Properties of estimators

Susan Thomas
IGIDR, Bombay

26 August, 2008

Recap

- Every estimator generates estimates based on the sample. Estimates are rv, which have a distribution.
- Statistical inference is about understanding the distribution of estimators for a given sample size.
- Inference is based on statistical theory: Law of Large Numbers (LLN) and the Central Limit Theorem (CLT).
- For example, CLT: as $n \to \infty$, distribution of sample mean generated by any distribution with finite mean μ and variance σ^2 tends to $N(\mu, \sigma^2/n)$.
- Theory also tells us the sampling distribution of some MLE statistics. For eg., the LR $\sim \chi^2(n)$.
- All sampling distributions have limiting distributions.

Recap

- Inference involves testing a null hypothesis: H_0
- Estimators are selected based on the kind of errors they minimise:
  1. Type I error: rejecting H_0 when it is true. (Leads to *size* of the test of the estimator.)
  2. Type II error: accepting H_0 when it is false. (Leads to the *power* of the test of the estimator.)
- Estimators are based on minimising the errors. Some features of a good test of an estimator are tests of: unbiasedness, consistency, power.

Attributes used to compare estimators

- “Finte sample properties” of an estimator: can be used to compare estimators independent of sample size.
- “Asymptotic properties”: features of the estimator that are not known in finite sample.

Finite sample properties of estimators

- Unbiased estimators: An estimator of parameter θ is unbiased if its sampling distribution mean is the population parameter itself.

$$E(\hat{\theta}) = \theta$$

  - Every estimator $\hat{\theta}$ can be written as

$$\hat{\theta} = \theta + B$$

    where $B(\hat{\theta})$ is the bias, where $B(\hat{\theta}) = E(\hat{\theta} - \theta)$.
  - Then the condition of unbiased estimator can also be written as:

$$E(\hat{\theta} - \theta) = \mathrm{Bias}(\hat{\theta}|\theta) = 0$$

- Efficient unbiased estimators: One estimator, $\hat{\theta}_1$ of θ is more efficient than another estimator, $\hat{\theta}_2$ if the sampling distribution variance of $\hat{\theta}_1$ is less than the variance of $\hat{\theta}_2$

$$\mathrm{var}(\hat{\theta}_1) < \mathrm{var}(\hat{\theta}_2)$$

MSE of an estimator

- Mean squared error of estimators, MSE, is defined as:

$$\text{MSE} = E(\hat{\theta} - \theta)^2$$

- Then:

$$\begin{aligned}
\text{var}(\hat{\theta}) = E(\hat{\theta} - E(\hat{\theta}))^2 &= E(\hat{\theta} - (\theta + B))^2 \\
\text{var}(\hat{\theta}) &= E(\hat{\theta} - \theta - B)^2 \\
&= E(\hat{\theta} - \theta)^2 - 2E(\hat{\theta} - \theta)B + B^2 \\
\text{var}(\hat{\theta}) &= E(\hat{\theta} - \theta)^2 - B(\hat{\theta})^2 \\
\text{or } E(\hat{\theta} - \theta)^2 = \text{MSE}(\hat{\theta}) &= \text{var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}$$

- Usually, the estimator is selected based on *minimum* MSE.
- If the estimator is unbiased, then MSE($\hat{\theta}$) = var($\hat{\theta}$).

Example: Comparing estimators for μ of a normal distribution

- Two estimators:
  1. $\hat{\theta}_1$ = first observation in the sample of size n.
  2. $\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Biasedness:
  1. $E(\hat{\theta}_1) = E(x_1) = \mu$
  2. $E(\hat{\theta}_2) = E(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \mu$

  Both estimators are unbiased.
- Efficiency:
  1. $\text{var}(\hat{\theta}_1) = \text{var}(x_1) = \sigma^2$
  2. $\text{var}(\hat{\theta}_2) = \text{var}(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(x_i) = \sigma^2/n$

  $\text{var}(\hat{\theta}_2) < \text{var}(\hat{\theta}_1)$.
  $\hat{\theta}_2$ is the more efficient estimator.
- Statistical theory defines lower bounds on the minimum variance that an unbiased estimator can achieve for a parameter.
  This is the **Cramer-Rao** bound.

The Cramer-Rao lower bound

- For a rv x which has a density distribution that satisfies *some regularity conditions*:
  1. $f(x)$ has continuous second derivatives.
  2. θ is not at the boundary of possible parameter values
  3. The *range* of x does not depend upon θ.
  4. Conditions on the third derivative of $\ln L$ that allow the calculation of the Taylor series, and the truncation of the Taylor series beyond the second derivative.
- Then, the variance of an unbiased estimator of θ will always be greater than, or equal to,

$$[\mathbf{I}(\theta)]^{-1} = \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \right] \right)^{-1}$$

- Where $I(\theta)$ is called the Fisher Information value.
- And the variance bound is the inverse of the Fisher Information value.

Interpretation of the Fisher Information

- For $l(\theta) = \ln L(\theta)$, that has a first derivative in θ of $l'(\hat{\theta})$ and second derivative of $l''(\hat{\theta})$, the Taylor expansion is:

$$l(\theta) = l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2 + \ldots$$

- At the top of the likelihood function, $l'(\hat{\theta}) = 0$.

$$l(\theta) \approx l(\hat{\theta}) + \frac{1}{2} l''(\hat{\theta})(\theta - \hat{\theta})^2$$

- The behaviour of $l(\theta)$ in the neighbourhood of $\hat{\theta}$ is largely determined by $l''(\hat{\theta})$, a measure of the local curvature of l.

Interpretation of the Fisher Information

- If the dataset and the model are very strong, as we move away from θ_n, $I(\theta)$ would drop off sharply.
  This is a case where there is a lot of information, i.e. I is large.
- The variance of the estimator $\hat{\theta}$ will be small if it's Fisher information value $I(\hat{\theta})$ is large. Ie, $\hat{\theta}$ is more efficient.
- If $I(\hat{\theta})$ is small instead, it means that the likelihood function has a slow flat top where we didn't really know $\hat{\theta}_n$ from other values around it.
  This makes for a less efficient estimator.
- The larger the Fisher Information, the easier it is to identify an efficient estimator for θ.

Example: Cramer-Rao bound for the Bernoulli MLE

- We know that the Bernoulli likelihood function is:

$$\ln L(\theta) = \sum_{i=1}^{n} y_i \ln \theta + \sum_{i=1}^{n} (1 - y_i) \ln (1 - \theta)$$

- Calculating the Fisher Information for this:

$$\begin{aligned}
\partial \ln L / \partial \theta &= \left(\frac{\sum_{i=1}^{n} y_i}{\theta} - \frac{\sum_{i=1}^{n} (1 - y_i)}{1 - \theta} \right) \\
\partial^2 \ln L / \partial \theta^2 &= \left(- \frac{\sum_{i=1}^{n} (1 - y_i)}{(1 - \theta)^2} - \frac{\sum_{i=1}^{n} y_i}{\theta^2} \right) \\
\mathbf{I}(\theta) &= -E \left[- \frac{\sum_{i=1}^{n} (1 - y_i)}{(1 - \theta)^2} - \frac{\sum_{i=1}^{n} y_i}{\theta^2} \right] \\
&= \left[\frac{1}{\theta(1 - \theta)} \right]
\end{aligned}$$

- Cramer-Rao bound sets the variance as $I(\theta)^{-1} = \theta(1 - \theta)$
- The Cramer-Rao lower bound does not apply to the Bernoulli when $\theta = 1$.

Example: Cramer-Rao bound for the Poisson Distribution

- If x is distributed as Poisson, $P(\theta)$, where

$$f(x) = \frac{e^{-\theta}\theta^{x}}{x!}$$

  What is the Cramer-Rao lower bound for a θ estimator?
- $\ln L = -n\theta + \left(\sum_{i=1}^{n} x_i\right) \ln\theta - \sum_{i=1}^{n} \ln\left(x_i!\right)$
- $\partial \ln L / \partial\theta = -n + \frac{\sum_{i=1}^{n} x_i}{\theta}$
- $\partial^2 \ln L / \partial\theta^2 = -\frac{\sum_{i=1}^{n} x_i}{\theta^2}$
- Cramer-Rao bound, $\mathrm{I}(\theta)^{-1} = -E\left(-\frac{\sum_{i=1}^{n} x_i}{\theta^2}\right)^{-1}$
- $E(x_i) = \theta$ for a Poisson distributed rv.
- Variance = CR bound = θ/n

Implications of the Cramer-Rao lower bound

- If a likelihood function can be defined for the rv, such that the $\ln L(\theta)$ is differentiable in θ, then we can calculate the lowest possible variance for any estimator for θ.
- If there exists an estimator that achieves the Cramer-Rao lower bound, then it is the most efficient estimator for θ.
- If there exists a *linear* estimator (ie, a linear function of the data) which has minimum variance among linear unbiased estimator, it is called the *best linear unbiased estimator*, BLUE.

Properties of the Maximum Likelihood Estimator, MLE

Properties of MLE

1. Consistency:
$$\text{plim}\hat{\theta}_{mle} = \theta$$
2. Asymptotic normality: this is derived from the CLT where the sampling distribution of the MLE is:
$$\hat{\theta}_{mle} \sim N[\theta, [\mathrm{I}(\theta)]^{-1}]$$
3. Asymptotic efficiency: MLE achieves the Cramer-Rao lower bound of variance for an estimator.
4. Invariance: the MLE of a function of θ is the function evaluated at θ_{mle}.

Consistency of the MLE

- Under fairly general conditions, the MLE is consistent.
- Wald's consistency theorem : the MLE is consistent if l_t satisfies certain regularity conditions, θ is restricted to lie in a compact space, and the model is asymptotically identified.
- How can consistency of the MLE break?
  1. With models where the number of parameters rises with n,
  2. With models which have characteristics that are not identified asymptotically.

Asymptotic efficiency of MLE

- The maths: If θ_n is the MLE of θ based on a ramdom sample of size n from the distribution of x, and if $\mathrm{I}(\theta)$ is the Fisher information, then if we define:

$$Z_n = \frac{(\hat{\theta}_n - \theta)}{\sqrt{1/n\mathrm{I}(\theta)}}$$

Given that $f(x)$ follows the Cramer-Rao regularity conditions,

$$\lim_{n \to \infty} P(Z_n \leq x) = \Phi(x) \qquad \forall x \in R$$

- The english: For rv x drawn from $f(x)$ that satisfies the Cramer-Rao smoothness conditions, and a large sample size n, the sampling distribution of the MLE is approximately gaussian with mean θ and variance equal to the Cramer-Rao lower bound.
- Ie, the MLE is *asymptotically efficient*.

Example: Efficiency of the Bernoulli MLE

- Y is a sample of size n drawn from a Bernoulli distribution with parameter, p.
- The MLE for p is $\hat{p} = \sum_{i=1}^{n} y_i$.
- Then according to the CR bound, variance for $\hat{p}_{mle}$ is $(nI(\theta))^{-1} = \hat{p}(1 - \hat{p})/n$
- Then, the $100(1 - \alpha)\%$ confidence interval is:

$$\hat{p} \pm z(\alpha/2)\sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$