Asymptotic theory of the MLE. Fisher information

Let $X_1, ..., X_n$ i.i.d. $\sim f(x|\theta)$, where θ is real. We will assume that the pdf/pmf $f(x|\theta)$ has continuous derivatives with respect to θ. As usual, we have the data: $X_1 = x_1, ..., X_n = x_n$. Recall that the likelihood function is the joint pdf/pmf, which is 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, of 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:

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

From that, we can see that an important role 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 pioneer of studying the 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 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 the **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:

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

These definitions and results lead to the following main

Theorem.

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

Corollary. An approximate $(1-\alpha)100\%$ for θ CI 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 given below, note the specific order in which the calculations are performed to find the MLE.

Example 1. Independent Bernoulli trials: $X_1, ..., X_n$ i.i.d. $\overset{d}{=}$ $X \sim \mathcal{B}(p)$. Here the common pmf is given by

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

so that the log-likelihood is

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

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

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

Note that the first score allows to set up the ML equation. Denote as usual

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

By the rules of summation,

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

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

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

Next we will check **Rule 1:**

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

Note that, although in this case Rule 1 is satisfied trivially, in other, less trivial cases, **it can be a useful source of additional information!**

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

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

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

$$\mathbf{Var}\left(X\left(\frac{1}{p}+\frac{1}{1-p}\right)\right)=\mathbf{Var}\left(\frac{X}{p(1-p)}\right)=\frac{\mathbf{Var}\,X}{(p(1-p))^2}=\frac{p(1-p)}{(p(1-p))^2}=\frac{1}{p(1-p)}.$$

Now, let us check that the second formula in **Rule 2** gives the same result, only somewhat easier, as it uses simpler properties of expectations, not variances:

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

Finally, having found the MLE $\hat{p}$ and the Fisher information $I(p)$. we can construct the $(1-\alpha)100\%$ CI in the usual way:

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

Our next example will bee dealing with the statistical model already discussed in the last Example 8 of the previous Lecture. Please, review it before reading the next example!

Example 2. Hardy-Weinberg equilibrium. Recall that in this model we are dealing with the data $X_1=x_1, X_2=x_2, X_2=x_3$ having trinomial distribution, with the probabilities p_1, p_2, p_2 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. We have done this, but in a somewhat complicated manner. Now we will see how our new methods can greatly simplify the solution.

Note that in the previous the example of i.i.d. Bernoulli random variables, it was sufficient to deal mostly with the marginal pmf $f(x|p)$. That was the convenience provided by the i.i.d. structure. Since in the trinomial model, the responses x_1, x_2, x_3 are not i.i.d., we will deal with their joint pmf, or the likelihood, $f(x_1, x_2, x_3|p_1(\theta), p_2(\theta), p_3(\theta))$. For the rest, the methods will be very 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

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

The corresponding log-likelihood is therefore

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

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

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 $x_1 + x_2 + x_3 = n$, it can be simplified (by excluding variable $x_1 = n - x_2 - x_3$) as

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

By differentiating one more time the log-likelihood, 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 that the MLE equals

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

Let now verify **Rule 1.** Since in it, we want to take the expectation of the score, we need to view it as a random variable. In other words, we view each count x_i as a realization 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 immediately 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, we found again, but much simpler, 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),

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

Next, using the first formula for the Fisher information $I(\theta)$, the formula (2) for the 1st score, and the just found expression (5), 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,

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

and

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

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

Consider briefly one more example.

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

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

As is often the case with the i.i.d. data, it is sufficient to look at the marginal likelihood $f(x|\theta)$. Here

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

Thus, the 1st and scores are, respectively

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

We can set the MLE equation:

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

Next, checking out Rule 1,

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

provides a useful relation

$$\mathbf{E}|X| = \theta. \tag{6}$$

From this relation, it follows directly 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 (6) again, we get

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

On the over hand, 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

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

and therefore

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

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

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