Anna Janicka

Mathematical Statistics 2020/2021
Lecture 5

1. Fisher Information, the Information Inequality and Estimator Efficiency

In order to be able to solve the problem of finding the MVUE estimators introduced in the last lecture, we will need to refer to the following definition:

Definition 1. *If a statistical model with observations $X_1, X_2, \ldots, X_n$ and probability f_θ fulfills the following regularity conditions:*

(1) *Θ is an open 1-dimensional set;*
(2) *The support of the distribution $\{x : f_\theta(x) > 0\}$ does not depend on θ;*
(3) *The derivative $\frac{df_\theta}{d\theta}$ exists;*

we can define **Fisher information** *(Information) for sample $X_1, X_2, \ldots, X_n$:*

$$I_n(\theta) = \mathbb{E}_\theta \left(\frac{d \ln f_\theta(X_1, X_2, \ldots, X_n)}{d\theta} \right)^2.$$

Note that in the above definition, f_θ may mean both a density function and a probability mass function, and that we do not assume independence of observations. For the special case when $X_1, X_2, \ldots, X_n$ are IID, we can write

$$I_n(\theta) = nI_1(\theta),$$

where $I_1(\theta)$ is the information connected with one observation.

In most cases, calculating Fisher Information from the definition may be computationally complicated (the formula in the expected value is compound). In such cases, one can use an alternative formula for I_n, which works in case of twice differentiable functions:

$$I_n(\theta) = -\mathbb{E}_\theta \left(\frac{d^2 \ln f_\theta(X_1, X_2, \ldots, X_n)}{d\theta^2} \right).$$

The Fisher Information describes the amount of knowledge about the distribution (the value of distribution parameters) that may be derived from a sample of size n. We can see that the larger the absolute value of the second derivative of the log of the probability function (i.e., the more steep the probability function), the larger the Fisher Information. Therefore, if the density around θ is flat, then information from a single observation or a small sample will not allow us to differentiate among possible values of θ. If the density around θ is steep, the sample contributes a lot of knowledge leading to θ identification.

Examples of calculations:

(1) For the Poisson distribution $Poiss(\theta)$, we have $f_\theta(x) = \frac{\theta^x}{x!}e^{-\theta}$, and the logarithm $\ln f_\theta(x) = -\theta + x \ln \theta - \ln(x!)$, so that

$$I_1(\theta) = \mathbb{E}_\theta \left(\frac{d \ln f_\theta(x)}{d\theta} \right)^2 = \sum_{x=0}^{\infty} \left(\frac{x}{\theta} - 1 \right)^2 \frac{\theta^x}{x!} e^{-\theta} = \sum_{x=0}^{\infty} \frac{1}{\theta^2} (x-\theta)^2 \frac{\theta^x}{x!} e^{-\theta} = \frac{1}{\theta^2} \mathrm{Var}_\theta(X) = \frac{1}{\theta},$$

or, alternatively,

$$I_1(\theta) = -\mathbb{E}_\theta \left(\frac{d^2 \ln f_\theta(x)}{d\theta^2} \right) = -\sum_{x=0}^{\infty} \left(-\frac{x}{\theta^2} \right) \frac{\theta^x}{x!} e^{-x} = \sum_{x=1}^{\infty} \frac{\theta^{x-2}}{(x-1)!} e^{-\theta} = \sum_{x=0}^{\infty} \frac{\theta^{x-1}}{x!} e^{-\theta} = \frac{1}{\theta} \sum_{x=0}^{\infty} \frac{\theta^x}{x!} e^{-\theta} = \frac{1}{\theta}.$$

(2) For an exponential distribution $Exp(\theta)$, we have $f_\theta(x) = \theta e^{-\theta x}$ for $x > 0$, and $\ln f_\theta(x) = \ln \theta - \theta x$, so that

$$I_1(\theta) = \mathbb{E}_\theta \left(\frac{d \ln f_\theta(x)}{d\theta} \right)^2 = \int_0^\infty \left(\frac{1}{\theta} - x \right)^2 \theta e^{-\theta x} dx = \mathrm{Var}_\theta(X) = \frac{1}{\theta^2},$$

or, alternatively,

$$I_1(\theta) = -\mathbb{E}_\theta\left(\frac{d^2 \ln f_\theta(x)}{d\theta^2}\right) = -\int_0^\infty \left(-\frac{1}{\theta^2}\right)\theta e^{-\theta x} dx = \frac{1}{\theta^2}.$$

(3) A uniform distribution over the interval $(0, \theta)$ does not fulfill the conditions which allow calculating Fisher Information (the support of the distribution depends on θ!). Therefore, although one can perform the calculations figuring in the definition of Fisher Information (calculate the expected value of....), the result will not have the usual meaning connected with the formula. Therefore, performing the calculations is pointless.

We have said that Fisher Information describes the amount of knowledge conveyed by a sample of size n. One may prove a strong result: this characteristic leads to the identification of the minimum variance of an unbiased estimator for a given distribution function, in the words of the

Theorem 1. The Cramér-Rao Information Inequality *Let $X = (X_1, X_2, \ldots, X_n)$ be observations from a joint distribution with density $f_\theta(x)$, where $\theta \in \Theta \subseteq \mathbb{R}$. If:*

- *$T(X)$ is a statistic with a finite expected value, and $E_\theta T(X) = g(\theta)$;*
- *Fisher information is well defined, $I_n(\theta) \in (0, \infty)$;*
- *All f_θ have the same support;*
- *The order of differentiating $d/d\theta$ and and integrating $\int \ldots dx$ may be reversed.*

Then, for any θ:

$$\mathrm{Var}_\theta T(X) \geq \frac{(g'(\theta))^2}{I_n(\theta)}.$$

As a special case, for $g(\theta) = \theta$, we get that for any unbiased estimator $\hat{\theta}(X)$ of θ, we have

$$\mathrm{Var}_\theta \hat{\theta}(X) \geq \frac{1}{I_n(\theta)}.$$

The implications of the above theorem are sound: the MSE of an unbiased estimator (i.e., the variance of this estimator) cannot be lower than a given function of n, $I_n(\theta)$, which depends on the distribution. Therefore, if the variance of an estimator is equal to the lower bound of the information inequality, then this estimator is MVUE.

Examples:

(1) In the Poisson model, we have that $\bar{X}$ is the MVUE of θ. We know that $I_n(\theta) = nI_1(\theta) = \frac{n}{\theta}$, and at the same time we have

$$\frac{1}{I_n(\theta)} = \frac{\theta}{n} = \mathrm{Var}_\theta(\bar{X}).$$

(2) In the exponential model, we have that $\bar{X}$ is the MVUE of $\frac{1}{\theta}$. We know that $I_n(\theta) = nI_1(\theta) = \frac{n}{\theta^2}$, and at the same time we have $\mathrm{Var}_\theta(\bar{X}) = \frac{1}{n}\mathrm{Var}X = \frac{1}{n\theta^2}$. We are estimating a function of θ: $g(\theta) = \frac{1}{\theta}$, for which $g'(\theta) = -\frac{1}{\theta^2}$, so that we have

$$\frac{(-1/\theta^2)^2}{I_n(\theta)} = \frac{1/\theta^4}{1/(n \cdot \theta^2)} = \frac{1}{n\theta^2} = \mathrm{Var}_\theta(\bar{X}).$$

Unfortunately, the lower bound from the Information Inequality is not always attained (depending on the distribution). This signifies that if an estimator has a variance exceeding the lower bound of the inequality, it is not yet proof that this estimator is not MVUE. We may encounter such a situation when dealing with the estimator of parameter θ in the exponential model. Based on the properties of the Gamma distribution, one can show that $\frac{1}{\bar{X}}$ (the Method of Moments and ML Estimator) is a biased estimator of θ, with $\mathbb{E}_\theta\left(\frac{1}{\bar{X}}\right) = \frac{n}{n-1}\theta$. On this basis, we can construct an unbiased estimator of θ as $\frac{n-1}{n\bar{X}}$. It can be shown that this latter estimator is MVUE, although its variance is higher than the bound in the Cramér-Rao Inequality. This is because in the case of the exponential distribution, the bound is never attained.

Based on the definition of the Fisher Information and the Cramér-Rao Inequality, we can describe the extent to which an estimator uses the knowledge conveyed by the data sample, by introducing the following concept of *efficiency*.

Definition 2. *The* **efficiency** *of an unbiased estimator $\hat{g}(x)$ of $g(\theta)$ is*

$$\mathrm{ef}(\hat{g}) = \frac{(g'(\theta))^2}{\mathrm{Var}_\theta(\hat{g}) \cdot I_n(\theta)}.$$

The relative efficiency of two unbiased estimators, $\hat{g}_1$ and $\hat{g}_2$ is

$$\mathrm{ef}(\hat{g}_1, \hat{g}_2) = \frac{\mathrm{Var}_\theta(\hat{g}_2)}{\mathrm{Var}_\theta(\hat{g}_1)} = \frac{\mathrm{ef}(\hat{g}_1)}{\mathrm{ef}(\hat{g}_2)}.$$

The definition of the efficiency of an estimator is based on the constatation that the smaller the variance of an (unbiased) estimator, the better it uses data. If an estimator has a variance as small as the Information Inequality bound, this means that it makes optimal use of the data. If the assumptions of the Cramér-Rao Inequality are fulfilled (this depends on the distribution, not the estimator), then it follows that the efficiency of any unbiased estimator does not exceed 1. If the efficiency is equal to 1 (the estimator is **efficient**), then this means that the estimator is also MVUE. However, if the efficiency of an estimator is strictly less than 1, then this does not yet necessarily mean that the estimator is not MVUE (it may or may not be MVUE, depending on the distribution). Note that when calculated for distributions which do not fulfill the Information Inequality assumptions (for example, a uniform distribution over the interval $(0, \theta)$), the "efficiency" may be found to be greater than 1, but this result is meaningless (the formulas do not have the meaning they were defined to have).

Examples:

(1) In the Poisson model, the $\bar{X}$ estimator of θ is efficient.
(2) In the exponential model, the $\bar{X}$ estimator of $\frac{1}{\theta}$ is efficient.
(3) In the exponential model, the $\frac{1}{\bar{X}}$ estimator of θ is biased. However, as we have said above this bias may easily be eliminated by multiplying by a constant: $\frac{n-1}{n\bar{X}}$. This MVUE estimator is not efficient.

2. Asymptotic properties of estimators

Until now, we have not discussed the effect of sample size on the properties of estimators – i.e., apart from noting that the bias of the biased estimator of the variance ($\hat{S}^2$) tends to 0 when n tends to infinity, we have just performed analyses based on a fixed sample size. If an estimator has the desirable characteristics regardless of sample size – perfect. However, in many cases, estimators do not behave as well for small samples as we would like them to, i.e. they are not unbiased or not efficient. In this case, the question arises: what would happen, if instead of a small sample, we had a large sample at our disposal? This leads to the discussion of the so-called asymptotic properties of estimators. These properties are in most cases governed by different versions of limit theorems; the practical use is assessing the (approximate) properties of estimators for large samples, although usually it is very hard to say what sample is "large enough" for the approximations to be good.

In what follows, instead of considering estimators, we will be in fact considering sequences of estimators, based on larger and larger samples. I.e., if were to study the asymptotic properties of the empirical mean estimator, we would be in fact dealing with a sequence of estimators based on increasing samples: X_1, $\frac{X_1+X_2}{2}$, $\frac{X_1+X_2+X_3}{3}$, $\ldots$, $\frac{X_1+X_2+\ldots+X_n}{n}$, $\ldots$. For simplicity, however, in most cases we will omit the sample size annotation (and use the notation $\bar{X}$, etc.).

2.1. Asymptotic unbiasedness.

Definition 3. *We will say that an estimator $\hat{g}(X)$ of the value $g(\theta)$ is* **asymptotically unbiased**, *if $b(\theta) \xrightarrow[n\to\infty]{} 0$.*

Any unbiased estimator is also obviously asymptotically unbiased. The biased estimator of the variance is asymptotically unbiased (i.e., for large samples, it behaves approximately just as well as the unbiased estimator).

2.2. **Consistency.**

Definition 4. *Let $X_1, X_2, \ldots$ be an IID sample. Let $\hat{g}$ be a sequence of estimators of the value $g(\theta)$. $\hat{g}$ is* **consistent***, if for all $\theta \in \Theta$, for any $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}_\theta(|\hat{g}(X_1, X_2, \ldots, X_n) - g(\theta)| \leq \varepsilon) = 1$$

(i.e. $\hat{g}$ converges to $g(\theta)$ in probability).
$\hat{g}$ is **strongly consistent***, if for all $\theta \in \Theta$, we have*

$$\mathbb{P}_\theta\left(\lim_{n\to\infty} g(X_1, X_2, \ldots, X_n) = g(\theta)\right) = 1$$

(i.e. $\hat{g}$ converges to $g(\theta)$ almost surely).

Note that from the Glivenko-Cantelli theorem it follows that empirical cumulative distribution functions connected with samples increasing in size converge almost surely to the theoretical CDF, which means that the empirical distribution reflects the theoretical distribution for large samples. Therefore, we should expect (strong) consistency from all sensible estimators – if an estimator is not consistent, then this means it does not fulfill this minimal requirement and therefore should not be used.

Verification of consistency is usually not too hard in practice. First of all, in many cases it can be derived on the base of the Laws of Large Numbers. Second, it can be checked from the definition – for example, with the use of a version of the Chebyshev inequality[1]:

$$\mathbb{P}(|\hat{g}(X) - g(\theta)| \geq \varepsilon) \leq \frac{\mathbb{E}(\hat{g}(X) - g(\theta))^2}{\varepsilon^2}.$$

Given that the MSE of an estimator is

$$MSE(\theta, \hat{g}) = \mathbb{E}_\theta(\hat{g}(X) - g(\theta))^2 = \mathrm{Var}_\theta \hat{g} + b^2(\theta),$$

we get a *sufficient condition* for consistency:

$$\lim_{n\to\infty} MSE(\theta, \hat{g}) = 0.$$

In other words, if we show that the MSE of an estimator tends to 0 as sample size increases, this means that the estimator is consistent.[2]

Examples:

(1) For any family of distributions with an expected value: the sample mean $\bar{X}$ is a consistent estimator of the expected value $\mu(\theta) = \mathbb{E}_\theta(X_1)$. Convergence (strong) may easily be derived from the Laws of Large Numbers (strong).
(2) For distributions having a variance: $\hat{S}^2$ and S^2 are consistent estimators of the variance $\sigma^2(\theta) = \mathrm{Var}_\theta(X_1)$. Convergence (strong) also stems from the Laws of Large Numbers applied to the sum of squares of the random variables in the sample.

Note that consistency is not equivalent to unbiasedness. An estimator may be consistent but biased (for example, the biased estimator of the variance), as well as unbiased but not consistent (e.g. an estimator of the mean which uses just the first observation in the sample, $T_n(X_1, X_2, \ldots, X_n) = X_1$ as an estimator of $\mu(\theta) = \mathbb{E}_\theta(X_1)$).

[1]The formula is derived from the basic Chebyshev inequality in the same way as the Chebyshev-Bienaymé inequality (the latter states that $\mathbb{P}(|X - \mathbb{E}X| \geq \varepsilon) \leq \frac{\mathbb{E}(X-\mathbb{E}X)^2}{\varepsilon^2} = \frac{\mathrm{Var}X}{\varepsilon^2}$).

[2]Note that an estimator may be consistent even if the MSE does not tend to 0, as this is not a necessary condition of consistency.