Fisher information and Cramer-Rao bound

Contents

1 Score and Fisher information

Let X be a rv (or a rve) with pd $f(x;\theta)$ indexed by an unknown parameter $\theta \in \Theta \subset \mathbb{R}$. The question of interest here is the following: how much information can be obtained about θ as X get observed?

To answer this question, let's assume that f is differentiable with respect to θ, and define the ***score*** function associated with f to be

$$S(\theta, x) := \partial_\theta \log f(x;\theta) = \frac{\partial_\theta f(x;\theta)}{f(x;\theta)}.$$

Obverse that, by definition, for any given $\theta_0 \in \Theta$,

$$S(\theta_0, x) = \lim_{\epsilon \to 0} \frac{\frac{1}{\epsilon}\left[f(x, \theta_0 + \epsilon) - f(x;\theta_0)\right]}{f(x;\theta_0)}.$$

Thus, we can interpret this score as the relative (instantaneous) rate of change of $\theta \mapsto f(x;\theta)$ at the point θ_0. In particular, a function f that varies rapidly in the neighborhood of θ_0 should produce a large (absolute) score, and, in opposite, if f is flat the score should be small. In other words, a large value of $|S(\theta_0, x)|$ indicate that we can easily distinguish θ_0 from its neighboring values.

A remarkable property of the score function is given in the following proposition.

Proposition 1.1. *Suppose that*

(I) the support of f, i.e. the set $\{x : \ f(x;\theta) > 0\}$, does not depend on θ, and

(II) the operations of integration (or summation) and differentiation by θ can be interchanged in $\int f(x;\theta)dx$. Thus, $\partial_\theta \int f(x;\theta)dx = \int \partial_\theta f(x;\theta)dx$. (more about this condition can be found here)

Then

$$E_\theta S(\theta, X) = 0, \ \forall \theta.$$

Assumptions (I) and (II), given above, are both known to *hold for the exponential family*. And from now on, *unless otherwise stated*, we will always assume that (I) and (II) hold.

Taking the square (of S) and averaging we obtain $I_X(\theta)$:

$$I_X(\theta) := E_\theta \left[S^2(\theta, X)\right] = Var_\theta \left[S(\theta, X)\right]$$

which is known as the (expected) ***Fisher information*** (**FI**) that X contains about θ.

The Fisher information attempts to *quantify the average sensitivity of the random variable X to the value of the parameter θ*. If small changes in θ result in large changes in the values of X, then observing the latter can tell us a lot about θ.

In this case the FI would be quite large. In other words, FI attempts to quantify how easy one can guess the θ that produced the observed X.

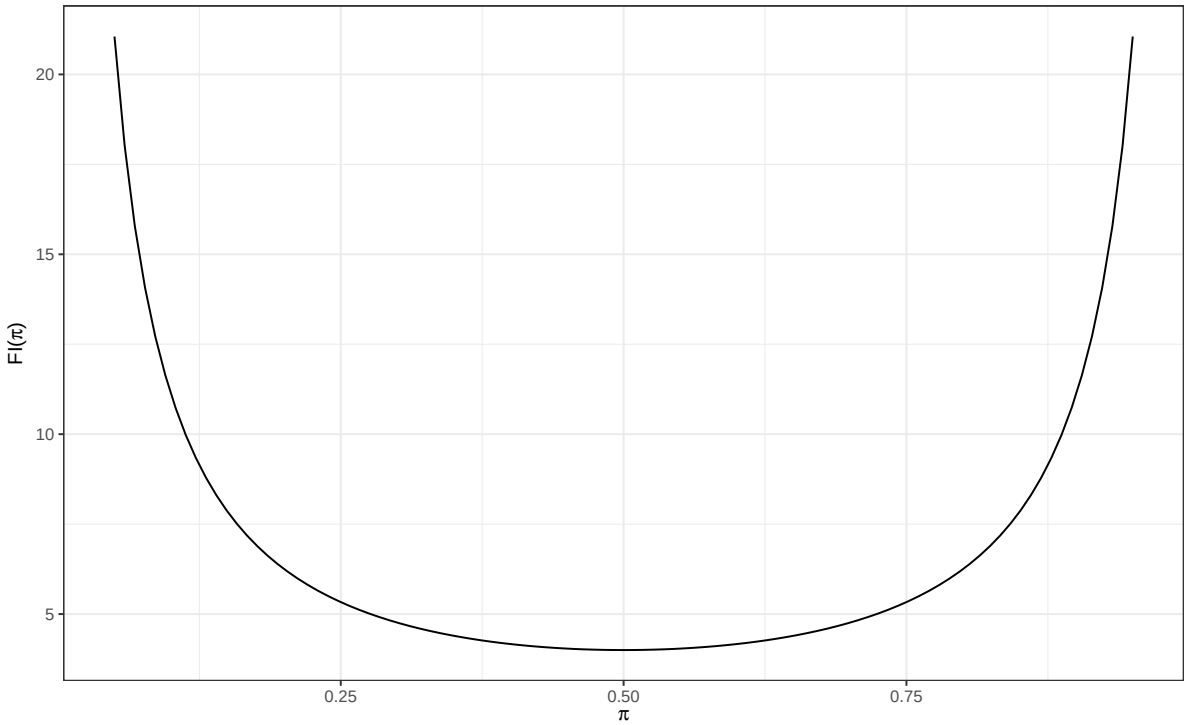
Remark. Note that the "X" in $I_X(\theta)$ is only a symbol used to indicate that the FI corresponds to the rv X. It does not mean that the FI it self is random! □

Example 1.1 (Calculating Fisher Information 1)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0, 1)$.

$$
\begin{aligned}
f(x;\pi) &= \pi^x (1-\pi)^{1-x},\ x = 0, 1 \\
S(\pi, x) &= \partial_\pi \left\{ x \log(\pi) + (1-x) \log(1-\pi) \right\} = \frac{x-\pi}{\pi(1-\pi)} \\
I(\pi) &= E_\pi \left[S^2(\pi, X) \right] = \frac{1}{\pi(1-\pi)}.
\end{aligned}
$$

In this case, the FI is the reciprocal of the variance. This is not a unusual situation. Actually, as we will see latter, the FI is typically inversely proportional to the variance. The greater the variation (thus the smaller the FI), the more difficult it is to recover the parameter of interest.


20
15
10
5
FI(π)
0.25
0.50
0.75
π

- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ (the parameter of interest) and $\sigma^2 > 0$.

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
$$S(\mu, x) = \partial_\mu \left\{ -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2 \right\} = \frac{x-\mu}{\sigma^2}$$
$$I(\mu) = E\left[S^2(\mu, X)\right] = \frac{1}{\sigma^2}.$$

In this case, the FI about μ does not depend on μ but only on σ^2. It decreases with σ^2. □

Proposition 1.2. *Assume thta $f(x;\theta)$ is twice differentiable, with respect to θ, and double integration (or summation) and differentiation under the integral sign can be interchanged, thus $\partial^2_\theta \int f(x;\theta)dx = \int \partial^2_\theta f(x;\theta)dx$. Then,*

$$I_X(\theta) = -E_\theta \left[\partial_\theta S(\theta, X)\right].$$

This is equivalent to say that $I_X(\theta) = -E_\theta \left[\partial^2_\theta \log f(X;\theta)\right]$.

To see why this proposition is true, consider the following calculation where we have simplify the notation by writing

S and f instead of $S(\theta, X)$ and $f(X;\theta)$, respectively.

$$\partial_\theta S = \partial_\theta \frac{\partial_\theta f}{f} = \frac{\partial_\theta^2 f \times f - (\partial_\theta f)^2}{f^2} = \frac{\partial_\theta^2 f}{f} - S^2$$

But $E_\theta\left(\frac{\partial_\theta^2 f(X;\theta)}{f(X;\theta)}\right) = \int \partial_\theta^2 f(x;\theta)dx = \partial_\theta^2 \int f(x;\theta)dx = 0$. Thus, $E_\theta(\partial_\theta S) = -E_\theta S^2$.

Example 1.2 (Calculating Fisher Information 2)**.**

- Bernoulli distribution: $X \sim Be(\pi)$ with $\pi \in (0,1)$.

$$I(\pi) = -E[\partial_\pi S(\pi, X)] = -\frac{1}{\pi^2(1-\pi)^2} E\left[-\pi(1-\pi) - (\pi(1-\pi))'(X-\pi)\right] = \frac{1}{\pi(1-\pi)}.$$

- Normal distribution: $X \sim N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$.

$$I(\mu) = -E[\partial_\mu S(\mu, X)] = -E\left[\frac{-1}{\sigma^2}\right] = \frac{1}{\sigma^2}. \square$$

2 FI contained in a statistic

The above definitions of the score and FI can be directly applied to any statistics. In fact, let $T \equiv T(\boldsymbol{X}) \equiv T(X_1, \ldots, X_n)$ be a statistic whose pd is given by $h_n(t;\theta)$. The score associated with h_n and its corresponding FI are given by

$$\begin{aligned} S(\theta, T) &:= \partial_\theta \log h_n(T;\theta) \\ I_T(\theta) &:= E\left[S^2(\theta, T)\right] = Var\{S(\theta, T)\} \end{aligned}$$

$I_T(\theta)$ is the (Fisher) information about θ that we can extract form T.
Assuming the interchangeability of integration and differentiation twice, this FI can also be expressed as

$$I_T(\theta) = -E\left[\partial_\theta S(\theta, T)\right] = -E\left[\partial_\theta^2 \log h_n(T;\theta)\right].$$

Example 2.1.

- Let $X_i, i = 1, \ldots, n$, be an iid sample from a Bernoulli distribution $Be(\pi)$. Let define the statistic $T = \sum_{i=1}^n X_i$.

Since $T \sim Bin(n, \pi)$, the pd of T is given by $h_n(t;\pi) = p_\pi(T = t) = C_n^t \pi^t (1-\pi)^{n-t}$. It follows that

$$I_T(\pi) = E\left[S^2(\theta, T)\right] = E\left[\frac{T - n\pi}{\pi(1-\pi)}\right]^2 = \frac{n}{\pi(1-\pi)}.$$

- Let $X_i, i = 1, \ldots, n$, be an iid sample from a Normal distribution $N(\mu, \sigma^2)$. Since $\overline{X}_n \sim N(\mu, \sigma^2/n)$, it follows that

$$I_{\overline{X}_n}(\mu) = \frac{n}{\sigma^2}. \square$$

An important fact about FI is its ***additivity***. Let T_1 and T_2 be two statistics with pd h_1 and h_2, and with FI $I_{T_1}(\theta)$ and $I_{T_2}(\theta)$, respectively. If T_1 and T_2 are *independent*, i.e. if $h(t_1, t_2; \theta) = h_1(t_1;\theta)h_2(t_2;\theta)$, $\forall t$, with h beging the joint pd of (T_1, T_2), then

$$\begin{aligned} I_{(T_1,T_2)}(\theta) &= E(\partial_\theta \log h(T_1, T_2; \theta))^2 = E(\partial_\theta \log h_1(T_1;\theta) + \partial_\theta \log h_2(T_2;\theta))^2 \\ &= E(\partial_\theta \log h_1(T_1;\theta))^2 + E(\partial_\theta \log h_2(T_2;\theta))^2 + 2E(\partial_\theta \log h_1(T_1;\theta)\partial_\theta \log h_2(T_2;\theta)) \\ &= I_{T_1}(\theta) + I_{T_2}(\theta). \end{aligned}$$

As consequence we have the following result.

Proposition 2.1. *Let* $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be an iid sample with joint pd* $f_n(\boldsymbol{x}; \theta) = \prod_i f(x_i, \theta)$, $I_{X_i}(\theta) = E(\partial_\theta \log f(X_i, \theta))^2$ *be the FI contained in* X_i *about* θ, *and* $I_{\boldsymbol{X}}(\theta) = E(\partial_\theta \log f_n(\boldsymbol{X}, \theta))^2$ *be the FI contained in* $\boldsymbol{X}$ *about* θ. *Then,* $I_{\boldsymbol{X}}(\theta) = nI_{X_i}(\theta)$.

Attention. From now on, if no confusion is possible, we reserve the notation I_n for the FI contained in the entire sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, i.e. $I_n = I_{\boldsymbol{X}_n}$, and reserve the notation I for the FI contained in one sample unit, i.e. $I = I_{X_i} = I_{X_1}$. We can therefore write the equality above as $I_n(\theta) = nI(\theta)$.

For any statistic $\boldsymbol{T} = (T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_d(\boldsymbol{X}))$, $d \geq 1$, it can be shown that

$$0 \leq I_{\boldsymbol{T}}(\theta) \leq I_n(\theta).$$

Thus, the information on θ contained in any statistic $\boldsymbol{T}$, derived from a sample $\boldsymbol{X}$, cannot exceed the information, on θ, contained in the sample $\boldsymbol{X}$ itself .

It may happen that $I_{\boldsymbol{T}}(\theta) = I_n(\theta)$, in which case $\boldsymbol{T}$ *contains the same amount of information about* θ *as the whole sample*. Such a statistic is called a ***sufficient statistic***. Formally, the (original) definition of a sufficient statistic is as follows (this definition applies to both the single-parameter and multi-parameter cases).

Definition 2.1 (Sufficiency)**.** A statistic $\boldsymbol{T}$ is sufficient for $\boldsymbol{\theta}$ if the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{T}$ does not depend on $\boldsymbol{\theta}$.

Put another way, given a sufficient statistic $\boldsymbol{T}$ for $\boldsymbol{\theta}$, the sample $\boldsymbol{X}$ provides no additional information about $\boldsymbol{\theta}$. A sufficient statistic is particularly interesting if it is of *smaller dimension than the sample size*; i.e. $d << n$.

Example 2.2.

- From the examples 1.1 and 2.1, we learned that, for the Bernoulli distribution, $I_{\sum_{i=1}^n X_i}(\pi) = I_n(\pi)$. Thus, the one-dimensional statistic $\sum_{i=1}^n X_i$ contains as much information about π as the whole sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. So, $\sum_{i=1}^n X_i$ is sufficient for π.
- The same remark applies to the normal distribution where, from the examples 1.1 and 2.1, we can see that $I_{\overline{X}_n}(\mu) = I_n(\mu)$ and so $\overline{X}_n$ is sufficient for μ. □

A sufficient statistic is not unique. In fact, if $U = k(T)$, T is a sufficient statistic, and k is bijective (or at least one-to-one), then U is also sufficient. As an example, if we consider the case of the normal distribution, we can say that $\sum_{i=1}^n X_i$ is sufficient for μ since we known that $\overline{X}_n$ is.

The following result, known as the ***Factorization Theorem***, makes it very easy to identify sufficient statistics.

Theorem 2.1 (Factorization Theorem). *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. A statistic $\boldsymbol{T}(\boldsymbol{X})$ is sufficient for $\boldsymbol{\theta}$ if and only if f_n can be factorized as*

$$f_n(\boldsymbol{x}; \boldsymbol{\theta}) = \phi(\boldsymbol{T}(\boldsymbol{x}); \boldsymbol{\theta}) h(\boldsymbol{x}), \ \forall \boldsymbol{x}, \boldsymbol{\theta}.$$

where ϕ is a function that depends on $\boldsymbol{x}$ only through $\boldsymbol{T}(\boldsymbol{x})$ and h is a function that does not depend on $\boldsymbol{\theta}$.

Example 2.3. Let $X_i, i = 1, \ldots, n$, be an iid sample from a Bernoulli distribution $Be(\pi)$. The joint pd of $(X_1, \ldots, X_n)$ is

$$\prod_{i=1}^{n} \pi^{x_i} (1 - \pi)^{1 - x_i} = \pi^{\sum_i x_i} (1 - \pi)^{n - \sum_i x_i}.$$

It follows that $\sum_{i=1}^{n} X_i$ is sufficient for π.

As a consequence of this theorem, we have the following result that allows one to obtain sufficient statistics when data come from an exponential family.

Proposition 2.2. *If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an iid sample from a J-parameter exponential family ($J \geq 1$) with pd*

$$h(x) \exp\big(\boldsymbol{\eta}^t(\boldsymbol{\theta}) \boldsymbol{T}(x) - B(\boldsymbol{\theta})\big),$$

then the statistic $\sum_{i=1}^{n} \boldsymbol{T}(X_i)$ *is sufficient for* $\boldsymbol{\theta}$.

Example 2.4.

- Let $X_1, \ldots, X_n$ be an iid sample from $f(x;\theta) = \theta x^{\theta-1}$, where $x \in (0,1)$ and $\theta > 0$. Let's show that $\prod_{i=1}^{n} X_i$ is sufficient for θ. To see this, we can write $f(x;\theta)$ as

$$f(x;\theta) = I(0 < x < 1)x^{-1} \exp(\theta \log x + \log \theta),$$

  which is an exponential family with natural statistic $T = \log X$. It follows that $L = \sum_i \log(X_i)$ is sufficient for θ. And since $\prod_{i=1}^{n} X_i = \exp(L)$ is a bijective function of L, it is also sufficient for θ. □

- In the case of $N(\mu, \sigma^2)$, with a unknown μ and σ^2, based on the above proposition and what we have learned previously about this distribution, we can conclude that $(\sum_i X_i, \sum_i X_i^2)$ is sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$. Thus, we can conclude that $(\overline{X}, S^2)$ is sufficient for $\boldsymbol{\theta}$.

3 FI and re-parametrization

We have already seen that statistical models can be parameterized in different ways. It is important to realize that FI depends on the chosen parameterization.

Proposition 3.1 (FI re-parametrization)**.** *Let $\eta : \theta \mapsto \eta(\theta)$ be a real differentiable function of θ. Denote by $f^*(x;\eta)$ the pd of X parameterized with η so that $f^*(x;\eta) = f(x;\theta)$, $\forall x$. Let $I(\theta) = E[\partial_\theta \log f(X;\theta)]^2$ be the FI about θ (when the parameterization in θ is used) and $I(\eta) = E\left[\partial_\eta \log f^*(X;\eta)\right]^2$ be the FI about η (when the parameterization in η is used). Then $I(\theta) = I(\eta) \times (\partial_\theta \eta(\theta))^2$.*

The proof of this proposition is straightforward and is a direct consequence of the chain rule :

$$I(\theta) = E[\partial_\theta \log f(X;\theta)]^2 = E[\partial_\theta \log f^*(X;\eta)]^2 = E\left[\partial_\theta \eta(\theta) \partial_\eta \log f^*(X;\eta)\right]^2 = (\eta'(\theta))^2 I(\eta).$$

In the following, for a given model $\{f(x;\theta);\ \theta \in \Theta\}$, every time we write $I(\eta(\theta))$, we will be referring to the FI with the reparameterization $\theta \mapsto \eta(\theta)$. □

Example 3.1.

- Let $X \sim N(\mu, \sigma^2)$. $f(x; \mu, \sigma^2) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

$$I(\sigma) = -E\left[\partial_\sigma^2 \log f\right] = -E\left[\partial_\sigma \left\{ -\frac{1}{\sigma} + \frac{(X-\mu)^2}{\sigma^3} \right\}\right] = -E\left[\frac{1}{\sigma^2} - 3\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{2}{\sigma^2}.$$

And, with the (re)parametrization $\theta = \sigma^2$,

$$I(\sigma^2) = -E[]\partial_\theta^2 \log f] = -E\left[\partial_\theta \left\{ -\frac{1}{2\theta} + \frac{1}{2}\frac{(X-\mu)^2}{\theta^2} \right\}\right] = -E\left[\frac{1}{2\theta^2} - \frac{(X-\mu)^2}{\theta^3}\right] = \frac{1}{2\sigma^4}$$

We observe that

$$I(\sigma^2) = \frac{1}{4\sigma^2} I(\sigma).$$

This last result can be obtained directly by applying Proposition 3.1. Indeed, we have that

$$I(\sigma^2) = I(\sigma) \times \left(\partial_t \sqrt{t} \Big|_{t=\sigma^2} \right)^2 = I(\sigma) \left(\frac{1}{2\sqrt{\sigma^2}} \right)^2.$$

Or, equivalently,

$$I(\sigma) = I(\sigma^2) \times \left(\partial_t \, t^2 \Big|_{t=\sigma} \right)^2 = I(\sigma^2) \times (2\sigma)^2.$$

- Let $X \sim Pois(\theta)$. $f(x;\theta) = \dfrac{\theta^x}{x!}e^{-\theta}$, $x = 0, 1, \ldots$ and $\theta > 0$. Direct calculation leads to

$$I(\theta) = -E\{\partial_\theta^2 \log f\} = -E\left\{-\frac{X}{\theta^2}\right\} = \frac{1}{\theta}$$

Now, let's consider the parametrization with $\eta = \log(\theta)$: $f^*(x;\eta) = \dfrac{e^{x\eta}}{x!}e^{-e^\eta}$, $\eta \in (-\infty, \infty)$.

$$I(\eta) = -E\left\{\partial_\eta^2 \log f^*\right\} = -E\{-e^\eta\} = e^\eta$$

Thus, $I(\log(\theta)) = \theta$. This same result can be obtained by directly applying Proposition 3.1 as follows:

$$I(\log(\theta)) = \frac{I(\theta)}{((\log(\theta))')^2} = \frac{1/\theta}{1/\theta^2} = \theta. \square$$

This last example suggests that, in Poisson model, it is easier to estimate $\eta = \log(\theta)$ than θ when the latter is large. Let's check this out. A natural estimator of θ is $\hat{\theta} := \overline{X}$ and a natural estimator of η is $\hat{\eta} := \log(\hat{\theta}) = \log(\overline{X})$. We have that $MSE(\hat{\theta}) = \dfrac{\theta}{n}$ and, by first order Taylor polynomial approximation, i.e. $\log(\hat{\theta}) \approx \log(\theta) + (\hat{\theta} - \theta)\dfrac{1}{\theta}$, we can write $MSE(\hat{\eta}) \approx \dfrac{1}{n\theta}$. These mean-square errors cannot be compared, as they are of different scales/magnitudes. However,

we can see that as θ increases, the MSE performance of $\hat{\theta}$ becomes worse and worse compared to that of $\hat{\eta}$. The following simulation confirms this fact.

```r
th <- 10 # 20, 40, 80, 100, ...
eta <- log(th)

hat.th <- replicate(5000, rpois(100, th) |> mean())
hat.eta <- log(hat.th)

# relative MSE
mse(hat.th, th) / mse(hat.eta, eta)
```

```
[1] 99.9
```

```r
# relative "mean absolute relative error"
mean(abs(hat.th / th - 1)) / mean(abs(hat.eta / eta - 1))
```

```
[1] 2.3
```

Remark. We have seen that, in general, $I(g(\theta)) \neq I(\theta)$. We may ask the question what happen with FI when the data itself, or a statistic from it, are transformed. The answer depends on the type of transformation used. For example in the case of strictly monotonic and differentiable transformation of the data, FI does change. More precisely, if $U = k(T)$, where k is a differentiable and strictly monotonic function that does not depend on θ, then $I_T(\theta) = I_U(\theta)$. This is a direct consequence of the *Change of Variable(s) Formula*: $f_T(T;\theta) = f_U(U;\theta)\,|k'(U)|$. □

4 Information Inequality: The Cramer-Rao Lower bound (CRLB)

We will develop a lower bound for the variance of any given statistic, which can be mainly used (i) as a benchmark for comparing estimator performance, and (ii) to find the MVUE (Minimum Variance Unbiased Estimator). The bound we are interested in is called the Cramer-Rao Lower Bound (**CRLB**), and is given in the following theorem.

Theorem 4.1 (Information Inequality)**.** *Let $\boldsymbol{X} \equiv \boldsymbol{X}_n = (X_1, \ldots, X_n)$ be an iid sample with joint pd $f_n(\boldsymbol{x};\theta)$, $\theta \in \Theta$. Assume that assumptions (I) and (II) as given above (see Proposition 1.1) hold for $f_n(\boldsymbol{x};\theta)$, the joint dp of $\boldsymbol{X}$. Let $T \equiv T(\boldsymbol{X})$ be a statistic. Assume that (III) $\partial_\theta E_\theta(T)$ exists and can be obtained by differentiating under the integral sign. i.e., $\partial_\theta \int T(\boldsymbol{x}) f_n(\boldsymbol{x};\theta) d\boldsymbol{x} = \int T(\boldsymbol{x}) \partial_\theta f_n(\boldsymbol{x};\theta) d\boldsymbol{x}$. Then*

$$Var_\theta(T) \geq \frac{\left(\partial_\theta E_\theta(T)\right)^2}{I_n(\theta)}, \; \forall \theta \in \Theta.$$

The inequality above is a direct consequence of the Cauchy-Schwarz inequality. The proof goes as follows. Let $S_n \equiv S(\theta; \boldsymbol{X}) = \partial_\theta \log f_n(\boldsymbol{X}; \theta)$ be the Score associated with f_n. By the Cauchy-Schwarz inequality,

$$[Cov(T, S_n)]^2 = [E(TS_n) - E(T)E(S_n)]^2 = [E(TS_n)]^2 \leq Var(T)Var(S_n) = Var(T)I_n(\theta).$$

The final result is the consequence of the fact that $E(TS_n) = \int T(\boldsymbol{x})\partial_\theta \log f_n(\boldsymbol{X}; \theta) f_n(\boldsymbol{x}; \theta) d\boldsymbol{x} = \int T(\boldsymbol{x})\partial_\theta f_n(\boldsymbol{x}; \theta) d\boldsymbol{x} = \partial_\theta E(T)$.

As a corollary of the theorem above, we can say that, if $\hat{\theta}$ is an *unbiased* estimator of θ, then $Var_\theta(\hat{\theta}) \geq \frac{1}{I_n(\theta)}$. Or, more generally, if $\hat{\delta}$ is an unbiased estimator of $g(\theta)$, then

$$Var_\theta(\hat{\delta}) \geq \frac{\left(g^{'}(\theta)\right)^2}{I_n(\theta)} = \frac{1}{I_n(g(\theta))}.$$

The right hand side of this inequality is called the ***CRLB*** *for* $g(\theta)$. This bound is the minimum possible variance that any unbiased estimator of $g(\theta)$ can achieve. To put it more precisely, in a *parametric model in which Assumption (III) above is fulfilled for any statistic* T *(including* $T = 1$*)*, the variance of any unbiased estimator of $g(\theta)$ is at least equal to the $CRLB(g(\theta)) := I_n^{-1}(g(\theta))$. Note that Assumption (III) *holds for the exponential family for any statistic* T.

An *unbiased* estimator $\hat{\delta}$ of $g(\theta)$ is called ***efficient*** if its variance equals the CRLB for $g(\theta)$, i.e. if $Var\left(\hat{\delta}\right) = I_n^{-1}(g(\theta))$;

otherwise its (absolute) efficiency is defined to be

$$Eff(\hat{\delta}) := \frac{CRLB(g(\theta))}{Var(\hat{\delta})} = \frac{\left(g^{'}(\theta)\right)^2}{I_n(\theta)Var(\hat{\delta})}.$$

$Eff(\hat{\delta}) \in [0,1]$, and equals 1 if and only if $\hat{\delta}$ is efficient.

By definition, an efficient estimator is (1) unbiased and (2) its variance is uniformly lower than (or equal to) the variance of any other unbiased estimator. Thus, *an efficient estimator, when it exists, is the uniformly minimum variance unbiased estimator (MVUE)*. Note that *efficiency is a stronger requirement than being MVUE*. In fact, there are many cases where the CRLB is not attainable and where a MVUE exists.

Example 4.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$ (exponential family).

- Suppose that μ is our parameter of interest. We have seen that $I_n(\mu) = n/\sigma^2$. So for any unbiased estimator $\hat{\mu}_n$ of μ,

  $$Var(\hat{\mu}_n) \geq \frac{\sigma^2}{n}.$$

  Now, since $\overline{X}_n$ is an unbiased estimator of μ and $Var(\overline{X}_n) = \frac{\sigma^2}{n}$, we conclude that $\overline{X}_n$ is efficient for μ. And so, $\overline{X}_n$ is the MVUE of μ.

- Suppose that σ^2 is our parameter of interest and μ is known. We known that $I_n(\sigma^2) = \frac{n}{2\sigma^4}$. So for any unbiased estimator $\hat{\sigma}_n^2$ of σ^2,

$$Var(\hat{\sigma}_n^2) \geq 2\sigma^4/n.$$

  On the other hand, we known that $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$ is an unbiased estimator of σ^2. And, using the fact that $E(X-\mu)^4 = 3\sigma^4$, we have that $Var\left(\tilde{\sigma}_n^2\right) = n^{-1}Var(X-\mu)^2 = n^{-1}\left(E(X-\mu)^4 - \sigma^4\right) = 2\sigma^4/n$. So, we conclude that $\tilde{\sigma}_n^2$ is efficient for σ^2. And so it is the MVUE of σ^2.

- Suppose that μ^2 is our parameter of interest. By the information inequality, with $g : \mu \mapsto \mu^2$, we have that

$$Var(\hat{\delta}) \geq \frac{(2\mu)^2}{n/\sigma^2} = \frac{4\mu^2\sigma^2}{n}.$$

  for any unbiased estimator $\hat{\delta}$ of μ^2. But, for now, we cannot say if this limit is attainable or not and thus if there is an efficient estimator for μ^2 or not.

Example 4.2 (Importance of assumptions)**.** Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $Unif(0, \theta)$, $\theta > 0$. Thus $f(x;\theta) = \frac{1}{\theta}$, $0 < x < \theta$. Since $\partial_\theta \log f(x;\theta) = -1/\theta$, if we apply Theorem 4.1, we could conclude that, for any unbiased estimator $\hat{\theta}$ of θ,

$$Var(\hat{\theta}) \geq \frac{\theta^2}{n}.$$

But, we learned earlier that the estimator $\hat{\theta}_2 = \frac{n+1}{n} X_{(n)}$ is unbiased and its variance is $\frac{\theta^2}{n(n+2)}$, which is uniformly smaller than θ^2/n ! The problem here is that the support of f depends on θ, which means that the required Assumption (III) is not fulfilled. □

5 Efficiency in exponential families

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from a one-parameter exponential family with pd (in canonical form, with η begin the canonical parameter) : $f(x;\eta) = h(x) \exp\left(\eta T(x) - A(\eta)\right)$. For $T = T(X_1)$, we have seen that $E(T) = A^{'}(\eta)$ and $Var(T) = A^{''}(\eta)$. The FI contained in T about η is

$$I(\eta) = Var\big(\partial_\eta \log f\big) = Var(T) = A^{''}(\eta).$$

Note that this result can also be obtained from

$$I(\eta) = -E\Big(\partial_\eta^2 \log f\Big) = -E\Big(-A^{''}(\eta)\Big) = A^{''}(\eta).$$

Now, let $\overline{T} = n^{-} \sum_i T(X_i)$. We have that, $E\left(\overline{T}\right) = E(T) = A^{'}(\eta)$, and

$$Var\left(\overline{T}\right) = \frac{Var(T)}{n} = \frac{A^{''}(\eta)}{n} = \frac{\left(A^{''}(\eta)\right)^2}{I_n(\eta)}.$$

This demonstrates that $\overline{T}$ is efficient for $A^{'}(\eta)$.

Let's now consider the original parameterization with $\theta : f(x;\theta) = h(x)\exp\left(\eta(\theta)T(x) - B(\theta)\right)$. Remember that, in terms of θ, we can express $E(T)$ and $Var(T)$ as $E(T) = B^{'}(\theta)/\eta^{'}(\theta)$ and $Var(T) = \partial_\theta E(T)/\eta^{'}(\theta)$. The FI contained in T about θ is

$$I(\theta) = Var(\partial_\theta \log f) = Var\left(\eta^{'}(\theta)T\right) = \eta^{'}(\theta)\partial_\theta E(T).$$

Again, we have that $E\left(\overline{T}\right) = E(T)$, and

$$Var(\overline{T}) = \frac{Var(T)}{n} = \frac{\partial_\theta E(T)}{n\eta^{'}(\theta)} = \frac{\left(\partial_\theta E(T)\right)^2}{I_n(\theta)}.$$

So, $\overline{T}$ is efficient for $B^{'}(\theta)/\eta^{'}(\theta)$.

Thus, we can conclude that, *in an exponential family, the sample mean of the natural statistic is always efficient for the*

expected value of the latter.

Example 5.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by

$$\begin{aligned} f(x;\theta) &= \frac{1}{\theta} e^{-\frac{x}{\theta}} I(x > 0) \\ &= I(x > 0) \exp\left(-\frac{1}{\theta} x - \log(\theta) \right) \qquad \text{(exponential family)} \end{aligned}$$

Since $E(X_1) = \theta$, $n^{-1} \sum_{i=1}^{n} X_i$ is efficient for θ. □

6 CRLB Attainment

A natural question to ask is under what conditions a given unbiased estimator, say $T(\boldsymbol{X})$, of $g(\theta)$ can attain the CRLB? It turns out that the CRLB is achieved only when the definition of the estimator $T(\boldsymbol{X})$ has the special form given in the following theorem.

Theorem 6.1 (CRLB Attainment). *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample with a joint pd $f_n(\boldsymbol{x}; \theta), \theta \in \Theta$. Suppose that Assumption (III), as given above, holds for any statistic. Then, there exists an efficient estimator $T(\boldsymbol{X})$ for $g(\theta)$ if and only if*

$$\partial_\theta \log f_n(\boldsymbol{x}; \theta) = a_n(\theta)[T(\boldsymbol{x}) - g(\theta)], \ \forall \theta \in \Theta, \tag{1}$$

for some function $a_n(\theta) \neq 0$, $\forall \theta \in \Theta$. In addition, if $f_n(\boldsymbol{x}; \theta)$ satisfies (1), then (i) $a_n(\theta)$ equals $I_n(\theta)/g^{'}(\theta)$, and (ii) $f_n(\boldsymbol{x}; \theta)$ is a one-parameter exponential family.

Let's proof this result. Let $S_n = \partial_\theta \log f_n(\boldsymbol{X}; \theta)$ be the Score associated with f_n. Remember that $E(S_n) = 0$, $Var(S_n) = I_n(\theta)$ and, for any statistic T, $Cov(S_n, T) = \partial_\theta E(T)$. Suppose that $T(\boldsymbol{X})$ is an efficient estimator of $g(\theta)$. Then, by definition, (i) $E(T) = g(\theta)$, and (ii) $Var(T) = (g'(\theta))^2/I_n(\theta)$, $\forall \theta \in \Theta$. This implies that $Var(T)Var(S_n) = Cov^2(S_n, T)$. Since Cauchy-Schwarz inequality become an equality only in the case of linear dependence, we conclude that $\exists a \equiv a_n(\theta) \neq 0$ and $b \equiv b_n(\theta)$, such that $S_n = aT + b$. But since $E(S_n) = 0$, we have that $b = -ag(\theta)$. Thus, $S_n = a(T - g(\theta))$. Conversely, suppose that, $\forall \theta \in \Theta$, $\exists a \equiv a_n(\theta) \neq 0$ such that $S_n = a(T - g(\theta))$. The fact that $E(S_n) = 0$ and $Var(S_n) = I_n(\theta)$ implies that $E(T) = g(\theta)$ and $I_n(\theta) = a^2 Var(T)$, receptively. And $Cov(S_n, T) = \partial_\theta E(T)$ implies that $\partial_\theta E(T) = aVar(T)$, which in turn implies that $a = I_n(\theta)/g^{'}(\theta)$. Thus, $Var(T) = (g'(\theta))^2/I_n(\theta)$, which concludes the proof.

The CRLB attainment theorem leads to an explicit constructive procedure for deriving the (efficient) MVUE of $g(\theta)$

when it exists. Namely, put

$$T = g(\theta) + \frac{g'(\theta)}{I_n(\theta)} \partial_\theta \log f_n(\boldsymbol{X}; \theta).$$

If the expression on the right hand side of the equality above *does not depend on* θ, i.e. T as defined above is a statistic, then T is efficient for $g(\theta)$, and so it is also the MVUE.

Example 6.1. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from the exponential distribution. Thus, for some $\theta > 0$, the pd of X_i is given by $f(x; \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$, for $x > 0$.

- Let's try to find the efficient estimator for θ. To do so, put

$$T = \theta + \partial_\theta \log f_n(\boldsymbol{X}; \theta) / I_n(\theta).$$

We have that $\partial_\theta \log f_n(\boldsymbol{X}; \theta) = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i$, and $\partial_\theta^2 \log f_n(\boldsymbol{X}; \theta) = \frac{n}{\theta^2} - \frac{2}{\theta^3} \sum_{i=1}^n X_i$. So, $I_n(\theta) = \frac{n}{\theta^2}$, and thus

$$T = \theta + \frac{\theta^2}{n} \left(-\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^n X_i \right) = n^{-1} \sum_{i=1}^n X_i.$$

$\rightarrow n^{-1} \sum_i X_i$ is the desired estimator.

- Let's now try to find an efficient estimator for $\delta = \frac{1}{\theta}$. Following the same procedure, let

$$\begin{aligned} T &= \frac{1}{\theta} + \frac{\left(\frac{1}{\theta}\right)'}{\frac{n}{\theta^2}} \left(-\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} X_i \right) \\ &= \frac{2}{\theta} - \frac{1}{\theta^2} n^{-1} \sum_{i=1}^{n} X_i. \end{aligned}$$

  The latter is not a statistic, so there is no efficient estimator for $1/\theta$. Note, however, that this does not mean that there is no MVUE. □

7 Multiparameter case

The above theory can be extended to the case of parametric models with a pd $f(x;\boldsymbol{\theta})$ that depends on several parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. We assume in the sequel that the following two regularity conditions are satisfied :

(I) The set $\{x : \ f(x;\boldsymbol{\theta}) > 0\}$ does not depend on $\boldsymbol{\theta}$, and

(II) The operations of integration (or summation) and differentiation by θ_j can be interchanged in $\int f(x;\boldsymbol{\theta})dx$. i.e., $\partial_{\theta_j}\int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j = 1,\ldots,d$.

- The ***Score vector*** of f is defined as the gradient of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$, i.e.

$$\boldsymbol{S} := \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta}) = (S_1,\ldots,S_d)^t,$$

  where $S_j = \partial_{\theta_j}\log f(X;\boldsymbol{\theta})$ is the score for θ_j. It is easy to see that $E(S_j) = 0$, $\forall j$. Thus, $E(\boldsymbol{S}) = \boldsymbol{0}$.

- The ***FI matrix*** contained in X about $\boldsymbol{\theta}$ is defined as

$$\boldsymbol{I}(\boldsymbol{\theta}) := E(\boldsymbol{S}\boldsymbol{S}^t) = Var\left(\boldsymbol{S}\right).$$

  Thus, $\boldsymbol{I}(\boldsymbol{\theta}) = \left[I_{jk}(\boldsymbol{\theta})\right]_{1\le j,k\le d}$, where $I_{jk}(\boldsymbol{\theta}) := E\left(S_j \times S_k\right) = Cov\left(S_j, S_k\right)$.

- Any FI matrix is symmetric and positive semidefinite. Actually, if the considered model is irredundant (i.e. not overparameterized), then its FI matrix is positive definite and therefore invertible. From now on, we'll always assume that the FI matrices we are considering are invertible.

- If $f(x;\boldsymbol{\theta})$ is twice differentiable and double integration and differentiation under the integral sign can be

interchanged, i.e., $\partial_{\theta_k\theta_j} \int f(x;\boldsymbol{\theta})dx = \int \partial_{\theta_k\theta_j} f(x;\boldsymbol{\theta})dx$, $\forall j,k = 1,\ldots,d$, then $I_{jk}(\boldsymbol{\theta}) = -E\left(\partial_{\theta_k} S_j\right)$. Thus,

$$\boldsymbol{I}(\boldsymbol{\theta}) = -E\left(\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})\right),$$

where $\boldsymbol{\nabla}^2_{\boldsymbol{\theta}} \log f(X;\boldsymbol{\theta})$ denotes the Hessian of $\boldsymbol{\theta} \mapsto \log f(X;\boldsymbol{\theta})$.

- To be more explicit about the formulas given above. Let's consider the special case of a two-parameter model with $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $S_1 = \partial_{\theta_1} \log f(X;\boldsymbol{\theta})$, and $S_2 = \partial_{\theta_2} \log f(X;\boldsymbol{\theta})$. Under the regularity assumptions stated above, we can write the FI matrix in any of the following equivalent expressions:

$$\boldsymbol{I}(\theta_1,\theta_2) = E\begin{pmatrix} S_1^2 & S_1S_2 \\ S_1S_2 & S_2^2 \end{pmatrix} = \begin{pmatrix} Var(S_1) & Cov(S_1,S_2) \\ Cov(S_1,S_2) & Var(S_2) \end{pmatrix} = -E\begin{pmatrix} \partial_{\theta_1}S_1 & \partial_{\theta_2}S_1 \\ \partial_{\theta_1}S_2 & \partial_{\theta_2}S_2 \end{pmatrix}.$$

- Let $\boldsymbol{X} = (X_1,\ldots,X_n)$ be an iid sample with joint density $f_n(\boldsymbol{x};\boldsymbol{\theta})$. The Score vector of f_n is defined as $\boldsymbol{S}_n = \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X};\boldsymbol{\theta})$, and the FI matrix contained in $\boldsymbol{X}$ about $\boldsymbol{\theta}$ is given by

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left(\boldsymbol{S}_n\boldsymbol{S}_n^t\right).$$

It can be shown that $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$.

Example 7.1. Normal distribution $N(\mu, \sigma^2)$ with pd $f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.

$$
\begin{aligned}
&S_1 := \partial_\mu \log f = \frac{X-\mu}{\sigma^2} \quad \text{and} \quad S_2 := \partial_{\sigma^2} \log f = -\frac{1}{2\sigma^2} + \frac{(X-\mu)^2}{2\sigma^4}, \text{ so} \\
&I_{11} := -E(\partial_\mu S_1) = -E\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2} \ , \quad I_{12} := -E(\partial_{\sigma^2} S_1) = -E\left(-\frac{X-\mu}{\sigma^4}\right) = 0, \\
&I_{21} := -E(\partial_\mu S_2) = 0 \text{ , and } \quad I_{22} := -E(\partial_{\sigma^2} S_2) = -E\left(\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6}\right) = \frac{1}{2\sigma^4}. \\
\Rightarrow &\boldsymbol{I}(\mu, \sigma^2) := \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \text{ and } \boldsymbol{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}. \square
\end{aligned}
$$

Proposition 7.1 (FI reparametrization – Multiparameter case)**.** *If we rewrite our model in terms of some other parameter $\boldsymbol{\eta} : \boldsymbol{\theta} \mapsto \boldsymbol{\eta}(\boldsymbol{\theta}) = (\eta_1(\boldsymbol{\theta}), \ldots, \eta_d(\boldsymbol{\theta}))$, and denote by $\boldsymbol{I}(\boldsymbol{\eta}) \equiv \boldsymbol{I}(\boldsymbol{\eta}(\boldsymbol{\theta}))$ the FI matrix of the new model/parametrization, then the FI for the two parametrizations (original and new) are related by*

$$
\boldsymbol{I}(\boldsymbol{\theta}) = \dot{\boldsymbol{\eta}}^t(\boldsymbol{\theta}) \boldsymbol{I}(\boldsymbol{\eta}) \dot{\boldsymbol{\eta}}(\boldsymbol{\theta}), \tag{2}
$$

where $\dot{\boldsymbol{\eta}}$ denotes the the $d \times d$ Jacobian matrix of $\boldsymbol{\eta}$, whose $(j,k)-$th element is $\partial_{\theta_k} \eta_j(\boldsymbol{\theta})$.

Example 7.2. Normal distribution $N(\mu, \sigma^2)$.

$$\boldsymbol{I}(\mu, \sigma) = \dot{\boldsymbol{\eta}}^t(\mu, \sigma)\boldsymbol{I}(\mu, \sigma^2)\dot{\boldsymbol{\eta}}(\mu, \sigma),$$

where $\dot{\boldsymbol{\eta}} := (\eta_1, \eta_2)$, with $\eta_1 : (\mu, \sigma) \mapsto \mu$ and $\eta_2 : (\mu, \sigma) \mapsto \sigma^2$.

$$\dot{\boldsymbol{\eta}} = \begin{pmatrix} \partial_\mu \eta_1 & \partial_\sigma \eta_1 \\ \partial_\mu \eta_2 & \partial_\sigma \eta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix}.$$

$$\Rightarrow \boldsymbol{I}(\mu, \sigma) = \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^3} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2\sigma \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \square$$

Theorem 7.1 (Information Inequality – Multiparameter CRLB)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample with joint pd $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. Assume that assumptions (I) and (II) as given above (see the beginning of the current section) hold for $f_n(\boldsymbol{x}; \boldsymbol{\theta})$. Let $\boldsymbol{T} \equiv (T_1(\boldsymbol{X}), \ldots, T_p(\boldsymbol{X}))$ be a statistic. Asuume that (III) $\partial_{\theta_k} E_{\boldsymbol{\theta}}(T_j)$ exists and can be obtained by differentiating under the integral sign. If $\boldsymbol{T}$ is an unbiased estimator of $\boldsymbol{g}(\boldsymbol{\theta}) = \big(g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta})\big)$, i.e. if $E_{\boldsymbol{\theta}}(T_j) = g_j(\boldsymbol{\theta})$, $\forall j, \boldsymbol{\theta}$, then*

$$Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})),$$

where $\boldsymbol{I}_n^{-1}(\boldsymbol{g}(\boldsymbol{\theta})) = \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\dot{\boldsymbol{g}}^t(\boldsymbol{\theta})$, $\boldsymbol{I}_n^{-1}$ is the inverse matrix of $\boldsymbol{I}_n$, and $\dot{\boldsymbol{g}}$ is the $p \times d$ Jacobian matrix of $\boldsymbol{g}$, whose $(j,k)-th$

element is $\partial_{\theta_k} g_j(\boldsymbol{\theta})$. *In particular, if* $\boldsymbol{T}$ *is unbiased for* $\boldsymbol{\theta}$, *then* $Var_{\boldsymbol{\theta}}(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})$.

Above, the notation $\boldsymbol{A} \succeq \boldsymbol{B}$ means that $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite matrix. Consequently, writing $Var(\boldsymbol{T}) \succeq \boldsymbol{I}_n^{-1}$ is equivalent to say that $Var(\boldsymbol{a}^t \boldsymbol{T}) \geq \boldsymbol{a}^t \boldsymbol{I}_n^{-1} \boldsymbol{a}$, $\forall \boldsymbol{a}$.

In simpler terms, the CRLB in the case of multiple parameters tells us how accurately we can estimate several parameters, or even functions of these parameters, **simultaneously**. To see an example of the above inequality, let's consider again the case of $N(\mu, \sigma^2)$, and say that we are interested in estimating $\eta = g(\mu, \sigma) \in \mathbb{R}$. Let $\hat{\eta}$ be any estimator of η. According to Theorem 7.1, we have that

$$\begin{aligned} Var(\hat{\eta}) \geq & \begin{pmatrix} \partial_\mu g(\mu,\sigma) & \partial_\sigma g(\mu,\sigma) \end{pmatrix} \boldsymbol{I}_n^{-1}(\mu,\sigma) \begin{pmatrix} \partial_\mu g(\mu,\sigma) \\ \partial_\sigma g(\mu,\sigma) \end{pmatrix} \\ = & \frac{\sigma^2}{n} \big(\partial_\mu g(\mu,\sigma)\big)^2 + \frac{\sigma^2}{2n} (\partial_\sigma g(\mu,\sigma))^2. \end{aligned}$$

For example, if we are interested in estimating the coefficient of variation (CV), then $g(\mu, \sigma) = \sigma/\mu$, and the CRLB tells us that

$$Var(\hat{\eta}) \geq \frac{\sigma^2}{n} \left(-\frac{\sigma}{\mu^2}\right)^2 + \frac{\sigma^2}{2n} \left(\frac{1}{\mu}\right)^2 = \frac{\sigma^2}{n\mu^2} \left(\frac{\sigma^2}{\mu^2} + \frac{1}{2}\right).$$

Thus, no unbiased estimator of the CV can have a variance smaller than the right hand side of the inequality above.

Or, in other words, $\frac{\sigma^2}{n\mu^2}\left(\frac{\sigma^2}{\mu^2}+\frac{1}{2}\right)$ is the best possible precision (smallest variance) that any unbiased estimator of the CV can achieve, when data is normally distributed.

To see an interesting implication of the Multiparameter CRLB, let's consider the two-parameters case ($d = 2$), with $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Let's write down the FI matrix and its inverse as follows

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = n \begin{pmatrix} Var(S_1) & Cov(S_1, S_2) \\ Cov(S_2, S_1) & Var(S_2) \end{pmatrix} \equiv n \begin{pmatrix} I_{11} & I_{12} \\ I_{12} & I_{22} \end{pmatrix}.$$

$$\boldsymbol{I}_n^{-1}(\boldsymbol{\theta}) = \frac{1}{n(I_{11}I_{22} - I_{12}^2)} \begin{pmatrix} I_{22} & -I_{12} \\ -I_{12} & I_{11} \end{pmatrix}.$$

Assume that θ_1 is our *primary parameter of interest* and θ_2 is a *nuisance parameter*. In the normal model, for example, we might be interested only in the mean μ, while σ^2 is a a "nuisance", which is there only to make the model correct. In such a context, we can think of two situations:

- θ_2 *is known*: in this case θ_1 is the only (unknown) parameter of our model, and the (univariate) information

inequality (see Theorem 4.1) tells us that, for any unbiased estimator T_1 of θ_1,

$$Var(T_1) \geq \frac{1}{nI_{11}}.$$

- θ_2 *is unknown*: in this case, by applying Theorem 7.1 with $g(\theta_1, \theta_2) = \theta_1$, the information inequality tells us that, for any unbiased estimator T_1 of θ_1,

$$Var(T_1) \geq (1,0) I_n^{-1}(\boldsymbol{\theta})(1,0)^t = \frac{1}{nI_{11}^*},$$

where $I_{11}^* = \dfrac{I_{11}I_{22} - I_{12}^2}{I_{22}} = I_{11}\left(1 - Cor^2(S_1, S_2)\right)$.

Observe that $I_{11}^* \leq I_{11}$, with equality if and only if S_1 and S_2 are uncorrelated (i.e. *diagonal FI matrix*).
$\rightarrow$ *nuisance parameter(s) cause loss of information*, resulting in an increase in the lower bound of the variance (of other model parameters).

Example 7.3. Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an iid sample from $N(\mu, \sigma^2)$, where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. We have seen that

$$\boldsymbol{I}_n(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

Notice that this is a *diagonal* matrix.

- Suppose that μ is our parameter of interest. The information inequality tells us that $Var(\hat{\mu}) \geq \frac{\sigma^2}{n}$, for any unbiased estimator $\hat{\mu}$ of μ, *whether* σ^2 *is known or not*. Since $E(\overline{X}_n) = \mu$ and $Var(\overline{X}_n) = \sigma^2/n$, $\overline{X}_n$ is efficient for μ, whether σ^2 is known or not.
- Suppose that σ^2 is our parameter of interest. The information inequality tells us that $Var(\hat{\sigma}^2) \geq \frac{2\sigma^4}{n}$, for any unbiased estimator $\hat{\sigma}^2$ of σ^2, *whether* μ *is known or not*. Let $\tilde{\sigma}_n^2 = n^{-1}\sum_i (X_i - \mu)^2$. We have seen that $E(\tilde{\sigma}_n^2) = \sigma^2$ and $Var(\tilde{\sigma}_n^2) = 2\sigma^4/n$, so $\tilde{\sigma}_n^2$ is efficient for σ^2 *when* μ *is known*. When μ is unknown, $\tilde{\sigma}_n^2$ is not an estimator. In this case, it is "natural" to estimate σ^2 by $S_n^2 = \frac{1}{n-1}\sum_i (X_i - \overline{X}_n)^2$. This latter is unbiased but (marginally) not efficient since $Var(S_n^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n}$. Note that it can be shown that S_n^2 is the MVUE of σ^2. □

Example 7.4. Let $\boldsymbol{X}_i = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution $Mult(m, (\pi_1, \pi_2, \pi_3))$, with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

where $m \geq 1$, $\boldsymbol{x} = (x_1, x_2)$, and $\boldsymbol{\pi} = (\pi_1, \pi_2)$ is unknown; with $x_1 = 0, \ldots, m$, $x_2 = 0, \ldots, m$, $x_3 = m - x_1 - x_2$, $\pi_1 \in (0,1)$, $\pi_2 \in (0,1)$, and $\pi_3 = 1 - \pi_1 - \pi_2$. We have that

$$S_1 := \partial_{\pi_1} \log f(\boldsymbol{X}; \boldsymbol{\pi}) = \frac{X_1}{\pi_1} - \frac{X_3}{\pi_3} \text{ and } S_2 := \partial_{\pi_2} \log f(\boldsymbol{X}; \boldsymbol{\pi}) = \frac{X_2}{\pi_2} - \frac{X_3}{\pi_3}$$

The score vector is given by $\boldsymbol{S} = (S_1, S_2)^t$, and, knowing that $E(X_k) = m\pi_k$, $k = 1, 2, 3$, the FI matrix is given by

$$\boldsymbol{I}(\pi_1, \pi_2) = -E\begin{pmatrix} \partial_{\pi_1} S_1 & \partial_{\pi_2} S_1 \\ \partial_{\pi_1} S_2 & \partial_{\pi_2} S_2 \end{pmatrix} = m\begin{pmatrix} \frac{1}{\pi_1} + \frac{1}{\pi_3} & \frac{1}{\pi_3} \\ \frac{1}{\pi_3} & \frac{1}{\pi_2} + \frac{1}{\pi_3} \end{pmatrix}.$$

And its inverse is given by

$$\boldsymbol{I}^{-1}(\pi_1, \pi_2) = m^{-1}(\pi_1\pi_2\pi_3)\begin{pmatrix} \frac{1}{\pi_2} + \frac{1}{\pi_3} & -\frac{1}{\pi_3} \\ -\frac{1}{\pi_3} & \frac{1}{\pi_1} + \frac{1}{\pi_3} \end{pmatrix} = m^{-1}\begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}.$$

Suppose that π_1 is our parameter of interest and π_2 is a nuisance parameter. According to the FI matrix above and the information inequality, we can say that, for any unbiased estimator $\hat{\pi}_1$ of π_1,

$$Var(\hat{\pi}_1) \geq n^{-1}m^{-1}\left(\frac{1}{\pi_1} + \frac{1}{\pi_3}\right)^{-1} = \frac{\pi_1(1-\pi_1-\pi_2)}{nm(1-\pi_2)}, \text{ if } \pi_2 \text{ is known.}$$
$$Var(\hat{\pi}_1) \geq \frac{\pi_1(1-\pi_1)}{nm}, \text{ if } \pi_2 \text{ is unknown. } \square$$

Theorem 7.2 (CRLB Attainment–Multiparameter case)**.** *Under the regularity conditions stated above, if the "statistic" $\boldsymbol{T}$ as*

defined by

$$\boldsymbol{T} = \boldsymbol{g}(\boldsymbol{\theta}) + \dot{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{I}_n^{-1}(\boldsymbol{\theta})\boldsymbol{\nabla}_{\boldsymbol{\theta}} \log f_n(\boldsymbol{X};\boldsymbol{\theta}),$$

do not depend on $\boldsymbol{\theta}$*, then* $\boldsymbol{T}$ *is efficient for* $\boldsymbol{g}(\boldsymbol{\theta})$.

Example 7.5. Consider the trinomial distribution of our previous example. Let's see if we can find an efficient estimator for $\boldsymbol{\pi}^t = (\pi_1, \pi_2)$. We have that

$$\boldsymbol{\pi} + \boldsymbol{I}_n^{-1}(\boldsymbol{\pi})\boldsymbol{\nabla}_{\boldsymbol{\pi}} \log f_n(\boldsymbol{X};\boldsymbol{\pi}) = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix} + n^{-1}m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix} \begin{pmatrix} \frac{\sum_i X_{1i}}{\pi_1} - \frac{\sum_i X_{3i}}{\pi_3} \\ \frac{\sum_i X_{2i}}{\pi_2} - \frac{\sum_i X_{3i}}{\pi_3} \end{pmatrix} = \begin{pmatrix} \frac{\sum_i X_{1i}}{nm} \\ \frac{\sum_i X_{2i}}{nm} \end{pmatrix}.$$

So $\left(\frac{\overline{X}_1}{m}, \frac{\overline{X}_2}{m}\right)$ is efficient for (π_1, π_2). □