4.5

Central Limit Theorem

Let $\{X_n\}$ be a sequence of independent random variables. Let $S_n = \sum_{i=1}^{n} X_i$. In laws of large numbers we considered convergence of $\frac{S_n}{n}$ to $E\left(\frac{S_n}{n}\right)$ which is a constant either in *probability* (in case of WLLN) or *almost surely* (in case of SLLN). Here we consider some different situations, namely, $\frac{S_n}{n} \xrightarrow{d} Z$, where Z is a normal variate. If the sequence $\frac{S_n}{n} \xrightarrow{d} Z$, $\frac{S_n}{n}$ is said to follow the **central limit theorem** (CLT) or **normal convergence**. In this module we consider different Central Limit Theorems.

Definition: A sequence of independent r.vs $\{X_i\}$ with mean $E(X_i) = \mu_i$ and $V(X_i) = {\sigma_i}^2 \ \forall \ i$ is said to follow **Central Limit Theorem** (CLT) under certain conditions, if the random variable $S_n = X_1 + X_2 + \cdots + X_n$ is **asymptotically normal (AN)** with mean μ and variance σ^2 where $\mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} {\sigma_i}^2$.

Notation: $S_n \sim AN(\mu, \sigma^2)$. Read as S_n follows **asymptotically normal** with mean μ and variance σ^2.

Note:

1. S_n is asymptotically normal means S_n *follows normal distribution as* $n \longrightarrow \infty$.
2. If $Z_n = \frac{(S_n - \mu)}{\sigma}$, then Z_n follows asymptotically standard normal with mean 0 and variance 1 and we write $\boldsymbol{Z_n \sim AN(0, 1)}$.

Variations of the CLT

The following are some variations of the CLT which are stated without proof.

Theorem 1 (De Moivre-Laplace CLT) : If $\{X_n\}$ is a sequence of Bernoulli trials with constant probability of success equal to p, then the distribution of the r.v. $S_n = X_1 + \cdots + X_n$ where X_i's are independent, is asymptotically normal ($i.e.$,

S_n is $AN(np, np(1-p))$

Theorem 2 (Lindeberg-Levy CLT) : This CLT theorem is for i.i.d.r.vs.

If $\{X_i\}$ is a sequence of i.i.d.r.vs with mean $E(X_i) = \mu_1$ and variance $V(X_i) = {\sigma_1}^2$ for all i, then the sum $S_n = X_1 + \cdots + X_n$ is asymptotically normal with mean $\mu = n\mu_1$ and variance $\sigma^2 = n{\sigma_1}^2$.

Theorem 3 (Liapounoff's CLT): This CLT theorem is for independent but not identically distributed random variables.

Let $\{X_i\}$ be a sequence of independent random variables with mean $E(X_i) = \mu_i$ and variance $V(X_i) = {\sigma_i}^2 \ \forall \ i$. Let us assume that third absolute moment, say ${\rho_i}^3$ of X_i about its mean exists $i.e.$, ${\rho_i}^3 = E\{|X_i - \mu_i|^3\}$ for $i = 1, 2, \ldots, n$ is finite. Let $\rho^3 = \sum_{i=1}^{n} {\rho_i}^3, \mu = \sum_{i=1}^{n} \mu_i$ and $\sigma^2 = \sum_{i=1}^{n} {\sigma_i}^2$. If $\lim_{n \to \infty} \frac{\rho}{\sigma} = 0$, then the sum $S_n = \sum_{i=1}^{n} X_i$ is $AN(\mu, \sigma^2)$.

Example 1: If $\{X_i\}$ are i.i.d.r.vs with p.m.f $(X_i = \pm 1) = \frac{1}{2}$, find the asymptotic distribution of $S_n = \sum_{i=1}^{n} X_i$.

Solution: Here $E(X_i) = 1.\frac{1}{2} - 1.\frac{1}{2} = 0$ and

$$V(X_i) = E\left({X_i}^2\right) = 1^2.\frac{1}{2} + 1^2.\frac{1}{2} = 1$$

Let $S_n = X_1 + \cdots + X_n$. Then $E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = 0$ and

$V(S_n) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} 1 = n$.

Since mean and variance exist for $\{X_i\}$, by Lindeberg-Levy CLT, $S_n \sim AN(0, n)$ or $\frac{S_n}{\sqrt{n}} \sim AN(0,1)$.

Example 2: If $\{X_i\}$ are i.i.d. with $E(X_i) = 0$, $V(X_i) = \sigma^2$, $0 < \sigma^2 < \infty$ and $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then show that for any $\epsilon > 0$

$$P(\overline{X_n} \geq \epsilon) = \frac{\sigma}{\epsilon\sqrt{n}} \frac{1}{\sqrt{2\pi}} e^{-\frac{n\epsilon^2}{2\sigma^2}} \text{ as } n \longrightarrow \infty.$$

Solution: Let $S_n = X_1 + \cdots + X_n$. Then $E(S_n) = \sum_{i=1}^{n} E(X_i) = 0$ and

$V(S_n) = \sum_{i=1}^{n} V(X_i) = n\sigma^2$. Since $\{X_i\}$ are i.i.d with finite mean and variance, then we have $S_n \sim AN[0, n\sigma^2]$ (by Lindeberg-Levy CLT).

Let $\overline{X_n} = \frac{S_n}{n}$. Then $E(\overline{X_n}) = \frac{1}{n}E(S_n) = \frac{0}{n} = 0$ and

$$V(\overline{X_n}) = \frac{1}{n^2}V(S_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus, $\overline{X_n} \sim AN\left[0, \frac{\sigma^2}{n}\right]$.

We have $P(\overline{X_n} \geq \epsilon) = P\left(\frac{\overline{X_n} - 0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{\epsilon - 0}{\frac{\sigma}{\sqrt{n}}}\right) = P\left(Z \geq \frac{\sqrt{n}\epsilon}{\sigma}\right)$ where $Z \sim N(0,1)$

$$= 1 - P\left(Z \leq \frac{\sqrt{n}\epsilon}{\sigma}\right)$$

$$= 1 - \Phi\left(\frac{\sqrt{n}\,\epsilon}{\sigma}\right), \text{where } \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}t^2} dt$$

$$\Longrightarrow P(\overline{X_n} \geq \epsilon) = 1 - \Phi\left(\frac{\sqrt{n}\,\epsilon}{\sigma}\right) \text{ for } \epsilon > 0 \quad \ldots.(1)$$

But $1 - \Phi(z) = \frac{1}{z}\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$ … (2)

(A result in normal distribution)

From (1) and (2), we have

$$P(\overline{X_n} \geq \epsilon) = 1 - \Phi\left(\frac{\sqrt{n}\epsilon}{\sigma}\right) \Longrightarrow P(\overline{X_n} \geq \epsilon) = \frac{\sigma}{\epsilon\sqrt{n}} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{n\epsilon^2}{2\sigma^2}}$$

Example 3: Examine if CLT holds for the sequence $\{X_k\}$ with p.m.f $P(X_k = \pm 2^k) = 2^{-(2k+1)}$, $(X_k = 0) = 1 - 2^{-2k}$.

Solution: Since it is a non identically distributed sequence of r.vs, for CLT to hold, we have to verify the Liapounoff's condition.

We have $\mu_k = E(X_k) = 2^k \cdot 2^{-(2k+1)} - 2^k \cdot 2^{-(2k+1)} = 0$,

$\sigma_k{}^2 = V(X_k) = E(X_k{}^2) = 2^{2k} \cdot 2^{-(2k+1)} + 2^{2k} \cdot 2^{-(2k+1)} = 1$ and

$\rho_k{}^3 = E\{|X_k - 0|^3\} = E(|X_k|^3) = 2^{3k} \cdot 2^{-(2k+1)} + 2^{3k} \cdot 2^{-(2k+1)}$

$= 2 \cdot 2^{3k} \cdot 2^{-(2k+1)} = 2^k$

Further, we have

$$\mu = \sum_{k=1}^{n} \mu_k = 0\,,$$

$$\sigma^2 = \sum_{k=1}^{n} \sigma_k{}^2 = \sum_{i=1}^{n} 1 = n\,,$$

$$\rho^3 = \sum_{k=1}^{n} \rho_k{}^3 = \sum_{k=1}^{n} 2^k = 2 + 2^2 + \cdots + 2^n = 2(2^n - 1) \text{ and}$$

$$\frac{\rho^3}{(\sigma^2)^{\frac{3}{2}}} = \frac{2(2^n-1)}{n^{\frac{3}{2}}}$$

Thus, $\lim_{n \to \infty} \frac{\rho^3}{(\sigma^2)^{\frac{3}{2}}} = \lim_{n \to \infty} \frac{2(2^n-1)}{n^{\frac{3}{2}}} = \infty$. Thus, the Liapounoff's condition is not satisfied and hence we cannot say that CLT holds for $\{X_k\}$.

Example 4: Examine if CLT holds for the sequence $\{X_k\}$ with p.m.f

$$\boldsymbol{P(X_k = \pm k^{\alpha}) = \frac{1}{2}.k^{-2\alpha}, P(X_k = 0) = 1 - k^{1-2\alpha}, \alpha < \frac{1}{2}.}$$

Solution: Since it is a non identically distributed of r.vs, for CLT to hold, we have to verify the Liapounov's condition.

We have $\mu_K = E(X_k) = k^{\alpha}.\frac{1}{2}.k^{-2\alpha} - k^{\alpha}.\frac{1}{2}.k^{-2\alpha} = 0$,

$\sigma_k^2 = V(X_k) = E(X_k^2) = k^{2\alpha}.\frac{1}{2}.k^{-2\alpha} + k^{2\alpha}.\frac{1}{2}.k^{-2\alpha} = \frac{1}{2} + \frac{1}{2} = 1$ and

$$\rho_k^3 = E\{|X_k - 0|^3\} = E\{|X_k|^3\} = k^{3\alpha}.\frac{1}{2}.k^{-2\alpha} + k^{3\alpha}.\frac{1}{2}k^{-2\alpha}$$

$$= \frac{1}{2}.k^{\alpha} + \frac{1}{2}k^{\alpha} = k^{\alpha}$$

Further, we have

$$\mu = \sum_{k=1}^{n} \mu_k = \sum_{k=1}^{n} 0 = 0 \ ,$$

$$\sigma^2 = \sum_{k=1}^{n} \sigma_k^2 = \sum_{k=1}^{n} 1 = n \quad \text{and} \quad \rho^3 = \sum_{k=1}^{n} \rho_k^3 = \sum_{k=1}^{n} k^{\alpha} = 1^{\alpha} + 2^{\alpha} + \ldots + n^{\alpha}$$

Note that $\rho^3 \leq n.n^{\alpha} = n^{\alpha+1}$

Now $\lim_{n\to\infty} \frac{\rho^3}{(\sigma^2)^{3/2}} \leq \lim_{n\to\infty} \frac{n^{\alpha+1}}{n^{3/2}} = \lim_{n\to\infty} n^{\alpha-\frac{1}{2}} = 0$, if $\alpha < \frac{1}{2}$

Thus $\lim_{n\to\infty} \frac{\rho^3}{(\sigma^2)^{3/2}} = 0$ if $\alpha < \frac{1}{2}$

Therefore, CLT holds for the sequence $\{X_k\}$.

Applications of central Limit Theorem:

In case of Bernoulli, Binomial and Poisson distributions, evaluation of probabilities using p.m.f. are tedious. Using normal approximation for large samples to these distributions, the probabilities can be easily evaluated.

(a) Let $\{X_n\}$ be a sequence of i.i.d Bernoulli variate $i.e., B(1,p)$.
Let $S_n = X_1 + \cdots + X_n$

Then $S_n \sim B(n,p)$, where $E(S_n) = np$ and $V(S_n) = np(1-p) = npq$

By Lindeberg Levy CLT for large n, $S_n \sim AN\left(E(S_n), V(S_n)\right)$

$$\Longrightarrow S_n \sim AN\left(np, np(1-p)\right) \qquad \ldots (1)$$

Let $Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}$

Then from (1), $Z_n \xrightarrow{d} Z$ where Z is $N(0,1)$

Thus, $\lim_{n\to\infty} P\left(a \le \frac{S_n - np}{\sqrt{npq}} \le b\right) = P(a \le Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}z^2} dz$

and the RHS can be evaluated using standard normal tables for given real numbers a and b.

(b) Let $\{X_n\}$ be a sequence of i.i.d Binomial variates . $e, B(r,p)$.
Let $S_n = X_1 + \cdots + X_n$

Then $E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} rp = nrp$

$(\because E(X_i) = rp \ \forall\ i)$

and $V(S_n) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i)$ $\qquad (\because \ X_i s \ are \ independent)$

$$= \sum_{i=1}^{n} rp(1-p) = nrp(1-p)$$

$$\left(\because \quad V(X_i) = rp(1-p)\right)$$

Thus $E(S_n) = nrp$ and $V(S_n) = nrp(1-p)$

By Lindberg – Levy CLT, for large n, we have

$$S_n \sim AN\left(nrp,\ nrp(1-p)\right)$$

Let $Z_n = \frac{S_n - nrp}{\sqrt{nrp(1-p)}}$

Then $Z_n \xrightarrow{d} Z$ where $Z \sim N(0,1)$

Thus, $\lim_{n\to\infty} P\left(a \le \frac{S_n - nrp}{\sqrt{nrp(1-p)}} \le b\right) = P(a \le Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}z^2} dz$

and the RHS can be evaluated using standard normal tables for given real numbers a and b.

(c) Let $\{X_n\}$ be a sequence of i.i.d Poisson variates . $e.$, $P(\lambda)$. Let $S_n = \sum_{i=1}^{n} X_i$

Here $E(X_i) = V(X_i) = \lambda \ \forall\ i$. Then $E\left(S_n\right) = \sum_{i=1}^{n} E\left(X_i\right) = n\lambda$

and $V\left(S_n\right) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V\left(X_i\right) = n\lambda$ $\qquad (\because\ X_i\ s\ are\ independent)$

Thus, by Lindeberg – Levy CLT, for large n, $S_n \sim AN(n\lambda, n\lambda)$

Let $Z_n = \frac{S_n - n\lambda}{\sqrt{n\lambda}}$, Then $Z_n \xrightarrow{d} Z$, where $Z \sim N(0,1)$

The probabilities $\lim_{n\to\infty} P\left(a \le \frac{S_n - n\lambda}{\sqrt{n\lambda}} \le b\right) = P(a \le Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}z^2} dz$

and the RHS can be evaluated using standard normal for given real numbers a and b.

Example 5: A sample of 100 items is taken at random from a batch known to contain 40% defectives. What is the probability that the sample contains

(i) at least 44 defetives,

(ii) exactly 44 defectives?

Solution:

Let $X_i = \begin{cases} 1 \quad , \quad if\ the\ i^{th}\ item\ is\ defective \\ 0 \ , \quad if\ the\ i^{th}\ item\ is\ nondefective \end{cases}$, for $i = 1, 2, \ldots$

It is given that $P(defective) = P(X_i = 1) = 40\% = 0.4$

Then X_i follows Bernoulli distribution $i.e.$, $B(1, 0.4)$

Let $S_n = X_1 + \cdots + X_n$. Then $S_n \sim B(n, p)$

Since $n = 100$ and $p = 0.4$, $S_n \sim B(100, 0.4)$

Since n is large, computation of probabilities using binomial formula is difficult. Hence, by CLT, we use normal approximation to compute the probabilities of S_n instead of binomial distribution.

Here $E(S_n) = np = 100.(0.4) = 40$ and

$$V(S_n) = np(1-p) = 100 \times 0.4 \times 0.6 = 24$$

Let $Z_n = \frac{S_n - E(S_n)}{\sqrt{V(S_n)}} = \frac{S_n - 40}{\sqrt{24}} = \frac{S_n - 40}{4.9}$

Then by Lindeberg Levy CLT, we have $Z_n \xrightarrow{d} Z$, where Z is $N(0, 1)$.

(i) It should be noted that the continuous normal distribution is approximating the discrete binomial distribution so that the continuity correction has to be taken into account in determining the various probabilities. So finding the probability of at least 44 defectives in a sample of 100 items requires finding the area under the normal curve from $\mathbf{43.5}$ **to** $\mathbf{100.5}$

Therefore, the probability of at least 44 defectives is given by

$$P(43.5 < S_n < 100.5) = P\left(\frac{43.5-40}{4.9} < Z < \frac{100.5-40}{4.9}\right)$$

$$= P(0.7143 < Z < 12.347)$$

$$= P(0 < Z < 12.347) - P(0 < Z < 0.7143)$$

$$= 0.5 - 0.2624$$

(See the standard normal distribution table)

$$= 0.2376$$

(ii) The probability of exactly 44 defectives is

$$P(S_n = 44) = P(43.5 < S_n < 44.5)$$

$$= P\left(\frac{43.5-40}{4.9} < Z < \frac{44.5-40}{4.9}\right)$$

$$= P(0.7143 < Z < 0.9184)$$

$$= P(0 < Z < 0.9184) - P(0 < Z < 0.7143)$$

$$= 0.3208 - 0.2624$$ (See table)

$$= 0.0584$$

Note: Using the binomial distribution, $P(S_n \geq 44) = \sum_{k=44}^{100} \binom{100}{k} (0.4)^k (0.6)^{100-k}$

and $P(S_n = 44) = \binom{100}{44} (0.4)^{44}(0.6)^{56} = 0.0576$ (Using **Binomial tables**).

As can be seen by comparing the answers, both sets of answers are remarkably close.

Example 6: Let $X_1, X_2, \ldots$ be i.i.d. Poisson variables with parameter λ. Use CLT to estimate $P(120 \leq S_n \leq 160)$,where $S_n = X_1 + \cdots + X_n$, $\lambda = 2$ and $n = 75$

Solution: Since $X_i\ s$ are i.i.d $P(\lambda), E(X_i) = \lambda = V(X_i)$ for $i = 1, 2, \ldots n$

$$\therefore\ E(S_n) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \lambda = n\lambda \quad \text{and}$$

$$V(S_n) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = n\lambda$$

Hence by Lindberg – Levy CLT, for large n, we have

$S_n \sim AN(n\lambda,\ n\lambda) = AN(150, 150)$. After applying the continuity correction, the required probability is

$p = P(119.5 \leq S_n \leq 160.5) = P\left(\frac{119.5-150}{\sqrt{150}} \leq Z \leq \frac{160.5-150}{\sqrt{150}}\right)$,
where $Z \sim N(0, 1)$

$$= P(-2.45 \leq Z \leq 0.82)$$

$$= P(-2.45 \leq Z \leq 0) + P(0 \leq Z \leq 0.82)$$

$$= P(0 \leq Z \leq 2.45) + P(0 \leq Z \leq 0.82)$$

$$= 0.4929 + 0.2938 \quad \text{(From standard normal table)}$$

$$= 0.7868$$