Recap:

1. Convergence concepts for a sequence of random variables:

   (i) Convergence in prob:
   
   \( X_n \to X \) in prob. if for every \( \varepsilon > 0 \),
   
   \( \lim_{n \to \infty} P(|X_n - X| > \varepsilon) = 0 \)

   (ii) Convergence in distribution:
   
   \( X_n \to X \) in dist. if
   
   \( \lim_{n \to \infty} F_{X_n}(x) = F_X(x) \), for all continuity points \( x \) of \( F_X \).

   \( F_{X_n} = \text{CDF of } X_n \)
   \( F_X = \text{CDF of } X \)

   (iii) Convergence in prob. \( \Rightarrow \) Convergence in dist. \( \not\Rightarrow \) in general.
Slutsky's theorem:
If \( X_n \xrightarrow{d} X \) and \( Y_n \xrightarrow{p} a \) (\( a \) is just a number), then
a) \( Y_nX_n \xrightarrow{d} aX \)

b) \( Y_n+X_n \xrightarrow{d} X+a \)

Weak law of large numbers
If \( X_1, \ldots, X_n \) are i.i.d. random variables with \( E[X_i] = \mu, \ Var(X_i) = \sigma^2 < \infty \) then,
\[
\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu
\]

Example: Toss a coin repeatedly.
\( X_i = 1 \) if the \( i \)th toss results in head
\( = 0 \) o.w.
\( X_i \sim Ber(\mu) \) \( E[X_i] = \mu \) \( Var(X_i) = \mu(1-\mu) < \infty \)

By weak law of large number \( \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{p} \mu \)
\( \frac{1}{n} \sum_{i=1}^{n} X_i \) = average number of heads in \( n \) tosses

Central limit theorem:
Let \( X_1, \ldots, \ldots \) be a sequence of random variables which are i.i.d. and whose MGF is bounded near 0. Let \( E[X_i] = \mu, \ Var(X_i) = \sigma^2 > 0 \).
Define \( \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \) \( \mu \)
\( Z_n = \sqrt{n} \left( \overline{X}_n - \mu \right) \)

\( \xrightarrow{d} N(0,1) \)
If we have a sequence of random variables \( X_1, \ldots \) such that \( \sqrt{n} \left( \frac{X_n - \mu}{\sigma} \right) \xrightarrow{d} N(0,1) \)

or equivalently, \( \sqrt{n} (X - \mu) \xrightarrow{d} N(0, \sigma^2) \)

then, is it possible to find the asymptotic distribution of \( \sqrt{n} (g(\bar{X}_n) - g(\mu)) \)

where \( g \) is any function?

**Delta theorem:**

Let \( Y_n \) be a sequence of random variables that satisfies \( \sqrt{n} (Y_n - \Theta) \xrightarrow{d} N(0, \sigma^2) \). For a given function \( g \), suppose \( g'(\Theta) \) exists and is not 0. Then

\[
\sqrt{n} \left( g(Y_n) - g(\Theta) \right) \xrightarrow{d} N\left(0, \sigma^2 \left[ g'(\Theta) \right]^2 \right)
\]

If \( g'(\Theta) = 0 \) and \( g''(\Theta) \) exists and is nonzero, then

\[
\sqrt{n} \left( g(Y_n) - g(\Theta) \right) \xrightarrow{d} \frac{g''(\Theta)}{2} \chi^2
\]

**Intuition:**

\[
g(x) = g(\Theta) + (x - \Theta) g'(\Theta)
\]

\[
g(Y_n) \approx g(\Theta) + (Y_n - \Theta) g'(\Theta)
\]

\[
\sqrt{n} \left( g(Y_n) - g(\Theta) \right) \approx \sqrt{n} \left( Y_n - \Theta \right) g'(\Theta)
\]

\[
\sqrt{n} (Y_n - \Theta) \xrightarrow{d} N(0, \sigma^2) \text{ for large } n \text{ approximately}
\]

\[
\Rightarrow \sqrt{n} (g(Y_n) - g(\Theta)) \xrightarrow{d} N\left(0, \left[ g'(\Theta) \right]^2 \sigma^2 \right) \text{ for large } n \text{ approximately}
\]
Example: \( X_1, \ldots, X_n \) are iid \( \text{Be}(\alpha, \beta) \) by CLT

\[ (n^1/2)(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \]

\[ g(x) = e^{x^2} \]

Find the approximate distribution of \( g(\bar{X}_n) \):
We reduce the data in our real life applications, i.e. we start with \((x_1, \ldots, x_n)\) a \(n\)-variate data vector and end up using \(T(x_1, \ldots, x_n)\) which is a smaller variate function.

For example
\[
T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \text{or} \quad T(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2
\]

\[\Phi(x) \rightarrow x_{\overline{x}}\]

So,
\[
\overline{x} = (x_1, \ldots, x_n) \sim F(x|\theta), \quad T(\overline{x}) \text{ is a function of } \overline{x}. \text{ How much reduction of } \overline{x} \text{ is allowable without losing any information on } \theta? \]

Sufficiency principle:

**Definition:** Let \(\overline{x} = (x_1, \ldots, x_n)\) and \(\overline{x} \sim F(x|\theta)\).

\(T(\overline{x})\) is known to be a sufficient statistic for \(\theta\) if the conditional distribution of \(\overline{x} | T(\overline{x})\) is independent of \(\theta\).

Intuitively, \(T(\overline{x})\) contains the same information about \(\theta\) that \(\overline{x}\) contains.
Example: \( X_1, X_2, X_3 \sim \text{Ber}(p) \).

\[
P(X_1 = x) = p^x (1-p)^{1-x}, \quad \text{if} \quad x = 0, 1
\]

\[
T(X_1, X_2, X_3) = X_1 + X_2 + X_3
\]

Let's look at the distribution of \( (X_1, X_2, X_3) | X_1 + X_2 + X_3 = t \).

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | X_1 + X_2 + X_3 = t) = 0 \quad \text{if} \quad t \neq x_1 + x_2 + x_3
\]

When \( t = x_1 + x_2 + x_3 \)

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | X_1 + X_2 + X_3 = t) = \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(X_1 + X_2 + X_3 = t)}
\]

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{P(X_1 = x_1) P(X_2 = x_2) P(X_3 = x_3)}{P(X_1 + X_2 + X_3 = t)}
\]

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{p^{x_1} (1-p)^{1-x_1} p^{x_2} (1-p)^{1-x_2} p^{x_3} (1-p)^{1-x_3}}{P(X_1 + X_2 + X_3 = t)}
\]

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{p^{x_1} (1-p)^{1-x_1} p^{x_2} (1-p)^{1-x_2} p^{x_3} (1-p)^{1-x_3}}{P(X_1 + X_2 + X_3 = t)}
\]

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{p^{x_1+x_2+x_3} (1-p)^{3-(x_1+x_2+x_3)}}{P(X_1 + X_2 + X_3 = t)}
\]

\[
P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = \frac{p^{x_1+x_2+x_3} (1-p)^{3-(x_1+x_2+x_3)}}{\binom{3}{t} p^t (1-p)^{3-t}}
\]
\[
\theta = \frac{p^3 (1-p)^{3-t}}{(3-t) p^3 (1-p)^{3-t}} = \frac{1}{(3-t)}
\]

Thus \((X_1, X_2, X_3) \mid X_1 + X_2 + X_3\) is independent of \(p\).

Hence by definition of sufficient statistic \(X_1 + X_2 + X_3\) is a sufficient statistic for \(p\).

\[
X_1, X_2, X_3 \sim \text{Bin}(p)
\]

<table>
<thead>
<tr>
<th>cases</th>
<th>prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>000</td>
<td>((1-p)^3)</td>
</tr>
<tr>
<td>001</td>
<td>((1-p)^2p)</td>
</tr>
<tr>
<td>100</td>
<td>(p(1-p)^2)</td>
</tr>
<tr>
<td>010</td>
<td>(p^2(1-p))</td>
</tr>
<tr>
<td>110</td>
<td>(p^2(1-p))</td>
</tr>
<tr>
<td>011</td>
<td>(p^3)</td>
</tr>
<tr>
<td>101</td>
<td>(p^3)</td>
</tr>
<tr>
<td>111</td>
<td>(p^3)</td>
</tr>
</tbody>
</table>

Therefore, \(\Omega = \{000, 001, 100, 010, 110, 011, 101, 111\}\).

\[
\begin{align*}
\mathcal{A}_0 & = \{000\}, \quad \mathcal{A}_0^1 = \{001, 100, 010\}, \\
\mathcal{A}_2^0 & = \emptyset, \quad \mathcal{A}_2^0 = \{110, 011, 101\}, \quad \mathcal{A}_3^0 = \{111\}
\end{align*}
\]

Within \(\mathcal{A}_t\), \(X_1 + X_2 + X_3 = t\), and within each \(\mathcal{A}_t\) for any sample point the prob. in \(p^t (1-p)^{3-t}\).
Goal: Find sufficient statistic from any situation when joint density/p.m.f. of $X = (X_1, \ldots, X_n)$ is known.

Theorem (Factorization theorem):

Let $X$ have joint p.d.f. (or p.m.f.) $f_\theta(x)$, where $\theta$ is the unknown parameter. A statistic $T(X)$ is sufficient for $\theta$ if and only if $f_\theta(x)$ can be expressed as

$$f_\theta(x) = g(T(X), \theta) h(x),$$

where $h(x)$ is a function of $x$ independent of $\theta$.

Proof: Only if part:

$$P(X = x) = \sum_t P(X = x \mid T(X) = t) P(T(X) = t)$$

There is only one $t$ for which the $P(X = x \mid T(X) = t)$ is positive.

$$P(X = x) = \underbrace{P(X = x \mid T(X) = t)}_{\text{independent of } \theta} \underbrace{P(T(X) = t)}_{\text{independent of } \theta}$$

$$= g(T(X), \theta) h(x)$$
If \( X = (x_1, x_2, x_3) \)
\[
(x_1, x_2, x_3) = (x_1, x_2, x_3) \implies x_1 + x_2 + x_3 = x_1 + x_2 + x_2
\]

In the if part we assume
\[
f_\theta(x) = g(T(x), \theta) h(x)
\]
and we have to show that \( X \mid T(x) \) is independent of \( \theta \).

\[
P(T(x) = t) = \sum_{x \in \mathcal{X}_t} f_\theta(x)
\]
where \( \mathcal{X}_t = \{ (x_1, x_2, \ldots, x_n) : T(x_1, x_2, \ldots, x_n) = t \} \)
\[
= \sum_{x \in \mathcal{X}_t} g(T(x), \theta) h(x)
\]
\[
= \sum_{x \in \mathcal{X}_t} g(t, \theta) h(x)
\]
\[
= g(t, \theta) \sum_{x \in \mathcal{X}_t} h(x)
\]
\[
P(X = x \mid T(x) = t) = \begin{cases} 0 & \text{if } t \neq T(x) \\ \frac{P(X = x)}{P(T(x) = t)} & \text{if } t = T(x) \end{cases}
\]
\[
\frac{P(X = x)}{P(T(x) = t)} = \frac{g(T(x), \theta) h(x)}{g(t, \theta) \sum_{x \in \mathcal{X}_t} h(x)} = \frac{g(t, \theta) h(x)}{g(t, \theta) \sum_{x \in \mathcal{X}_t} h(x)}
\]