Central Limit Theorem (CLT)

Previously, from the weak LLN applied to

$$\frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i ,$$

iid $E[X_i] = \mu$
$var(X_i) = \sigma^2$

the pdf of $\frac{S_n}{n}$ becomes more concentrated around μ, that is,

$$P\left[\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right] \to 0 \quad \text{as } n \to +\infty.$$

Note: This convergence is called convergence in probability

$$\frac{S_n}{n} \xrightarrow{P} \mu \quad \text{as } n \to +\infty.$$

What is the form of the limiting distribution?

Focus: limiting Gaussian distributions.
Motivation: The condition under which the limiting distribution is Gaussian are weak requiring primarily that each individual R.V. contributes only negligibly to the overall sum of a large number of RVs.

ex. 1: generation of a current by propagating electrons
By the CLT, $i(t) \sim \mathcal{N}(m, \sigma^2)$

ex. 2: radar signal returned from a rain cloud is modeled as Gaussian distributed.

Consider a set of iid RVs $X_1, \ldots, X_n$, each with the mean $E[X_i] = \mu$ and variance $E[X_i] = \sigma^2$.

From the last lecture:

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

has the mean $E[S_n] = \mu$ and variance

$var(S_n) = \sigma^2 \cdot n$

CLT states that if S_n is properly normalized, then its cdf converges to the cdf of $N(0,1)$.

<u>Theorem</u> (CLT): Let S_n be a sum of iid RVs with the mean μ and variance σ^2. Define RV

$$S_n^* = \frac{S_n - E[S_n]}{\sigma_{S_n}} .$$

Then the cdf of S_n^* converges to:

$$F_{S_n^*}(\alpha) \xrightarrow{n\to+\infty} \underbrace{\int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt}_{\text{Gaussian cdf at } \alpha.} \qquad (*)$$

<u>Comment</u>: $S_1, S_2, \ldots, S_n, \ldots$ is a sequence of RVs. Therefore, $S_1^*, S_2^*, \ldots, S_n^*, \ldots$ is also a sequence of RVs. CLT states the result on convergence of the cdf S_n^* to another cdf. In probability theory this convergence is known as <u>convergence in distribution</u>.

Let $Z \sim N(0,1)$ with $F_Z(\alpha) = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$, then $S_n^* \xrightarrow{d} Z$ as $n \to +\infty$. This is described by equation (1).

"in distribution"

To prove the CLT, we need another result known as continuity theorem.

<u>Theorem</u> (Continuity)

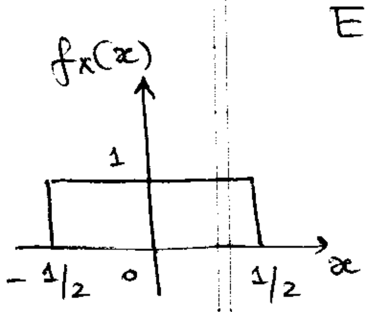
$$X_n \xrightarrow{d} X \text{ as } n \to +\infty \quad \text{iff} \quad M_{X_n}(j\nu) \to M_X(j\nu) \text{ as } n \to +\infty .$$

We will omit the proof, but illustrate it using the following 2 examples.

<u>ex</u>. 1 (uniform components)

Let $X_1, \ldots\ldots, X_n$ be unif. $([-\frac{1}{2}, \frac{1}{2}])$

$E[X_i] = 0$, $\quad var = \dfrac{1}{12}$



Form $S_n = \sum_{i=1}^{n} X_i$ and then

$$S_n^* = \frac{\sum_{i=1}^{n} X_i - E[S_n]}{\sigma_{S_n}}$$

standard deviation of S_n

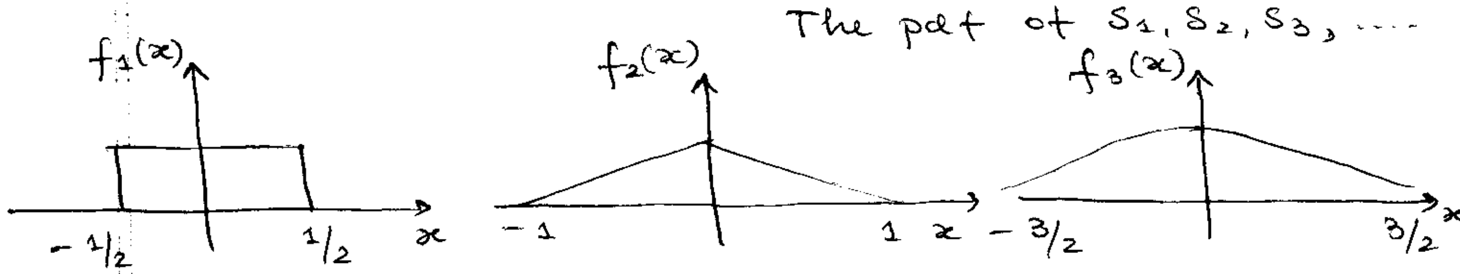
Note that in the Sequence $S_1^*, S_2^*, \ldots\ldots, S_n^*, \ldots\ldots$ each RV has equal mean and variance:

$E[S_n^*] = 0$ and $var(S_n^*) = 1$.

Thus, $S_n^* = \sqrt{\frac{12}{n}} \underbrace{\sum_{i=1}^{n} X_i}$

consider the sum

The pdf of $S_1, S_2, S_3, \ldots$



The pdf is "spreading" + "becoming smoother."

The characteristic function of S_n^*:

$$M_{S_n^*}(jv) \triangleq E\left[e^{jv\sqrt{\frac{12}{n}}\sum_{i=1}^{n} X_i} \right]$$

$$\overset{iid}{=} \Big\{ \underbrace{E\left[e^{jv\sqrt{12/n}\cdot X_1} \right]}_{M_X\left(jv\sqrt{\frac{12}{n}}\right)} \Big\}^n = M_X^n\left(jv\sqrt{\tfrac{12}{n}}\right)$$

$$= \underbrace{\left(\frac{\sin\sqrt{\frac{3}{n}}\,v}{\sqrt{\frac{3}{n}}\,v} \right)}_{\text{sinc}(\cdot)}^{n}$$

$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + o(x^5)$

$$\overset{\text{Taylor series expansion}}{\cong} \left(\frac{\sqrt{\frac{3}{v}}\,v - \left(\sqrt{\frac{3}{n}}\,v\right)^3 \frac{1}{3!} + \cdots\cdots}{\sqrt{\frac{3}{n}}\,v} \right)^n = \left(1 - \frac{1}{6}\cdot\frac{3}{n}v^2 + \cdots\right)^n$$

Since $\lim_{n\to+\infty}\left(1 - \frac{x}{n}\right)^n = e^{-x}$,

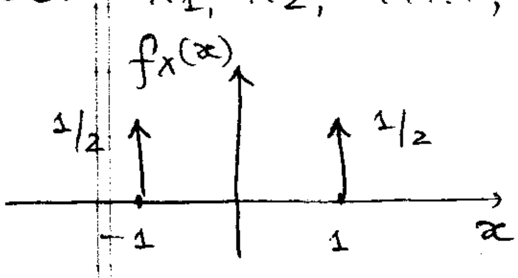
$$\lim_{n\to+\infty}\left(1 - \frac{v^2}{2n}\right)^n = e^{-\frac{v^2}{2}}$$

Therefore, $M_{S_n^*}(jv) \xrightarrow{n\to+\infty} e^{-v^2/2}$ and by continuity property

$$S_n^* = \sqrt{\frac{12}{n}}\sum_{i=1}^{n} X_i \xrightarrow[\text{as } n\to+\infty]{d} \text{Gaussian R.V. } N(0,1)$$

<u>ex. 2</u> (binary valued)

Let $X_1, X_2, \ldots\ldots, X_n, \ldots\ldots$ be iid binary $\{-1, +1\}$



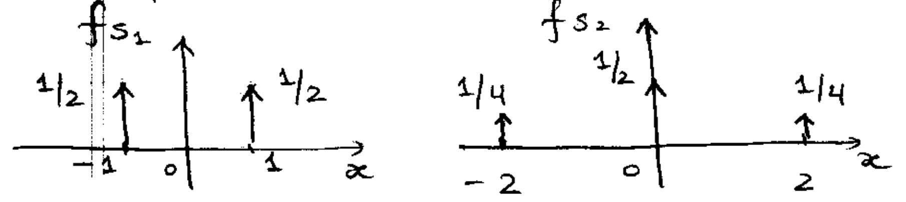
Form $S_n = \sum_{i=1}^{n} X_i$ and

$$S_n^* = \frac{S_n - E[S_n]}{\sigma_{S_n}}$$

$E[X_i] = 0, \quad var(X_i) = 1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 1$

Then $E[S_n] = 0, \quad var(S_n) = n$

The pdfs of $S_1, S_2, S_3, \ldots$



The characteristic function of S_n^*:

$$M_{S_n^*}(jv) = E\left[e^{jv \frac{\sum_{i=1}^n X_i}{n}}\right] \underset{\text{iid}}{=} M_X^n\left(\frac{jv}{n}\right)$$

$$= \left[\cos\left(\frac{v}{n}\right)\right]^n$$

Taylor series expansion:

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \ldots$$

Then,

$$M_{S_n^*}(jv) = \left\{1 - \frac{v^2}{2n} + o\left(\frac{v^2}{n}\right)\right\}^n$$

Note $\boxed{\lim_{m \to +\infty} \left(1 - \frac{1}{m}\right)^m = e^{-1}}$

Then $$\lim_{n \to +\infty} M_{S_n^*}(jv) = e^{-\frac{v^2}{2}}.$$

By continuity property

$$F_{S_n^*}(\alpha) \xrightarrow{\text{as } n \to +\infty} \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt,$$

that is, $\underset{\text{as } n \to +\infty}{S_n^* \xrightarrow{d} Z}$. $Z \hookrightarrow N(0,1)$

Application of CLT

ex. Let $X_i = \begin{cases} 1, & \text{if } A \text{ occurs}, \quad P(A) = p \\ 0, & \text{if } A \text{ does not occur} \quad P(A^c) = 1-p \end{cases}$

indicator function of event A ; assume iid

Form the sample mean $M_n = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i = \frac{n(A)}{n}$

Find : n such that $n(A)/n$ is within $\overbrace{\pm 0.01p}^{1\% \text{ of } p}$ of p with probability at least 0.95.

Form $S_n = \sum_{i=1}^{n} X_i$, $E[S_n] = n \cdot p$
$\text{var}(S_n) = np(1-p)$.

$$P\left[\left| \frac{n(A)}{n} - p \right| \leq 0.01p \right]$$

$$= P\left[-0.01p \leq \frac{\overbrace{n(A)}^{S_n} - np}{n} \leq 0.01p \right]$$

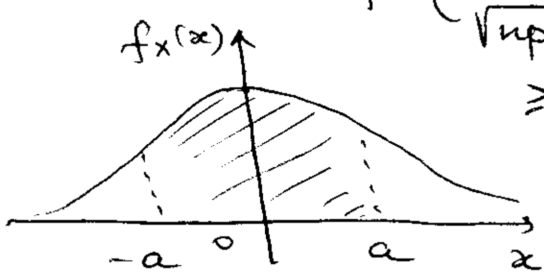
$$= P\left[\frac{-0.01pn}{\sqrt{np(1-p)}} \leq S_n^* \leq \frac{0.01pn}{\sqrt{np(1-p)}} \right],$$

apply normalization to obtain S_n^*

where $S_n^* = \dfrac{n(A) - np}{\sqrt{np(1-p)}}$

$$\cong \int_{-\frac{0.01pn}{\sqrt{np(1-p)}}}^{\frac{0.01pn}{\sqrt{np(1-p)}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt = 1 - 2\,\text{erfc}\left(\frac{0.01pn}{\sqrt{np(1-p)}} \right) \geq 0.95$$

for large n



Check :
$0.01 \cdot \sqrt{\frac{np^2}{p(1-p)}} \approx 2$, $n \geq 4 \times 10^4 \frac{1-p}{p}$
$p \geq 10^{-2}$, $n \gtrsim 4 \times 10^6$.