Statistics 3858 : Maximum Likelihood Estimators - Regular Case Large Sample Theory

1 Regular Case

The regular maximum likelihood case is essentially the case where calculus methods apply in order to calculate the MLE. These are written for the case of iid sampling, that is for iid random variables. The statistical model has finite dimensional parameter space Θ and each distribution in the model has pdf (or pmf) $f(\cdot, \theta)$. The function (with argument x)

$$\frac{\partial \log(f(x;\theta))}{\partial \theta}$$

is called the *score function.*

Smoothness Condition : Regularity Condition 1 Assume for any other $\theta \in \Theta$ (except on the boundary) assume that f satisfies

- in the case of a pdf, in terms of integrals this is an interchange of integration and differentiation

$$\int_R \left(\frac{\partial f(x;\theta)}{\partial \theta} \right) dx = \frac{\partial}{\partial \theta} \int_R f(x;\theta) dx$$

- in the pmf case, interpreting $P(X = x; \theta) = f(x; \theta)$

$$\sum_x \left(\frac{\partial f(x;\theta)}{\partial \theta} \right) = \frac{\partial}{\partial \theta} \sum_x f(x;\theta)$$

End of Assumption 1

This assumption is the smoothness condition alluded to in Theorem 8.5 A in Rice.

In the continuous and discrete case the this assumption can be rewritten in terms of interchanging differentiation and integration, or in terms of interchanging differentiation and summation. This assumption tells us a very useful property of the score function, in particular its expectation at θ_0 the *true value of the parameter*, that is

$$\mathrm{E}_{\theta_0} \left(\frac{\partial \log(f(X;\theta))}{\partial \theta} \right) \Bigg|_{\theta=\theta_0} = \mathrm{E}_{\theta_0} \left(\frac{\partial \log(f(X;\theta_0))}{\partial \theta} \right) = 0 \tag{1}$$

- in the case of a pdf, in terms of integrals this is an interchange of integration and differentiation

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

Evaluate this derivative at $\theta = \theta_0$

$$\begin{aligned} \int_R \frac{\partial f(x;\theta)}{\partial\theta}\frac{1}{f(x;\theta)} f(x;\theta_0)dx\Big|_{\theta=\theta_0} &= \int_R \frac{\partial f(x;\theta_0)}{\partial\theta}\frac{1}{f(x;\theta_0)} f(x;\theta_0)dx \\ &= \int_R \frac{\partial f(x;\theta_0)}{\partial\theta}dx \\ &= \frac{\partial}{\partial\theta}\int_R f(x;\theta)dx\Big|_{\theta=\theta_0} \\ &= \frac{\partial 1}{\partial\theta} = 0 \end{aligned}$$

Where are we using Assumption 1?

- in the pmf case, interpreting $P(X = x;\theta) = f(x;\theta)$

$$\begin{aligned} \mathrm{E}_{\theta_0}\left(\frac{\partial \log(f(X;\theta))}{\partial\theta}\right) &= \sum_x \left(\frac{\partial \log(f(x;\theta))}{\partial\theta}\right) f(x;\theta_0) \\ &= \sum_x \frac{\partial f(x;\theta)}{\partial\theta}\frac{1}{f(x;\theta)} f(x;\theta_0) \end{aligned}$$

Evaluate this derivative at $\theta = \theta_0$

$$\begin{aligned} \sum_x \frac{\partial f(x;\theta)}{\partial\theta}\frac{1}{f(x;\theta)} f(x;\theta_0)\Big|_{\theta=\theta_0} &= \sum_x \frac{\partial f(x;\theta_0)}{\partial\theta}\frac{1}{f(x;\theta_0)} f(x;\theta_0) \\ &= \sum_x \frac{\partial f(x;\theta_0)}{\partial\theta} \\ &= \frac{\partial}{\partial\theta}\sum_x f(x;\theta)\Big|_{\theta=\theta_0} \\ &= \frac{\partial 1}{\partial\theta} = 0 \end{aligned}$$

Example

Consider the exponential distribution with parameter λ. Then for $x > 0$

$$\log\left(f(x;\lambda) = \log(\lambda) - \lambda x\right.$$

and hence

$$\frac{\partial \log\left(f(x;\lambda)\right.}{\partial\lambda} = \frac{1}{\lambda} - x\ .$$

Then

$$\mathrm{E}_{\lambda_0}\left(\frac{\partial \log\left(f(X;\lambda)\right.}{\partial\lambda}\right) = \mathrm{E}_{\lambda_0}\left(\frac{1}{\lambda} - X\right) = \frac{1}{\lambda} - \frac{1}{\lambda_0}\ .$$

Notice this expectation is 0 if and only if $\lambda = \lambda_0$, that is if λ happens to be the true parameter.

End of Example

Notice that

$$\mathrm{E}_{\theta_0}\left(\frac{\partial \log(f(X;\theta))}{\partial \theta}\right) = 0$$

if and only if $\theta = \theta_0$. Thus for every θ_0, except possibly on the boundary of Θ, we have

$$\mathrm{E}_{\theta_0}\left(\frac{\partial \log(f(X;\theta_0))}{\partial \theta}\right) = 0 \ .$$

Since this holds for every θ_0 we can simply change notation in this last line and notice that for every θ,

$$\mathrm{E}_{\theta}\left(\frac{\partial \log(f(X;\theta))}{\partial \theta}\right) = 0 \ .$$

Smoothness Condition : Regularity Condition 2 For any $\theta \in \Theta$ (except on the boundary) assume that f satisfies Assumption 1 and

- in the case of a pdf, in terms of integrals this is an interchange of integration and differentiation

$$\frac{\partial}{\partial \theta}\int_R \left(\frac{\partial \log(f(x;\theta))}{\partial \theta} f(x;\theta)\right) dx = \int_R \frac{\partial}{\partial \theta}\left\{\left(\frac{\partial \log(f(x;\theta))}{\partial \theta} f(x;\theta)\right)\right\} dx$$

- in the pmf case, interpreting $P(X = x;\theta) = f(x;\theta)$

$$\frac{\partial}{\partial \theta}\sum_x \left(\frac{\partial \log(f(x;\theta))}{\partial \theta} f(x;\theta)\right) = \sum_x \frac{\partial}{\partial \theta}\left(\frac{\partial \log(f(x;\theta))}{\partial \theta} f(x;\theta)\right)$$

The left hand side of the regularity condition 2 in other notation is

$$\frac{\partial}{\partial \theta}\mathrm{E}_{\theta}\left(\frac{\partial \log(f(X;\theta))}{\partial \theta}\right)$$

and the condition says the integration or summation sign can be interchanged with the differentiation operation.

Theorem 1 (Theorem 8.5 A in Rice) *Under Assumption 1 and some additional technical assumptions (not specified in this course so the variance below is finite) the mle from an iid sample is consistent, that is the MLE $\hat{\theta}_n$ converges in probability to the population parameter value θ_0.*

Proof (Sketch of the main idea)

Let $\ell(\theta)$ be the log likelihood. Therefore

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(f(X_i;\theta))$$

Therefore

$$\frac{1}{n}\ell(\theta) \to \mathrm{E}_{\theta_0}\left(\log(f(X;\theta))\right) \equiv \Lambda(\theta)$$

in probability as $n \to \infty$. Call this limit function, with argument θ, $\Lambda(\theta)$. This limit function Λ has the property that

$$\frac{\partial \Lambda(\theta_0)}{\partial \theta} = \frac{\partial \mathrm{E}_{\theta_0}\left(\log(f(X;\theta))\right)}{\partial \theta} = \mathrm{E}_{\theta_0}\left(\frac{\partial \log(f(X;\theta_0))}{\partial \theta}\right) = 0$$

(Here we use Assumption 1). It can be concluded that this limit function has a maximum at θ_0 (the other unspecified assumptions are for this part, and this would require a careful mathematical analysis beyond our course) and that

$$\hat{\theta}_n = \text{argmax } \ell(\theta)$$

converges in probability to

$$\theta_0 = \text{argmax } \Lambda(\theta) \ .$$

End of proof for Theorem 1

See the R script file *Thm-8-5-2A-Rice.txt* for some simulation plots of $\frac{1}{n}\ell(\theta)$ and $\Lambda(\theta)$ in the case of the exponential θ family of distributions. If $X_i, i = 1, \ldots, n$ are iid exponential, θ then

$$\frac{1}{n}\ell(\theta) = \frac{1}{n}\sum_{i=1}^{n}\{\log(\theta) - \theta X_i\} = \log(\theta) - \theta \bar{X}_n \ .$$

Notice θ is the argument of this function. There is one specific value of θ that corresponds to the true value of the parameter. If θ_0 is the true value of the parameter then by the Law of Large Numbers, applied to the r.v.s $\log(\theta) - \theta X_i$

$$\frac{1}{n}\ell(\theta) \to \log(\theta) - \frac{\theta}{\theta_0} \equiv \Lambda(\theta)$$

in probability as $n \to \infty$.

We now use Assumption 2 to obtain a formula for the variance of the score r.v., that is

$$\mathrm{Var}_{\theta_0}\left(\frac{\partial \log(f(X;\theta_0))}{\partial\theta}\right)$$

Recall that from Assumption 1 the expected value of the score r.v at the true θ_0 is 0; see equation (1). Thus we need to evaluate

$$\mathrm{E}_{\theta_0}\left\{\left(\frac{\partial \log(f(X;\theta_0))}{\partial\theta}\right)^2\right\} .$$

Since the argument θ_0 in the partial derivative is the same as the parameter value for the distribution for the expectation we can drop the subscript 0 notation and work with

$$\mathrm{Var}_{\theta_0}\left(\frac{\partial \log(f(X;\theta_0))}{\partial\theta}\right) = \mathrm{E}_{\theta}\left\{\left(\frac{\partial \log(f(X;\theta))}{\partial\theta}\right)^2\right\} .$$

Theorem 2 (Lemma 8.5 A Rice) *Let*

$$I(\theta) = E_{\theta}\left\{\left(\frac{\partial \log(f(X;\theta))}{\partial\theta}\right)^2\right\}$$

Under Assumptions 1 and 2 (smoothness assumptions on f wrt θ) then

$$I(\theta) = -E_{\theta}\left\{\frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2}\right\}$$

Remark

Lemma A is useful because it is often easier to calculate $I(\theta)$ using the second expectation rather than calculating $I(\theta)$ = the variance of the score r.v. directly. Sometimes the direct calculation is not very easy but the second method is surprisingly easy, for example the Gamma family.

End of Remark

Proof of Theorem 2 [do this in terms of the pdf for convenience]

Recall

$$\frac{\partial}{\partial\theta}\int_R f(x;\theta)dx = \frac{\partial 1}{\partial\theta} = 0$$

Under the Assumptions, differentiating once more wrt θ we also have

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\left\{\frac{\partial}{\partial\theta}\int_R f(x;\theta)dx\right\} \\
&= \frac{\partial}{\partial\theta}\int_R \left\{\frac{\partial f(x;\theta)}{\partial\theta}\right\}dx \\
&= \frac{\partial}{\partial\theta}\int_R \left\{\frac{1}{f(x;\theta)}\frac{\partial f(x;\theta)}{\partial\theta}f(x;\theta)\right\}dx \\
&= \frac{\partial}{\partial\theta}\int_R \left\{\frac{\partial \log(f(x;\theta))}{\partial\theta}f(x;\theta)\right\}dx \\
&= \int_R \frac{\partial}{\partial\theta}\left\{\frac{\partial \log(f(x;\theta))}{\partial\theta}f(x;\theta)\right\}dx \\
&= \int_R \left\{\frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2}f(x;\theta) + \frac{\partial \log(f(x;\theta))}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta}\right\}dx
\end{aligned}$$

$$
\begin{aligned}
&= \int_R \left\{ \frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2} f(x;\theta) + \frac{\partial \log(f(x;\theta))}{\partial\theta} \frac{1}{f(x;\theta)} \frac{\partial f(x;\theta)}{\partial\theta} f(x;\theta) \right\} dx \\
&= \int_R \left\{ \frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2} f(x;\theta) + \left(\frac{\partial \log(f(x;\theta))}{\partial\theta} \right)^2 f(x;\theta) \right\} dx \\
&= \int_R \left\{ \frac{\partial^2 \log(f(x;\theta))}{\partial\theta^2} f(x;\theta) \right\} dx + I(\theta) \\
&= \mathrm{E}_\theta \left\{ \frac{\partial^2 \log(f(X;\theta))}{\partial\theta^2} \right\} + I(\theta)
\end{aligned}
$$

Therefore $I(\theta)$ is the negative of the last integral, proving Theorem 2

End of proof for Theorem 2

$I(\theta)$ is called *Fisher's Information* or sometimes simply *information*.

2 When Will the Smoothness Conditions Fail?

This is not easy to answer. However there is an easy case when they will fail. They fail if the support of $f(\cdot;\theta)$ is not the same for every θ in the interior of the parameter space. In our various parametric families in this course, otherwise the Assumptions will hold.

Our prototype example is the Uniform family. Specifically consider the Uniform(0,θ) statistical model for iid sampling, where $\theta \in \Theta = (0, \infty)$. Then

$$f(x;\theta) = \frac{1}{\theta} \mathrm{I}(0 < x < \theta) \ .$$

In particular we will determine if (1) holds or if Assumption 1 holds in this case. The answer will be no for both. First consider

$$\mathrm{E}_{\theta_0} \left(\frac{\partial \log(f(X;\theta_0))}{\partial\theta} \right) \ .$$

Recall

$$\frac{\partial \log(f(x;\theta_0))}{\partial\theta} = \lim_{\Delta\to 0} \frac{\log(f(x;\theta_0+\Delta)) - \log(f(x;\theta_0))}{\Delta} \ .$$

If $0 < x < \theta_0$ then for Δ sufficiently close to 0 we also have $x < \theta_0 + \Delta$. Therefore we are not calculating the log of 0 and hence

$$\frac{\partial \log(f(x;\theta_0))}{\partial\theta} = \lim_{\Delta\to 0} \frac{-\log(\theta_0+\Delta) + \log(\theta_0)}{\Delta} = -\frac{1}{\theta_0} \ .$$

Thus

$$
\begin{aligned}
\mathrm{E}_{\theta_0} \left(\frac{\partial \log(f(X;\theta_0))}{\partial\theta} \right) &= \int_0^{\theta_0} \left\{ -\frac{1}{\theta_0} \times \frac{1}{\theta_0} \right\} dx \\
&= -\frac{1}{\theta_0} \neq 0 \ .
\end{aligned}
$$

Thus (1) does not hold.

To consider Assumption 1 similar we have

$$
\begin{aligned}
\int_0^{\theta_0} \left\{ \frac{\partial f(x;\theta_0)}{\partial\theta} \right\} dx &= \int_0^{\theta_0} \left\{ -\frac{1}{\theta_0^2} \right\} dx \\
&= -\frac{1}{\theta_0} \neq 0
\end{aligned}
$$

while

$$\frac{\partial}{\partial\theta}\int_R f(x;\theta)dx = \frac{\partial 1}{\partial\theta} = 0 \ .$$

Thus Assumption 1 fails to hold.

3 Check that the Assumptions Hold for Parametric Models with Common Support for all θ

Binomial Model

For x integer valued from 0 to n

$$\begin{aligned}\frac{\partial \log(f(x;\theta))}{\partial\theta} &= \frac{\partial(n-x)\log(1-\theta)+x\log(\theta)}{\partial\theta} \\ &= -\frac{(n-x)}{1-\theta}+\frac{x}{\theta}\end{aligned}$$

Then

$$\begin{aligned}\mathrm{E}_\theta\left(\frac{\partial \log(f(X;\theta))}{\partial\theta}\right) &= \mathrm{E}_\theta\left(-\frac{(n-X)}{1-\theta}+\frac{X}{\theta}\right) \\ &= -\frac{(n-n\theta)}{1-\theta}+\frac{n\theta}{\theta} \\ &= -n\frac{(1-\theta)}{1-\theta}+n\frac{\theta}{\theta}=0\end{aligned}$$

Next

$$\begin{aligned}\frac{\partial^2 \log(f(x;\theta)}{\partial\theta^2} &= \frac{\partial\left\{-\frac{(n-x)}{1-\theta}+\frac{x}{\theta}\right\}}{\partial\theta} \\ &= -\frac{(n-x)}{(1-\theta)^2}-\frac{x}{\theta^2}\end{aligned}$$

$$\begin{aligned}\mathrm{E}_\theta\left\{\frac{\partial^2 \log(f(X;\theta)}{\partial\theta^2}\right\} &= \mathrm{E}_\theta\left\{-\frac{(n-X)}{(1-\theta)^2}-\frac{X}{\theta^2}\right\} \\ &= -\frac{(n-n\theta)}{(1-\theta)^2}-\frac{n\theta}{\theta^2} \\ &= -n\left\{\frac{1}{(1-\theta)}+\frac{1}{\theta}\right\} \\ &= -\frac{n}{\theta(1-\theta)}\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var}_\theta\left\{\frac{\partial \log(f(X;\theta)}{\partial\theta}\right\} &= \mathrm{Var}_\theta\left\{-\frac{(n-X)}{1-\theta}+\frac{X}{\theta}\right\} \\ &= \mathrm{Var}_\theta\left\{X\left[\frac{1}{1-\theta}+\frac{1}{\theta}\right]\right\}\end{aligned}$$

$$\begin{aligned} &= \text{Var}_\theta \left\{ X \frac{1}{\theta(1-\theta)} \right\} \\ &= \frac{n}{\theta(1-\theta)} \end{aligned}$$

See the additional file FisheInfoComments.pdf for additional information on Fisher's Information. It resolves the calculation of Fisher's informaiton for the Binomial and Bernoulli experiments.

End of Example

The student should verify these conditions for the exponential model, the normal model, the Poisson model and the geometric model.

4 More Examples

Gamma Model

In this example it is not so easy to show directly that the Assumptions for the smoothness or regularity conditions hold. But is makes certain calculations much easier to obtain.

For $x > 0$, $f(x;\alpha,\lambda) > 0$ and

$$\begin{aligned} \frac{\partial \log(f(x;\alpha,\lambda))}{\partial \alpha} &= \frac{\partial \left\{ \alpha \log(\lambda) + (\alpha-1)\log(x) - \log(\Gamma(\alpha)) - \lambda x \right\}}{\partial \alpha} \\ &= \log(\lambda) + \log(x) - \frac{d \log(\Gamma(\alpha))}{d\alpha} \\ \frac{\partial \log(f(x;\alpha,\lambda))}{\partial \lambda} &= \frac{\partial \left\{ \alpha \log(\lambda) + (\alpha-1)\log(x) - \log(\Gamma(\alpha)) - \lambda x \right\}}{\partial \lambda} \\ &= \frac{\alpha}{\lambda} - x \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \log(f(x;\alpha,\lambda))}{\partial \alpha^2} &= \frac{\partial \left\{ \log(\lambda) + \log(x) - \frac{d \log(\Gamma(\alpha))}{d\alpha} \right\}}{\partial \alpha} \\ &= -\frac{d^2 \log(\Gamma(\alpha))}{d\alpha^2} \\ \frac{\partial^2 \log(f(x;\alpha,\lambda))}{\partial \lambda^2} &= \frac{\partial \left\{ \frac{\alpha}{\lambda} - x \right\}}{\partial \lambda} \\ &= -\frac{\alpha}{\lambda^2} \\ \frac{\partial^2 \log(f(x;\alpha,\lambda))}{\partial \lambda \partial \alpha} &= \frac{\partial \left\{ \frac{\alpha}{\lambda} - x \right\}}{\partial \alpha} \\ &= \frac{1}{\lambda} \end{aligned}$$

In particular the Fisher's information matrix is easy to calculate since all these second order derivatives are constant wrt x and hence the expectations are these constants.

$$-I(\alpha,\lambda) = \begin{pmatrix} \text{E}\left(\frac{\partial^2 \log(f(X;\alpha,\lambda)}{\partial \alpha^2} \right) & \text{E}\left(\frac{\partial^2 \log(f(X;\alpha,\lambda)}{\partial \lambda \partial \alpha} \right) \\ \text{E}\left(\frac{\partial^2 \log(f(X;\alpha,\lambda)}{\partial \lambda \partial \alpha} \right) & \text{E}\left(\frac{\partial^2 \log(f(X;\alpha,\lambda)}{\partial \lambda^2} \right) \end{pmatrix} = \begin{pmatrix} -\frac{d^2 \log(\Gamma(\alpha))}{d\alpha^2} & \frac{1}{\lambda} \\ \frac{1}{\lambda} & -\frac{\alpha}{\lambda^2} \end{pmatrix}$$

Also we have

$$\begin{aligned} 0 &= \mathrm{E}\left(\frac{\partial \log(f(X;\alpha,\lambda)}{\partial \alpha}\right) \\ &= \mathrm{E}\left(\log(\lambda) + \log(X) - \frac{d\log(\Gamma(\alpha))}{d\alpha}\right) \end{aligned}$$

and hence

$$\log(\lambda) - \frac{d\log(\Gamma(\alpha))}{d\alpha} + \mathrm{E}\left(\log(X)\right) = 0$$

giving us the expression for $\mathrm{E}(\log(X))$.

End of Example

Normal Model

$$\log(f(x;\mu,\sigma^2)) = \log(\sqrt{2\pi}) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x-\mu)^2$$

$$\begin{aligned} \frac{\partial \log(f(x;\mu,\sigma^2))}{\partial \mu} &= \frac{1}{\sigma^2}(x-\mu) \\ \frac{\partial \log(f(x;\mu,\sigma^2))}{\partial (\sigma^2)} &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x-\mu)^2 \end{aligned}$$

and

$$\begin{aligned} \frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\ \frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial (\sigma^2)^2} &= \frac{1}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3}(x-\mu)^2 \\ \frac{\partial^2 \log(f(x;\mu,\sigma^2))}{\partial \mu \partial \sigma^2} &= -\frac{1}{(\sigma^2)^2}(x-\mu) \end{aligned}$$

$$-I(\mu,\sigma^2) = \begin{pmatrix} \mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial \mu^2}\right) & \mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial \mu \partial (\sigma^2)}\right) \\ \mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial \mu \partial (\sigma^2)}\right) & \mathrm{E}\left(\frac{\partial^2 \log(f(X;\mu,\sigma^2))}{\partial (\sigma^2)^2}\right) \end{pmatrix} = \begin{pmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{2\sigma^4} \end{pmatrix}$$

The student should verify that

$$\mathrm{E}\left(\frac{\partial \log(f(X;\mu,\sigma^2))}{\partial \mu}\right) = 0$$

and

$$\mathrm{E}\left(\frac{\partial \log(f(X;\mu,\sigma^2))}{\partial \sigma^2}\right) = 0$$

End of Example

5 Limit Normal Distribution for MLE in the Regular Case

The result is given in Rice Theorem 8.5B page 277. We will restate in a form that will be a little easier to read with multidimensional parameter spaces.

The true parameter value θ_0 is in the interior of the parameter space Θ. This is needed so that derivatives make sense. For example in the Binomial model $\Theta = [0, 1]$, but we will not allow θ_0 to be equal to 0 or 1.

Theorem 3 (Rice Theorem 8.5B, p 277) *Assume that the regularity Assumptions 1 and 2 hold. Let $\hat{\theta}_n$ be the MLE. Then*

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \to N(0, I^{-1}(\theta_0))$$

in distribution as $n \to \infty$.

In the multidimensional parameter case $I(\theta)$ is a matrix, and $I^{-1}(\theta)$ is a covariance matrix (in various texts it is called the variance matrix, the covariance matrix, the variance-covariance matrix). The student should also recall the multivariate normal distribution from Chapter 3. We then have the limiting marginal normal distributions for the components of $\hat{\theta}_n$. In particular for dimension 2, write

$$I^{-1}(\theta) = \begin{pmatrix} I^{-1}_{1,1} & I^{-1}_{1,2} \\ I^{-1}_{2,1} & I^{-1}_{2,2} \end{pmatrix}$$

where in this notation the elements are the (j, k)-th elements of the matrix I^{-1}, not 1 over the element (j, k) of the matrix I. This is a common notation in computer science and engineering. When we need to emphasize the role of θ in these we write for example $I^{-1}_{1,1}(\theta)$. Writing $\hat{\theta}_n = (\hat{\theta}_{1,n}, \hat{\theta}_{2,n})$ we then obtain the limit marginal distributions

$$\sqrt{n}\left(\hat{\theta}_{1,n} - \theta_0\right) \to N(0, I^{-1}_{1,1}(\theta_0))$$

and

$$\sqrt{n}\left(\hat{\theta}_{2,n} - \theta_0\right) \to N(0, I^{-1}_{2,2}(\theta_0))$$

We can use these to help us construct marginal confidence intervals.

Finally as before, in practice if we do not know the value θ_0 we use an estimate, the MLE, and obtain the following approximate normal distributions for the estimator

$$\sqrt{n}\left(\hat{\theta}_{1,n} - \theta_0\right) \approx N(0, I^{-1}_{1,1}(\hat{\theta}_n))$$

and

$$\sqrt{n}\left(\hat{\theta}_{2,n} - \theta_0\right) \approx N(0, I^{-1}_{2,2}(\hat{\theta}_n))$$

when n is large. This is the method that we will use when it applies. Another method is discussed after an outline of the proof of Theorem 3, and is useful as it is a by product of many iterative methods.

This is a result not proven in this course, but is similar to a result related to the Central Limit Theorem (CLT) in which one replaces the population variance term in the denominator with the consistent estimator of σ^2, S_n^2 = sample variance.

Notice we need both components of θ to calculate $I^{-1}(\theta)$, and to calculate for example $I^{-1}_{1,1}(\hat{\theta}_n)$.

Proof of Theorem 3; outline

We also only sketch this in the case on θ being real (Θ of dimension 1).

Recall the log likelihood is

$$\ell(\theta) = \sum_{i=1}^{n} \log(f(X_i;\theta))$$

and the X_is are iid $f(\cdot;\theta_0)$. Therefore for any given θ, $\frac{\partial \log(f(X_i;\theta)}{\partial\theta}$ are iid mean 0 and variance $I(\theta_0)$. Using a Taylor's series of order 1 we obtain

$$0 = \ell'(\hat{\theta}_n) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta}_n - \theta_0) \tag{2}$$

By using the method in the proof of the CLT

$$\begin{aligned} W_n = \frac{1}{\sqrt{n}}\ell'(\theta_0) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log(f(X_i;\theta_0))}{\partial\theta} \\ &\to N(0, I(\theta_0)) \end{aligned}$$

in distribution as $n \to \infty$. Dividing both sides of (2) by $\sqrt{n}$ giving

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}}\ell'(\theta_0) + \frac{1}{\sqrt{n}}\ell''(\theta_0)(\hat{\theta}_n - \theta_0) \\ &= \frac{1}{\sqrt{n}}\ell'(\theta_0) + \left(\frac{1}{n}\ell''(\theta_0)\right)\sqrt{n}(\hat{\theta}_n - \theta_0) \end{aligned}$$

Also notice that

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i;\theta_0)}{\partial\theta^2}$$

which is an average of iid random variables. Assuming these have finding variances then the Law of Large Numbers applies, and then

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i;\theta_0)}{\partial\theta^2} \to \mathrm{E}_{\theta_0}\left\{\frac{\partial^2 \log(f(X;\theta_0)}{\partial\theta^2}\right\} = -I(\theta_0)$$

where the convergence is in probability as $n \to \infty$.

Therefore

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \approx \frac{1}{\left(\frac{1}{n}\ell''(\theta_0)\right)}\frac{1}{\sqrt{n}}\ell'(\theta_0) \approx \frac{-1}{I(\theta_0)}W_n \to N(0, I(\theta_0)^{-1}) \ .$$

End of Proof

There is a second method to approximate or estimate $I(\theta_0)$. Earlier we discussed using $I(\hat{\theta}_n)$. In the proof near the end we used

$$\frac{1}{n}\ell''(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i;\theta_0)}{\partial\theta^2} \to \mathrm{E}_{\theta_0}\left\{\frac{\partial^2 \log(f(X;\theta_0)}{\partial\theta^2}\right\} = -I(\theta_0)$$

While it is beyond the mathematical tools in our course it is also true that

$$\frac{1}{n}\ell''(\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log(f(X_i;\hat{\theta}_n)}{\partial\theta^2} \to -I(\theta_0)$$

in probability as $n \to \infty$.

Many iterative or numerical methods for optimizing output something called the Hessian, which is the second derivative. It usually refers to $-\ell''(\hat{\theta}_n)$, but may also refer to

$$-\frac{1}{n}\ell''(\hat{\theta}_n) \ .$$

One has to look at the documentation to see which is used for a particular program, or else run a test case to determine which is used. Recall also that many of the numerical methods minimize a function, so that for those methods we use negative log likelihood.

In the discussion below we assume that the Hessian refers to $-\ell''(\hat{\theta}_n)$. In that case the Hessian is approximately $nI(\theta_0)$. Also the standard error of an estimator of a maximum likelihood estimator $\hat{\theta}_n$ is

$$\text{Var}(\hat{\theta}_n) \approx \frac{1}{nI(\theta_0)} \approx \frac{1}{nI(\hat{\theta}_n)} \approx \frac{1}{-\ell''(\hat{\theta}_n)} = \frac{1}{\text{Hessian}}$$

Any of these expressions are used depending on the numerical procedures used and the form of $I(\theta)$. $I(\theta_0)$ is known in some hypothesis testing situations, but usually we need to use either $\frac{1}{nI(\hat{\theta}_n)}$ or $\frac{1}{\text{Hessian}}$.

In the non regular case one may have non normal limit theorems. For example consider the Uniform(0,θ) model with $\theta > 0$. We have found the MLE to be $\hat{\theta}_n = \max\{X_1, \ldots, X_n\} = X_{(n)}$. Let F_n be the cdf of $X_{(n)}$ and let F be the cdf of the Uniform(0,θ) distribution. Then for $0 < x < \theta_0$ (recall material about order statistics)

$$F_n(x) = F(x)^n \ .$$

It is not possible to a obtain a normal distribution limit from this. We will not study this type of problem further in this course. However for a fixed value of θ_0, for example 2, we can do a computer simulation of $X_{(n)}$. Do this for various sample size n and see that the histogram of $X_{(n)}$ does not look of the of a normal distribution.