Chapter 2. Central Limit Theorem.

Central limit theorem, or DeMoivre-Laplace Theorem, which also implies the weak law of large numbers, is the most important theorem in probability theory and statistics. For independent random variables, Lindeberg-Feller central limit theorem provides the best results. Throughout this chapter, random variables shall not take values in ∞ or $-\infty$ with positive chance.

§ 2.1. Convergence in distribution and characteristic functions.

Convergence in distribution, which can be generalized slightly to weak convergence of measures, has been introduced in Section 1.2. This section provides a more detailed description.

(i). *Definition, basic properties and examples.*

Recall that in Section 1.3, we have already defined convergence in distribution for a sequence of random variables. Here we present the same definition in terms of weak convergence of their distributions. We first note that a function F is a cdf if and only if it is right continuous, nondecreasing with $F(t) \to 1$ and 0 when $t \to \infty$ and $-\infty$, respectively.

Definition. A sequence of distribution function F_n is called converging to another distribution function F_∞ *weakly*, if

(1) $F_n(t) \to F_\infty(t)$ for every continuity points of F_∞; or

(2), $\liminf_n F_n(B) \geq F_\infty(B)$ for every open set B in $(-\infty, \infty)$; or

(3) $\limsup_n F_n(C) \leq F_\infty(C)$ for every closed set C in $(-\infty, \infty)$; or

(4) $\int g(x)dF_n(x) \to \int g(x)dF_\infty(x)$ for every continuous function g.

Here $F_n(A)$ is defined as $\int_A dF_n(x) = \int 1_{x\in A}dF_n(x)$ for any Borel set A. The above four claims are equivalent to each other, as proved in Section 1.3.

Remark. If F_∞ is continuous, the inequalities in (2) and (3) are actually equalities. On the other hand, if X_n all takes integer values, then $X_n \to X$ in distribution is equivalent to $P(X_n = k) \to P(X = k)$ for all integer values k.

Remark. (Sheffe's Theorem) Suppose X_n has density function $f_n(\cdot)$ and $f_n(t) \to f(t)$ for every finite t and f *is a density function.* Then, $X_n \to X$ in distribution, where X has density f. This can be shown quite straightforwardly as follows:

$$2 = \int \liminf_n (f_n + f - |f_n(x) - f(x)|)dx \leq \liminf_n \int (f_n(x) + f(x) - |f_n(x) - f(x)|)dx$$
$$= \liminf_n \Big(2 - \int |f_n(x) - f(x)|dx\Big) = 2 - \limsup_n \int |f_n(x) - f(x)|dx.$$

Certainly, for any Borel set B,

$$P(X_n \in B) - P(X \in B) = \int_B (f_n(x) - f(x))dx \leq \int |f_n(x) - f(x)|dx \to 0.$$

□ In the above proof, we have used Fatou lemma with Lebesgue measure. In fact, the monotone convergence theorem, Fatou lemma and dominated convergence theorem that we have established with probability measure all hold with σ-finite measures, including Lebesgue measure.

Remark. (Slutsky's Theorem) Suppose $X_n \to X_\infty$ in distribution and $Y_n \to c$ in probability. Then, $X_nY_n \to cX_\infty$ in distribution and $X_n + Y_n \to X_n - c$ in distribution.

We leave the proof as an exercise.

In the following, we provide some classical examples about convergence in distribution, only to show that there are a variety of important limiting distributions besides the normal distribution as the limiting distribution in CLT.

Example 2.1. (Convergence of maxima and extreme value distributions) Let $M_n = \max_{1\le i\le n} X_i$ where X_i are iid r.v.s with c.d.f. $F(\cdot)$. Then,

$$P(M_n \le t) = P(X_1 \le t)^n = F(t)^n.$$

As $n \to \infty$, the limiting distribution of properly scaled M_n, should it converge, should only be related with the right tail of the distribution of $F(\cdot)$, i.e., the $F(x)$ when x is large. The following are some examples.

(a). $F(x) = 1 - x^{-\alpha}$ for some $\alpha > 0$ and all large x. Then, for any $t > 0$,

$$P(M_n/n^{1/\alpha} < t) = (1 - n^{-1}t^{-\alpha})^n \to e^{-t^{-\alpha}}$$

(b). $F(x) = 1 - |x|^\beta$ for $x \in [-1, 0]$ and some $\beta > 0$. Then, for any $t < 0$,

$$P(n^{1/\beta} M_n \le t) = (1 - n^{-1}|t|^\beta)^n \to e^{-|t|^\beta}$$

(c). $F(x) = 1 - e^{-x}$ for $x > 0$, i.e., X_i follows exponential distribution. Then for all t,

$$P(M_n - \log n \le t) \to e^{-e^{-t}}$$

These limiting distributions are called extreme value distributions.

Example 2.2. (Birthday problem) Suppose $X_1, X_2, ...$ are iid with uniform distribution on the integers $\{1, 2, ..., N\}$ with $n < N$ and , Let

$$T_N = \min\{k : \text{ there exists a } j < k \text{ such that } \{X_j = X_k\}\ \}.$$

Then, for $k \le N$,

$$\begin{aligned} P(T_N > k) &= P(\ X_1, ..., X_k \text{ all take different values }) \\ &= \prod_{j=2}^{k} \Big(1 - P(\ X_j \text{ takes one of the values of } X_1, .., X_{j-1})\Big) \\ &= \prod_{j=2}^{k} (1 - \frac{j-1}{N}) = \exp\{\sum_{j=1}^{k-1} \log(1 - j/N)\} \end{aligned}$$

Then, for any fixed $x > 0$, as $N \to \infty$,

$$P(T_N/N^{1/2} > x) = P(T_N > N^{1/2}x) \approx \exp\{\sum_{1\le j<N^{1/2}x} \log(1 - j/N)\}$$

$$\approx \exp\{-\sum_{1\le j<N^{1/2}x} j/N\} \approx \exp\{-(1/N)N^{1/2}x(N^{1/2}x + 1)/2\} \approx \exp\{-x^2/2\}$$

In other words, $T_N/N^{1/2}$ converges in distribution to a distribution $F(t) = 1 - \exp(-t^2/2)$ for $t \ge 0$.

Suppose now $N = 365$. By this approximation, we have $P(T_{365} > 22) \approx .5153$ and $P(T_{365} > 50) \approx .0326$, meaning that, with 22 (50) people there is about half (3%) probability that all of them have different birthday.

Example 2.3. (Law of rare events) Suppose there are totally n flights worldwide each year, and each flight has chance p_n to have an accident, independent of rest flights. There is on average λ accidents a year worldwide. The distribution of the number of accidents is $B(n, p_n)$ with np_n close to λ. Then this distribution approximates Poisson distribution with mean λ, namely,

$$Bin(n, p_n) \to \mathcal{P}(\lambda) \qquad \text{if} \quad n \to \infty \text{ and } np_n \to \lambda > 0.$$

Proof. For any fixed $k \geq 0$, and $n \geq k$

$$
\begin{aligned}
P(Bin(n, p_n) = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{n!}{k!(n-k)!} \frac{(np_n)^k}{n^k} \frac{(1-p_n)^n}{(1-p_n)^k} \\
&= \frac{1}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \frac{(np_n)^k e^{n\log(1-p_n)}}{(1-p_n)^k} \\
&\to \frac{\lambda^k e^{-\lambda}}{k!}, \qquad \text{as } n \to \infty.
\end{aligned}
$$

□

EXAMPLE 2.4. (THE SECRETARY/MARRIAGE PROBLEM) Suppose there are n secretary to be interviewed one by one and, right after each interview, you must make immediate decision of "hire or fire" the interviewee. You observe only the relative ranks of the interviewed candidates. What is the optimal strategy is maximize the chance of hiring the best of the n candidates? (Assume no ties of performance.)

One type of strategy is to give up the first m candidates, whatever their performance in the interview. Afterwards, the one that outperforms all previous candidates is hired. In other words, starting from $m+1$-th interview, the first candidate that outperforms the first m candidates is hired. Or else you settle with the last candidate. The chance that the k-th best among all n candidates is hired is

$$
\begin{aligned}
P_k &= \sum_{j=m+1}^{n} P(\text{ the } k\text{-th best is the } j\text{-th interviewee and is hired}) \\
&= \sum_{j=m+1}^{n} \frac{1}{n} P(\text{the best among first } j-1 \text{ appears in the first } m, \\
&\quad \text{the } j\text{-th candidate is the } k\text{-th best, and the } k-1 \text{ best all appear after the } j\text{-th candidate.}) \\
&\approx \sum_{j=m+1}^{n} \frac{m}{j-1} \times \frac{1}{n} \times (\frac{n-j}{n})^{k-1}
\end{aligned}
$$

Let $n \to \infty$, and $m \approx nc$ where c is the percentage of the interviews to be given up. Then the probability of hiring the k-th best

$$
P_k \approx c \sum_{j=m}^{n} \frac{1}{j}(1-j/n)^{k-1} \approx c \int_c^1 \frac{(1-x)^{k-1}}{x} dx = cA_k, \quad \text{say.}
$$

Since $A_{k+1} = A_k - (1-c)^k/k$, for $k \geq 1$, and $A_1 = -\log c$, it follows that

$$
P_k \to c\Big(-\log c - \sum_{j=1}^{k-1} \frac{(1-c)^j}{j} \Big), \qquad \text{as} \qquad n \to \infty.
$$

In particular, $P_1 \to -c\log c$. The function $c\log c$ is maximized at $c = 1/e = 0.368$. The best strategy is to give up the first 36.8% of the interviews and then hire the best to date. The chance of hiring the best overall is also 36.8%. The chance of hiring the last person is also c. This phenomenon is also called $1/e$ law. □

You might please formulate this problem in terms of a sequence of random variables.

(ii). *Some theoretical results about convergence in distribution.*

(a). FATOU LEMMA Suppose $X_n \geq 0$ and $X_n \to X_\infty$ in distribution. Then $E(X_\infty) \leq \liminf_n E(X_n)$.

Proof. Write

$$
E(X_\infty) = \int_0^\infty P(X_\infty \geq t)dt \leq \int_0^\infty \liminf_n P(X_n \geq t)dt = \liminf_n \int_0^\infty P(X_n \geq t)dt \leq \liminf_n E(X_n).
$$

□

The dominated convergence theorem also holds with convergence in distribution, which is left as an exercise.

(b). CONTINUOUS MAPPING THEOREM: $X_n \to X_\infty$ in distribution and $g(\cdot)$ is a continuous function. Then, $g(X_n) \to g(X_\infty)$ in distribution.

Proof. For any bounded continuous function f, $f(g(\cdot))$ is still bounded continuous function. Hence $E(f(g(X_n))) \to E(f(g(X_\infty)))$, proving that $g(X_n) \to g(X_\infty)$ in distribution. □

(c). Tightness and convergent subsequences.

In studying the convergence of a sequence of numbers, it is very useful that boundedness of the sequence, guarantees a convergent subsequence. The same is true for uniformly bounded monotone functions, such as, for example, distribution functions. This is the following Helly's Selection theorem, which is useful in studying weak convergence of distributions.

HELLY'S SELECTION THEOREM. A sequence of cumulative distribution functions F_n always contains a subsequence, say F_{n_k}, that converges to a function, say F_∞, which is nondecreasing and right continuous, at every continuity point of F_∞. If $F_\infty(-\infty) = 0$ and $F_\infty(\infty) = 1$. Then, F_∞ is a distribution function and F_{n_k} converges to F weakly.

Proof Let $t_1, t_2, ...$ be all rational numbers. In the sequence $F_n(t_1), n \geq 1$, there is always a convergent subsequence. Denote one of them as, say $n_k^{(1)}, k = 1, 2, ...$ Among this subsequence there is again a further subsequence, denoted as $n_k^{(2)}, k = 1, 2, ...$, with $n_1^{(2)} > n_1^{(1)}$, such that $F_{n_k^{(2)}}(t_2)$ is convergent. Repeat this process of selection infinitely. Let $n_k = n_1^{(k)}$ be the first element of the k-th sub-sub-sequence. Then, for any fixed m, $\{n_k : k \geq m\}$ is always a subsequence of $\{n_k^{(l)} : k \geq 1\}$ for all $l \leq m$. Hence F_{n_k} is convergent on every rational number. Denote the limit as $F^*(t_l)$ on every rational t_l. Monotonicity of F_{n_k} implies the monotonicity of F^* on rational numbers. Define, for all t, $F_\infty(t) = \inf\{F^*(t_l) : t_l > t,\ t_l \text{ are rational}\}$. Than, F_∞ is right continuous and non-decreasing. The right continuity of F_n ensures that, if s is a continuity point of F_∞, $F_{n_k}(s) \to F_\infty(s)$. □

Not all sequence of distributions F_n would converge weakly to a *distribution function.* The easiest example is $F_n(\{n\}) = F_n(n) - F_n(n-) = 1$, i.e., $P(X_n = n) = 1$. Then, $F_n(t) \to 0$ for all $t \in (-\infty, \infty)$. If F_n all have little probability mass near ∞ or $-\infty$, then the convergence to a function which is not a distribution function can be avoided. A sequence of distribution functions F_n is called *tight* if, for any $\epsilon > 0$, there exists a $M > 0$ such that $\limsup_{n\to\infty}(1 - F_n(M) + F_n(M) < \epsilon$; Or, in other words,

$$\sup_n (1 - F_n(x) + F_n(-x)) \to 0 \qquad \text{as } x \to \infty.$$

PROPOSITION. Every tight sequence of distribution functions contains a a subsequence that weakly converges to a distribution function.

Proof Repeat the proof Helly's Selection Theorem. The tightness ensures the limit is a distribution function. □

(iii). *Characteristic functions.*

Characteristic function is one of the most useful tools in developing theory about convergence in distribution. The technical details of characteristic functions involve some knowledge of complex analysis. We shall view them as only a tool and try not to elaborate the technicalities.

1°. Definition and examples.

For a r.v. X with distribution F, its characteristic function is

$$\psi(t) = E(e^{itX}) = E(cos(tX) + isin(tX)) = \int e^{itx} dF(x), \qquad t \in (-\infty, \infty)$$

where $i = \sqrt{-1}$.

Some basic properties are:

$$\psi(0)=1; \qquad |\psi(\cdot)|\le 1; \qquad \psi(\cdot) \text{ is continuous on } (-\infty,\infty)$$

If ψ is characteristic function of X, then $e^{itb}\psi(at)$ is characteristic function of $aX+b$.

Product of characteristic functions is still a characteristic function. And the characteristic function of $X_1+...+X_n$ is the product of those of $X_1,...,X_n$.

The following table lists some characteristic functions for some commonly used distributions:

Distribution	Density/Probability function	characteristic function (of t)
Degenerate	$P(X=a)=1$	e^{iat}
Binomial $Bin(n,p)$	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\ \ k=0,1,...,n$	$(pe^{it}+1-p)^n$
Poisson $\mathcal{P}(\lambda)$:	$P(X=k)=\lambda^k e^{-\lambda}/k!,\ \ k=0,1,...$	$\exp(\lambda(e^{it}-1))$
Normal $N(\mu,\sigma^2)$:	$f(x)=e^{-(x-\mu)^2/(2\sigma^2)}/\sqrt{2\pi\sigma^2},\ \ x\in(-\infty,\infty)$	$e^{i\mu t-\sigma^2t^2/2}$
Uniform $Unif[0,1]$:	$f(x)=1,\ \ x\in[0,1]$	$(e^{it}-1)/(it)$
Gamma :	$f(x)=\lambda^\alpha x^{\alpha-1}e^{-\lambda x}/\Gamma(\alpha),\ \ x>0$	$(1-it/\lambda)^{-\alpha}$
Cauchy:	$f(x)=1/[\pi(1+x^2)],\ \ x\in(-\infty,\infty)$	$e^{-\lvert t\rvert}$

2°. Levy's inversion formula.

PROPOSITION Suppose X is r.v. with characteristic function $\psi(\cdot)$. Then, for all $a<b$,

$$\lim_{n\to\infty}\frac{1}{2\pi}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\psi(t)dt = P(a<X<b)+\frac{1}{2}(P(X=a)+P(X=b)).$$

Proof. The proof uses Fubini's theorem to interchange the the expectation with the integration and the fact that $\int_0^\infty sin(x)/xdx=\pi/2$. We omit the proof.

The above theorem clearly implies that two different distribution cannot have same characteristic function, as formally presented in the following corollary.

Corollary. There is one-to-one correspondence between distribution functions and characteristic functions.

3°. Levy's continuity theorem.

Theorem 2.1 LEVY'S CONTINUITY THEOREM. *Let F_n, F_∞ be cdf with characteristic function ψ_n,ψ_∞. Then,*

(a). If $F_n\to F_\infty$ weakly, the $\psi_n(t)\to\psi(t)$ for every t.

(b). If $\psi_n(t)\to\psi(t)$ for every t, and $\psi(\cdot)$ is continuous at 0, then $F_n\to F$ weakly, where F is a cdf with characteristic function ψ.

Proof. Part (a) directly follows from the definition of convergence in distribution since e^{itx} is a continuous function of x for every t. Proof of part (b) uses the Levy inversion formula. We omit the details.

REMARK. Levy's continuity theorem enables us to show convergence of distribution through pointwise convergence of characteristic functions. This shall be our approach to establish the central limit theorem.

DIY EXERCISES:

Exercise 2.1. Prove Slutsky's Theorem.

Exercise 2.2. (DOMINATED CONVERGENCE THEOREM) Suppose $X_n\to X_\infty$ in distribution and $|X_n|\le Y$ with $E(Y)<\infty$. Show that $E(X_n)\to E(X_\infty)$.

Exercise 2.3. Suppose X_n is independent of Y_n, and X is independent of Y. Use characteristic functions to show that, if X_n converges to X in distribution and Y_n converges to Y in distribution and , then $X_n + Y_n$ converges in distribution to $X + Y$.

Exercise 2.4. Suppose a r.v. X has characteristic function ψ. Show that

$$P(X = x) = \lim_{C \to \infty} \int_{-C}^{C} e^{itx} \psi(t) dt.$$