Lecture 4
"LLN's, CLT's, and the LIL"

"*Ab uno disce omnes* - not!"[*]

Today we will study the asymptotic behavior of sample means. We will focus on independent but not necessarily identically distributed observations. Results will come in two flavors:

LLN's	Laws of Large Numbers	$\hat{\mu} \to \mu$
CLT's	Central Limit Theorems	$\sqrt{n}(\hat{\mu} - \mu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$.

It seems that these results are very restrictive since they concern sample means, but we will eventually see that this gets us a long way.

(1) *LLN's* The simplest form of LLN and the only one we will *prove* is

WLLN:: (Chebyshev) Let $Z_1, \ldots$ be independent r.v.'s with means $\mu_1, \ldots$ and variances $\sigma_1^2 \ldots$. If $n^{-2} \sum \sigma_i^2 \to 0$ as $n \to \infty$, then $\bar{Z} = \hat{\mu} \to \bar{\mu}$.

Pf::

$$P[|\bar{Z} - \bar{\mu}| \geq \varepsilon] \leq \frac{E(\bar{Z} - \bar{\mu})^2}{\varepsilon^2} = \frac{\frac{1}{n^2} \sum \sigma_i^2}{\varepsilon^2} \to 0 \qquad \square$$

Remark:: If $\sigma_i = \sigma^2$, then $\frac{1}{n^2} \sum \sigma_i^2 = \sigma^2/n$ and Chebyshev's inequality suggests that the *lhs* probability $\to 0$ at rate $1/n$. So in this case we could say $\bar{Z} - \mu = O_p(1/n)$ rather than the weaker $\bar{Z} - \mu = o_p(1)$. A stronger result is the following:

SLLN: (Kolmogorov) For $\{Z_i, \mu_i, \sigma_i^2\}$ as above, if $\sum \sigma_i^2/i^2 < \infty$, then $\bar{Z} \to \bar{\mu}$ a.s.

Pf:: See Breiman.

Remark:: If $\sigma_i = \sigma_0^2$, we have $\sum \sigma_i^2/i^2 = \sigma_0^2 \sum i^{-2} = \sigma_0^2 \pi^2/6$. However, $\sigma_0 < \infty$ is really stronger than we need, in the iid case as the next result shows.

Thm:: (Kolmogorov) For $\{Z_i\}$ iid, iff $EZ_1 = \mu$, then $\bar{Z} \to \mu$ a.s.

Pf:: (Somewhat sketchy à la Whittle, p. 127, of the "in probability" version of the "if" part, see Breiman for the full proof.)

Let $\phi(t)$ be the cf of Z which since Z has a first moment has an expansion of the form

$$\phi_{Z_1}(t) = 1 + i\mu t + o(t) = e^{i\mu t + o(t)}$$

and thus (recall that $lim(1 + a/n)^n = e^a$),

$$\phi_{\bar{Z}}(t) = [\phi(t/n)]^n \quad = \quad e^{i\mu t + no(t/n)} \to e^{i\mu t} \qquad \square$$

An important special case, or application if you will, of the $SLLN$ involves taking

$$Z_i = I_{[X_i, \infty)}(x)$$

for iid. r.v.'s $X_i \sim F$ and fixed x. Since

$$EZ_i = P(X_i \leq x) = F(x)$$

we can infer immediately that

$$F_n(x) = \frac{1}{n} \sum I_{[X_i, \infty)}(x) \to F(x) \quad \text{a.s.}$$

This can be strengthened in the following way

Thm:: (Glivenko-Cantelli) $P(\sup_x |F_n(x) - F(x)| \to 0) = 1$.

[*]The Latin phrase is, I believe, from Virgil's *Aenid* and roughly translated means "from one example, all is revealed" so it is a phrase which succinctly captures the antithesis of statistical thinking.

Pf:: (á la Ferguson) Let $\varepsilon > 0$, and find an integer $k > 1/\varepsilon$ and numbers $-\infty = x_0 < x_1 \leq \cdots \leq x_{k-1} < x_k = \infty$ such that

$$F(x_j^-) \leq j/k \leq F(x_j) \qquad j = 1, \ldots, k-1.$$

[Interpret $F(x^-) = P(X < x)$]. Note that if $x_{j-1} < x_j$, then $F(x_j^-) - F(x_{j-1}) \leq \varepsilon$. From the *SLLN,*

$$F_n(x_j) \to F(x_j) \;\; a.s.$$

and

$$F_n(x_j^-) \to F(x_j^-) \;\; a.s.$$

for $j = 1, \ldots, k-1$. Hence,

$$\Delta_n = \max\{|F_n(x_j) - F(x_j)|, \;\; |F_n(x_j^-) - F(x_j^-)|, \;\; j = 1, \ldots, k-1\} \to 0$$

Now let x be arbitrary and find j such that $x_{j-1} < x \leq x_j$. Then

$$F_n(x) - F(x) \leq F_n(x_j^-) - F(x_{j-1}) \leq F_n(x_j^-) - F(x_j^-) + \varepsilon$$

and

$$F_n(x) - F(x) \geq F_n(x_{j-1}) - F(x_j^-) \geq F_n(x_{j-1}) - F(x_{j-1}) - \varepsilon.$$

Thus

$$\sup_x |F_x(x) - F(x)| \leq \Delta_n + \varepsilon \to \varepsilon \;\; a.s.$$

And since this holds for all $\varepsilon > 0$ the result follows. □

Vapnik (1999) calls this result "the most important result in the foundation of statistics."

(2) *CLT's*

In this section we would like to refine the analysis of the LLN's and explore the behavior of "normalized" sums of independent r.v.'s. This is particularly important for the asymptotic theory of testing. It is not good enough to have convergence results we would like to know the limiting *distributions* of possible test statistics, if we are to compute critical values.

Some preliminary results

Consider first the following very simple property of the normal (Gaussian) distribution. Suppose $Z_1, \ldots, Z_n \ldots$ are iid $\mathcal{N}(0,1)$, or $Z \sim \mathcal{N}(0, I_n)$, we know that $\alpha' Z \sim \mathcal{N}(0, \alpha'\alpha)$ so, e.g., if we take $\alpha = n^{-1/2} 1_n$, we have

$$n^{-1/2}(Z_1 + \cdots + Z_n) \sim \mathcal{N}(0,1)$$

or

$$\sqrt{n}\bar{Z} \sim \mathcal{N}(0,1)$$

This result is exact. What happens when the original Z's aren't quite normal?

Thm:: Suppose $X_1, \ldots$ are iid r.v.'s with $EX_1 = \mu$ $VX_1 = \sigma^2$, then $\sqrt{n}(\bar{X}_n - \mu)/\sigma \rightsquigarrow \mathcal{N}(0,1)$.

Pf:: Existence of moments μ, σ^2 implies the moment expansion of the cf of X_1

$$\phi_{X_1}(t) = \exp\{i\mu t - \frac{1}{2}\sigma^2 t^2 + o(t^2)\}.$$

Note that this is the cumulant version of the moment expansion from L1. Thus $S_n = X_1 + \cdots + X_n$ has $\phi_{S_n}(t) = \phi^n_{X_1}(t)$ and $U_n = \sqrt{n}(\bar{X}_n - \mu)/\sigma = \sqrt{n}(S_n/n - \mu)/\sigma$ has cf,

$$\begin{aligned}
\phi_{U_n}(t) &= Ee^{itU_n} \\
&= \phi^n_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\exp\{-\frac{i\mu t\sqrt{n}}{\sigma}\} \\
&= \left[\exp\left\{\frac{i\mu t}{\sigma\sqrt{n}} - \frac{\frac{1}{2}\sigma^2 t^2}{\sigma^2 n} + o\left(\frac{t^2}{\sigma^2 n}\right)\right\}\right]^n \exp\{-\frac{i\mu t\sqrt{n}}{\sigma}\} \\
&= \exp\left\{-\frac{1}{2}t^2 + no(t^2/n)\right\} \\
&\to \exp\{-\frac{1}{2}t^2\}
\end{aligned}$$

And the result follows by the uniqueness of the cf.

Why normal? (cf Breiman § 9.)

If $\sqrt{n}(\bar{X}_n - \mu) \rightsquigarrow X$, what must X look like? Consider

$$Z_{2n} = \frac{X_1 + \cdots + X_n + X_{n+1} + \ldots X_{2n}}{\sqrt{2n}}$$

Clearly $Z_{2n} \rightsquigarrow X$, but

$$\begin{aligned}
Z_{2n} &= \frac{X_1 + \cdots + X_n}{\sqrt{2n}} + \frac{X_{n+1} + \ldots X_{2n}}{\sqrt{2n}} \\
&= Z_{an} + Z_{bn}
\end{aligned}$$

where Z_{an} and $Z_{bn} \rightsquigarrow X/\sqrt{2}$. This is a very special property: sums of independent r.v.'s have the same df as the summands. A general characterization of such distinctions would require us to delve into stable laws, instead we simply give the following result.[2]

Thm:: If X_1, X_2 and $X = (X_1+X_2)/\sqrt{2}$ all have the same df with X_1 and X_2 independent and $EX^2 < \infty$, then X is $\mathcal{N}(0, \sigma^2)$.

Pf:: Iterating the hypothesis m times we have

$$X \stackrel{\mathcal{D}}{=} \frac{X_1 + \cdots + X_n}{\sqrt{n}} \qquad \text{for} \quad n = 2m$$

But by the previous result the rhs tends to $\mathcal{N}(0, \sigma^2)$.

Extending our basic iid result to more general cases is the subject of a vast literature. We will mention some results for the inid case leaving the dependent case for later.

The most straightforward result is

Thm:: Lyapunov (1901). Let $X_1, \ldots$ be independent with $EX_i = 0$, $EX_i^2 = \sigma_i^2 < \infty$, $E|X_i|^3 < \infty$ and $S_n^2 = \sum \sigma_i^2$. If

$$\overline{\lim} S_n^{-3} \sum_{i=1}^n E|X_i|^3 = 0,$$

then

$$S_n^{-1} \sum X_i \rightsquigarrow \mathcal{N}(0, 1).$$

[2] A nice mechanical exposition of this idea was "invented" by Galton(1877), cf. the dicussion of Stigler (1989, "The Invention of Correlation," Stat. Sci). Galton imagined making a two-level version of his quincunx: you would pour balls in the top as usual but at the first level you would get a bunch of little Gaussian hills, and then you would release these hills letting the balls pass through another set of pins. The accumulation of balls at the bottom would again look normal, showing that the resulting process was equivalent to not interupting the balls in the first place.

Pf:: Breiman (p. 187) calls this proof "very humdrum using cfs," and does it in about one page. He uses the moment expansion argument, as above.

A considerably deeper result is the following due to Lindeberg and Feller.

Thm:: As in the previous theorem, consider $X_1, \ldots$ with μ_i, σ_i, S_n^2. Suppose that

$$\frac{\sigma_n^2}{S_n^2} \to 0 \quad \text{with } S_n \to \infty \quad \text{as } n \to \infty$$

then, $S_n^{-1} \sum (X_i - \mu_i) \rightsquigarrow \mathcal{N}(0,1)$ *iff*

$$S_n^{-2} \sum_{i=1}^{n} \int_{|t-\mu_i| > \varepsilon S_n} (t-\mu_i)^2 dF_i(t) \to 0$$

for each $\varepsilon > 0$. The latter condition is usually called the Lindeberg Condition.

Proof:: (Delicate, Chung is quite good on this.)

Remarks:: The great thing is the iff. The Lindeberg Condition is essentially needed to rule out the possibility that any one of the summands has variability that dominates the others.

We will now give a number of examples from econometrics to illustrate this result.

Example Regression à la OLS

Consider the simplest through-the-origin model

$$y_i = X_i \beta + u_i \quad u_i \sim \text{iid} \quad F,$$

and the OLS estimator,

$$\hat{\beta} - \beta = (x'x)^{-1} x'u$$

Let $z_i = x_i u_i$ so $Ez_i = 0$ and $Vz_i = x_i^2 \sigma^2$ and

$$S_n^2 = \sigma^2 \sum x_i^2 \equiv \sigma^2 Q_n$$

Consider

$$S_n^{-2} \sum E[z_i^2 I(|z_i| \geq \varepsilon S_n)] = S_n^{-2} \sum x_i^2 E u_i^2 I(|u_i| \geq \varepsilon |x_i|^{-1} S_n)$$

Thus, if $\max x_i / S_n \to 0$, we have satisfied the Lindeberg Condition. We now provide some details of this argument, to keep things a bit simpler we will assume that $Eu_i^{2+\delta} < M$. Note that

$$\begin{aligned}
\int u^2 I(|u| \geq \eta) dF &\leq \int u^2 (|u|/\eta)^\delta I(|u|) \geq \eta) dF \\
&= \eta^{-\delta} \int u^{2+\delta} I(|u|) \geq \eta) dF \\
&\leq \eta^{-\delta} \int u^{2+\delta} dF \\
&\leq \eta^{-\delta} M
\end{aligned}$$

So

$$S_n^{-2} \sum x_i^2 E u_i^2 I(|u_i| \geq \eta_i) \leq S_n^{-2} \sum x_i^2 \eta_i^{-\delta} M$$

where $\eta_i = S_n/|x_i|$ so $\eta_i^{-\delta} = (|x_i|/S_n)^\delta$ so $\max\{|x_i|/S_n\} \to 0$ implies the rhs $\to 0$.

There is usually some tradeoff between F-conditions and X-conditions. Here we are trying to be general about X and the $2+\delta$-moment is stronger than absolutely necessary on F. Now we have

(i) $Q_n \to \infty$
(ii) $\max |x_i|/Q_n \to 0$
(iii) $Eu_i^{2+\delta} < M \qquad \forall_i$

If we assume in addition,

(iv) $n^{-1} Q_n \to Q_0$

which implies

$$\sqrt{n}(\hat{\beta}-\beta) = (x'x/n)^{-1} n^{-1/2} x'u$$

The Lindeberg-Feller CLT implies $n^{-1/2}(x'u) \rightsquigarrow \mathcal{N}(0, \sigma^2(x'x/n))$ and by Slutsky it follows that

$$\begin{aligned} \sqrt{n}(\hat{\beta}-\beta) &\rightsquigarrow \mathcal{N}(0, \sigma^2(x'x/n)^{-1}) \\ &\rightsquigarrow \mathcal{N}(0, \sigma^2 Q_0^{-1}). \end{aligned}$$

Suppose we replace (iv) with an alternative specification in which Q_n grows more rapidly than n. A simple example is the following important example of a linear trend

$$\begin{aligned} x_i = i \Rightarrow \sum_{i=1}^{n} x_i^2 &= 1+4+9+\cdots+n^2 = \frac{n(n+1)(2n+1)}{6} \\ &= O(n^3) \end{aligned}$$

Note that

$$\max x_i^2/Q_n = n^2 \cdot O(n^{-3}) = O(n^{-1}) \to 0.$$

so there is hope in the sense that we do have a Lindeberg Condition. But clearly normalizing by $\sqrt{n}$ won't work since

$$Var(\frac{1}{\sqrt{n}} \sum x_i u_i) = \sigma^2 Q_n = O(n^2)$$

and

$$n^{-1} \sum x_i^2 = Q_n = O(n^2)$$

so

$$\sqrt{n}(\hat{\beta}-\beta) \rightsquigarrow \mathcal{N}(0, \sigma^2 Q_n^{-1}) \to 0$$

But we can replace $\sqrt{n}$ by $n^{3/2}$ to obtain

$$n^{3/2}(\hat{\beta}-\beta) = \left(\frac{1}{n^3}\sum x_i^2\right)^{-1} n^{-3/2} \sum x_i u_i$$

now the first factor tends to 3 and the second factor $\rightsquigarrow \mathcal{N}(0, \sigma^2/3)$ so by Slutsky,

$$n^{3/2}(\hat{\beta}-\beta) \rightsquigarrow \mathcal{N}(0, 3\sigma^2).$$

Two other semi-pathological examples

(1.): Consider $x_i = 1/\sqrt{i}$ so $\sum x_i^2 = \sum 1/i \approx \log n$ then

$$\frac{\max x_i}{\sum x_i^2} = \frac{1}{\log n} \to 0$$

so here $\sqrt{\log n}(\hat{\beta}-\beta) \rightsquigarrow \mathcal{N}(0, \sigma^2)$. This is *very slow* convergence. What is the intuition here? Ulike the previous case of the linear trend where new observations are increasingly informative about the trend parameter, β, here new observations are less informative as we go further out in the sequence, but just barely informative enough so that the Lindeberg condition is satisfied and thus we get limiting normality.

(2.): Now consider $x_i = 1/i$ so $\sum x_i^2 = \sum 1/i^2 = \frac{\pi^2}{6}$ so $\sum x_i^2 \not\to \infty$ and

$$\max |x_i| / \sum x_i^2 = 6/\pi^2 \not\to 0$$

so now the Lindeberg condition fails to hold and consequently we have no CLT and therefore no limiting normality for $\hat{\beta}$. In this case the x_i's go to 0 too quickly, and x_1 exerts an effect which is never dominated by the remaining elements of the sequence.

CLTs for Dependent Sequences

In dependent cases the CLT situation becomes much more complicated and there are lots of special circumstances. Of course if there isn't too much dependence then we would expect that we would still see CLT behavior. The problem is how to make this precise. One way is α-mixing.

Given a sequence $X_1, X_2, \cdots$ and sets $A \in \sigma(X_1, \cdots, X_k)$ and $B \in \sigma(X_{k+n}, X_{k+n+1}, \cdots)$ for $k \geq 1$ and $n \geq 1$, then if there exists a sequence of real numbers $\alpha_n \to 0$ such that

$$|P(A \cap B) - P(A)P(B)| \leq \alpha_n$$

then the sequence $\{X_n\}$ is α- mixing. If $\alpha_n = 0$ for $n > m$ then the sequence is said to be m-dependent, and this is an important special case.

Recall that if the distribution of the random vector $(X_n, X_{n+1}, \cdots X_{n+j})$ doesn't depend upon n, then it is said to be stationary. A proof of the following result is given in Billingsley (Thm 27.5, p 316, ed I).

CLT for α-mixing sequences Suppose $X_1, X_2, \cdots$ is stationary and α-mixing with $\alpha_n = \mathcal{O}(n^{-5})$, $\mathbb{E}X_n = 0$ and $\mathbb{E}X_n^{12} < \infty$. Set $S_n = X_1 + \cdots + X_n$, if

$$n^{-1}V(S_n) \to \sigma^2 \equiv \mathbb{E}X_1^2 + \sum_{k=1}^{\infty} \mathbb{E}X_1 X_{k+1}$$

converges absolutely with $\sigma^2 > 0$, then $S_n/(\sigma\sqrt{n}) \rightsquigarrow \mathcal{N}(0,1)$.

In many dependent situations, asymptotics boil down to "finding the martingale" and then applying a martingale CLT. What's a martingale?

Definition Let $X_1, X_2, \cdots$ be a squence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{F}_1, \mathcal{F}_2, \cdots$ be a sequence of σ-fields in $\mathcal{F}$ then the sequence $\{(X_n, \mathcal{F}_n) : \ n = 1, 2, \cdots\}$ is a martingale if:
(a) $\mathcal{F}_n \subset \mathcal{F}_{n-1}$
(b) X_n is measureable $\mathcal{F}_n$
(c) $\mathbb{E}|X_n| < \infty$
(d) $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ with probability one.

Gambling is the usual example with $\mathcal{F}_n$ representing (wealth) information at step n. The following result generalizes the Lindeberg-Feller CLT to an important class of dependent settings. We can decompose a martingale, $\{X_n\}$ as

$$X_n = X_0 + \sum_{j=1}^{n} \xi_i$$

where $\xi_j = X_j - X_{j-1}$ is called a martigale difference sequence and has the property that $E(\Xi_{n+1}|\mathcal{F}_n) = 0$. Thus, $EX_n = EX_0$ and provided that $EX_n^2 < \infty$, the ξ_j's are square integrable and uncorrelated, so

$$EX_n^2 = EX_0^2 + \sum EX_j^2.$$

To see this, assume wlog that $X_0 = 0$ and consider for $j \leq k \leq n$,

$$\begin{aligned} E(\xi_j \xi_{k+1} &= EE(\xi_j \xi_{k+1}|\mathcal{F}_k) \\ &= E\xi_j E(\xi_{k+1}|\mathcal{F}_k) \\ &= 0 \end{aligned}$$

so increments are uncorrelated and we can write,

$$\begin{aligned} EX_n^2 &= E(\sum \xi_j)^2 \\ &= \sum \sum E\xi_j \xi_k \\ &= \sum E\xi_j^2 + 2\sum_{j \neq k} \xi_j \xi_k \\ &= \sum E\xi_j^2. \end{aligned}$$

Thus we see that the variances of the martingale increments are summable. The following generalization of the Lindeberg-Feller theorem is usually attributed to Paul Levy.

Martingale CLT Suppose that for each n, $X_{n1}, X_{n2}, \cdots$ is a martingale with respect to the filtration $\mathcal{F}_{n1}, \mathcal{F}_{n2}, \cdots$ Define $\xi_{nk} = X_{nk} - X_{n,k-1}$, and suppose that $\sigma_{nk}^2 = \mathbb{E}(\xi_{nk}^2 | \mathcal{F}_{n,k-1}) < \infty$. If $\sum_{k=1}^{\infty} \sigma_{nk}^2 \to \sigma^2$ and for every $\epsilon > 0$,

$$\sum_{k=1}^{\infty} \mathbb{E}\xi_{nk}^2 I_{(|\xi_{nk}| \geq \epsilon)} \to 0$$

then $S_n = \sum_{k=1}^{\infty} \xi_{nk} \rightsquigarrow \mathcal{N}(0, \sigma^2)$.

(3) *LIL's*

Another remarkable class of results are the Laws of the Iterated Logarithm. Here is the most basic version.

Thm:: Let $Y_1, Y_2, \ldots$ be iid $r.v.$s with mean 0 and variance 1, then

$$\lim_{n\to\infty} \sup \frac{Y_1 + \ldots + Y_n}{\sqrt{n \log \log n}} = \sqrt{2} \qquad a.s.$$

Rem:: Since we know that $S_n/\sqrt{n} \rightsquigarrow \mathcal{N}(0,1)$ under these conditions it seems that the extra factor $\sqrt{\log \log n}$ is just what is needed to keep $S_n/\sqrt{n \log \log n}$ within a compact interval eventually.

Slutsky's Lemma implies $Z_n = S_n/\sqrt{n \log \log n} \to 0$ so Z_n is eventually in $[-\varepsilon, \varepsilon]$ for any $\varepsilon > 0$. But LIL says that Z_n is in $(\sqrt{2} \pm \varepsilon)$. What gives? The fact is that both are possible.

Implications:: Suppose $EY_i = \mu$ and $VY_i = 1$, then

$$P(\bar{Y}_n - 2/\sqrt{n} \leq \mu \leq \bar{Y}_n + 2/\sqrt{n}) \to \Phi(2) - \Phi(-2) \cong .95$$

Suppose $\mu = 0$, then $\mu = 0$ is contained in the interval iff

$$|Z_n| \leq \frac{2}{\sqrt{n \log \log n}}$$

The LIL implies that $Z_n \in (\sqrt{2} - \varepsilon, \sqrt{2} + \varepsilon)$ infinitely often. Since $2/\sqrt{\log \log n}$ is near zero we can be sure that $\mu = 0$ is outside the interval infinitely often.

This is clarified by thinking of a large number of independent experiments each of which compute a sequence of CI's as $n \to \infty$. At each n, 5% of them will exclude the true value $\mu = 0$, but as $n \to \infty$ we may expect that this is not the same 5%, this is what is predicted by the LIL. Think of many econometricians making intervals and as $n \to \infty$. There would always be 5% who had a bad interval, but the composition of this 5% would fluctuate.

Note also that the whole paradox (if it really is a paradox) is based on the idea that $2/\sqrt{\log \log n}$ is close to zero and it is nonsensical if this quantity is larger than $\sqrt{2}$. Thus, for example for $2/\sqrt{\log \log n} < 1$ requires that n be at least 10^{23}.

How good is the CLT?

As a final result for this lecture we offer a classical result on the accuracy of the CLT.

Thm:: (Berry-Esséen) Let $\{X_i\}$ be iid $r.v.$'s with mean μ and variance $\sigma^2 > 0$, then

$$\sup_t |F_n(t) - \Phi(t)| \leq \frac{33}{4} \; \frac{E|X-\mu|^3}{\sigma^3\sqrt{n}}$$

where $F_n(t) = P(Z_n \leq t)$ and

$$Z_n = \frac{n^{-1/2}\sum(X_i - \mu)}{\sigma}.$$

Remark:: This can be extended to *nid* cases painfully. See Serfling for references. Note that the accuracy of the approximation is $O(1/\sqrt{n})$ but the crucial, and remarkable, fact is the nature of the constant, which has been variously improved since the work of Berry and Esséen in the 1940's.

References

Vapnik, V. (1999) *The Nature of Statistical Learning Theory*, Springer.