Limits of Random Variables

- What happens to the sum of random variables $X_1, X_2, \dots, X_n$ as the number of random variables n increases?

- The answer depends on how we normalize the sum. For instance, if $X_1, \dots, X_n$ are i.i.d., then our intuition should tell us that their average $\frac{1}{n}\sum_{i=1}^{n} X_i$ should approach $\mathbb{E}[X]$ as n increases. However, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i$ will behave more like a Gaussian random variable (with mean $\mathbb{E}[X]$).

- We will now try to make these notions precise.

- Formally, an infinite sequence of random variables $X_1, X_2, \dots$ is specified by a collection of joint CDFs (joint PMFs for the discrete case and joint PDFs for the continuous case) for every possible finite subset of random variables.
  - → We will focus on i.i.d. sequences of random variables, meaning that every possible finite subset is i.i.d.

- <u>Weak Law of Large Numbers</u>: Let $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean of an i.i.d. sequence of random variables $X_1, X_2, \ldots$ with mean $\mathbb{E}[X_i] = \mu < \infty$. For any choice of $\epsilon > 0$,

$$\boxed{\lim_{n\to\infty} \mathbb{P}[\,|M_n - \mu| > \epsilon\,] = 0}$$

→ Intuition: For any tolerance $\epsilon > 0$, the sample mean M_n eventually lands in the interval $[\mu-\epsilon, \mu+\epsilon]$.

→ How fast does this converge? We need addition assumptions.

* Assuming $\mathrm{Var}[X_i] = \sigma^2 < \infty$, $\mathbb{P}[|M_n - \mu| > \epsilon] < \frac{\sigma^2}{n\epsilon^2}$.

* Assuming $a < X < b$, $\mathbb{P}[|M_n - \mu| > \epsilon] < 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$

* Assuming $X_i \sim \mathrm{Gaussian}(\mu, \sigma^2)$, $\mathbb{P}[|M_n - \mu| > \epsilon] < 2\exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$

- <u>Strong Law of Large Numbers</u>: $\boxed{\mathbb{P}[\lim_{n\to\infty} M_n = \mu] = 1}$ Same setup as Weak Law.

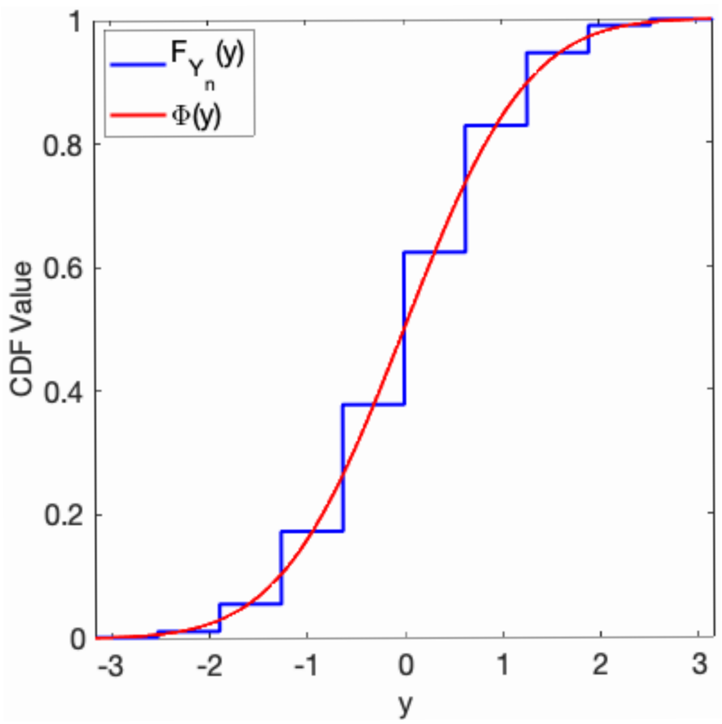
→ Intuition: Eventually, the sample mean M_n converges exactly to the true mean μ.

- **Central Limit Theorem:** For any i.i.d. sequence $X_1, X_2, \ldots$ with finite means $\mathbb{E}[X_i] = \mu$ and variances $\text{Var}[X_i] = \sigma^2$, let $S_n = \sum_{i=1}^{n} X_i$. Then, the CDF of $Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$ converges to the standard normal CDF for any value of y, $\boxed{\lim_{n\to\infty} F_{Y_n}(y) = \Phi(y).}$ ← CDF for Gaussian(0,1)

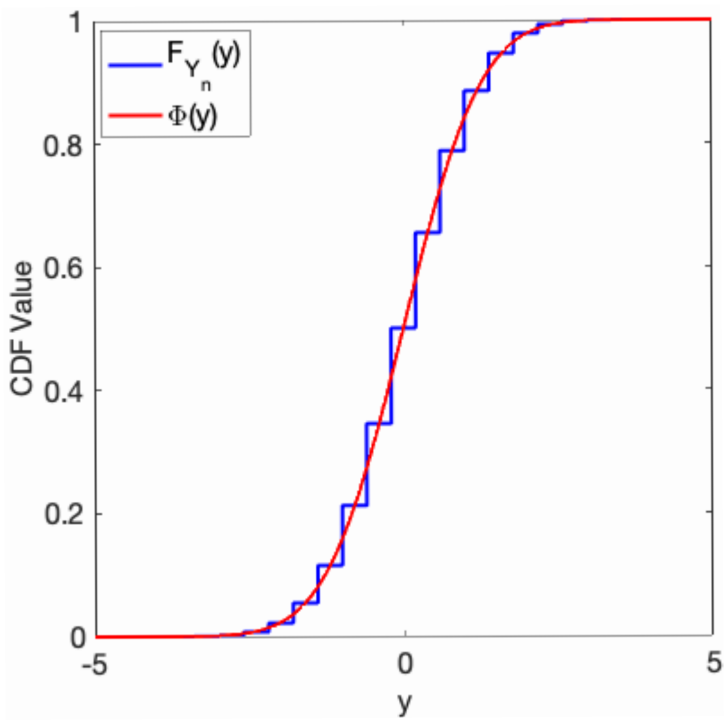
→ Intuition: $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$ looks like a Gaussian random variable for large n. In other words, the sum of many small, independent effects looks Gaussian (eventually).

→ Example: $X_1, X_2, \ldots$ i.i.d. Bernoulli$(\frac{1}{2})$ and Y_n defined as above.

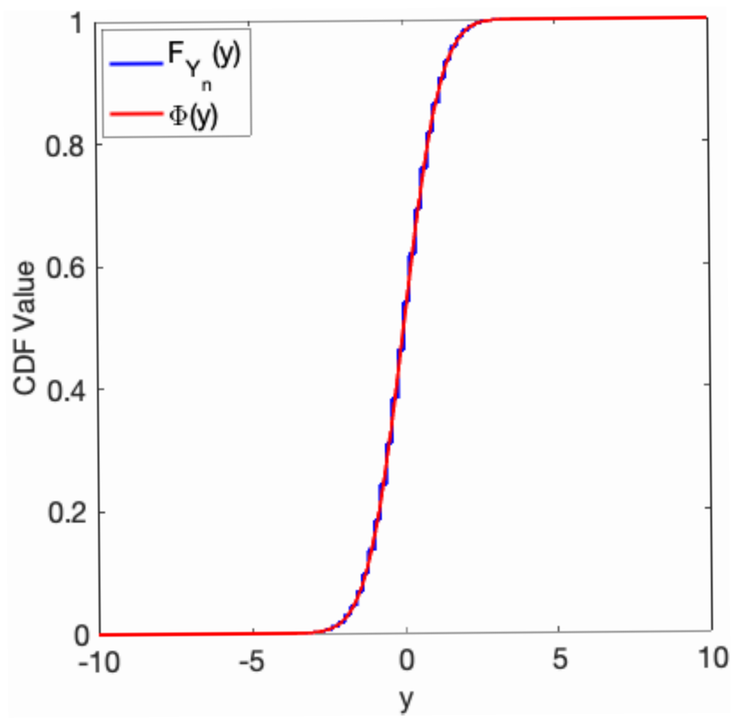
$n = 10$



$n = 25$



$n = 100$



- The Central Limit Theorem is often used to justify approximating the distribution of a sum as Gaussian.

- Example: Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$ and variance $\text{Var}[X_i] = \sigma^2$. Let $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Using the Central Limit Theorem, approximate the following:

→ $\mathbb{P}[M_n \leq c]$. We know $\mathbb{E}[M_n] = \mu$ and $\text{Var}[M_n] = \frac{\sigma^2}{n}$.

Approximate distribution as $M_n \sim \text{Gaussian}(\mu, \frac{\sigma^2}{n})$.

$$\mathbb{P}[M_n \leq c] = F_{M_n}(c) \approx \Phi\left(\frac{c-\mu}{\sqrt{\sigma^2/n}}\right)$$

→ $$\mathbb{P}[|M_n - \mu| > \epsilon] = \mathbb{P}[M_n < \mu - \epsilon] + \mathbb{P}[M_n > \mu + \epsilon]$$

$$\approx \Phi\left(\frac{\mu - \epsilon - \mu}{\sqrt{\sigma^2/n}}\right) + 1 - \Phi\left(\frac{\mu + \epsilon - \mu}{\sqrt{\sigma^2/n}}\right)$$

$$= \Phi\left(-\frac{\epsilon\sqrt{n}}{\sigma}\right) + 1 - \Phi\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)$$

$$= 2Q\left(\frac{\epsilon\sqrt{n}}{\sigma}\right)$$

← This will be useful for confidence intervals and significance testing.

Standard Normal Complementary CDF

$Q(z) = \Phi(-z) = 1 - \Phi(z)$