ex. $E(X) = 1$, X non-negative $\rightarrow$

$P(X \geq 100) \leq \frac{1}{100}$ The inequality is sharp, meaning that the upper bound $\frac{E(X)}{t}$ on $P(X \geq t)$ is attainable, ⊛

---

ex. $E(X) = 1$, X - nonnegative $\rightarrow$ put probability 0.99 on $X = 0$ and 0.01 on 100

⊛ but most of the time (i.e., for most distributions) it's a crude upper bound.

(30 May 19)

Can apply Markov inequality to the r.v. $Y = (X - E(X))^2$ to get

Chebyshev Inequality → X r.v. with $V(X)$ existing

for every $t \geq 0$,

$$P\left[|X - E(X)| \geq t\right] \leq \frac{V(X)}{t^2}$$

(attributed to Pafnuty Chebyshev (1821 - 1894), also a Russian mathematician, one of whose Ph.D. students was Markov)

Ex. $E(X) = \mu$, $V(X) = \sigma^2$

Chebyshev says $P\left[\left|\frac{X-\mu}{\sigma}\right| \geq 3\right] \leq \frac{1}{3^2} = \frac{1}{9}$,

so no more than $\frac{1}{9} = 11\%$ of the probability in any distribution with finite variance can be more than 3 SDs away from the mean (recall for Normal dist. this prob. is 0.3%)

This upper bound is also sharp, but for most distributions it's (also) crude (as with the Markov bound).

Back to $\bar{X}_n$

$X_i \overset{IID}{\sim}$ some dist. with mean $E(X_i) = \mu$ $(i = 1, \ldots, n)$ and variance $V(X_i) = \sigma^2 < \infty$

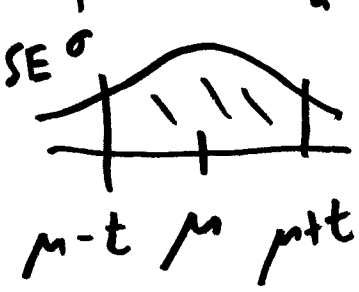
We have already shown that if $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

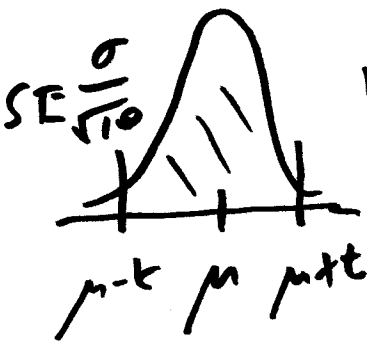
then $E(\bar{X}_n) = \mu$ and $V(\bar{X}_n) = \frac{\sigma^2}{n}$ for all $n = 1, 2, \ldots$

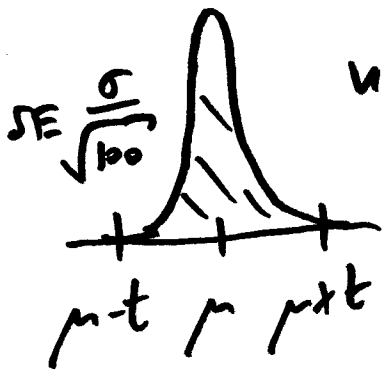
Chebyshev then gives $P(|\bar{X}_n - \mu| \geq t) \leq \frac{\sigma^2}{nt^2}$ for all $t > 0$

This can be rewritten $P(|\bar{X}_n - \mu| < t) \geq 1 - \frac{\sigma^2}{nt^2}$ $\quad \epsilon > 0$, ϵ

PDF of $\bar{X}_n$







⋮

This suggests a way to quantify how close a r.v. like $\bar{X}_n$ is to a constant like μ:

Def. A sequence $Z_1, Z_2, \ldots$ of r.v. is said to converge in probability to a constant b if for all $\epsilon > 0$, $\lim_{n \to \infty} P(|Z_n - b| < \epsilon) = 1$;

this is denoted $Z_n \xrightarrow{P} b$.

This is immediate consequence of Chebyshev & this definition is

(Weak) Law of Large Numbers

$X_i \overset{IID}{\sim}$ a dist. with mean μ and variance $\sigma^2 < \infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

($\bar{X}_n$ is consistent for μ)

$\rightarrow \bar{X}_n \xrightarrow{p} \mu$

This result has a long history: the Italian mathematician Girolamo Cardano (1501-1576) asserted it without proof; Jacob Bernoulli (1655-1705) proved it for $(X_i|\theta) \overset{IID}{\sim}$ Bernoulli(θ) (it took him 20 years to find a correct proof, published posthumously in 1713; Bernoulli thought that this theorem proved the existence of God); Siméon Denis Poisson named it the Law of Large Numbers in 1837.

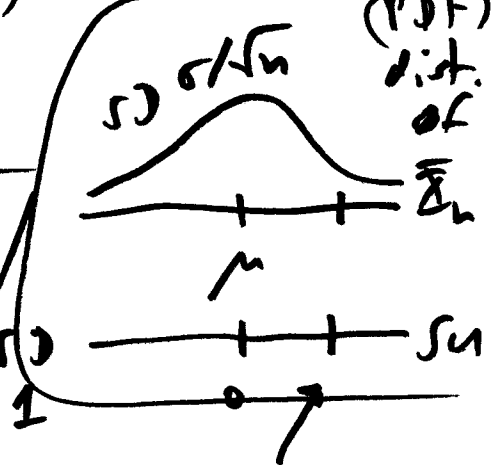
Corollary If $Z_n \xrightarrow{p} b$ and $g(z)$ is continuous at $z = b$ then $g(Z_n) \xrightarrow{p} g(b)$.

Central Limit Theorem (CLT)

Example $X_i \overset{IID}{\sim} N(\mu, \sigma^2)$, $\sigma < \infty$ $(i = 1, \ldots, n)$

we know that $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has mean μ, variance $\frac{\sigma^2}{n}$ and is normally distributed, so that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \quad \text{for all } n = 1, 2, \ldots$$



Q: Does something like this work for other choices of $X_i \overset{IID}{\sim}$?

A: Yes: it's the most famous result in all of probability:

Central Limit Theorem (CLT)

$X_i \overset{IID}{\sim}$ (any) dist. with mean μ and finite variance $0 < \sigma^2 < \infty$,

$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to$ for (large) n ? $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$

Careful statement

Def. $X_1, X_2, \ldots$ a sequence of r.v.; let F_n be the CDF of X_n

→ if there exists a CDF F^* such that $\lim_{n\to\infty} F_n(x) = F^*(x)$ for all x at which $F^*(x)$ is continuous, then people say that $X_n \xrightarrow{D} F^*$ ("X_n converges in distribution to F^*")

CLT $X_i \overset{IID}{\sim}$ (any) dist. with mean μ and variance $0 < \sigma^2 < \infty$, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

→ $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1)$.

the CLT

---

also has a long history: it was

first demonstrated for $X_i \overset{IID}{\sim} \text{Bernoulli}(\theta)$ by the French/British mathematician Abraham de Moivre (1667 - 1754) in 1733; almost forgotten until revived by the French mathematician Pierre-Simon de Laplace (1749 - 1827) in 1812; almost forgotten again until 1901, when the Russian mathematician Aleksandr Lyapunov gave a more general proof; even more general proof provided by JW Lindeberg (Finnish mathematician (1876 - 1932)) and independently by Paul Lévy (French mathematician (1886 - 1971)) in the early 1920s. (26 Aug 19) CLT name due to Hungarian-American mathematician (1887-1985) George Pólya in 1920

Example Contaminated water supply:

X = arsenic concentration

Y = lead concentration (same units) (e.g. ppb, 30)

Interest focuses on $R = \frac{X}{X+Y}$

(proportion of contamination due to lead)

$E(R) = E\left(\frac{X}{X+Y}\right)$ difficult to calculate.

(IID)

Simulation approach Randomly sample n pairs (X_i, Y_i) from the joint PDF of (X, Y), calculate $R_i = \frac{X_i}{X_i + Y_i}$ and

$\bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$ ← good Monte Carlo (simulation) estimate of $E(R)$.

Q: How big does n need to be to achieve a desired accuracy target?

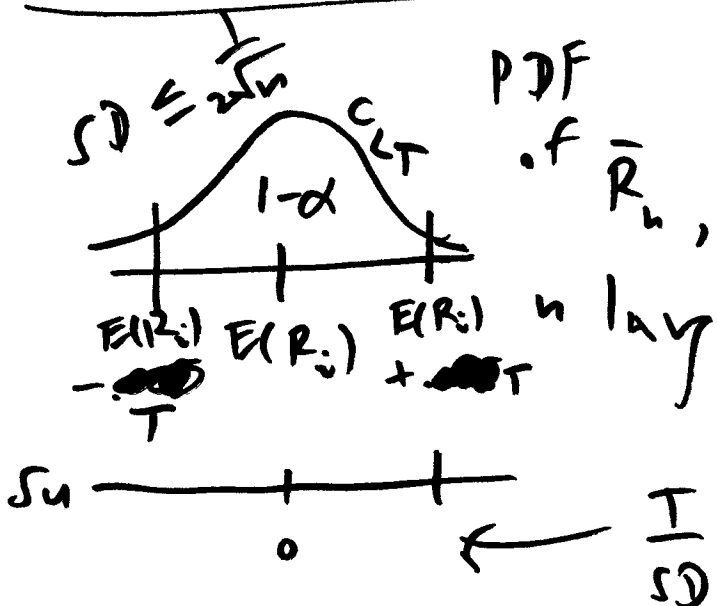
By definition

$$|R_i| = \left|\frac{I_i}{\bar{I}_i + I_i}\right| \leq 1$$; can show that

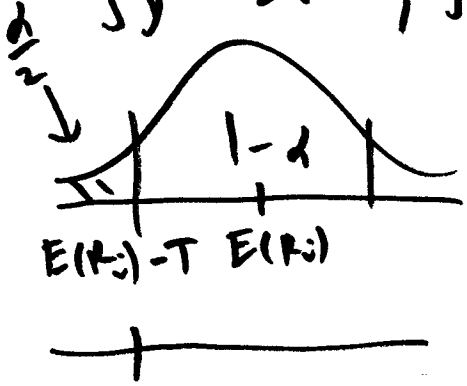
as a result $V(R_i) \leq \frac{1}{4}$.

CLT says that dist. of $\bar{R}_n$ will be close to Normal for large n, with mean $E(R_i)$ and variance $\frac{V(R_i)}{n} \leq \frac{1}{4n}$

Suppose we want $\bar{R}_n$ to differ from $E(R_i)$ by no more than some tolerance T with probability at least $(1-\alpha)$ ...



$SD \le \frac{1}{2\sqrt{n}}$, so $\frac{1}{SD} \ge 2\sqrt{n}$. and



$$\frac{-T}{SD} \le 2T\sqrt{n}$$

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) = \frac{[E(\bar{R}_n) - T] - E(\bar{R}_n)}{SD} = \frac{-T}{SD} \le 2T\sqrt{n}$$

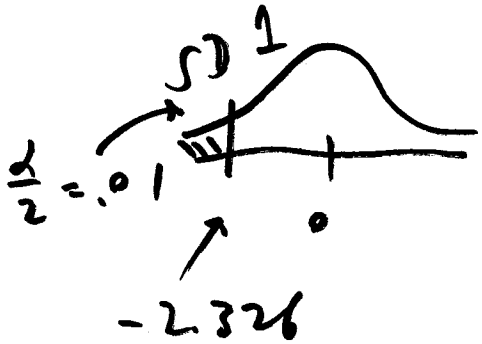
from which $n \ge \left[\frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{2T}\right]^2$

For instance, set $T = 0.005$ ($\frac{1}{2}$ of 1%)

and $\alpha = .02$ to get

$$n \ge \left[\frac{-2.326}{2(.005)}\right]^2 \doteq 54,119$$

simulation replications needed



Case study: Escalators in the London Underground (redacted)

the Delta method

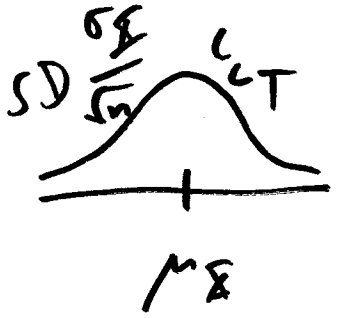
The CLT says that if $X_i \overset{IID}{\sim}$ (any) dist. with finite mean μ_X and finite variance σ_X^2, then

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

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



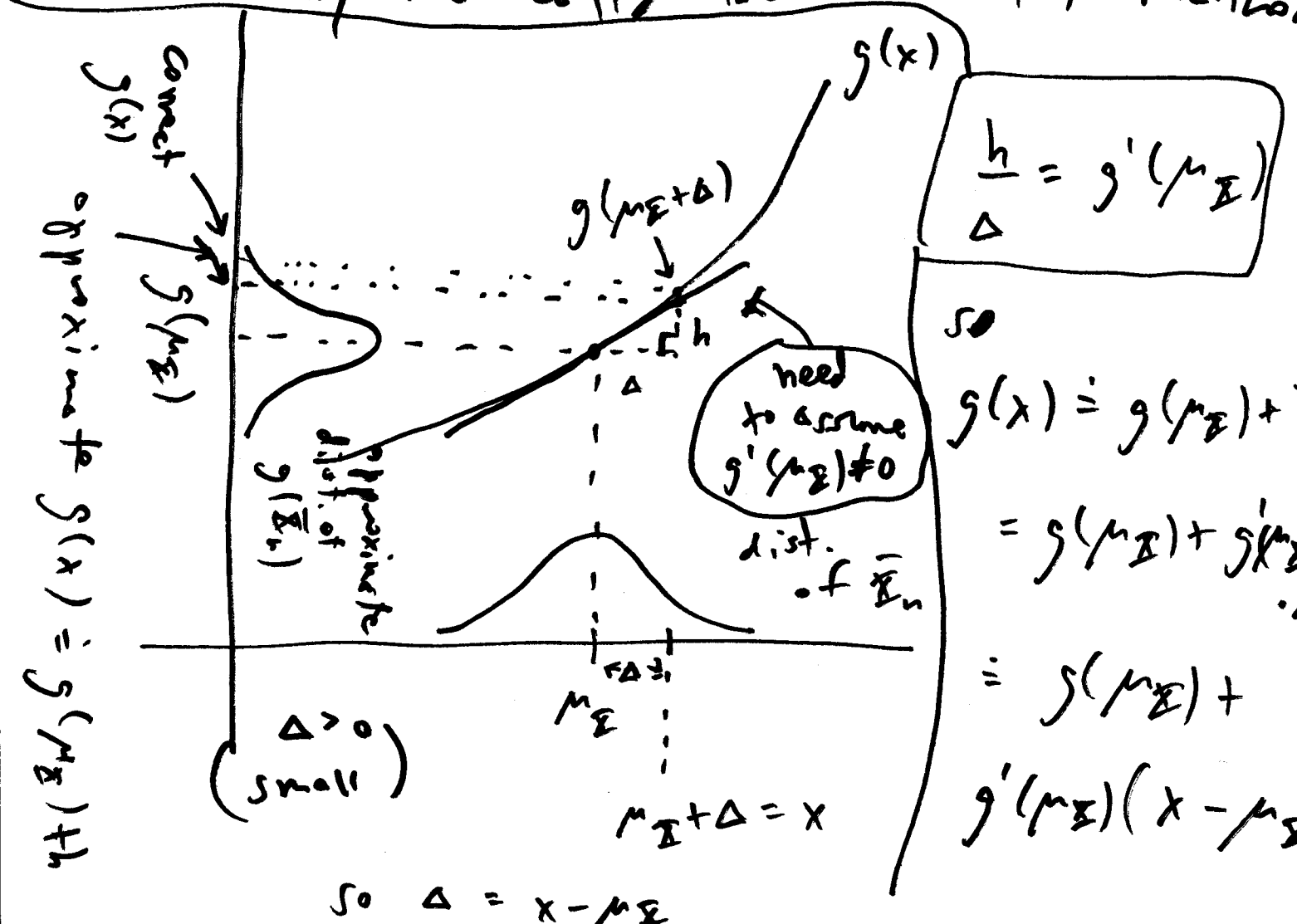
Question: If $g(x)$ is a sufficiently "nice" function, is there a comparable result for $g(\bar{X}_n)$?

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 $x = \mu_X$:

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

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



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

so

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

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

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

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

(constant; r.v.; constant; constant)

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

$$= g(\mu_X) + g'(\mu_X)[E(\bar{X}_n) - \mu_X] \quad (E(\bar{X}_n) \to 0)$$

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

and

(constant; constant; r.v.)

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

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

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

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

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

This works for <u>any</u> 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), $W = g(V)$

$$\rightarrow E(W) \doteq g(\mu_V) \text{ and}$$

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

Δ method (part 1)

provided $g'(v)$ is continuous and

$g'(\mu_V) \neq 0$

Δ method (part 2)

Moreover, if $V \sim$ Normal then $W = g(V) \sim$ Normal also.

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 $\{X_i, i = 1, 2, \ldots\}$ 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.

Complications: seasonal effects (ignored here)

The rate of service is defined to be $g(\mu_X) = \frac{1}{\mu_X}$, which would naturally be estimated by $g(\bar{X}_n) = \frac{1}{\bar{X}_n}$.

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

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

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

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

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

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

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

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

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

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

Specific Calculation

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

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

$= \mu_X$, $= \sigma_X^2$

$(X_i|\lambda)$ has PDF

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

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

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

(discrete or continuous)

Fancy version of Δ-method

$Y_1, Y_2, \ldots$ sequence of 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 $a_n(T_n - \theta) \xrightarrow{D} F^*$,

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

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

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

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

In this context the theorem says that

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

(28 Aug 17)

A little bit more about the continuity correction

Tay-Sachs case study, revisited

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

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

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

$i = 1, \ldots, n = 5$

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

$(i = 1, \ldots, n)$

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

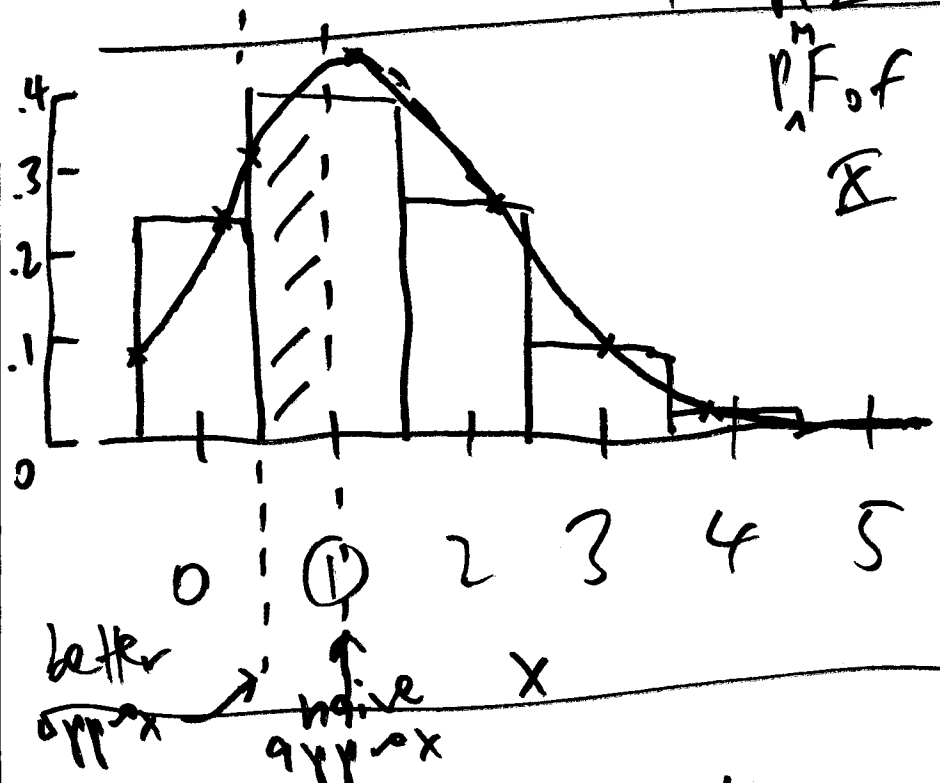
$\mu_X = E(X) = np = 1.25$ and SD

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

On day 1 of this class we worked out that $p(1 \text{ or more T-s landing}) = P(\bar{X} \geq 1)$

$$1 - p(\text{no T-s landing}) = 1 - (1-p)^n \doteq 0.76$$

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



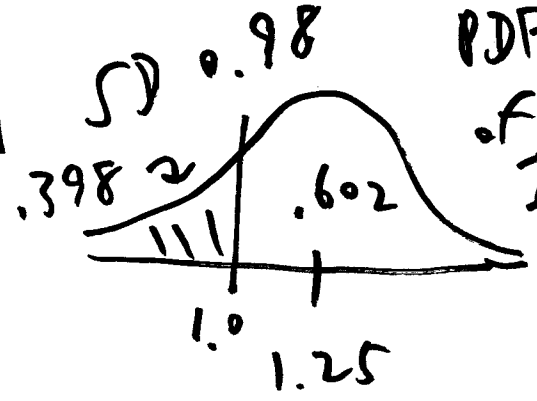
PDF of $\bar{X}$

Naive Normal approximation, from CLT:

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

$$= 1 - 0.398$$

$$\doteq \boxed{0.602}$$ (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 ($\hat{M}$ PF) bars:

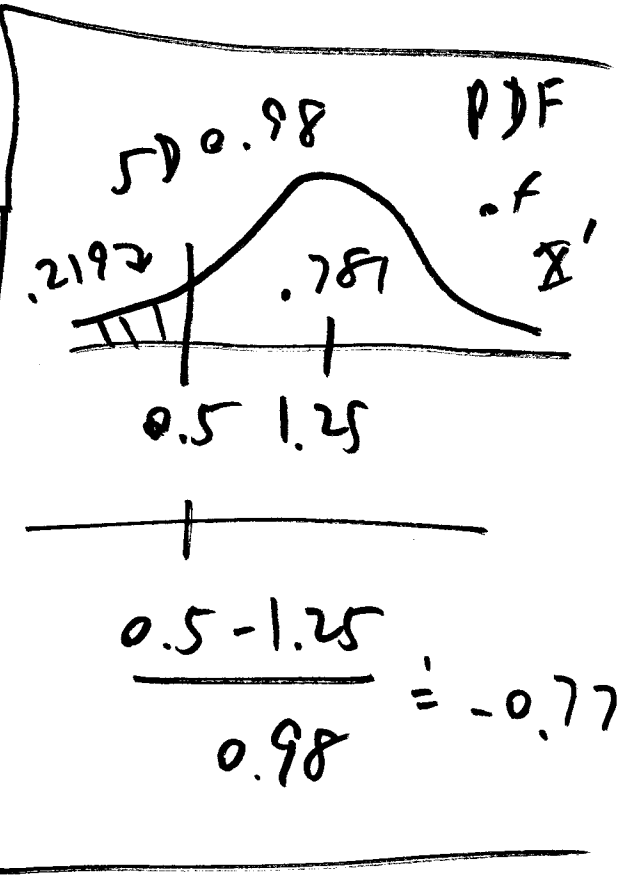
Normal approximation with continuity correction

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

$$\doteq 1 - .219$$

$$\doteq 0.781$$

(correct answer 0.76; much better approx.)



(4 Jan 19)

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_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 $(X_i \mid \theta) \overset{IID}{\sim}$ from some dist. depending on θ.

Q: Does this process have a memory?

Example, revisited

Machine with θ dial from 0 to 1, produces IID Bernoulli(θ) trials X_i.

Recall that 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 $(X_i | \theta) \overset{IID}{\sim}$ 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 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) Markov chain if for $n = 1, 2, \ldots$; (b) any real number; and for all possible sequences of states $x_1, x_2, \ldots$

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

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

In other words, the only thing you need to know to simulate where the Markov chain is going next is where it is now. (25 Aug 19)

(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 (countably or uncountably).

(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 →

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

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

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

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

Def. A Markov chain with a finite state space is called a finite Markov chain.

Def. Suppose you have a finite Markov chain with k possible states numbered $1, \ldots, k$ (k integer ≥ 2) → $\{P(X_{n+1} = j \mid X_n = i), i, j = 1, \ldots, k, n = 1, 2, \ldots\}$ is called the transition distribution 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. If (time-homogeneous) (DS) (bad name)

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.

$$P = \begin{array}{c} \\ 1 \\ 2 \\ \vdots \\ k \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & k \end{array} \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ \vdots & \vdots & & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix} \end{array}$$

$k \times k$

(from) state

(to) state

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

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

matrix versus quaternion

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

~~definition/theorem~~

Example Gene inheritance is Markovian: all we need to know to predict your genetic makeup at birth is the genetic story of your parents

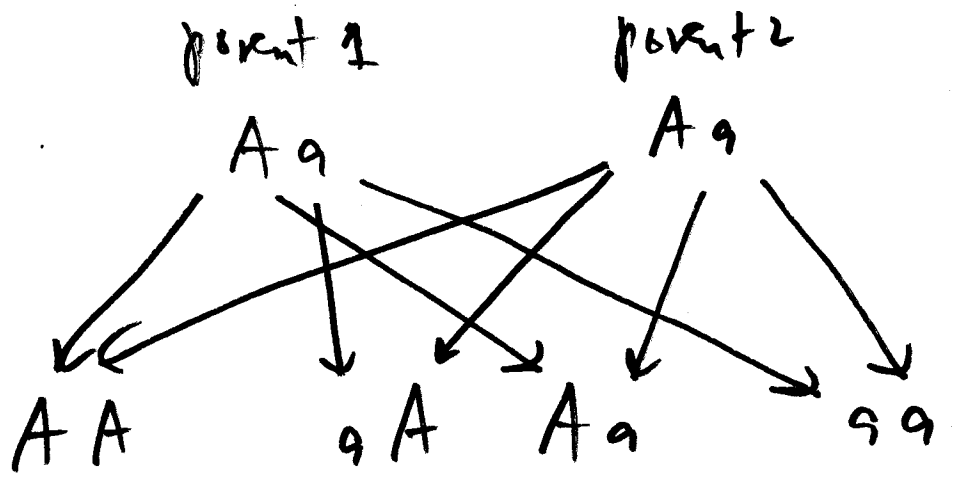
(your grandparents, ..., are irrelevant).

~~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

$$\{\underset{\text{from parent 1}}{\text{allele 1}}, \underset{\text{from parent 1}}{\text{allele 2}}, \underset{\text{from parent 2}}{\text{allele 1}}, \underset{\text{from parent 2}}{\text{allele 2}}\},$$ 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 $\{aa, aa\}$.

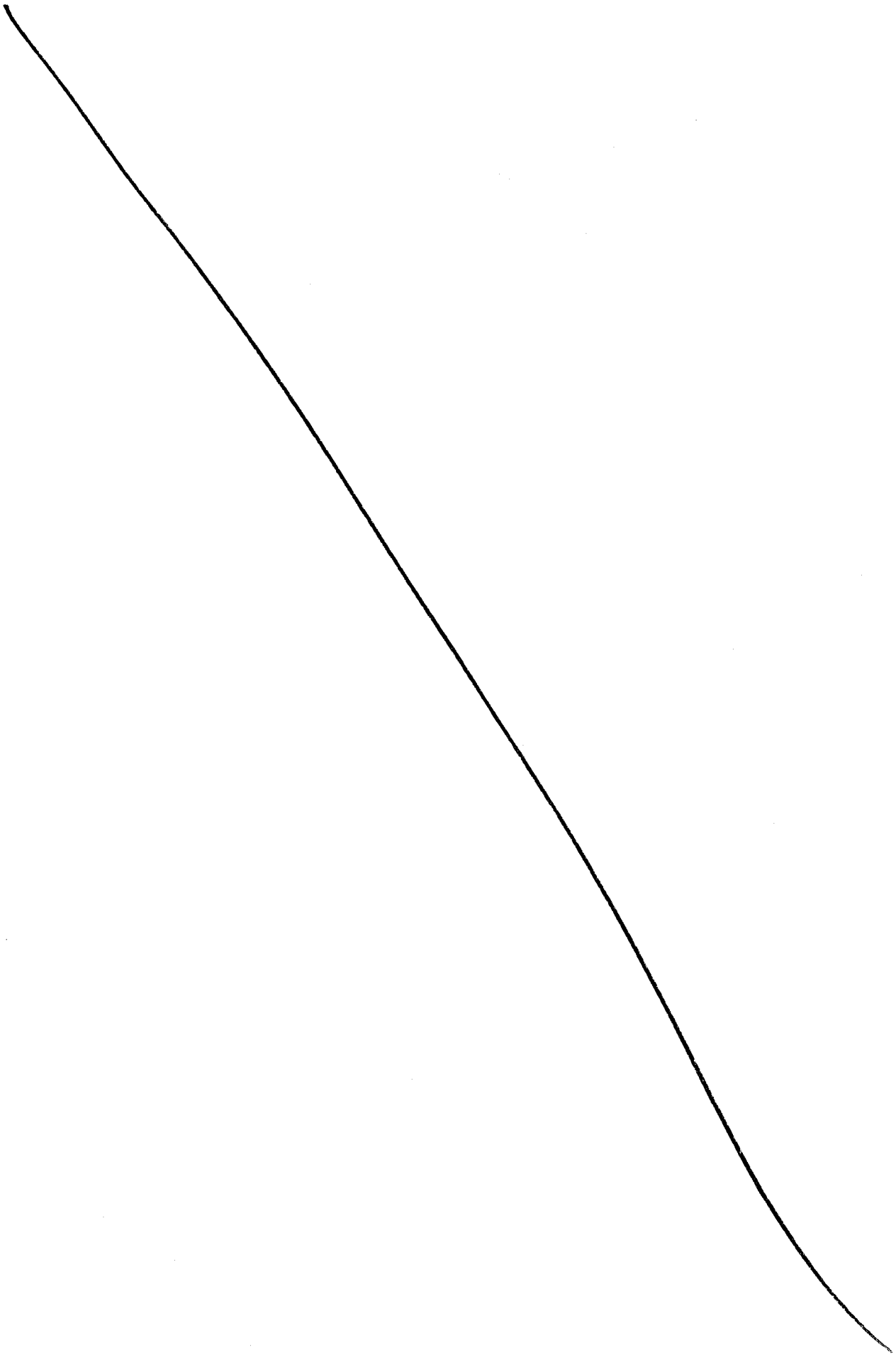


One possible inheritance sequence

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

Transition matrix

from ↓ →	{AA, AA}	{AA, Aa}	{AA, aa}	{Aa, Aa}	{Aa, aa}	{aa, aa}
{AA, AA}	1	0	0	0	0	0
{AA, Aa}	$\frac{1}{4}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0
{AA, aa}	0	0	0	1	0	0
{Aa, Aa}	$\frac{1}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{16}$
{Aa, aa}	0	0	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
{aa, aa}	0	0	0	0	0	1



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

from ↓ / to →	...	-2	-1	0	1	2	...
-2	...	p_2	p_3	0	0	0	...
-1	...	p_1	p_2	p_3	0	0	...
0	...	0	p_1	p_2	p_3	0	...
1	...	0	0	p_1	p_2	p_3	...
2	...	0	0	0	p_1	p_2	...
⋮	⋱	⋮	⋮	⋮	⋮	⋮	⋱

$= \underline{\underline{P}}$

This is an example of a band matrix, 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 tridiagonal.

Moreover, all of the main diagonal entries are the same (p_2); 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_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 $-\infty$. $(p_1, p_2, p_3) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$?

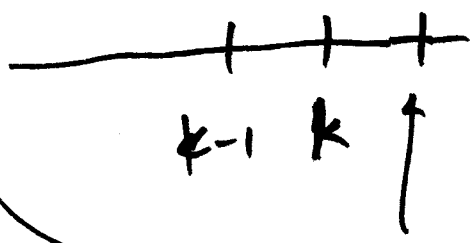
Can show that as $n \to \infty$ every integer is visited infinitely many times, and the expected time you must wait for the chain to return to 0 (having started there) is also infinite.

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

Bounded random walk

Restrict the Markov chain to $S = \{-k, -k+1, \ldots, -1, 0, 1, \ldots, k-1, k\}$ for some integer $k \geq 1$.

Q: what to do at the boundaries?



one idea would be to wrap around: if you try to move to $(k+1)$, interpret that as a move to $-k$; if you try to move to $-(k+1)$, move to $+k$.

(move.2)

Transition matrix with $k=2$

from \ to	-2	-1	0	1	2
-2	p_2	p_3	0	0	p_1
-1	p_1	p_2	p_3	0	0
0	0	p_1	p_2	p_3	0	(M[k,k])
1	0	0	p_1	p_2	p_3
2	p_3	0	0	p_1	p_2

(move.2)

R demo

Another idea: keep trying until you make a legal move

(move.1)

Back to a general finite Markov chain

Let $p_{ij}^{(m)} = P($ chain moves from i to j in m steps $)$

Theorem $= P(X_{n+m} = j \mid X_n = i)$

Finite Markov chain with stationary transition distributions & transition matrix $\underline{\underline{P}} \rightarrow p_{ij}^{(m)}$ is just the (i,j) entry of the matrix $\underline{\underline{P}}^m$ which is called the m-step transition matrix of the Markov chain.

Genetic example, continued

$\{AA, Aa\}$ has the property that once the chain is in that state, it can't

to anywhere else; so does {99, 99}

This occurs for a state i when $p_{ii} = 1$.

Def Any state with $p_{ii} = 1$ is called an absorbing state. Notice that in this genetic Markov chain, states 1, 2, 4 all have positive probability of moving to state 1 in 2 steps, and the same is true of moving to state 6 in 2 steps. It follows that, if the chain is run long enough (simulating many generations), it will either end up

in state {AA, AA} or in state {aa, aa}. (341)

Markov chains that settle down to a single stable long-run distribution are especially important in contemporary Bayesian computation; the long-run stable distribution is called the equilibrium distribution of the chain.

Important note on terminology

JS call this distribution the stationary dist. of the chain, but this choice is unfortunate because they've already used stationary to mean something else:

DS: If $P(X_{n+1} = j \mid X_n = i)$ is the same for all n, DS say that the transition distribution is stationary; other people call this time-homogeneous.

I'll use equilibrium distribution for the long-run behavior of Markov chains that settle down into a stable long-run story.

Where should the Markov chain start?

You can either initialize a Markov chain to a deterministic value, or you can start it off by making a random draw from what's called the initial distribution of the Markov chain:

Def Any vector $\underset{\sim}{v}$ of non-negative numbers that add up to 1 is called a probability vector; any such vector whose components specify that a Markov chain will be in each possible state at time 1 is referred to as the initial distribution of the chain.

So: After 1 timestep (iteration), the probability dist. over the Markov chain's possible states is $\underset{\sim}{v}$; after 2 iterations the chain's dist. is $\underset{\sim}{v}\underset{=}{P}$; after $(m+1)$ iterations its dist. is $\underset{\sim}{v}\underset{=}{P}^m$; it would be nice if $\underset{\sim}{v}\underset{=}{P}^m$ converged to a unique dist. as $m \to \infty$: this would

be its equilibrium distribution. (344)

Notice something interesting: if we choose $\underset{\sim}{v}$ so that $\underset{\sim}{v}\underline{\underline{P}} = \underset{\sim}{v}$, then

$\underset{\sim}{v}\underline{\underline{P}}^2 = (\underset{\sim}{v}\underline{\underline{P}})\underline{\underline{P}} = \underset{\sim}{v}\underline{\underline{P}} = \underset{\sim}{v}$, $\underset{\sim}{v}\underline{\underline{P}}^3 = (\underset{\sim}{v}\underline{\underline{P}}^2)\underline{\underline{P}}$

$= \underset{\sim}{v}\underline{\underline{P}} = \underset{\sim}{v}$; and so $\lim_{m\to\infty} \underset{\sim}{v}\underline{\underline{P}}^m = \underset{\sim}{v}$

Def. Markov chain with transition matrix $\underline{\underline{P}}$ → any probability vector $\underset{\sim}{v}$ such that $\underset{\sim}{v}\underline{\underline{P}} = \underset{\sim}{v}$ is (an) equilibrium dist. for the Markov chain

under additional conditions on $\underline{\underline{P}}$, such an equilibrium dist. will be unique (we won't fully pursue that here).

How find $\underset{\sim}{v}$ so that $\underset{\sim}{v}\,\underline{\underline{P}} = \underset{\sim}{v}$?

In linear algebra this is an example of an eigenvalue/eigenvector problem:

Def.] Given a square matrix $\underset{k\times k}{\underline{\underline{P}}}$,

any vector $\underset{k\ 1}{\underset{\sim}{v}_R}$ satisfying $\underset{k\times k}{\underline{\underline{P}}}\ \underset{\sim 1}{v_R} = \lambda_R \underset{k\ 1}{\underset{\sim}{v}_R}$

is called a right eigenvector of $\underline{\underline{P}}$ with

(right) eigenvalue λ_R, and any vector $\underset{1\ k}{\underset{\sim}{v}_L}$

satisfying $\underset{1\ k}{\underset{\sim}{v}_L}\ \underset{k\times k}{\underline{\underline{P}}} = \lambda_L \underset{1\ k}{\underset{\sim}{v}_L}$ is called

a left eigenvector of $\underline{\underline{P}}$ with (left)

eigenvalue λ_L.

So, given a transition matrix $\underset{=}{P}_{k=k}$ for a Markov chain, an equilibrium dist. for the chain can be found by computing the left eigenvector $\underset{1\sim L}{V}_{k}$ where eigenvalue is 1, if such a vector exists. Most computer routines for eigenanalysis only give you right eigenvectors, but notice that if

$$\underset{1\sim L}{V}_{k}\ \underset{=}{P}_{k=k} = \underset{1\sim L}{V}_{k}$$

then $\left(\underset{1\sim L}{V}_{k}\ \underset{=}{P}_{k=k}\right)^{T} = \underset{k\sim 1}{V}^{T}_{L}$ ← transpose

$$= \underset{k=k}{P}^{T}\ \underset{k\sim 1}{V}^{T}_{L} = \underset{k\sim 1}{V}^{T}_{L}$$

so we can just eigendecompose $\underset{=}{P}^{T}$ instead of $\underset{=}{P}$.

$\underset{\sim L}{V}\ \underset{=}{P} = \underset{\sim L}{V}$

Genetic example, continued

R's routine eigen gives the following results: P^T has two eigenvectors whose eigenvalues are 1: $\begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$, corresponding to the two absorbing states.

This suggests that there's an entire family of equilibrium distributions of the form $(p, 0, 0, 0, 0, 1-p)^T$ for $0 \le p \le 1$; and [Wa] can be used to verify this conjecture.

So the earlier guess is also correct: after many generations either one of {AA, AA} or {aa, aa} will be absorbing.

There is a special case in which a unique stationary distribution exists.

Theorem If you can find a positive integer $m \geq 1$ such that every element of $\underset{=}{P}^m$ is strictly positive, then $\lim_{n \to \infty} \underset{=}{P}^n$ is a matrix with all rows equal to the unique stationary dist $\underset{\sim}{v}$, and no matter what the chain's initial distribution is, its distribution after n steps converges to $\underset{\sim}{v}$ as $n \to \infty$.