Recap: Modes of convergence

- $X_n \xrightarrow{d} X$ iff
$$F_n(x) \to F(x), \ \ \forall x \ \ \text{where } F \text{ is continuous}$$
- $X_n \xrightarrow{P} X$ iff
$$\lim_{n\to\infty} P\left[|X_n - X| > \epsilon\right] = 0, \ \forall \epsilon > 0$$
- $X_n \xrightarrow{r} X$ iff
$$E\left[|X_n - X|^r\right] \to 0 \qquad \text{as } \ n \to \infty$$
- $X_n \xrightarrow{a.s} X$ iff
$$P[X_n \to X] = 1 \ \ \text{or} \ \ P[\limsup |X_n - X| > \epsilon] = 0$$

Recap

- We have the following relations among different modes of convergence

$$X_n \xrightarrow{r} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

$$X_n \xrightarrow{a.s.} X \;\Rightarrow\; X_n \xrightarrow{P} X \;\Rightarrow\; X_n \xrightarrow{d} X$$

- All the implications are one-way and we have seen counter examples
- In general, almost sure convergence does not imply convergence in r^{th} mean and vice versa

Recap

- Given X_i are iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Weak law of large numbers: $\frac{S_n}{n} \xrightarrow{P} \mu$
- strong law of large numbers: $\frac{S_n}{n} \xrightarrow{a.s.} \mu$
- Central Limit Theorem: $\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$

Recap

- Take X_i iid, $EX_i = 0$, $\text{Var}(X_i) = 1$, $n = 1, 2, \cdots$
- $S_n = \sum_{i=1}^{n} X_i$
- Strong law of large numbers implies

$$\frac{S_n}{n} \stackrel{a.s.}{\rightarrow} 0$$

- Central Limit Theorem implies

$$\frac{S_n}{\sqrt{n}} \stackrel{a.s.}{\rightarrow} \mathcal{N}(0, 1)$$

Recap: Characteristic Function

- Given rv X, its characteristic function, ϕ_X, is defined by

$$\phi_X(u) = E\left[e^{iuX}\right] = \int e^{iux}\, dF_X(x) \quad (i = \sqrt{-1})$$

- Since $|e^{iux}| \leq 1$, ϕ_X exists for all random variables
  - ϕ is continuous; $|\phi(u)| \leq \phi(0) = 1$; $\phi(-u) = \phi^*(u)$
  - If $Y = aX + b$, $\phi_Y(u) = e^{iub}\phi_X(ua)$
  - If $E|X|^r < \infty$, ϕ would be differentiable r times and

$$\phi^{(r)}(u) = E[(iX)^r e^{iuX}]$$

Recap

- Let $\mu_r = E[X^r]$ and let $\nu_r = E[|X|^r]$
- If ν_r is finite, then

$$\phi_X(u) = \sum_{s=0}^{r-1} \mu_s \, \frac{(iu)^s}{s!} + \rho(u) \; \mu_r \, \frac{(iu)^r}{r!}$$

  where $|\rho(u)| \leq 1$ and $\rho(u) \to 1$ as $u \to 0$
- If all moments exist, then

$$\phi_X(u) = \sum_{s=0}^{\infty} \mu_s \, \frac{(iu)^s}{s!}$$

Recap

- We denote by ϕ_F characteristic function of df F
- Let F_n be a sequence of distribution functions
- **Continuity theorem**
  - If $F_n \to F$ then $\phi_{F_n} \to \phi_F$
  - If $\phi_{F_n} \to \psi$ and ψ is continuous at zero, then ψ would be characteristic function of some df, say, F, and $F_n \to F$

- Given X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^n X_i$
- Let $\tilde{S}_n = \frac{S_n - ES_n}{\sqrt{\text{var}(S_n)}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$
- **(Lindberg-Levy) Central Limit Theorem**

$$\lim_{n\to\infty} P\left[\tilde{S}_n \le x\right] = \lim_{n\to\infty} P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \ \ \forall x$$

Proof:

- Without loss of generality let us assume $\mu = 0$.
- We use characteristic function of $\tilde{S}_n$ for the proof.
- Let ϕ be the characteristic function of X_i. Then

$$\phi_{S_n}(t) = (\phi(t))^n \quad \text{and} \quad \phi_{\tilde{S}_n}(t) = \left(\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n$$

- Recall that we can expand ϕ in a Taylor series

$$\phi(u) = \sum_{s=0}^{r-1} \mu_s \, \frac{(iu)^s}{s!} + \rho(u) \, \mu_r \, \frac{(iu)^r}{r!}, \quad \rho(u) \to 1, \text{ as } u \to 0$$

- Here we assume: $EX_i = 0$ and $EX_i^2 = \sigma^2$

$$\phi(t) = 1 \, + \, 0 \, - \, \frac{1}{2} \, \rho(t) \, \sigma^2 \, t^2$$

$$\begin{aligned}
\phi\left(\frac{t}{\sigma\sqrt{n}}\right) &= 1 - \frac{1}{2} \, \rho\left(\frac{t}{\sigma\sqrt{n}}\right) \, \sigma^2 \, \frac{t^2}{\sigma^2 n} \\
&= 1 - \frac{1}{2} \, \frac{t^2}{n} \, + \, \frac{1}{2} \, \frac{t^2}{n} \, \left(1 - \rho\left(\frac{t}{\sigma\sqrt{n}}\right)\right) \\
&= 1 - \frac{1}{2} \, \frac{t^2}{n} \, + \, o\left(\frac{1}{n}\right)
\end{aligned}$$

- Hence we get

$$\begin{aligned}\lim_{n\to\infty} \phi_{\tilde{S}_n}(t) &= \lim_{n\to\infty} \left(\phi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n \\ &= \lim_{n\to\infty} \left(1 - \frac{1}{2}\,\frac{t^2}{n} \;+\; o\left(\frac{1}{n}\right)\right)^n \\ &= e^{-\frac{t^2}{2}}\end{aligned}$$

  which is the characteristic function of standard normal
- By Continuity theorem, distribution function of $\tilde{S}_n$ converges to that of standard Normal rv

$$\lim_{n\to\infty} P\left[\tilde{S}_n \leq x\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt, \;\; \forall x$$

- What CLT says is that sums of iid random variables, when appropriately normalized, would always approach the Gaussian distribution.
- It allows one to approximate distribution of sums of independent rv's
- Let X_i be iid and $S_n = \sum_{i=1}^{n} X_i$

$$P[S_n \leq x] = P\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{x - n\mu}{\sigma\sqrt{n}}\right] \approx \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

- Thus, S_n is well approximated by a normal rv with mean $n\mu$ and variance $n\sigma^2$, if n is large

Example

- Twenty numbers are rounded off to the nearest integer and added. What is the probability that the sum obtained differs from true sum by more than 3.
- A reasonable assumption is round-off errors are independent and uniform over $[-0.5,\ 0.5]$
- Take $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid.
- Then Z represents the error in the sum.

- $Z = \sum_{i=1}^{20} X_i$, $X_i \sim U[-0.5,\ 0.5]$, X_i iid
- $EX_i = 0$ and $\text{Var}(X_i) = \frac{1}{12}$.
- Hence, $EZ = 0$ and $\text{Var}(Z) = \frac{20}{12} = \frac{5}{3}$

$$
\begin{aligned}
P[|Z| \leq 3] &= P[-3 \leq Z \leq 3] \\
&= P\left[\frac{-3}{\sqrt{\frac{5}{3}}} \leq \frac{Z - EZ}{\sqrt{\text{Var}(Z)}} \leq \frac{3}{\sqrt{\frac{5}{3}}}\right] \\
&\approx \Phi\left(\frac{3}{\sqrt{\frac{5}{3}}}\right) - \Phi\left(\frac{-3}{\sqrt{\frac{5}{3}}}\right) \\
&\approx \Phi(2.3) - \Phi(-2.3) \\
&= 0.9893 - 0.0107 \approx 0.98
\end{aligned}
$$

- Hence probability that the sum differs from true sum by more than 3 is 0.02

- We can approximate binomial rv with Gaussian for large n
- Binomial random variable with parameters n, p is a sum of n independent Bernoulli variables:
  $S_n = \sum_{i=1}^{n} X_i; \;\; X_i \in \{0, \; 1\}$, $P[X_i = 1] = p$, X_i ind
- Hence we can approximate distribution of S_n by

$$
\begin{aligned}
P[S_n \leq x] &= P\left[\frac{S_n - np}{\sqrt{np(1-p)}} \leq \frac{x - np}{\sqrt{np(1-p)}}\right] \\
&\approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)
\end{aligned}
$$

- For large n, binomial rv is like a Gaussian rv with mean np and variance $np(1-p)$
- The approximation is quite good in practice

- S_n be binomial with parameters n, p

$$P[S_n \leq x] \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

- For example, with $p = 0.95$

$$P[S_{110} \leq 100] \approx \Phi\left(\frac{100 - 110 * 0.95}{\sqrt{110 * 0.05 * 0.95}}\right) \approx \Phi(-1.97) = 0.025$$

- Since S_n is integer-valued, the LHS above is same for all x between two consecutive integers; but RHS changes
- To get a good approximation, to calculate $P[S_n \leq m]$ one uses $P[S_n \leq m + 0.5]$ in the above approximation formula

- CLT allows one to get rate of convergence of law of large numbers
- Let X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^{n} X_i$
- By Law of large numbers, $\frac{S_n}{n} \to \mu$.
- Now, by CLT

$$\begin{aligned} P\left[\left|\frac{S_n}{n} - \mu\right| > \epsilon\right] &= P\left[|S_n - n\mu| > n\epsilon\right] \\ &= P\left[\left|\frac{S_n - n\mu}{\sigma\sqrt{n}}\right| > \frac{n\epsilon}{\sigma\sqrt{n}}\right] \\ &\approx 1 - \left(\Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right) - \Phi\left(-\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right) \\ &= 2\left(1 - \Phi\left(\frac{n\epsilon}{\sigma\sqrt{n}}\right)\right) \end{aligned}$$

(because $\Phi(-x) = (1 - \Phi(x))$)

Example: Opinion Polls

- let p denote the fraction of population that prefers product A to product B
- We want to estimate p
- We conduct a sample survey by asking n people
- We want to make a statement such as
$$p = 0.34 \pm 0.07 \textit{ with a confidence of } 95\%$$
- Here, the 0.34 would be the sample mean. The other two numbers can be fixed using CLT

- $X_i \in \{0,\ 1\}$ iid, $EX_i = p$, $S_n = \sum_{i=1}^n X_i$
- Now, by CLT, we have

$$
\begin{aligned}
P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] &= P\left[|S_n - np| > n\epsilon\right] \\
&= 2\left(1 - \Phi\left(\frac{n\epsilon}{\sqrt{np(1-p)}}\right)\right)
\end{aligned}
$$

- Suppose we want to satisfy

$$
P\left[\left|\frac{S_n}{n} - p\right| > \epsilon\right] = \delta
$$

- We can calculate any one of ϵ, δ or n given the other two using the earlier equation.
- But we need value of p for it!

- Fortunately, $\sqrt{p(1-p)}$ does not change too much with p
- It attains its maximum value of 0.5 at $p = 0.5$
- It is 0.458 at $p = 0.3$ and is 0.4 at $p = 0.2$
- One normally fixes this variance as 0.5 or 0.45 to calculate the sample size, n.
- There are other ways of handling it

- We have

$$P\left[\left|\frac{S_n}{n}-p\right|>\epsilon\right]=2\left(1-\Phi\left(\frac{\epsilon\sqrt{n}}{\sqrt{p(1-p)}}\right)\right)$$

- Suppose $n=900$ and $\epsilon=0.025$.
  Let us approximate $\sqrt{p(1-p)}=0.45$. Then

$$2\left(1-\Phi\left(\frac{0.025*30}{0.45}\right)\right)=2(1-\Phi(1.66))\approx 0.1$$

- If we took $\sqrt{p(1-p)}=0.5$ we get the value as 0.14
- If we use Chebyshev inequality with variance as 0.5 we get the bound as 0.8
- If we change ϵ to 0.05, then at variance equal to 0.5 the probability becomes about 0.02 while the Chebyshev bound would be about 0.2

Confidence intervals

- Let X_i iid, $EX_i = \mu$, $\text{Var}(X_i) = \sigma^2$, $S_n = \sum_{i=1}^{n} X_i$.
- Using CLT, we get

$$P\left[\left|\frac{S_n}{n} - \mu\right| > c\right] = 2\left(1 - \Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- If the RHS above is δ, then we can say that $\frac{S_n}{n} \in [\mu - c,\ \mu + c]$ with probability $(1 - \delta)$
- This interval is called the $100(1-\delta)\%$ confidence interval.

$$P\left[\left|\frac{S_n}{n}-\mu\right| > c\right] = 2\left(1-\Phi\left(\frac{c\sqrt{n}}{\sigma}\right)\right)$$

- Suppose $c = \frac{1.96\sigma}{\sqrt{n}}$
- Then

$$P\left[\left|\frac{S_n}{n}-\mu\right| > \frac{1.96\sigma}{\sqrt{n}}\right] = 2\left(1-\Phi\left(1.96\right)\right) = 0.05$$

- Denoting $\bar{X} = \frac{S_n}{n}$, the 95% confidence interval is $\left[\bar{X} - \frac{1.96\sigma}{\sqrt{n}},\ \bar{X} + \frac{1.96\sigma}{\sqrt{n}}\right]$
- One generally uses an estimate for σ obtained from X_i
- In analyzing any experimental data the confidence intervals or the variance term is important

central limit theorem

- CLT essentially states that sum of many independent random variables behaves like a Gaussian random variable
- It is very useful in many statistics applications.
- We stated CLT for iid random variables.
- While independence is important, all rv need not have the same distribution.
- Essentially, the variances should not die out.

- We have been considering sequences: $X_n, \ n = 1, 2, \cdots$
- We have so far considered only the asymptotic properties or limits of such sequences.
- Any such sequence is an example of what is called a random process or stochastic process
- Given n rv, they are completely characterized by their joint distribution.
- How doe we specify or characterize an infinite collection of random variables?
- We need the joint distribution of every finite subcollection of them.

Markov Chains

- Let $X_n,\ n = 0, 1, \cdots$ be a sequence of discrete random variables taking values in S.
  Note that S would be countable
- We say it is a Markov chain if

$$P[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1} \cdots X_0 = x_0] = P[X_{n+1} = x_{n+1} | X_n = x_n], \forall x$$

- We can write it as

$$P[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} \cdots X_0] = P[X_{n+1} = x_{n+1} | X_n = x_n], \forall x_i$$

- Conditioned on X_n, X_{n+1} is independent of $X_{n-1}, X_{n-2}, \cdots$
- We think of X_n as state at n
- For a Markov chain, given the current state, the future evolution is independent of the history of how you reached the current state

Example

- Let X_i be iid discrete rv taking integer values.
- Let $Y_0 = 0$ and $Y_n = \sum_{i=1}^{n} X_i$
- $Y_n, \ n = 0, 1, \cdots$ is a Markov chain with state space as integers
- Note that $Y_{n+1} = Y_n + X_{n+1}$ and X_{n+1} is independent of $Y_0, \cdots, Y_n$.

$$P[Y_{n+1} = y | Y_n = x, Y_{n-1}, \cdots] = P[X_{n+1} = y - x]$$

- Thus, Y_{n+1} is conditionally independent of $Y_{n-1}, \cdots$ conditioned on Y_n

- In this example, we can think of X_n as the number of people or things arriving at a facility in the n^{th} time interval.
- Then Y_n would be total arrivals till end of n^{th} time interval.
- Number of packets coming into a network switch, number people joining the queue in a bank, number of infections till date are all Markov chains.
- This is a useful model for many dynamic systems or processes

- The Markov property is: given current state, the future evolution is independent of the history of how we came to current state.
- It essentially means the current state contains all needed information about history
- We are considering the case where states as well as time are discrete.
- It can be more general and we discuss some of them

Transition Probabilities

- Let $\{X_n,\ n = 0, 1, \cdots\}$ be a Markov Chain with (countable) state space S

$$Pr[X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} \cdots X_0] = Pr[X_{n+1} = x_{n+1} | X_n = x_n], \forall x$$

  (Notice change of notation)
- Define function $P : S \times S \to [0,\ 1]$ by

$$P(x, y) = Pr[X_{n+1} = y | X_n = x]$$

- P is called the state transition probability function. It satisfies
  - $P(x, y) \geq 0,\ \forall x, y \in S$
  - $\sum_{y \in S} P(x, y) = 1, \forall x \in S$
- If S is finite then P can be represented as a matrix

- The state transition probability function is given by

$$P(x, y) = Pr[X_{n+1} = y | X_n = x]$$

- In general, this can depend on n though our notation does not show it
- If the value of that probability does not depend on n then the chain is called homogeneous
- For a homogeneous chain we have

$$Pr[X_{n+1} = y | X_n = x] = Pr[X_1 = y | X_0 = x], \ \forall n$$

- In this course we will consider only homogeneous chains

Initial State Probabilities

- Let $\{X_n\}$ be a Markov Chain with state space S
- Define function $\pi_0 : S \to [0,\ 1]$ by

$$\pi_0(x) = Pr[X_0 = x]$$

- It is the pmf of the rv X_0
- Hence it satisfies
  - $\pi_0(x) \geq 0,\ \forall x \in S$
  - $\sum_{x \in S} \pi_0(x) = 1$
- From now on, without loss of generality, we take $S = \{0, 1, \cdots\}$

- Let X_n be a (homogeneous) Markov chain
- Then we have

$$\begin{aligned} Pr[X_0 = x_0, X_1 = x_1] &= Pr[X_1 = x_1 | X_0 = x_0] \; Pr[X_0 = x_0], \forall x_0, x_1 \\ &= P(x_0, x_1)\pi_0(x_0) = \pi_0(x_0)P(x_0, x_1) \end{aligned}$$

- Now we can extend this as

$$\begin{aligned} Pr[X_0 = x_0, X_1 = x_1, X_2 = x_2] &= Pr[X_2 = x_2 | X_1 = x_1, X_0 = x_0] \cdot \\ &\qquad Pr[X_0 = x_0, X_1 = x_1] \\ &= Pr[X_2 = x_2 | X_1 = x_1] \cdot \\ &\qquad Pr[X_0 = x_0, X_1 = x_1] \\ &= P(x_1, x_2) \; P(x_0, x_1)\pi_0(x_0) \\ &= \pi_0(x_0)P(x_0, x_1)P(x_1, x_2) \end{aligned}$$

- This calculation is easily generalized to any number of time steps

$$
\begin{aligned}
Pr[X_0 = x_0, \cdots X_n = x_n] &= Pr[X_n = x_n | X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \cdot \\
&\qquad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= Pr[X_n = x_n | X_{n-1} = x_{n-1}] \cdot \\
&\qquad Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1}, \cdots X_0 = x_0] \\
&= P(x_{n-1}, x_n) Pr[X_{n-1} = x_{n-1} | X_{n-2} = x_{n-2}] \cdot \\
&\qquad Pr[X_{n-2} = x_{n-2}, \cdots X_0 = x_0] \\
&\vdots \\
&= \pi_0(x_0) P(x_0, x_1) \cdots P(x_{n-1}, x_n)
\end{aligned}
$$

- We showed

$$Pr[X_0 = x_0, \cdots X_n = x_n] = \pi_0(x_0)P(x_0, x_1) \cdots P(x_{n-1}, x_n)$$

- This shows that the transition probabilities, P, and initial state probabilities, π_0, completely specify the chain.
- They give us the joint distribution of any finite subcollection of the rv's
- Suppose you want joint distribution of $X_{i_1}, \cdots X_{i_k}$
- Let $m = \max(i_1, \cdots, i_k)$
- We know how to get joint distribution of $X_0, \cdots, X_m$.
- The joint distribution of $X_{i_1}, \cdots X_{i_k}$ is now calculated as a marginal distribution from the joint distribution of $X_0, \cdots, X_m$