Finite and large sample properties of MLE

Contents

We now turn our attention to some appealing mathematical properties of the maximum likelihood estimators. We will start with key finite sample properties before discussing, in the next section, some asymptotic features.

1 Finite sample properties

1.1 Efficiency

Without loss of generality, we consider here the one parameter case. Suppose that an efficient estimator $\hat{\delta} \equiv \hat{\delta}(\boldsymbol{X})$ of θ exists. By the CRLB attainment theorem (assuming the required assumptions are met), $\hat{\delta}$ must satisfy the equation

$$S_n(\theta, \boldsymbol{X}) = I_n(\theta)\left(\hat{\delta} - \theta\right), \tag{1}$$

where $I_n(\theta) = Var\left(S_n^2(\theta, \boldsymbol{X})\right) = -E\left(\partial_\theta^2 \ell_n(\theta|\boldsymbol{x})\right)$ is the Fisher information, contained in the sample, about θ. Now, let $\hat{\theta}$ be the MLE of θ, then $\hat{\theta}$ satisfies $S_n(\hat{\theta}, \boldsymbol{X}) = 0$. Thus, provided that $I_n(\hat{\theta}) > 0$, the MLE $\hat{\theta}$ coincides with the efficient estimator $\hat{\delta}$.

Theorem 1.1 (MLE and efficiency)**.** *If an efficient estimator exists, then the maximum likelihood method of estimation will produce it.*

In other words, $\hat{\theta}$ is efficient $\implies \hat{\theta}$ is the MLE. The opposite of this statement is not true; as a counter-example, see the example above with $Unif[0, \theta]$.

1.2 Invariance to re-paramerization

Theorem 1.2. *A MLE is invariant with respect to any bijective transformation. That is, if $g : \Theta \longrightarrow \Lambda$ is bijective, then*

$$\hat{\theta} \text{ is a MLE of } \theta \Leftrightarrow g(\hat{\theta}) \text{ is a MLE of } g(\theta).$$

To see why this is the case, let $L_n(\theta)$ be the likelihood with the parametrization θ, i.e. $L_n(\theta) = \prod_i f(x_i, \theta)$, and $L_n^*(\eta)$ be the likelihood with the parametrization $\eta = g(\theta)$, i.e. $L_n^*(\eta) = \prod_i f^*(x_i, \eta)$, where $f^*(x, \eta) := f(x, g^{-1}(\eta))$. Put $\hat{\eta} = g(\hat{\theta})$, and let η be any point in Λ and $\theta = g^{-1}(\eta)$. Since $\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta)$, we have that,

$$L_n^*(\hat{\eta}) = L_n(g^{-1}(\hat{\eta})) = L_n(\hat{\theta}) \geq L_n(\theta) = L_n(g^{-1}(\eta)) = L_n^*(\eta).$$

We conclude that $L_n^*(\hat{\eta}) \geq L_n^*(\eta)$, $\forall \eta \in \Lambda$, and so $\hat{\eta}$ is the MLE of η.

Example 1.1. Let X_i, $i = 1, \ldots, n$, be an iid sample from $Ber(\pi)$, $\pi \in (0, 1)$. We have seen that the likelihood function

is given by $L_n(\pi|\boldsymbol{x}) = \pi^{\sum_{i=1}^n x_i}(1-\pi)^{n-\sum_{i=1}^n x_i}$, and the MLE of π is $\hat{\pi} = \overline{X}_n$.
Now, put $\eta = \log(\pi/(1-\pi))$; with this parametrization, the likelihood function becomes

$$L_n^*(\eta|\boldsymbol{x}) = \left(\frac{1}{1+e^{-\eta}}\right)^{\sum_{i=1}^n x_i}\left(\frac{1}{1+e^{\eta}}\right)^{n-\sum_{i=1}^n x_i}.$$

One can check, by taking the logarithm and then differentiate, that the maximizer of this likelihood is $\hat{\eta} = \log(\sum_i X_i/(n-\sum_i X_i))$. Hence, the MLE of $\eta = \log(\pi/(1-\pi))$ is $\hat{\eta} = \log(\hat{\pi}/(1-\hat{\pi}))$. The same result can be obtained directly (without any calculation) by simply applying the above theorem. □

This simple version of the invariance of MLE is not always useful because many of the functions we are interested in are not bijective. For example, in the case of $Ber(\pi)$, we can't apply the result above to get the MLE of $\pi(1-\pi)$.
Now, whether g is bijective or not, if $\hat{\theta}$ is the MLE of θ then it is reasonable to use $g(\hat{\theta})$ as an estimator for $g(\theta)$. *Even if g is not bejective, some authors refer to $g(\hat{\theta})$ as the MLE of $g(\theta)$*, although, according to the (classical) definition of MLE, we can't really call it so because we can't necessarily express the pd as a genuine function of $g(\theta)$.

Exercise 1.1. Let X_i, $i = 1,\ldots,n$, an iid sample from $N(\mu,\sigma^2)$. Propose a "good" estimator for the coefficient of variation $cv = \sigma/\mu$ ($\mu \neq 0$ and σ are unknown). Justify your choice.

2 Large sample properties

We will show that the MLE is ***often***: (1) consistent, (2) asymptotically normal, and (3) asymptotically efficient.

For (1) to hold, we need some regularity conditions:

- We observe X_i, $i = 1, \ldots, n$, an iid sample from a pd $f(x;\theta)$, $\theta \in \Theta$.
- The model is identifiable; that is, if $f(x;\theta) = f(x;\tilde{\theta})\ \forall x$, then $\theta = \tilde{\theta}$.
- The model is correctly specified. *We denote by θ_0 is the true parameter value.*
- Θ is a compact set (closed and bounded) and $\theta \mapsto f(x;\theta)$ is continuous on Θ.
- $|f(X;\theta)| \leq d(X)$, $\forall \theta \in \Theta$, and $E_{\theta_0}(d(X)) < \infty$.

For (2)-(3) to hold, we need in addition to the above conditions the next assumptions:

- The support of f is independent of θ.
- θ_0 is in the interior of Θ.
- f is twice continuously differentiable in θ and $\int f(x,\theta)dx$ can be differentiated two times under the integral sign.
- The Fisher information satisfies $0 < I(\theta_0) < \infty$.
- In a neighborhood of θ_0, $\left|\partial^3_\theta \log f(x;\theta)\right| \leq M(x)\ \forall \theta$ and $E_{\theta_0}(M(X)) < \infty$.

These assumptions are sufficient (to prove consistency and asymptotic normality) but not necessary. More general and weaker conditions can be found in the literature.

2.1 Consistency

Theorem 2.1 (Consistency of MLEs)**.** *Under the regularity assumptions stated above,* $\hat{\theta}_n \xrightarrow{p} \theta_0$.

We will not prove this result here, but will only sketch out why this is happening.

First, by definition, the MLE $\hat{\theta}_n$ is the maximizer of $\bar{\ell}_n(\theta) = n^{-1} \sum_{i=1}^{n} \log f(X_i; \theta)$ which is the log-likelihood function normalized by $1/n$ (of course, this does not affect maximization). Second, by the law of large numbers (WLLN),

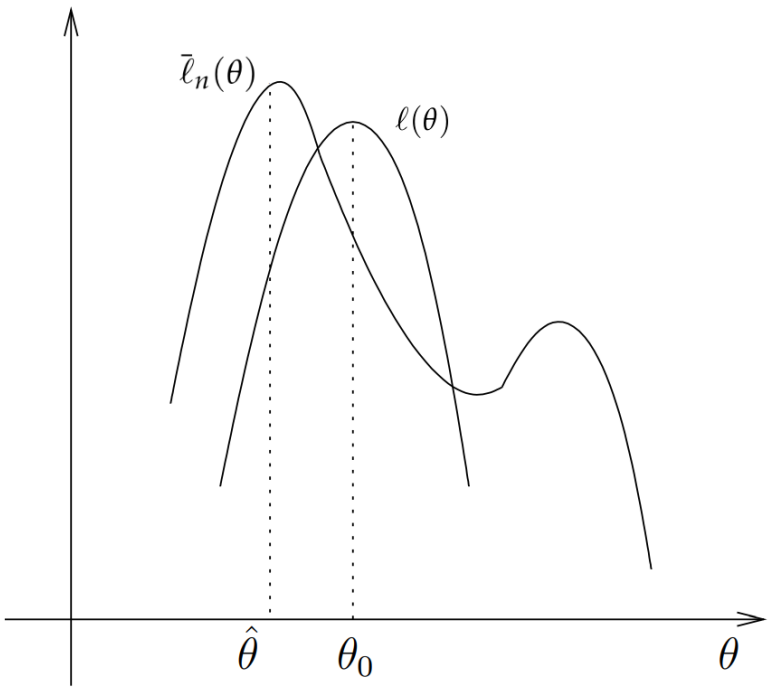
$$\bar{\ell}_n(\theta) \xrightarrow{p} \ell(\theta) := E_{\theta_0}(\log f(X; \theta)), \ \forall \theta.$$

The expectation operator E is indexed by θ_0 to explicitly point out that the expectation is evaluated using the true parameter θ_0, i.e. acknowledging that the pd of X is $f(x; \theta_0)$. Third, by Jensen's inequality, for any $\theta \neq \theta_0$,

$$\ell(\theta) - \ell(\theta_0) = E_{\theta_0}\left(\log \frac{f(X;\theta)}{f(X;\theta_0)}\right) < \log E_{\theta_0}\left(\frac{f(X;\theta)}{f(X;\theta_0)}\right) = 0.$$

So, θ_0 is the maximizer of $\ell(\theta)$.

To sum up, we know that $\theta_0 = \arg\max \ell(\theta)$, $\hat{\theta}_n = \arg\max \bar{\ell}_n(\theta)$, and $\bar{\ell}_n(\theta) \xrightarrow{p} \ell(\theta)$, $\forall\theta$. So, we (intuitively) expect $\hat{\theta}_n$ to approach θ_0 as the sample size increases.



In fcat, the assumed hypothesis guarantees that this is indeed the case. Thus, $\hat{\theta}_n \xrightarrow{p} \theta_0$.

2.2 Asymptotic normality and asymptotic efficiency

Theorem 2.2 (Asymptotic normality of MLEs)**.** *Under the regularity assumptions stated above,*

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

where $I(\theta_0)$ is the Fisher information evaluated at θ_0.

The above result can be expressed as $\hat{\theta} \sim_a N(\theta_0, I_n^{-1}(\theta_0))$, where $I_n(\theta) = nI(\theta)$. And thus, *the MLE $\hat{\theta}$ is asymptotically efficient* for θ_0 (according to the definition given previously).

The proof uses Taylor's theorem, CLT, and Slutsky's theorem. In fact, Taylor expansion of $\theta \mapsto \bar{\ell}_n'(\theta)$ around θ_0 yields to $0 = \bar{\ell}_n'(\hat{\theta}) \approx \bar{\ell}_n'(\theta_0) + (\hat{\theta} - \theta_0)\bar{\ell}_n''(\theta_0)$. So,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\frac{1}{\bar{\ell}_n''(\theta_0)}\sqrt{n}\bar{\ell}_n'(\theta_0).$$

$\sqrt{n}\bar{\ell}_n'(\theta_0) = \sqrt{n}\big(n^{-1}\sum_i \partial_\theta \log f(X_i;\theta_0) - 0\big) \xrightarrow{d} N(0, I(\theta_0))$ by the CLT and the fact that $\partial_\theta \log f(X_i;\theta_0)$ has mean zero and variance $I(\theta_0)$. For the denominator, by WLLN, we have that $-\bar{\ell}_n''(\theta_0) = -n^{-1}\sum_i \partial^2_{\theta^2} \log f(X_i;\theta_0) \xrightarrow{p} I(\theta_0)$. The proof is completed by applying Slutsky's Theorem.

Attention: To simplify the notations, henceforth, we suppress the subscript 0 in θ_0, and write $\hat{\theta} \xrightarrow{p} \theta$, which must be understood as $\hat{\theta}$ converges to the true value of θ whatever this one is. In the same way, we write $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I^{-1}(\theta))$.

The results stated above (consistency, asymptotic normality and asymptotic efficiency) can be extended to the multiparameter case.

If $\hat{\boldsymbol{\theta}}$ is $d-$dimensional MLE of $\boldsymbol{\theta}$, then, under some regularity assumptions (similar to those stated above),

- $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$,
- $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$, where $\boldsymbol{I}^{-1}$ is the inverse of the Fisher information matrix, and
- $\hat{\boldsymbol{\theta}}$ is asymptotically efficient for $\boldsymbol{\theta}$.

Example 2.1.

- Let X_i, $i = 1, \ldots, n$, be an iid sample from $Bin(m, \pi)$, $\pi \in (0, 1)$. We have that

$$
\begin{aligned}
L_n(\pi|\boldsymbol{x}) &= \left(\prod_{i=1}^{n} C_m^{x_i} \right) \pi^{\sum_{i=1}^n x_i} (1-\pi)^{\sum_{i=1}^n (m - x_i)} \\
\ell_n(\pi|\boldsymbol{x}) &= \left(\sum_{i=1}^{n} x_i \right) \log(\pi) + \left(nm - \sum_{i=1}^{n} x_i \right) \log(1-\pi) + const, \\
S_n(\pi, \boldsymbol{x}) &= \partial_\pi \ell_n(\pi|\boldsymbol{x}) = \frac{\sum_i x_i}{\pi} - \frac{nm - \sum_i x_i}{1-\pi}, \\
\partial^2_\pi \ell_n(\pi|\boldsymbol{x}) &= \partial_\pi S_n(\pi, \boldsymbol{x}) = -\frac{\sum_i x_i}{\pi^2} - \frac{nm - \sum_i x_i}{(1-\pi)^2}.
\end{aligned}
$$

  Hence, the MLE of π is $\hat{\pi} = \dfrac{\sum_i X_i}{nm} = \dfrac{\overline{X}_n}{m}$. This later is consistent and asymptotically normal. More precisely

$$\hat{\pi} \sim_a N(\pi, I_n^{-1}(\pi)),$$

  where $I_n(\pi) := -E(\partial_{\pi^2} \ell_n(\pi|\boldsymbol{x})) = \dfrac{nm}{\pi(1-\pi)}$.

  Note that the asymptotic distribution of $\hat{\pi}$, as given above, can also be obtained by applying the CLT directly to $\overline{X}_n$.

- Let $\boldsymbol{X_i} = (X_{1i}, X_{2i}, X_{3i})$, $i = 1, \ldots, n$, be iid rve from the trinomial distribution with joint pd

$$f(\boldsymbol{x}; \boldsymbol{\pi}) = \frac{m!}{x_1! x_2! x_3!} \pi_1^{x_1} \pi_2^{x_2} \pi_3^{x_3},$$

We have seen the the MLE of (π_1, π_2) is $(\overline{X}_{1n}/m, \overline{X}_{2n}/m)$. We have also seen that

$$I^{-1}(\pi_1, \pi_2) = m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix}.$$

Hence,

$$\sqrt{n} \begin{pmatrix} \hat{\pi}_1 - \pi_1 \\ \hat{\pi}_2 - \pi_2 \end{pmatrix} \xrightarrow{d} N_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, m^{-1} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) \end{pmatrix} \right).$$

Again, the same result can be obtained by applying the CLT directly on $\overline{\boldsymbol{X}}_{\boldsymbol{n}} = (\overline{X}_{1n}, \overline{X}_{2n})$. □

The theoretical results presented above can be extended to the situation where the parameter of interest is $\boldsymbol{g}(\boldsymbol{\theta})$ rather than $\boldsymbol{\theta}$. In fact, let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Then, *whether* $\boldsymbol{g} : \mathbb{R}^d \mapsto \mathbb{R}^p$ *is bijective or not*,

- By the continuous mapping theorem, $\boldsymbol{g}(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{\theta})$;

- By the Delta method, $\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} N_p\big(\mathbf{0}, \boldsymbol{I}^{-1}(g(\theta))\big)$, where $\boldsymbol{I}^{-1}(g(\theta)) = \dot{g}(\theta)\boldsymbol{I}^{-1}(\theta)\dot{g}^t(\theta)$;
- $g(\hat{\theta})$ is asymptotically efficient for $g(\theta)$.

2.3 Observed Fisher information

We have seen that the MLE $\hat{\theta}$ of θ is asymptotically normal, i.e. $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N_d(\mathbf{0}, \boldsymbol{I}^{-1}(\theta))$. Equivalently, we can write that

$$\sqrt{\boldsymbol{I}_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}) \ ,$$

where $\mathbb{1}$ is the identity matrix and $\sqrt{\boldsymbol{I}_n(\theta)}$ is the *square-root matrix* of the FI

$$\boldsymbol{I}_n(\theta) = -E\left\{\left[\partial_{\theta_j\theta_k}\ell_n(\theta)\right]_{j,k}\right\} = n\boldsymbol{I}(\theta) \ , \text{with } \boldsymbol{I}(\theta) = -E\left\{\left[\partial_{\theta_j\theta_k}\log f(X;\theta)\right]_{j,k}\right\}.$$

To use this asymptotic normality in practical inference, $\boldsymbol{I}(\theta)$ must be estimated. The most obvious estimator of $\boldsymbol{I}(\theta)$ is $\boldsymbol{I}(\hat{\theta})$. Since $\hat{\theta} \xrightarrow{p} \theta$, $\boldsymbol{I}(\hat{\theta}) \xrightarrow{p} \boldsymbol{I}(\theta)$, provided that $\theta \mapsto \boldsymbol{I}(\theta)$ is a continuous function. In this case, Slutsky's theorem ensures that

$$\sqrt{\boldsymbol{I}_n(\hat{\theta})}(\hat{\theta} - \theta) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

The disadvantage of this approach is that, in practice, it is not always easy to calculate $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ because of the difficulties of working out the expectations.

Let's define the matrix

$$\boldsymbol{J}_n(\boldsymbol{\theta}) := -\boldsymbol{\nabla}^2 \ell_n(\boldsymbol{\theta}) = -\left[\partial_{\theta_j\theta_k} \ell_n(\boldsymbol{\theta})\right]_{j,k} = -\left[\sum_{i=1}^{n} \partial_{\theta_j\theta_k} \log f(X_i;\boldsymbol{\theta})\right]_{j,k}.$$

$\boldsymbol{J}_n(\boldsymbol{\theta})$ is the Hessian of the negative log-likelihood. This matrix is known as *the sample Fisher information* or *the observed Fisher information*. It can always be calculated as long as the second partial derivatives can be calculated. Observe that $\boldsymbol{I}_n(\boldsymbol{\theta}) = E\left(\boldsymbol{J}_n(\boldsymbol{\theta})\right)$. The law of large numbers guarantees that $n^{-1}\boldsymbol{J}_n(\boldsymbol{\theta}) \xrightarrow{p} \boldsymbol{I}(\boldsymbol{\theta})$. So,

$$\sqrt{\boldsymbol{J}_n(\boldsymbol{\theta})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

Again Slutsky's theorem can be applied to show that

$$\sqrt{\boldsymbol{J}_n(\hat{\boldsymbol{\theta}})}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N_d(\mathbf{0}, \mathbb{1}).$$

To sum up, ***for large sample sizes***, one can use $\boldsymbol{I}_n(\boldsymbol{\theta})$, $\boldsymbol{I}_n(\hat{\boldsymbol{\theta}})$, $\boldsymbol{J}_n(\boldsymbol{\theta})$ and $\boldsymbol{J}_n(\hat{\boldsymbol{\theta}})$ interchangeably.

Example 2.2. Let X_i, $i = 1, \ldots, n$, be an iid sample from the pd

$$f(x;\theta) = \frac{1+\theta x}{2} I(-1 \leq x \leq 1);\ -1 < \theta < 1.$$

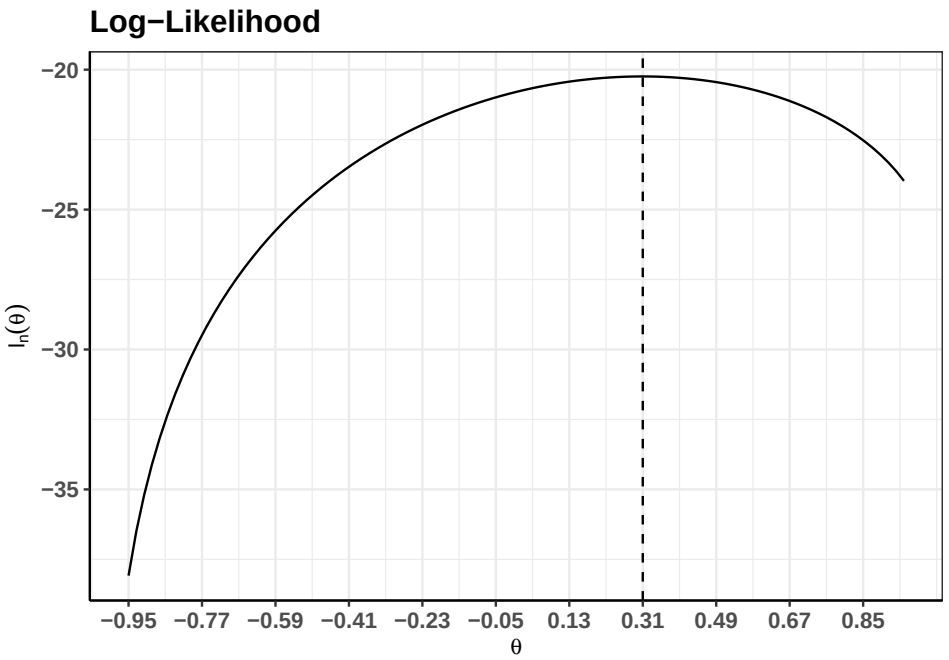
The log-likelihood, the Score and the observed FI are given by

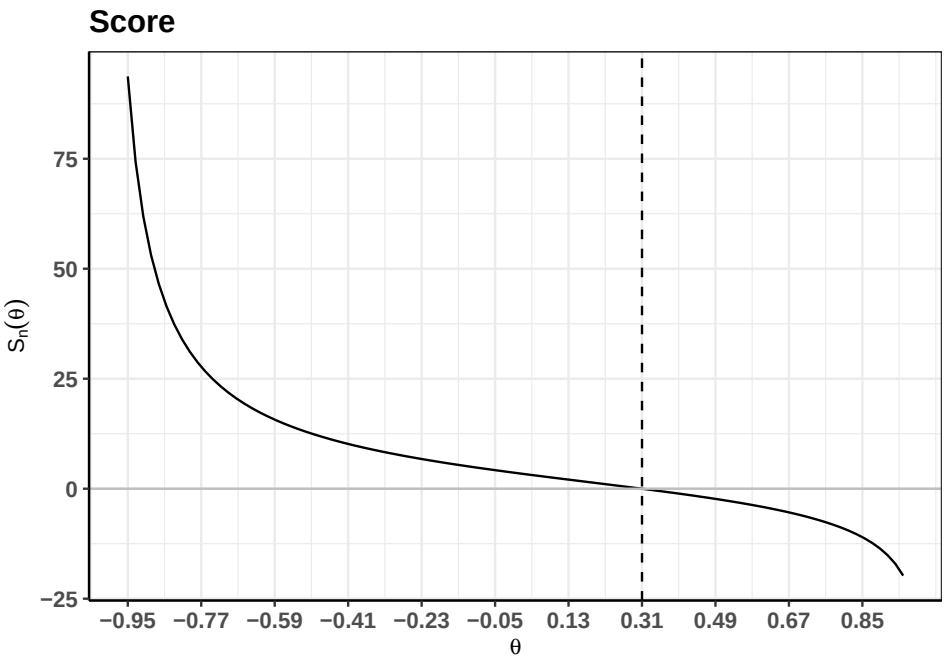
$$\begin{aligned}
\ell_n(\theta|x) &= \sum_{i=1}^{n} \log(1+\theta x_i) - n\log(2), \\
S_n(\theta, x) &= \partial_\theta \ell_n(\theta|x) = \sum_{i=1}^{n} \frac{x_i}{1+\theta x_i}, \\
J_n(\theta) &= -\partial_\theta S_n(\theta, x) = -\partial_\theta^2 \ell_n(\theta|x) = \sum_{i=1}^{n} \frac{x_i^2}{(1+\theta x_i)^2}.
\end{aligned}$$

As a function of θ, the log-likelihood is *continuous and strictly concave* ($J_n > 0$), so there is a unique MLE, say $\hat{\theta}$, of θ. Now, although the likelihood equation of this model *cannot be solved explicitly* to get the analytic expression of $\hat{\theta}$, the theory tells us that $\sqrt{I_n(\theta)}(\hat{\theta} - \theta) \xrightarrow{d} N(0,1)$, where $I_n(\theta) = E(J_n(\theta))$.

The following data are simulated from the f above with $\theta = 0.5$ (hereafter, we'll pretend that we don't know the θ that generated the data and see how maximum likelihood behaves).

```r
x <- c(0.9852, 0.0450, -0.6123, -0.7518, -0.2824, 0.7085, -0.0711, 0.9625,
       -0.4746, 0.1617, -0.4592, -0.3113, 0.6800, -0.6694, 0.1512, -0.7048,
       0.3421, -0.9658, 0.9809, -0.1205, 0.4730, -0.1665, 0.9956, 0.8720,
       0.9849, -0.7650, 0.4528, 0.2190, 0.9611, -0.0257)
```





By inspecting the graphs above, we can see that $\hat{\theta} = 0.31$ (we will see later how to obtain this result numerically).

As we discussed above, we can estimate $I_n(\theta)$ using $I_n(\hat{\theta})$ or $J_n(\hat{\theta})$. To use the former, we must first derive the expression of I_n, which can be done by using the polynomial division. In fact,

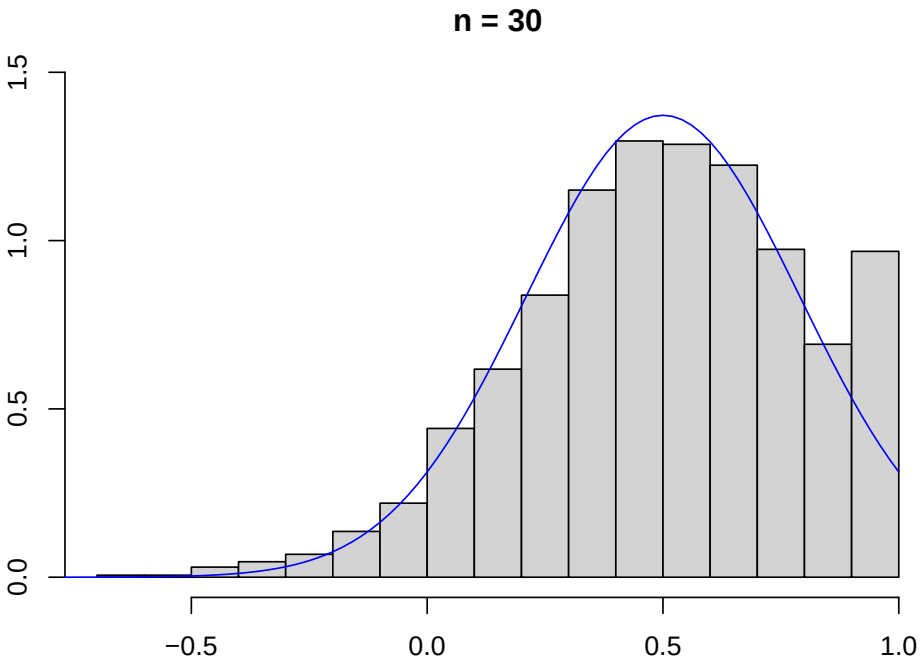
$$\begin{aligned} I_n(\theta) &= nE\left(\frac{X^2}{(1+\theta X)^2}\right) = \frac{n}{2}\int_{-1}^{1}\left(\frac{1}{\theta}x - \frac{1}{\theta^2}\right)dx + \frac{n}{2\theta^2}\int_{-1}^{1}\frac{1}{1+\theta x}dx \\ &= -\frac{n}{\theta^2} + \frac{n}{2\theta^3}\log\left(\frac{1+\theta}{1-\theta}\right). \end{aligned}$$

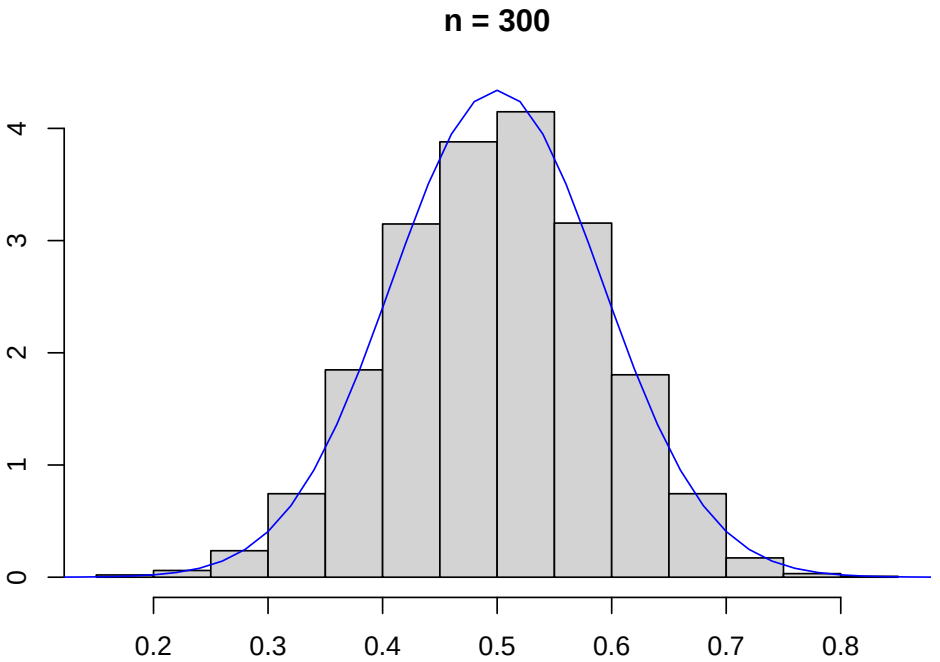
In our case, $n = 30$, $\hat{\theta} = 0.31$, $I_{30}(\hat{\theta}) = 10.619$ and $1/I_{30}(\hat{\theta}) = 0.094$. Hence, based on the observed data, we conclude that $\hat{\theta}_{30} \sim_a N(\theta, 0.094)$.

The second method for estimating $I_n(\theta)$ is much simpler and consists in using the observed FI which is given here by $J_{30}(\hat{\theta}) = 11.846$. On this basis, we can write that $\hat{\theta}_{30} \sim_a N(\theta, 0.084)$.

Note that the true (but *unknown*) asymptotic variance of $\hat{\theta}_{30}$ is actually $1/I_{30}(0.5) = 0.085$.

To see how the asymptotic normal approximation works in practice, we repeat the data generation procedure 5000 times and calculate $\hat{\theta}$ each time. We did this for $n = 30$ and $n = 300$, respectively. The graph below shows the histogram of the simulated $\hat{\theta}$ and the curve of the true asymptotic normal density, i.e. $N(0.5, I_n^{-1}(0.5))$, in blue.


n = 30
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0


n = 300
0
1
2
3
4
0.2
0.3
0.4
0.5
0.6
0.7
0.8