the Delta method

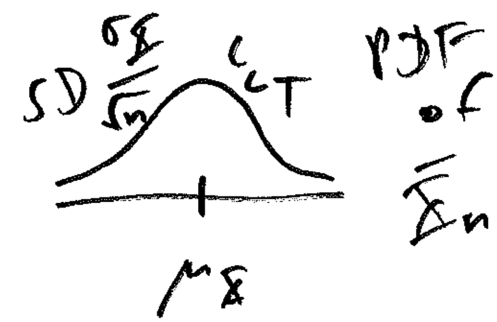
the CLT says that if

$X_i \overset{IID}{\sim}$ any dist. with finite mean μ_X and finite variance σ_X^2, then

the distribution of $\dfrac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}}$ for large n

is approximately standard normal, where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

This is equivalent to saying that



$$\bar{X}_n \,\dot\sim\, N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

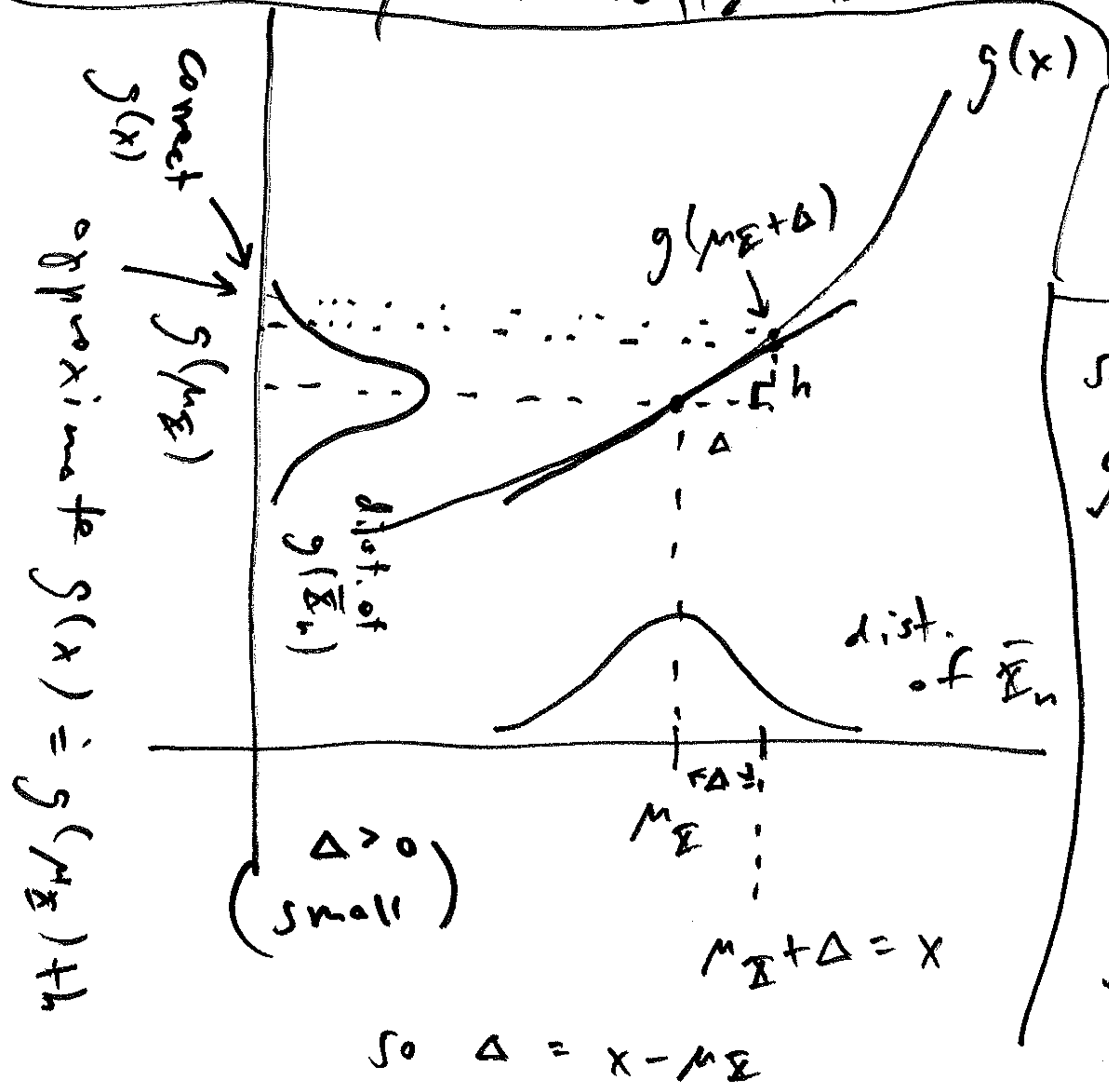
Question: If $g(x)$ is a sufficiently "nice" function, is there a comparable result ~~for~~ for $g(\bar{X}_n)$?

Answer: Yes, via a Taylor-series-based approach called the Delta method

$\bar{X}_n$ should be close to μ_X for large n (that's the (weak) Law of Large Numbers); this suggests making a two-term Taylor expansion of $g(\bar{X}_n)$ around the point

$$x = \mu_X : \quad g(\bar{X}_n) \doteq g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)$$

This is why it's called the Δ (Delta) - Method



$$\frac{h}{\Delta} = g'(\mu_X)$$

so

$$g(x) \doteq g(\mu_X) + h$$

$$= g(\mu_X) + g'(\mu_X) \cdot \Delta$$

$$= g(\mu_X) + g'(\mu_X)(x - \mu_X)$$

$$g(\bar{X}_n) \doteq g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X) \quad \text{so}$$

r.v.

constant

$$E[g(\bar{X}_n)] \doteq E\left[g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)\right]$$

$$= g(\mu_X) + g'(\mu_X)\left[\cancel{E(\bar{X}_n) - \mu_X}^{\,0}\right]$$

$$\text{so } E[g(\bar{X}_n)] \doteq g(\mu_X) = g[E(\bar{X}_n)]$$

and

constant

$$V[g(\bar{X}_n)] \doteq V\left[g(\mu_X) + g'(\mu_X)(\bar{X}_n - \mu_X)\right]$$

r.v.

$$= [g'(\mu_X)]^2 \cdot V(\bar{X}_n - \mu_X)$$

$$\text{so } V[g(\bar{X}_n)] \doteq [g'(\mu_X)]^2 \, V(\bar{X}_n)$$

$$V[g(\bar{X}_n)] \doteq [g'(\mu_X)]^2 \frac{\sigma_X^2}{n}$$

There's one hidden assumption in this calculation: $g'(\mu_{\bar{X}}) \neq 0$.

This works for any r.v. with finite variance, not just $\bar{X}_n$:

V any r.v. with finite variance σ_V^2 (and therefore finite mean μ_V), $W = g(V)$

$\rightarrow$ $E(W) \doteq g(\mu_V)$ and

$V(W) \doteq [g'(\mu_V)]^2 \sigma_V^2$,

$\boxed{\Delta \text{ method}}$ part 1

provided $g'(v)$ is continuous and

$g'(\mu_V) \neq 0$

Moreover, if V is Normal then $W = g(V)$ is Normal also

$\boxed{\Delta \text{ method}}$ part 2

Example A bank typically has a single queue (line) at which customers arrive to transact banking business.

Let X_i = time customer i waits from reaching the head of the queue until served. To be completely realistic, the dist. of X_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 X_i from week to week only in that time slot; now the $\{X_i, i = 1, 2, ..\}$ form a stationary stochastic process with fixed (non-time-varying) finite $E(X_i) = \mu_X > 0$

and fixed (non-time-varying) finite

$V(X_i) = \sigma_X^2$. 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

complication: seasonal effects (ignored here)

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

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

$E(\bar{X}_n) = \mu_X$

$V(\bar{X}_n) = \frac{\sigma_X^2}{n}$

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

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

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

$\bar{X}_n \,\dot\sim$ Normal by CLT

so Δ-method says $g(\bar{X}_n) = \frac{1}{\bar{X}_n} \,\dot\sim$ Normal also,

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

$\sigma_X^2 / (n \mu_X^4)$

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

Specific calculation

Under some plausible assumptions, we've seen that $(X_i \mid \lambda) \stackrel{IID}{\sim} \text{Exponential}(\lambda)$ may be a reasonable model for waiting times.

$E(X_i) = \frac{1}{\lambda} = \mu_X$, $V(X_i) = \frac{1}{\lambda^2} = \sigma_X^2$

$(X_i \mid \lambda)$ has PDF $f_{X_i}(x_i \mid \lambda) = \lambda e^{-\lambda x_i} I(x_i > 0)$

so $\frac{1}{\bar{X}_n}$ should (for large n) be approximately Normal with mean $\frac{1}{\frac{1}{\lambda}} = \lambda$

and SD $\frac{\sigma_X}{\mu_X^2 \sqrt{n}} = \frac{\frac{1}{\lambda}}{\left(\frac{1}{\lambda}\right)^2 \sqrt{n}} = \frac{\lambda}{\sqrt{n}}$.

Fancy version of Δ-method

$Y_1, Y_2, \ldots$ sequence of (discrete or continuous) r.v.; F^* continuous CDF;

θ a real number; $a_1, a_2, \ldots \uparrow \infty$ positive sequence

$g(\cdot)$ a real-valued function of a real variable such that $g'(\cdot)$ is continuous and $g'(\theta) \neq 0$; then if $a_n(Y_n - \theta) \xrightarrow{D} F^*$,

$$a_n\left[\frac{g(Y_n) - g(\theta)}{|g'(\theta)|}\right] \xrightarrow{D} F^* \text{ also}$$

Typical application: $X_1, X_2, \ldots$ IID

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

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

In this context the theorem says that

$$\text{if } \frac{\bar{X}_n - \mu_X}{\sigma_X/\sqrt{n}} \;\dot\sim\; N(0,1) \xrightarrow{\text{then}} \frac{g(\bar{X}_n) - g(\mu_X)}{|g'(\mu_X)|\,\sigma_X/\sqrt{n}} \text{ is also } \dot\sim\; N(0,1)$$

(28 Aug 17)

A little bit more about the continuity correction

Tay-Sachs case study, revisited

X = # T-S babies in family of $n = 5$ children, both parents carriers so that

$P(\text{T-S baby}) = \frac{1}{4} = p$ $\boxed{X \sim \text{Binomial}(n, p)}$

But also let $T_i = \begin{cases} 1 & \text{if child } i \text{ is T-S baby} \\ 0 & \text{else} \end{cases}$ $i = 1, \dots, n = 5$

Then $(T_i) \overset{IID}{\sim} \text{Bernoulli}(p)$ $(i = 1, \dots, n)$ and $X = \sum_{i=1}^{n} T_i$

So by the CLT the dist. of X should be approximately Normal with mean

$\mu_X = E(X) = np = 1.25$ and SD

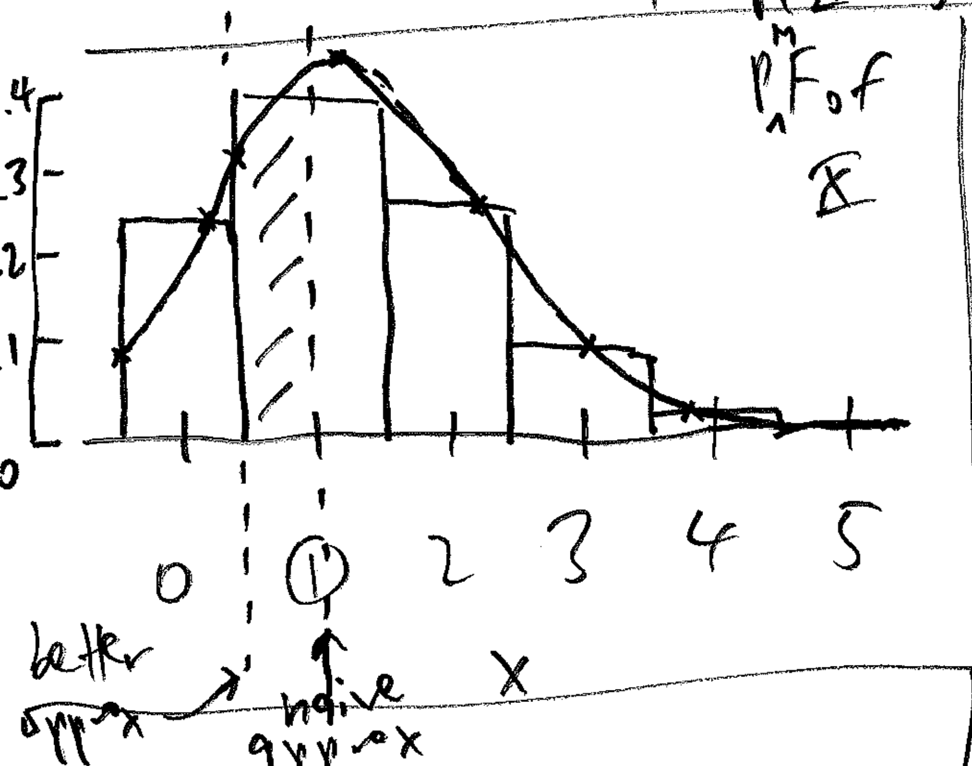
$$\sigma_X = \sqrt{V(X)} = \sqrt{np(1-p)} \doteq 0.98$$

On day 1 of this class we worked out

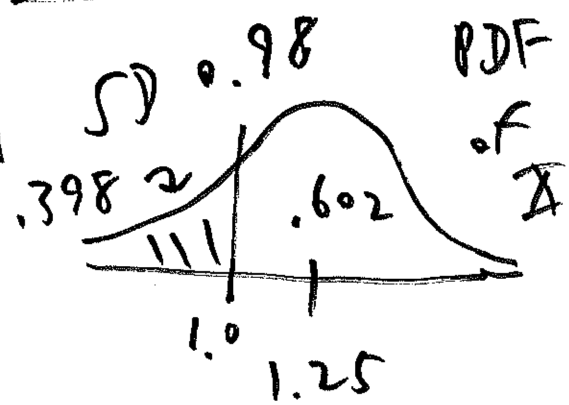
that $P($1 or more T-S babies$) = P(X \geq 1)$

$1 - P($no T-S babies$) = 1 - (1-p)^n \doteq 0.76$

$= 1 - P(X = 0)$



Naive Normal approximation, from CLT:



$$P(X \geq 1) \doteq 1 - P(X' < 1)$$

$$= 1 - 0.398$$

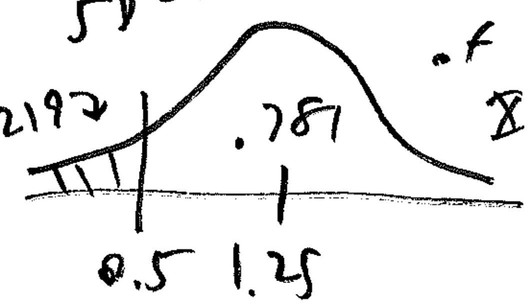
$$\doteq 0.602 \quad \text{(quite a bad approximation)}$$

$$\frac{1.0 - 1.25}{0.98} \doteq -0.26$$

Improved approximation obtained by paying attention to the edges of the histogram (PMF) bars:

Normal approximation with <u>continuity correction</u>

SD 0.98 PDF of X'



$$\frac{0.5-1.25}{0.98} \doteq -0.77$$

$$P(X \geq 1) \doteq 1 - P(X' < 0.5)$$

$$\doteq 1 - .219$$

$$\doteq 0.781$$ (correct answer 0.76; much better approx.)

Markov Chains	Recall the definition of a <u>stochastic process</u>:

Def. A sequence of rvs $X_1, X_2, \ldots$ is called a stochastic process with discrete time parameter $t = 1, 2, \ldots$.

X_1 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 θ such that $(X_i \mid \theta) \stackrel{IID}{\sim}$ from some dist. depending on θ. $\boxed{\text{Q:}}$ Does this process have a memory?

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

An IID process $(X_i \mid \theta) \stackrel{IID}{\sim}$ is called a <u>white-noise (stochastic) process</u> or a <u>white noise time series</u>.

Q: What's the next level of complexity for discrete-time stochastic processes up from white noise? A: Allow X_{n+1} to depend on X_n but not on $X_{n-1}, X_{n-2}, \ldots$ (i.e., let the process have a short-term memory, 1 time period back in the past).

From now on, I'll suppress the dependence of the process on θ in the notation.

Def. A discrete-time stochastic process is a (first-order) <u>Markov chain</u> 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 <u>next</u> is <u>where it is now</u>.

(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 <u>state space</u> 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 PF of a Markov chain with finite S:

Consequences

① $(X_1, X_2, \ldots)$ finite Markov chain →

Def. A Markov chain with a finite state space is called a <u>finite</u> Markov chain.

$$P(X_1 = x_1, \ldots, X_n = x_n) =$$

$$P(X_1 = x_1) \cdot P(X_2 = x_2 \mid X_1 = x_1) \cdot$$

$$P(X_3 = x_3 \mid X_2 = x_2) \cdot \ldots$$

$$P(X_n = x_n \mid X_{n-1} = x_{n-1}).$$

Def. Suppose you have a finite Markov chain with k possible states numbered $1, \ldots, k$ (k integer ≥ 2) $\rightarrow \{P(X_{n+1} = j \mid X_n = i),$ $i, j = 1, \ldots, k,\ n = 1, 2, \ldots\}$ are called the <u>transition distribution</u> of the 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 (DS) (time-homogeneous). If the Markov chain does have a stationary transition distribution, then the probabilities

$$p_{ij} \triangleq p(X_{n+1}=j \mid X_n=i)$$

completely characterize the Markov chain's behavior.

Can arrange the p_{ij} in a matrix called the transition matrix.

to state

$$\underset{k\times k}{P} = \text{from state}\;\begin{array}{c|cccc} & 1 & 2 & \cdots & k \\ \hline 1 & p_{11} & p_{12} & \cdots & p_{1k} \\ 2 & \vdots & & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ k & p_{k1} & p_{k2} & \cdots & p_{kk} \end{array}$$

All of the elements of $\underline{P}$ are non-negative (they're probabilities), and all of the row sums are 1 (because the chain has to go somewhere), i.e.

$$\sum_{j=1}^{k} p_{ij} = 1 \text{ for all } i = 1, \ldots, k.$$

matrix versus quaternion

Def.

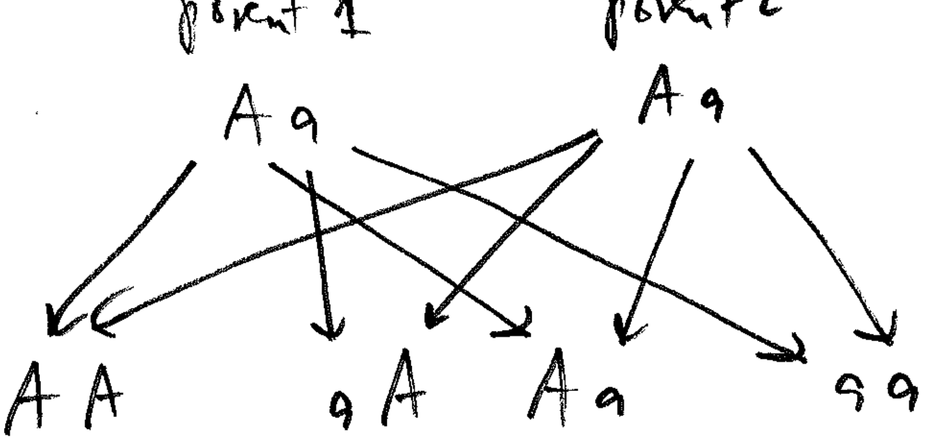
A square matrix $\underset{k \times k}{\underline{P}}$ with non-negative entries and all row sums equal to 1 is called a stochastic matrix.

~~(Dominant/recessive)~~

Example: Gene inheritance is Markovian: all we need to know to predict you is the genetic story of your parents (your grand parents, ..., are irrelevant).

~~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

{allele 1 from parent 1, allele 2 from parent 1, allele 1 from parent 2, allele 2 from parent 2}, 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,

each (A or a) with probability $\frac{1}{2}$.

Transition matrix

$$
\begin{array}{c|c}
 & \text{to} \rightarrow \quad \{AA,AA\} \quad \{AA,Aa\} \quad \{AA,aa\} \quad \{Aa,Aa\} \quad \{Aa,aa\} \quad \{aa,aa\} \\
\begin{array}{c} \text{from} \downarrow \\ \{AA,AA\} \\ \{AA,Aa\} \\ \{AA,aa\} \\ \{Aa,Aa\} \\ \{Aa,aa\} \\ \{aa,aa\} \end{array} &
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{2} & 0 & \frac{1}{4} & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 \\
\frac{1}{16} & \frac{1}{4} & \frac{1}{8} & \frac{1}{4} & \frac{1}{4} & \frac{1}{16} \\
0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\
0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\end{array}
$$

Example (random walk) You're watching a particle move around on the integers $S = \{ \ldots, -2, -1, 0, 1, 2, \ldots \}$ 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 —— p_3,

and it stays where it is with prob p_2,

where $0 < p_i < 1$ and $\sum_{i=1}^{3} p_i = 1$. This is clearly a Markov chain (why?);

what is its transition matrix?

$$
\begin{array}{c|c}
\text{from} \downarrow \ \text{to} \rightarrow & \begin{array}{cccccc} \cdots & -2 & -1 & 0 & 1 & 2 \ \ \cdots \end{array} \\
\hline
\begin{array}{c} \vdots \\ -2 \\ -1 \\ 0 \\ 1 \\ 2 \\ \vdots \end{array} &
\left[\begin{array}{ccccccc}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & p_2 & p_3 & 0 & 0 & 0 & \cdots \\
\cdots & p_1 & p_2 & p_3 & 0 & 0 & \cdots \\
\cdots & 0 & p_1 & p_2 & p_3 & 0 & \cdots \\
\cdots & 0 & 0 & p_1 & p_2 & p_3 & \cdots \\
\cdots & 0 & 0 & 0 & p_1 & p_2 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array} \right]
\end{array}
= \underline{\underline{P}}
$$

This is an example of a <u>band matrix</u>, in which the only non-zero entries are on the main diagonal and 1 diagonal either way from the main diagonal; since there are only 3 non-zero diagonals, $\underline{\underline{P}}$ is said to be <u>tridiagonal</u>.

Moreover, all of the main diagonal entries are the same (p_2); all of the entries 1 diagonal ~~above~~ below are also the same (p_1); and all of the entries 1 diagonal above are also the same (p_3).

Such matrices are called Toeplitz (named after Otto Toeplitz, (1881-1940) (died of tuberculosis at 58), a German mathematician who was fired by the Nazis from his university position in 1935 for being Jewish).

Q: 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: Suppose, for example, that $(p_1, p_2, p_3) = (0.1, 0.3, 0.6)$. Then you would expect the particle to drift off to $+\infty$. Similarly, $(p_1, p_2, p_3) = (0.5, 0.25, 0.25)$ should yield a drift to $-\infty$. $(p_1, p_2, p_3) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$?

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.