Asymptotic theory of the MLE. Fisher information

Let $X_1, ..., X_n$ be i.i.d. random variables, with a common pdf/pmf $f(x|\theta)$, where θ is an unknown real parameter. We will assume that $f(x|\theta)$ has two continuous derivatives with respect to θ. As usual, assume that the data is given, $X_1 = x_1, ..., X_n = x_n$. Recall that *likelihood function* is the *joint* pdf/pmf of $X_1, ..., X_n$ viewed as a function of θ. Due to independence,

$$\text{lik}(\theta) = f(x_1, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

To find the MLE $\hat{\theta}_n$ of θ, we maximize the likelihood or, equivalently, the log-likelihood,

$$\ln \text{lik}(\theta) = \sum_{i=1}^{n} \ln f(x_i|\theta) \quad \to \quad \max_{\theta}.$$

Since, by our assumption, the likelihood is differentiable, we can set up the ML equation:

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{n} \ln f(x_i|\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \ln f(x_i|\theta) = 0.$$

Obviously, an important role here is played by the function $\frac{\partial}{\partial\theta} \ln f(x_i|\theta)$. Later, we will see that an equally important role is played by the second derivative, $\frac{\partial^2}{\partial\theta^2} \ln f(x_i|\theta)$. Fisher – the famous English statistician who pioneered the study of MLE – proposed to call

$$\frac{\partial}{\partial\theta} \ln f(x_i|\theta) \quad = \quad \textbf{the 1st score}, \qquad \frac{\partial^2}{\partial\theta^2} \ln f(x_i|\theta) \quad = \quad \textbf{the 2nd score}.$$

These two functions have some important properties, which are far from obvious. Here are some of these properties, without proofs, but with some illustrating examples.

Rule 1: The expected value of the first score is 0.

$$\mathbf{E}\left(\frac{\partial}{\partial\theta} \ln f(X_i|\theta)\right) = 0.$$

Definition 2. The variance of the first score is denoted

$$I(\theta) = \mathbf{Var}\left(\frac{\partial}{\partial\theta} \ln f(X_i|\theta)\right),$$

and is called **Fisher information about the unknown parameter** θ, contained in a single observation X_i.

Rule 2: The Fisher information can be calculated in two different ways:

$$I(\theta) = \mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f(X_i|\theta)\right) = -\mathbf{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(X_i|\theta)\right). \tag{1}$$

The theory of MLE established by Fisher results in the following main

Theorem 1. For large n, the MLE $\hat{\theta}$ is asymptotically normally distributed,

$$\hat{\theta}_n \quad \overset{d}{\approx} \quad \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right).$$

Corollary 1. An approximate $(1-\alpha)100\%$ confidence interval (CI) for θ based on the MLE $\hat{\theta}_n$ is given by

$$\hat{\theta}_n \pm z(\alpha/2)\sqrt{\frac{1}{nI(\hat{\theta}_n)}}.$$

In the examples presented below, one should pay attention to the specific order in which the calculations are performed to find the MLE and the corresponding CI.

Example 1. Independent Bernoulli trials. Let $X_1 = x_1, ..., X_n = x_n$ be observed values of i.i.d. random variables, each with the same distribution as a single Bernoulli trial, $X \sim \mathcal{B}(p)$. Here the common pmf is given by

$$f(x|p) = p^x(1-p)^{1-x},$$

so that

$$\ln f(x|p) = x\ln p + (1-x)\ln(1-p).$$

It is convenient to start by calculating the **1st and 2nd scores.** In this case,

$$\frac{\partial}{\partial p}\ln f(x|p) = \frac{x}{p} - \frac{1-x}{1-p}, \quad \frac{\partial^2}{\partial p^2}\ln f(x|p) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.$$

Denote

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The first score allows one to set up the ML equation. By the rules of summation, it becomes

$$\sum_{i=1}^{n}\frac{\partial}{\partial p}\ln f(x_i|p) = \sum_{i=1}^{n}\left(\frac{x_i}{p} - \frac{1-x_i}{1-p}\right) = \frac{n\bar{x}}{p} - \frac{n-n\bar{x}}{1-p} =$$

$$n\left(\frac{\bar{x}}{p} - \frac{1-\bar{x}}{1-p}\right) = n\frac{\bar{x}-p}{p(1-p)} = 0,$$

from which the MLE $\hat{p} = \bar{x}$ can be found. Of course, this result is already known to us!

Next, we can check **Rule 1:**

$$\mathbf{E}\left(\frac{\partial}{\partial p}\ln f(X|p)\right) = \mathbf{E}\left(\frac{X}{p} - \frac{1-X}{1-p}\right) = \frac{p}{p} - \frac{1-p}{1-p} = 1 - 1 = 0.$$

In this particular case, **Rule 1** is obvious. In other less trivial cases **it will be a source of useful information!**

The next step is to find the **Fisher information.** Our equation (1) gives two different formulas for the Fisher information. Usually, the second formula, i.e., the right-hand side of (1), is **numerically easier**. Here, we will verify that both formulas produce the same result. It is highly recommended to use **both formulas**, as **it may provide a valuable further information!**

Using the well known property of the variance $\mathbf{Var}(a + bX) = b^2\mathbf{Var}\,X$, we get

$$I(p) = \mathbf{Var}\left(\frac{\partial}{\partial p}\ln f(X|p)\right) = \mathbf{Var}\left(\frac{X}{p} - \frac{1-X}{1-p}\right) = \mathbf{Var}\frac{X}{p(1-p)} = \frac{1}{p(1-p)}.$$

Let us check next that the second formula in **Rule 2** gives the same result – only easier – as it only uses properties of expectations:

$$I(p) = -\mathbf{E}\left(\frac{\partial^2}{\partial p^2}\ln f(X|p)\right) = \mathbf{E}\left(\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right) = \frac{p}{p^2} + \frac{1-p}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Finally, having found the MLE, $\hat{p} = \bar{x}$, and the Fisher information $I(p)$, we can construct the $(1-\alpha)100\%$ CI using the **Corollary 1**:

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

Our next example will be dealing with the statistical model already discussed in the **Example 8** of the previous Lecture. It is recommended to review it before reading the following.

Example 2. Hardy-Weinberg equilibrium. In this model, the observed data $X_1 = x_1, X_2 = x_2, X_2 = x_3$ comes from *trinomial distribution*, with probabilities p_1, p_2, p_3 of corresponding outcomes satisfying the following equations:

$$p_1(\theta) = (1-\theta)^2, \quad p_2(\theta) = 2\theta(1-\theta), \quad p_3(\theta) = \theta^2.$$

Based on the data, we need to find the MLE $\hat{\theta}$ and construct a confidence interval for the unknown parameter θ. We will see how the above discussed methods can significantly simplify the solution.

Note that in the previous example of i.i.d. Bernoulli random variables, it was sufficient to deal with the marginal pmf $f(x|p)$. This convenience was provided by

the i.i.d. structure. Since in the trinomial model, the responses X_1, X_2, X_3 *are not* independent, we will deal with their *joint* pmf, or the likelihood, $f(x_1, x_2, x_3|\theta)$. For the rest, the methods are similar to the previous example.

For briefness, denote $\mathbf{x} = (x_1, x_2, x_3)$. As we know, the likelihood function is given by

$$f(\mathbf{x}|\theta) = \frac{n!}{x_1!x_2!x_3!} p_1(\theta)^{x_1} p_2(\theta)^{x_2} p_3(\theta)^{x_3}.$$

Therefore, the corresponding log-likelihood is

$$\ln f(\mathbf{x}|\theta) = \ln \frac{n!}{x_1!x_2!x_3!} + x_1 \ln(1-\theta)^2 + x_2 \ln 2\theta(1-\theta) + x_3 \ln \theta^2 =$$

$$\ln \frac{n!}{x_1!x_2!x_3!} + x_2 \ln 2 + (2x_1 + x_2) \ln(1-\theta) + (2x_3 + x_2) \ln \theta.$$

Let us calculate the **1st score**,

$$\frac{\partial}{\partial \theta} \ln f(\mathbf{x}|\theta) = -\frac{2x_1 + x_2}{1-\theta} + \frac{2x_3 + x_2}{\theta}.$$

Since $x_1 + x_2 + x_3 = n$, the 1st score can be simplified, by excluding the variable $x_1 = n - x_2 - x_3$:

$$\frac{\partial}{\partial \theta} \ln f(\mathbf{x}|\theta) = -\frac{2n - (2x_3 + x_2)}{1-\theta} + \frac{2x_3 + x_2}{\theta} = \frac{(2x_3 + x_2) - 2n\theta}{\theta(1-\theta)}. \tag{2}$$

By differentiating the log-likelihood once more, we get the **2-nd score**:

$$\frac{\partial^2}{\partial \theta^2} \ln f(\mathbf{x}|\theta) = -\frac{2n - (2x_3 + x_2)}{(1-\theta)^2} - \frac{2x_3 + x_2}{\theta^2} =$$

$$-\frac{2n}{(1-\theta)^2} - (2x_3 + x_2) \left(\frac{1}{\theta^2} - \frac{1}{(1-\theta)^2} \right). \tag{3}$$

Now we can set up the ML equation. By (2), it becomes

$$\frac{\partial}{\partial \theta} \ln f(\mathbf{x}|\theta) = \frac{(2x_3 + x_2) - 2n\theta}{\theta(1-\theta)} = 0,$$

from which we easily find the MLE:

$$\hat{\theta} = \frac{2x_3 + x_2}{2n}.$$

Next, let us now check out **Rule 1.** Since it deals with the expectation of the 1st score, we need to view it as a random variable. (Recall that the observed counts x_i are realizations of the corresponding random variable X_i). According to **Rule 1**,

$$\mathbf{E}\left(\frac{\partial}{\partial\theta}\ln f(\mathbf{X}|\theta)\right)=\mathbf{E}\left(\frac{(2X_3+X_2)-2n\theta}{\theta(1-\theta)}\right)=0.$$

It follows that

$$\mathbf{E}(2X_3+X_2)=2n\theta, \tag{4}$$

or equivalently

$$\mathbf{E}\hat{\theta}=\mathbf{E}\left(\frac{2X_3+X_2}{2n}\right)=\theta.$$

Thus, in a simpler way, we found again that the MLE $\hat{\theta}$ is an *unbiased estimator* of θ.

Next, let us calculate the Fisher information using the right-hand side of **Rule 2**.

$$I(\theta)=-\mathbf{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(\mathbf{X}|\theta)\right).$$

By (3)–(4),

$$I(\theta)=\frac{2n}{(1-\theta)^2}+2n\theta\left(\frac{1}{\theta^2}-\frac{1}{(1-\theta)^2}\right)=\frac{2n}{\theta}+\frac{2n(1-\theta)}{(1-\theta)^2}=\frac{2n}{\theta(1-\theta)}. \tag{5}$$

This results in the same $(1-\alpha)100\%$ CI for θ, as in the previous **Lecture**.

Let us check out **Rule 2** in (1). Using the just found expression (5) and the formula (2) for the 1st score, we get

$$I(\theta)=\frac{2n}{\theta(1-\theta)}=\mathbf{Var}\left(\frac{2X_3+X_2-2n\theta}{\theta(1-\theta)}\right)=\frac{\mathbf{Var}(2X_3+X_2)}{\theta^2(1-\theta)^2}.$$

Equivalently,

$$\mathbf{Var}(2X_3+X_2)=2n\theta(1-\theta).$$

This gives also the *mean squared error* of the MLE $\hat{\theta}$:

$$\mathbf{E}(\hat{\theta}-\theta)^2=\mathbf{Var}\hat{\theta}=\mathbf{Var}\frac{2X_3+X_2}{2n}=\frac{\mathbf{Var}\,(2X_3+X_2)}{(2n)^2}=\frac{\theta(1-\theta)}{2n}. \tag{6}$$

Example 3. Double exponential, or Laplace distribution. Let $X_1,...,X_n$ be i.i.d., with the common pdf

$$f(x|\theta)=\frac{1}{2\theta}e^{-\frac{|x|}{\theta}},\quad \theta>0.$$

With the i.i.d. data, it is always sufficient to look at the marginal pdf/pmf $f(x|\theta)$. Here

$$\ln f(x|\theta)=-\ln 2-\ln\theta-\frac{|x|}{\theta}.$$

Thus, the 1st and 2nd scores are, respectively

$$\frac{\partial}{\partial\theta}\ln f(x|\theta) = -\frac{1}{\theta} + \frac{|x|}{\theta^2}, \quad \text{and} \quad \frac{\partial^2}{\partial\theta^2}\ln f(x|\theta) = \frac{1}{\theta^2} - \frac{2|x|}{\theta^3}.$$

We can set up the MLE equation:

$$\sum_{i=1}^{n}\left(-\frac{1}{\theta} + \frac{|x_i|}{\theta^2}\right) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n}|x_i|}{\theta^2} = 0 \quad \Longrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|x_i|.$$

Next, by checking out **Rule 1**,

$$\mathbf{E}\left(-\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = \mathbf{E}\left(-\frac{1}{\theta} + \frac{|X|}{\theta^2}\right) = 0,$$

we find that

$$\mathbf{E}|X| = \theta. \tag{7}$$

From this relation, it immediately follows that the MLE,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}|X_i|,$$

is an *unbiased estimator* of θ.

Next, we can find the Fisher information. By using (7) again, we get

$$I(\theta) = -\mathbf{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(X|\theta)\right) = -\mathbf{E}\left(\frac{1}{\theta^2} - \frac{2|X|}{\theta^3}\right) = -\left(\frac{1}{\theta^2} - \frac{2\theta}{\theta^3}\right) = \frac{1}{\theta^2}.$$

Thus, an approximate $(1-\alpha)100\%$ CI for θ is given by

$$\hat{\theta} \pm \frac{z(\alpha/2)}{\sqrt{nI(\hat{\theta})}} = \hat{\theta} \pm \frac{z(\alpha/2)\hat{\theta}}{\sqrt{n}} = \hat{\theta}\left(1 \pm \frac{z(\alpha/2)}{\sqrt{n}}\right).$$

We can also find the variance of the MLE. Note that by the definition of $I(\theta)$,

$$\frac{1}{\theta^2} = I(\theta) = \mathbf{Var}\left(\frac{\partial}{\partial\theta}\ln f(X|\theta)\right) = \mathbf{Var}\left(-\frac{1}{\theta} + \frac{|X|}{\theta^2}\right) =$$

$$\mathbf{Var}\left(\frac{|X|}{\theta^2}\right) = \frac{\mathbf{Var}\,|X|}{\theta^4}.$$

This tells us that

$$\mathbf{Var}|X| = \theta^2.$$

Thus, the variance and the *mean squared error* of the unbiased MLE $\hat{\theta}$ is

$$\mathbf{E}(\hat{\theta}-\theta)^2 = \mathbf{Var}\,\hat{\theta} = \mathbf{Var}\left(\frac{1}{n}\sum_{i=1}^{n}|X_i|\right) = \frac{\mathbf{Var}|X|}{n} = \frac{\theta^2}{n}. \tag{8}$$