

Example of achieving CRLB, informal hypothesis-testing, large-sample ML, consistency and efficiency

Lecture #10: Tuesday, 3 February 2005
Lecturer: Prof. Charles Elkan
Scribe: Sourav Bandyopadhyay
Reviewer: Max Chang

1 Achieving the Cramer-Rao Lower Bound

First we remember that the CRLB states that there exists an estimator $g(x)$ whose variance is $var[g(x)] = \frac{1}{var[s(x, \theta)]}$ if and only if (iff) the score function can be written as:

$$s(x, \theta) = \frac{\partial}{\partial \theta} \ln(p(x, \theta)) = b(\theta)[h(x) - \theta]$$

if this is true then $g(x) = h(x)$ is the MVUE.

1.1 Lemma:

$$var[h(x)] = \frac{1}{b(\theta)}$$

Proof:

$$\begin{aligned} var[g(x)] &= \frac{1}{var[s(x, \theta)]} \\ 1 &= var[h(x)]var[s(x, \theta)] \\ &= cov[h(x), s(x, \theta)]^2 \\ var[s(x, \theta)] &= var[b(\theta)[h(x) - \theta]] \\ &= b(\theta)^2 var[h(x) - \theta] \\ &= b(\theta)^2 var[h(x)] \\ var[h(x)] &= \frac{var[s(x, \theta)]}{b(\theta)^2} \\ &= \frac{1}{b(\theta)^2 var[h(x)]} \\ var[h(x)]^2 &= \frac{1}{b(\theta)^2} \\ var[h(x)] &= \frac{1}{b(\theta)} \end{aligned}$$

2 A Cramer-Rao Example

Let $x = (x_1 \dots x_n)$ be the result of n independent coin flips, with success probability θ . As usual, $p(x, \theta) = \theta^{m(x)}(1 - \theta)^{(n - m(x))}$ where $m(x)$ is the number of successes observed. So

$$\ln(p(x, \theta)) = m(x)\ln(\theta) + (n - m(x))\ln(1 - \theta)$$

$$s(x, \theta) = \frac{\partial}{\partial \theta} \ln(p(x, \theta)) = \frac{m(x)}{\theta} - \frac{n - m(x)}{\ln(1 - \theta)} = \frac{n}{\theta(1 - \theta)} \left(\frac{m(x)}{n} - \theta \right)$$

where we choose $g(x) = \frac{m(x)}{n}$. So this $g(x)$ is an MVUE and its variance is $\frac{\theta(1-\theta)}{n}$.

3 Now a Digression

Note that if you have n binary trials with success probability p , then the expected number of successes is np and the variance is npq . Often p and n are unknown, but p is small and n is large. In this case the variance and the expectation are approximately equal. This fact can be used to test informally whether the observed number of successes in two different scenarios is significantly different.

For example, suppose there were three abductions of children by strangers in California last year, and six this year. The observed rate has doubled. Is this a terrible crime wave?

The answer is no. Let the null hypothesis be that the true expected number per year is $np = 3$, with random variability. Under this hypothesis, the standard deviation is around $\sqrt{3} = 1.7$. About $2/3$ of years will have a rate within \pm one standard deviation of the mean, and about 95% within \pm two standard deviations. In this application, about one year out of every three the number will be zero, or five or more, without any change in the underlying rate.

Even with a very large sample, the number of information-rich examples may still be very low. For example, there are over five million children in California but very little information is available about whether there has been a change in the probability of abduction.

4 Notes on testing a hypothesis

- Which null hypothesis you choose should depend on your point of view, and can change your final conclusion. Here, should H_0 be that $np = 3$, or that $np = 4.5$, where 4.5 is our best guess of the true rate, assuming that the true rate is constant? Which H_0 to choose is a real-world question, not a technical mathematical one.
- Once you have chosen H_0 , the mathematical question is "what is the probability of either the observed outcome, or a more extreme outcome?" The definition of "more extreme" depends on the real-world scenario.
- The probability defined in (2) is called the p -value. Your final conclusion is based on comparing the p -value to a threshold. Which threshold you use is again a real-world question, not a mathematical one.

5 Fisher Information

The variance of the score function is a formalization of the concept "amount of information" from the 1920s that predates Shannon's famous notion of entropy (1948). Of course, the two are related.

Fisher information is additive, because variances are additive. If the sample (i.e. training set) is a set of iid observations, then the total information is n times the information provided by each observation.

6 Large-sample Maximum-Likelihood

Let $p^*(x_i, \theta)$ be the distribution followed by a single element of large element of a large iid sample of size n . Then we have:

$$l(x, \theta) = \ln(p^*(x, \theta)) = \sum_i \ln(p^*(x_i, \theta))$$

Given any θ we can think of $l(x, \theta)$ as a function of x , i.e. a random variable. It's a different random variable for each θ . So for each n , let $\hat{\theta}_n$ be the MLE. Assuming that the MLE is not a "corner case" solution, because it maximizes the log likelihood, $\hat{\theta}_n$ is a solution of the equation $\frac{\partial}{\partial \theta} l(x, \theta) = 0$. Remember that this is the score function called $s(x, \theta)$ before.

We are going to prove that the MLE is essentially an ideal estimator as $n \rightarrow \infty$. More precisely, with probability one

- The MLE tends towards the true θ
- The variance of the MLE tends towards the Cramer-Rao lower bound.

7 Consistency

Definition: The sequence $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \dots$ is consistent if for all θ , $\hat{\theta}_n(x) \rightarrow \theta$. That is that our estimate converges toward the true θ . Remember that each $\hat{\theta}_n$ is a function of x . It's too much to ask that convergence be true for all x . There are weak and strong versions of the definition using different probabilistic conditions on x . Note:

- The sequence $\hat{\theta}_n$ can be consistent even though each $\hat{\theta}_n$ is *not* unbiased.
- Conversely, the sequence can fail to be consistent, even though each $\hat{\theta}_n$ is *is* unbiased.
- A sequence can be consistent, but still converge very slowly, e.g. if each estimator throws away some useful information.

7.1 Theorem

For large n , MLEs

- are consistent
- variance goes to Cramer-Rao lower bound of $\frac{1}{\text{var}[s(x_1 \cdots x_n, \theta)]}$

Note on Cramer-Rao lower bound.

$$\begin{aligned} \text{var}[s(x_1 \cdots x_n, \theta)] &= \sum_{i=1}^n \text{var}[s(x_i, \theta)] \\ &= n \cdot \text{var}[s(x_1, \theta)] \end{aligned}$$

Lemma:

$$\begin{aligned} s(x_1 \cdots x_n, \theta) &= \frac{\partial}{\partial \theta} \ln(p(x_1 \cdots x_n, \theta)) \\ &= \frac{\partial}{\partial \theta} \sum \ln(p^*(x_i, \theta)) \\ &= \sum \frac{\partial}{\partial \theta} \ln(p^*(x_i, \theta)) \\ &= \sum s^*(x_i, \theta) \\ \text{var}[s(x_1 \cdots x_n, \theta)] &= \sum_{i=1}^n \text{var}[s^*(x_i, \theta)] \\ &= n \cdot \text{var}[s^*(x_1, \theta)] \end{aligned}$$

Note: $\text{var}[s^*(x_i, \theta)]$ is called the Fisher Information Content (FI), which is additive.

7.2 What we mean by MLE

Given n , let $\hat{\theta}$ be the MLE. So:

$$\hat{\theta} = \text{argmax } l(x_1 \cdots x_n, \theta)$$

and

$$s(x_1 \cdots x_n, \theta) = \frac{\partial}{\partial \theta} l(x_1 \cdots x_n, \theta) = 0$$

for $\theta = \hat{\theta}$

Next week we will prove that for large n , MLEs are consistent and have variance only slightly above the Cramer-Rao lower bound. This second property is called efficiency. We shall use several intermediate results.

- The Taylor expansion of the score function, $s(x_1 \cdots x_n, \theta)$ around θ .
- The weak law of large numbers: Let $x_1 \cdots x_n$ be iid, random variables with mean μ and variance σ^2 . Let $S_n = \sum_{i=1}^n X_i$ let $\epsilon > 0$. Then:

$$P\left(\left|\frac{S_n}{n} - \mu\right| > \epsilon\right) \rightarrow 0$$

as $n \rightarrow \infty$

- The central limit theorem. Using the same assumptions as the weak law of large numbers:

$$P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq z\right) \rightarrow \Phi(z)$$

Where $\Phi(z)$ is the cumulative $N(0, 1)$ distribution.