\theta(W(\underset{\sim}{X}), \frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta)]^2 \leq \text{Var} W(\underset{\sim}{X}) \text{Var}\left(\frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta)\right),$$

$$[\tau'(\theta)]^2 \leq \text{Var}_\theta W(\underset{\sim}{X}) I_{\underset{\sim}{X}}(\theta).$$

$\square$

Remark 1. *Equality in* **A**. *is achieved iff*

$$Y = aX + b$$

for some constants a, b. Moreover, if $EY = 0$, then $E(aX + b) = 0$ forces $b = -aEX$ so that

$$Y = a(X - EX)$$

for some constant a. Applying this to the proof of CRLB with $X = W(\underset{\sim}{X}), Y = \frac{\partial}{\partial\theta} \log f(\underset{\sim}{X} \mid \theta)$ tells us that

$$Var_\theta W(\underset{\sim}{X}) = \frac{\{\tau'(\theta)\}^2}{I_{\underset{\sim}{X}}(\theta)}$$

iff

$$\frac{\partial}{\partial\theta}\log f(\underset{\sim}{X} \mid \theta) = a(\theta)[W(\underset{\sim}{X}) - \tau(\theta)] \tag{1}$$

for some function $a(\theta)$. (1) *is true only when* $f(\underset{\sim}{x} \mid \theta)$ *is a* 1*pef and* $W(\underset{\sim}{X}) = cT(\underset{\sim}{X}) + d$ *for some* c, d *where* $T(\underset{\sim}{X})$ *is the natural sufficient statistic of the* 1*pef.*

4 Asymptotic Efficiency

Let $\underset{\sim}{X}_n = (X_1, X_2, \ldots, X_n)$. Given a sequence of estimators $W_n = W_n(\underset{\sim}{X}_n)$. If $E(W_n) = \tau(\theta)$ for all n, then $\{W_n\}$ is asymptotically efficient if

$$\lim_{n\to\infty}\frac{\text{Var}_\theta W_n}{V_n(\theta)} = 1.$$

where

$$V_n(\theta) = \frac{\{\tau'(\theta)\}^2}{I_{\underset{\sim}{X}_n}(\theta)}$$

What if $\text{Var}_\theta W_n = \infty$ or if W_n is biased?
<u>An alternative definition:</u> A sequence of estimators $\{W_n\}$ is asymptotically normal if

$$\frac{W_n - \tau(\theta)}{\sqrt{V_n(\theta)}} \xrightarrow{d} \text{N}(0, 1).$$

as $n \to \infty$. $\{W_n\}$ is asymptotically efficient for estimating $\tau(\theta)$ if $W_n \sim \text{AN}(\tau(\theta), V_n(\theta))$.
Example: Observe $X_1, X_2, \ldots, X_n$ iid Poisson(λ).

- **Estimation of** $\tau(\lambda) = \lambda$:
  $E\bar{X} = \lambda$. Does $\bar{X}$ achieve the CRLB? Yes !

$$\text{Var}(\bar{X}) = \frac{\text{Var}(X_1)}{n} = \frac{\lambda}{n}$$
$$\text{CRLB} = \frac{\{\tau'(\lambda)\}^2}{I_{\mathbf{X}}(\lambda)} = \frac{1}{n/\lambda} = \frac{\lambda}{n}$$

  Alternative: Check condition for exact attainment of CRLB.

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log f(\mathbf{X} \mid \lambda) = \sum_{i=1}^{n}\frac{\partial}{\partial\lambda}\log f(X_i \mid \lambda) &= \sum_i\left(\frac{X_i}{\lambda} - 1\right) \\ &= \frac{n}{\lambda}(\bar{X} - \lambda)\end{aligned}$$

Note: Since $\bar{X}$ attains the CRLB (for all), it must be the best unbiased estimator of λ.
Showing that an estimator attains the CRLB is one way to show it is best unbiased. (But see later remark.)

- **Estimation of $\tau(\lambda) = \lambda^2$:** Define $W = T(T-1)/n^2$ where $T = \sum_{i=1}^n X_i$. $EW = \lambda^2$ (see calculations below) and W is a function of the CSS T. Thus W is best unbiased for λ^2. Does W achieve the CRLB? No !!! Note that

$$\begin{aligned} \text{CRLB} &= \frac{\{\tau'(\lambda)\}^2}{I_{\mathbf{X}}(\lambda)} = \frac{(2\lambda)^2}{n/\lambda} = \frac{4\lambda^3}{n}. \\ \text{Var}(W) &= \frac{4\lambda^3}{n} + \frac{2\lambda^2}{n^2} \quad \text{(see calculations below).} \end{aligned}$$

Alternative: Show condition for achievement of CRLB fails.
As show earlier:

$$\frac{\partial}{\partial\lambda} \log f(\mathbf{X} \mid \lambda) = \sum_i \left(\frac{X_i}{\lambda} - 1\right) = \frac{T}{\lambda} - n$$

The CRLB is attained iff there exists $a(\lambda)$ such that

$$\frac{T}{\lambda} - n = a(\lambda)\left(\frac{T(T-1)}{n^2} - \lambda^2\right).$$

But the left side is linear in T and the right side is quadratic in T, so that no multiplier $a(\lambda)$ can make them equal for all possible values of $T = 0, 1, 2, \ldots$.

Remark 2. *This situation is not unusual. The best unbiased estimator often fails to achieve the CRLB. But W is asymptotically efficient:*

$$\lim_{n\to\infty} \frac{Var(W)}{CRLB} = \lim_{n\to\infty} \frac{\frac{4\lambda^3}{n} + \frac{2\lambda^2}{n^2}}{\frac{4\lambda^3}{n}} = \lim_{n\to\infty}\left(1 + \frac{1}{2n\lambda}\right) = 1.$$

Calculations: Suppose $Y \sim \text{Poisson}(\xi)$. The factorial moments of the Poisson follow simple pattern:

$$\begin{aligned} EY &= \xi \\ EY(Y-1) &= \xi^2 \\ EY(Y-1)(Y-2) &= \xi^3 \\ EY(Y-1)(Y-2)(Y-3) &= \xi^4 \\ \cdots & \end{aligned}$$

Proof of one case:

$$\begin{aligned} EY(Y-1)(Y-2) &= \sum_{i=0}^{\infty} i(i-1)(i-2)\frac{\xi^i e^{-\xi}}{i!} \\ &= \xi^3 \sum_{i=3}^{\infty} \frac{\xi^{i-3}e^{-\xi}}{(i-3)!} = \xi^3 \sum_{i=0}^{\infty} \frac{\xi^i e^{-\xi}}{j!} = \xi^3 \end{aligned}$$

From the factorial moments, we can calculate everything else. For example:

$$\begin{aligned} \text{Var}(Y(Y-1)) &= E[\{Y(Y-1)\}^2] - [EY(Y-1)]^2 \\ &= E[\{Y^2(Y-1)^2\}] - [\xi^2]^2 \\ &= E[\langle Y\rangle_4 + 4\langle Y\rangle_3 + 2\langle Y\rangle_2] - \xi^4 \\ &= [\xi^4 + 4\xi^3 + 2\xi^2] - \xi^4 = 4\xi^3 + 2\xi^2 \end{aligned}$$

where $\langle Y\rangle_k \equiv Y(Y-1)(Y-2)\cdots(Y-k+1)$.
In our case $T \sim \text{Poisson}(\lambda)$ so that substituting $\xi = n\lambda$ in the above results leads to

$$\begin{aligned} ET(T-1) &= (n\lambda)^2 = n^2\lambda^2 \\ \text{Var}[T(T-1)] &= 4(n\lambda)^3 + 2(n\lambda)^2 = 4n^3\lambda^3 + 2n^2\lambda^2 \end{aligned}$$

so that $W = T(T-1)/n^2$ satisfies:

$$\begin{aligned} EW &= \lambda^2 \\ \text{Var}(W) &= \frac{4\lambda^3}{n} + \frac{2\lambda^2}{n^2}. \end{aligned}$$

4.1 An asymptotically inefficient estimator

Example: Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x \mid \alpha) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}, \quad x > 0.$$

For this pdf $EX = \text{Var}(X) = \alpha$. Clearly $E\bar{X} = \alpha$. Thus $\bar{X} = \text{MOM}$ estimator of α. Is it asymptotically efficient? No. (verified below).
Note: This is 1pef with natural sufficient statistic $T = \sum_{i=1}^{n} \log X_i$. Since T is complete, $E(\bar{X} \mid T)$ is the UMVUE of α. Since $\bar{X}$ is not a function of T, we know $\text{Var}(\bar{X}) > \text{Var}[E(\bar{X} \mid T)]$. But $\text{Var}[E(\bar{X} \mid T)] \geq \text{CRLB}$. Thus, without calculation, we know that $\bar{X}$ cannot achieve the CRLB for any value of n. We now show it does not achieve it asymptotically either.
Note that

$$\text{Var}\bar{X} = \frac{\text{Var}(X_1)}{n} = \frac{\alpha}{n}.$$

And,

$$I_{\underset{\sim}{X}_n}(\alpha) = nI_{X_1}(\alpha) = n\left[\frac{\Gamma''(\alpha)\Gamma(\alpha) - \{\Gamma'(\alpha)\}^2\}}{\{\Gamma(\alpha)\}^2}\right]$$

by a routine calculation. Hence

$$\text{CRLB} = \frac{1}{nI_{X_1}(\alpha)}.$$

Thus

$$\frac{\text{Var}(\bar{X})}{\text{CRLB}} = \alpha I_{X_1}(\alpha)$$

which does not depend on n. Since $\bar{X}$ does not achieve CRLB for any n, we know $\alpha I_{X_1}(\alpha) > 1$. Thus

$$\lim_{n\to\infty} \frac{\text{Var}(\bar{X})}{\text{CRLB}} = \alpha I_{X_1}(\alpha) > 1$$

so that $\bar{X}$ is not asymptotically efficient. The function $\alpha I_{X_1}(\alpha)$ is a non-negative decreasing function with

$$\lim_{\alpha\to 0} \alpha I_{X_1}(\alpha) = \infty \qquad \lim_{\alpha\to\infty} \alpha I_{X_1}(\alpha) = 1.$$

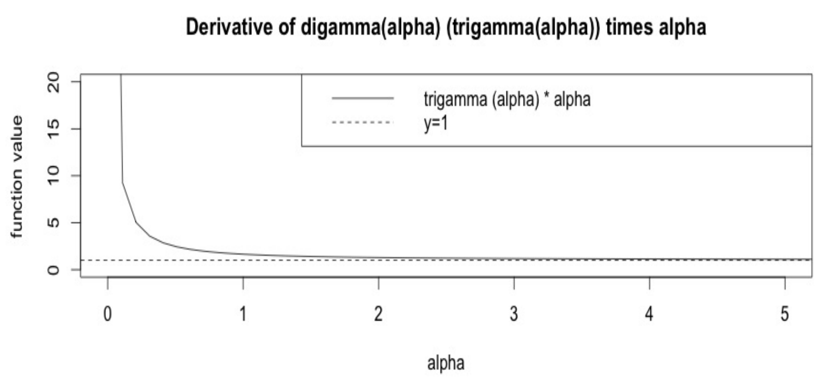

Figure 1: Plot of $\alpha I_{X_1}(\alpha)$, where $I_{X_1}(\alpha)$ is called the trigamma function (derivative of digamma function: $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$)

When α is small, $\bar{X}$ is horrible. When α is large, $\bar{X}$ is pretty good.
General Comment: For regular families, the MLE is asymptotically efficient. (MOM is inefficient in general). Thus

$$\lim_{n\to\infty} \frac{\mathrm{Var}W_n}{\mathrm{CRLB}(n)}$$

essentially compares the variance of W_n with that of the MLE in large samples.

5 Fisher Information, CRLB, Asymptotic distribution of MLE's in the multi parameter case

Notation: $\underset{\sim}{X} \sim f(\underset{\sim}{x} \mid \theta)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ and

$$\frac{\partial}{\partial \theta} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{pmatrix}$$

and $S_{p\times 1}$ is the vector of scores

$$\frac{\partial}{\partial \theta} \log f(\underset{\sim}{X} \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \log f(\underset{\sim}{X} \mid \theta) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \log f(\underset{\sim}{X} \mid \theta) \end{pmatrix}$$

Define

$$(p \times p)\,\text{matrix} \quad I_{\underset{\sim}{X}}(\theta) = E(S_{p\times 1} S'_{1\times p})$$

Note that S is evaluated at θ and the expectation is taken under the distribution indexed by the same parameter θ. For a vector or matrix, we define the expected values in this way:

$$E\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} EY \\ ZZ \end{pmatrix} \quad E\begin{pmatrix} W & X \\ Y & Z \end{pmatrix} = \begin{pmatrix} EW & EX \\ EY & EZ \end{pmatrix}$$

5.1 Properties

1. $E_\theta S_{p\times 1} = 0_{p\times 1}$.

2. $I_{\underset{\sim}{X}}(S) = \text{Cov}(S)$, the variance-covariance matrix of S

3. If $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ has independent components, then

$$I_{\underset{\sim}{X}}(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta) + \cdots + I_{X_n}(\theta).$$

4. If $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ are iid, then

$$I_{\underset{\sim}{X}}(\theta) = nI_{X_1}(\theta).$$

5. $I_{\underset{\sim}{X}}(\theta) = E\Big(- \frac{\partial^2}{\partial\theta^2} \log f(\underset{\sim}{X} \mid \theta)\Big)$ where we define

$$\frac{\partial^2}{\partial\theta^2} \log f(\underset{\sim}{X} \mid \theta) = \left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\underset{\sim}{X} \mid \theta)\right)$$

which is the $p \times p$matrix whose (i, j) entry is

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\underset{\sim}{X} \mid \theta).$$

5.2 Asymptotic distribution of MLE (of θ)

If $\hat{\theta}_n = \hat{\theta}_n(X_1, X_2, \ldots, X_n)$ is the sequence of MLE's (based on progressively larger samples), then

$$\hat{\theta}_n \sim \text{AN}(\theta, (I_{\underset{\sim}{X}}(\theta))^{-1})$$

where AN now stands for asymptotically multivariate normal. This means

$$\hat{\theta}_n \sim \text{N}(\theta, (I_{\underset{\sim}{X}}(\theta))^{-1})$$

for large n.
Recall: In iid case $I_{\underset{\sim}{X}}(\theta) = nI_{X_1}(\theta)$.

Estimate $I_{\underset{\sim}{X}}(\theta)$ by $I_{\underset{\sim}{X}}(\hat{\theta}_n)$ or

$$-\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log f(\underset{\sim}{X} \mid \theta)\right)\Bigg|_{\theta=\hat{\theta}_n}$$

5.3 Multi-parameter CRLB

$\mathbf{X}$ has joint pdf (pmf) $f(\mathbf{x} \mid \theta)$ which is a regular family. $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$. If $EW(\mathbf{X}) = \tau(\theta)$ where $\tau(\theta) \in \mathbb{R}$ is differentiable function of $\theta_i, i = 1, \ldots, p$, then

$$\mathrm{Var}(W(\mathbf{X}) \geq g' I^{-1} g$$

where $g \equiv \frac{\partial \tau(\theta)}{\partial \theta}_{p\times 1}$ and $I \equiv I_{\mathbf{X}}(\theta)_{p\times p}$.

Special Case: $W(\mathbf{X}) = \hat{\theta}_i$ with $\tau(\theta) = \theta_i$. That is, $\hat{\theta}_i$ is an unbiased estimate of θ_i. Now that vector g has $g_i = 1$ and $g_j = 0$ for $j \neq i$, and the CRLB gives

$$\mathrm{Var}(\hat{\theta}_i) \geq (I^{-1})_{ii}$$

where the right hand side is the ith diagonal element of I^{-1}.

Weaker result: Suppose we knew θ_j for all $j \neq i$. By fixing θ_j for $j \neq i$ at the known values, we get a one-parameter family and the CRLB for the one-parameter case gives

$$\mathrm{Var}(\hat{\theta}_i) \geq I_{ii}^{-1} = \frac{1}{I_{ii}} = \frac{1}{E\left(\frac{\partial}{\partial \theta_i} \log f(\mathbf{X} \mid \theta)\right)^2}$$

But, since $(I^{-1})_{ii} \geq I_{ii}^{-1}$,

$$\mathrm{Var}(\hat{\theta}_i) \geq (I^{-1})_{ii} \geq I_{ii}^{-1}$$

where the upper lower bound is the best you can do if you are estimating θ_i and all the other parameters are unknown, and the lower lower bound is the best you can do when all the other parameters are known.

Example: $\mathrm{N}(\mu, \sigma^2 = \xi)$ distribution.

$$f(x \mid \mu, \xi) = \frac{1}{\sqrt{2\pi\xi}} e^{-(x-\mu)^2/(2\xi)}.$$

Note that

$$l = \log f = -\frac{1}{2}\log(2\pi\xi) - \frac{(x-\mu)^2}{2\xi}$$

and

$$\frac{\partial}{\partial \theta} \log f(X \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial \mu} \log f \\ \frac{\partial}{\partial \xi} \log f \end{pmatrix} = \begin{pmatrix} \frac{x-\mu}{\xi} \\ -\frac{1}{2\xi} + \frac{(x-\mu)^2}{2\xi^2} \end{pmatrix}$$

and

$$I(\theta) = -E \begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \mu \partial \xi} \\ \frac{\partial^2 l}{\partial \xi \partial \mu} & \frac{\partial^2 l}{\partial \xi^2} \end{pmatrix} = -E \begin{pmatrix} \frac{-1}{\xi} & \frac{-(X-\mu)}{\xi^2} \\ \frac{-(X-\mu)}{\xi^2} & \frac{1}{2\xi^2} - \frac{(X-\mu)^2}{\xi^3} \end{pmatrix} = \begin{pmatrix} \frac{1}{\xi} & 0 \\ 0 & \frac{1}{2\xi^2} \end{pmatrix}$$

Hence

$$I^{-1} = \begin{pmatrix} \xi & 0 \\ 0 & 2\xi^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

For an unbiased estimate of $\mu(E_{\mu,\sigma^2} W = \mu)$, $\mathrm{Var}(W) \geq \frac{\sigma^2}{n}$ (achieved by $W = \bar{X}$).
For an unbiased estimate of σ^2, $\mathrm{Var}(W) \geq \frac{2\sigma^4}{n}$ (not achieved exactly) S^2 is best unbiased and $S^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1}$ so that $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$.
The limiting distribute of the MLE is given by

$$\begin{pmatrix} \bar{X} \\ \hat{\sigma}^2 \end{pmatrix} \sim \mathrm{AN}\left(\begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix} \right)$$

<u>Note:</u>

$$\mathrm{Var}\left(\frac{1}{n} \sum (X_i - \mu)^2 \right) = \frac{2\sigma^4}{n}$$
$$\mathrm{E}\left(\frac{1}{n} \sum (X_i - \mu)^2 \right) = \sigma^2.$$

achieves the CR-bound, but not legitimate estimator if μ is unknown.
Example: Gamma(α, β) Recall the digamma function $\psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. Note that

$$f(x \mid \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$
$$l = \log f = -\log\Gamma(\alpha) - \alpha\log\beta + (\alpha - 1)\log x - x/\beta.$$

Then

$$\frac{\partial}{\partial\theta} \log f(X \mid \theta) = \begin{pmatrix} \frac{\partial}{\partial\alpha} \log f \\ \frac{\partial}{\partial\beta} \log f \end{pmatrix} = \begin{pmatrix} -\psi(\alpha) - \log\beta + \log X \\ -\frac{\alpha}{\beta} + \frac{X}{\beta^2} \end{pmatrix}$$

and

$$I(\theta) = -E \begin{pmatrix} \frac{\partial^2 l}{\partial\alpha^2} & \frac{\partial^2 l}{\partial\alpha\partial\beta} \\ \frac{\partial^2 l}{\partial\beta\partial\alpha} & \frac{\partial^2 l}{\partial\beta^2} \end{pmatrix} = -E \begin{pmatrix} -\psi'(\alpha) & \frac{-1}{\beta} \\ \frac{-1}{\beta} & \frac{\alpha}{\beta^2} - \frac{2X}{\beta^3} \end{pmatrix} = \begin{pmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & \frac{\alpha}{\beta^2} \end{pmatrix}$$

Hence

$$I(\theta)^{-1} = \frac{\beta^2}{\alpha\psi'(\alpha) - 1} \begin{pmatrix} \frac{\alpha}{\beta^2} & -\frac{1}{\beta} \\ -\frac{1}{\beta} & \psi'(\alpha) \end{pmatrix} = \frac{1}{\alpha\psi'(\alpha) - 1} \begin{pmatrix} \alpha & -\beta \\ -\beta & \beta^2\psi'(\alpha) \end{pmatrix}$$

CRLB for unbiased estimator of β is given by

$$\mathrm{Var}(\hat{\beta}) \geq \frac{1}{n}(I^{-1}(\theta))_{22} \geq \frac{1}{n}\{I(\theta))_{22}\}^{-1}.$$

Note that

$$\frac{1}{n}(I^{-1}(\theta))_{22} = \frac{\beta^2}{\alpha n} \cdot \frac{\psi'(\alpha)}{\psi'(\alpha) - 1/\alpha}, \quad \frac{1}{n}\{I(\theta))_{22}\}^{-1} = \frac{\beta^2}{\alpha n}.$$

If α is known the lower lower bound is achieved

$$\begin{aligned} E\left(\frac{\bar{X}}{\alpha}\right) &= \beta \\ \mathrm{Var}\left(\frac{\bar{X}}{\alpha}\right) &= \frac{1}{\alpha^2}\frac{\mathrm{Var}(X)}{n} = \frac{\alpha\beta^2}{n\alpha^2} = \frac{\beta^2}{\alpha n}. \end{aligned}$$

If α must be estimated, there is a variance penalty which does not vanish asymptotically ($n \to \infty$).

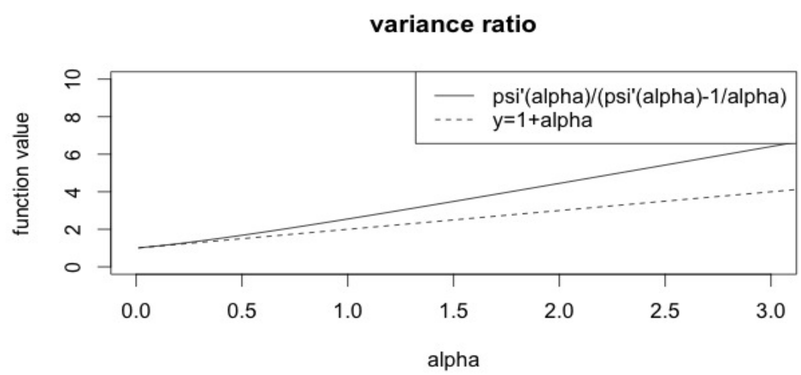

Figure 2: Plot of $\frac{\psi'(\alpha)}{\psi'(\alpha)-1/\alpha}$, showing that it does not become asymptotically 1