

The Score Function and Cramer-Rao Lower Bound

Lecture #9: Tuesday, 1 February 2005
Lecturer: Prof. Charles Elkan
Scribe: Max Chang
Reviewer: Sourav Bandyopadhyay

1 Expectation and Variance of the Score Function

As noted in an earlier lecture, the likelihood function is defined as $p(x, \theta)$. Since it is often preferable to work with logarithms, we use the log likelihood function $l(x, \theta) = \ln p(x, \theta)$. The score function is the derivative of the log likelihood function with respect to θ .

$$s(x, \theta) = \frac{\partial}{\partial \theta} l(x, \theta) = \frac{1}{p(x, \theta)} \frac{\partial}{\partial \theta} p(x, \theta)$$

Generally, given x we want to find a local maximum for $p(x, \theta)$ by guessing θ such that $p(x, \theta)$ is high and $\frac{\partial}{\partial \theta} p(x, \theta) = 0$. Hence for fixed x , the score function says which values of θ are best: the optimum score is zero and any non-zero score is less desirable.

To calculate the expected value of the score function, we integrate over all values of x :

$$\begin{aligned} E[s(x, \theta)] &= \int_{x \in X} s(x, \theta) p(x, \theta) dx \\ &= \int_{x \in X} \frac{1}{p(x, \theta)} \frac{\partial}{\partial \theta} p(x, \theta) p(x, \theta) dx \\ &= \int_{x \in X} \frac{\partial}{\partial \theta} p(x, \theta) dx \\ &= \frac{\partial}{\partial \theta} \int_{x \in X} p(x, \theta) dx \quad * \\ &= \frac{\partial}{\partial \theta} 1 = 0 \end{aligned}$$

* Moving the derivative outside the integral can be done as long as the limits of integration are fixed, i.e. they do not depend on θ . We can prove this easily for the discrete case:

$$\begin{aligned} \sum_{x \in X} \frac{\partial}{\partial \theta} p(x_i, \theta) dx &= \frac{\partial}{\partial \theta} p(x_1, \theta) + \frac{\partial}{\partial \theta} p(x_2, \theta) + \dots + \frac{\partial}{\partial \theta} p(x_n, \theta) \\ &= \frac{\partial}{\partial \theta} [p(x_1, \theta) + p(x_2, \theta) + \dots + p(x_n, \theta)] \\ &= \frac{\partial}{\partial \theta} \sum_{x \in X} p(x_i, \theta) \end{aligned}$$

Since the expectation of the score function is 0, the variance is simply: $\text{var}[s(x, \theta)] = E[s(x, \theta)^2]$.

2 The Cramer-Rao Theorem

Sometimes we can't find an MVUE, but we can find an unbiased estimator. In this case we'd like to know how good its variance is. One way to do this is to compare it to some lower bound. The result we'll see now gives such a lower bound.

Suppose that the score function has small variance, for some θ . This means that all x have scores close to zero, so whatever the x that we observe, it doesn't provide much information about the value of θ . Hence every estimator of theta based on x is likely to be bad.

More specifically, the smaller the variance of $s(x, \theta)$, the bigger the variance of any unbiased estimator $g(x)$, including the MVUE.

Theorem 2.1. Suppose the family of distributions P_θ is defined by a density function $p(x, \theta)$ where $\theta \in \mathbb{R}$. Let $g(x)$ be any unbiased estimator of θ . Then

$$\text{var}[g(x)] \geq \frac{1}{\text{var}[s(x, \theta)]}$$

Intuitively, the idea is to look at $p(x, \theta)$ and evaluate $\text{var}[\frac{\partial}{\partial \theta} \ln p(x, \theta)]$. For each θ , know that you cannot find an estimator better (unbiased and with lower variance) than $\frac{1}{\text{var}[s(x, \theta)]}$.

Proof

$$\begin{aligned} \int g(x) p(x, \theta) dx &= \theta \\ \frac{\partial}{\partial \theta} \int g(x) p(x, \theta) dx &= 1 \\ \int g(x) \frac{\partial}{\partial \theta} p(x, \theta) &= 1 \\ \int g(x) s(x, \theta) p(x, \theta) dx &= 1 \\ E[g(x) s(x, \theta)] &= 1 \end{aligned}$$

Consider the covariance of $g(x)$ and $s(x, \theta)$.

$$\begin{aligned} \text{cov}[g(x), s(x, \theta)] &= E[(g(x) - \theta) (s(x, \theta) - 0)] \\ &= E[g(x) s(x, \theta) - \theta s(x, \theta)] \\ &= E[g(x) s(x, \theta)] - 0 \\ &= 1 \end{aligned}$$

Reminder, $cov[f(x), f(x)] = var[f(x)]$ and $cov[v(x), w(x)]^2 \leq var[v(x)] \times var[w(x)]$.

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

3 Achieving the Cramer-Rao Lower Bound

When does $var[g(x)] = \frac{1}{var[s(x, \theta)]}$?

Theorem 3.1. There exists an estimator $g(x)$ who variance is $\frac{1}{var[s(x, \theta)]}$ iff the score function can be written $s(x, \theta) = \frac{\partial}{\partial \theta} \ln(p(x, \theta)) = b(\theta)[h(x) - \theta]$, where $h(x)$ is an unbiased estimator of θ . In this case, $h(x)$ is an MVUE with variance $\frac{1}{b(\theta)}$.

Proof

Use the fact that $var[g(x)] \times var[s(x, \theta)] = cov[g(x), s(x, \theta)]^2$ iff $g(x) = \lambda s(x, \theta) + v$. Intuitively, this means that the covariance is maximized when $g(x)$ and $s(x, \theta)$ are linearly related, where the constant b is allowed to depend on θ .

$$\begin{aligned} s(x, \theta) - E[s(x, \theta)] &= b(\theta) (g(x) - E[g(x)]) \\ s(x, \theta) &= b(\theta) g(x) - \theta \end{aligned}$$

$g(x)$ is the MVUE.

Example 3.1. Let $x = (x_1, x_2, \dots, x_n)$ be the result of n independent coin flips with success probability θ . We define the statistic $m(x) = \sum x_i$. The probability distribution function is $p(x, \theta) = \theta^{m(x)}(1 - \theta)^{n - m(x)}$. Thus, we have the log likelihood function $\ln(p(x, \theta)) = m(x) \ln(\theta) + (n - m(x)) \ln(1 - \theta)$ and the score function $s(x, \theta) = \frac{\partial}{\partial \theta} \ln(p(x, \theta))$.

$$\begin{aligned} s(x, \theta) &= \frac{m(x)}{\theta} - \frac{(n - m(x))}{1 - \theta} \\ &= \frac{(1 - \theta)m(x) - \theta(n - m(x))}{\theta(1 - \theta)} \\ &= \frac{(1 - \theta)\frac{m}{n} - \theta(1 - \frac{m}{n})}{\theta(1 - \theta)} \\ &= \frac{n}{\theta(1 - \theta)} \left(\frac{m}{n} - \theta \right) \end{aligned}$$

So $\frac{m}{n}$ is the MVUE.