P

Statistics Applied to Economics

Degree in Economics

F.Tusell

Dpto. Economía Aplicada III (Estadística y Econometría)

Curso 2011–2012

Universidad del País Vasco / Euskal Herriko Unibertsitatea

Índice I

Consistency
Efficiency
Suficiency

Consistency (I) (reminder: probability limits)

- We say that the limit in probability of a sequence or random variables $\{Z_n\}$ is Z if for any $\epsilon > 0$, $\eta > 0$ there is N such that for $n > N$:

$$P(|Z_n - Z| < \epsilon) \geq 1 - \eta$$

- In plain English: if taking sufficiently advanced terms of $\{Z_n\}$ we can be within ϵ of Z with probability as close to 1 as we wish.
- Compare with usual notion of limit in mathematical analysis.
- Usual notation is $Z_n \xrightarrow{p} Z$ or $\text{plim}(Z_n) = Z$.

Consistency (II)

- $\hat{\theta}_n$ denotes an estimator of θ based on a sample of size n. For instance, we might have

$$\hat{\theta}_n = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- $\hat{\theta}_n$ is *consistent* if $\hat{\theta}_n \xrightarrow{P} \theta$
- In plain English: if by increasing the sample size we can obtain arbitrary precision with as close to 1 confidence as we choose.
- In general, consistency is the very least we ask for. (We want to be rewarded for our effort in sampling!)

Consistency (III)

- We can usually show consistency by using; i) The laws of large numbers, or ii) Tchebycheb inequality, among other ways.
- Consistency does not imply unbiasedness.

How can we have consistency and not unbiasedness?

Think of $\hat{\theta}_n$ taking the true value θ with probability $1 - \frac{1}{n}$ and the value n with probability $\frac{1}{n}$.

Consistency vía Tchebychev inequality

Example: consistency of $\hat{\lambda} = \overline{X}$ as estimator of λ of a $\mathcal{P}(\lambda)$.

- We know $E[\hat{\lambda}] = \lambda$ and $\mathrm{Var}(\hat{\lambda}) = \lambda/n$.
- Then (Tchebycheff),

$$P(|\hat{\lambda} - \lambda| < \underbrace{k\sqrt{\lambda/n}}_{\epsilon}) \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Make your pick of $1 - \eta$ as close to 1 as desired; whatever the implied k, we only have to choose n large enough to make ϵ as small as we wish.

Unbiasedness + variance$\rightarrow 0 \implies$ consistency

- Again, simple application of Tchebychev's inequality.
- Unbiasedness implies $E(\hat{\theta}_n) = \theta$.

$$P(|\hat{\theta}_n - \theta| < \underbrace{k\sigma_n}_{\epsilon} \geq \underbrace{1 - 1/k^2}_{1-\eta}$$

- Let $1 - \eta$ be as close to 1 as desired; whatever the implied k, ϵ can be made small for large n, as $\sigma_n \rightarrow 0$.
- If both variance and bias decrease to zero, we also have consistency.

Consistency of moment estimators

- Moment estimators are usually consistent.
- Sketch of argument for a particular case:

$$m = \alpha_1(\hat{\theta}) = \overline{X}$$

- **If** $\alpha_1(\hat{\theta})$ has a continuous inverse function, $\hat{\theta} = \alpha_1^{-1}(\overline{X})$.
- Now, convergence of $\overline{X}$ to m (law of large numbers) entails convergence of $\hat{\theta}$ to θ:

$$\text{plim}(\hat{\theta}) = \alpha_1^{-1}(\text{plim}(\overline{X})) = \alpha_1^{-1}(m) = \theta$$

- Notice: if $\alpha_q^{-1}()$ were not continuous, $\overline{X}$ could be very close of m and $\alpha_q^{-1}(\overline{X})$ **not** close to $\alpha_q^{-1}(m) = \theta$.

Consistency is not everything!

- Consistency is an asymptotic property. It tells us what happen when the sample size goes to infinity.
- In practice, we may be limited to small samples, and then the consistency property offers little confort.
- Example: (artificial). In a $\mathcal{P}(\lambda)$,

$$\hat{\lambda}_n = \begin{cases} 0 & \text{if } n < 10^5. \\ \overline{X} & \text{if } n \geq 10^5. \end{cases}$$

  would be consistent (but pretty bad for sample sizes n below 10^5!).
- Consistency is reassuring, but we need to check for realistic sample sizes (often through simulation).

Efficiency

- Among estimators which are both unbiased, it makes sense to chose the one with smallest variance.
- For $\hat{\theta}_1$ and $\hat{\theta}_2$ both unbiased estimators of θ, we define *efficiency of* $\hat{\theta}_1$ *relative to* $\hat{\theta}_2$ as:

$$\frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}$$

- Assume Var(θ_1) were the lowest attainable. Then, any estimator with efficiency 1 relative to $\hat{\theta}_1$ will be called **efficient**.
- But, how do we find a $\hat{\theta}_1$ which cannot be improved upon?

The Cramer-Rao bound

- It turns out that we do have a universal yardstick, under *regularity* conditions (more on that later)
- For any unbiased $\hat{\theta}$ based on n observations under regularity conditions:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)};$$

  this is the celebrated Cramer-Rao lower bound.
- $I(\theta)$ is the so-called Fisher information contained in one observation, and is defined as:

$$I(\theta) = E\left(\frac{\partial \log f(x;\theta)}{\partial \theta}\right)^2$$

Intuition for Fisher information

- Why is $I(\theta)$ a measure of information?
- Imagine a given (fixed) x;

$$\left(\frac{\partial \log f(x;\theta)}{\partial \theta} \right)^2$$

measures how fast $\log f(x;\theta)$ changes in response to changes in θ.
- If $\log f(x;\theta)$ were very flat, close values of θ would have similar likelihood, and we would be very uncertain about the "true" θ.
- If $\log f(x;\theta)$ changes fast, it gives much information about θ.
- If we average the derivative over possible values of X we have Fisher information.

Efficient estimators and the Cramer-Rao bound

- Under regularity conditions, if

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{nI(\theta)};$$

the Cramer-Rao lower bound implies the unbiased $\hat{\theta}$ cannot be improved upon by any other unbiased estimator. It is then called **efficient.**
- We *know* what the optimum is before we start.
- No fear that there is a better estimator that just didn't occur to us!

The Cramer-Rao bound: historical notes

- Harald Crameŕ (1892-1985), swedish statistician, author of the extremely influential *Mathematical Methods of Statistics* (1946), still a good reading.
- C.R.Rao (1920-), a distinguished indian statistician. Aside from the Cramer-Rao bound, other contributions like the celebrated Rao-Blackwell theorem (in the same vein than the Cramer-Rao bound, but more powerful).
- The original publications date of 1945 (Rao) and 1946 (Cramer).

What are those regularity conditions?

- Basically,
  1. The support of the distribution does not depend on the parameter. Example of violation: $U(0,\theta)$.
  2. The log likelihood function "sufficiently smooth": differentiable and order of integration and differentiation interchangeable:

$$\frac{\partial}{\partial \theta} E(\log f(x,\theta)) = E\left(\frac{\partial \log f(x,\theta)}{\partial \theta} \right)$$

- Failure of these conditions render unusable the Cramer-Rao bound.

A trick to compute the Cramer-Rao bound.

- It turns out that

$$E\left(\frac{\partial \log f(X,\theta)}{\partial\theta}\right)^2 = -E\left(\frac{\partial^2 \log f(X,\theta)}{\partial\theta^2}\right)$$

- Either expression can be used to compute Fisher's information (the denominator of the Cramer-Rao bound).
- Usually best the second derivative, but sometimes looking at the first we can easily compute its mean value.

The Cramer-Rao bound: examples (I)

We know $\overline{X}$ is unbiased for λ in a $\mathcal{P}(\lambda)$. Its variance is λ/n. Is there anything better?

$$\begin{aligned}
\log f(X,\lambda) &= -\lambda + X\log(\lambda) - \log(X!) \\
\frac{\partial \log f(X,\lambda)}{\partial\lambda} &= -1 + X/\lambda = \left(\frac{X-\lambda}{\lambda}\right) \\
E\left(\frac{X-\lambda}{\lambda}\right)^2 &= \frac{1}{\lambda}
\end{aligned}$$

The Cramer-Rao is

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{1}{n\frac{1}{\lambda}} = \frac{\lambda}{n}$$

so $\overline{X}$ is optimal in the unbiased class.

The Cramer-Rao bound: examples (II)

- We might have missed the fact that:

$$E\left(\frac{X-\lambda}{\lambda}\right)^2 = \frac{1}{\lambda};$$

- In that case, taking the second derivative of

$$\left(\frac{X-\lambda}{\lambda}\right)$$

would have readily given us $1/\lambda$.

The Cramer-Rao bound: examples (III)

- Consider estimation of p in a binary distribution.
- Moment and MLE is $\hat{p} = \overline{X}$ with variance $p(1-p)/n$.
- We have,

$$\begin{aligned}
\log f(X,p) &= X\log(p) + (1-X)\log(1-p) \\
\frac{\partial \log f(X,p)}{\partial p} &= \frac{X}{p} - \frac{1-X}{1-p} \\
E\left(\frac{X}{p} - \frac{1-X}{1-p}\right)^2 &= E\left(\frac{X-p}{p(1-p)}\right)^2 = \frac{1}{p(1-p)}
\end{aligned}$$

- The CR bound is then,

$$\mathrm{Var}(\hat{p}) \geq \frac{1}{n\frac{1}{p(1-p)}} = \frac{p(1-p)}{n}$$

and $\hat{p} = \overline{X}$ is efficient.

Some facts about the Cramer-Rao bound

- The CR bound may not be attainable.
- What it says is that we can do no better...
- ...not that we can do as well.
- Hence, estimators with efficiency 1 as defined previously, may not exist.
- In general, the MLE reaches the CR lower bound, at least asymptotically.

The concept of sufficiency (I)

- To obtain estimators, we have made use of a *statistic*, a function of the sample.
- Are we losing something?
- Or, could we do better looking individually at each sample value, rather than to a summarizing function?
- Loose idea: when a statistic "squeezes all the juice" out of a sample, it is sufficient.
- We have to formalize this "squeezing" property.

The concept of sufficiency (II)

- If given a statistic $S = S(\vec{X})$ the conditional density (or probability)
$$f(\vec{X}|S) = \frac{f_{\vec{X}}(\vec{X} : \theta)}{f_S(S; \theta)}$$
is independent of θ, $S(\vec{X})$ is said to be **sufficient** for θ.
- Motivation: if once we know $S = S(\vec{X})$ the density (or probability) of the sample values does not depend on θ, *knowing those individual sample values cannot be of help in determining* θ.
- All information about θ is then contained in $S = S(\vec{X})$.

The concept of sufficiency (III)

- Let $X_1, \ldots, X_n \sim \mathcal{P}(\lambda)$. Let $S = X_1 + \cdots + X_n$. We know $S \sim \mathcal{P}(n\lambda)$. Then
$$\begin{aligned} f(\vec{X}|S) &= \frac{f_{\vec{X}}(\vec{X} : \lambda)}{f_S(S; \lambda)} \\ &= \frac{\prod_{i=1}^{n} e^{-\lambda} \lambda^{X_i} / X_1!}{e^{-n\lambda}(n\lambda)^S / S!} \\ &= \frac{S!}{X_1! X_2! \ldots X_n!} n^{-S} \end{aligned}$$
- Therefore, S (or any other 1-1 function of S) is sufficient for λ.

The concept of sufficiency (IV)

- As a further example, let's consider the ordered sample $X_{(1)}, \ldots, X_{(n)}$.
- If sampled values are i.i.d., values may arise in any order.
- *Given* $X_{(1)}, \ldots, X_{(n)}$, any order is equally likely, with probability $1/n!$, whichever the parameter(s) of the distribution may be.
- Therefore, $X_{(1)}, \ldots, X_{(n)}$ is always a sufficient statistic, although of little interest (it doesn't "compact" information).

The factorization theorem (I)

- If we can decompose the joint density (or probability) as a product,
$$f_{\vec{X}}(\vec{X} : \theta) = g(S(\vec{X}); \theta) \times h(\vec{X})$$
where $h(\vec{X})$ does **not** depend on θ, then S is sufficient.
- Quite easy to prove.
- Quite practical; we only have to see which function (or functions) of the sample "carry with them" the parameter θ.

The factorization theorem (II)

- Take the Poisson case again. We have,
$$\begin{aligned} f_{\vec{X}}(\vec{X}; \lambda) &= \prod_{i=1}^{n} e^{-\lambda} \lambda^{X_i} / X_i! \\ &= \underbrace{e^{-n\lambda} \lambda^{X_1 + \ldots + X_n}}_{g(S,\lambda)} \times \underbrace{\prod_{i=1}^{n} (1/X_i!)}_{h(\vec{X})} \end{aligned}$$
- Clearly, $S = X_1 + \ldots + X_n$ is sufficient.

The factorization theorem (III)

- MLE have "built in" sufficiency.
- Using the factorization theorem, to maximize the left hand side of
$$f_{\vec{X}}(\vec{X} : \theta) = g(S(\vec{X}); \theta) \times h(\vec{X})$$
as a function of θ**,** we only need $g(S(\vec{X}); \theta)$;
- The term $h(\vec{X})$ is just a constant in the likelihood function.

Some ill-behaved distributions

- Most distributions in common us have sufficient statistics for their parameters.
- This is not always the case. Consider the Cauchy distribution (aka t_1) with location θ:

$$f_X(x:\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$$

- If you use the factorization theorem to look for sufficient statistics,

$$f_{\vec{X}}(\vec{X}:\theta) = g(S(\vec{X});\theta) \times h(\vec{X})$$

  hard as you may try, you will at least need the ordered sample (which is always a sufficient statistic).
- No further reduction is possible.