5 Likelihood inference and first order asymptotics

- 5.1 Why asymptotics
- 5.2 Convergence concepts in asymptotics
- 5.3 Consistency and asymptotic normality of MLE
- 5.4 Additional comments on asymptotic properties of MLE
- 5.5 Delta method

5.1 Why asymptotics

So far we have focused on obtaining unbiased estimators for parameters with very strong finite-sample optimality properties. The existence of unbiased estimators with uniformly minimum variance, the so-called UMVUE class of estimators, needed the sufficient statistics of the statistical model to be complete.

Also, the derivation of the UMVUE is tricky sometimes, especially when the Cramér–Rao bound is not attainable.

Fortunately, methods such as the MLE are easier and more universally applicable and they share some of the optimality properties of the UMVUE asymptotically, i.e., when the sample size n is large.

Let us now recall the definition of the Likelihood Function.

Let X denote the variable of interest with pdf $f(x, \theta)$, where $\theta \in \mathbb{R}^k$. Next we denote by $(X_1, X_2, \ldots, X_n)$ a random sample of size n drawn on X. The pdf of a random sample $f(x_1, x_2, \ldots, x_n, \theta)$ can be obtained from the knowledge we have about f and, therefore, it depends on θ.

Then the Likelihood Function is mathematically equivalent to the joint density function of the sample. However, instead of being a function of the X_i's, we treat it as a function of θ after the observed values $X_i = x_i$ of the sample have been substituted.

We will always assume that X_i's are mutually independent. Then the Likelihood Function can be obtained:

X_i's iid

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

There are essentially two different settings when trying to derive MLEs.

1) Regular case:

MLE $\arg\max_{\theta} L(X,\theta)$

This case occurs when the log-likelihood function is differentiable with respect to the parameter, the MLE will be a stationary point, i.e., we need to solve the Equation

$$V(\mathbf{x};\theta) = \frac{\partial}{\partial\theta}\log L(\mathbf{x},\theta) = 0$$

$\frac{\partial^2}{\partial\theta^2}\log L(X,\theta)\Big|_{\hat\theta} < 0$

Then we need to investigate if this stationary point $\hat\theta$ indeed delivers a maximum. Typically we need to investigate the sign of the second derivative of the log-likelihood function at the stationary point.

When θ is a vector rather than a scalar parameter, we need to set the gradient of $\log L(\mathbf{x},\theta)$ to zero to find the stationary point and again include further arguments (by investigating the Hessian matrix) to argue that the stationary point maximises the log-likelihood.

$x^T H x < 0$ negative definite

Example 5.36

Let $X_1, X_2, \ldots, X_n$ be a sample (i.i.d.) from observations with support in $[0, 1]$ with the density function

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 < x < 1.$$

Here $\theta > 0$ is an unknown parameter to be estimated. Find the MLE of θ.

Solution:

1. Write down the Likelihood function:

$$L(\mathbf{x};\theta) = \theta^n \left(\prod_{i=1}^{n} x_i \right)^{\theta-1}. \quad = \prod_{i=1}^{n} \theta x_i^{\theta-1}$$

2. Write down the log-Likelihood function:

$$\ln L(\mathbf{x};\theta) = n\ln\theta + (\theta - 1)\sum_{i=1}^{n}\ln x_i.$$

3. Calculate the partial derivative and set it to 0 to find a stationary point:

$$\frac{\partial}{\partial\theta}\ln L(\mathbf{X};\theta) = \frac{n}{\theta} + \sum_{i=1}^{n}\ln x_i = 0$$

which gives the solution

$$\hat{\theta}_{mle} = \frac{-n}{\sum_{i=1}^{n}\ln x_i}.$$

Note that, as expected, $\hat{\theta} > 0$ is positive.

4. Find the second partial derivative

$$\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{x};\theta) = -\frac{n}{\theta^2}.$$

At the stationary point we get the value

$$-\frac{n}{\hat{\theta}^2} < 0,$$

which implies that the stationary point $\hat{\theta}$ delivers a maximum to the log-likelihood function, i.e., it is the MLE.

Exercise 5.20 (at lecture)

Let $X \sim \text{Poisson}(\lambda)$, where parameter $\lambda > 0$ and pmf

$$f(x, \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \{0, 1, 2, \dots\}.$$

Let $(X_1, \dots, X_n)$ denote a size n random sample drawn on X. Derive $\hat{\lambda}$, the MLE of λ.

$\hat{\lambda} = \bar{X}$

1) Irregular Case:

Sometimes the likelihood function $L(\mathbf{x}, \theta)$ is not smooth (and is possibly discontinuous) with respect to the parameter and it is not possible to calculate derivatives for each value of θ. In this case, we need to examine $L(\mathbf{x}, \theta)$ directly to find the argument $\hat{\theta}$ that maximises it.

Example 5.37

Typically, irregular cases happen when the support of the distribution depends on the unknown parameter. Consider deriving the MLE for the parameter θ of the continuous uniform distribution in $[0, \theta]$ by using a sample of n observations from this distribution. Show that the MLE is equal to the maximal of the observations in the sample, i.e., $\hat{\theta} = x_{(n)}$.

Solution

$$L(x,\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{(x_i,\infty)}(\theta)$$

Clearly the right-hand side of the support depends on the parameter. We have investigated $L(\mathbf{x},\theta)$ before. Using indicator function notation, it was equal to:

$$L(\mathbf{x},\theta) = \frac{1}{\theta^n} I_{(x_{(n)},\infty)}(\theta)$$

Considered as a function of θ, this is a discontinuous function. It is equal to zero when $\theta < x_{(n)}$. From $x_{(n)}$ onwards, $L(\mathbf{x},\theta)$ just coincides with the function $\frac{1}{\theta^n}$. Since this is monotonically decreasing in θ, the maximum of $L(\mathbf{x},\theta)$ over the interval $[0,\infty)$ is attained at $\hat{\theta} = x_{(n)}$.

That is, the maximal of the n observations in the sample is the MLE of θ in this example.

$$\prod_{i=1}^{n} I_{(-\infty, x_i)}(\theta) = I_{(-\infty, x_{(1)})}(\theta)$$

Exercise 5.21 (at lecture)

Let $X_1, X_2, \ldots, X_n$ be a sample from the density

$$f(x;\theta) = \theta x^{-2} I_{[\theta,\infty)}(x) \qquad \theta < x < \infty$$

where $\theta > 0$. Find the MLE of θ.

Solution:

The likelihood function is:

$$L(\mathbf{X};\theta) = \theta^n \prod_{i=1}^{n} x_i^{-2} I_{[\theta,\infty)}(x_{(1)})$$

Note that now the smallest of the n observations $(x_{(1)})$ comes into play. Indeed, since now θ is the left-hand side end of the support, we need all observations to be not smaller than θ, which logically implies that even the smallest of them should be not smaller than θ.

We consider L as a function of θ after the sample has been substituted. When θ moves on the positive half-axis, this function first grows monotonically (when θ moves between 0 and $x_{(1)}$) and then drops to zero onward since the indicator becomes equal to zero. Hence L is a discontinuous function of θ and its maximum is attained at $x_{(1)}$. This means that $\hat{\theta}_{mle} = X_{(1)}$.

5.2 Convergence concepts in asymptotics

We realised that finding the UMVUE for a fixed sample size n could be difficult in some cases especially when the CR bound is not attainable.

Finding them requires some skill, and there is no easy-to-follow constructive algorithm for their determination.

On the other hand, the MLEs are typically easy to construct by following a general recipe of optimising either directly the Likelihood or the log-likelihood function, i.e., by following an easy general recipe.

Both the regular and the irregular case for deriving MLE are clearly formulated algorithmically and typically are easy to deal with.

It should be pointed out that sometimes the MLE could be biased or, even if unbiased, could not attain the CR bound when outside the exponential family setting.

Yet, it is simpler to work with the MLEs and, as shown in the many examples below, usually the UMVUE is just a bias-corrected MLE.

Lets consider some examples to confirm this statement. These examples have either already been discussed and are just summarised here, or are left as an exercise for you.

- UMVUE for the variance of Bernoulli trials was $\bar{X}(1-\bar{X})\frac{n}{n-1}$ whereas the MLE is $\bar{X}(1-\bar{X})$.
- UMVUE for the endpoint θ of uniform $(0,\theta)$ distribution was $\frac{n+1}{n}X_{(n)}$ whereas the MLE is $X_{(n)}$.
- UMVUE for the probability of no occurrence based on n independent Poisson random variables: $(1-\frac{1}{n})^{n\bar{X}}$ whereas the MLE is $\exp(-\bar{X})$.

The bias-correction itself tends to be negligible as the sample size increases. Therefore the UMVUEs are either MLEs or almost MLEs. Hence, it is justified to look for a strong backing of the properties of MLEs in a general setting.

This can be done using asymptotic arguments, i.e. by looking at the performance of MLEs when $n \to \infty$, i.e. by letting the amount of information become arbitrarily large.

> Statistical folklore says then that "nothing can beat the MLE asymptotically".

When trying to defend the MLE on asymptotic grounds, we need some concepts about convergence of random variables and random vectors (since the MLE is a random variable in the univariate case and a random vector in the multivariate case).

We briefly discuss some stochastic convergence concepts first. An estimator T_n of the parameter θ is said to be:

i) consistent (or *weakly consistent*) if

$$\lim_{n\to\infty} P_\theta(|\, T_n - \theta \,| > \epsilon) = 0$$

$$\lim_{n\to\infty} P_\theta(|T_n - \theta| < \varepsilon) = 1$$

for all $\theta \in \Theta$ and for every fixed $\epsilon > 0$. We denote this by $T_n \xrightarrow{P} \theta$ and say that T_n tends to θ in probability.

In other words, considering that the complementary probability $P_\theta(|\, T_n - \theta \,| \leq \epsilon)$ tends to one when $n \to \infty$, we can say that the consistency implies that the probability of the estimator being in an ϵ neighbourhood of the true unknown parameter tends to one no matter how small the $\epsilon > 0$, as long as the sample size is large enough.

ii) strongly consistent if

$$P_\theta\left\{\lim_{n\to\infty} T_n = \theta\right\} = 1 \qquad \text{for all } \theta \in \Theta$$

iii) mean-square consistent if

$$\text{MSE}_\theta(T_n) \xrightarrow{n\to\infty} 0 \qquad \text{for all } \theta \in \Theta$$

$\text{Var}(T_n) + b_n(T_n)^2 \to 0$ as $n \to \infty$

Example 5.38

For a random sample of size n from a $N(\theta, \sigma^2)$ population, the sample mean $\bar{X}_n$ is proposed as an estimator of θ. In this example we will show that $\bar{X}_n$ is a consistent estimator of θ.

Solution:

We know that the exact distribution of $\bar{X}_n$ is $N(\theta, \frac{1}{n}\sigma^2)$. Using standardisation, this implies that

$$Z = \sqrt{n}\frac{\bar{X} - \theta}{\sigma}$$

is standard normally distributed $(Z \sim N(0, 1))$.

If Φ denotes the cdf of the standard normal distribution then we have

$$\begin{aligned} P_\theta(|\bar{X}-\theta| > \epsilon) &= 2P_\theta\left(\frac{\sqrt{n}(\bar{X}-\theta)}{\sigma} > \frac{\sqrt{n}\epsilon}{\sigma}\right) \\ &= 2P\left(Z > \frac{\sqrt{n}\epsilon}{\sigma}\right) \\ &= 2\left(1-\Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right)\right) \end{aligned}$$

Since $\Phi(\frac{\sqrt{n}\epsilon}{\sigma}) \to 1$ as $n \to \infty$ we get the consistency statement

$$P_\theta(|\bar{X}-\theta| > \epsilon) \to 0.$$

Hence, $\bar{X}_n \xrightarrow{P} \theta$ and $\bar{X}_n$ is consistent for θ.

Remark 5.13

It is important to note that if the estimator is mean-square consistent then it is also consistent. This relationship probably has the most important practical consequence.

The reason is that, most often, we are interested in weak consistency and a common method that often works in proving it, is by showing mean-square consistency first.

Note also that strong consistency implies weak consistency.

To justify the relation between mean-square consistency and consistency we can use the **Chebyshev Inequality**.

The Chebyshev Inequality states that for any random variable X and any $\epsilon > 0$ it holds for the k-th moment:

$$P(|X| > \epsilon) \leq \frac{\mathbb{E}(|X|^k)}{\epsilon^k}.$$

Applying this inequality for X being $T_n - \theta$ and $k = 2$ we get

$$0 \leq P(|T_n - \theta| > \epsilon) \leq \frac{MSE_\theta(T_n)}{\epsilon^2}.$$

Therefore, if an estimator T_n is mean-square consistent and the right-hand side tends to zero, the left-hand side will also tend to zero thus implying consistency.

Example 5.39 (more discussion at lecture)

Let $X \sim \text{Uniform}(0,1)$ with pdf $f(x) = 1$ for $0 < x < 1$.

Solution:

Let parameter $\theta \in (0,1)$. Additionally, express $\bar{X}_n$ in power sum notation, namely: $\bar{X}_n = \frac{s_1}{n}$ where $s_1 = \sum_{i=1}^{n} X_i$. Then

$$\text{MSE} = E\left[\left(\frac{s_1}{n} - \theta\right)^2\right]$$

is the second raw moment of $(\frac{s_1}{n} - \theta)$. The MSE can be written as:

$$\theta^2 - 2\theta\acute{\mu}_1 + \frac{(n-1)\acute{\mu}_1^2}{n} + \frac{\acute{\mu}_2}{n}$$

which is the solution expressed in terms of the moments of the population of X, namely $\acute{\mu}_1(X)$ and $\acute{\mu}_2(X)$.

Now, we evaluate the MSE as follows:

$$\theta^2 - \theta + \frac{1}{12n} + \frac{1}{4}.$$

Then the right-hand side of Chebyshev Inequality is

$$\frac{1}{\epsilon^2}(\theta^2 - \theta + \frac{1}{12n} + \frac{1}{4}),$$

when $X \sim Uniform(0,1)$. By taking the limits on both sides of Chebyshev Inequality:

$$\lim_{n\to\infty} P(|\bar{X}_n - \theta| > \varepsilon) \leq \frac{\theta^2 - \theta + \frac{1}{4}}{\varepsilon^2}$$

It follows from the definition of convergence in probability that

$$\bar{X}_n \xrightarrow{P} \frac{1}{2}$$

and ensures, due to uniqueness, that $\bar{X}_n$ cannot converge in probability to any other point in the parameter space. $\bar{X}_n$ is a consistent estimator of $\theta_0 = \frac{1}{2}$.

Additionally, we have $E[X] = 1/2$. Thus, the sample mean $\bar{X}_n$ is a consistent estimator of the population mean.

The above example is suggestive of a more general result encapsulated in a set of theorems known as the Law of Large Numbers.

Theorem 5.19 (Law of Large Numbers)

Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of mutually independent and identically distributed random variables with finite mean μ. The sample mean:

$$\bar{X}_n \xrightarrow{p} \mu.$$

There is one more form of convergence of random variables. It is called convergence in distribution and is the weakest form of convergence.

It follows from any of the three convergences discussed above. Not surprisingly it is called a weak convergence (or convergence in distribution).

Assume that the sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ have cumulative distribution functions $F_1, F_2, \ldots, F_n, \ldots$ respectively. Assume the continuous random variable X has a cdf F and that it holds for each argument $x \in R$ that

$$\lim_{n\to\infty} F_n(x) = F(x).$$

Then we say that the sequence of random variables $\{X_n\}, n = 1, 2, \ldots$ converges weakly (or in distribution) to X and denote this fact by

$$X_n \xrightarrow{d} X.$$

In practice, however, the distribution of a sequence of random variables is often not available. In this case, if the variables in the sequence are used to form sums and averages, the limiting distribution can often be derived by applying the Central Limit Theorem.

Theorem 5.20 (Central Limit Theorem)

Let the random variables in the sequence $\{X_n\}_{n=1}^{\infty}$ be independent and identically distributed, each with finite mean μ and finite variance σ^2. Then the random variable

$$\frac{S_n - \mu n}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}$$

converges in distribution to a random variable $Z \sim N(0,1)$ where

$$S_n = \sum_{k=1}^{n} X_i = n\bar{X}.$$

Example 5.40

Let $X \sim \text{Uniform}(0,1)$, the Uniform distribution on the interval $(0,1)$ with pdf $f(x) = 1$ for $0 < x < 1$. Let $\bar{X}_{10}$ denote the sample mean of a random sample of size $n = 10$ collected on X. Now suppose that we wish to obtain the probability:

$$p = P\left(\frac{1}{6} < \bar{X}_{10} < \frac{4}{6}\right).$$

Solution: The mean μ and the variance σ^2 of X are, respectively:

$$\mathbb{E}(X) = \int_0^1 x \cdot 1 \mathrm{d}x = \frac{1}{2}$$

$$Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \int_0^1 x^2 \cdot 1 \mathrm{d}x - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

As the conditions of the Central Limit Theorem are satisfied, the asymptotic distribution of $\bar{X}_{10}$ is:

$$\bar{X}_{10} \overset{a}{\sim} N\left(\frac{1}{2}, \frac{1}{120}\right).$$

We may therefore use this asymptotic distribution to find an approximate solution for p.

$$
\begin{aligned}
p &= P\left(\frac{\frac{1}{6} - \frac{1}{2}}{\sqrt{\frac{1}{120}}} < \frac{\bar{X}_{10} - \frac{1}{2}}{\sqrt{\frac{1}{120}}} < \frac{\frac{4}{6} - \frac{1}{2}}{\sqrt{\frac{1}{120}}} \right) \\
&= P\Big(-3.651 < Z < 1.826 \Big) \\
&= \Phi(1.826) - \Phi(-3.651) \\
&= 0.9659
\end{aligned}
$$

where $\Phi(x)$ is the cdf of a standard normal evaluated at x.

```
(p <- pnorm(1.826) - pnorm(-3.651))
```

```
#> [1] 0.9659443
```

5.3 Consistency and asymptotic normality of MLE

There exist sets of regularity conditions that, if satisfied, permit us to make relatively straightforward statements about the asymptotic properties of the MLE. The following theorem summarises the basic statement about asymptotic properties of MLEs.

Theorem 5.21

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be i.i.d. from $f(x, \theta), \theta \in \Theta \in \mathbb{R}^1$, Θ– open interval. Assume, following regularity conditions are satisfied:

1) $$\frac{\partial f}{\partial \theta}(x, \theta), \frac{\partial^2 f}{\partial \theta^2}(x, \theta), \frac{\partial^3 f}{\partial \theta^3}(x, \theta)$$ exist for all x and all $\theta \in \Theta$.

2) $$\frac{\partial}{\partial \theta} \int f(x, \theta) dx = \int \frac{\partial}{\partial \theta} f(x, \theta) dx$$

and

$$\frac{\partial^2}{\partial \theta^2} \int f(x, \theta) dx = \int \frac{\partial^2}{\partial \theta^2} f(x, \theta) dx$$

3) For the information in a single observation $I(\theta)$ we have

$$0 < I(\theta) = E_\theta\left(\frac{\partial \ln f}{\partial \theta}(x, \theta)^2\right) < \infty$$

for all $\theta \in \Theta$.

Theorem 5.22 (Theorem cont.)

4)

$$\left|\frac{\partial^3 \ln f}{\partial \theta^3}(x,\theta)\right| \leq H(x)$$

for all $\theta \in \Theta$ with $E_\theta H(X) = \int H(x) f(x,\theta) dx \leq C$, C not depending on $\theta \in \Theta$.

Let θ_0 be the "true" value of θ. Then the MLE $\hat{\theta}_n$ of θ_0 is strongly consistent and asymptotically normal, i.e.

a) $P_{\theta_0}(X : \hat{\theta}_n \to \theta_0) = 1$

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

For large n, this can also be written roughly as:

$$\hat{\theta} \approx N(\theta_0, \frac{1}{nI(\theta_0)})$$

Proof.

a) Consistency

First, we notice that

$$\frac{1}{n}\log L(X,\hat{\theta}_n) \geq \frac{1}{n}\log L(X,\theta_0) \tag{8}$$

holds due to the definition of $\hat{\theta}_n$ as a maximizer of $\log L$ where $L(.,.)$ denotes the joint density of n independent identically distributed (i.i.d.) observations, each with a density $f(.,.)$.

On the other hand, Jenson's inequality implies

$$\begin{aligned}\mathbb{E}_{\theta_0}\left[\log \frac{L(X,\theta)}{L(X,\theta_0)}\right] &< \log \mathbb{E}_{\theta_0}\left[\frac{L(X,\theta)}{L(X,\theta_0)}\right] \\ &= \log \int \cdots \int \frac{L(X,\theta)}{L(X,\theta_0)} L(X,\theta_0) dX \\ &= \log 1 \\ &= 0\end{aligned}$$

which implies that

$$\mathbb{E}_{\theta_0}[\log L(X,\theta)] - \mathbb{E}_{\theta_0}[\log L(X,\theta_0)] < 0$$

or equivalently

$$\mathbb{E}_{\theta_0}\left[\frac{1}{n}\log L(X,\theta)\right] < \mathbb{E}_{\theta_0}\left[\frac{1}{n}\log L(X,\theta_0)\right].$$

Since

$$\frac{1}{n}\log L(X,\theta) = \frac{1}{n}\sum_{i=1}^{n}\log f(X_i,\theta)$$

is an *average* and averages converge with probability one to the true means (the Law of Large numbers) it implies that for a fixed $\theta \neq \theta_0$ we should have

$$P_{\theta_0}\left\{\lim_{n\to\infty}\frac{1}{n}\log L(X,\theta) < \lim_{n\to\infty}\frac{1}{n}\log L(X,\theta_0)\right\} = 1 \qquad (9)$$

Comparing (8) and (9) we see that we need to have

$$P_{\theta_0}(X : \hat{\theta}_n \to \theta_0) = 1.$$

and this means strong consistency!

b) Asymptotic normality:

Now $\hat{\theta}_n$ is known to be "near" θ_0 due to the consistency statement.

Since

$$0 = \frac{\partial}{\partial\theta} \log L(\mathbf{X}, \theta)\Big|_{\theta=\hat{\theta}_n} = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i, \theta)\Big|_{\theta=\hat{\theta}_n}$$

holds, after Taylor expansion around θ_0 we get

$$\begin{aligned} 0 = & \sqrt{n} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i, \hat{\theta}_n) \\ = & \sqrt{n} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i, \theta_0) + \sqrt{n}(\hat{\theta}_n - \theta_0) \sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f(X_i, \theta_0) \\ & + \sqrt{n} \frac{(\hat{\theta}_n - \theta_0)^2}{2} \sum_{i=1}^{n} \frac{\partial^3}{\partial\theta^3} \log f(X_i, \theta_0) \end{aligned}$$

Here, we write $\eta_i H(X_i)$ for each summand that involved third order derivatives with $|\eta_i| < 1$.

Then upon a recollection of terms we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(-\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\partial}{\partial\theta} \log f(x_i, \theta_0)) / [\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log f(x_i, \theta_0)]}{1 + \frac{1}{2}(\hat{\theta}_n - \theta_0) \frac{\frac{1}{n} \sum_{i=1}^n \eta_i H(x_i)}{\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2} \log f(x_i, \theta_0)}}.$$

Now we just have to use the following facts:

i) consistency which implies

$$\hat{\theta}_n - \theta_0 \xrightarrow{w.p.1} 0.$$

ii) the Law of large numbers regarding the term

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)\right] \to -I_{X_1}(\theta),$$

iii) the central limit theorem regarding the term

$$-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta_0) \xrightarrow{d} N(0, I_{X_1}(\theta))$$

iv) the uniform bound assumption 4 of the theorem

$$\frac{1}{n}\sum_{i=1}^{n}|\eta_i H(X_i)| \leq \frac{1}{n}\sum_{i=1}^{n} H(X_i) \to \mathbb{E}_{\theta_0} H(X_i) \leq C$$

v) and the property that

$$\mathrm{Var}(aX) = a^2 \mathrm{Var}(X).$$

to finish the argument that

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

□

In this example we will check the regulatory conditions introduced in the asymptotic normality theorem (I will never get you to undertake such an onerous task).

Example 5.41

Let θ_0 be the "true" value of θ. We will show that MLE $\hat{\theta}$ of θ_0 is strongly consistent and asymptotically normal. We will consider the following pdf

$$f(x, \theta_0) = \theta_0 x^{\theta_0 - 1},$$

for $\theta_0 > 0$ and $0 < x < 1$.

Solution:

1. $\frac{\partial f}{\partial \theta}(x,\theta)$, $\frac{\partial^2 f}{\partial \theta^2}(x,\theta)$, $\frac{\partial^3 f}{\partial \theta^3}(x,\theta)$ exist for all x and all $\theta \in \Theta$?

Here we have

$$\frac{\partial f}{\partial \theta}(x,\theta) = x^{\theta_0 - 1} + \theta_0 x^{\theta_0 - 1} \log(x)$$
$$\frac{\partial^2 f}{\partial \theta^2}(x,\theta) = \theta_0 x^{\theta_0 - 1} \log^2(x) + 2x^{\theta_0 - 1} \log(x),$$
$$\frac{\partial^3 f}{\partial \theta^3}(x,\theta) = \theta_0 x^{\theta_0 - 1} \log^3(x) + 3x^{\theta_0 - 1} \log^2(x)$$

which exist for all $0 < x < 1$ and $\theta_0 > 0$.

2. $$\frac{\partial}{\partial\theta}\int f(x,\theta)dx = \int \frac{\partial}{\partial\theta} f(x,\theta)dx \quad \text{and} \quad \frac{\partial^2}{\partial\theta^2}\int f(x,\theta)dx = \int \frac{\partial^2}{\partial\theta^2} f(x,\theta)dx?$$

Here we have

$$\frac{\partial}{\partial\theta}\int f(x,\theta)dx = \int \frac{\partial}{\partial\theta} f(x,\theta)dx = x^{\theta}\log(x)$$

and

$$\frac{\partial^2}{\partial\theta^2}\int f(x,\theta)dx = \int \frac{\partial^2}{\partial\theta^2} f(x,\theta)dx = x^{\theta}\log^2(x)$$

Hence this is also satisfied.

3

$$0 < I(\theta) = E_\theta\left(\frac{\partial ln f}{\partial \theta}(x,\theta)^2\right) < \infty \quad \text{for all } \theta \in \Theta$$

We have that

$$I(\theta) = \frac{1}{\theta_0^2}$$

which is finite.

4

$$\left|\frac{\partial^3 \ln f}{\partial \theta^3}(x,\theta)\right| \leq H(x) \quad \text{for all } \theta \in \Theta$$

with

$$\mathbb{E}_\theta H(X) = \int H(x) f(x,\theta) dx \leq C,$$

C not depending on $\theta \in \Theta$.

Here we have

$$\left|\frac{\partial^3 ln f}{\partial \theta^3}(x,\theta)\right| = \frac{2}{\theta_0^3}$$

which is non-stochastic.

In conclusion, all regulatory conditions are satisfied and the MLE $\hat{\theta}$ is consistent for θ_0.

Additionally,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \theta_0^2)$$

and

$$\hat{\theta} \approx N\left(\theta_0, \frac{1}{n}\theta_0^2\right).$$

Remark 5.14

The statement of the above theorem can be extended to the *multivariate case.* This is, of course, a crucial step regarding practical applications of the maximum likelihood methodology since in the majority of cases, the parameter vector of interest is multi-dimensional.

Multidimensional case:

Let now $f(x, \theta)$, $\theta \in \Theta \in \mathbb{R}^p$. To formulate it, we need to extend the notion of Fisher information in a parameter-**vector** $\overrightarrow{\theta}$. For such a vector, we define a Fisher information **matrix in the whole sample** $I_{\mathbf{X}}(\overrightarrow{\theta})$ whose (i, j)-th element is defined as

$$\mathbb{E}\left(\frac{\partial}{\partial\theta_i}\log \mathbf{L}\frac{\partial}{\partial\theta_j}\log \mathbf{L}\right) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log \mathbf{L}\right),$$

for $i = 1, \ldots, p$ and $j = 1, \ldots, p$.

For simplicity of notation, we usually skip the arrow over the parameter even though we are in the multidimensional case

Then, for an inner point θ_0 (the "true value" of the parameter space Θ) we have under some regularity conditions on the density (similar to the ones listed in Theorem 5.21):

a)

$$P_{\theta_0}(X : \hat{\theta}_n \xrightarrow[n\to\infty]{} \theta_0) = 1$$

b)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_{X_1}^{-1}(\theta_0)) \qquad \text{(asymptotic normality)}$$

where $I_{X_1}^{-1}(\theta_0)$ is the information in **one** observation and that $I_{X_1}^{-1}(\theta_0) = nI_{\mathbf{X}}^{-1}(\theta_0)$ holds. Hence, result b) can be also written roughly as

$$\hat{\theta}_n \approx N(\theta_0, I_{\mathbf{X}}^{-1}(\theta_0)) \approx N\Big(\theta_0, \frac{1}{n} I_{X_1}^{-1}(\theta_0)\Big).$$

Even more can be said. It turns out that the limiting $p \times p$ symmetric variance-covariance matrix $I_{X_1}^{-1}(\theta_0)$ in b) is the *smallest possible*. This is to be interpreted in the sense that any other limiting matrix A_{θ_0} (related to another possible estimator) is such that the difference

$$A_{\theta_0} - I_{X_1}^{-1}(\theta_0) \tag{10}$$

is non-negative definite, that is, has non-negative eigenvalues. We also denote this as

$$A_{\theta_0} \geq I_{X_1}^{-1}(\theta_0).$$

5.4 Additional comments on asymptotic properties of MLE

We indicate how the above result can be interpreted as "asymptotic efficiency" of the MLE. For simplicity, start with a *one-dimensional parameter*. Formally, the asymptotic normality of the MLE and the form of the asymptotic variance show that

$$\lim_{n\to\infty} \mathrm{Var}(\hat{\theta}_{mle}).[nI_{X_1}(\theta)] = 1$$

for all $\theta \in \Theta$ which means that the MLE "asymptotically achieve the CR bound on variance".

In fact: there are additional obstacles in formulating such a claim but we will skip over these.

But for those interested:

First, the asymptotic normality claim is about *convergence in distribution* (which was the weakest type of convergence) and it does not immediately follow from this result that also the variances converge.

Second, the existence of the so-called *superefficiency phenomena* (first example of a superefficient estimator has been suggested by Hodges).

Superefficient estimators: it is possible to construct estimators for which the above limit can be even less than one *for a certain small number of θ values*.

But by imposing further reasonable regularity conditions and a suitable interpretation of the convergence to the asymptotic distribution, the obstacles can be overcome.

We finish our discussion with the words (since the above difficulties can be overcome):

> "Folklore says that MLE are asymptotically the best (*asymptotically efficient*) estimators meaning that they are asymptotically unbiased and with the smallest possible asymptotic variance".

The asymptotic efficiency in the case of *multi-dimensional* parameter vector is interpreted in a similar way. Using the fact that the MLE is asymptotically centered at the "true value" θ_0 and asymptotically is less spread around this true value than any other of its competitors because of (10) we can claim that the MLE is asymptotically efficient.

5.5.1 Invariance property of the MLE

The main theorem about asymptotic properties of MLE was related to the estimation of the parameter θ itself. Sometimes, a certain smooth function (a transformation) of the parameter θ is of interest to us, as we already had a chance to see earlier in this course. If we denote by $h(\theta)$ such a transformation, it is useful to know two things:

- the MLE of the new parameter $h(\theta)$
- the asymptotic distribution of the MLE of the new parameter $h(\theta)$

The answer to the first question shows one of the very useful properties of the MLE. The claim is that the MLE of $h(\theta)$ can be obtained by substitution (plug-in) of the MLE $\hat{\theta}$ of θ in the transformation formula.

That is, $h(\hat{\theta})$ is the MLE of $h(\theta)$. This is the invariance or better to say transformation invariance property of the MLE.

5.5.2 Delta method

Assume that the transformation $h(\theta)$ is smooth enough. Then we are also able to find the asymptotic distribution of $h(\hat{\theta})$. This is a very important result called "the delta method".

Since the transformation is assumed to be smooth, we can expand $h(\hat{\theta})$ around the true parameter θ_0 (Taylor Series):

$$h(\hat{\theta}_{mle}) = h(\theta_0) + (\hat{\theta}_{mle} - \theta_0)\frac{\partial h}{\partial \theta}(\theta_0) + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2 . \frac{\partial^2 h(\theta_0)}{\partial \theta^2} + \dots$$

or by rearranging and ignroing higher order term we obtain

$$h(\hat{\theta}_{mle}) - h(\theta_0) = (\hat{\theta}_{mle} - \theta_0)\frac{\partial h}{\partial \theta}(\theta_0) + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2 . \frac{\partial^2 h(\theta_0)}{\partial \theta^2}$$

From here we get the convergence in distribution:

$$\sqrt{n}(h(\hat{\theta}_{mle}) - h(\theta_0)) \xrightarrow{d} N\left(0, \left[\frac{\partial h}{\partial \theta}(\theta_0)\right]^2 I^{-1}(\theta_0)\right)$$

This result is called **"the delta method"**. Roughly, we shall also say that the distribution of $h(\hat{\theta}_{mle})$ can be approximated by

$$N\left(h(\theta_0), \frac{1}{n}\left[\frac{\partial h}{\partial \theta}(\theta_0)\right]^2 I^{-1}(\theta_0)\right).$$

The delta method has a version applicable for the case where $h(\theta)$ is a smooth transformation of a p–dimensional parameter-vector $\overrightarrow{\theta}$.

If we introduce the vector of partial derivatives

$$\nabla h(\overrightarrow{\theta}) = \left(\frac{\partial}{\partial \theta_1} h(\overrightarrow{\theta}), \ldots, \frac{\partial}{\partial \theta_p} h(\overrightarrow{\theta}) \right)^\top$$

then the distribution of $h(\hat{\overrightarrow{\theta}}_{mle})$ can be approximated by

$$N\left(h(\overrightarrow{\theta}_0), \nabla h(\overrightarrow{\theta}_0)^\top I_{\mathbf{X}}(\overrightarrow{\theta}_0)^{-1} \nabla h(\overrightarrow{\theta}_0) \right).$$

where $I_{\mathbf{X}}(\overrightarrow{\theta}) = n I_{X_1}(\overrightarrow{\theta})$ and hence $I_{\mathbf{X}}(\overrightarrow{\theta})^{-1} = \frac{1}{n} I_{X_1}(\overrightarrow{\theta})^{-1}$ in agreement with the one-dimesional case $p = 1$.

5.5.3 Delta method examples

Example 5.42 (more details at lecture)

For estimating the parameter $\sqrt{\lambda}$ of the Poisson (λ) distribution using MLE, we get

$$\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{d} N(0, 1/4).$$

Interesting: although the asymptotic normal distribution of the MLE for λ has a variance that **depends** on the (unknown) λ itself, the asymptotic normal distribution of the *transformed* $h(\lambda) = \sqrt{\lambda}$ has a *constant* variance of $1/4$ independent of λ. Such a transformation $h(.)$ it termed a variance stabilising transformation.

Variance stabilising transformations are actively sought after especially for the purpose of constructing confidence intervals with a more precise coverage accuracy.

Example 5.43 (Asymptotic distribution of a ratio estimator)

Suppose X and Y are bivariate normally distributed with a known 2×2 covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}$$

and a mean vector $(\mu_X, \mu_Y)'$. A sample of n observation pairs $(X_i, Y_i)', i = 1, 2, \ldots, n$ is given. We are interested in the asymptotic distribution of the MLE $\bar{X}/\bar{Y}$ of the ratio μ_X/μ_Y.

Solution:

First we note that $h(\mu_X, \mu_Y) = \mu_X/\mu_Y$ and

$$\frac{\partial h}{\partial \mu_X} = \frac{1}{\mu_Y} \quad \text{and} \quad \frac{\partial h}{\partial \mu_Y} = \frac{-\mu_X}{\mu_Y^2}.$$

From the first order Taylor expansion we have $E(\bar{X}/\bar{Y}) \approx \mu_X/\mu_Y$. The inverse of the information matrix is

$$I_n(\mu_X, \mu_Y)^{-1} = \frac{1}{n} \begin{Bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{Bmatrix}$$

hence by aplying the delta method we get

$$\sqrt{n}\left(\frac{\bar{X}}{\bar{Y}} - \frac{\mu_X}{\mu_Y}\right) \xrightarrow{d} N\left(0, \left(\frac{1}{\mu_Y}, \frac{-\mu_X}{\mu_Y^2}\right) \begin{Bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{Bmatrix} \begin{Bmatrix} 1/\mu_Y \\ \frac{-\mu_X}{\mu_Y^2} \end{Bmatrix}\right)$$

Completing the matrix multiplication we get for the asymptotic variance the expression reduced to

$$\frac{\mu_X^2}{\mu_Y^2}\left(\frac{\sigma_{11}}{\mu_X^2}+\frac{\sigma_{22}}{\mu_Y^2}-2\frac{\sigma_{12}}{\mu_X\mu_Y}\right).$$

Another way in which we can state the same result is to say that

$$\frac{\bar{X}}{\bar{Y}} \approx N\left(\frac{\mu_X}{\mu_Y}, \frac{1}{n}\frac{\mu_X^2}{\mu_Y^2}\left(\frac{\sigma_{11}}{\mu_X^2}+\frac{\sigma_{22}}{\mu_Y^2}-2\frac{\sigma_{12}}{\mu_X\mu_Y}\right)\right).$$

Note: it would have been quite difficult to get exact closed-form expression for the variance whereas the delta method is routinely applicable!

Exercise 5.22 (at lecture)

Let $X_1, X_2, \ldots, X_n$ be a sample from the density function:

$$f(x;\theta) = \theta x^{\theta-1} I_{(0,1)}(x)$$

where $\theta > 0$.

i) Find the MLE of $\tau(\theta) = \frac{\theta}{1+\theta}$

ii) State the asymptotic distribution of the MLE of $\tau(\theta)$ in i).

Exercise 5.23 (at lecture)

Consider n i.i.d. observations from a Poisson (λ) distribution. Suppose the parameter of interest is $\tau(\lambda) = \frac{1}{\lambda}$.

i) What is the MLE of $\tau(\lambda)$?

ii) What is its variance?

iii) What is its asymptotic variance?

Exercise 5.24 (at lecture)

Let $X_1, X_2, \ldots, X_n$ be a sample from normal distribution $N(\mu, \sigma^2)$ where μ is known and σ^2 is the parameter to be estimated.

i) Find the MLE and state its asymptotic distribution.

ii) Assume now that σ is to be estimated. Find the MLE and state its asymptotic distribution.

Exact and asymptotic distributions for the deviance

The deviance is defined as:

$$D(\theta) = -2\log\left(\frac{L(\mathbf{X},\theta)}{L(\mathbf{X},\hat{\theta}_{MLE})}\right)$$

and has an important role in constructing confidence intervals and testing hypotheses about unknown parameters. It is a function of the observations, as well as of the (unknown) parameters of interest.

For a fixed value of the parameters, the deviance is only a function of the observations (i.e. statistic). Its distribution is of great interest because of the applications mentioned above.

Example 5.44 (details at lecture)

For n i.i.d. observations from a $N(\mu, \sigma^2)$ with σ^2 known, the deviance is

$$D(\mu) = \frac{n(\bar{x} - \mu)^2}{\sigma^2}.$$

Suppose the true mean is μ_0 so that

$$\bar{X} \sim N\Big(\mu_0, \frac{\sigma^2}{n}\Big).$$

Then $D(\mu_0) \sim \chi^2(1)$ (chi-squared with one degree of freedom) and this is an *exact* (not asymptotic) result.

Chi square approximation of the deviance for "any" smooth model $f(x, \theta)$:

Assume that $\theta = \theta_0$ is the "true" value of the population parameter. Expand the log-likelihood in Taylor series around $\theta = \hat{\theta}_{mle}$:

$$logL(\mathbf{X}, \theta_0) = \log L(\mathbf{X}, \hat{\theta}_{mle}) + (\theta_0 - \hat{\theta}_{mle}) \frac{\partial \log L(\mathbf{X}, \hat{\theta}_{mle})}{\partial \theta} + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2 \frac{\partial^2 \log L(\mathbf{X}, \hat{\theta}_{mle})}{\partial \theta^2} + \dots$$

Because the second summand in the RHS vanishes at $\hat{\theta}_{mle}$, and by ignoring higher order terms, we get:

$$D(\theta_0) = -2 \log \frac{L(X, \theta_0)}{L(X, \hat{\theta}_{\text{mle}})} \approx (\hat{\theta}_{mle} - \theta_0)^2 \left\{ - \frac{\partial^2 \log L(\boldsymbol{X}, \hat{\theta}_{mle})}{\partial \theta^2} \right\}.$$

Since $\hat{\theta}_{\text{mle}} \approx \theta_0$ and by applying the law of large numbers we have:

$$D(\theta_0) \approx n(\hat{\theta}_{\text{mle}} - \theta_0)^2 I_{X_1}(\theta_0)$$

But since

$$\sqrt{n}(\hat{\theta}_{\text{mle}} - \theta_0) \xrightarrow{d} N\Big(0, I_{X_1}^{-1}(\theta_0)\Big)$$

we get

$$D(\theta_0) \xrightarrow{d} \chi_1^2$$

but as an *approximation!*

Using the results about the normal approximation to the distribution of MLE, we get that the deviance has an asymptotic χ^2 distribution with one degree of freedom. More generally, if the parameter $\theta_0 \in R^p$ then, asymptotically,

$$(\theta_0 - \hat{\theta}_{mle})^\top \left\{ - \frac{\partial^2 \log L(\mathbf{X}, \hat{\theta}_{mle})}{\partial\theta\partial\theta^\top} \right\} (\theta_0 - \hat{\theta}_{mle}) \sim \chi^2_p$$

This is called the Wald statistic.

Hence, we can now suggest a test of the null hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$ with an asymptotic level equal to α. Discussion of this example will be conitnued later in the course under the heading "Generalized likelihood ratio tests".

Example 5.45

In this example we apply the above results in the case of the exponential distribution. Assume that $x_1, x_2, \ldots, x_n$ are i.i.d. realisations from the density

$$f(x, \theta) = \frac{1}{\theta} \exp(-\frac{x}{\theta}), \quad x > 0$$

Solution:

Likelihood:

$$L(\mathbf{x}, \theta) = \frac{1}{\theta^n} \exp(-\sum_{i=1}^{n} x_i/\theta)$$

The MLE is $\hat{\theta} = \bar{X}$ and the **exact** deviance is given by:

$$\begin{aligned} D(\theta) &= -2\left[\log \frac{1}{\theta_0^n} \exp\Big(-\frac{\sum_{i=1}^n X_i}{\theta_0}\Big) - \log \frac{1}{\bar{X}^n} \exp(-n)\right] \\ &= 2n \log \theta_0 + 2n\frac{\bar{X}}{\theta_0} - 2n \log \bar{X} - 2n \\ &= 2n\left[\frac{\bar{X}}{\theta} - \ln(\frac{\bar{X}}{\theta}) - 1\right] \end{aligned}$$

We can use $D(\theta_0)$ as a test statistic for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ by comparing $D(\theta_0)$ with $\chi^2_{1,\alpha}$ for a given α. For instance, when $\alpha = 0.05$ we have $\chi^2_{1,0.05} = 3.84$.

The expected information (Fisher information) is $I_{\mathbf{X}}(\theta) = n/\theta^2$ and, hence, the chi square **approximation** to the deviance is:

$$D_{\text{approx}}(\theta) \approx \frac{n(\theta - \bar{x})^2}{\bar{x}^2}.$$

This looks a bit different to the exact value of $D(\theta)$. To see why they are yet very close, let us expand the exact deviance $D(\theta)$ using the Taylor series approximation:

$$\log(1+y) \approx y - \frac{y^2}{2} + \ldots \qquad \text{for} \qquad y = \frac{\bar{x}-\theta}{\theta}$$

when y is small, we can set

$$\begin{aligned} \log \frac{\bar{X}}{\theta_0} &= \log\Big(1 + \frac{\bar{X}-\theta_0}{\theta_0}\Big) \\ &\approx \frac{\bar{X}-\theta_0}{\theta_0} - \frac{1}{2}\Big(\frac{\bar{X}-\theta_0}{\theta_0}\Big)^2 + \ldots \end{aligned}$$

Then by substituing this into the equation for the deviance in $D(\theta_0)$, and after some obvious cancellations, we end up with the approximation

$$\frac{n(\theta - \bar{x})^2}{\theta^2}.$$

for the deviance statistic. This is, of course, asymptotically very close to the chi square approximation to the deviance $D_{\text{approx}}(\theta)$:

$$\frac{n(\theta - \bar{x})^2}{\bar{x}^2}$$

since the consistent MLE $\bar{x} \approx \theta$ for large n. Both statistics ($D(\theta_0)$ and $D_{approx}(\theta_0)$) can be used to test the hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.