# Minimum Variance Unbiased Estimation

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

# Optimality in the Decision Theory Framework

# Decision Theory Framework

Saw how point estimation can be seen as a game: Nature VS Statistician. The decision theory framework includes:

- A *family of distributions* $\mathcal{F}$, usually assumed to admit densities (frequencies) and a *parameter space* $\Theta \subseteq \mathbb{R}^p$ which parametrizes the family $\mathcal{F} = \{F_\theta\}_{\theta \in \Theta}$.
- A *data space* $\mathcal{X}$, on which the parametric family is supported.
- An *action space* $\mathcal{A}$, which represents the space of possible *actions* available to the statistician. In point estimation $\mathcal{A} \equiv \Theta$
- A *loss function* $\mathcal{L} : \Theta \times \mathcal{A} \to \mathbb{R}^+$. This represents the lost incurred when estimating $\theta \in \Theta$ by $\alpha \in \mathcal{A}$.
- A set $\mathcal{D}$ of *decision rules*. Any $\delta \in \mathcal{D}$ is a (measurable) function $\delta : \mathcal{X} \to \mathcal{A}$. In point estimation decision rules are simply estimators.

Performance of decision rules was to be judged by the risk they induce:

$$R(\theta, \delta) = \mathbb{E}_\theta[\mathcal{L}(\theta, \delta(\mathbf{X}))], \quad \boldsymbol{\theta} \in \Theta, X \sim F_\theta, \delta \in \mathcal{D}$$

# Optimality in Point Estimation

An optimal decision rule would be one that uniformly minimizes risk:

$$R(\theta, \delta_{\text{OPTIMAL}}) \leq R(\theta, \delta), \quad \forall \theta \in \Theta \ \& \ \forall \delta \in \mathcal{D}.$$

But such rules can very rarely be determined.

$\hookrightarrow$ optimality becomes a *vague* concept

$\hookrightarrow$ can be made precise in many ways...

Avenues to studying optimal decision rules include:

- **Restricting attention to global risk criteria rather than local**
  - $\hookrightarrow$ Bayes and minimax risk.
- **Focusing on restricted classes of rules** $\mathcal{D}$
  - $\hookrightarrow$ e.g. Minimum Variance Unbiased Estimation.
- **Studying risk behaviour asymptotically** $(n \to \infty)$
  - $\hookrightarrow$ e.g. Asymptotic Relative Efficiency.

# Uniform Optimality in Unbiased Quadratic Estimation

# Unbiased Estimators under Quadratic Loss

- **Focus on Point Estimation**

  1. Assume that $F_\theta$ is known up to the parameter $\theta$ which is unknown
  2. Let $(x_1, ..., x_n)$ be a realization of $\mathbf{X} \sim F_\theta$ which is available to us
  3. Estimate the value of $\theta$ that generated the sample given $(x_1, ..., x_n)$

- **Focus on Quadratic Loss**

  Error incurred when estimating $\theta$ by $\hat{\theta} = \delta(\mathbf{X})$ is

  $$\mathcal{L}(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

  giving MSE as risk $R(\theta, \hat{\theta}) = \mathbb{E}_\theta \|\theta - \hat{\theta}\|^2 = \text{Variance} + \text{Bias}^2$.

- **RESTRICT class of estimators (=decision rules)**

  Consider ONLY *unbiased* estimators: $\mathcal{D} := \{\delta : \mathcal{X} \to \Theta | \mathbb{E}_\theta[\delta(\mathbf{X})] = \theta\}$.

# Comments on Unbiasedness

- Unbiasedness requirement is one means of reducing the class of rules/estimators we are considering
  $\hookrightarrow$ Other requirements could be invariance or equivariance, e.g.

$$\delta(\mathbf{X}+\mathbf{c}) = \delta(\mathbf{X}) + \mathbf{c}$$

- Risk reduces to variance since bias is zero.
- Unbiased Estimators may not exist in a particular problem
- Unbiased Estimators may be silly for a particular problem
- However unbiasedness can be a reasonable/natural requirement in a wide class of point estimation problems.
- Not necessarily a sensible requirement
  $\hookrightarrow$ e.g. violates "likelihood principle"
- Unbiasedness can be defined for more general loss functions, but not as conceptually clear (and with tractable theory) as for quadratic loss.
  $\hookrightarrow$ $\delta$ is unbiased under $\mathcal{L}$ if $\mathbb{E}_\theta[\mathcal{L}(\theta',\delta)] \geq \mathbb{E}_\theta[\mathcal{L}(\theta,\delta)] \quad \forall\ \theta,\theta' \in \Theta$.

**Example (Unbiased Estimators Need not Exist)**

Let $X \sim$Binomial$(n, \theta)$, with $\theta$ unknown but $n$ known. We wish to estimate

$$\psi = \sin\theta$$

We require that our estimator $\delta(X)$ be unbiased, $\mathbb{E}_\theta[\delta] = \psi = \sin\theta$. Such an estimator satisfies

$$\sum_{x=0}^{n} \delta(x) \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad = \quad \sin\theta$$

but this cannot hold for all $\theta$, since the sine function cannot be represented as a finite polynomial.

The class of unbiased estimators in this case is empty.

## Example (Unbiased Estimators May Be "Silly")

Let $X \sim$Poisson$(\lambda)$. We wish to estimate the parameter

$$\psi = e^{-2\lambda}.$$

If $\delta(X)$ is an unbiased estimator of $\psi$, then

$$\begin{aligned}\sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} e^{-\lambda} = e^{-2\lambda} \implies \sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} &= e^{-\lambda} \\ \implies \sum_{x=0}^{\infty} \delta(x) \frac{\lambda^x}{x!} &= \sum_{x=0}^{\infty} (-1)^x \frac{\lambda^x}{x!}\end{aligned}$$

so that $\delta(X) = (-1)^X$ is the only unbiased estimator of $\psi$.

But $0 < \psi < 1$ for $\lambda > 0$, so this is clearly a silly estimator

# Comments on Unbiased Estimators

## Example (A Non-Trivial Example)

Let $X_1, \ldots, X_n$ be iid random variables with density

$$f(x;\mu) = e^{-(x-\mu)}, \quad x \geq \mu \in \mathbb{R}.$$

Two possible unbiased estimators are

$$\hat{\mu} = X_{(1)} - \frac{1}{n} \quad \& \quad \tilde{\mu} = \bar{X} - 1.$$

In fact, $t\hat{\mu} + (1-t)\tilde{\mu}$ is unbiased for any $t$. Simple calculations reveal

$$R(\mu, \hat{\mu}) = \mathrm{Var}(\hat{\mu}) = \frac{1}{n^2} \quad \& \quad R(\mu, \tilde{\mu}) = \mathrm{Var}(\tilde{\mu}) = \frac{1}{n}$$

so that $\hat{\mu}$ dominates $\tilde{\mu}$. Will it dominate any other unbiased estimator?

(note that $\hat{\mu}$ depends only on the one-dimensional sufficient statistic $X_{(1)}$)

# Unbiased Estimation and Sufficiency

**Theorem (Rao-Blackwell Theorem)**

*Let $\mathbf{X}$ be distributed according to a distribution depending on an unknown parameter $\theta$ and let $T$ be a sufficient statistic for $\theta$. Let $\delta$ be decision rule such that*

1. $\mathbb{E}_\theta[\delta(\mathbf{X})] = g(\theta)$ *for all* $\theta$
2. $Var_\theta(\delta(\mathbf{X})) < \infty$, *for all* $\theta$.

*Then $\delta^* := \mathbb{E}[\delta|T]$ is an unbiased estimator of $g(\theta)$ that dominates $\delta$, i.e.*

1. $\mathbb{E}_\theta[\delta^*(\mathbf{X})] = g(\theta)$ *for all* $\theta$.
2. $Var_\theta(\delta^*(\mathbf{X})) \leq Var_\theta(\delta(\mathbf{X}))$ *for all* $\theta$.

*Moreover, inequality is replaced by equality if and only if $\mathbb{P}_\theta[\delta^* = \delta] = 1$.*

- The theorem indicates that any candidate minimum variance unbiased estimator should be a function of the sufficient statistic.
- Intuitively, an estimator that takes into account aspects of the sample that are irrelevant with respect to $\theta$, can always be improved

**Proof.**

Since $T$ is sufficient for $\theta$, $\mathbb{E}[\delta|T=t]=h(t)$ is independent of $\theta$, so that $\delta^*$ is well-defined as a statistic (depends only on $\mathbf{X}$). Then,

$$\mathbb{E}_\theta[\delta^*(\mathbf{X})]=\mathbb{E}_\theta[\mathbb{E}[\delta(\mathbf{X})|T(\mathbf{X})]]=\mathbb{E}_\theta[\delta(\mathbf{X})]=g(\theta).$$

Furthermore, we have

$$\begin{aligned}\mathrm{Var}_\theta(\delta)=\mathrm{Var}_\theta[\mathbb{E}(\delta|T)]+\mathbb{E}_\theta[\mathrm{Var}(\delta|T)] &\geq \mathrm{Var}_\theta[\mathbb{E}(\delta|T)]\\ &= \mathrm{Var}_\theta(\delta^*)\end{aligned}$$

In addition, note that

$$\mathrm{Var}(\delta|T):=\mathbb{E}[(\delta-\mathbb{E}[\delta|T])^2|T]=\mathbb{E}[(\delta-\delta^*)^2|T]$$

so that $\mathbb{E}_\theta[\mathrm{Var}(\delta|T)]=\mathbb{E}_\theta(\delta-\delta^*)^2>0$ unless if $\mathbb{P}_\theta(\delta^*=\delta)=1$. $\square$

**Exercise**

Show that $\mathrm{Var}(Y)=\mathbb{E}[\mathrm{Var}(Y|\mathbf{X})]+\mathrm{Var}[\mathbb{E}(Y|\mathbf{X})]$ when $\mathrm{Var}(Y)<\infty$.

# The role of sufficiency and "Rao-Blackwellization"

# Unbiasedness and Sufficiency

- Any admissible unbiased estimator should be a function of a sufficient statistic
  - $\hookrightarrow$ If not, we can dominate it by its conditional expectation given a sufficient statistic.
- But is any function of a sufficient statistic admissible? (provided that it is unbiased)

Suppose that $\delta$ is an unbiased estimator of $g(\theta)$ and $T$, $S$ are $\theta$-sufficient.

- What is the relationship between $\mathrm{Var}_\theta(\underbrace{\mathbb{E}[\delta|T]}_{\delta^*_T}) \overset{?}{\gtreqless} \mathrm{Var}_\theta(\underbrace{\mathbb{E}[\delta|S]}_{\delta^*_S})$
- Intuition suggests that whichever of $T, S$ carries the least irrelevant information (in addition to the relevant information) should "win"
  - $\hookrightarrow$ More formally, if $T = h(S)$ then we should expect that $\delta^*_T$ dominate $\delta^*_S$.

# Unbiasedness and Sufficiency

**Proposition**

Let $\delta$ be an unbiased estimator of $g(\theta)$ and for $T, S$ two $\theta$-sufficient statistics define

$$\delta^*_T := \mathbb{E}[\delta | T] \quad \& \quad \delta^*_S := \mathbb{E}[\delta | S].$$

Then, the following implication holds

$$T = h(S) \implies \mathrm{Var}_\theta(\delta^*_T) \leq \mathrm{Var}_\theta(\delta^*_S)$$

1. Essentially this means that the best possible "Rao-Blackwellization" is achieved by conditioning on a minimal sufficient statistic.
2. This does not necessarily imply that for $T$ minimally sufficient and $\delta$ unbiased, $\mathbb{E}[\delta | T]$ has minimum variance.
   - $\hookrightarrow$ In fact it does not even imply that $\mathbb{E}[\delta | T]$ is admissible.

## Proof.

Recall the *tower property* of conditional expectation: if $Y = f(X)$, then

$$\mathbb{E}[Z|Y] = \mathbb{E}\{\mathbb{E}(Z|X)|Y\}.$$

Since $T = f(S)$ we have

$$\begin{aligned} \delta_T^* &= \mathbb{E}[\delta|T] \\ &= \mathbb{E}[\mathbb{E}(\delta|S)|T] \\ &= \mathbb{E}[\delta_S^*|T] \end{aligned}$$

The conclusion now follows from the Rao-Blackwell theorem. $\square$

## A mathematical remark

To better understand the tower property intuitively, recall that $\mathbb{E}[Z|Y]$ is the minimizer of $\mathbb{E}[(Z - \varphi(Y))^2]$ over all (measurable) functions $\varphi$ of $Y$. You can combine that with the fact that $\sqrt{\mathbb{E}[(X - Y)^2]}$ defines a Hilbert norm on random variables with finite variance to get geometric intuition.

# The role of completeness in Uniform Optimality

# Completeness, Sufficiency, Unbiasedness, and Optimality

**Theorem (Lehmann-Scheffé Theorem)**

*Let $T$ be a complete sufficient statistic for $\theta$ and let $\delta$ be a statistic such that $\mathbb{E}_\theta[\delta] = g(\theta)$ and $Var_\theta(\delta) < \infty$, $\forall \theta \in \Theta$. If $\delta^* := \mathbb{E}[\delta | T]$ and $V$ is any other unbiased estimator of $g(\theta)$, then*

1. $Var_\theta(\delta^*) \leq Var_\theta(V)$, $\forall \theta \in \Theta$
2. $Var_\theta(\delta^*) = Var_\theta(V) \implies \mathbb{P}_\theta[\delta^* = V] = 1$.

*That is, $\delta^* := \mathbb{E}[\delta | T]$ is the unique Uniformly Minimum Variance Unbiased Estimator of $g(\theta)$.*

- The theorem says that if a complete sufficient statistic $T$ exists, then the MVUE of $g(\theta)$ (if it exists) must be a function of $T$.
- Moreover it establishes that whenever $\exists$ UMVUE, it is unique.
- Can be used to examine whether unbiased estimators exist at all: if a complete sufficient statistic $T$ exists, but there exists no function $h$ with $\mathbb{E}[h(T)] = g(\theta)$, then no unbiased estimator of $g(\theta)$ exists.

## Proof.

To prove (1) we go through the following steps:

- Take $V$ to be any unbiased estimator with finite variance.
- Define its "Rao-Blackwellized" version $V^* := \mathbb{E}[V|T]$
- By unbiasedness of both estimators,

$$0 = \mathbb{E}_\theta[V^* - \delta^*] = \mathbb{E}_\theta[\mathbb{E}[V|T] - \mathbb{E}[\delta|T]] = \mathbb{E}_\theta[h(T)], \quad \forall \theta \in \Theta.$$

- By completeness of $T$ we conclude $\mathbb{P}_\theta[h(T) = 0] = 1$ for all $\theta$.
- In other words, $\mathbb{P}_\theta[V^* = \delta^*] = 1$ for all $\theta$.
- But $V^*$ dominates $V$ by the Rao-Blackwell theorem.
- Hence $\text{Var}_\theta(\delta^*) = \text{Var}_\theta(V^*) \leq \text{Var}_\theta(V)$.

For part (2) (the uniqueness part) notice that from our reasoning above

- $\text{Var}_\theta(V) = \text{Var}_\theta(\delta^*) \implies \text{Var}_\theta(V) = \text{Var}_\theta(V^*)$
- But Rao-Blackwell theorem says
  $\text{Var}_\theta(V) = \text{Var}_\theta(V^*) \iff \mathbb{P}_\theta[V = V^*] = 1$.

$\square$

# Completeness, Sufficiency, Unbiasedness, and Optimality

Taken together, the Rao-Blackwell and Lehmann-Scheffé theorems also suggest approaches to finding UMVUE estimators when a complete sufficient statistic $T$ exists:

1. Find a function $h$ such that $\mathbb{E}_\theta[h(T)] = g(\theta)$. If $\text{Var}_\theta[h(T)] < \infty$ for all $\theta$, then $\delta = h(T)$ is the unique UMVUE of $g(\theta)$.
   - $\hookrightarrow$ The function $h$ can be found by solving the equation $\mathbb{E}_\theta[h(T)] = g(\theta)$ or by an educated guess.
2. Given an unbiased estimator $\delta$ of $g(\theta)$, we may obtain the UMVUE by "Rao-Blackwellizing" with respect to the complete sufficient statistic:

**Example (Bernoulli Trials)**

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. What is the UMVUE of $\theta^2$?

- By the Neyman factorization theorem $T = X_1 + \ldots + X_n$ is sufficient,
- Since the distribution of $(X_1, ..., X_n)$ is a 1-parameter exponential family, $T$ is also complete.

## Example (Bernoulli Trials)

First suppose that $n = 2$. If a UMVUE exists, it must be of the form $h(T)$ with $h$ satisfying

$$\theta^2 = \sum_{k=0}^{2} h(k) \binom{2}{k} \theta^k (1-\theta)^{2-k}$$

It is easy to see that $h(0) = h(1) = 0$ while $h(2) = 1$. Thus, for $n = 2$, $h(T) = T(T-1)/2$ is the unique UMVUE of $\theta^2$.

For $n > 2$, set $\delta = \mathbf{1}\{X_1 + X_2 = 2\}$ and note that this is an unbiased estimator of $\theta^2$. By the Lehmann-Scheffé theorem, $\delta^* = \mathbb{E}[\delta|T]$ is the unique UMVUE estimator of $\theta^2$. We have

$$\begin{aligned}
\mathbb{E}[S|T=t] &= \mathbb{P}[X_1 + X_2 = 2 | T = t] \\
&= \frac{\mathbb{P}_\theta[X_1 + X_2 = 2, X_3 + \ldots + X_n = t-2]}{\mathbb{P}_\theta[T = t]} \\
&= \left\{\begin{matrix} 0 & \text{if } t \leq 1 \\ \binom{n-2}{t-2} / \binom{n}{t} & \text{if } t \geq 2 \end{matrix}\right\} = \frac{t(t-1)}{n(n-1)}.
\end{aligned}$$

# Lower Bounds for the Risk and Achieving them

# Variance Lower Bounds for Unbiased Estimators

- Often $\rightarrow$ minimal sufficient statistic exists but is not complete.
  $\hookrightarrow$ Cannot appeal to the Lehmann-Scheffé theorem in search of a UMVUE.
- However, if we could establish a *lower bound* for the variance as a function of $\theta$, than an estimator achieving this bound will be the unique UMVUE.

**The Aim**

For iid $X_1, ..., X_n$ with density (frequency) depending on $\theta$ unknown, we want to establish conditions under which

$$Var_\theta[\delta] \geq \phi(\theta), \quad \forall\theta$$

for any unbiased estimator $\delta$. We also wish to determine $\phi(\theta)$.

Let's take a closer look at this...

## Cauchy-Schwarz Bounds

**Theorem (Cauchy-Schwarz Inequality)**

*Let $U, V$ be random variables with finite variance. Then,*

$$Cov(U, V) \leq \sqrt{Var(U)Var(V)}$$

The theorems yields an immediate lower bound for the variance of an unbiased estimator $\delta_0$:

$$\mathrm{Var}_\theta(\delta_0) \geq \frac{\mathrm{Cov}^2_\theta(\delta_0, U)}{\mathrm{Var}_\theta(U)}$$

which is valid for any random variable $U$ with $\mathrm{Var}_\theta(U) < \infty$ for all $\theta$.

- The bound can be made tight be choosing a suitable $U$.
- However this is still not very useful as it falls short of our aim
    - The bound will be specific to $\delta_0$, while we want a bound that holds for any unbiased estimator $\delta$.
- Is there a smart choice of $U$ for which $\mathrm{Cov}_\theta(\delta_0, U)$ depends on $g(\theta) = \mathbb{E}_\theta(\delta_0)$ only? (and so is not specific to $\delta_0$)

# Optimizing the Cauchy-Schwartz Bound

Assume that $\theta$ is real and the following regularity conditions hold

**Regularity Conditions**

(C1) The support of $A := \{\mathbf{x} : f(\mathbf{x};\theta) > 0\}$ is independent of $\theta$

(C2) $f(\mathbf{x};\theta)$ is differentiable w.r.t. $\theta$, $\forall \theta \in \Theta$

(C3) $\mathbb{E}_\theta \left[\frac{\partial}{\partial\theta} \log f(\mathbf{X};\theta)\right] = 0$

(C4) For a statistic $T = T(\mathbf{X})$ with $\mathbb{E}_\theta|T| < \infty$ and $g(\theta) = \mathbb{E}_\theta T$ differentiable,

$$g'(\theta) = \mathbb{E}_\theta\left[T \frac{\partial}{\partial\theta} \log f(\mathbf{X};\theta)\right], \quad \forall\theta$$

To make sense of (C3) and (C4), suppose that $f(\cdot;\theta)$ is a density. Then

$$\frac{d}{d\theta}\int T(\mathbf{x}) f(\mathbf{x};\theta) d\mathbf{x} \overset{!}{=} \int T(\mathbf{x}) \frac{f(x;\theta)}{f(x;\theta)} \frac{d}{d\theta} f(\mathbf{x};\theta) dx = \mathbb{E}_\theta\left[T \frac{\partial}{\partial\theta} \log f(\mathbf{X};\theta)\right]$$

provided integration/differentiation can be interchanged.

## Theorem

*Let* $\mathbf{X} = (X_1, ..., X_n)$ *have joint density (frequency)* $f(\mathbf{x}; \theta)$ *satisfying conditions (C1), (C2) and (C3). If the statistic* $T$ *satisfies condition* $(C4)$, *then*

$$Var_\theta(T) \geq \frac{[g'(\theta)]^2}{I(\theta)}$$

*with* $I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2\right]$

## Proof.

By the Cauchy-Schartz inequality with $U = \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{\mathrm{Cov}_\theta^2\left(T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)}{\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)}$$

Since $\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] = 0$ we have $\mathrm{Var}_\theta\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right) = I(\theta)$.

Also, observe that

$$\begin{aligned}
\mathrm{Cov}_\theta\left(T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right) &= \mathbb{E}_\theta\left[T\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right] \\
&= \frac{d}{d\theta}\mathbb{E}_\theta[T] \\
&= g'(\theta)
\end{aligned}$$

which completes the proof. □

# The Cramér-Rao Lower Bound

When is the Cramér-Rao lower bound achieved?

$$\text{if} \qquad \mathrm{Var}_\theta[T] = \frac{[g'(\theta)]^2}{I(\theta)}$$

$$\text{then} \qquad \mathrm{Var}_\theta[T] = \frac{\mathrm{Cov}_\theta^2\left[T, \frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right]}{\mathrm{Var}_\theta\left[\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right]}$$

which occurs if and only if $\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)$ is a linear function of $T$ (correlation 1). That is, w.p.1:

$$\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta) = A(\theta)T(\mathbf{x}) + B(\theta)$$

Solving this differential equation yields, for all $\mathbf{x}$,

$$\log f(\mathbf{x};\theta) = A^*(\theta)T(\mathbf{x}) + B^*(\theta) + S(\mathbf{x})$$

so that $\mathrm{Var}_\theta(T)$ attains the lower bound if and only if the density (frequency) of $\mathbf{X}$ is a one-parameter exponential family as above

# The Cramér-Rao bound asymptotically

If we have many IID observations: $n \to \infty$, the Fisher information is:

$$I_{X_1,\dots,X_n} = nI_{X_1}(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X};\theta)\right)^2\right]$$

More generally, the Fisher information of two independent observations is the sum of the Fisher informations of each one.

**Definition**

The *asymptotic efficiency* of a sequence of estimators $\hat{\theta}_n$ defined on IID observations $X_1 \dots X_n$ is defined by the ratio:

$$\mathrm{var}(\hat{\theta}_n)/\left[nI_{X_1}(\theta)\right]^{-1}$$

The asymptotic efficiency measures whether a given estimator saturates the Cramer-Rao bound or falls short.

# Summary

Unbiasedness is one criteria we can follow to find a good estimator.

Rao-Blackwell*ising* an unbiased estimator with a sufficient statistic gives a better estimator (lower variance).

If there is a complete sufficient statistic, there might exist a unique unbiased estimator. (MVUE)

Recall that, besides exponential families, a complete and sufficient statistic rarely exists !

More generally, all estimators must obey the Cramer-Rao lower bound. If we can prove that an estimator saturates the Cramer-Rao bound, then that proves that it is optimal.

# The MLE dominates

From the results presented in this lecture, we see that the MLE is a great estimator:

- It automatically depends only on a minimally sufficient statistic: its already Rao-Blackwellised !
- If there is a complete sufficient statistic AND the MLE is unbiased, then it is the UMVE.

- Even without completeness, the MLE is asympotically:
  - Unbiased: $\mathbb{E}(\hat{\theta}) = \theta$.
  - Gaussian with variance: $(nI(\theta))^{-1}$.

  Asymptotically, it saturates the Cramer-Rao bound !

It is a great estimator **if the model is correctly specified** !