Stat 5102 Notes: Fisher Information and Confidence Intervals Using Maximum Likelihood

Charles J. Geyer

March 2, 2007

1 Web Page

This handout accompanies the web page

`http://www.stat.umn.edu/geyer/5102/examp/rlike.html`

which has computer examples of confidence intervals using maximum likelihood.

2 One Parameter

2.1 Asymptotics of Maximum Likelihood

If X_1, X_2, $\ldots$ are i. i. d. with density f_θ, then the log likelihood is

$$l_n(\theta) = \sum_{i=1}^n \log f_\theta(X_i)$$

(or perhaps the same thing except with additive terms not containing θ dropped). In "nice" situations, the maximum likelihood estimate (MLE) will be a stationary point, a zero of the first derivative

$$l'_n(\theta) = \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f_\theta(X_i) \tag{2.1}$$

also called the *score*. Let θ_0 denote the true unknown parameter value and expand the score in a Taylor series

$$l_n'(\theta) = l_n'(\theta_0) + l_n''(\theta_0)(\theta - \theta_0) + \tfrac{1}{2} l_n'''(t)(\theta - \theta_0)^2 \tag{2.2}$$

where t is some point between θ and θ_0. Plugging in the MLE $\hat{\theta}_n$ for θ in (2.2) and letting $\tilde{\theta}_n$ denote the t in (2.2) we get

$$0 = l_n'(\theta_0) + l_n''(\theta_0)(\hat{\theta}_n - \theta_0) + \tfrac{1}{2} l_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \tag{2.3}$$

We know the asymptotics of $l_n'(\theta_0)$ and $l_n''(\theta_0)$. They are given by the central limit theorem (CLT) and the law of large numbers (LLN). From (2.1) we see that the score is the sum of i. i. d. terms, further more we know from the differentiation under the integral sign identities

$$E_\theta\{l_n'(\theta)\} = 0 \tag{2.4a}$$

$$\operatorname{Var}_\theta\{l_n'(\theta)\} = -E\{l_n''(\theta)\} = I_n(\theta) \tag{2.4b}$$

that each term in (2.1) has mean zero and variance $I_1(\theta)$. Thus the CLT says

$$\frac{1}{\sqrt{n}} l'(\theta_0) \xrightarrow{\mathcal{D}} \operatorname{Normal}\bigl(0, I_1(\theta_0)\bigr) \tag{2.5a}$$

Similarly, the terms in

$$l_n''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)$$

are i. i. d. with mean $-I_1(\theta_0)$. Hence the LLN says

$$\frac{1}{n} l_n''(\theta) \xrightarrow{P} -I_1(\theta_0). \tag{2.5b}$$

These two equations tell us most of what we want to know about (2.3). From (2.5a) we see that we want to multiply (2.3) through by $1 \sqrt{n}$ to get asymptotics. That gives

$$0 = \frac{1}{\sqrt{n}} l_n'(\theta_0) + \frac{1}{n} l_n''(\theta_0) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{1}{2\sqrt{n}} l_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \tag{2.6}$$

We will assume that the remainder term in (2.6) is negligible, that is,

$$\frac{1}{2\sqrt{n}} l_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2 \xrightarrow{P} 0 \tag{2.7}$$

Then if we let Z denote a random variable having the distribution on the right-hand side of (2.5a) by two applications of Slutsky's theorem, we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{\frac{1}{\sqrt{n}} l_n'(\theta_0) + \frac{1}{2\sqrt{n}} l_n'''(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)^2}{-\frac{1}{n} l_n''(\theta_0)} \xrightarrow{\mathcal{D}} \frac{Z}{I_1(\theta_0)}$$

Now Z is normal with mean zero and variance $I_1(\theta_0)$. So $Z/I_1(\theta_0)$ is also normal with mean zero and variance $1/I_1(\theta_0)$. In short

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathcal{D}} \text{Normal}\bigl(0, I_1(\theta_0)^{-1}\bigr) \tag{2.8}$$

Equation is the key to likelihood inference. It gives the asymptotic distribution of the MLE, which we can use to make confidence intervals (more on this below).

The only requirements for (2.8) to hold are the differentiation under the integral sign identities (2.4a) and (2.4b) and assumption (2.7). These assumptions do hold in many, but not all, situations.

The assumption (2.7) is not unreasonable. By another use of the LLN

$$\frac{1}{n} l_n'''(\theta_0) \xrightarrow{P} E\{l_1'''(\theta_0)\}$$

So if the desired conclusion (2.8) is true, then we have, by Slutsky's theorem, (2.7) with $\tilde{\theta}_n$ replaced by θ_0. But $\tilde{\theta}_n$ is converging to θ_0 because it is between θ_0 and $\hat{\theta}_n$, which is converging to zero. So that is very close to proving what we want. However this "proof" is circular, because it assumes what we are trying to prove (2.8). Rather than a proof, it is a demonstration of plausibility, which is all we can do at this level. Investigation of exact conditions under which (2.8) holds is a topic for a more advanced theoretical statistics course.

We will just assume (2.8) holds, which it does in many models of practical interest. Rewriting (2.8) in "sloppy" fashion, it becomes

$$\hat{\theta}_n \approx \text{Normal}\bigl(0, I_n(\theta_0)^{-1}\bigr) \tag{2.9}$$

Recall that $I_n(\theta) = nI_1(\theta)$ so $I_n(\theta)^{-1} = 1/nI_1(\theta)$ and we have statistical precision varying as constant over $\sqrt{n}$ as we have become accustomed to seeing.

Either characterization (2.8) or (2.9) of the asymptotic distribution of the MLE is remarkable. We may have no closed-form expression for the MLE. We may only be able to calculate the MLE by letting a computer maximize the log likelihood. Nevertheless, we know the asymptotic distribution of $\hat{\theta}_n$ even though we have no formula for the MLE itself!

2.2 Observed and Expected Fisher Information

Equations (7.8.9) and (7.8.10) in DeGroot and Schervish give two ways to calculate the Fisher information in a sample of size n. DeGroot and Schervish don't mention this but the concept they denote by $I_n(\theta)$ here is only one kind of Fisher information. To distinguish it from the other kind, $I_n(\theta)$ is called *expected Fisher information.*

The other kind

$$J_n(\theta) = -l_n''(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i) \tag{2.10}$$

is called *observed Fisher information.* Note that the right hand side of our (2.10) is just the same as the right hand side of (7.8.10) in DeGroot and Schervish, except there is no expectation.

It is not always possible to calculate expected Fisher information. Sometimes you can't do the expectations in (7.8.9) and (7.8.10) in DeGroot and Schervish. But if you can evaluate the log likelihood, then you can calculate observed Fisher information. Even if you can't do the derivatives, you can approximate them by finite differences. From the definition of limit,

$$l_n'(\theta) \approx \frac{l_n(\theta + \epsilon) - l_n(\theta)}{\epsilon}$$

for small ϵ. Applying the same idea again gives approximate second derivatives

$$\begin{aligned} J_n(\theta) &\approx - \frac{l_n'(\theta + \epsilon) - l_n'(\theta)}{\epsilon} \\ &\approx - \frac{l_n(\theta + \epsilon) - 2 l_n(\theta) + l_n(\theta - \epsilon)}{\epsilon^2} \end{aligned} \tag{2.11}$$

Since the last approximation has no actual derivatives, it can be calculated whenever the log likelihood can be calculated. The formula is a bit messy for hand calculation. It's better to use calculus when possible. But this finite difference approximation is well suited for computers. Many computer statistical packages don't know any calculus but can do finite differences just fine.

The relation between observed and expected Fisher information is what should now be a familiar theme: consistent estimation.

If we write out what observed Fisher information actually is, we get

$$J_n(\theta) = - \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f(X_i | \theta) \tag{2.12}$$

Since X_1, X_2, ... are assumed to be independent and identically distributed, the terms of the sum on the right hand side of (2.12) are also independent and identically distributed and the law of large numbers says their average (not sum, we need to divide by n to get the average) converges to the expectation of one term, which by definition is $-I(\theta)$, that is,

$$\frac{1}{n} J_n(\theta) \xrightarrow{P} I(\theta) \tag{2.13}$$

or in sloppy notation

$$J_n(\theta) \approx n I(\theta) = I_n(\theta)$$

So we can (for large sample sizes) use $J_n(\theta)$ and $I_n(\theta)$ interchangeably.

2.3 Plug-In

Actually, we need another use of plug-in. We don't know θ (otherwise we wouldn't be trying to estimate it). Hence we don't know either $J_n(\theta)$ or $I_n(\theta)$. We know the functions J_n and I_n but we don't know the true value of the parameter θ where we should evaluate them. However (an old theme again) we do have a consistent estimator

$$\hat{\theta}_n \xrightarrow{P} \theta$$

which implies by the continuous mapping theorem (Slutsky for a single sequence) under the additional assumption that I_n is a continuous function

$$I(\hat{\theta}_n) \xrightarrow{P} I(\theta)$$

or

$$\frac{1}{n} I_n(\hat{\theta}_n) \xrightarrow{P} I(\theta). \tag{2.14a}$$

The analogous equation for observed Fisher information

$$\frac{1}{n} J_n(\hat{\theta}_n) \xrightarrow{P} I(\theta). \tag{2.14b}$$

doesn't quite follow from Slutsky and continuity of J_n; it really requires that (2.13) be replaced by a so-called uniform law of large numbers (which is way beyond the scope of this course). However, in "nice" problems both (2.14a) and (2.14b) are true, and so both can be used in the plug-in theorem to estimate asymptotic variance of the maximum likelihood estimator.

Using "sloppy" notation, either of the following approximations can be used to construct confidence intervals based on maximum likelihood estimators

$$\hat{\theta}_n \approx \text{Normal}\bigl(\theta, I_n(\hat{\theta}_n)^{-1}\bigr) \tag{2.15a}$$

The analogous equation for observed Fisher information

$$\hat{\theta}_n \approx \text{Normal}\bigl(\theta, J_n(\hat{\theta}_n)^{-1}\bigr) \tag{2.15b}$$

2.4 Confidence Intervals

The corresponding confidence intervals are

$$\hat{\theta}_n \pm c I_n(\hat{\theta}_n)^{-1/2} \tag{2.16a}$$

where c is the appropriate z critical value (for example, 1.96 for 95% confidence or 1.645 for 90% confidence). The analogous equation for observed Fisher information

$$\hat{\theta}_n \pm c J_n(\hat{\theta}_n)^{-1/2} \tag{2.16b}$$

Example 2.1 (Binomial, Again)**.**
We redo the binomial distribution. The log likelihood is

$$l_n(p) = x \log(p) + (n - x) \log(1 - p)$$

and two derivatives are

$$l_n'(p) = \frac{x}{p} - \frac{n - x}{1 - p}$$

and

$$l_n''(p) = -\frac{x}{p^2} - \frac{n - x}{(1 - p)^2} \tag{2.17}$$

We know from previous work with maximum likelihood that the MLE is $\hat{p}_n = x/n$. Plugging in $\hat{p}_n$ for p and writing $x = n\hat{p}_n$ in (2.17) and attaching a minus sign gives the observed Fisher information

$$\begin{aligned}
J_n(\hat{p}_n) &= \frac{x}{\hat{p}_n^2} + \frac{n - x}{(1 - \hat{p}_n)^2} \\
&= \frac{n\hat{p}_n}{\hat{p}_n^2} + \frac{n - n\hat{p}_n}{(1 - \hat{p}_n)^2} \\
&= n \left(\frac{1}{\hat{p}_n} + \frac{1}{1 - \hat{p}_n} \right) \\
&= \frac{n}{\hat{p}_n(1 - \hat{p}_n)}
\end{aligned}$$

The expected Fisher information calculation is very similar. Taking minus the expectation of (2.17) using $E(X) = np$ gives

$$I_n(p) = \frac{n}{p(1-p)}$$

and plugging in the consistent estimator $\hat{p}_n$ of p gives

$$I_n(\hat{p}_n) = \frac{n}{\hat{p}_n(1-\hat{p}_n)}$$

So in this problem $J_n(\hat{p}_n) = I_n(\hat{p}_n)$. Sometimes this happens, sometimes observed and expected information are different.

In this problem either of the confidence intervals (2.16a) or (2.16b) turns out to be

$$\hat{p}_n \pm c\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}$$

where c is the "z" critical value, which is the usual "plug-in" confidence interval taught in elementary statistics courses.

In our binomial example, we didn't really need the asymptotics of maximum likelihood to construct the confidence interval, since more elementary theory arrives at the same interval. But in complicated situations where there is no simple analytic expression for the MLE, there is no other way to get the asymptotic distribution except using Fisher information. An example is given on the web page given in Section 1.

3 Multiple Parameters

3.1 Observed and Expected Fisher Information Matrices

The story for maximum likelihood for multiple parameters is almost the same. If the parameter is a vector $\boldsymbol{\theta}$, then instead of one first derivative we have a vector of first partial derivatives, sometimes called

$$\nabla l_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_2} \\ \vdots \\ \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_d} \end{pmatrix} \tag{3.1}$$

the *gradient* vector, and instead of one second derivative we have a matrix of second partial derivatives

$$\nabla^2 l_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_d} \\ \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_1} & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_d \partial \theta_2} & \cdots & \frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_d^2} \end{pmatrix} \tag{3.2}$$

As in the one-parameter case, we have identities derived by differentiation under the integral sign. The multiparameter analog of (2.4a) is

$$E_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = 0, \tag{3.3a}$$

a vector equation,[1] which, if you prefer, can be written instead as d scalar equations

$$E_{\boldsymbol{\theta}}\left\{\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i}\right\} = 0, \qquad i = 1, \ldots, d. \tag{3.3b}$$

And the multiparameter analog of the equivalence of (7.8.9) and (7.8.10) in DeGroot and Schervish is

$$\operatorname{Var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\}, \tag{3.4a}$$

a matrix equation,[2] which, if you prefer, can be written instead as d^2 scalar equations

$$E_{\boldsymbol{\theta}}\left\{\frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial l_n(\boldsymbol{\theta})}{\partial \theta_j}\right\} = -E_{\boldsymbol{\theta}}\left\{\frac{\partial^2 l_n(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right\}, \qquad i, j = 1, \ldots, d. \tag{3.4b}$$

We generally prefer the vector and matrix equations (3.3a) and (3.4a) because they are much simpler to read and write, although we have to admit that this concise notation hides a lot of details.

As in the one-parameter case, expected Fisher information is defined as either side of (3.4a)

$$\begin{aligned} \mathbf{I}_n(\boldsymbol{\theta}) &= \operatorname{Var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} \\ &= -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\} \end{aligned} \tag{3.5}$$

[1] Recall that the mean of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ is a vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ having components that are the expectations of the components of the random vector, that is, when we write $\boldsymbol{\mu} = E(\mathbf{Y})$ we mean the same thing as $\mu_i = E(Y_i)$, $i = 1, \ldots, d$.

[2] Recall that the mean of a random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)$ is a matrix $\mathbf{M}$ with components m_{ij} that are the covariances of the components of the random vector, that is, when we write $\mathbf{M} = \operatorname{Var}(\mathbf{Y})$ we mean the same thing as $m_{ij} = \operatorname{Cov}(Y_i, Y_j)$, $i, j = 1, \ldots, d$.

The difference between the one-parameter and many-parameter cases is that in the first the Fisher information is a scalar and in the second it is a *matrix.*

Similarly, we define the observed Fisher information matrix to be the quantity we are taking the expectation of on the right hand side of (3.4a)

$$\mathbf{J}_n(\boldsymbol{\theta}) = -\nabla^2 l_n(\boldsymbol{\theta}) \tag{3.6}$$

We also have the multiparameter analog of equation (7.8.12) in DeGroot and Schervish

$$\mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta})$$

and we often write $\mathbf{I}(\boldsymbol{\theta})$ with no subscript instead of $\mathbf{I}_1(\boldsymbol{\theta})$.

3.2 Plug-In

The multiparameter analogs of (2.14a) and (2.14b)

$$\frac{1}{n}\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}) \tag{3.7a}$$

and

$$\frac{1}{n}\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}), \tag{3.7b}$$

where $\hat{\boldsymbol{\theta}}_n$ is the MLE, hold in "nice" situations (and as in the one-parameter situation we will be vague about exactly what features of a statistical model make it "nice" for maximum likelihood theory).

This allows us to use the natural plug-in estimators $\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)$ and $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$ which we can calculate in place of $\mathbf{I}_n(\boldsymbol{\theta})$ and $\mathbf{J}_n(\boldsymbol{\theta})$ which we can't calculate because we don't know the true value of the parameter $\boldsymbol{\theta}$.

3.3 Multivariate Convergence in Distribution

We didn't actually say what the convergence in probability in equations (3.7a) and (3.7b) means, but it is trivial. For any sequence of random vectors $\mathbf{Y}_1$ $\mathbf{Y}_2$, ... and any constant vector $\mathbf{a}$, the statement

$$\mathbf{Y}_n \xrightarrow{P} \mathbf{a}$$

contains no more and no less mathematical content than the d convergence in probability statements

$$Y_{ni} \xrightarrow{P} a_i, \qquad i = 1, \ldots, d,$$

where $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nd})$ and $\mathbf{a} = (a_1, \ldots, a_d)$.

The situation with convergence in distribution is quite different. The statement

$$\mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{Y}, \tag{3.8}$$

where $\mathbf{Y}$ is now a random vector, contains much more mathematical content than the d convergence in distribution statements

$$Y_{ni} \xrightarrow{\mathcal{D}} Y_i, \qquad i = 1, \ldots, d. \tag{3.9}$$

When we need to make the distinction, we refer to (3.8) as *joint* convergence in distribution and to (3.9) as *marginal* convergence in distribution. The vector statement (3.8) can actually be defined in terms of scalar statements, but not just d such statements. The joint convergence in distribution statement (3.8) holds if and only if

$$\mathbf{t}'\mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{Y}, \qquad \mathbf{t} \in \mathbb{R}^d.$$

What this means is that we must check an infinite set of convergence in distribution statements: for every constant random vector $\mathbf{t}$ we must have the scalar convergence in distribution $\mathbf{t}'\mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{Y}$.

However, we don't actually check an infinite set of statements (that would be tough). We usually just use the central limit theorem. And the univariate CLT quite trivially implies the multivariate CLT.

Theorem 3.1 (Multivariate Central Limit Theorem)**.** *If $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ is a sequence of independent, identically distributed random vectors having mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$ and*

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

is the sample mean for sample size n, then

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \mathbf{M}). \tag{3.10}$$

The trivial proof goes as follows, Let $\mathbf{Y}$ be a random vector having the distribution on the right-hand side of (3.10) so (3.10) can be rewritten

$$\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathbf{Y}. \tag{3.11}$$

Then for any constant vector $\mathbf{t}$, the scalar random variables $\mathbf{t}'\mathbf{X}_i$ are independent and identically distributed with mean $\mathbf{t}'\boldsymbol{\mu}$ and variance $\mathbf{t}'\mathbf{M}\mathbf{t}$. Hence the univariate CLT says

$$\sqrt{n}\left(\mathbf{t}'\overline{\mathbf{X}}_n - \mathbf{t}'\boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \mathbf{t}'\mathbf{M}\mathbf{t}),$$

which can be rewritten

$$\mathbf{t}'\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \text{Normal}(0, \mathbf{t}'\mathbf{M}\mathbf{t}), \tag{3.12}$$

But the distribution of $\mathbf{t}'\mathbf{Y}$ is $\text{Normal}(0, \mathbf{t}'\mathbf{M}\mathbf{t})$, so (3.12) can be rewritten

$$\mathbf{t}'\sqrt{n}\left(\overline{\mathbf{X}}_n - \boldsymbol{\mu}\right) \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{Y}$$

since this is true for arbitrary vectors $\mathbf{t}$ this means (3.11) holds.

3.4 Asymptotics of Maximum Likelihood

With multivariate convergence theory in hand, we can now explain the asymptotics of multiparameter maximum likelihood. Actually it look just like the uniparameter case. You just have to turn scalar quantities into vectors or matrices as appropriate.

In "nice" situations (again being vague about what "nice" means) the multiparameter analogs of (2.15a) and (2.15b) are

$$\hat{\boldsymbol{\theta}}_n \approx \text{Normal}\big(\boldsymbol{\theta}, \mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big) \tag{3.13a}$$

and

$$\hat{\boldsymbol{\theta}}_n \approx \text{Normal}\big(\boldsymbol{\theta}, \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big) \tag{3.13b}$$

The only difference between these equations and the earlier ones being some boldface type. Here the MLE $\hat{\boldsymbol{\theta}}_n$ is a vector because the parameter it estimates is a vector, and the Fisher information matrix, either $\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)$ or $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$ as the case may be is (as the terminology says) a matrix, this means the inverse operation denoted by the superscript -1 is a matrix inverse. Matrix inversion is hard when done by hand, but we will generally let a computer do it, do it's really not a big deal.

3.5 Confidence Intervals

Confidence intervals analogous to (2.16a) and (2.16b) are a bit tricky. When the parameter is a vector it doesn't fit in an "interval" which is a one-dimensional object. So there are two approaches.

- We can generalize our notion of *confidence interval* to multidimensional random sets, which are called *confidence sets* or *confidence regions.* Some theory courses cover this generalization, but I have never seen it actually applied in an actual application.

- What users actually do in multiparameter situations is to focus on confidence intervals for single parameter or for scalar functions of parameters.

So we will concentrate on *linear* scalar functions of the parameters, of the form $\mathbf{t}'\boldsymbol{\theta}$, which, of course, we estimate by $\mathbf{t}'\hat{\boldsymbol{\theta}}_n$. A special case of this is when $\mathbf{t}$ has all components zero except for $t_j = 1$. Then $\mathbf{t}'\boldsymbol{\theta}$ is just complicated notation for the j-th component θ_j, and, similarly, $\mathbf{t}'\hat{\boldsymbol{\theta}}_n$ is just complicated notation for the j-th component $\hat{\theta}_{nj}$.

So with that said, the confidence intervals for $\mathbf{t}'\theta$ analogous to (2.16a) and (2.16b) are

$$\mathbf{t}'\hat{\boldsymbol{\theta}}_n \pm c\sqrt{\mathbf{t}'\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{t}} \tag{3.14a}$$

where c is the appropriate z critical value (for example, 1.96 for 95% confidence or 1.645 for 90% confidence) and

$$\mathbf{t}'\hat{\boldsymbol{\theta}}_n \pm c\sqrt{\mathbf{t}'\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\mathbf{t}} \tag{3.14b}$$

When we specialize to $\mathbf{t}$ with only one nonzero component $t_j = 1$ we get

$$\hat{\theta}_{nj} \pm c\sqrt{\Bigl(\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\Bigr)_{jj}} \tag{3.15}$$

and the similar interval with $\mathbf{J}_n$ replacing $\mathbf{I}_n$.

It may not be obvious what the notation in (3.15) for the asymptotic variance $\Bigl(\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\Bigr)_{jj}$ means, so we explain it in words. First you invert the Fisher information matrix, and then you take the jj component of the inverse Fisher information matrix. This can be very different from taking the jj component of the Fisher information matrix, which is a scalar, and inverting that.