Sufficient statistics, MLE and Cramér–Rao inequality

Marianna Bolla, Prof, DSc.
Institute of Mathematics, BME

December 19, 2020

To explain the theorems of C.R. Rao, first we introduce parametric statistical spaces and the notion of a sufficient statistic that is very useful in parameter estimation.

1 Basic notions, statistics

We take an i.i.d. sample $X_1, \dots, X_n$ from a population with distribution $\mathbb{P}_\theta$, where θ is unknown parameter, and it is in the *parameter space* Θ, so $\theta \in \Theta$. For example, if $\mathbf{X} := (X_1, \dots, X_n)$ follow Poisson distribution, then the parameter, now denoted by λ is in the parameter space $\Theta = (0, \infty)$. The *sample space* is the set of all possible n-tuples $(x_1, \dots, x_n)$ that are possible *realizations* of the sample. For fixed *simple size* n, let $\mathcal{X} \subset \mathbb{R}^n$ denote the *sample space*, that is the set of all possible realizations. In the Poisson case, it is $\mathcal{X} = \{0, 1, 2, \dots\}^n$.

Definition 1 *The likelihood function for* $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{X}$ *and* $\theta \in \Theta$ *is* $L_\theta(\mathbf{x}) = \mathbb{P}_\theta(\mathbf{X} = \mathbf{x}) = \prod_{i=1}^n \mathbb{P}_\theta(X_i = x_i) = \prod_{i=1}^n p_\theta(x_i)$ *in the discrete, and* $L_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$ *in the absolutely continuous case, where* $p_\theta(x)$ *is the probability mass function (p.m.f.) in the discrete, and* $f_\theta(x)$ *is the probability density function (p.d.f.) in the continuous case.*

It is customary to call $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ a *statistical space*, where $\mathcal{B}$ contains the σ-algebra of the Borel sets within $\mathcal{X}$, and

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

is the parametric set of probability measures such that

$$\mathbb{P}_\theta((X_1, \dots, X_n) \in B), \quad B \in \mathcal{B}$$

is the induced probability that can be calculated (with summation or integration) by means of the likelihood function, for any $\theta \in \Theta$.

Now we organize the sample entries into a *statistic* $T := T(X_1, \ldots, X_n) = T(\mathbf{X})$ such that, by this compression, we would not loose any information for the parameter.

Definition 2 *The statistic $T(\mathbf{X})$ is sufficient for θ if the distribution of $\mathbf{X}$ conditioned on $T(\mathbf{X})$ does not depend on θ.*

It means that T contains all the information that can be retrieved from the sample for the parameter.

Theorem 1 (Neyman–Fisher factorization) *The statistic $T(\mathbf{X})$ is sufficient for θ if and only if the likelihood function can be factorized as*

$$L_\theta(\mathbf{x}) = g_\theta(T(\mathbf{x})) \cdot h(\mathbf{x}), \quad \forall \theta \in \Theta, \quad \mathbf{x} \in \mathcal{X}$$

with some measurable, nonnegative real functions g and h.

Sufficient statistics are many, even based on the same sample and for the same parameter (e.g., the ordered sample is such). A sufficient statistic is *minimal* if it is the function of any other sufficient statistic. Minimal sufficient statistic always exists, and it is unique up to equivalence.

Definition 3 *The statistic $T = T(\mathbf{X})$ is complete for θ if $\mathbb{E}_\theta g(T) = 0$, $\forall \theta$ implies that $g(T(\mathbf{X})) = 0$ almost surely (with probability 1).*

Theorem 2 *If a statistic is sufficient and complete, then it is also minimal sufficient.*

Definition 4 *The $\mathbb{P}_\theta$ distribution belongs to the exponential family if its p.m.f. or p.d.f. has the form*

$$c(\theta) \cdot \exp\left[\sum_{j=1}^{k} a_j(\theta) \cdot T_j(x)\right] \cdot h(x), \qquad \forall \theta \in \Theta,$$

where $k = \dim(\Theta)$, c and a_j's are measurable functions on Θ, whereas T_j's and h are real measurable functions.

Theorem 3 *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. sample from the $\mathbb{P}_\theta$ distribution that belongs to the exponential family, $\theta \in \Theta \subset \mathbb{R}^k$. Then*

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right) \tag{1}$$

is sufficient for θ. (It is an easy consequence of the Neyman–Fisher factorization.)

When $a_j(\theta) = \theta_j$ $(j = 1, \ldots, k)$, then $\theta = (\theta_1, \ldots, \theta_k)$ is called *canonical parameter*, and the above statistic $T(\mathbf{X})$ is called *canonical statistic*.

Theorem 4 (P. Halmos) *If the parameter space $\Theta \subset \mathbb{R}^k$ contains a k-dimensional parallelepiped, then the statistic $T(\mathbf{X})$ of (1) is also complete, so it is minimal sufficient (in exponential family).*

2 Theory and methods of point estimation

Let $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ be parametric statistical space, $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. We want to estimate θ, or its measurable function $\psi(\theta)$ by means of the statistic $T(\mathbf{X})$ on the basis of the i.i.d. sample $\mathbf{X} = (X_1, \ldots, X_n)$. The point estimator is sometimes denoted by $\hat{\theta}$ or $\hat{\psi}$. Criteria for the 'goodness' of a point estimator:

- $T(\mathbf{X})$ is an **unbiased** estimator of $\psi(\theta)$, if $\mathbb{E}_\theta(T(\mathbf{X})) = \psi(\theta), \quad \forall\theta \in \Theta$.
- $T(\mathbf{X}_n)$ is an **asymptotically unbiased** estimator of $\psi(\theta)$, if

$$\lim_{n\to\infty} \mathbb{E}_\theta(T(\mathbf{X}_n)) = \psi(\theta), \quad \forall\theta \in \Theta.$$

- Let T_1 and T_2 be both unbiased estimators of $\psi(\theta)$. T_1 is **at least as efficient** than T_2, if $\mathbb{D}^2_\theta(T_1) \leq \mathbb{D}^2_\theta(T_2)$, $\forall\theta \in \Theta$. An unbiased estimator is **efficient**, if it is at least as efficient than any other unbiased estimator. An efficient estimator is sometimes called *minimum variance unbiased estimator*.

  Efficient estimator does not always exist, but if yes, then it is unique (with probability 1).

- $T(\mathbf{X}_n)$ is a weakly/strongly/mean square **consistent** estimator of $\psi(\theta)$, if $\forall\theta \in \Theta$:
  $T(\mathbf{X}_n) \to \psi(\theta)$ in probability/almost surely/mean square as $n \to \infty$.

Basic statistics:

- *Sample mean*: $\bar{X} = \dfrac{1}{n}\sum_{i=1}^n X_i$. (Sometimes $\bar{X}_n$, $\bar{x}$, $\bar{x}_n$.)
- *Steiner's Theorem*: $\dfrac{1}{n}\sum_{i=1}^n (x_i - c)^2 = \dfrac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2 + (\bar{x} - c)^2$.
- *Empirical variance*: $S^2 = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 = \overline{X^2} - \bar{X}^2$.

- *Corrected empirical variance*: $S^{*2} = \dfrac{n}{n-1}S^2 = \dfrac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.
- *Standard Error of Mean*: $\bar{X}\sqrt{n}/S^*$.
- *k-th empirical moment*: $M_k = \dfrac{1}{n}\sum_{i=1}^n X_i^k$. *Centered* version: $M_k^c = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^k$. ($S^2 = M_2^c = M_2 - M_1^2$.)
- (Location and variability of data) *Skewness*: $M_3^c/(M_2^c)^{3/2}$. *Kurtosis*: $M_4^c/(M_2^c)^2 - 3$.
- *Empirical covariance*: $C = \dfrac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \dfrac{1}{n}\sum_{i=1}^n X_i Y_i - \bar{X}\bar{Y}$.
- *Empirical correlation coefficient*: $R = \frac{C}{S_X S_Y} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^n X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^n Y_i^2 - n\bar{Y}^2\right)}}$.
- *Order statistics*: $X_1^* \le X_2^* \le \cdots \le X_n^*$ (neither independent, nor identically distributed).
- *Sample range*: $X_n^* - X_1^*$.
- *Empirical median*: X_{k+1}^* (if $n = 2k+1$), and $(X_k^* + X_{k+1}^*)/2$ (if $n = 2k$).
- *Proposition* (Steiner in L_1-norm): $\min_{c\in\mathbb{R}} \dfrac{1}{n}\sum_{i=1}^n |x_i - c| = \dfrac{1}{n}\sum_{i=1}^n |x_i - m|$.
- *Empirical c.d.f.*: $F_n^*(x) := \frac{\sum_{i=1}^n I(X_i < x)}{n}$ (stochastic process, x is the time).

Theorem 5 (Glivenko–Cantelli) *:* $\sup_{x\in\mathbb{R}} |F_n^*(x) - F(x)| \to 0$*, almost surely* ($n \to \infty$)*.*

It is sometimes called fundamental theorem of statistics, as it guarantees that the true distribution can be concluded from a large enough sample.

Examples of 'good' estimators:

- the sample mean $\bar{X}$ is always unbiased estimator of the pupulation mean $\mathbb{E}(X_1)$;
- the empirical variance is asymptotically unbiased, whereas, the corrected empirical variance is unbiased estimator of the population variance $\mathrm{Var}\,(X_1)$;

- the above are also consistent in all the three meanings (provided the first/second/fourth population moments exist).

Methods of point estimation:

- **Maximum Likelihood Estimation (MLE)**: given the sample, the MLE of θ is $\hat{\theta}$ if it maximizes the likelihood function. By common sense, in case of a discrete distribution, the MLE is a possible parameter value, for which having the actual sample is the most likely. However, $\hat{\theta} = T(\mathbf{X})$ is a statistic, and it is asymptotically unbiased and strongly consistent estimator of θ.

- **Method of moments**: if $\dim(\theta) = k$, then we find the first k empirical moments of the $\mathbb{P}_{(\theta_1,\dots,\theta_k)}$ distribution. If, vice versa, θ_j can be expressed by the first k moments, then the same function of the empirical moments gives $\hat{\theta}_j$, for $j = 1, \dots, k$.

3 Fisher information, Cramér–Rao inequality, Rao–Blackwell theorem

We want to give a lower bound for the variance of an unbiased estimator if $\dim(\Theta) = 1$.

Definition 5 (Fisher information) *The Fisher information contained in the i.i.d. sample* $\mathbf{X} = (X_1, \dots, X_n) \sim \mathbb{P}_\theta$ *is*

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln L_\theta(\mathbf{X}) \right)^2 \geq 0, \quad \theta \in \Theta.$$

Note that $I_n(\theta) = nI_1(\theta)$ under the regularity conditions below, and then we also have that $I_n(\theta) = \operatorname{Var}_\theta \left(\frac{\partial}{\partial \theta} \ln L_\theta(\mathbf{X}) \right)$ and

$$I_1(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ln p_\theta(X_1) \right), \quad I_1(\theta) = -\mathbb{E}_\theta \left(\frac{\partial^2}{\partial \theta^2} \ln f_\theta(X_1) \right)$$

that gives an interesting relation to the Shannon entropy.

Theorem 6 (Cramér–Rao inequality) *In the above setup, let* $T(\mathbf{X})$ *be unbiased estimator of the differentiable parameter function* $\psi(\theta)$*, and suppose*

that $\operatorname{Var}_\theta(T) < \infty$ $(\forall \theta \in \Theta)$. *Further, the following regularity conditions hold,* $\forall \theta \in \Theta$*:*

$$\frac{\partial}{\partial \theta} \int L_\theta(\mathbf{x})\, d\mathbf{x} = \int \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x} \quad \textit{and} \quad \frac{\partial}{\partial \theta} \int T(\mathbf{x}) L_\theta(\mathbf{x})\, d\mathbf{x} = \int T(\mathbf{x}) \frac{\partial}{\partial \theta} L_\theta(\mathbf{x})\, d\mathbf{x}.$$

Then $\operatorname{Var}_\theta(T) \geq \frac{(\psi'(\theta))^2}{I_n(\theta)} = \frac{(\psi'(\theta))^2}{nI_1(\theta)}$, $\qquad \forall \theta \in \Theta.$

The Cramér–Rao inequality can be extended to biased estimates. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. sample from the $\mathbb{P}_\theta$ distribution, and the *bias* of the statistic $T(\mathbf{X})$ with respect to the differentiable parameter function $\psi(\theta)$ is

$$b_T(\theta) = \mathbb{E}_\theta(T) - \psi(\theta), \qquad \theta \in \Theta, \quad \dim(\Theta) = 1$$

which is also differentiable with respect to θ. Let $\operatorname{Var}\theta^2(T) < +\infty$, $\forall \theta \in \Theta$. Then under the usual regularity conditions,

$$\operatorname{Var}\theta^2(T) \geq \frac{(\psi'(\theta) + b_T'(\theta))^2}{I_n(\theta)}, \qquad \forall \theta \in \Theta.$$

The statement easily follows if we apply the Cramér–Rao inequality for the parameter function $\psi(\theta) + b_T(\theta)$. Observe that T is an unbiased estimator of it.

The Cramér–Rao inequality can as well be extended to **multidimensional parameter spaces**. Under some regularity conditions, it gives unified lower bound for the covariance matrix of every unbiased estimator (for a given parameter function) based merely on a quantity, called Fisher information matrix, which can be calculated from the underlying distribution as a function of the parameter.

Let $(\mathcal{X}, \mathcal{B}, \mathcal{P})$ be dominated, identifiable, parametric statistical space, and $X_1, \ldots, X_n$ be i.i.d. (univariate or multivariate) sample from the $\mathbb{P}_\theta$ distribution, where $\theta \in \Theta \subset \mathbb{R}^k$ is the parameter space. Suppose, we want to estimate the function $\psi(\theta) = (\psi_1(\theta), \ldots, \psi_k(\theta))$ of the parameter, where $\psi : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one function (often it is the identity, when the parameter θ itself is to be estimated).

- Based on the above i.i.d. sample, construct the k-dimensional statistic $\mathbf{T}(X_1, \ldots, X_n)$, briefly $\mathbf{T} = (T_1, \ldots T_k)$ which is an unbiased estimator of $\psi(\theta)$ in the sense that

  $$\mathbb{E}_\theta \mathbf{T} = \psi(\theta), \quad \forall \theta \in \Theta,$$

  where the expectation of the random vector $\mathbf{T}$ is the vector $(\mathbb{E}_\theta T_1, \ldots, \mathbb{E}_\theta T_k)'$ and $'$ denotes the transposition (the vectors are column vectors now).

- Based on the $\mathbb{P}_\theta$ distribution itself, calculate the Fisher information matrix of the 1-element sample:

$$\mathbf{I}_1(\theta) = \mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right) \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right)' = \operatorname{Var}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(X) \right),$$

  where $X \sim \mathbb{P}_\theta$ and f_θ is the p.d.f. of the $\mathbb{P}_\theta$ distribution (if it is absolutely continuous, and use p_θ for the p.m.f. if $\mathbb{P}_\theta$ is discrete). Here under Var of a random vector its covariance matrix is understood, whereas under the derivative with respect to the vector θ of the scalar valued function g the column vector of the derivatives with respect to the components of θ, i.e., the gradient vector is understood, that is $\frac{\partial}{\partial \theta} g = (\frac{\partial}{\partial \theta_1} g, \dots, \frac{\partial}{\partial \theta_k} g)'$. The information matrix of the n-element sample is defined by $\mathbf{I}_n(\theta) = n\mathbf{I}_1(\theta)$, provided the regularity condition

$$\mathbb{E}_\theta \left(\frac{\partial}{\partial \theta} \ln f_\theta(\mathbf{X}) \right) = \mathbf{0}$$

  holds, where $\mathbf{0}$ is the zero vector. In fact, this condition follows from some simpler ones (we can "differentiate through" the $\int$ or $\sum$). In this case, $\mathbf{I}_n(\theta)$ is the $k \times k$ covariance matrix of the k-dimensional random vector $\frac{\partial}{\partial \theta} \ln \prod_{i=1}^n f_\theta(X_i)$.

- The Cramér–Rao inequality states the following relation between 1 and 2. Denote by $\mathbf{S} = (s_{ij})$ the $k \times k$ matrix of entries $s_{ij} = \frac{\partial}{\partial \theta_j} \psi_i(\theta)$. Then for the covariance matrix of any unbiased estimator (for $\psi(\theta)$), the inequality

$$\operatorname{Var}_\theta \mathbf{T} \geq \frac{1}{n} \mathbf{S} \mathbf{I}_1^{-1}(\theta) \mathbf{S}' = \mathbf{S} \mathbf{I}_n^{-1}(\theta) \mathbf{S}'$$

  holds, where the inequality means that the difference of the left and right hand side matrices is positive semidefinite. Here $\mathbf{S}$ also depends on θ, however, if θ itself is estimated, then $\mathbf{S}$ is the identity matrix and does not appear in the formula above.

Note that if a statistic attains the Cramér–Rao bound, then it is surely efficient. However, the bound cannot always be attained for all $\theta \in \Theta$. Even though, an efficient estimator can exist, and it can be concluded by the following theorem.

Theorem 7 (Rao–Blackwell theorem) *Let $\mathbf{X} = (X_1, \dots, \mathbf{X}_n)$ be an i.i.d. sample from the $\mathbb{P}_\theta$ distribution, $\theta \in \Theta \subset \mathbb{R}^k$ ($k > 1$ can be). Let $T(\mathbf{X})$ be a sufficient statistic and $S(\mathbf{X})$ be an unbiased estimator for $\psi(\theta)$. Then one*

can construct an unbiased estimator $U = g(T)$ for $\psi(\theta)$, that is at least as efficient as S. The construction of U (called 'blackwellization'):

$$U := \mathbb{E}_\theta(S|T) = g(T(\mathbf{X})), \quad \forall\theta \in \Theta.$$

Note that if T is also complete (so minimal sufficient), then blackwellizing any unbiased S (for the same parameter function) with it results in the same (unique) U. Consequently, this U will be the efficient estimator for the given parameter function. The message of the Rao–Blackwell theorem is: find the efficient estimator among the functions of the minimal sufficient statistic.

The Rao–Blackwell theorem can also be extended to biased estimators as follows. Let S be a biased estimator of $\psi(\theta)$ and blackwellize it with the sufficient statistic T. Then the so obtained U has the same bias as S and

$$R_\theta(U) \leq R_\theta(S), \quad \forall\theta \in \Theta,$$

where $R_\theta(T) = \mathbb{E}_\theta(T - \psi(\theta))^2$ is the *squared risk* of estimating the parameter function $\psi(\theta)$ with the statistic T. Observe that by the Steiner's equality

$$R_\theta(T) = \operatorname{Var}_\theta^2(T) + b_T^2(\theta).$$

4 Examples

1. Let $X_1, \ldots, X_n$ be i.i.d. sample from Poisson distribution with parameter λ.

   $$L_\lambda(\mathbf{x}) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \left(\lambda^{\sum_{i=1}^n x_i} e^{-n\lambda}\right) \cdot \left(\prod_{i=1}^n \frac{1}{x_i!}\right) = g_\lambda(\sum_{i=1}^n x_i) \cdot h(\mathbf{x}),$$

   so $\sum_{i=1}^n X_i$ is sufficient statistic for λ, akin to its one-to-one function $\bar{X}$.

   To find the MLE,

   $$\ln L_\lambda(\mathbf{x}) = \ln\left[\prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right] = \ln\lambda \sum_{i=1}^n x_i - \sum_{i=1}^n \ln x_i! - \lambda n.$$

   Differentiating with respect to λ, the likelihood equation is

   $$\frac{\partial \ln L_\lambda(\mathbf{x})}{\partial\lambda} = \frac{1}{\lambda}\sum_{i=1}^n x_i - n = 0.$$

The solution is $\hat{\lambda} = \bar{x}$, which indeed gives a local and global maximum. So $T(\mathbf{X}) = \bar{X}$ is the MLE of λ, provided it is not 0, i.e., not all the sample entries are zero at the same time (it can happen with positive, albeit 'small' probability).

This distribution belongs to the exponential family, as

$$p_\lambda = \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} e^{x \ln \lambda} \frac{1}{x!}, \quad x = 0, 1, 2, \ldots,$$

where $\ln \lambda$ is the canonical parameter and $\sum_{i=1}^n X_i$ is the canonical sufficient statistic. It is also complete, so making it unbiased, $\bar{X}$ will be the efficient, unbiased estimator for λ, based on the Rao–Blackwell theorem. It also implies, that starting with any other unbiased estimator, e.g., with $\sum_{i=1}^n a_i X_i$, where $\sum_{i=1}^n a_i = 1$, and 'blackwellizing' it with $\sum_{i=1}^n X_i$, we always get $\bar{X}$.

$I_1(\lambda) = \frac{1}{\lambda}$, $I_n(\lambda) = \frac{n}{\lambda}$, so the information bound for λ is $\frac{1}{I_n(\lambda)} = \frac{\lambda}{n} = \operatorname{Var}(\bar{X})$, and it is attained by $\bar{X}$. So $\bar{X}$ is the efficient estimator of λ, for this reason too.

2. Let $X_1, \ldots, X_n$ be i.i.d. sample from exponential distribution with parameter λ). Then

$$L_\lambda(\mathbf{x}) = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^n x_i},$$

that is $g_\lambda(T(\mathbf{x}))$, and $h(\mathbf{x}) = 1 \cdot I_{(0,\infty)}$. Therefore, $\sum_{i=1}^n X_i$ is sufficient akin to $\bar{X}$ or $\frac{1}{\bar{X}}$. It is also complete, so making it unbiased, $\frac{n-1}{n} \frac{1}{\bar{X}}$ will be the efficient, unbiased estimator for λ, based on the Rao–Blackwell theorem. However, it can be shown that it does not reach the information bound. Consequently, no unbiased estimator reaches the bound for λ. What is this bound?

$$\mathbb{E}_\lambda \left(\frac{\partial}{\partial \lambda} (\ln f_\lambda(X) \right) = \mathbb{E}_\lambda \left(\frac{1}{\lambda} - X) \right) = 0.$$

therefore

$$I_1(\lambda) = \operatorname{Var}_\lambda \left(\frac{1}{\lambda} - X) \right) = \operatorname{Var}_\lambda(X) = \frac{1}{\lambda^2}.$$

So

$$I_n(\lambda) = \frac{n}{\lambda^2}.$$

If $\psi(\lambda) = \lambda$, then the information bound is $\frac{1}{I_n(\lambda)} = \frac{\lambda^2}{n}$, whereas

$$\operatorname{Var}_\lambda \left(\frac{n-1}{n} \frac{1}{\bar{X}} \right) = \frac{\lambda^2}{n-2}$$

which is larger than the bound (though, asymptotically approaches it as $n \to \infty$.

However, if we consider the parameter function $\psi(\lambda) = \frac{1}{\lambda}$, this is the expectation of the underlying distribution. For it, $\bar{X}$ is an unbiased estimator and reaches the bound, which is

$$\frac{\psi'(\lambda)^2}{I_n(\lambda)} = \frac{[-\frac{1}{\lambda^2}]^2}{\frac{n}{\lambda^2}} = \frac{1}{n\lambda^2}.$$

Indeed, $\operatorname{Var}(\bar{X}) = \frac{1/\lambda^2}{n}$ that is the same as the above bound. So $\bar{X}$ is efficient estimator of $\frac{1}{\lambda}$.

As for the MLE of λ,

$$\ln L_\lambda(\mathbf{x}) = \ln \left[\prod_{i=1}^n \lambda e^{-\lambda x_i} \right] = n \ln \lambda - \lambda \sum_{i=1}^n x_i,$$

from which, after differentiating, we get that $\hat{\lambda} = 1/\bar{x}$, that gives a local and global maximum. Consequently, $T(\mathbf{X}) = 1/\bar{X}$ is the MLE of λ with probability 1 ($\bar{X}$ can be 0 only with probability 0).

3. Let $X_1, \ldots, X_n$ be i.i.d. sample from normal (Gaussian) distribution with unknown parameter $\theta = (\mu, \sigma^2)$. Then

$$\begin{aligned} L_\theta(\mathbf{x}) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right) = \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{1}{2\sigma^2} \left[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \right] \right). \end{aligned}$$

It is $g_\theta(T(\mathbf{x}))$, where $T(\mathbf{X}) = (\bar{X}, S^2)$ sufficient for θ, and $h(\mathbf{x}) = 1$. Obviously, $(\bar{X}, S^{*2})$ or $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ are also sufficient. It also belongs to the exponential family.

To find MLE,

$$\begin{aligned} \ln L_\theta(\mathbf{x}) &= \ln \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \sum_{i=1}^n \left[-\ln(\sqrt{2\pi\sigma^2}) - \frac{(x_i-\mu)^2}{2\sigma^2} \right] = \\ &= -\frac{n}{2}(\ln(2\pi) + \ln \sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \end{aligned}$$

Taking partial derivatives,

$$\frac{\partial \ln L_\theta(\mathbf{x})}{\partial \mu} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) = 0 \Longrightarrow \hat{\mu} = \bar{x}.$$

and

$$\frac{\partial \ln L_\theta(\mathbf{x})}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0.$$

Since the solution $\hat{\mu} = \bar{x}$ does not depend on the actual value of σ^2 substituting it to the second equation, we get that $\hat{\sigma^2} = S_n^2$, that is only asymptotically unbiased for σ^2. The Hessian at $(\bar{x}, s_n^2)$ is:

$$H = \begin{pmatrix} -\frac{n}{s_n^2} & 0 \\ 0 & -\frac{n}{2(s_n^2)^2} \end{pmatrix},$$

which is negative definite, so we indeed have a local and global maximum here.

If both parameters are unknown that the information bound cannot be attained for θ. Howewer, if the variance is known (σ_0^2), then the bound can be attained for μ. Indeed,

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu)^2}{2\sigma_0^2}}.$$

Therefore,

$$I_1(\mu) = \mathrm{Var}_\mu^2\left(\frac{\partial}{\partial \mu}\ln f_\mu(X)\right) = \mathrm{Var}_\mu^2\left(\frac{2(X-\mu)}{2\sigma_0^2}\right) = \frac{1}{\sigma_0^4}\mathrm{Var}_\mu^2(X-\mu) = \frac{1}{\sigma_0^4}\sigma_0^2 = \frac{1}{\sigma_0^2} \neq 0.$$

So

$$I_n(\mu) = n\frac{1}{\sigma_0^2},$$

and it is $\frac{1}{/}\mathrm{Var}_\mu^2(\bar{X})$.

4. Let $X_1, \ldots, X_n$ be i.i. sample from continuous uniform distribution on $[a, b]$. Here $\theta = (a, b)$.

$$L_\theta(\mathbf{x}) = \prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{(b-a)^n}, \quad \text{if} \quad x_1, \ldots, x_n \in [a, b],$$

and 0, otherwise. $L_\theta(\mathbf{x}) = (b-a)^{-n} I(x_1^* \geq a,\, x_n^* \leq b) = g_\theta(x_1^*, x_n^*)$ and $h(\mathbf{x}) = 1$. So the pair (X_1^*, X_n^*) is sufficient for (a, b). It also gives the

MLE, as we maximize the likelihood on the constraint that $[a, b]$ should contain all the sample entries.

This distribution does not belong to the exponential family, as its support depends on the parameters. Therefore, the moment estimate of the parameters is not the same as the MLE, in contrast to the first three exampes that belonged to the exponential family.