Mathematical Statistics

Anna Janicka

Lecture VI, 26.03.2018

PROPERTIES OF ESTIMATORS, PART II

Plan for Today

1. Fisher information
2. Information inequality
3. Estimator efficiency
4. Asymptotic estimator properties
   - consistency
   - *asymptotic normality*
   - *asymptotic efficiency*

Fisher information

If a statistical model with obs. $X_1, X_2, \ldots, X_n$ and probability f_θ fulfills regularity conditions, i.e.:

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 $X_1, X_2, \ldots, X_n$:

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

we do not assume independence of $X_1, X_2, \ldots, X_n$

Fisher information – what does it mean?

- It is a measure of how much a sample of size *n* can tell us about the value of the unknown parameter θ (on average).
- If the density around θ is flat, then information from a single observation or a small sample will not allow to differentiate among possible values of θ. If the density around θ is steep, the sample contributes a lot of info leading to θ identification.

Fisher Information – cont.

Some formulae:

- if the distribution is continuous

$$I_n(\theta) = \int_X \left(\frac{\frac{df_\theta(x)}{d\theta}}{f_\theta(x)} \right)^2 f_\theta(x) dx$$

- if the distribution is discrete

$$I_n(\theta) = \sum_{x \in X} \left(\frac{\frac{dP_\theta(x)}{d\theta}}{P_\theta(x)} \right)^2 P_\theta(x)$$

- if f_θ is twice differentiable

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

Fisher information – cont. (2)

- If the sample consists of independent random variables from the same distribution, then

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

where $I_1(\theta)$ is Fisher information for a single observation

Fisher Information – examples

- Exponential distribution exp(λ)

$$I_1(\lambda) = \ldots = \frac{1}{\lambda^2}$$

- Poisson distribution Poiss(θ)

$$I_1(\theta) = \ldots = \frac{1}{\theta}$$

Information Inequality (Cramér-Rao)

Let $X=(X_1, X_2, \ldots, X_n)$ be observations from a joint distribution with density $f_\theta(x)$, where $\theta \in \Theta \subseteq \mathrm{R}$. 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 densities f_θ have the same support
- The order of differentiating ($d/d\theta$) and integrating $\int \ldots dx$ may be reversed.

Then, for any θ:

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

Information inequality – implications

- The MSE of an unbiased estimator (= the variance) cannot be lower than a given function of n and θ.
- If the MSE of an estimator is equal to the lower bound of the information inequality, then the estimator is MVUE.
- If $\hat{\theta}(X)$ is an unbiased estimator of θ, then

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

Information inequality – examples

- In the Poisson model, $\hat{\theta} = \overline{X}$ is MVUE(θ)

$$Var_\theta(\overline{X}) = \frac{\theta}{n}$$

- In the exponential model, $\overline{X}$ is MVUE($1/\lambda$)

$$Var_\lambda(\overline{X}) = \frac{1}{n\lambda^2}$$

- The Cramér-Rao inequality is not always optimal. In the exponential model, $\hat{\lambda} = 1/\overline{X}$ is a biased estimator of λ.

$$\tilde{\lambda} = \frac{n-1}{n\overline{X}}$$

is an unbiased estimator, which is also MVUE(λ), although its variance is *higher* than the bound in the Cramér-Rao inequality.

Efficiency

The efficiency of an unbiased estimator $\hat{g}(X)$ of $g(\theta)$ is:

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

Relative efficiency of unbiased estimators $\hat{g}_1(X)$ and $\hat{g}_2(X)$:

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

Efficiency and the information inequality

- If the information inequality holds, then for any unbiased estimator

$$\mathrm{ef}(\hat{g}) \leq 1$$

- If $\hat{g}$ = MVUE(g), then it is possible that $\mathrm{ef}(\hat{g}) = 1$, but it is also possible that

$$\mathrm{ef}(\hat{g}) < 1$$

- If $\mathrm{ef}(\hat{g}) = 1$, then the estimator is **efficient**.

Efficiency – examples

- In the Poisson model, $\hat{\theta} = \overline{X}$ is efficient.
- In the exponential model, $\overline{X}$ is an efficient estimator of $1/\lambda$.
- In the exponential model, $\hat{\lambda} = \frac{n-1}{n\overline{X}}$

is not an efficient estimator of λ, although it is MVUE(λ).

- In a uniform model U$(0, \theta)$, for the MLE(θ) we get ef >1 (that is because the assumptions of the information inequality are not fulfilled)

Asymptotic poperties of estimators

- Limit theorems describing estimator properties when $n \to \infty$
- In practice: information on how the estimators behave for large samples, *approximately*
- Problem: usually, there is no answer to the question what sample is large enough (for the approximation to be valid)

Consistency

Let $X_1, X_2, \ldots, X_n, \ldots$ be an IID sample (of independent random variables from the same distribution) . Let $\hat{g}(X_1, X_2, \ldots, X_n)$ be a sequence of estimators of the value $g(\theta)$.

$\hat{g}$ is a **consistent** estimator, if for all $\theta \in \Theta$, for any $\varepsilon > 0$:

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

(i.e. $\hat{g}$ converges to $g(\theta)$ in probability)

Strong consistency

Let $X_1, X_2, \ldots, X_n, \ldots$ be an IID sample (of independent random variables from the same distribution). Let $\hat{g}(X_1, X_2, \ldots, X_n)$ be a sequence of estimators of the value $g(\theta)$.

$\hat{g}$ is **strong consistent**, if for any $\theta \in \Theta$:

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

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

Consistency – note

From the Glivenko-Cantelli theorem it follows that empirical CDFs converge almost surely to the theoretical CDF. Therefore, we should expect (strong) consistency from all sensible estimators.

Consistency = minimal requirement for a sensible estimator.

Consistency – how to verify?

- From the definition: for example with the use of a version of the Chebyshev inequality:

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

Given that the MSE of an estimator is

$$MSE(\theta, \hat{g}) = E_\theta(\hat{g}(X) - g(\theta))^2$$

we get a sufficient condition for consistency:

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

- From the LLN

Consistency – examples

- For any family of distributions with an expected value: the sample mean $\overline{X}_n$ is a consistent estimator of the expected value $\mu(\theta)=E_\theta(X_1)$. Convergence from the SLLN.
- For distributions having a variance:
  $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ and $\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$
  are consistent estimators of the variance $\sigma^2(\theta)=\mathrm{Var}_\theta(X_1)$. Convergence from the SLLN.

Consistency – examples/properties

- An estimator may be unbiased but unconsistent; eg. $T_n(X_1, X_2, ..., X_n)=X_1$ as an estimator of $\mu(\theta)=E_\theta(X_1)$.

- An estimator may be biased but consistent; eg. the biased estimator of the variance or any unbiased consistent estimator + 1/n.


Warsaw University
Faculty of Economic Sciences

Asymptotic normality

$\hat{g}(X_1, X_2, \ldots, X_n)$ is an **asymptotically normal** estimator of $g(\theta)$, if for any $\theta \in \Theta$ there exists $\sigma^2(\theta)$ such that, when $n \to \infty$

$$\sqrt{n}\left(\hat{g}(X_1, X_2, \ldots, X_n) - g(\theta)\right) \xrightarrow{D} N(0, \sigma^2(\theta))$$

We are looking at convergence in distribution, i.e. for any a

$$\lim_{n\to\infty} P_\theta\left(\frac{\sqrt{n}}{\sigma(\theta)}\left(\hat{g}(X_1, X_2, \ldots, X_n) - g(\theta)\right) \le a\right) = \Phi(a)$$

in other words, the distribution of $\hat{g}(X_1, X_2, \ldots, X_n)$ is for large n similar to $N(g(\theta), \frac{\sigma^2}{n})$

Asymptotic normality – properties

- An asymptotically normal estimator is consistent (not necessarily strongly).
- Has a *similar* condition to unbiasedness – the expected value of the asymptotic distribution equals $g(\theta)$ (but the estimator *does not need to* be unbiased).
- **Asymptotic variance** defined as $\sigma^2(\theta)$ or $\sigma^2(\theta)/n$ – the variance of the asymptotic distribution

Asymptotic normality – what it is not

- For an asymptotically normal estimator we usually have:

$$E_\theta \hat{g}(X_1, X_2, \ldots, X_n) \xrightarrow{n\to\infty} g(\theta)$$

$$n \operatorname{var} \hat{g}(X_1, X_2, \ldots, X_n) \xrightarrow{n\to\infty} \sigma^2(\theta)$$

but these properties needn't hold, because convergence in distribution does not imply convergence of moments.

Asymptotic normality – example

- Let $X_1, X_2, \ldots, X_n, \ldots$ be an IID sample from a distribution with mean μ and variance σ^2. On the base of the CLT, for the sample mean we have

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$$

In this case the asymptotic variance, σ^2/n, is equal to the estimator variance.

Asymptotic normality – how to prove it

In many cases, the following is useful:

Delta Method. Let T_n be a sequence of random variables such that for $n \to \infty$ we have

$$\sqrt{n}(T_n - \mu) \xrightarrow{D} N(0, \sigma^2)$$

and let $h: \mathrm{R} \to \mathrm{R}$ be a function differentiable at point μ such that $h'(\mu) \neq 0$. Then

$$\sqrt{n}\left(h(T_n) - h(\mu)\right) \xrightarrow{D} N(0, \sigma^2 (h'(\mu))^2)$$

μ, σ^2 are functions of θ

usually used when estimators are functions of statistics T_n, which can be easily shown co converge on the base of CLT

Asymptotic normality – examples cont.

In an exponential model: $MLE(\lambda) = \frac{1}{\overline{X}}$

From CLT, we get

$$\sqrt{n}(\overline{X} - \frac{1}{\lambda}) \xrightarrow{D} N(0, \frac{1}{\lambda^2})$$

so from the Delta Method for $h(t)=1/t$:

$$\sqrt{n}(\frac{1}{\overline{X}} - \lambda) \xrightarrow{D} N(0, \frac{1}{\lambda^2} \cdot (-\frac{1}{(1/\lambda)^2})^2)$$

so $\frac{1}{\overline{X}}$ is an asymptotically normal (and consistent) estimator of λ.

Asymptotic efficiency

For an asymptotically normal estimator $\hat{g}(X_1, X_2, ..., X_n)$ of $g(\theta)$ we define **asymptotic efficiency** as

$$\text{as.ef}(\hat{g}) = \frac{\left(g'(\theta)\right)^2 n}{\sigma^2(\theta) \cdot I_n(\theta)},$$

where $\sigma^2(\theta)/n$ is the asymptotic variance, i.e. for $n \to \infty$

$$\sqrt{n}\left(\hat{g}(X_1, X_2, ..., X_n) - g(\theta)\right) \xrightarrow{D} N(0, \sigma^2(\theta))$$

modification of the definition of efficiency to the limit case, with the asymptotic variance in place of the normal variance

$$\text{as.ef}(\hat{g}) = \frac{\left(g'(\theta)\right)^2}{\sigma^2(\theta) \cdot I_1(\theta)}$$

Relative asymptotic efficiency

Relative asymptotic efficiency for asymptotically normal estimators

$\hat{g}_1(X)$ and $\hat{g}_2(X)$

$$\text{as.ef}(\hat{g}_1, \hat{g}_2) = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)} = \frac{\text{as.ef}(\hat{g}_1)}{\text{as.ef}(\hat{g}_2)}$$

Note. A less (asymptotically) efficient estimator may have other properties, which will make it preferable to a more efficient one.

Relative asymptotic efficiency – examples. Is the mean better than the median?

Depends on the distribution!

a) normal model N(μ, σ^2):

$$\sqrt{n}\left(\overline{X}-\mu\right)\xrightarrow{D}N(0,\sigma^2)$$

$$\sqrt{n}\left(\text{mêd}-\mu\right)\xrightarrow{D}N(0,\tfrac{\pi\sigma^2}{2})$$

$$\text{as.ef}(\text{mêd},\overline{X})=\tfrac{2}{\pi}<1$$

b) Laplace model Lapl(μ, λ)

$$\sqrt{n}\left(\overline{X}-\mu\right)\xrightarrow{D}N(0,\tfrac{2}{\lambda^2})$$

$$\sqrt{n}\left(\text{mêd}-\mu\right)\xrightarrow{D}N(0,\tfrac{1}{\lambda^2})$$

$$\text{as.ef}(\text{mêd},\overline{X})=2>1$$

c) some distributions do not have a mean...

Theorem: For a sample from a continuous distribution with density *f*(*x*), the sample median is an asymptotically normal estimator for the median *m* (provided the density is continuous and ≠0 at point *m*):

$$\sqrt{n}\left(\text{mêd}-m\right)\xrightarrow{D}N(0,\tfrac{1}{4(f(m))^2})$$