# Lecture 15. Convergence in Distribution, Continuous Mapping Theorem, Delta Method

11/7/2011

**Approximation using CLT (Review)**

The way we typically use the CLT result is to approximate the distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ by that of a standard normal. Note that if $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is exactly a $N(0,1)$ random variable, then $X_n$ is exactly a $N(\mu, \sigma^2/n)$ random variable for any $n$. Consequently, if $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximately a $N(0,1)$ random variable, then it makes sense to use $N(\mu, \sigma^2/n)$ as an approximating distribution for $\bar{X}_n$ .

Similarly, one can also use $N(n\mu, n\sigma^2)$ to approximate the sum of $n$ i.i.d. random variables: $S_n := \sum_{i=1}^n X_i$. For example, one could use $N(np, np(1-p))$ to approximate the a binomial distribution: $Bin(n,p)$ because the binomial distribution is the sum of $n$ i.i.d. Bernoulli random variables.

This approximation is not good unless $n$ is sufficient large. How large is sufficiently large is a good question. It depends on the underlying distribution from which the random sample $(X_1, ..., X_n)$ is drawn.

For the normal approximation of $Bin(n,p)$, a useful rule of thumb is that $n > 30$ is large enough for $Bin(n,p)$ to be approxiamted by $N(np, np(1-p))$ well enough. (Question: does this depend on $p$?)

How does one evaluate the approximation quality? One needs to have a measure of distance on the space of probability measures. A popular distance is the Kolmolgorov-Smirnov distance:

$$KS(F_n, F) = \sup_{x \in R} |F_n(x) - F(x)|, \tag{1}$$

where $F_n$ and $F$ are two cdf's. There are many other distances that one can use, but the KS distance is the most common, perhaps for its simplicity. Now that we have a measure of approximation quality, we can do the following computer exercise:

(1) For a fixed $n$, and some very large $S$ (say $S = 1000000$), generate $S$ random $n$-

samples: $(X_{1,s}, ..., X_{n,s}) : s = 1, ..., S$ from a population distribution $F_X$ with known mean $\mu$ and variance $\sigma^2$.

(2) Compute $\sqrt{n}(\bar{X}_{n,1} - \mu)/\sigma, ..., \sqrt{n}(\bar{X}_{n,S} - \mu)/\sigma$. Note that this is an i.i.d. sample of size $S$ from the distribution $F_{\sqrt{n}(\bar{X}_{n,1}-\mu)/\sigma}(\cdot)$.

(3) Estimate the cdf of $\sqrt{n}(\bar{X}_{n,1} - \mu)/\sigma$ by $\hat{F}_{\sqrt{n}(\bar{X}_{n,1}-\mu)/\sigma}(x) = S^{-1} \sum_{s=1}^{S} 1\{\sqrt{n}(\bar{X}_{n,1} - \mu)/\sigma \leq x\}$.

(4) Estimate the KS distance between $\hat{F}_{\sqrt{n}(\bar{X}_{n,1}-\mu)/\sigma}(x)$ and $\Phi(x)$ by taking many grid points on $R$.

(5) Change $n$ and repeat to see how large $n$ needs to be for the KS distance to be smaller than your tolerance level.

Obviously, different $F_X$ will require different $n$. Try $Bern(p)$ for different $p$. Try $t$-distributions with different degrees of freedom, and then try other familar distributions. You will get a sense about the applicability of the central limit theorem.

**Convergence in Distribution**

The CLT is a special case of a sequence of random variables "converge in distribution" to a random variable.

**Definition 1.** A sequence of random variables or vectors $\{Y_n\}_{n=1}^{\infty}$ **converges in distribution** to a random variable $Y$, if

$$\lim_{n\to\infty} \Pr(Y_n \leq y) = \Pr(Y \leq y), \tag{2}$$

for **all points of continuity** of $F_Y(\cdot)$.

*Remark.* (1) Equivalently, (2) can be replaced by:

$$\lim_{n\to\infty} F_{Y_n}(y) = F(y).$$

(2) Convergence in distribution is denoted $\to_d$: $Y_n \to_d Y$.

(3) If $Y_n \to_d Y$, we say $Y_n$ has an asymptotic/limiting distribution with cdf $F_Y(y)$.

(4) The concept of convergence in distribtion involves the distributions of random variables only, not the random variable themselves. e.g. suppose the CLT conditions hold:

$$\sqrt{n}(\bar{X}_n - \mu)/\sigma \to_d Z,$$

where $Z \sim N(0, 1)$. It is equally true that $\sqrt{n}(\bar{X}_n - \mu)/\sigma \to_d -Z$, because $-Z$ has the same distribution as $Z$.

Because of this, it is OK to simply write $Y_n \to_d N(0, 1)$ when $Y_n \to_d Z$ for $Z \sim N(0, 1)$.

(5) In the definition above, the convergence of the cdf is only required to hold for continuous points of the cdf of the limiting random variable. This is important when the limiting random variable is not a continuous random variable.

e.g. $Y_n$ has *pdf*:

$$f_{Y_n}(y) = \begin{cases} n/2 & x \in (0, 1/n) \\ n/2 & x \in (1, 1+1/n) \\ 0 & otherwise. \end{cases}$$

Then $Y_n \to_d Y$ for $Y \sim Bern(1/2)$, and $F_{Y_n}(y) \to F_Y(y)$ for all $y \in R/\{0, 1\}$.

**Continuous Mapping Theorem**

**Theorem 1** (CMT)**.** *If $h : R^m \to R^p$ is a continuous function, $Y_n$, $Y$ are $R^m$-valued random vectors, and $Y_n \to_d Y$. Then*

$$h(Y_m) \to_d h(Y).$$

e.g. If $Y_n \to_d Z \sim N(0, 1)$, then $Y_n^2 \to_d Z^2 \sim \chi^2(1)$.
If $\sqrt{n}(\bar{X}_n - \mu)/\sigma \to_d Z \sim N(0, 1)$, then $\sqrt{n}(\bar{X}_n - \mu) \to_d \sigma Z \sim N(0, \sigma^2)$.

If $Y_n := (Y_{1n}, Y_{2n})' \to_d (Z_1, Z_2)' =: Z \sim N(0, I_2)$, then $Y_n' Y_n \to_d Z'Z \sim \chi^2(2)$.

The CMT requires joint convergence (in distribution) of all elements of the vector $Y_n$. The following is NOT true: "if $Y_{1,n} \to_d Y_1$, $Y_{2,n} \to_d Y_2$ then $h(Y_{1,n}, Y_{2,n}) \to_d h(Y_1, Y_2)$".

However, the following is true: "if $Y_{1,n} \to_d Y_1$, $Y_{2,n} \to_p r$ then $h(Y_{1,n}, r) \to_d h(Y_1, r)$, where $r$ is a constant vector". This is due to the Lemma 1 below. In Lemma 1, "$Y_{n,2} \to_p r$" can be replaced by "$Y_{n,2} \to_d r$", because for a constant vector/scalar $r$, "$Y_{n,2} \to_p r$" is equivalent to "$Y_{n,2} \to_d r$" as we proved in class.

**Lemma 1.** *Supposes $\{Y_{n,1}\}_{n=1}^{\infty}$ and $\{Y_{n,2}\}_{n=1}^{\infty}$ are two sequences of random vectors/variables, and as $n \to \infty$, $Y_{n,1} \to_d Y_1$ and $Y_{n,2} \to_p r$ for a random variable $Y_1$ and a constant vector/scalar $r$. Then*

$$Y_n := \begin{pmatrix} Y_{1,n} \\ Y_{2,n} \end{pmatrix} \to_d \begin{pmatrix} Y_1 \\ r \end{pmatrix} =: Y.$$

*Proof.* (scalar case only) The set of points of continuity of the cdf of $Y$ is $\{(y_1, y_2) \in R^2 : y_1 \in C(F_{Y_1}), y_2 \neq r\}$, where $C(F_{Y_1})$ is the set of points of continuity of the cdf of $Y_1$. Consider

$(y_1, y_2) \in \{(y_1, y_2) \in R^2 : y_1 \in C(F_{Y_1}), y_2 < r\}$. Then

$$\begin{aligned} F_{Y_n}(y_1, y_2) &= \Pr(Y_{1,n} \leq y_1, Y_{2,n} \leq y_2) \\ &\leq \Pr(Y_{2,n} \leq y_2) \\ &= \Pr(Y_{2,n} - r \leq y_2 - r) \\ &= \Pr(r - Y_{2,n} > r - y_2) \\ &\leq \Pr(|Y_{2,n} - r| > r - y_2) \\ &\to_p 0 \\ &= F_Y(y_1, y_2). \end{aligned} \tag{3}$$

Consider $(y_1, y_2) \in \{(y_1, y_2) \in R^2 : y_1 \in C(F_{Y_1}), y_2 > r\}$. Then

$$\begin{aligned} F_{Y_n}(y_1, y_2) &= \Pr(Y_{1,n} \leq y_1, Y_{2,n} \leq y_2) \\ &= \Pr(Y_{1,n} \leq y_1) - \Pr(Y_{1,n} \leq y_1, Y_{2,n} > y_2) \\ &\leq \Pr(Y_{1,n} \leq y_1) \\ &\to \Pr(Y_1 \leq y_1) \\ &= \Pr(Y_1 \leq y_1, r \leq y_2) \\ &= F_Y(y_1, y_2). \end{aligned} \tag{4}$$

Also,

$$\begin{aligned} F_{Y_n}(y_1, y_2) &= \Pr(Y_{1,n} \leq y_1) - \Pr(Y_{1,n} \leq y_1, Y_{2,n} > y_2) \\ &\geq \Pr(Y_{1,n} \leq y_1) - \Pr(Y_{2,n} > y_2) \\ &= \Pr(Y_{1,n} \leq y_1) - \Pr(Y_{2,n} - r > y_2 - r) \\ &\geq \Pr(Y_{1,n} \leq y_1) - \Pr(|Y_{2,n} - r| > y_2 - r) \\ &\to \Pr(Y_1 \leq y_1) - 0 \\ &= F_Y(y_1, y_2). \end{aligned} \tag{5}$$

The above two displays imply that $F_{Y_n}(y_1, y_2) \to F_Y(y_1, y_2)$.

Now that we have shown for any $(y_1, y_2) \in \{(y_1, y_2) \in R^2 : y_2 \neq r\}$, $F_{Y_n}(y_1, y_2) \to F_Y(y_1, y_2)$. By the definition of convergence in distribution, $Y_n \to_d Y$. □

The vector case of the above lemma can be proved using the Cramér-Wold Device, the CMT, and the scalar case proof above. The Cramér-Wold device is a device to obtain the convergence in distribution of random vectors from that of real random variables. The the-

orem is given below without proof (– the proof is straightforward using mgf's/characteristic functions).

**Theorem 2** (Cramér-Wold Device)**.** *Suppose* $\{Y_n\}_{n=1}^{\infty}$ *is a sequence of random* $k$*-vectors that satisfies* $c'Y_n \to_d c'Y$ *as* $n \to \infty$ *for all* $c \in R^k$*. Then* $Y_n \to_d Y$*.*

The proof of Lemma 1 for the vector case is left as an exercise.

**Example 1.** t-statistic with estimated variance. Consider a random sample $\{X_1, ..., X_n\}$ drawn from a population distribution with mean $\mu$ and variance $\sigma^2 > 0$. The distribution of the following random variable is often of interest: $t_n = \sqrt{n}(\bar{X}_n - \mu)/S_X$, where $S_X^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. Suppose that the mean $\mu$ is known constant, then $t_n$ is a statistic and it is often called the $t$-statistic. What is the limiting distribution of $t_n$?

We have learned from the CLT that $t_n \times S_X/\sigma \to_d Z \sim N(0,1)$. We have shown in previous lectures that $S_X^2 \to_p \sigma^2$. Then by Lemma 1 above, we have

$$\begin{pmatrix} t_n \times S_X/\sigma \\ S_X^2 \end{pmatrix} \to_d \begin{pmatrix} Z \\ \sigma^2 \end{pmatrix}. \tag{6}$$

Let $h(x,y) = \sigma x/\sqrt{y}$. Then, $h$ is continuous and $t_n = h(t_n \times S_X/\sigma, S_X^2)$. Thus, the CMT applies and give us

$$t_n \to_d h(Z, \sigma^2) = Z \to N(0,1). \tag{7}$$

**Example 2.** the asymptotic distribution of the variance estimator. Consider a random sample $\{X_1, ..., X_n\}$ drawn from a population distribution with mean $\mu$ and variance $\sigma^2$ and finite fourth moment: $E|X|^4 < \infty$. We know that $\sigma^2$ is the probability limit of $S_X^2$. What is the limiting distribution of $\sqrt{n}(S_X^2 - \sigma^2)$?

$$\begin{aligned}
\sqrt{n}(S_X^2 - \sigma^2) &= \frac{\sqrt{n}}{n-1}\sum_{i=1}^n [(X_i - \bar{X}_n)^2 - \frac{n-1}{n}\sigma^2] \\
&= \frac{\sqrt{n}}{n-1}\sum_{i=1}^n [(X_i - \mu)^2 - (\bar{X}_n - \mu)^2 - \sigma^2 + \sigma^2/n] \\
&= \frac{\sqrt{n}}{n-1}\sum_{i=1}^n [(X_i - \mu)^2 - \sigma^2] - \frac{n\sqrt{n}}{n-1}(\bar{X}_n - \mu)^2 + \sqrt{n}\sigma^2/(n-1) \\
&= \frac{n}{n-1}\cdot\frac{1}{\sqrt{n}}\sum_{i=1}^n [(X_i - \mu)^2 - \sigma^2] - \frac{\sqrt{n}}{n-1}(\sqrt{n}(\bar{X}_n - \mu))^2 + \sqrt{n}\sigma^2/(n-1).
\end{aligned}$$

First, by the CLT, $\sqrt{n}(\bar{X}_n - \mu) \to_d Z_{21} \sim N(0, \sigma^2)$. Because $\sqrt{n}/(n-1) \to 0$, by the Lemma 1 above,

$$\begin{pmatrix} \sqrt{n}(\bar{X}_n - \mu) \\ \sqrt{n}/(n-1) \end{pmatrix} \to_d \begin{pmatrix} Z_1 \\ 0 \end{pmatrix}.$$

By the CMT,

$$\frac{\sqrt{n}}{n-1}(\sqrt{n}(\bar{X}_n - \mu))^2 + \sqrt{n}\sigma^2/(n-1) \to_d 0 \cdot Z_1 + 0 \cdot \sigma^2 = 0. \tag{8}$$

Second, let $Y_i = (X_i - \mu)^2$. Then $(Y_1, ..., Y_n)$ is an i.i.d. sample from a population distribution with mean $\sigma^2$ and variance $E((X_i - \mu)^2 - \sigma^2)^2$. By the CLT, we have

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} [(X_i - \mu)^2 - \sigma^2] \to_d Z_2 \sim N(0, E((X_i - \mu)^2 - \sigma^2)^2).$$

Clearly, $n/(n-1) \to 1$. Using Lemma 1 and CMT in the same way as above, we have

$$\frac{n}{n-1} \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [(X_i - \mu)^2 - \sigma^2] \to_d Z_2. \tag{9}$$

Using (8) and (9), Lemma 1 and CMT, we can conclude that

$$\sqrt{n}(S_X^2 - \sigma^2) \to_d Z_2. \tag{10}$$

**Example 3.** the **Delta Method**. Suppose $\hat{\theta}_n$ is an estimator of $\theta$ and $\hat{\theta}_n$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \Sigma). \tag{11}$$

Suppose $g : R^{d_\theta} \to R^k$ is continuously differentiable in a neighborhood of $\theta$. Then, the delta method says:

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to_d N(0, G\Sigma G'), \tag{12}$$

where $G = \frac{\partial g(\theta)}{\partial \theta'}$.

Proving that the Delta method works is a simple application of Lemma 1 and CMT. Because $g(\cdot)$ is continuously differentiable in a neighborhood of $\theta$, we can do a mean-value expansion of $g(\hat{\theta}_n)$ around $\theta$:

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) = \frac{\partial g(\bar{\theta}_n)}{\partial \theta'}[\sqrt{n}(\hat{\theta}_n - \theta)], \tag{13}$$

where $\bar{\theta}_n$ lies on the line segment connectiong $\hat{\theta}_n$ and $\theta$. Notice that $\sqrt{n}(g(\hat{\theta}_n) - g(\theta))$ is a product of two components: $\frac{\partial g(\bar{\theta}_n)}{\partial \theta'}$ and $[\sqrt{n}(\hat{\theta}_n - \theta)]$. We already know that the second component converges in distribution to $N(0, \Sigma)$. We need to show the limiting distribution (or probability limit) of the first component.

The convergence (11) implies that

$$(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \cdot [\sqrt{n}(\hat{\theta}_n - \theta)] \to_d 0 \cdot N(0, \Sigma) = 0. \tag{14}$$

Or equivalently: $\hat{\theta}_n - \theta \to_p 0$. Because $\bar{\theta}_n$ lies in between $\hat{\theta}_n - \theta$, $||\bar{\theta}_n - \theta|| \leq ||\hat{\theta}_n - \theta||$. Thus for any $\varepsilon > 0$,

$$\Pr(||\bar{\theta}_n - \theta|| > \varepsilon) \leq \Pr(||\hat{\theta}_n - \theta|| > \varepsilon) \to 0. \tag{15}$$

Therefore, $\bar{\theta}_n \to_p \theta$. This and the Slutsky Theorem (you can use the CMT here, too) implies that

$$\frac{\partial g(\bar{\theta}_n)}{\partial \theta'} \to_p \frac{\partial g(\theta)}{\partial \theta'} = G. \tag{16}$$

By (11) and (16) and Lemma 1, the convergences in the two equations hold jointly. Then by (13) and the CMT,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to_d G \cdot N(0, \Sigma) = N(0, G\Sigma G'). \tag{17}$$

**The $O_p$ and $o_p$ notation.**

**Definition 2** (Bounded in Probability)**.** A sequence of random variables $\{Y_n\}$ is bounded in probability, if

$$\lim_{B \to \infty} \lim_{n \to \infty} \Pr(|Y_n| > B) = 0.$$

*Remark.* (1) A sequence of uniformly bounded random variables ($\exists B \, |Y_n| < B \, \forall n$) is bounded in probability.

(2) A sequence of random variables that converges in distribution to a random variable $Y$ is bounded in probability.

$$\begin{aligned}
\lim_{B \to \infty} \lim_{n \to \infty} \Pr(|Y_n| > B) &= \lim_{B \to \infty} \lim_{n \to \infty} [\Pr(Y_n > B) + \Pr(Y_n < -B)] \\
&= \lim_{B \to \infty} \lim_{n \to \infty} [1 - F_{Y_n}(B) + \Pr(Y_n < -B)] \\
&\leq \lim_{B \to \infty} \lim_{n \to \infty} [1 - F_{Y_n}(B) + F_{Y_n}(-B)] \\
&= \lim_{B \to \infty} [1 - F_Y(B) + F_Y(-B)] \\
&= 0,
\end{aligned} \tag{18}$$

where the last equality holds by the properties of cdf's.

(3) A sequence of random variables that converges in probability to a constant $r$ is bounded in probability. (Excercise)

**Definition 3** ($O_p$)**.** We say $Y_n = O_p(X_n)$ if and only if $Y_n/X_n$ is bounded in probability as $n \to \infty$.

*Remark.* The definition above gives us a notation for "bounded in probability": $Y_n = O_p(1)$ if and only if $Y_n$ is bounded in probablility as $n \to \infty$.

Another useful notation is the little $o_p$:

**Definition 4** ($o_p$)**.** We say $Y_n = o_p(X_n)$ if and only if $Y_n/X_n \to_p 0$ as $n \to \infty$.

*Remark.* The definition above gives us a new notation for "convergence in probability": $Y_n \to_p r$ if anly only if $Y_n - r = o_p(1)$, or sometimes written as $Y_n = r + o_p(1)$.

*Remark.* Loosely, one can say that $Y_n$ is of smaller "stochastic order" than $X_n$ when $Y_n = o_p(X_n)$.