- Some properties of sum of random variables:

  - Say $X_1, \cdots, X_n$ are i.i.d R.V. (identitical independent distributed.)

  - $S = X_1 + \cdots + X_n$, $\qquad \overline{X} = S/n$. (sample mean)

  - Let's denote: $\mathbb{E}(X_i) = \mu$
    $Var(X_i) = \sigma^2$. $\qquad Var(X) = \mathbb{E}(\underline{(X-\mu)}^2)$.

$$\mathbb{E}(S) = \mathbb{E}(X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$$

$$= n \cdot \mu.$$

$$\mathbb{E}(\overline{X}) = \mathbb{E}(S)/n = \mu.$$

$$Var(S) = \mathbb{E}((S - n\mu)^2)$$

$$= \mathbb{E}\left(\left[\sum_{i=1}^n (X_i - \mu)\right]^2\right)$$

$$= \mathbb{E}\left(\sum_{i=1}^n \sum_{j=1}^n (X_i - \mu)(X_j - \mu)\right)$$

$$= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}\left((X_i - \mu)(X_j - \mu)\right).$$

$i \neq j$, then $\mathbb{E}((X_i - \mu)(X_j - \mu)) = \mathbb{E}(X_i - \mu) \cdot \mathbb{E}(X_j - \mu) = 0 \cdot 0$
$= 0.$

because X_i and X_j are indep.
or rather, $X_i - \mu$ and $X_j - \mu$ are indep.

$$= \sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) = \sum_{i=1}^n Var(X_i)$$

$$= n \cdot \sigma^2$$

In fact, if all X_i are indep, then

$\ast \qquad Var(\sum_i X_i) = \sum_i Var(X_i)$.

$Var(\overline{X}) = Var(S/n) = Var(S)/n^2$
$= (n \cdot \sigma^2)/n^2 = \sigma^2/n$

<u>Thm</u>: (Law of large numbers).

say $X_1, X_2, \cdots$ a seq of iid R.V.

then. "$\lim_{n \to \infty} \underbrace{\frac{X_1 + \cdots + X_n}{n}}_{\text{R.V.}} = \underbrace{\mu}_{}$." ($=\mu_n$)

constant, or a R.V. that $= \mu$ with propability one..

? what does it mean, for a sequence of R.V. μ_n to converge?

• **Thm**: (Central Limit Thm).

Say $X_1, X_2, \cdots$ iid R.V. $\quad E(X_i) = \mu, \quad Var(X_i) = \sigma^2.$

$$\mu_n = \frac{X_1 + \cdots + X_n}{n}$$ (again, μ_n is a R.V.).

Then $$\left(\frac{\mu_n - \mu}{\sqrt{Var(\mu_n)}} = \right) \frac{\mu_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{?} N(0, 1)$$

mean, std dev $= \sqrt{var}$

informally: the sample average $\mu_n \sim N(\mu, \sigma/\sqrt{n})$.

Hypothesis Testing and p-value.

• **Setup**: { • observe certain data
  • want to figure out the mechanism that generate these data.

• we assume there is a <u>fixed model</u> that produces these data, but with some unknown parameters.

$\theta \in \Theta$
 parameter space.

[for example, normal distribution, $N(\mu, \sigma)$, it has 2 parameters μ, σ. $\quad \theta = (\mu, \sigma) \quad \Theta = \mathbb{R} \times \mathbb{R}_{>0}$.]

• For each $\theta \in \Theta$, we have a probability $\mathbb{P}_\theta$

given an event $E \subset \Omega$, we can ask $\mathbb{P}_\theta(E)$. probality of E. it is a fcn of θ.

[Ex: $X \sim N(\mu, \sigma)$. $\quad \mathbb{P}_{\mu,\sigma}(X < 1) = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \cdot dx$]

• Given the output (data), try to infer what the input (parameters) are like. --- " Inferences "

- <u>Hypothesis</u> : (statement about the param.) H_0 : <u>null</u> hypothesis $\theta \in \Theta_0$

  H_1 : alternative hypothesis. $\theta \in \Theta_1 \iff \theta \notin \Theta_0$

  $\Theta = \Theta_0 \sqcup \Theta_1$ decomposition into 2 disjoint sets.

  let $\mathcal{X}$ denote the "range" ("possible values") of R.V. X, then

- <u>Test</u> : (about data, or R.V.) a <u>test</u> is a subset $R \subset \mathcal{X}$,

  R is known as the "re'ection re ion"

- We say, the test reject the null hypothesis, if

  $X \in R$.

- <u>power function</u> :

$$\beta(\theta) = \mathbb{P}_\theta(X \in R).$$

= if we had model param θ, the probability we reject H_0.

- <u>level α</u> :

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta).$$

(maximum false rejection rate.)

(the smaller α is, the better the test).