5 Lecture 5: Likelihood Inference. First order asymptotics

5.1 Why asymptotics

We realized 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 art, and there is no easy to follow constructive algorithm for their determination. On the other hand, the MLE's are typically easy to construct by following a general recipe of optimizing either directly the Likelihood or the log-likelihood function, i.e., by following an easy general recipe. 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.

Note: It is easy to work with exponential families because their structure directly helps us identify a **minimal sufficient and complete** statistic: once the function $d(x)$ in the definition of the exponential family has been identified, we know that $T = \sum_{i=1}^{n} d(X_i)$ is complete and minimal sufficient. The Lehmann-Scheffe theorem can be used to construct UMVUE in such families.

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

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

The bias-correction itself tends to be negligible as the sample size increases. Therefore the UMVUE$'s$ are either MLE$'s$ or "almost" MLE's. Hence, it is justified to look for a strong backing of the properties of MLE$'s$ in a general setting. This can be done using asymptotic arguments, i.e. by looking at the performance of MLE$'s$ when $n \longrightarrow \infty$, i.e. by letting the amount of information become arbitrarily large. Statistical folklore says then that "nothing can beat the MLE asymptotically".

5.2 Convergence concepts in asymptotics

We remind some stochastic convergence concepts first.

An estimator T_n of the parameter θ is said to be:

i) *consistent* (or *weakly consistent*) if

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

for all $\theta \in \Theta$ and for every fixed $\epsilon > 0$. We denote this by $T_n \xrightarrow{P} \theta$.

ii) *strongly consistent* if $P_\theta\{\lim_{n\to\infty} T_n = \theta\} = 1$ for all $\theta \in \Theta$.

iii) *mean-square consistent* if $\mathrm{MSE}_\theta(T_n) \longrightarrow_{n\to\infty} 0$ for all $\theta \in \Theta$.

It is important to note that **if the estimator is mean-square consistent then it is also consistent**. This relation has probably 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. To justify the relation between mean-square consistency and consistency we can use the **Chebyshev Inequality.** It states that for any random variable X and any $\epsilon > 0$ it holds for the k-th moment:

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

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 RHS tends to zero, the LHS will also tend to zero thus implying consistency.

Also, **strong consistency implies weak consistency.**

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 $X_1, X_2, .., X_n, ..$ have cumulative distribution functions $F_1, F_2, \ldots, F_n, ..$ 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 $X_n \xrightarrow{d} X$.

5.3 Consistency and asymptotic normality of MLE.

The basic statement about asymptotic properties of MLE follows.

Theorem 5.1. *Let* $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ *be i.i.d. from* $f(x, \theta), \theta \in \Theta \in R^1, \Theta-$ *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) $\frac{\partial}{\partial \theta} \int f(x, \theta)dx = \int \frac{\partial}{\partial \theta} f(x, \theta)dx$; $\frac{\partial^2}{\partial \theta^2} \int f(x, \theta)dx = \int \frac{\partial^2}{\partial \theta^2} f(x, \theta)dx$

3) $0 < I(\theta) = E_\theta(\frac{\partial ln f}{\partial \theta}(x, \theta)^2) < \infty$ *for all* $\theta \in \Theta$

4) $|\frac{\partial^3 ln f}{\partial \theta^3}(x, \theta)| \leq H(x)$ *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) $\sqrt{n}(\hat{\theta_n} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$

Proof a) i) First step: notice that

$$\frac{1}{n}\log L(X,\hat{\theta_n}) \geq \frac{1}{n}\log L(X,\theta_0) \tag{9}$$

where $L(.,.)$ denotes the joint density of n independent identically distributed (i.i.d.) observations, each with a density $f(.,.)$.

ii) Notice that by Jensen's Inequality:

$$E_{\theta_0}[\log\frac{L(X,\theta)}{L(X,\theta_0)}] < \log E_{\theta_0}[\frac{L(X,\theta)}{L(X,\theta_0)}] = 0$$

which implies

$$E_{\theta_0}[\frac{1}{n}\log L(X,\theta)] < E_{\theta_0}[\frac{1}{n}\log L(X,\theta_0)]$$

The Law of Large numbers implies then that for a fixed $\theta \neq \theta_0$ we should have

$$P_{\theta_0}\{\lim_{n\to\infty}\frac{1}{n}\log L(X,\theta) < \lim_{n\to\infty}\frac{1}{n}\log L(X,\theta_0)\} = 1 \tag{10}$$

Comparing (9) and (10) we see that we need to have $P_{\theta_0}(\hat{\theta}_n \to \theta_0) = 1$.

b) Since

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

holds, after Taylor expansion around θ_0 and recollection of terms we get

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

Here η_i are intermediate values: $|\eta_i| < 1$. Now we just have to use:

-the Law of large numbers regarding the term $[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)]$,

-the central limit theorem regarding the term $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i,\theta_0)$

-and the uniform bound assumption 4 of the theorem

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

to finish the argument.

Note: 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 predominant majority of cases, the parameter vector of interest is multi-dimensional. Let now $f(x,\theta), \theta \in \Theta \in R^p$.

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

$$E(\frac{\partial}{\partial\theta_i}log\mathbf{L}\frac{\partial}{\partial\theta_j}log\mathbf{L}) = -E(\frac{\partial^2}{\partial\theta_i\partial\theta_j}log\mathbf{L}), i = 1,2,\ldots,p; j = 1,2,\ldots,p.$$

(*For simplicity of notation, we 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 5.1):

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

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

Here $I_{X_1}^{-1}(\theta_0)$ 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

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

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

$$A_{\theta_0} - I_{X_1}^{-1}(\theta_0) \tag{11}$$

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 the case of 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 some additional obstacles in formulating such a claim. First of all, 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. The second obstacle is the existence of the so-called *superefficiency phenomena* (first example of a superefficient estimator has been suggested by Hodges). The examples of superefficient estimators show that 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*. Nevertheless, by imposing some further reasonable regularity conditions and a suitable interpretation of the convergence to the asymptotic distribution, the above obstacles can be overcome. The details are subtle and will not be discussed in our course. 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 (11) we can claim that the MLE is asymptotically efficient.

5.5 Delta method

5.5.1 Invariance property of MLE.

The main theorem about asymptotic properties of MLE was related to the estimation of the parameter θ itself. Sometimes, 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:

- what is the MLE of the new parameter $h(\theta)$.
- what is 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.

Let us now assume in addition 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". Let us discuss it below.

5.5.2 Delta method

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

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

From here we get the convergence in distribution:

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

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

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

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

$$\nabla h(\overrightarrow{\theta}) = (\frac{\partial}{\partial \theta_1} h(\overrightarrow{\theta}), \dots, \frac{\partial}{\partial \theta_p} h(\overrightarrow{\theta}))'$$

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

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

(Note that $I_{\mathbf{X}}(\overrightarrow{\theta}) = nI_{X_1}(\overrightarrow{\theta})$ 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 Examples

Example 1. (details at lecture). For estimating the parameter $\sqrt{\lambda}$ of the Poisson (λ) distribution using MLE, we get $\sqrt{n}(\sqrt{\bar{X}} - \sqrt{\lambda}) \xrightarrow{d} N(0, \frac{1}{4})$. There is an interesting additional observation that should be made in relation to this example. Although the asymptotic normal distribution of the MLE for the parameter λ has a variance that **depends** on the (unknown) parameter λ itself, the asymptotic normal distribution of the *transformed* parameter $h(\lambda) = \sqrt{\lambda}$ has a *constant* variance of $\frac{1}{4}$ independent of the value of λ. Such a transformation $h(.)$ that makes the asymptotic variance independent of the parameter, is called **variance stabilising transformation.** Variance stabilising transformations are actively sought after sometimes, especially for the purpose of constructing confidence intervals with a more precise coverage accuracy.

Example 2 (exponential transformation) Consider $h(\theta) = e^{\theta}$. Assume $X_1, X_2, \ldots, X_n$ are i.i.d. (not necessarily normal) with $E(X_i) = \theta, Var(X_i) = \sigma^2, i = 1, 2, \ldots, n$. Here $h'(\theta) = e^{\theta}$. The delta method tells us that the distribution of $e^{\bar{X}}$ can be approximated by **normal** with mean e^{θ} and variance $\frac{1}{n}\sigma^2 e^{2\theta}$. Using the data, this variance could be estimated by $\frac{1}{n} S^2 e^{2\bar{X}}$ where S^2 is just the sample variance.

Example 3 (Asymptotic distribution of a ratio estimator, this is a variant of the Example 5.5.27 in CB) Suppose X and Y are bivariate normally distributed with a known 2×2 covariance matrix $\Sigma = (\sigma_{ij}, i = 1, 2, j = 1, 2)$ 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 $\frac{\mu_X}{\mu_Y}$.

Solution: First we note that $h(\mu_X, \mu_Y) = \frac{\mu_X}{\mu_Y}$ and $\frac{\partial}{\partial \mu_X} = \frac{1}{\mu_Y}, \frac{\partial}{\partial \mu_Y} = \frac{-\mu_X}{\mu_Y^2}$. From the first order Taylor expansion we have $E(\frac{\bar{X}}{\bar{Y}}) \approx \frac{\mu_X}{\mu_Y}$. The inverse of the information matrix is

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

(WHY(!)) hence aplying the delta method we get

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

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

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

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

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

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

5.5.4 Exact and asymptotic distributions for the deviance

The deviance $D(\theta) = -2\log\left(\frac{L(\mathbf{X},\theta)}{L(\mathbf{X},\hat{\theta}_{MLE})}\right)$ has an important role in constructing confidence intervals and testing hypotheses about unknown parameters. It is a function of the observations, as well 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.

Examples (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 $\bar{X} \sim N(\mu_0, \sigma^2/n)$. Then $D(\mu_0) \sim \chi^2(1)$(chi-squared with one d.f.) (this is an *exact* (not asymptotic) result).

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

  $\text{logL}(\mathbf{X}, \theta_0) = \text{logL}(\mathbf{X}, \hat{\theta}_{mle}) + (\theta_0 - \hat{\theta}_{mle})\frac{\partial \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta} + \frac{1}{2}(\hat{\theta}_{mle} - \theta_0)^2 . \frac{\partial^2 \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta^2} + \ldots$

  Because the second summand in the RHS vanishes at $\hat{\theta}_{mle}$, ignoring higher order terms, we get: $D(\theta_0) \approx (\hat{\theta}_{mle} - \theta_0)^2\{-\frac{\partial^2 \text{logL}(\mathbf{X},\hat{\theta}_{mle})}{\partial\theta^2}\}$.

  Using the results about the normal approximation to the distribution of MLE, we get 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,

  $$(\theta_0 - \hat{\theta}_{mle})'\{-\frac{\partial^2 \text{logL}(\mathbf{X}, \hat{\theta}_{mle})}{\partial\theta\partial\theta'}\}(\theta_0 - \hat{\theta}_{mle}) \sim \chi^2_p$$

  Hence, we can easily 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. (How (?)) Discussion of this example will be continued later in the course under the heading "Generalized likelihood ratio tests".

- As a further example, we apply the above results in the case of the exponential distribution. Assume that $x_1, x_2, .., x_n$ are i.i.d. realizations from the density $f(x, \theta) = \frac{1}{\theta}\exp(-\frac{x}{\theta}), x > 0$. Then $L(\mathbf{x}, \theta) = \frac{1}{\theta^n}\exp(-\sum_{i=1}^n x_i/\theta)$; the MLE is $\hat{\theta} = \bar{x}$ and the exact deviance is easily seen to be $D(\theta) = 2n[\frac{\bar{x}}{\theta} - \ln(\frac{\bar{x}}{\theta}) - 1]$. The expected information (Fisher information) is $I_{\mathbf{X}}(\theta) = n/\theta^2$ and hence the chi square approximation to the deviance is $D_{approx}(\theta) \approx \frac{n(\theta-\bar{x})^2}{\theta^2}$. Comparing this with the exact value

$D(\theta)$ we see that if we expand $\log(1+y) \approx y - \frac{y^2}{2}$ for $y = \frac{\bar{x}-\theta}{\theta}$ in the exact formula, we would get the above chi square approximation for the deviance. Both statistics ($D(\theta_0)$ and $D_{approx}(\theta_0)$) can be used to test the hypothesis $H_0 : \theta = \theta_0$ versus $H_a : \theta \neq \theta_0$.