§ 2.3. Ramification, extension and application.

We examine various types of CLT, beginning with the most original of all, de Moivre's calculation of binomial probabilities. The rates of convergence of CLT is discussed in the form of universal bound and local approximations. We show some application of CLT in statistical analysis, where it plays a fundamental role. The exposition of this section only serves the purpose of introducing related result without attempting to produce proofs.

(i). *Miscellaneous central limit theorems.*

Theorem 2.6 (De Moivre-Laplace local CLT) *Suppose $X, X_1, ..., X_n, ...$ are iid Bernoulli r.v.s, i.e., $P(X=1) = p = 1-q = 1-P(X=0)$. Let $S_n = \sum_{i=1}^n X_i$ (as always) and $x = x_{n,k} = (k-np)/\sqrt{npq} = o(n^{1/6})$. Then,*

$$P(S_n = k) = \frac{\phi(x)}{\sqrt{npq}}(1 + a_n(x))$$

where

$$|a_n(x)| \le \frac{A}{\sqrt{n}}\Big(\frac{1}{\sqrt{n}} + |x| + |x|^3\Big)$$

for some constant A.

Remark. Notice that rate of convergence is about $1/\sqrt{n}$. The proof can be carried out by brute force calculation without appealing to characteristic functions. As

$$P(S_n = k) = P(Bin(n,p) = k) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k},$$

it is not difficult to see that the key is the approximation of factorials, which is, via Sterling's formula:

$$n! = \sqrt{2\pi}\sqrt{n}\Big(\frac{n}{e}\Big)^n e^{\epsilon_n}, \quad \text{with } \frac{1}{12n+1} < \epsilon_n < \frac{1}{12n}.$$

(Actually the formula was first outlined in De Moivre's *Doctrine of Chances* with constant coefficient c and Sterling (1730) identified c as $\sqrt{2\pi}$.)

The above local CLT leads to the (global) De Moivre-Laplace CLT

Theorem 2.7 (De Moivre-Laplace CLT) *Suppose $X, X_1, ..., X_n, ...$ are iid with Bernoulli r.v.s, i.e., $P(X=1) = p = 1-q = 1-P(X=0)$. Then, for any two integers $x_n \le y_n$ such that $x_n = np + o(n^{2/3})$ and $y_n = np + o(n^{2/3})$.*

$$P(x_n \le S_n \le y_n) \approx \Phi\Big(\frac{y_n - np + 1/2}{\sqrt{npq}}\Big) - \Phi\Big(\frac{x_n - np + 1/2}{\sqrt{npq}}\Big) \tag{2.7}$$

As a result,

$$\frac{S_n - np}{\sqrt{npq}} \to N(0,1). \tag{2.8}$$

(2.7) provides theoretical support of normal approximation of binomial distribution *with continuity correction.* (2.8) is the standard CLT for binomial random variables.

Theorem 2.8 (CLT for U-statistics) *Suppose $X, X_1, ..., X_n, ...$ are iid r.v.s. Suppose g is a bivariate symmetric function (e.g. $g(x,y) = \min(x,y)$) such that $g(X_1, X_2)$ has mean 0 and $b(X_1) \equiv E(g(X_1,X_2)|X_1)$ has positive variance σ^2. Let*

$$U_n = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} g(X_i, X_j)$$

be the so-called U-statistic. Then,

$$\frac{n^{1/2}U_n}{2\sigma} \to N(0,1).$$

U-statistic is common and has broad applications in statistical analysis. Decompose the U-statistic into

$$U = \frac{2}{n}\sum_{j=1}^{n} b(X_j) + \frac{2}{n(n-1)} \sum_{1\le i<j\le n} \{g(X_i,X_j) - b(X_i) - b(X_j)\} \tag{2.9}$$

The second term on the left hand side is a *degenerate* U-statistic, which converges to 0 at a faster rate than $n^{-1/2}$. Then the CLT follows from the standard CLT for iid r.v.s. The decomposition (2.9) is the most useful when dealing with U-statistics.

Theorem 2.9 (CLT FOR MARTINGALES) *Suppose $(X_n, \mathcal{F}_n), n \ge 1$ is a sequence of martingales with mean 0 and variance σ_n^2, $n \ge 1$. $\mathcal{F}_0$ is the trivial σ-algebra. Suppose Lindeberg condition holds and, moreover,*

$$\frac{1}{s_n^2}\sum_{j=1}^{n} E|E(X_j^2|\mathcal{F}_{j-1}) - \sigma_j^2| \to 0. \tag{2.10}$$

Then, $S_n/s_n \to N(0,1)$.

The proof is actually analogous to the proof of Lindeberg-Feller CLT. The only difference is that X_is are now not independent. Therefore $E(e^{itS_n/s_n}) \ne \prod_{j=1}^n E(e^{itX_j/s_n})$. However,

$$E(e^{itS_n/s_n}) = E\Big(\prod_{j=1}^{n} E(e^{itX_j/s_n}|\mathcal{F}_{j-1})\Big).$$

Then, the fact that $E(X_j|\mathcal{F}_{j-1}) = 0$ and that $\mathrm{var}(X_j|\mathcal{F}_{j-1})$ are close enough to σ_j^2 by (2.10) ensure the proof goes through.

Theorem 2.10 (HALL-HEYDE CLT FOR ARRAYS) *Suppose $(X_{n,j}, \mathcal{F}_{n,j}), 1 \le j \le n$ is a sequence of r.v.s with finite positive variances and $X_{n,j} \in \mathcal{F}_{n,j}$ and $\mathcal{F}_{n,j} \subseteq \mathcal{F}_{n,j+1}$. Let $Z_n = \max\{X_{n,j} : 1 \le j \le n\}$. If*

$$\sum_{j=1}^{n} X_{n,j}^2 \to a^2 > 0, \quad Z_n \to 0, \ \textit{in probability}, \quad \sup_{n\ge 1} E(Z_n^2) < \infty, \tag{2.11}$$

$$\sum_{j=1}^{n} E(X_{n,j}|\mathcal{F}_{n,j-1}) \to 0 \quad \textit{and} \quad \sum_{j=1}^{n}\{E(X_{n,j}|\mathcal{F}_{n,j-1})\}^2 \to 0, \quad \textit{in probability}$$

Then, $S_n/a \equiv \sum_{j=1}^n X_{n,j}/a \to N(0,1)$.

Condition (2.11) can be replaced by a Lindeberg condition with condition expectation:

$$\sum_{j=1}^{n} E(X_{n,j}^2 1_{\{|X_{n,j}|>\epsilon\}}|\mathcal{F}_{n,j-1}) \to 0 \qquad \text{in probability}$$

plus

$$\sum_{j=1}^{n} \mathrm{var}(X_{n,j}|\mathcal{F}_{n,j-1}) \to a^2 \qquad \text{in probability.}$$

And the central limit theorem still hold.

(ii). *Rates of convergence*

So far, CLTs (except for De Moivre-Laplace local CLT) only claims the limit is normal distribution. Berry-Esseen bound claims the rate is about $1/\sqrt{n}$, under ideal circumstances.

Theorem 2.11 (Berry (1941)-Esseen (1942)) *Suppose $X_1, X_2, ...$ are independent with mean 0 and variance $\sigma_n^2 > 0$. Let $2 < \delta \le 3$. Then, there exists a universal constant C_δ such that*

$$\sup_x |P(S_n/s_n < x) - \Phi(x)| \le C_\delta \frac{\sum_{j=1}^n E(|X_j|^\delta)}{s_n^\delta}$$

In particular,

$$\sup_x |P(S_n/s_n < x) - \Phi(x)| \le C_3 \frac{\sum_{j=1}^n E|X_i|^3}{s_n^3}.$$

If $X, X_1, X_2, ...$ are iid with mean 0 and variance $\sigma^2 > 0$. Then, there exists a universal constant c_δ such that

$$\sup_x |P(S_n/\sqrt{n\sigma^2} < x) - \Phi(x)| \le \frac{c_\delta}{n^{\delta/2}} \frac{E(|X|^\delta)}{\sigma^\delta},$$

and, in particular,

$$\sup_x |P(S_n/\sqrt{n\sigma^2} < x) - \Phi(x)| \le \frac{c_3}{n^{1/2}} \frac{E(|X|^3)}{\sigma^3}.$$

Note that all the above inequalities hold for *all* n, (not in the sense of taking limit), and that the universality of the constants means they are all all distributions of X. The proofs generally involve characteristic function and can be rather technical. The above theorem implies CLT when $\sum_{j=1}^n E(|X_j|^\delta)/s_n^\delta \to 0$. It is the most useful when the $\delta = 3$, as it specifies the the fastest rate of convergence. One of the highly attractive problems in probability theory is to find the best possible C_δ and c_δ. Esseen (1945) proves $0.4097 \le c_3 \le 7.5$, and $c_3 \le 0.7975$ (Van Beek, 1972), 0.7655 (Shiganov, 1986) and the best thus far, 0.7056 (Shevtsova, 2007).

Berry-Esseen bound is a rate of uniform convergence. Another similar type of approximation, called Edgeworth expansion, confines on the rate of convergence at a fixed x but provides explicit expression of the coefficients at all orders.

Theorem 2.12 (Edgeworth expansion) *Suppose $X, X_1, ...$ are iid with mean 0 and positive variance σ^2, satisfying the Cramer condition (nonlattice distribution of X). Then,*

$$P(S_n/\sqrt{n\sigma^2} < x) = \Phi(x) + \frac{\kappa_3}{6\sqrt{n}} \Phi^{(3)}(x) + O(1/n).$$

where $\kappa_3 = E(X^3)/\sigma^3$ is the skewness of X.

The above theorem is a simplified form of Edgeworth expansion, which can be of arbitrary order of the series. We provide a heuristic understanding as follows. For simplicity, assume X has density f, cdf F, mean 0 and variance 1 and all finite moments. Then, its characteristic function is

$$\psi(t) = E(e^{itX}) = e^{\log E(e^{itX})} = e^{\sum_{j=1}^\infty \kappa_j (it)^j/j!}$$

where κ_j is the j-th derivative of $\log E(e^{itX})$ at $t = 0$, which is called j-th cumulants of X. $\kappa_1 = E(X) = 0$, $\kappa_2 = \mathrm{var}(X) = \sigma^2 = 1$ and $\kappa_3 = E(X^3)$,... Let f_n be the density of $S_n/\sqrt{n}$. The characteristic function of $S_n/\sqrt{n}$ is

$$\begin{aligned}
& \int e^{itx} f_n(x)dx = E(e^{itS_n/\sqrt{n}}) = \psi(t/\sqrt{n})^n = e^{n \log \psi(t/\sqrt{n})} = e^{n \sum_{j=2}^\infty \kappa_j (it/\sqrt{n})^j/j!} \\
= \;& e^{-t^2/2} e^{n \sum_{j=3}^\infty \kappa_j (it/\sqrt{n})^j/j!} = e^{-t^2/2} (e^{\kappa_3 (it)^3/6n^{-1/2} + O(1/n)}) \\
= \;& e^{-t^2/2} (1 + \kappa_3 (it)^3/6n^{-1/2} + O(1/n)) \\
= \;& \int e^{itx} \phi(x) dx + \frac{\kappa_3}{6\sqrt{n}} \int e^{itx} (-1)^3 \frac{d^3}{dx^3} \phi(x) dx + O(1/n),
\end{aligned}$$

since $(it)^k \int e^{itx} f(x)dx = \int e^{itx} (-1)^m f^{(m)}(x)dx$. As a result, it should hold that

$$f_n(x) \approx \phi(x) - \frac{\kappa_3}{6\sqrt{n}} \phi^{(3)}(x) + O(1/n).$$

And through integration,

$$P(S_n/\sqrt{n} < x) = \Phi(x) - \frac{\kappa_3}{6\sqrt{n}}(x^2-1)\phi(x) + O(1/n).$$

(iii). CLT for processes.

Suppose we have random processes or functions $\{\xi_n(t) : t \in [0,1]\}$. The weak convergence of r.v.s or their distributions or measures can be extended to the define weak convergence of processes or their measures, with finite dimensional convergence plus certain tightness condition on processes (stochastic equi-continuity). We skip the details.

Theorem 2.13. (Donsker's invariance principle, 1951) *Suppose $X_1, X_2, ...$ are iid with mean 0 and variance $\sigma^2 > 0$. Define $S_0 = 0$ and $S_t = S_{[t]} + (t-[t])X_{[t]+1}$ for $t \geq 0$. Here [] means the integer part. Then,*

$$\{\frac{S_{nt}}{\sqrt{n}\sigma} : t \in [0,1]\} \to \{B_t : t \in [0,1]\}$$

"in distribution" or weakly, where $B_t : t \geq 0$ is Brownian motion.

For fixed t, the classical CLT implies $S_{nt}/\sqrt{n\sigma^2}$ converges to $N(0,t)$. It is not hard to generalize it to convergence of finite dimension.

Brownian motion on $[0,\infty)$ is can be regarded as r.v.s with taking values of continuous functions on $[0,\infty)$. It induces a probability measure on the space of continuous functions on $[0,\infty)$, which is often called Wiener measure.

Theorem 2.14. Weak convergence of empirical distribution *Suppose $X_1, ..., X_n, ...$ are iid with continuous cdf F. Let $F_n(t) = (1/n)\sum_{j=1}^n 1_{\{X_i \leq t\}}$ be the empirical distribution. Then,*

$$\{\sqrt{n}(F_n(t) - F(t)) : t \in (-\infty,\infty)\} \to \{(1-F(t))B_{F(t)/\{1-F(t)\}} : t \in (-\infty,\infty)\}$$

which is a so-called Brownian bridge.

As one application of this theorem,

$$\sqrt{n}\sup_x |F_n(x) - F(x)| \to max_{t\in[0,1]}(1-t)|B_{t/(1-t)}| \quad \text{in distribution.}$$

Observe that $\sup_x |F_n(x) - F(x)|$ is the Kolmogorov-Smirnov statistic.

(iii). *Some applications.*

Example 2.8 Asymptotic normality of quantile estimation Suppose $X_1, ...X_n$ are iid with positive density $f(t)$ at p-th quantile $F^{-1}(p)$. Let $X_{(1)} < X_{(2)} < ... < X_{(n)}$ be the order statistics. Suppose $k = k_n$ is such that $k/n \to p$. Then,

$$\sqrt{n}(X_{(k)} - F^{-1}(k/n)) \to N(0, \sigma^2),$$

where $\sigma^2 = p(1-p)/f^2(F^{-1}(p))$.

Proof. Let $t_n = \sigma t/\sqrt{n} + F^{-1}(k/n)$. Then, $F(t_n) \approx k/n \approx p$ and

$$k - nF(t_n) \approx k - n(k/n + f(F^{-1}(p))\sigma t/\sqrt{n}) \approx -\sqrt{n}f(F^{-1}(p))\sigma t$$

Write

$$P(\frac{\sqrt{n}(X_{(k)} - F^{-1}(k/n))}{\sigma} \leq t) = P(X_{(k)} \leq \sigma t/\sqrt{n} + F^{-1}(k/n)) = P(\sum_{j=1}^n 1_{\{X_j \leq t_n\}} \geq k)$$

$$= P\Big(\frac{\sum_{j=1}^n 1_{\{X_j \leq t_n\}} - nF(t_n)}{\sqrt{nF(t_n)(1-F(t_n))}} \geq \frac{k - nF(t_n)}{\sqrt{nF(t_n)(1-F(t_n))}}\Big)$$

$$\approx 1 - \Phi\Big(\frac{k - nF(t_n)}{\sqrt{nF(t_n)(1-F(t_n))}}\Big) \approx 1 - \Phi\Big(\frac{-\sqrt{n}f(F^{-1}(p))\sigma t}{\sqrt{np(1-p)}}\Big) \approx \Phi(t).$$

Example 2.9. Renewal processes Suppose $X, X_1, X_2, ...$ are iid positive random variables with mean μ and variance $\sigma^2 > 0$. Let $N_t = \max\{n : S_n \le t\}$, which is called renewal process. Then, as $t \to \infty$,

$$\frac{\mu N_t - t}{\sigma\sqrt{t/\mu}} \to N(0,1).$$

Notice that $\{N_t < k\}$ is the same as $S_k > t$. Then,

$$\begin{aligned} & P\Big(\frac{\mu N_t - t}{\sigma\sqrt{t/\mu}} < x\Big) = P(N_t < x\sigma\sqrt{t/\mu^3} + t/\mu) \\ = & \quad P(S_k > t) \quad k \text{ is the integer part of } x\sigma\sqrt{t/\mu^3} + t/\mu \\ = & \quad P(\frac{S_k - k\mu}{\sqrt{k\sigma^2}} > \frac{t - k\mu}{\sqrt{k\sigma^2}}) \approx 1 - \Phi\Big(\frac{t-k\mu}{\sqrt{k\sigma^2}}\Big) \approx \Phi(x). \end{aligned}$$

Example 2.10. Estimating functions. Suppose $X_1, ..., X_n$ are iid following a distribution with density $f(x;\theta)$ with the parameter θ. Let $g(X_i;\theta)$ be such that $E(g(X_i;\theta)) = 0$. An estimator of θ based on observed values of $X_1, ..., X_n$, denoted as $\hat\theta_n$, can be defined as the root of

$$\sum_{j=1}^n g(X_j;\theta) = 0.$$

Under regularity conditions,

$$\sqrt{n}(\hat\theta_n - \theta) \to N(0, \sigma^2)$$

where σ^2 can be estimated by the so-called sandwich formula:

$$(1/n)\sum_{j=1}^n \dot g(X_j;\hat\theta)\sum_{j=1}^n g^2(X_j;\hat\theta)\sum_{j=1}^n \dot g(X_j;\hat\theta).$$

where $\dot g$ is derivative of g with respect to θ.
When $g(x;\theta) = \dot f(x;\theta)/f(x;\theta)$, the estimator $\hat\theta$ is the maximum likelihood estimator. And

$$\sqrt{n}(\hat\theta - \theta) \to N(0, I^{-1}(\theta))$$

where

$$I(\theta) = -E(\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)) = E\{(\frac{\partial}{\partial\theta}\log f(X;\theta))^2\}$$

is the Fisher information. As an application, one might use the Wald test statistic

$$(\hat\theta - \theta_0)^2/\hat{\text{var}}(\hat\theta) \sim \chi_1^2, \quad \text{approximately},$$

to test the hypothesis of $\theta = \theta_0$, where $\hat{\text{var}}(\hat\theta)$ is the estimator of the variance of $\hat\theta$.

(v) *Stable laws*

When the r.v.s are iid but do not have finite second moment, the partial sum after proper normalization may still converge in distribution, but the limiting distribution can be very different from normal distribution.

Theorem 2.15 *Suppose $X, X_1, ...$ are iid such that the limit of $P(X > x)/P(|X| > x)$ as $x \to \infty$ exists and $P(|X| > x) = x^{-\alpha}L(x)$ with $0 < \alpha < 2$, where L satisfies $L(tx)/L(x) \to 1$ as $x \to \infty$ for all fixed $t > 0$, and is called a slowly varying function. Then, $(S_n - b_n)/a_n$ converges to a nondegenerate distribution called stable law, where $b_n = nE(X1_{\{|X| \le a_n\}})$ and $a_n = \inf x : P(|X| > x) \le 1/n$.*

A typical example is that X follows Cauchy distribution, corresponding to $\alpha = 1$. The limiting distribution is still Cauchy, which is one of the stable laws. (Normal distribution is also a stable law with $\alpha = 2$.)