Fisher Information and Cramer-Rao Bound

Maura Mezzetti

Department of Economics and Finance
Università Tor Vergata

Outline

Point Estimation

"What! you have solved it already?"
"Well, that would be too much to say. I have discovered a suggestive fact, that is all."

Dr. Watson and Sherlock Holmes
The Sign of Four

Fisher Information

Estimation:

Information about the parameter is obtained from a sample of data coming from the underlying probability distribution

How much information can a sample of data provide about the unknown parameter?

Fisher Information

Estimation:

Information about the parameter is obtained from a sample of data coming from the underlying probability distribution

How much information can a sample of data provide about the unknown parameter?

This section introduces such a measure for information

This information measure can be used to find bounds on the variance of estimators, and it can be used to approximate the sampling distribution of an estimator obtained from a large sample, and further be used to obtain an approximate confidence interval in case of large sample

Fisher Information

Estimation:

Information about the parameter is obtained from a sample of data coming from the underlying probability distribution

How much information can a sample of data provide about the unknown parameter?

This section introduces such a measure for information

This information measure can be used to find bounds on the variance of estimators, and it can be used to approximate the sampling distribution of an estimator obtained from a large sample, and further be used to obtain an approximate confidence interval in case of large sample

Intuitive explanation of Fisher Information

$logL(\theta)$ is the log-likelihood function where θ is the parameter of interest

The *observed Fisher information* is the curvature at the peak of this function

Observed Fisher Information: $-\frac{\delta^2 logL(\theta)}{\delta^2 \theta}(\hat{\theta}_{MLE})$

intuitively tells us how peaked the likelihood function is or how *well* we know the parameter after data has been collected

Intuitive explanation of Fisher Information

$logL(\theta)$ is the log-likelihood function where θ is the parameter of interest

The *observed Fisher information* is the curvature at the peak of this function

Observed Fisher Information: $-\frac{\delta^2 logL(\theta)}{\delta^2 \theta}(\hat{\theta}_{MLE})$

intuitively tells us how peaked the likelihood function is or how *well* we know the parameter after data has been collected

A log-likelihood which is not terribly peaked is somewhat spread out, and we don't really have much confidence in what θ is after having collected data and conversely, a very peaked likelihood implies we have a great deal of "confidence" of the precise value of θ

Intuitive explanation of Fisher Information

$logL(\theta)$ is the log-likelihood function where θ is the parameter of interest

The *observed Fisher information* is the curvature at the peak of this function

Observed Fisher Information: $-\frac{\delta^2 logL(\theta)}{\delta^2 \theta}(\hat{\theta}_{MLE})$

intuitively tells us how peaked the likelihood function is or how *well* we know the parameter after data has been collected

A log-likelihood which is not terribly peaked is somewhat spread out, and we don't really have much confidence in what θ is after having collected data and conversely, a very peaked likelihood implies we have a great deal of "confidence" of the precise value of θ

Intuitive explanation of Fisher Information

Expected Fisher Information $E_\theta\left(-\frac{\delta^2 logL(\theta)}{\delta^2\theta}\right)$

applies the same concept except we average out the data, and we treat θ as a constant

It tells us on average how curved or peaked the likelihood function will be after the data has been collected, for a prescribed value of θ

Intuitive explanation of Fisher Information

Expected Fisher Information $E_\theta\left(-\frac{\delta^2 logL(\theta)}{\delta^2\theta}\right)$

applies the same concept except we average out the data, and we treat θ as a constant

It tells us on average how curved or peaked the likelihood function will be after the data has been collected, for a prescribed value of θ

In the multi-dimensional setting, we simply take the Hessian as opposed to the second derivative to measure curvature

Intuitive explanation of Fisher Information

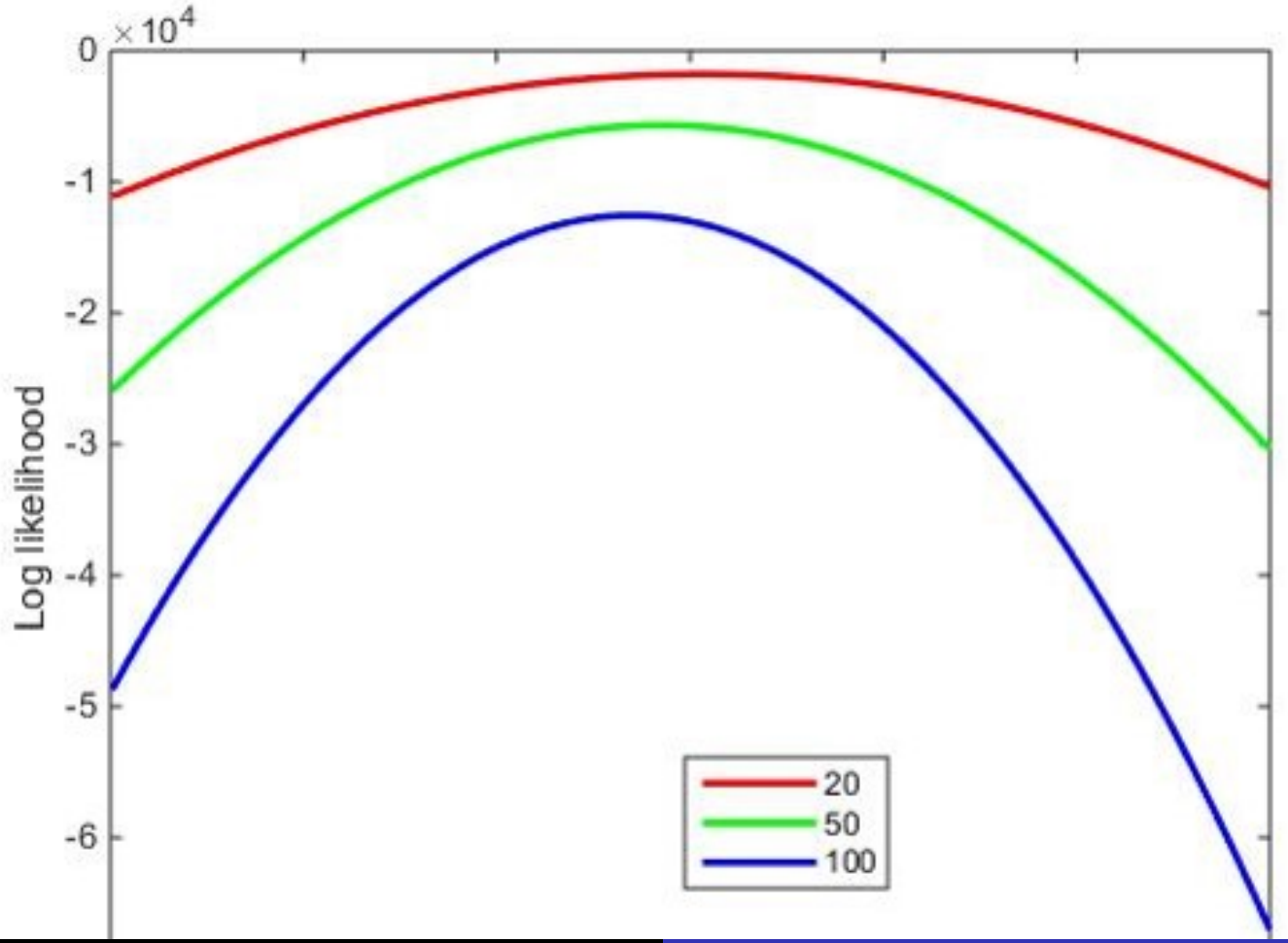
Expected Fisher Information $E_\theta\left(-\frac{\delta^2 logL(\theta)}{\delta^2\theta}\right)$

applies the same concept except we average out the data, and we treat θ as a constant

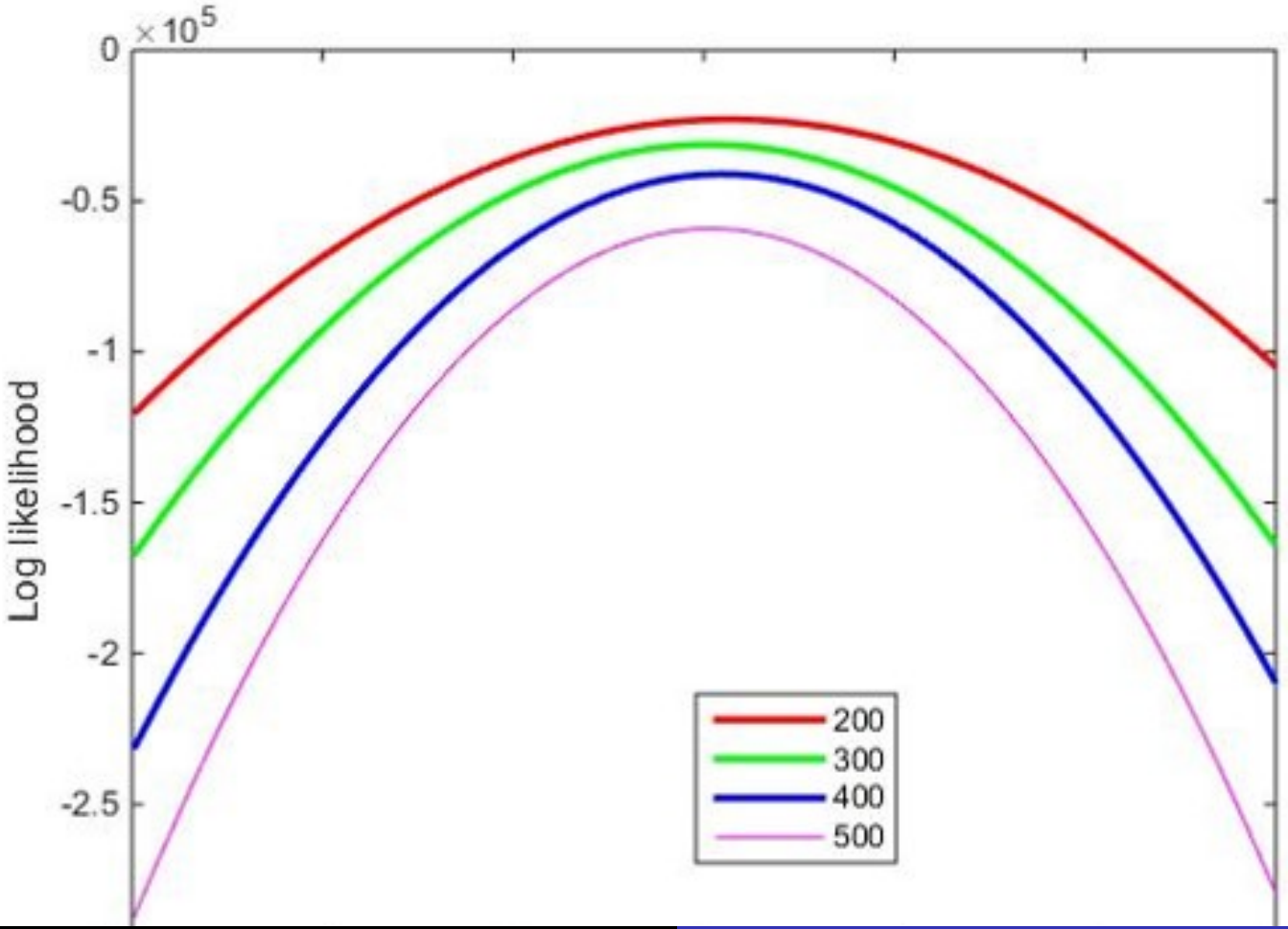
It tells us on average how curved or peaked the likelihood function will be after the data has been collected, for a prescribed value of θ

In the multi-dimensional setting, we simply take the Hessian as opposed to the second derivative to measure curvature

Log likelihood for the population mean in the Gaussian case - varying the sample size



Log likelihood for the population mean in the Gaussian case - varying the sample size



The score vector

The first derivative of the log-likelihood function is called Fisher's score function, and is denoted by

$$u(\theta) = \frac{\partial log L(\theta|x)}{\partial \theta}$$

The score vector

Note that the score is a vector of first partial derivatives, one for each element of θ.

$$u(\theta) = \begin{pmatrix} \frac{\partial log L(\theta|x)}{\partial \theta_1} \\ \vdots \\ \frac{\partial log L(\theta|x)}{\partial \theta_k} \end{pmatrix}$$

The score vector

Important results for the score vector:

Let $\hat{\theta}_{MLE}$ be the MLE for θ

$$u(\hat{\theta}_{MLE}) = 0$$

Under some regular conditions (see Casella Berger):

$$E[u(\theta)] = 0$$

The score vector

Important results for the score vector:

Let $\hat{\theta}_{MLE}$ be the MLE for θ

$$u(\hat{\theta}_{MLE}) = 0$$

Under some regular conditions (see Casella Berger):

$$E[u(\theta)] = 0$$

under some regular conditions (see Casella Berger):

$$E[u^2(\theta)] = E\left[\left(\frac{\partial logL(\theta|x)}{\partial\theta}\right)^2\right] = -E\left[\frac{\partial^2 logL(\theta|x)}{\partial\theta^2}\right]$$

The score vector

Important results for the score vector:

Let $\hat{\theta}_{MLE}$ be the MLE for θ

$$u(\hat{\theta}_{MLE}) = 0$$

Under some regular conditions (see Casella Berger):

$$E[u(\theta)] = 0$$

under some regular conditions (see Casella Berger):

$$E[u^2(\theta)] = E\left[\left(\frac{\partial logL(\theta|x)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 logL(\theta|x)}{\partial \theta^2}\right]$$

Properties of the score vector

Theorem

Let $X_1, \ldots, X_n$ with pdf $f(x|\theta)$ and likelihood function $L(\theta|\mathbf{x})$ derivable with respect to θ at least two times (and some regular conditions are satisfied (see Casella Berger):). Then

$$E[u(\theta)] = E\left[\frac{\partial \log L(\theta|\mathbf{X})}{\partial \theta}\right] = 0$$

Moreover,

$$E[u^2(\theta)] = E\left[\left(\frac{\partial \log L(\theta|\mathbf{X})}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L(\theta|\mathbf{X})}{\partial \theta^2}\right]$$

Remarks

- If $\theta \in \mathbb{R}$ then $E[u(\theta)]$ is a number
- If $\theta \in \mathbb{R}^k$ then $E[u(\theta)]$ is a k-dim vector
- If $\theta \in \mathbb{R}$ then $E[u(\theta)^2] = Var(u(\theta))$ is a number
- If $\theta \in \mathbb{R}^k$ then $E[u(\theta)^2]$ is a $k \times k$-dim matrix called variance covariance matrix of the score function:

$$E[u(\theta)^2] = E\left(\begin{array}{c} \frac{\partial \log L(\theta|X)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\theta|X)}{\partial \theta_k} \end{array} \right) \left(\frac{\partial \log L(\theta|X)}{\partial \theta_1} \dots \frac{\partial \log L(\theta|X)}{\partial \theta_k} \right)$$

The score function

Measures the sensitivity of the likelihood to changes of the parameters for given X

Plays an important role in many applications of ML estimation

The score function

Measures the sensitivity of the likelihood to changes of the parameters for given X

Plays an important role in many applications of ML estimation

The ML estimatex of θ is the value $\hat{\theta}_{MLE}$ that makes the realization of the score equal to its expected value at the true θ (under regularity conditions)

The score function

Measures the sensitivity of the likelihood to changes of the parameters for given X

Plays an important role in many applications of ML estimation

The ML estimatex of θ is the value $\hat{\theta}_{MLE}$ that makes the realization of the score equal to its expected value at the true θ (under regularity conditions)

Proof

Since $f(x|\theta)$ is a pdf: $\int_{-\infty}^{\infty} f(x|\theta)dx = 1$
Taking the derivative:

$$\begin{aligned}
\frac{\partial \int_{-\infty}^{\infty} f(x|\theta)dx}{\partial \theta} &= \frac{\partial 1}{\partial \theta} \\
\int_{-\infty}^{\infty} \frac{\partial f(x|\theta)dx}{\partial \theta} &= 0 \\
\int_{-\infty}^{\infty} \frac{\partial f(x|\theta)}{\partial \theta} \frac{f(x|\theta)}{f(x|\theta)} dx &= \int_{-\infty}^{\infty} \frac{\partial log f(x|\theta)}{\partial \theta} f(x|\theta)dx = 0 \\
E\left(\frac{\partial log f(x|\theta)}{\partial \theta}\right) &= 0
\end{aligned}$$

Proof

$$\begin{aligned}
\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\frac{\partial log\, f(x|\theta)}{\partial\theta}f(x|\theta)dx &= 0 \\
\int_{-\infty}^{\infty}\frac{\partial^2 log\, f(x|\theta)}{\partial\theta^2}f(x|\theta)dx &+ \int_{-\infty}^{\infty}\frac{\partial log\, f(x|\theta)}{\partial\theta}\frac{\partial log\, f(x|\theta)}{\partial\theta}f(x|\theta)dx = 0 \\
\int_{-\infty}^{\infty}\left(\frac{\partial log\, f(x|\theta)}{\partial\theta}\right)^2 f(x|\theta)dx &= -\int_{-\infty}^{\infty}\frac{\partial^2 log\, f(x|\theta)}{\partial\theta^2}f(x|\theta)dx \\
E[u^2(\theta)] &= -E\left[\frac{\partial^2 logL(\theta|x)}{\partial\theta^2}\right]
\end{aligned}$$

The score function for the Geometric distribution

Considering for n observations from a Geometric distribution:

$$p(x|\pi) = \pi(1-\pi)^x$$

The score function is

$$u(\pi) = \frac{\partial logL(\pi|x)}{\partial \pi} = \frac{dlogL(\pi|x)}{d\pi}$$

$$u(\pi) = \frac{n}{\pi} - \frac{\sum_i x_i}{1-\pi}$$

The Information Matrix

- The Fisher Information (sometimes simply called information) is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ if its distribution.
- $$I(\theta) = E\left(\frac{\partial}{\partial\theta} log(f(x;\theta)\right)^2\right)$$
- $$I_n(\theta) = E\left(\frac{\partial}{\partial\theta} log(L(\theta|x))\right)^2\right)$$
- The observed Information or observed Fisher information
$$J_n(\theta) = \left(\frac{\partial}{\partial\theta} log(L(\hat{\theta}|x))\right)^2$$

The Information Matrix

- Fisher Information matrix represents the variance covariance matrix of the score function

$$I_n(\theta) = var\left(u(\theta)\right)$$

- $$I_n(\theta) = -E\left(\frac{\partial^2 logL(\theta|x)}{\partial^2 \theta}\right)$$
- that is (see *Casella and Berger* for conditions):

$$Var\left(\frac{\partial logL(\theta|x)}{\partial \theta}\right) = -E\left(\frac{\partial^2 logL(\theta|x)}{\partial^2 \theta}\right)$$

The Information Matrix

Let us denote the joint pdf of $(X_1, \ldots, X_n)$ as

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- $logL(x|\theta) = \sum_{i=1}^{n} log\, f(x_i|\theta)$
- $E\left(\frac{\partial logL_n(x|\theta)}{\partial \theta}\right) = 0$
- $I_n(\theta) = -E_\theta\left(\frac{\partial^2 logL_n(x|\theta)}{\partial^2 \theta}\right) = -E_\theta\left(\sum_{i=1}^{n} \frac{\partial^2 log\, f(x|\theta)}{\partial^2 \theta}\right) = nI(\theta)$

The Information Matrix

- $I(\theta) = -E_\theta \left(\frac{\partial^2 log\, f(x|\theta)}{\partial^2 \theta} \right)$
- $I_n(\theta) = -E_\theta \left(\frac{\partial^2 log\, L_n(x|\theta)}{\partial^2 \theta} \right) = -E_\theta \left(\sum_{i=1}^{n} \frac{\partial^2 log\, f(x|\theta)}{\partial^2 \theta} \right) = nI(\theta)$

Fisher Information

Definition

Given X with pdf $f(x|\theta)$, the quantity

$$E[u^2(\theta)] = E\left[\left(\frac{\partial \log f(\theta|x)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log f(\theta|X)}{\partial \theta^2}\right]$$

is called Fisher Information of X and denoted with $I(\theta)$.

Definition

Given $X_1, \ldots, X_n$ with pdf $f(x|\theta)$ and likelihood $L(\theta|\mathbf{X})$ the quantity

$$E[u^2(\theta)] = E\left[\left(\frac{\partial \log L(\theta|\mathbf{x})}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \log L(\theta|X)}{\partial \theta^2}\right]$$

Independent and identically distributed r.v.

Theorem Let $X_1, \ldots, X_n$ i.i.d. r.v. with pdf $f(x|\theta)$. Then $I_n(\theta) = nI(\theta)$.

Proof

The likelihood is

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

and

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Now:

$$I_n(\theta) = -E_\theta\left(\frac{\partial^2 \log L(\theta|X)}{\partial^2\theta}\right) = -E_\theta\left(\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i|\theta)}{\partial^2\theta}\right) = nI(\theta)$$

Independent and identically distributed r.v.

Theorem Let $X_1, \ldots, X_n$ i.i.d. r.v. with pdf $f(x|\theta)$. Then $I_n(\theta) = nI(\theta)$.

Proof

The likelihood is

$$L(\theta|x) = \prod_{i=1}^{n} f(x_i|\theta)$$

and

$$\log L(\theta|x) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Now:

$$I_n(\theta) = -E_\theta \left(\frac{\partial^2 \log L(\theta|X)}{\partial^2 \theta} \right) = -E_\theta \left(\sum_{i=1}^{n} \frac{\partial^2 \log f(X_i|\theta)}{\partial^2 \theta} \right) = nI(\theta)$$

The Fisher Information for the Geometric distribution

Considering for n observations from a Geometric distribution:

$$p(x|\pi) = \pi(1-\pi)^x$$

The score function is

$$\frac{dlogL(\pi|x)}{d\pi} = \frac{n}{\pi} - \frac{\sum_i x_i}{1-\pi}$$

The second derivative:

$$\frac{d^2 logL(\pi|x)}{d\pi^2} = -\frac{n}{\pi^2} - \frac{\sum_i x_i}{(1-\pi)^2}$$

Since $E(x) = \frac{1-\pi}{\pi}$, the Fisher information will be:

$$E\left(-\frac{d^2 logL(\pi|x)}{d\pi^2}\right) = \frac{n}{\pi^2} - \frac{n}{\pi(1-\pi)} = \frac{n}{\pi^2(1-\pi)}$$

Information Matrix for Bernoulli distribution

Let us calculate Fisher Information for a Bernoulli(θ) distribution

$$
\begin{aligned}
f(x|\theta) &= \theta^x(1-\theta)^{1-x} \qquad x \in 0,1 \\
l(x,\theta) &= x log\theta + (1-x) log(1-\theta) \\
\frac{\partial l(x,\theta)}{\partial \theta} &= \frac{x}{\theta} - \frac{1-x}{1-\theta} \\
\frac{\partial^2 l(x,\theta)}{\partial \theta^2} &= -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \\
I(\theta) &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} \\
I_n(\theta) &= \frac{n}{\theta(1-\theta)}
\end{aligned}
$$

The Fisher Information for the Geometric distribution

Considering for n observations from a Geometric distribution:

$$p(x|\pi) = \pi(1-\pi)^x$$

The score function is

$$\frac{dlogL(\pi|x)}{d\pi} = \frac{n}{\pi} - \frac{\sum_i x_i}{1-\pi}$$

The Fisher information:

$$\frac{d^2 logL(\pi|x)}{d\pi^2} = -\frac{n}{\pi^2} - \frac{\sum_i x_i}{(1-\pi)^2}$$

Since $E(x) = \frac{1-\pi}{\pi}$

$$E\left(-\frac{d^2 logL(\pi|x)}{d\pi^2}\right) = \frac{n}{\pi^2} - \frac{n}{\pi(1-\pi)} = \frac{n}{\pi^2(1-\pi)}$$

Information Matrix for Gaussian distribution

Let us calculate Fisher Information for an $N(\mu, \sigma^2)$ distribution where σ^2 known

$$
\begin{aligned}
f(x|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
l(x,\mu) &= -\frac{1}{2} log\left(2\pi\sigma^2\right) - \frac{(x-\mu)^2}{2\sigma^2} \\
\frac{\partial l(x,\mu)}{\partial \mu} &= \frac{(x-\mu)}{\sigma^2} \\
\frac{\partial^2 l(x,\mu)}{\partial \mu^2} &= -\frac{1}{\sigma^2} \\
I(\theta) &= \frac{1}{\sigma^2} \\
I_n(\theta) &= \frac{n}{\sigma^2}
\end{aligned}
$$

Information Matrix for Gaussian distribution

- $\frac{\partial^2}{\partial \mu^2} l(\mu, \sigma^2 | x) = -\frac{n}{\sigma^2}$
- $\frac{\partial^2}{\partial \mu \partial \sigma^2} l(\mu, \sigma^2 | x) = -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)$
- $\frac{\partial^2}{\partial \sigma^4} l(\mu, \sigma^2 | x) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2$

Information Matrix for Gaussian distribution

$$I_n(\mu, \sigma^2) = - \begin{bmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & -\frac{n}{2\sigma^4} \end{bmatrix}$$

$$I_n(\mu, \sigma^2)^{-1} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

Asymptotic normality of MLE

Under some regular conditions (see Casella Berger)

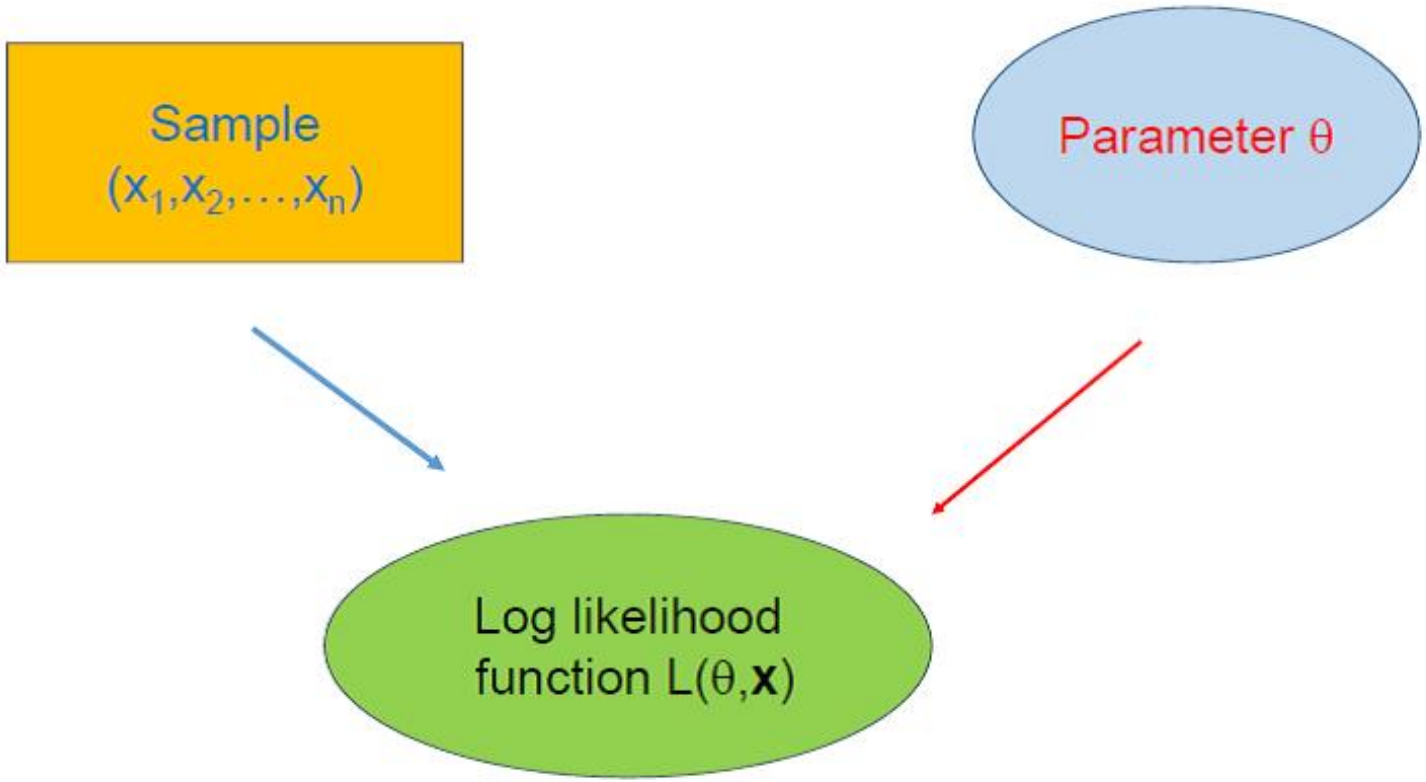
$$\sqrt{n}(\hat{\theta} - \theta_0) \longrightarrow_d N\left(0, \frac{1}{I(\theta_0)}\right)$$

$$(\hat{\theta} - \theta_0) \longrightarrow_d N\left(0, \frac{1}{I_n(\theta_0)}\right)$$

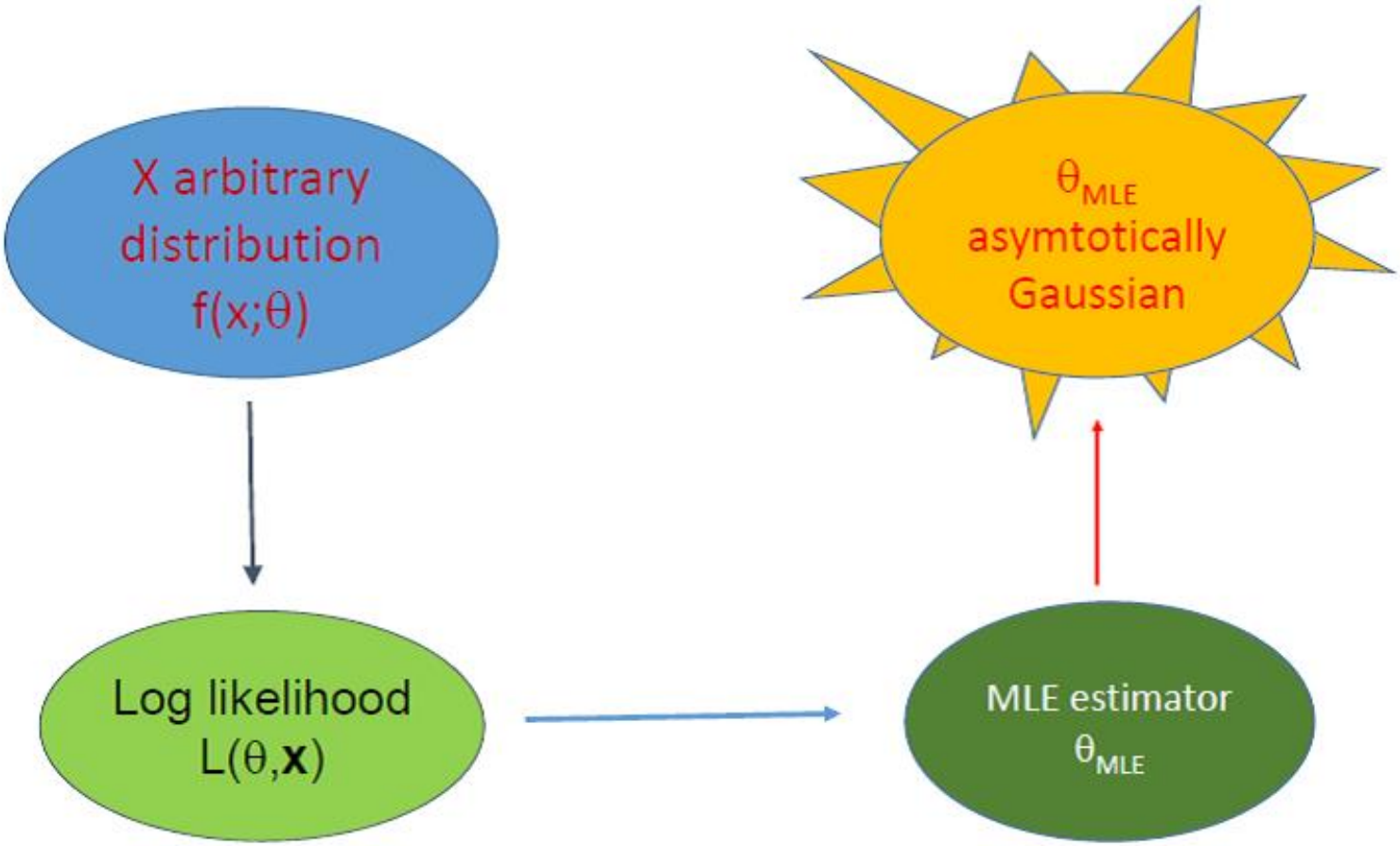
To remember about the likelihood function

- The MLE does not give the "most probable" value of θ. It gives the value θ under which the sample is the most likely: i.e., the likelihood is maximised.
- MLE is not magic: all the problems of inference from sample remain with us.
- For example: I tossed a coin 10 times and got 9 heads. Using this data, the MLE gives $\hat{p} = 0.9$. MLE does not eliminate sampling noise, or give us the truth. Its just a decent estimator.

the (log-)likelihood function depends on two type of arguments



Asymptotic normality of MLE



Non-differentiable likelihood

- $(X_1, \ldots, X_n)$ iid $U(0, \theta) \quad \theta > 0$
- $(X_1, \ldots, X_n)$ iid exponential location parameter with pdf $f(x|\theta) = exp(-(x - \theta)), \quad if\ \theta \geq 0$
- $(X_1, \ldots, X_n)$ iid $U(\theta - 2, \theta + 1/2)$

Cramér-Rao inequality

Theorem: Let $X_1, X_2, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}, \theta)$ and let $W(\mathbf{X}) = W(X_1, \ldots, X_n))$ be an estimator where $E_\theta W(\mathbf{X})$ is a differentiable function of θ. Suppose the joint pdf $f(\mathbf{x}|\theta) = f(x_1, \ldots, x_n|\theta)$ satisfies:

$$\frac{d}{d\theta} \int \ldots \int h(\mathbf{x}) f(\mathbf{x}|\theta) dx_1, \ldots, dx_n = \int \ldots \int h(\mathbf{x}) \frac{\partial}{\partial \theta} f(\mathbf{x}|\theta) dx_1, \ldots, dx_n$$

for any function $h(\mathbf{x})$ with $E_\theta |h(\mathbf{X})| < \infty$. Then:

$$Var_\theta \left(W(\mathbf{X})\right) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X})\right)^2}{E_\theta \left(\left(\frac{\partial}{\partial \theta} I(\mathbf{X}|\theta)\right)^2\right)}$$

$$Var_\theta \left(W(\mathbf{X})\right) \geq \frac{\left(\frac{d}{d\theta} E_\theta W(\mathbf{X})\right)^2}{nI(\theta)}$$

Cramér-Rao inequality: Corollary

Theorem: Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample with pdf $f(x, \theta)$ and let $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ be an unbiased estimator where $f(x|\theta)$ satisfies some regularity conditions, then:

$$Var_\theta\left(T(\mathbf{X})\right) \geq \frac{1}{nE_\theta\left(\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right)}$$

$$Var_\theta\left(T(\mathbf{X})\right) \geq \frac{1}{nI(\theta)}$$

Cramér-Rao inequality: Remarks

- Although Cramér-Rao Lower Bound is stated for continuous random variables, it also applies to discrete random variables. The main key in the proof that allows interchange of integration and differentiation needs to be modified. Even thought $p(x|\theta)$ is a probability distribution function and not differentiable in x, it is differentiable in θ
- Even if Cramér-Rao Theorem is applicable, this does not guarantees that the lower bound is reached by any estimator. Some conditions guarantee the existence of estimator reaching lower bound (exponential family does)
- If we cannot find an estimator that attains the lower bound, how can we asses the existence of the best estimator?

Cramér-Rao inequality: Remarks

If an unbiased estimator attains the Cramér-Rao Lower Bound , it must be the minimum-variance unbiased estimator

The converse is not always true

Not all the MVU estimators attain the CRLB

Cramér-Rao inequality: Remarks

If an unbiased estimator attains the Cramér-Rao Lower Bound , it must be the minimum-variance unbiased estimator

The converse is not always true

Not all the MVU estimators attain the CRLB

An estimator that is unbiased and attains the CRLB is said to be efficient

Cramér-Rao inequality: Remarks

If an unbiased estimator attains the Cramér-Rao Lower Bound , it must be the minimum-variance unbiased estimator

The converse is not always true

Not all the MVU estimators attain the CRLB

An estimator that is unbiased and attains the CRLB is said to be efficient

Rao-Blackwell Theorem

Theorem: Rao-Blackwell : Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = E(W|T)$. Then:

$$E_\theta \phi(T) = \tau(\theta)$$

and

$$Var_\theta \phi(T) \leq Var_\theta W, \qquad \forall \theta$$

that is: $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$

Theorem: If W is a best unbiased estimator of $\tau(\theta)$, then W is unique.

Rao-Blackwell Theorem: Proof

Since T is sufficient for θ, $h(t) = E(W|T = t)$ does not depend on θ and so $W^\star = h(T)$ is a statistic with $E_\theta(W^\star) = E_\theta(E(W|T)) = E_\theta(W) = g(\theta)$, i.e. $W^\star$ is an unbiased estimator of $g(\theta)$. We note also that

$$\begin{aligned} Var_\theta(W) &= Var_\theta(E(W|T)) + E_\theta(Var(W|T)) \\ &\geq Var_\theta(E(W|T)) = Var_\theta(W^\star) \end{aligned}$$

and so $Var_\theta(W) \geq Var_\theta(W^\star)$

Asymptotic Distribution

Let us consider the MLE $\hat{\theta}$ of θ, to make notations clear, let us assume the true value of θ is θ_0. We shall prove that as the sample size n is very big, the distribution of MLE estimator is approximately normal with mean θ_0 and variance $1/[nI(\theta_0)]$. Since this is merely a limiting result, which holds as the sample size tends to infinity, we say that the MLE is *asymptotically unbiased* and refer to the variance of the limiting normal distribution as *the asymptotic variance* of the MLE.

Asymptotic Distribution

Theorem Let $X_1, \ldots, X_n$ be a sample of size n from a distribution for which the pdf is $f(x|\theta)$, with θ the unknown parameter. Assume that the true value of θ is θ_0, and the MLE of θ is $\hat{\theta}$. Then the probability distribution of $\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0)$ tends to a standard normal distribution. In other words, the asymptotic distribution of $\hat{\theta}$ is

$$N\left(\theta_0, \frac{1}{nI(\theta_0)}\right)$$

Reminder: Taylor Expansion

Taylors Theorem

If f is a function continuous and n times differentiable in an interval $[x, x+h]$, then there exists some point in this interval, denoted by $x + \lambda h$ for some $\lambda \in [0, 1]$, such that

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h}{2}f''(x) + \ldots + \frac{h^{n-1}}{(n-1)!}f^{(n-1)}(x) + \\ &+ \ldots + \frac{h^n}{n!}f^{(n)}(x+\lambda x) \end{aligned}$$

Asymptotic normality of MLE: Proof

Since MLE $\hat{\theta}$ is maximizer of $logL(\theta) = \sum_{i=1}^{n} log\, f(X_i|\theta)$, we have $(logL)'(\hat{\theta}) = 0$
By Taylor expansion we have:

$$0 = (logL)'(\hat{\theta}) = (logL)'(\theta_0) + (logL)''(\theta_0)(\hat{\theta} - \theta_0)$$

so that

$$\sqrt{n}(\hat{\theta} - \theta_0) = -\frac{\sqrt{n}(logL)'(\theta_0)}{(logL)''(\theta_0)}$$

$$(logL)'(\theta_0) = \sum_i \frac{\partial log\, f(x|\theta_0)}{\partial \theta}$$

Asymptotic normality of MLE: Proof

For the central limit theorem

$$\frac{1}{\sqrt{n}} log L'(\theta_0) \longrightarrow N(0, I(\theta_0))$$

For the WLLN

$$L''_n(\theta_0) \longrightarrow I_n(\theta_0)$$

Asymptotic normality of MLE:Proof

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

$$(\hat{\theta} - \theta_0) \longrightarrow_d N\left(0, \frac{1}{I_n(\theta_0)}\right)$$

Delta Method

Let $Y_n, \quad n = 1, 2, \ldots,$ be a sequence of random variables such that

$$\sqrt{n}\,(Y_n - \theta) \longrightarrow_d N(0, \sigma^2)$$

Furthermore let $g()$ be a twice differentiable real function defined on R such that $g'(\theta) \neq 0$. Then

$$\sqrt{n}\,(g(Y_n) - g(\theta)) \longrightarrow_d N(0, \left(g'(\theta)\right)^2 \sigma^2)$$

Approximation of Mean and Variance

Let X a r.v. with $E(X) = \mu_X$, and $Var(X) = \sigma^2_X$.

Theorem

The following approximations hold:

$$E(g(X)) \approx g(\mu)$$

and

$$Var(g(X)) \approx [g'(\mu)]^2 \sigma^2_X$$

The theorem follows from the Taylor approximation of the function g in the point μ:

$$g(x) = g(\mu) + g'(\mu)(x - \mu)$$

Approximation of Mean and Variance

Let X a r.v. with $E(X) = \mu_X$, and $Var(X) = \sigma_X^2$.

Theorem

The following approximations hold:

$$E(g(X)) \approx g(\mu)$$

and

$$Var(g(X)) \approx [g'(\mu)]^2 \sigma_X^2$$

The theorem follows from the Taylor approximation of the function g in the point μ:

$$g(x) = g(\mu) + g'(\mu)(x - \mu)$$

Numerical algorithms

- Sometimes it is not possible to find an explicit solution of the *likelihood equations* and so we have to use iterative algorithms to maximize $l(\theta)$, as the Newton-Raphson or the Fisher-scoring, which at any iteration update the parameter θ in appropriate way until convergence.
- Let $\theta^{(t)}$ denote the value of θ after the $t - th$ iteration of the algorithm. Then at the $(t+1) - th$ iteration **the Newton-Raphson** updates θ as:

$$\theta^{(t+1)} = \theta^{(t)} + \left[\mathbf{J}(\theta^{(t)})\right]^{-1} s(\theta^{(t)})$$

where $s(\theta^{(t)})$ is the score vector, i.e. the vector of the first derivatives of $l(\theta)$, and $[\mathbf{J}(\theta)]$ is the observed information.

Numerical algorithms

- The **Fisher-scoring** is a variant of the Newton-Raphson that updates θ as

$$\theta^{(t+1)} = \theta^{(t)} + \left[\mathbf{I_n}(\theta^{(t)})\right]^{-1} s(\theta^{(t)})$$

  where $\mathbf{I_n}(\theta^{(t)})$ is the expected information matrix or Fisher information.
- Through these algorithm it is usually possible to find a local maximum of $I(\theta)$, but, in general, there is not guarantee that it corresponds also to a global maximum. At this regard, a crucial point is that of the choice of the starting value of the algorithm, $\theta^{(0)}$.