Statistics 358b : Comments on Fisher's Information

February 5, 2014

We showed under the regularity conditions, f being pdf (or pmf and summation accordingly)

$$\begin{aligned}
\mathrm{E}_{\theta_0}\left[\frac{\partial \log\left(f(X;\theta_0)\right)}{\partial\theta}\right] &= \int_R \left[\frac{\partial \log\left(f(x;\theta_0)\right)}{\partial\theta}\right] f(x;\theta_0)dx \\
&= \int_R \left[\frac{1}{f(x,\theta_0)}\frac{\partial f(x;\theta_0)}{\partial\theta}\right] f(x;\theta_0)dx \\
&= \int_R \frac{\partial f(x;\theta_0)}{\partial\theta}dx \\
&= \left.\frac{\partial \int_R f(x;\theta)dx}{\partial\theta}\right|_{\theta=\theta_0} \\
&= \frac{\partial(1)}{\partial\theta} \\
&= 0
\end{aligned}$$

Notice the integration is done with respect to θ_0, and this final statement is therefore true for all θ_0. Thus in our statement we can also write

$$\int_R \left[\frac{\partial \log\left(f(x;\theta)\right)}{\partial\theta}\right] f(x;\theta)dx = 0 \tag{1}$$

for all θ.

Define $\mathrm{I}(\theta)$ in both the 1-D and higher dimensional parameter case, that is as a matrix.

Re : Lemma 8.5.2 A proof.

Remark : Equation (1) is true for all θ but it is not true that

$$\frac{\partial\Lambda(\theta)}{\partial\theta} = \mathrm{E}_{\theta_0}\left[\frac{\partial \log\left(f(X;\theta)\right)}{\partial\theta}\right] = \int_R \left[\frac{\partial \log\left(f(x;\theta)\right)}{\partial\theta}\right] f(x;\theta_0)dx$$

is equal to 0 for all θ. In each integral the integrand is a product of two functions, one with *parameters* θ and θ_0 (not necessarily matching) and in the other integral with *parameters* θ and θ (necessarily matching). If we calculate the

derivative is this second integral, which is valid, it is not the case that the derivative is equal to 0. However if we differentiate (1) with respect to θ, it is the case that this derivative is equal to 0.

The correct method is therefore to differentiate (1).

For real θ

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathrm{E}_\theta\left[\frac{\partial\log\left(f(X;\theta)\right)}{\partial\theta}\right] &= \frac{\partial}{\partial\theta}\int_R\left[\frac{\partial\log\left(f(x;\theta)\right)}{\partial\theta}\right]f(x;\theta)dx\\ &= \frac{\partial}{\partial\theta}0=0\end{aligned}$$

This last line is not true if we calculate

$$\mathrm{E}_{\theta_0}\left[\frac{\partial\log\left(f(X;\theta)\right)}{\partial\theta}\right]=\int_R\left[\frac{\partial\log\left(f(x;\theta)\right)}{\partial\theta}\right]f(x;\theta_0)dx$$

even though the differentiation is valid.

Remark : A variance matrix (also called a variance-covariance or covariance matrix) for the vector X is given by

$$\mathrm{var}(X)=\begin{bmatrix}\mathrm{Cov}(X_1,X_1) & \mathrm{Cov}(X_1,X_2) & \dots & \mathrm{Cov}(X_1,X_d)\\ \mathrm{Cov}(X_2,X_1) & \mathrm{Cov}(X_2,X_2) & \dots & \mathrm{Cov}(X_2,X_d)\\ \vdots & & & \\ \mathrm{Cov}(X_d,X_1) & \mathrm{Cov}(X_d,X_2) & \dots & \mathrm{Cov}(X_d,X_d)\end{bmatrix}$$

If we have a row vector $X=(X_1,X_2,\dots,X_d)$ (a $1\times d$ matrix) then X^tX is a $d\times d$ matrix. The variance matrix is thus the component or element wise matrix of expectations of $(X-\mu)^t(X-\mu)$ where μ is the row vector of means of X_i, that is

$$\mathrm{var}(X)=\mathrm{E}\left[(X-\mu)^t(X-\mu)\right]\ .$$

If one interprets X as a column vector then one has to interpret

$$\mathrm{var}(X)=\mathrm{E}\left[(X-\mu)(X-\mu)^t\right]$$

so one needs to be consistent with the interpretation of row or column random vectors.

Comments on calculation of Fisher's information matrix
In the asymptotic normality of the MLE we use the property that

$$\frac{1}{n}\ell''(\theta_0) \to -\mathrm{I}(\theta_0)$$

Thus we interpret, after taking expectations, in the iid case,

$$\begin{aligned} -I(\theta) &= \lim_{n\to\infty} \frac{1}{n}\mathrm{E}(\ell''(\theta)) \\ &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial^2 log(f(X_i;\theta)}{\partial\theta^2}\right) \\ &= \mathrm{E}\left(\frac{\partial^2 log(f(X;\theta)}{\partial\theta^2}\right) \end{aligned}$$

This is the interpretation that we wish to use. It is also the one that makes Theorem 8.5B page 277 Rice correct.
The later interpretation of

$$-I(\theta) = \lim_{n\to\infty} \frac{1}{n}\mathrm{E}(\ell''(\theta)) \tag{2}$$

is the one that that is actually correct, but in the iid case this simplifies to the definition given in Rice. In the case of Binomial(n,θ) we need to use (2).

To help us with this clarification consider the following example.

Example : Bernoulli and Binomial MLE
Consider the following two related models and sampling from these

- Bernoulli parameter $\theta \in [0,1]$ and iid sampling $Y_1,\ldots,Y_n$
- Binomial, sample size n and parameter $\theta \in [0,1]$ and one observation X

In both there is a sample size n, but in the binomial there is only one observation X. However $Y_1 + \ldots + Y_n$ and X both have the same distribution, Binomial(n,θ). Below for the Bernoulli model data we also write $X = Y_1 + \ldots + Y_n$.
The pmf's are

- Bernoulli :
$$f_1(k;\theta) = (1-\theta)^{1-k}\theta^k \ , \ k = 0,1 \ .$$
- Binomial :
$$f_2(k;\theta) = \binom{n}{k}(1-\theta)^{n-k}\theta^k \ , \ k = 0,1,\ldots,n \ .$$

We will use a subscript 1 and 2 to distinguish these two models and relevant calculations below. Notice in the Bernoulli sampling, the sample size n is reflected in the increasing number of observations, that is $Y_1, \ldots, Y_n$. However in the Binomial sampling game, there is only one observation X and the sample size n is hidden in the pmf of X. For the iid Bernoulli sampling the number of r.v.s observed $Y_1, \ldots, Y_n$ tends to infinity. For the Binomial sampling experiment there is only one r.v., so the number of r.v.s does not tend to infinity. As such we need to interpret the limit distribution or the approximate normal distribution accordingly.

The log likelihood functions are

- Bernoulli :

$$\ell_1(\theta) = \sum_{i=1}^{n} \{(1 - Y_i)\log(1-\theta) + Y_i \log(\theta)\} = (n - X)\log(1-\theta) + X\log(\theta)$$

- Binomial : omitting the terms that do not involve θ, that is $\binom{n}{X}$

$$\ell_2(\theta) = (n - X)\log(1-\theta) + X\log(\theta)$$

Thus we see these two models and data have essentially the same log likelihood, and both yield the MLE

$$\hat{\theta}_n = \frac{X}{n} = \bar{X}_n \ .$$

In both cases we would obtain by the Central Limit Theorem, or more specifically by the method of Moment Generating Functions used in the proof, that

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \theta(1-\theta))$$

in distribution as $n \to \infty$.

For the Bernoulli model the student should verify that Fisher's information is

$$I_1(\theta) = \frac{1}{\theta(1-\theta)}$$

Now consider the Binomial model. There is only 1 observation of X. The second derivative of the log pmf is is exactly the same the as the second derivative of the log lilkelihood.

$$\begin{aligned} \ell_2''(\theta) &= \frac{\partial^2 \left[(n - X)\log(1-\theta) + X\log(\theta)\right]}{\partial \theta^2} \\ &= -\frac{(n-X)}{(1-\theta)^2} + n\frac{X}{\theta^2} \end{aligned}$$

This has expectation

$$\mathrm{E}_\theta(\ell_2''(\theta)) = -\frac{n}{\theta(1-\theta)}$$

and hence

$$\mathrm{E}_\theta\left(\frac{\partial^2 f_2(X;\theta)}{\partial \theta^2}\right) = -\frac{n}{\theta(1-\theta)}$$

- Bernoulli :
$$\mathrm{E}_\theta(\ell_1''(\theta)) = -\frac{n}{\theta(1-\theta)} = -nI_1(\theta)$$
- Binomial :
$$\mathrm{E}_\theta(\ell_2''(\theta)) = -\frac{n}{\theta(1-\theta)}$$

Thus we interpret
$$I_2(\theta) = \frac{1}{n}\frac{n}{\theta(1-\theta)} = \frac{1}{\theta(1-\theta)} \ .$$
This make the limiting normal distribution with variance 1 over Fisher's information correct.

End of Example

The problem in getting the *correct* Fisher's information, in the sense of the correct variance for the Normal approximation, is that for the Binomial sampling model the sample size n is *hidden* in the pmf of X. For all other iid sampling models the marginal pmf or pdf does not include the sample size n.

The actual definition of Fisher's information may also be written as
$$nI(\theta) = -\mathrm{E}_\theta(\ell_1''(\theta))$$
This calculation will be correct for all statistical models of the type in our course.

Following the discussion above another useful property is used in many statistical packages. In the regular case the MLE is calculated by solving

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0$$

This may be done for example by the Newton-Raphson iterative method

$$\hat{\theta}_{(k+1)} = \hat{\theta}_{(k)} - \left(\frac{\partial^2 \ell(\hat{\theta}_{(k)})}{\partial \theta^2} \right)^{-1} \frac{\partial \ell(\hat{\theta}_{(k)})}{\partial \theta}$$

where $\hat{\theta}_{(k)}$ is the k-th term in the iterative solution approximation for $\hat{\theta}$.

The matrix of second order partial derivatives

$$H(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta^2}$$

is called the Hessian matrix. It may be output by such a package, specifically the matrix $H(\hat{\theta}_{k_{END}})$, that is evaluated at the end point of the iteration. Then, based on some additional property of convergence for random variables, related to the Law of Large Numbers,

$$-\frac{1}{n} H(\hat{\theta}_{k_{END}}) \approx I(\hat{\theta}).$$

Thus the iterative method, when it outputs the Hessian, also yields an approximation to Fisher's information. This is particularly useful when the expectation calculation is not easily obtained.

Another useful method to approximate Fisher information is to use the so called observed Fisher's information, that is replace θ by $\hat{\theta}$ in the formula for Fisher's information. The observed Fisher's information, or observed information, is given by $\mathrm{I}(\hat{\theta})$.