The CLT says that if \( X_i \sim \text{any} \) dist. with finite mean \( \mu_X \) and finite variance \( \sigma^2_X \), then the distribution of \( \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \) is approximately normal, where \( \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \).

This is equivalent to saying that

\[
\frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1)
\]

A question: If \( g(x) \) is a sufficiently "nice" function, is there a comparable result for \( g(\bar{X}_n) \)?

Answer: Yes, via a Taylor-series-based approach called the Delta Method.
This is why it's called the \( \Delta \) (Delta) Method.

So \( \Delta = x - x - 1 \)

\[ m \Delta + \Delta = x \]

\[ \Delta = \frac{b}{y} \]

\[ g(x) = \frac{1}{y} + \frac{1}{y} \cdot (x - y) \]

\[ \frac{\Delta}{b} = \frac{1}{y} \]

\[ g(x) = g(x) + g'(x) \Delta + \frac{1}{2} g''(x) \Delta^2 \]

This suggests making a two-term Taylor expansion of \( g(x) \) around the point \( x \).

(That's the weak law of large numbers.)
\[ g(\overline{X}_n) = g(\mu_X) + g'(\mu_X) (\overline{X}_n - \mu_X) \] 

\[ \text{constant} \quad \overrightarrow{\text{rv.}} \] 

\[ E[g(\overline{X}_n)] = E[g(\mu_X) + g'(\mu_X) (\overline{X}_n - \mu_X)] \]

\[ = g(\mu_X) + g'(\mu_X) \left[ E(\overline{X}_n) - \mu_X \right] \]

\[ \text{so} \quad E[g(\overline{X}_n)] = g(\mu_X) = g \left[ E(\overline{X}_n) \right] \]

\[ \text{and} \quad \overrightarrow{\text{constant}} \]

\[ \text{constant} \quad \overrightarrow{\text{rv.}} \]

\[ \text{var} \left[ g(\overline{X}_n) \right] = \text{var} \left[ g(\mu_X) + g'(\mu_X) (\overline{X}_n - \mu_X) \right] \]

\[ = \left[ g'(\mu_X) \right]^2 \text{var} (\overline{X}_n - \mu_X) \]

\[ \text{so} \quad \text{var} \left[ g(\overline{X}_n) \right] = \left[ g'(\mu_X) \right]^2 \frac{\sigma^2}{n} \]
There's one hidden assumption in this calculation: \( g'(\mu_V) \neq 0 \).

This works for any \( R.V. \), not just \( X_n \):

For any \( R.V. \) with finite variance \( \sigma_V^2 \) (and therefore finite mean \( \mu_V \)), \( W = g(V) \)

\[ E(W) = g(\mu_V) \quad \text{and} \]

\[ V(W) = \left[ g'(\mu_V) \right]^2 \sigma_V^2, \]

provided \( g'(V) \) is continuous and \( g'(\mu_V) \neq 0 \).

Moreover, if \( V \sim \text{Normal} \), then \( W = g(V) \sim \text{Normal} \) also.
Example 3.16. A bank typically has a single queue (line) at which customers arrive to transact banking business.

Let $\xi_i = $ time customer $i$ waits from reaching the head of the queue until served. To be completely realistic, the dist. of $\xi_i$ would vary by day of week and time of day, so pick a single time slot (e.g. Tue 10-10.15am) and observe the $\xi_i$ from week to week only in that time slot; now the $\{\xi_i, i = 1, 2, \ldots\}$ form a stationary stochastic process with fixed (non-time-varying) $E(\xi) = \mu$. 

and fixed (non-time-varying) finite

\[ V(X_i) = \sigma^2. \] 

Gather data over many

weeks and form \( \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \)

for large \( n \). The rate of service

is defined to be \( g(\mu_X) = \frac{1}{\mu_X} \), which

would naturally be estimated by \( g(\bar{X}_n) = \frac{1}{\bar{X}_n} \).

\[ E(\bar{X}_n) = \mu_X \]

\[ V(\bar{X}_n) = \frac{\sigma^2}{n} \]

\[ g(x) = \frac{1}{x} = x^{-1} \]

\[ g'(x) = -\frac{1}{x^2} \]

\[ g'(\mu_X) = -\frac{1}{\mu_X^2} \]

\( \bar{X}_n \sim \text{Normal} \) by CLT

so \( \Delta \)-method says \( g(\bar{X}_n) = \frac{1}{\bar{X}_n} \sim \text{Normal} \)

also, with mean \( g(\mu_X) = \frac{1}{\mu_X} \) and variance

\[ \left( g'(\mu_X) \right)^2 = \frac{1}{\mu_X^4} \neq 0 \]

\[ \sigma^2 / (n \mu_X^4) \]
Specific calculation

Under some plausible assumptions, we'll see that $(X_i | \theta) \overset{\text{iid}}{\sim} \text{Exponential}(\theta)$ may be a reasonable model for waiting times.

\[ E(X_i) = \frac{1}{\theta} \quad \text{Var}(X_i) = \frac{1}{\theta^2} \]

Thus, $(X_i | \theta) \overset{\text{d}}{\sim} \text{Exp}(\theta)$ for $\theta > 0$

so $\frac{1}{\bar{X}_n}$ should (large) for

be approximately Normal with mean $\frac{1}{\frac{1}{\theta}}$

and SD

\[ \frac{\sqrt{\text{Var}(X_i)}}{n^{1/2}} = \frac{\theta}{\sqrt{n}} \]

(continuous)

Feynman version

$Z_1, Z_2, \ldots$ sequence of $\text{Exp}(\theta)$ of $\Delta$-method $F^{*}$ continuous CDF,

$\Theta$ a real number; $\alpha_1, \alpha_2, \ldots \uparrow \theta$ positive sequence
$g(\cdot)$ a function of a real variable such that $g'(\cdot)$ is continuous and $g'(\theta) \neq 0$; then if $a_n (\overline{X}_n - \theta) \xrightarrow{D} F^*$,

$$\frac{a_n \left[ g(\overline{X}_n) - g(\theta) \right]}{|g'(\theta)|} \xrightarrow{D} F^*$$

Typical application:

$$\overline{X}_1, \overline{X}_2, \ldots \text{IID}$$

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i; \quad \theta = \mu_X; \quad a_n = \frac{\sqrt{n}}{\sigma_X};$$

$$F^* = F$$, the standard normal CDF.

In this context, the theorem says that

if

$$\frac{\overline{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \sim N(0, 1) \quad \text{then} \quad g(\overline{X}_n) - g(\mu_X)$$

is also $\sim N(0, 1)$.
A little bit more about the continuity correction

Toy Sachs case study, revisited

\[ \mathbf{X} \sim \text{Binomial}(n, p) \]

\[ f(\mathbf{X} = k) = \binom{n}{k} p^k (1-p)^{n-k} \]

in family of \( n = 5 \) children, both parents carriers so that

\[ P(\text{T-5 baby}) = \frac{1}{4} = p \]

But also let \( T_i := \begin{cases} 1 & \text{if child } i \text{ is T-5 baby} \\ 0 & \text{else} \end{cases} \)

Then \((T_i) \sim \text{Bernoulli}(p)\) and \[ \mathbf{X} = \sum_{i=1}^{n} T_i \]

So by the CLT the dist. of \( \mathbf{X} \) should be approximately normal with mean

\[ \mu_X = E(\mathbf{X}) = np = 1.25 \text{ and } \sigma^2 = np(1-p) = 0.625 \]
\[ \delta \bar{X} = \sqrt{\text{Var}(\bar{X})} = \sqrt{np(1-p)} = 0.98 \]

On day 1 of this class we worked out

that \[ P(1 \text{ or more T-S babies}) = P(\bar{X} \geq 1) \]

\[ 1 - P(\text{no T-S babies}) = 1 - (1-p)^n = 0.76 \]

\[ = 1 - P(\bar{X} = 0) \]

Naive Normal approximation, from CLT:

\[ \mu \quad \sigma \quad PDF \]

\[ \bar{X} \quad 0.98 \quad 0.398 \]

\[ 1.0 - 1.25 \]

\[ 0.98 - 0.26 \]

\[ 0.602 \text{ (quite a bad approximation)} \]
Improved approximation obtained by paying attention to the edge of the histogram (VF) bars:

Normal approximation with continuity correction

\[ P(X \geq 1) = 1 - P(X < 0.5) \]

\[ = 1 - 0.219 \]

\[ = 0.781 \] (much better approx.)

Markov Chains Recall the definition of a stochastic process:
Def. A sequence of rvs \(X_1, X_2, \ldots\) is called a stochastic process with discrete time parameter \(t = 1, 2, \ldots\). \(X_0\) is the initial state of the process; \(X_n, n \geq 1\) is the state of the process at time \(t = n\). The simplest possible discrete-time stochastic process is an IID sequence of rvs \((X_1, X_2, \ldots)\).

Suppose that there's a parameter \(\theta\) such that \((X_i | \theta)\) IID from some dist. depending on \(\theta\). 

Q: Does this process have a memory?
Example, machine with $\theta$ dial from 0 to 1, produces IID Bernoulli($\theta$) trials $X_i$. The process $(X_1, X_2, \ldots)$ does have a memory if $\theta$ is unknown to you; the information that 17 out of the first 20 trials were successes helps you predict $X_21$, because it's reasonable to conclude from $X_1, \ldots, X_{20}$ that $\theta$ is around $\frac{17}{20} = 0.85$, so $X_{21}$ will be probably a success. But the process $(X_i; \Theta), i = 1, 2, \ldots$ has no memory once $\theta$ is known: information about
The first n trials is irrelevant to your prediction of \( \Xi_{n+1} \) if you know \( \Theta \).

\[ \text{An IID process } (X_i; \Theta) \text{ is called a white-noise (stochastic) process or a white noise time series.} \]

Q: What's the next level of complexity for discrete-time stochastic processes up from white noise? 

A: Allow \( \Xi_{n+1} \) to depend on \( \Xi_n \) but not on \( \Xi_{n-1}, \Xi_{n-2}, \ldots \) (i.e., let the process have a short-term memory, \( \Theta \) time period back in the past).
From now on, I'll suppress the dependence of the process on $t$ in the notation.

**Def.** A stochastic process is a \textit{discrete-time} \textit{(first-order) Markov chain} if for $n = 1, 2, \ldots$, $b$ any real number, and for all possible sequences of states $x_1, x_2, \ldots$

$$P(X_{n+1} \leq b \mid X_1 = x_1, \ldots, X_n = x_n)$$

$$= P(X_{n+1} \leq b \mid X_n = x_n).$$

In other words, the only thing you need to know to simulate where the Markov chain is going next is where it is now.
(Can define higher-order Markov chains with memory of 2 or more time periods; we won't pursue that here.)

**Def.**

The set of values a Markov chain can take on is called its state space \( S \), which may be finite or infinite.

(Can also have Markov chains unfolding in continuous time, e.g. \( X_t \) = stock price at time \( t \) = seconds, milliseconds, microseconds,...; we also won't pursue that here.)

It's easy to write down the joint PMF of a Markov chain with finite
Def. A Markov chain with a finite state space is called a finite Markov chain.

Suppose you have a finite Markov chain with \( k \) possible states numbered \( 1, 2, \ldots, \).

\[
\pi(x) = P(X_1 = x) P(X_2 = x_2 | X_1 = x) \ldots P(X_n = x_n | X_{n-1} = x_{n-1})
\]

\[
\pi(x_1, x_2, \ldots) \text{ finite Markov chain}
\]

**Definition**

- A Markov chain with a finite state space is called a finite Markov chain.
If \( P(X_{n+1} = j \mid X_n = i) \) is the same for all \( n \), the transition distribution is said to be stationary. If the Markov chain does not have a stationary transition distribution, then the probability \( p_{ij} = P(X_{n+1} = j \mid X_n = i) \) completely characterize the Markov chain's behavior. Can arrange the \( p_{ij} \) in a matrix called the transition matrix.
All of the elements of $P$ are non-negative (they're probabilities), and all of the row sums are 1 (because the chain has to go somewhere), i.e.
\[
\sum_{j=1}^{k} p_{ij} = 1 \quad \text{for all } i = 1, \ldots, k.
\]

A square matrix $P$ with non-negative entries and all row sums equal to 1 is called a stochastic matrix.

Example: Gene inheritance is Markovian.

Gene inheritance is Markovian, all we need to know to predict you is the genetic story of your parents.
Suppose that a gene of interest to you has two alleles, A and a. Then a state in the Markov chain is of the form:

\[
\begin{pmatrix}
\text{allele 1} & \text{allele 2} & \text{allele 1} & \text{allele 2} \\
\text{from point 1} & \text{from point 2} & \text{from point 1} & \text{from point 2}
\end{pmatrix}
\]

For example, \{Aa, Aa\}. Ignoring order (because it's irrelevant in inheritance), there are 6 possible states: \{AA, AA\}, \{AA, Aa\}, \{AA, aa\}, \{Aa, Aa\}, \{Aa, aa\}, and \{aa, aa\}.
One possible inheritance sequence:

Offspring gets A or a from parent 1 and A or a (independently) from parent 2 (stora) each with probability $\frac{1}{2}$.

<table>
<thead>
<tr>
<th>State</th>
<th>AA, AA</th>
<th>AA, Aa</th>
<th>AA, aa</th>
<th>Aa, Aa</th>
<th>Aa, qa</th>
<th>qa, qa</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA, AA</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA, Aa</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AA, aa</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Aa, Aa</td>
<td>$\frac{1}{16}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{16}$</td>
</tr>
<tr>
<td>Aa, qa</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
<td>0</td>
</tr>
<tr>
<td>qa, qa</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>qa, qa</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Example (random walk) You're watching a particle move around on the integers \( \mathbb{Z} = \{ \ldots, -2, -1, 0, 1, 2, \ldots \} \) over time; here are the rules:
wherever it is at time \( t = n \), it moves left 1 unit with prob \( p_1 \),
right 1 unit with \( p_2 \), and it stays where it is with prob \( p_3 \),
where \( p_1 + p_2 + p_3 = 1 \).
This is clearly a Markov chain (why?); what is its transition matrix?
This is an example of a **banded matrix**, in which the only non-zero main entry are on the diagonal and 1 diagonal either way from the main diagonal; since there are only 3 main diagonals, $P$ is said to be tridiagonal.
Moreover, all of the main diagonal entries are the same \( (p_1) \); all of the entries 1 diagonal below are also the same \( (p_1) \); and all of the entries 1 diagonal above are also the same \( (p_2) \).

Such matrices are called **Toeplitz** matrices (named after Otto Toeplitz, 1880–1940, a German mathematician who was tuberculosis-ridden and 58 years old, died 1940).

Start this process, which is called a **random walk**, at 0 & let it go; where is the particle likely to be at time \( n \), \( n \) large?
A: I suppose, for example, that
\((q_1, q_2, q_3) = (0.1, 0.3, 0.6)\).

Then you would expect the particle to drift off to \(\infty\). Similarly,

\((q_1, q_2, q_3) = (0.5, 0.25, 0.25)\) should yield a drift to \(-\infty\).

Can show that as \(n \to \infty\) every integer is visited infinitely many times, and the expected time you must wait for the chain to return to 0 (having started there) is also infinite.

The infinite random walk evidently has “too much freedom” to move around to get interesting results; let’s bound it.