Massimo Guidolin
Massimo.Guidolin@unibocconi.it
Dept. of Finance

Università Commerciale
Luigi Bocconi

STATISTICS/ECONOMETRICS PREP COURSE – PROF. MASSIMO GUIDOLIN

SECOND PART, LECTURE 2: MODES OF CONVERGENCE AND POINT ESTIMATION

OVERVIEW

1) "As $T \to \infty$"???? What is that?
2) Convergence in probability and weak law of large numbers
3) Almost sure convergence and strong law of large numbers
4) Convergence in distribution and the central limit theorem
5) Estimators and estimates; methods of estimation
6) Method of moments
7) Maximum likelihood estimation (MLE)
8) Hints to Bayesian estimation
9) Evaluating estimators: MSE, UMVUE, Consistency

+ Bonus track: an introduction to the delta method

Modes of Convergence in Econometrics

- We introduce the idea/abstraction of allowing the sample size to approach infinity and investigates the behavior of certain sample quantities as this happens
  - Although the notion of an infinite sample size is a theoretical artefact, it can often provide us with some useful approximations
- We are concerned with three types of convergence:
  - ❶ Convergence in probability, the weakest type (easy to check)
  - ❷ Almost sure convergence, a stronger mode
  - ❸ Convergence in distribution
- Definition [CONVERGENCE IN PROB.]: A generic sequence of random variables, $X_1, X_2, \ldots$, converges in probability to a random variable X if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \epsilon) = 0 \text{ or, equivalently, } \lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1$$

Modes of Convergence in Econometrics

- Why the need of that "generic sequence"?
- Differently from lecture 1, X_1, X_2, ... are typically NOT independent and identically RVs (random variables), i.e., not a random sample
- The distribution of X_n may change as the subscript changes

- One famous example of convergence in probability that requires IID-ness and the concerns sample mean is:
- Key Result 4 [WEAK LAW OF LARGE NUMBERS, WLLN]: Let X_1, X_2, ... be IID RVs with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| < \epsilon) = 1, \text{ or } \bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu$$

  - In words, the (WLLN) elegantly states that, under general condi-tions, the sample mean approaches the population mean as $n \to \infty$
  - Notice that this version requires that the sequence of RVs be IID, although more general versions exist

Modes of Convergence in Econometrics

- Note that if $X_1, X_2, \ldots$ converges in probability to a random variable X and $h(\cdot)$ is a continuous function, then $h(X_1), h(X_2), \ldots$ converges in probability to h(X)
  - No problem if X is not a random variable, but a constant, say *a*
- A type of convergence that is stronger than convergence in probability is almost sure convergence (sometimes confusingly known as convergence with probability 1)

- Definition [ALMOST SURE CONVERGENCE]: A generic sequence of random variables, $X_1, X_2, \ldots$, converges almost surely to a random variable X if, for every $\epsilon > 0$,

$$\Pr\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1 \text{ or, equivalently, } X_n \overset{a.s.}{\to} X$$

  - Note: $\Pr\left(\lim_{n\to\infty} |X_n - X| < \epsilon\right) = 1$ different from $\lim_{n\to\infty} \Pr(|X_n - X| < \epsilon) = 1$
  - To understand the meaning of a.s. convergence, think of the meaning of RV: a RV is a real-valued function defined on a sample space S

Modes of Convergence in Econometrics

- If a sample space S has elements denoted by s, then $X_n(s)$ and $X(s)$ are all functions defined on S
- The definition states that X_n converges to X almost surely if the functions $X_n(s)$ converge to $X(s)$ for all $s \in S$ except perhaps for $s \in N$, where $N \subset S$ and $P(N) = 0$
- The key is that $N \subset S$ and $P(N) = 0$: what does it mean in practice? If you want to understand, read the following example
  - Example: Let the sample space S be the closed interval [0, 1] with the uniform probability distribution. Define random variables $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0, 1)$, $s^n \longrightarrow 0$ as $n \longrightarrow \infty$ and $X_n(s) \longrightarrow s = X(s)$
  - However, $X_n(1) = 2$ for every n so $X_n(1)$ does not converge to $1 = X(1)$. But since the convergence occurs on the set [0, 1) and $P([0, 1)) = 1$, X_n converges to X almost surely

- Convergence almost surely, being the stronger criterion, implies convergence in probability; the converse is false

Modes of Convergence in Econometrics

  - The stronger analog of the WLLN uses almost sure convergence:

- Key Result 5 [STRONG LAW OF LARGE NUMBERS, SLLN]: Let $X_1, X_2, \ldots$ be IID RVs with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Then for every $\epsilon > 0$,

$$\Pr(\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon) = 1, \text{ or } \bar{X}_n \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu$$

  - Also this SLLN requires that the sequence of RVs be IID, although more general versions exist

- Definition [CONVERGENCE IN DISTRIBUTION]: A generic sequence of random variables, $X_1, X_2, \ldots,$ converges in distribution to a random variable X if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$

  - Here $F(\cdot)$ is the CDF of a random variable; we also write $X_n \to^D X$
  - Note that although we talk of a sequence of random variables converging in distribution, it is really the cdfs that converge

Modes of Convergence in Econometrics

  - Is there a link between different modes of convergence? In general no , but when convergence results refer to RVs, some results exist

- Key Result 6: If the sequence of random variables, X_1, X_2, ... converges in probability to a random variable X, the sequence also converges in distribution to X
  - A similar statement is that if the sequence of RVs, X_1, X_2, ... , converges in probability to a constant (say, θ) if and only if the sequence also converges in distribution to θ; the statements are equivalent:

$$\Pr(|X_n - \theta| \geq \epsilon) = 0 \ \forall \epsilon > 0 \iff \Pr(X_n \leq x) = \begin{cases} 0 \text{ if } x < \theta \\ 1 \text{ if } x \geq \theta \end{cases}$$

- Applied to the sample mean, convergence in distribution originates one key result:
- Key Result 7 [CENTRAL LIMIT THEOREM, CLT]: Let X_1, X_2, ... be IID RVs with $E[X_i] = \mu$ and $Var[X_i] = \sigma^2 < \infty$. Define $\bar{Z}_n \equiv (\bar{X}_n - \mu)/(\sigma/\sqrt{n})$

The Central Limit Theorem

Then $$\lim_{n\to\infty} F_{\bar{Z}_n}(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = \Phi(z)$$

i.e., the standardized sample mean has a limiting N(0,1) distribution

- Proof is easy making assumptions on MGFs, but that is not necessary
- Starting from virtually no assumptions (other than IIDness and finite variances), we end up with normality!
- The point here is that normality comes from sums of "small" (finite variance), independent disturbances
- You will encounter this result in continuous time finance too
- While the CLT gives us a useful general approximation, we have no automatic way of knowing how good the approximation is in general
- With the availability of cheap computing power, the importance of approximations like the CLT is somewhat lessened

Methods to Find Point Estimators

  - The CLT gives conditions under which a standardized RV has a limit normal distribution; there are times when we are not interested in the distribution of the RV, but rather some function of the RV

- The Appendix at the end of this lecture provides one useful tool, the delta method which can be seen as a CLT extension
- We start by reviewing a number of alternative methods to find estimators; subsequently we evaluate them
  - The rationale behind point estimation is simple: when sampling is from a population described by a pdf or pmf $f(x;\theta)$, knowledge of θ yields knowledge of the entire population
- We call θ the parameter(s) of interest; when they are many, then θ is a Kx1 vector; $\hat{\theta}$ is then an estimator
  - In general, an estimator is a function of the data, formally a function of a random sample $X_1, X_2, \ldots, X_n$, and as such it is a sample statistic, i.e., $\hat{\theta} = W(X_1, X_2, \ldots, X_n)$

Methods to Find Point Estimators

- An estimator is a function of the sample, and THEREFORE IT IS A RANDOM VARIABLE WITH A DISTRIBUTION, while an estimate is the realized value of an estimator (that is, a number) that is obtained when a sample is actually taken
  - Notationally, when a sample is taken, an estimator is a function of the random variables $X_1, \dots, X_n$, while an estimate is a function of the realized values $x_1, \dots, x_n$

- Three methods to find estimators:
  1. Method of moments
  2. Maximum likelihood estimation
  3. Bayes' methods
- The method of moments is the oldest method of finding point estimators; it is easy and always delivers "some" estimator, unfortunately it is rarely the best possible one

METHOD OF MOMENTS

- Definition [METHOD OF MOMENTS]: Let $X_1, \ldots, X_n$ be a sample from a population with pdf or pmf $f(x; \theta_1, \ldots, \theta_K)$. Method of moments estimators are found by equating the first K sample moments to the corresponding K population moments, and solving the resulting system of simultaneous equations:

$$m_1 \equiv n^{-1} \sum_{i=1}^{n} X_i = \mu_1(\theta_1, \theta_2, \ldots, \theta_K)$$

$$m_2 \equiv n^{-1} \sum_{i=1}^{n} X_i^2 = \mu_2(\theta_1, \theta_2, \ldots, \theta_K)$$

$$\ldots$$

$$m_K \equiv n^{-1} \sum_{i=1}^{n} X_i^K = \mu_K(\theta_1, \theta_2, \ldots, \theta_K)$$

  - This is a system of K equations in K unknowns; if a solution $\boldsymbol{\theta}^*$ can be found, then the resulting vector $\boldsymbol{\theta}^*$ is the MM estimator
  - Why "if a solution can be found"? Some or all the K equations may be

METHOD OF MOMENTS

non-linear which you may know can be problematic...

- Because K is the both the number of parameters and the number of equations, this MM is called just-identified
- Example Suppose $X_1, \ldots, X_n$ are IID $N(\theta, \sigma^2)$. Call then $\theta_1 = \theta$ and $\theta_2 = \sigma^2$. Therefore:

$$m_1 \equiv n^{-1} \sum_{i=1}^{n} X_i = \theta \Longrightarrow \hat{\theta}^{MM} = \bar{X}$$

$$m_2 \equiv n^{-1} \sum_{i=1}^{n} X_i^2 = \theta_2 + \theta^2 \Longrightarrow \hat{\theta}_2^{MM} = n^{-1} \sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \frac{n-1}{n} S^2$$

- Time to switch to the most popular among estimation me-thods, one that (as you will discover in the Metrics sequence) also possesses a number of optimal properties
  - Recall that if $X_1, \ldots, X_n$ is an IID sample from a population with pdf or pmf $f(x; \theta_1, \ldots, \theta_K)$, then the likelihood function is defined as:

MAXIMUM LIKELIHOOD ESTIMATORS

$$L(\boldsymbol{\theta}; \mathbf{x}) \equiv L(\theta_1, \theta_2, \ldots, \theta_K; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_K)$$

- The likelihood function is just the joint PDF or PMF of the data, but interpreted to be a function of the K parameters given the data
  - The joint PDF or PFM is instead the opposite: a function of data, given the values of the parameters

- Definition [MAXIMUM LIKELIHOOD METHOD]: For a given sample **x**, let $\hat{\boldsymbol{\theta}}(\mathbf{x})$ be a parameter value at which L(θ; **x**) attains its maximum as a function of θ, with **x** held fixed; a maximum likelihood estimator (MLE) of θ based on a sample **X** is $\hat{\boldsymbol{\theta}}(\mathbf{X})$
- The intuition is simple: the MLE is the parameter "configuration" for which the observed sample is most likely
  - The maximum of L(θ; **x**) should be a global one
  - This is they key problem of MLE: one needs to maximize L(θ; **x**) and verify that the stationarity point one has found is actually global

Maximum Likelihood Estimators

- If the likelihood function is differentiable in the parameters, then MLEs are characterized as:

$$\hat{\boldsymbol{\theta}}(\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x}) \Leftrightarrow \frac{\partial L(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}'} = \mathbf{0}$$

- Note that the solutions to this system of K equations are only possible candidates for the MLE since the first derivatives being 0 is only a necessary condition for a maximum, not a sufficient condition
- Furthermore, the zeros of the first derivative locate only extreme points in the interior of the domain of a function
- Example Suppose $X_1, \ldots, X_n$ are IID $N(\theta, 1)$. Then the likelihood function is:

$$\frac{\partial L(\theta; \mathbf{x})}{\partial \theta} = \underbrace{\frac{1}{\sqrt[n]{2\pi}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i - \hat{\theta}^{MLE})^2\right]}_{>0 \text{ for all possible values of } \theta} \sum_{i=1}^{n}(x_i - \hat{\theta}^{MLE}) = 0$$

$$\Longrightarrow \sum_{i=1}^{n}(x_i - \hat{\theta}^{MLE}) = 0 \Longrightarrow \hat{\theta}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}_n$$

MAXIMUM LIKELIHOOD ESTIMATORS

– Moreover, this MLE represents an interior maximum not a minimum, as it can be verified from:

$$\left.\frac{\partial^2 L(\theta;\mathbf{x})}{\partial\theta^2}\right|_{\theta=\hat{\theta}^{MLE}} = \frac{1}{\sqrt[n]{2\pi}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\theta}^{MLE})^2\right]\underbrace{\left[\sum_{i=1}^{n}(x_i-\hat{\theta}^{MLE})\right]^2}_{=0}+$$

$$-\frac{1}{\sqrt[n]{2\pi}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\theta}^{MLE})^2\right] = -\frac{1}{\sqrt[n]{2\pi}}\exp\left[-\frac{1}{2}\sum_{i=1}^{n}(x_i-\hat{\theta}^{MLE})^2\right] < 0$$

– In most cases, it is easier to work with the natural logarithm of $L(\boldsymbol{\theta}; \mathbf{x})$, $\ln L(\boldsymbol{\theta}; \mathbf{x})$, known as the log likelihood function

– This is possible because the log function is strictly increasing on $(0, +\infty)$, which implies that the extrema of $L(\boldsymbol{\theta}; \mathbf{x})$ and $\ln L(\boldsymbol{\theta}; \mathbf{x})$ coincide

– Example Suppose $X_1, \ldots, X_n$ are IID Bernoulli(p). Then the log-lik fnct. is:

$$\ln L(p;\mathbf{x}) = \ln\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = \ln\left[p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}\right]$$

$$= \left(\sum_{i=1}^{n}x_i\right)\ln p + \left(n-\sum_{i=1}^{n}x_i\right)\ln(1-p)$$

Maximum Likelihood Estimators

– At this point, taking FOCs and solving, one has the estimator that you would expect, that is however a MLE:

$$\frac{\partial L(p;\mathbf{x})}{\partial p} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{(n-\sum_{i=1}^{n} x_i)}{1-p} = 0$$

$$\Longrightarrow \frac{1\text{-}\hat{p}}{\hat{p}} = \frac{n\text{-}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} = \frac{n}{\sum_{i=1}^{n} x_i} - 1 \Longrightarrow \frac{1}{\hat{p}} - 1 = \frac{n}{\sum_{i=1}^{n} x_i} - 1$$

$$\Longrightarrow \frac{1}{\hat{p}} = \frac{n}{\sum_{i=1}^{n} x_i} \Longrightarrow \hat{p}^{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}_n$$

- A useful property of MLEs is their invariance property: a distribution is indexed by a parameter θ, but the interest is in finding an estimator for some function of θ, say $\tau(\theta)$
- The invariance property of MLEs says that if θ-hat is the MLE of θ, then, if $\tau(\bullet)$ fulfils adequate conditions, $\tau(\theta\text{-hat})$ is the MLE of $\tau(\theta)$

PROPERTIES OF MLES: INVARIANCE

- Example: if θ is the mean of a normal distribution, the MLE of $\sin(\theta)$ is sin of the sample mean

- What are the conditions on $\tau(\cdot)$ alluded to above?

  ❶ The mapping $\theta \longrightarrow \tau(\theta)$ is one-to-one (i.e., for each value of θ there is a unique value of $\tau(\theta)$, and vice versa)

  - However, such a condition is not helpful in many cases: think of the case of volatility = square root of variance in finance!

  ❷ Under some regularity conditions ("smoothness") on $\tau(\cdot)$ then the invariance property always holds

  - Therefore the MLE of volatility is simply the square root of the MLE estimator of variance,
  - The MLE estimator of the Bernoulli variance p(1 – p) is $\hat{p}^{MLE}(1-\hat{p}^{MLE})$
  - Note that a function may be one-to-one even when it is not smooth

- Since we can apply more than one of the estimation methods to each problem, we need to choose between estimators

Evaluating Alternative Estimators

  - In many cases, different methods will lead to different estimators

- Three main criteria:
  1. Mean squared error
  2. Efficiency (best unbiased criterion)
  3. Consistency

- Definition [MEAN SQUARED ERROR]: The MSE is a finite-sample measure of the quality of an estimator W (short for $\hat{\theta} = W(X_1, X_2, \ ..., \ X_n)$) of a parameter θ is $E_\theta[(W - \theta)^2]$ and it measures the average squared difference between the estimator W and the parameter θ
- Importantly MSE may be decomposed as:

$$MSE(W) \equiv E_\theta[(W-\theta)^2] = E_\theta[W^2 - 2W\theta + \theta^2] = E_\theta[W^2] - 2\theta E_\theta[W] + \theta^2]$$
$$= Var_\theta[W^2] + \{E_\theta[W]\}^2 - 2\theta E_\theta[W] + \theta^2$$

MEAN SQUARED ERROR CRITERION

$$MSE(W) = Var_\theta[W] + \{\underbrace{E_\theta[W] - \theta}_{\text{bias}(W)}\}^2 = Var_\theta[W] + \{bias(W)\}^2$$

where the bias is the mean difference between the estimator W and the population parameter, θ

- Definition [UNBIASEDNESS]: An estimator W that has a bias of 0 is called unbiased and in this case MSE(W) = Var[W], it is only the uncertainty on the estimator that matters
  - An unbiased estimator with a large Var[W] is one that is on average correct, but that fluctuates a lot around the true θ, and such is not precise; such an estimator will also said to be inefficient
  - In Lecture 1 we have stated or proven that in the case of a random (IID) sample both sample mean and variance are unbiased and so:

$$E[\bar{X}] = \mu \; E[S^2] = \sigma^2 \Longrightarrow MSE[\bar{X}] = \frac{\sigma^2}{n}, \; MSE[S^2] = \frac{2\sigma^2}{n-1} \text{ (under normality)}$$

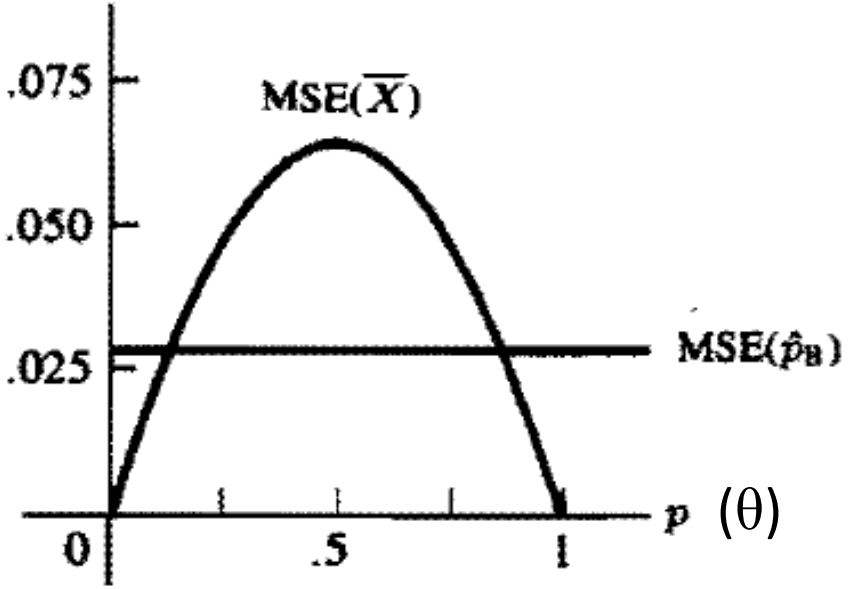
  - In general, since MSE is a function of the parameter, there will not

Mean Squared Error Criterion

be one "best" estimator: often, the MSEs of two estimators will cross each other, showing that each estimator is better (with respect to the other) in only a portion of the parameter space

– The reason that there is no one "best MSE" estimator is that the class of all estimators is too large a class

  - E.g., the estimator θ-hat = 17 cannot be beaten in MSE at $\theta = 17$ but is a terrible estimator otherwise



- One way to make the problem of finding a "best" estimator tractable is to limit the class of estimators; a popular way of restricting the class, is to consider only unbiased estimators
- If W_1 and W_2 are both unbiased estimators of a parameter θ, then their mean squared errors are equal to their variances, so we should choose the estimator with the smaller variance

UNIFORM MINIMUM VARIANCE UNBIASED

- Definition [BEST UNBIASED ESTIMATOR]: An unbiased estimator W^* for θ is the best unbiased, if for all possible values of θ, $Var_\theta[W^*] \leq Var_\theta[W]$ among all other unbiased estimators
- W^* is also called a uniform minimum variance unbiased estimator (UMVUE) of θ
- It is then legitimate to ask: how small can this minimum variance be? It seems a dumb question, but surprisingly, the following gives the answer:
- Key Result 8 [CRAMER-RAO LOWER BOUND (INEQUALITY)]: Let $X_1, X_2, ..., X_n$ be a sample from a PDF $f(\mathbf{x}; \theta)$ and let $W(X_1, X_2, ..., X_n)$ be an estimator with $\frac{d}{d\theta}E[W(\mathbf{X})] = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x};\theta)]d\mathbf{x}$ and $Var[W] < \infty$. Then

$$Var[W(\mathbf{X})] \geq \frac{\left(\frac{d}{d\theta}E[W(\mathbf{X})]\right)^2}{E\left\{\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2\right\}}$$

Fisher information

CRAMER-RAO LOWER BOUND

- In words, the key condition $\frac{d}{d\theta}E[W(\mathbf{X})] = \int_{-\infty}^{+\infty} \frac{\partial}{\partial\theta}[W(\mathbf{x})f(\mathbf{x};\theta)]d\mathbf{x}$ simply allows the interchange between integration and differentiation
  - Densities in the exponential class, such as the normal, will satisfy this
- As for Fisher's number, clearly it measures how sensitive the log-likelihood is to a change in θ; the more this quantity is sensitive, the lower is the bound, which makes sense: the data, as summarized by the log-likelihood will contain a lot of information on θ
- How do you use Cramer-Rao bound? Simple, if you have a unbiased estimator W and can verify that

  $$Var[W(\mathbf{X})] = \frac{\left(\frac{d}{d\theta}E[W(\mathbf{X})]\right)^2}{E\left\{\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2\right\}}$$

  then it must be UMVUE
- Notice that the converse does not hold though, there is no guarantee that the bound is sharp: the value of the Cramer-Rao Lower Bound may be strictly smaller than the variance of any unbiased estimator

- All the criteria considered thus far are finite-sample criteria

ASYMPTOTICS (CONSISTENCY)

- In contrast, we may consider asymptotic properties, properties describing the behavior of an estimator as the sample size becomes infinite
- The power of asymptotic evaluations is that, when we let the sample size become infinite, calculations simplify
- The property of consistency requires that the estimator converges to the "correct" value as $n \longrightarrow \infty$
  - However consistency (as well as all asymptotic properties) concerns a sequence of estimators rather than a single estimator, although it is common to speak of a "consistent estimator“
  - If we observe X_1, X_2 , ... according to a distribution $f(x;\theta)$, we can construct a sequence of estimators $W_n=W_n(X_1, X_2, ..., X_n)$ merely by performing the same estimation procedure for each sample size n
    - For example, $X\text{-bar}_1 = X_1$, $X\text{-bar}_2 = (X_1+X_2)/2$, $X_3\text{-bar} =(X_1+X_2+X_3)/3$, etc.

CONSISTENCY

- Definition [CONSISTENCY]: A sequence of estimators $W_n=W_n(X_1, X_2, ..., X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$,

$$\lim_{n\to\infty} \Pr\left(|W_n - \theta| < \epsilon\right) = 1$$

  - Informally, this says that as the sample size becomes infinite (and the sample information becomes better and better), the estimator will be arbitrarily close to the parameter with high probability
  - Or, turning things around, we can say that the probability that a consistent sequence of estimators misses the true parameter is small
  - An equivalent statement is: for every $\epsilon > 0$ and every $\theta \in \Theta$, a consistent sequence W_n will satisfy $\lim_{n\to\infty} \Pr\left(|W_n - \theta| \geq \epsilon\right) = 0$
  - This definition says that a consistent sequence of estimators converges in probability to the parameter θ it is estimating
  - Whereas the definition of convergence in probability dealt with one

CONSISTENCY

sequence of RVs with one probability structure, consistency deals with an entire family of probability structures, indexed by θ

- Problem 12 in the first exercise set makes you discover how one should examine consistency "the hard way", i.e., using the definition
- In practice, using Chebychev's Inequality there is a simpler way

$$\Pr\left(|W_n - \theta| \geq \epsilon\right) \leq \frac{E[(W_n - \theta)^2]}{\epsilon^2}$$

so that consistency obtains if for every $\theta \in \Theta$, $\lim_{n\to\infty} E[(W_n - \theta)^2] = 0$

- Furthermore, because $E[(W_n - \theta)^2] = Var[W_n] + \{E[W_n - \theta]\}^2$
  $= Var[W_n] + \{\text{bias}(W_n)\}^2$
  we obtain

- <u>Key Result 9</u>: A sequence of estimators $W_n = W_n(X_1, X_2, \ldots, X_n)$ is a consistent sequence of estimators of the parameter θ if, for every $\epsilon > 0$ and every $\theta \in \Theta$, (i) $\lim_{n\to\infty} \text{Var}[W_n] = 0$, (ii) $\lim_{n\to\infty} \text{bias}(W_n) = 0$ (this is just a sufficient condition)

CONSISTENCY OF ML ESTIMATORS

- Because the sample mean has no bias and has variance σ^2/n, clearly, the sufficient conditions are satisfied which shows consistency
- We close by showing that MLEs possess all these asymptotic properties

- Key Result 10: Let $X_1, X_2, \ldots, X_n$ be an IID sample and let θ^{ML} the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under some regularity conditions on the PDF $f(x;\theta)$, $\tau(\theta^{ML})$ is consistent
  - Another useful asymptotic property relates to efficiency, which is concerned with the asymptotic variance of an estimator
- Definition [ASYMPTOTIC EFFICIENCY]: A sequence of estimators $W_n = W_n(X_1, X_2, \ldots, X_n)$ is asymptotically efficient for a parameter θ if $\sqrt{n}(W_n - \theta) \xrightarrow{D} N(0, v(\theta))$ $\quad v(\theta) = \dfrac{1}{E\left\{\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2\right\}}$

  that is, the asymptotic variance of W_n reaches the Cramer-Rao lower bound

Asymptotic Normality of ML Estimators

- Key Result 11: Let $X_1, X_2, \ldots, X_n$ be an IID sample and let θ^{ML} the MLE of θ. Let $\tau(\theta)$ be a continuous function of θ. Under some regularity conditions on the PDF $f(x;\theta)$, $n^{1/2}(\tau(\theta^{ML}) - \tau(\theta)) \rightarrow N(0, v(\theta))$, where $v(\theta)$ is the Cramer-Rao Lower Bound
  - I.e., $\tau(\theta^{ML})$ is a consistent and asymptotically efficient estimator of $\tau(\theta)$
- These asymptotic formulas are important because they often provide us with approximate variances, provided we are ready to invoke expanding samples, $n \rightarrow \infty$
  - For any fnct $h(\theta)$, the variance of the ML of θ can be approximated as:

$$Var[h(\hat{\theta})] \simeq \frac{h'(\theta)}{E\left\{\left[\frac{\partial}{\partial\theta}\ln f(\mathbf{x};\theta)\right]^2\right\}} = \frac{h'(\theta)}{-E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(\mathbf{x};\theta)\right\}} \simeq \frac{h'(\hat{\theta})}{-E\left\{\frac{\partial^2}{\partial\theta^2}\ln f(\mathbf{x};\hat{\theta})\right\}}$$

  - The variance estimation process is a two-step procedure, (i) first we approximate Var[$h(\theta)$], (ii) then we estimate the resulting approximation, usually by substituting the ML of θ for θ

Bayesian Methods

- The Bayesian approach to statistics is fundamentally different from the classical approach that we have been taking
  - In the classical approach the parameter, θ, is thought to be an unknown, but fixed, quantity
  - A random sample $X_1, \ldots, X_n$ is drawn from a population indexed by θ and, based on the observed values in the sample, knowledge about the value of θ is obtained
- In the Bayesian approach θ is considered to be a quantity whose variation can be described by a probability distribution (called the prior distribution)
  - This is a subjective distribution, based on the experimenter's belief, and is formulated before the data are seen
- A sample is then taken from a population indexed by θ and the prior distribution is updated with this sample information
- The updated prior is called the posterior distribution

Bayesian Methods

– If we denote the prior distribution by $\pi(\theta)$ and the sampling distribution by $f(x; \theta)$, then the posterior distribution, the conditional distribution of θ given the sample, x, is

$$\pi(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}, \boldsymbol{\theta})}{\int f(\mathbf{x}, \boldsymbol{\theta}) d\mathbf{x}} = \frac{f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) d\mathbf{x}}$$

Marginal distribution of **x**

- The posterior distribution is now used to make statements about θ, which is still considered a random quantity
  – E.g., the mean of the posterior distribution can be used as a point estimate of θ
- Although lots of interesting financial econometrics is performed with Bayesian methods, we shall not pursue it
  – This reflects the presumed structure and contents of your econometrics sequence in the first year of the MSc.
  – Optional exams may be selected for you to get an exposure to the practice of Bayesian econometrics

Useful Notions Reviewed in This Lecture

- Let me give you a list to follow up on:
- Weak and Strong laws of large numbers
- The central limit theorem
- Converge in probability, almost-sure convergence, and convergence in distribution
- Estimates vs. estimators
- Method of moments estimators
- Maximum likelihood estimators and their properties
- Mean-Squared Error criterion
- Unbiased and efficient estimators estimators
- The Cramer-Rao lower bound
- Consistency and asymptotic normality of estimators
- Brief overview of Bayesian methods

Appendix: Review of the Delta Method

- Motivating example on estimating the odds ratio. Suppose we observe $X_1, X_2, \ldots, X_n$ independent Bernoulli(p) random variables. The typical parameter of interest is p, the success probability, but another popular parameter is $p/(1 - p)$, the odds
  - For example, if the data represent the outcomes of a medical treatment with p = 2/3, then a person has odds 2 : 1 of getting better
- As we would typically estimate the success probability p with the observed success probability p* = sample mean of outcomes (the proportion of successes), we might consider using p*/(1 - p*) as an estimate of $p/(1 - p)$
- What are the properties of this estimator? How might we estimate the variance of p*/(1 - p*)?

- The Delta Method allow you to obtain reasonable, approximate answers to these questions
- It is based on using a Taylor series approximation to approximate

APPENDIX: REVIEW OF THE DELTA METHOD

the mean and variance of a function of a RV and then apply CLT

- Taylor's theorem states that if a function g(x) has derivatives of order r, that is, $g^{(r)}(x) = d^r g(x)/x^r$ exists, then for any constant *a*, the Taylor polynomial of order r about *a* is:

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^r \quad \text{(e.g., if } g(x) = \ln x \text{ and } a = 2,\ T_2(x) = \ln 2 + \frac{x-2}{2} - \frac{(x-a)^2}{8})$$

and the remainder from the approximation, $g(x) - T_r(x)$, always tends to 0 faster than the highest-order explicit term

  - Generalizing to the multivariate case (when **Z** is a K x 1 random vector that collects K RVs), and forgetting the remainder, we have that a first-order expansion of g(**Z**) around the vector of means θ gives the approximation:

$$g(\mathbf{z}) \simeq g(\boldsymbol{\theta}) + \sum_{i=1}^{K} g_i'(\boldsymbol{\theta})(z_i - \theta_i)$$

  - At this point, taking expectations of both sides, note that

Appendix: Review of the Delta Method

$$E[g(\mathbf{z})] \simeq E[g(\boldsymbol{\theta})] + \sum_{i=1}^{K} g_i'(\boldsymbol{\theta}) \underbrace{E[(z_i - \theta_i)]}_{=0} = g(\boldsymbol{\theta})$$

– Therefore the variance of g(**Z**) can be approximated as:

$$Var[g(\mathbf{z})] = E[(g(\mathbf{z}) - E[g(\mathbf{z})])^2] \simeq E\left[\left(\sum_{i=1}^{K} g_i'(\boldsymbol{\theta}) E[(z_i - \theta_i)]\right)^2\right]$$

$$= \sum_{i=1}^{K} (g_i'(\boldsymbol{\theta}))^2 Var[z_i] + 2 \sum_{i>j}^{K} g_i'(\boldsymbol{\theta}) g_j'(\boldsymbol{\theta}) Cov[z_i, z_j]$$

– This approximation is useful because it gives us a variance formula for a general function, using only simple variances and covariances

– For instance, consider g(p*) = p*/(1 - p*), so that K=1 (univariate) and g'(p) = 1/[(1 - p)^2]. Then

$$Var\left[\frac{p^*}{1-p^*}\right] \underbrace{\simeq}_{\Uparrow} (g'(p))^2 Var[p] = \frac{1}{(1-p)^4} \underbrace{np(1-p)}_{\text{Var of } Bi(n,p)} = \frac{np}{(1-p)^3}$$

Appendix: Review of the Delta Method

- The delta method generalizes this intuition and the CLT to any function g(•) of a normal random variable in a formal way
- Let Y_n be a sequence of RVs that satisfies $n^{1/2}(Y_n - \mu) \longrightarrow N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

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

  - In words: any "smooth" (differentiable) function of a sequence of RVs to which a CLT applies, has its own approximate CLT with variance $[g'(\theta)]^2\sigma^2$

- Why to stop at the first-order? You do not have to. First, notice

$$g(z) \simeq g(\theta) + g'(\theta)(z-\theta) + \frac{1}{2}g''(\theta)(z-\theta)^2$$

$$E[g(z)] \simeq g(\theta) + g'(\theta)E[(z-\theta)] + \tfrac{1}{2}g''(\theta)E[(z-\theta)^2]$$

$$\Longrightarrow Var[g(z)] \simeq [g'(\theta)]^2 Var[(z-\theta)] + \tfrac{1}{4}[g''(\theta)]^2 Var[(z-\theta)^2] +$$
$$+ g'(\theta)g''(\theta)E\{[(z-\theta) - E[(z-\theta)]][(z-\theta)^2 - E[(z-\theta)^2]]\}$$

Appendix: Review of the Delta Method

  - Therefore there is also a second-order delta method that is particularly useful when $g'(\theta) = 0$:

- Let Y_n be a sequence of RVs that satisfies $n^{1/2}(Y_n - \mu) \longrightarrow N(0, \sigma^2)$ in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta) = 0$ and $g''(\theta)$ exists and is not 0. Then

$$\sqrt{n}\,(g(Y_n) - g(\theta)) \xrightarrow{D} \frac{[g''(\theta)]^2}{2}\sigma^2\chi_1^2$$

  - Oddly enough, this is a sort of "chi-squared" version of the CLT, in the sense that because of the second-order Taylor approximation, the distribution of the sequence $g(Y_n)$ does not converge to a Normal, but instead to a (non-central, with dislocation $g(\theta)$) chi-squared distribution with 1 d.f.

- As a last touch, one exercise helps you "discover" that

$$E\left[\frac{X}{Y}\right] \simeq \frac{\mu_X}{\mu_Y} \qquad Var\left[\frac{X}{Y}\right] \simeq \frac{\mu_X^2}{\mu_Y^2}\left(\frac{\sigma_X^2}{\mu_X^2} + \frac{\sigma_Y^2}{\mu_Y^2} - 2\frac{\sigma_{XY}}{\mu_X\mu_Y}\right)$$