**Stat 100a: Introduction to Probability.**

Outline for the day

1. CLT.
2. CIs.
3. Sample size calculations.
4. Random walks.
5. Reflection principle.
6. Ballot theorem.
7. Avoiding zero.
8. Chip proportions and induction.
9. Doubling up.
10. Examples.

hw3 is due Tue Jul28.

The computer project is due on Sun Jul26 8:00pm.

http://www.stat.ucla.edu/~frederic/100A/S20 .

Mean on exam2 was 79%, SD was 30%. Read directions. Cheating?

# 1. Central Limit Theorem (CLT), ch 7.4.

Sample mean $\overline{X_n} = \sum X_i / n$

iid: independent and identically distributed.

Suppose $X_1$, $X_2$ , etc. are iid with expected value $\mu$ and sd $\sigma$ ,

LAW OF LARGE NUMBERS (LLN):

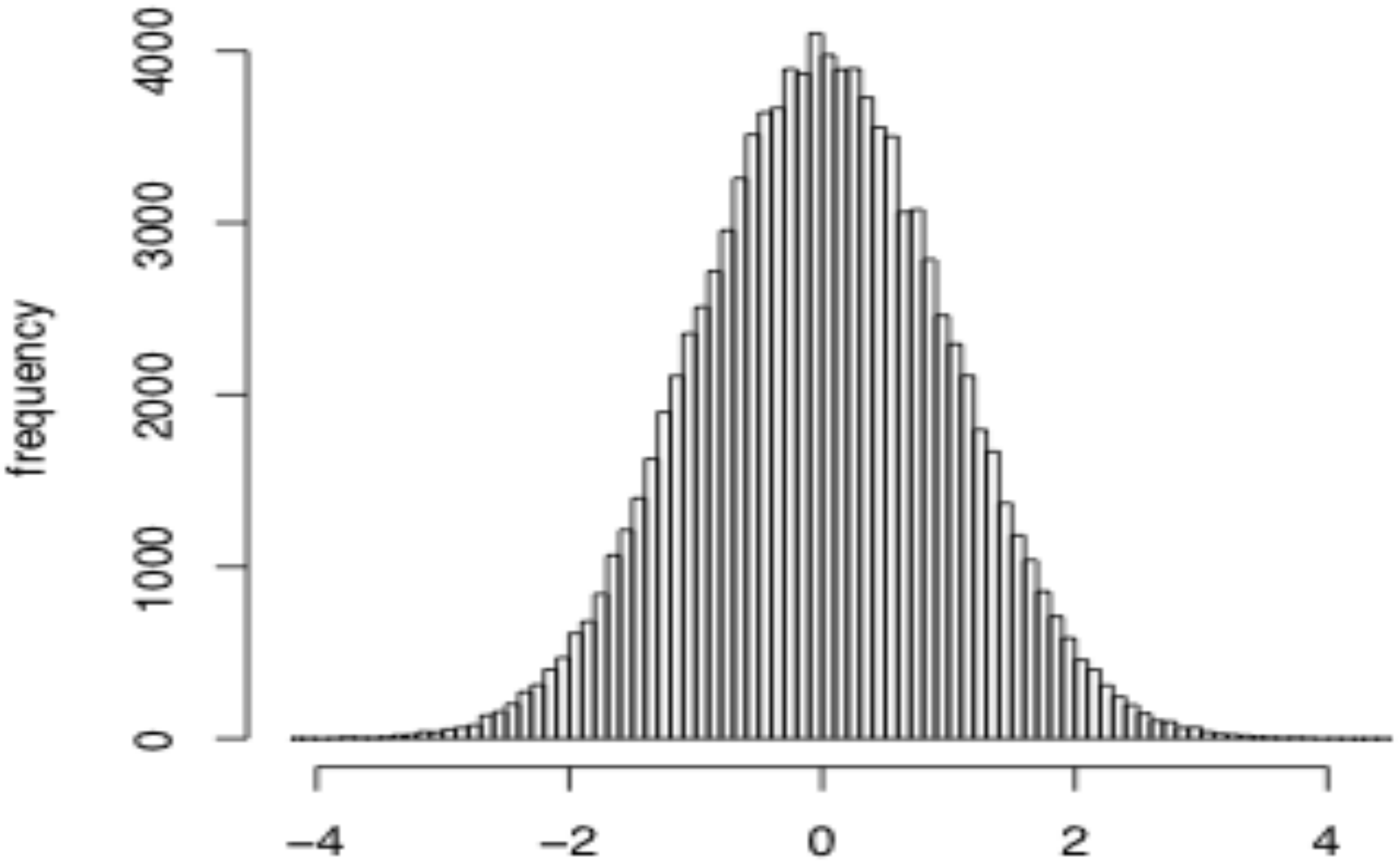
$\overline{X_n} \text{--->} \mu$ .

CENTRAL LIMIT THEOREM (CLT):

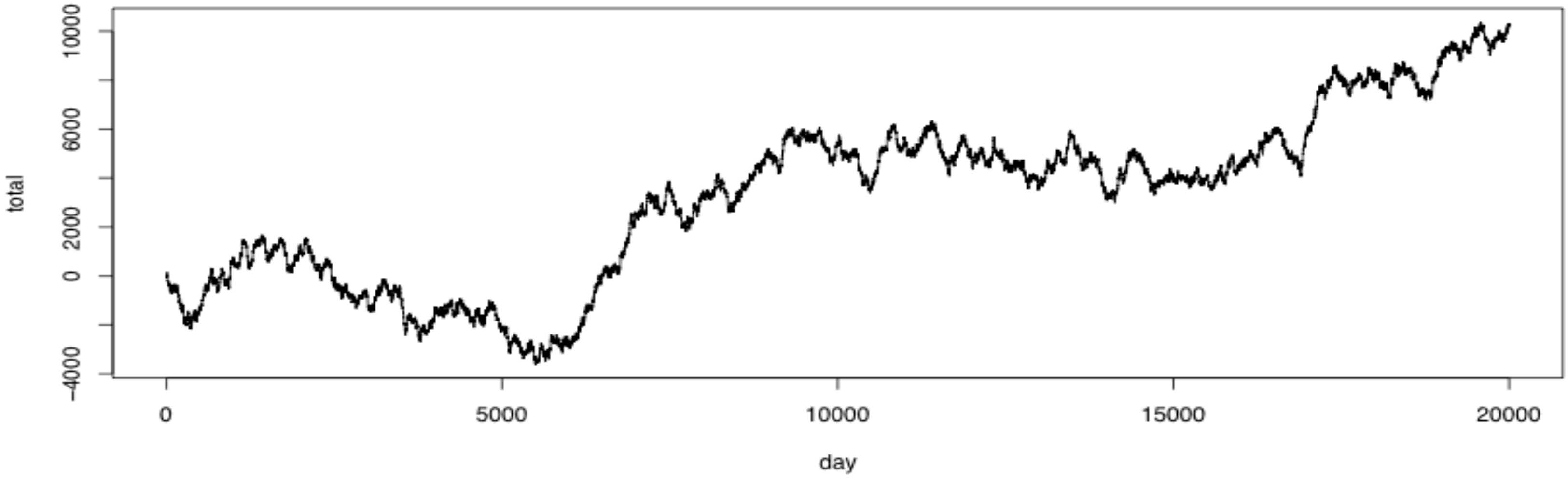
$(\overline{X_n} - \mu) \div (\sigma/\sqrt{n}) \text{--->}$ Standard Normal.

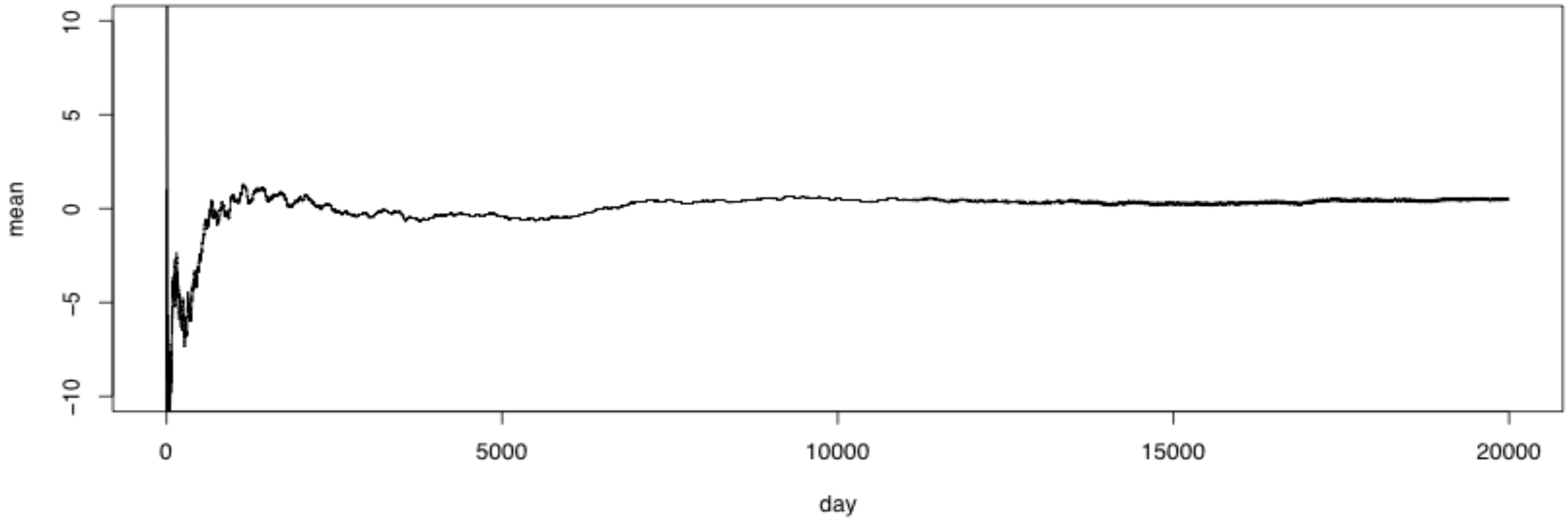
Useful for tracking results.

# 95% between -1.96 and 1.96

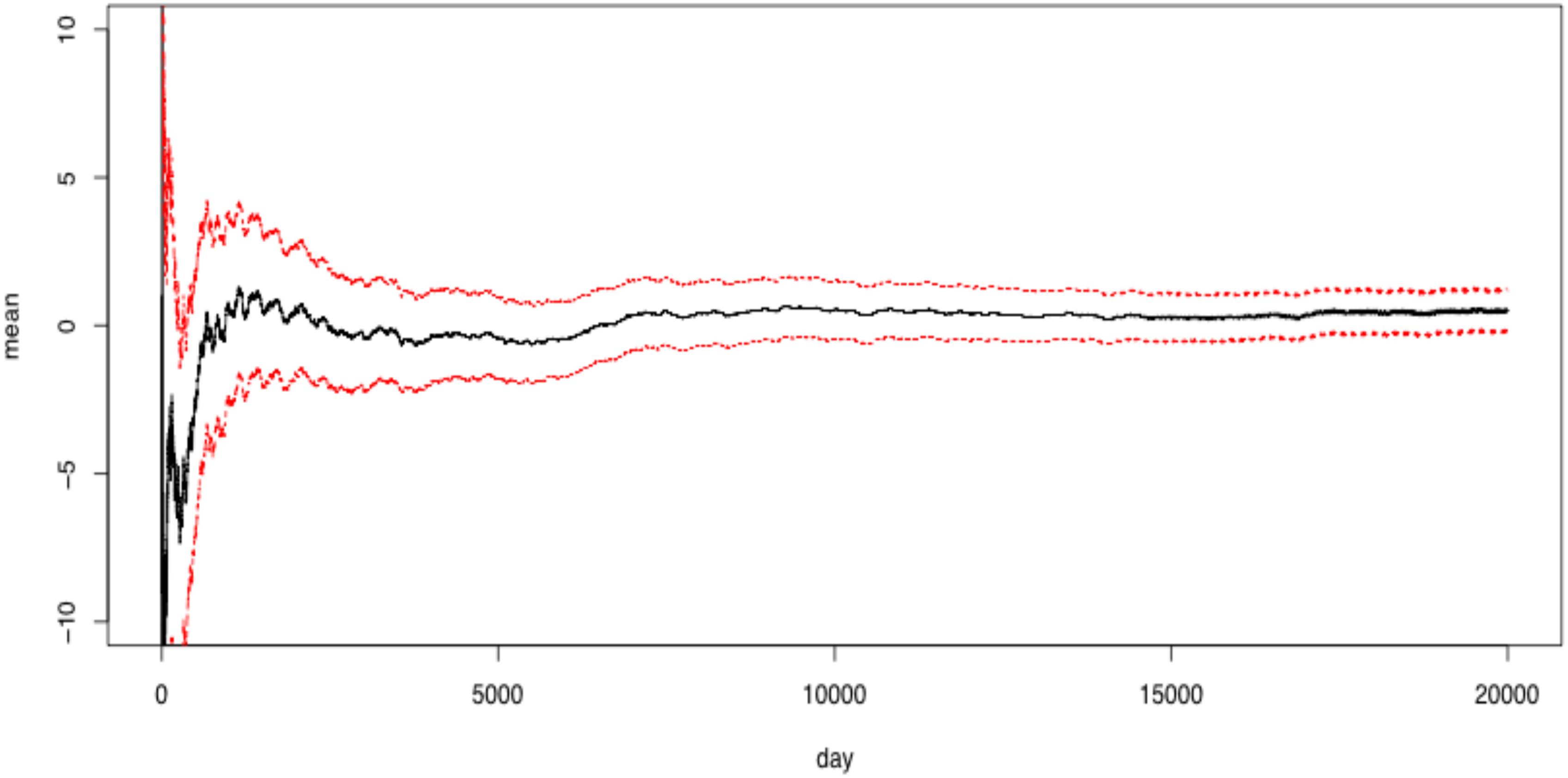


# Truth: -49 to 51, exp. value $\mu = 1.0$





Truth: uniform on -49 to 51. $\mu = 1.0$
Estimated using $\overline{X_n}$ +/- 1.96 $\sigma/\sqrt{n}$
= .95 +/- 0.28 in this example



Central Limit Theorem (CLT): if $X_1$, $X_2$ …, $X_n$ are iid with mean $\mu$ & SD $\sigma$, then

$(\overline{X_n} - \mu) \div (\sigma/\sqrt{n})$ ---> Standard Normal. (mean 0, SD 1).

In other words, $\overline{X_n}$ has mean $\mu$ and a standard deviation of $\sigma \div \sqrt{n}$.

Two interesting things about this:

(i) As n --> ∞, $\overline{X_n}$ --> ***normal.*** Even if $X_i$ are far from normal.

e.g. *average* number of pairs per hand, out of n hands. $X_i$ are 0-1 (Bernoulli).

$\mu$ = p = P(pair) = 3/51 = 5.88%. $\sigma = \sqrt{(pq)} = \sqrt{(5.88\% \times 94.12\%)}$ = 23.525%.

(ii) We can use this to find **a range** where $\overline{X_n}$ is likely to be.

About 95% of the time, a std normal random variable is within -1.96 to +1.96.

So 95% of the time, $(\overline{X_n} - \mu) \div (\sigma/\sqrt{n})$ is within -1.96 to +1.96.

So 95% of the time, $(\overline{X_n} - \mu)$ is within $-1.96\,(\sigma/\sqrt{n})$ to $+1.96\,(\sigma/\sqrt{n})$.

So 95% of the time, $\overline{X_n}$ is within $\mu - 1.96\,(\sigma/\sqrt{n})$ to $\mu + 1.96\,(\sigma/\sqrt{n})$.

**That is, 95% of the time, $\overline{X_n}$ is in the interval $\mu$ +/- 1.96 $(\sigma/\sqrt{n})$.**

**= 5.88% +/- 1.96(23.525%/$\sqrt{n}$). For n = 1000, this is 5.88% +/- 1.458%.**

**For n = 1,000,000 get 5.88% +/- 0.0461%.**

# Another CLT Example

Central Limit Theorem (CLT): if $X_1$ , $X_2$ …, $X_n$ are iid with mean $\mu$ & SD $\sigma$, then

$$(\overline{X_n} - \mu) \div (\sigma/\sqrt{n}) \text{ ---> Standard Normal. (mean 0, SD 1).}$$

In other words, $\overline{X_n}$ is like a draw from a normal distribution

with mean $\mu$ and standard deviation of $\sigma \div \sqrt{n}$.

That is, 95% of the time, $\overline{X_n}$ is in the interval $\mu$ +/- 1.96 $(\sigma/\sqrt{n})$.

Q. Suppose you average \$5 profit per hour, with a SD of \$60 per hour. If you play 1600 hours, let Y be your average profit over those 1600 hours. Find a range where Y is 95% likely to fall.

**A.** We want $\mu$ +/- 1.96 $(\sigma/\sqrt{n})$, where $\mu$ = \$5, $\sigma$ = \$60, and n=1600. So the answer is

\$5 +/- 1.96 x \$60 / $\sqrt{(1600)}$

= \$5 +/- \$2.94, or the range [\$2.06, \$7.94].

## 2. Confidence Intervals (CIs) for $\mu$, ch 7.5.

Central Limit Theorem (CLT): if $X_1$ , $X_2$ …, $X_n$ are iid with mean $\mu$ & SD $\sigma$, then

$$(\overline{X}_n - \mu) \div (\sigma/\sqrt{n}) \text{ ---> Standard Normal. (mean 0, SD 1).}$$

So, 95% of the time, $\overline{X}_n$ is in the interval $\mu$ +/- $1.96 (\sigma/\sqrt{n})$.

Typically you know $\overline{X}_n$ but not $\mu$. Turning the blue statement above around a bit means that 95% of the time, $\mu$ is in the interval $\overline{X}_n$ +/- $1.96 (\sigma/\sqrt{n})$.

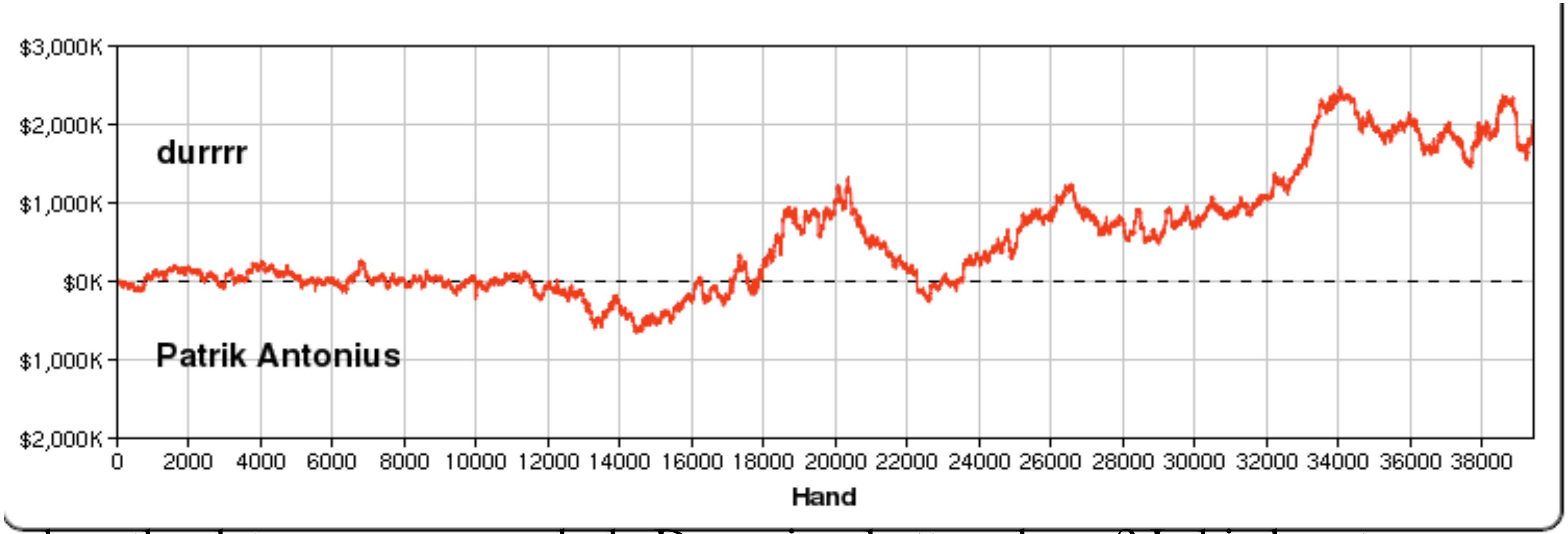
This range $\overline{X}_n$ +/- $1.96 (\sigma/\sqrt{n})$ is called a 95% confidence interval (CI) for $\mu$.

[Usually you don't know $\sigma$ and have to estimate it using the sample std deviation, s, of your data, and ( $\overline{X}_n - \mu$) $\div$ ($s/\sqrt{n}$) has a $t_{n-1}$ distribution if the $X_i$ are normal. For n>30, $t_{n-1}$ is so similar to normal though.]

$1.96 (\sigma/\sqrt{n})$ is called the *margin of error.*

The range $\overline{X}_n$+/- 1.96 ($\sigma/\sqrt{n}$) is a 95% confidence interval for $\mu$. 1.96 ($\sigma/\sqrt{n}$)

(from fulltiltpoker.com:)



Based on the data, can we conclude Dwan is a better player? Is his longterm avg. $\mu > 0$?

Over these 39,000 hands, Dwan profited $2 million. $51/hand. sd ~ $10,000.

95% CI for $\mu$ is $51 +/- 1.96 ($10,000 / √39,000) = $51 +/- $99 = (-$48, $150).

Results are inconclusive, even after 39,000 hands!

**3. Sample size calculation.** How many *more* hands are needed?

If Dwan keeps winning $51/hand, then we want n so that the margin of error = $51.

1.96 ($\sigma/\sqrt{n}$) = $51 means 1.96 ($10,000) / √n = $51, so n = [(1.96)($10,000)/($51)]$^2$ ~ 148,000, so about 109,000 *more* hands.

## **4. Random walks,** ch. 7.6**.**

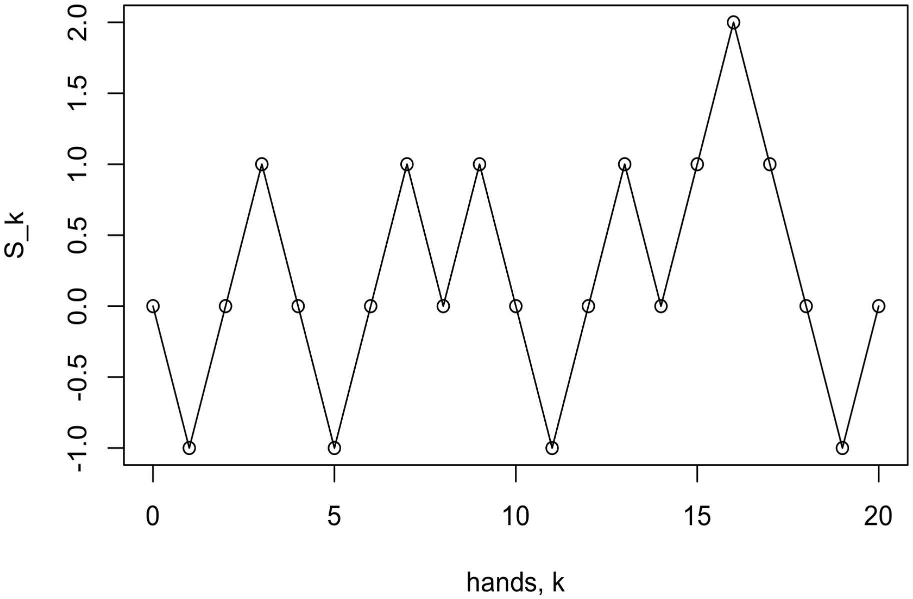
Suppose that $X_1, X_2, \ldots$, are iid,

and $S_k = X_0 + X_1 + \ldots + X_k$ for $k = 0, 1, 2, \ldots$.

The totals $\{S_0, S_1, S_2, \ldots\}$ form a *random walk*.

The classical *(simple)* case is when each $X_i$ is

1 or -1 with probability ½ each.



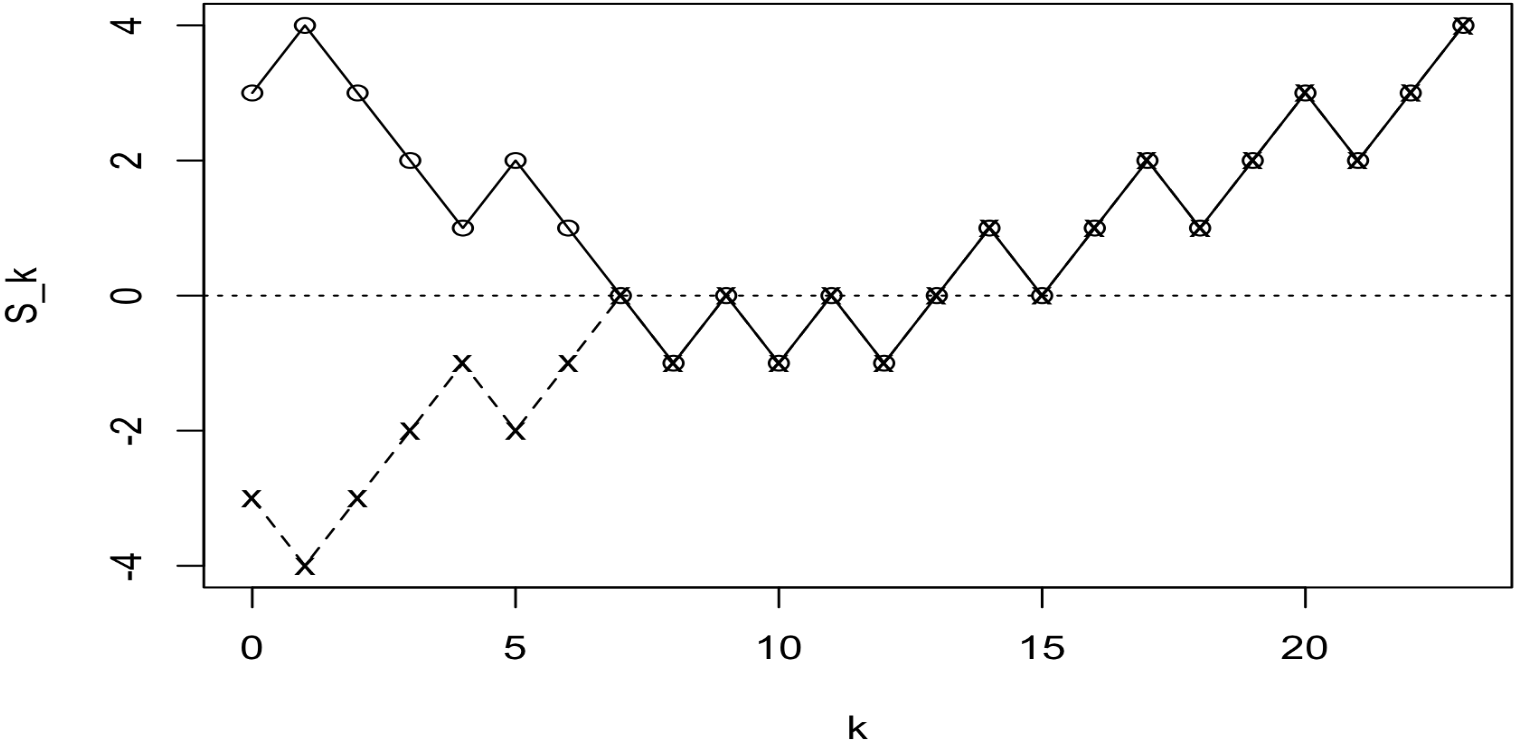
* *Reflection principle:* The number of paths from $(0,X_0)$ to $(n,y)$ that touch the x-axis = the number of paths from $(0,-X_0)$ to $(n,y)$, for any n,y, and $X_0 > 0$.

* *Ballot theorem:* In n = a+b hands, if player A won a hands and B won b hands, where a>b, and if the hands are aired in random order, P(A won more hands than B *throughout* the telecast) = (a-b)/n.

[In an election, if candidate X gets x votes, and candidate Y gets y votes, where x > y, then the probability that X always leads Y throughout the counting is (x-y) / (x+y).]

* For a simple random walk, $P(S_1 \neq 0, S_2 \neq 0, \ldots, S_n \neq 0) = P(S_n = 0)$, for any even n.

**5. Reflection Principle.** The number of paths from $(0,X_0)$ to (n,y) that touch the x-axis = the number of paths from $(0,-X_0)$ to (n,y), for any n,y, and $X_0 > 0$.



For each path from $(0,X_0)$ to (n,y) that touches the x-axis, you can reflect the first part til it touches the x-axis, to find a path from $(0,-X_0)$ to (n,y), and vice versa.

Total number of paths from $(0,-X_0)$ to (n,y) is easy to count: it's just C(n,a), where you go up *a* times and down *b* times.

[For example, to go from (0,-10) to (100, 20), you have to "profit" 30, so you go up a=65 times and down b=35 times, and the number of paths is C(100,65).

In general, $a-b = y - (-X_0) = y + X_0$. $a+b=n$, so $b = n-a$, $2a-n=y+X_0$, $a=(n+y+X_0)/2$].

**6. Ballot theorem.** In n = a+b hands, if player A won a hands and B won b hands, where a>b, and if the hands are aired in random order,

then P(A won more hands than B *throughout* the telecast) = (a-b)/n.

Proof: We know that, after n = a+b hands, the total difference in hands won is a-b.

Let x = a-b.

We want to count the number of paths from (1,1) to (n,x) that do not touch the x-axis.

By the reflection principle, the number of paths from (1,1) to (n,x) that **do** touch the x-axis equals the total number of paths from (1,-1) to (n,x).

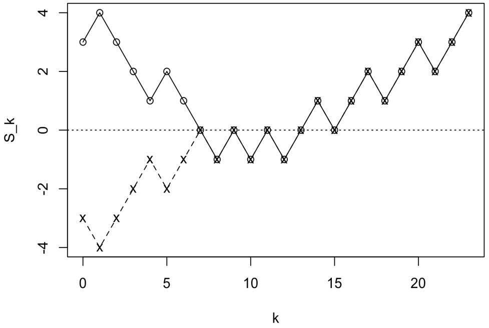
So the number of paths from (1,1) to (n,x) that **do not** touch the x-axis equals the number of paths from (1,1) to (n,x) minus the number of paths from (1,-1) to (n,x)

$= C(n-1,a-1) – C(n-1,a)$

$= (n-1)! / [(a-1)! (n-a)!] – (n-1)! / [a! (n-a-1)!]$

$= \{n! / [a! (n-a)!]\} [(a/n) - (n-a)/n]$

$= C(n,a) (a-b)/n.$



And each path is equally likely, and has probability 1/C(n,a).

So, P(going from (0,0) to (n,x) without touching the x-axis = (a-b)/n.

## 7. Avoiding zero.

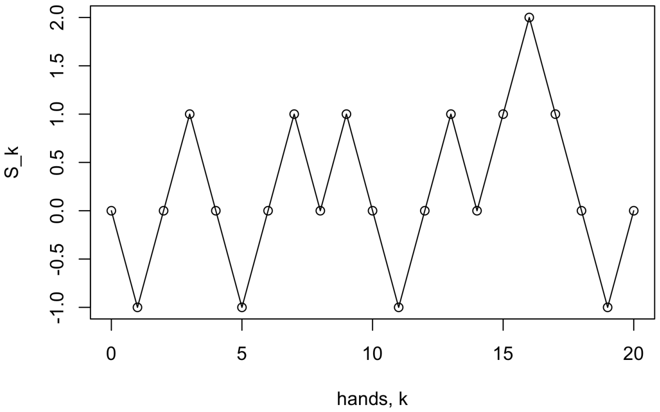
For a simple random walk, for any even # n, $P(S_1 \neq 0, S_2 \neq 0, \ldots, S_n \neq 0) = P(S_n = 0)$.

Proof. The number of paths from (0,0) to (n, j) that don't touch the x-axis at positive times

= the number of paths from (1,1) to (n,j) that don't touch the x-axis at positive times

= paths from (1,1) to (n,j) - paths from (1,-1) to (n,j) by the *reflection principle*

$= N_{n-1,j-1} - N_{n-1,j+1}$.



Let $Q_{n,j} = P(S_n = j)$. By the logic above,

$P(S_1 > 0, S_2 > 0, \ldots, S_{n-1} > 0, S_n = j) = ½[Q_{n-1,j-1} - Q_{n-1,j+1}]$.

Summing from j = 2 to ∞,

$P(S_1 > 0, S_2 > 0, \ldots, S_{n-1} > 0, S_n > 0)$

$= ½[Q_{n-1,1} - Q_{n-1,3}] + ½[Q_{n-1,3} - Q_{n-1,5}] + ½[Q_{n-1,5} - Q_{n-1,7}] + \ldots$ and these terms are eventually 0

$= (1/2)\, Q_{n-1,1}$.

Now note that $Q_{n-1,1} = P(S_n = 0)$, because to end up at (n, 0), you have to be at (n-1,1) and then go down, or at (n-1,-1) and then go up. So $P(S_n = 0) = (1/2)\, Q_{n-1,1} + (1/2)\, Q_{n-1,-1} = Q_{n-1,1}$.

Thus $P(S_1 > 0, S_2 > 0, \ldots, S_{n-1} > 0, S_n > 0) = ½\, P(S_n = 0)$. By the same arguments,

$P(S_1 < 0, S_2 < 0, \ldots, S_{n-1} < 0, S_n < 0) = 1/2\, P(S_n = 0)$.

So, $P(S_1 \neq 0, S_2 \neq 0, \ldots, S_n \neq 0) = P(S_n = 0)$.

**8. Chip proportions and induction**, Theorem 7.6.6.

P(win a tournament) is proportional to your number of chips.

Simplified scenario. Suppose you either go up or down 1 each hand, with prob. 1/2.

Suppose there are n chips, and you have k of them.

Let $p_k$ = P(win tournament given k chips) = P(random walk goes k -> n before hitting 0).

Now, clearly $p_0 = 0$. Consider $p_1$. From 1, you will either go to 0 or 2.

So, $p_1 = 1/2\ p_0 + 1/2\ p_2 = 1/2\ p_2$. That is, $p_2 = 2\ p_1$.

We have shown that $p_j = j\ p_1$, for j = 0, 1, and 2.

***(induction:)* Suppose that, for j = 0, 1, 2, …, m, $p_j = j\ p_1$.**

**We will show that $p_{m+1} = (m+1)\ p_1$.**

**Therefore, $p_j = j\ p_1$ for all j.**

That is, P(win the tournament) is prop. to your number of chips.

$p_m = 1/2\ p_{m-1} + 1/2\ p_{m+1}$. If $p_j = j\ p_1$ for $j \leq m$, then we have

$mp_1 = 1/2\ (m-1)p_1 + 1/2\ p_{m+1}$,

so $p_{m+1} = 2mp_1 - (m-1)\ p_1 = (m+1)p_1$.

**9. Doubling up.** Again, P(winning) = your proportion of chips.

Theorem 7.6.7, p152, describes another simplified scenario.

Suppose you either double each hand you play, or go to zero, each with probability 1/2.

Again, P(win a tournament) is prop. to your number of chips.

Again, $p_0 = 0$, and $p_1 = 1/2\ p_2 = 1/2\ p_2$, so again, $p_2 = 2\ p_1$.

We have shown that, for j = 0, 1, and 2, $p_j = j\ p_1$.

***(induction:)*** **Suppose that, for $j \leq m$, $p_j = j\ p_1$.**

**We will show that $p_{2m} = (2m)\ p_1$.**

**Therefore, $p_j = j\ p_1$ for all $j = 2^k$.** That is, P(win the tournament) is prop. to # of chips.

This time, $p_m = 1/2\ p_0 + 1/2\ p_{2m}$. If $p_j = j\ p_1$ for $j \leq m$, then we have

$mp_1 = 0 + 1/2\ p_{2m}$, so $p_{2m} = 2mp_1$. Done.

In Theorem 7.6.8, p152, you have k of the n chips in play. Each hand, you gain 1 with prob. p, or lose 1 with prob. q=1-p.

Suppose $0<p<1$ and $p\neq0.5$. Let $r = q/p$. Then P(you win the tournament) = $\mathbf{(1-r^k)/(1-r^n)}$.

The proof is again by induction, and is similar to the proof we did of Theorem 7.6.6.

## 10. Examples.

(Chen and Ankenman, 2006). Suppose that a \$100 winner-take-all tournament has $1024 = 2^{10}$ players. So, you need to double up 10 times to win. Winner gets \$102,400.

Suppose you have probability $p = 0.54$ to double up, instead of 0.5.

What is your expected profit in the tournament? (Assume only doubling up.)

Answer. P(winning) = $0.54^{10}$, so exp. return = $0.54^{10}$ (\$102,400) = \$215.89. So exp. profit = \$115.89.

What if each player starts with 10 chips, and you gain a chip with $p = 54\%$ and lose a chip with $p = 46\%$? What is your expected profit?

Answer. $r = q/p = .46/.54 = .852$. P(you win) = $(1-r^{10})/(1-r^{10240}) = 79.9\%$.

So exp. profit = .799(\$102400) - \$100 ~ \$81700.

**Random Walk example.**

Suppose you start with 1 chip at time 0 and that your tournament is like a simple random walk, but if you hit 0 you are done. P(you have not hit zero by time 47)?

We know that starting at 0, $P(Y_1 \neq 0, Y_2 \neq 0, \ldots, Y_{2n} \neq 0) = P(Y_{2n} = 0)$.

So, for a random walk starting at (0,0),

by symmetry $P(Y_1 > 0, Y_2 > 0, \ldots, Y_{48} > 0) = ½\ P(Y_1 \neq 0, Y_2 \neq 0, \ldots, Y_{2n} \neq 0)$

$= ½\ P(Y_{48} = 0) = ½$ Choose(48,24)$(½)^{48}$.

Also $P(Y_1 > 0, Y_2 > 0, \ldots, Y_{48} > 0) = P(Y_1 = 1, Y_2 > 0, \ldots, Y_{48} > 0)$

= P(start at 0 and win your first hand, and then stay above 0 for at least 47 more hands)

= P(start at 0 and win your first hand) x P(from (1,1), stay above 0 for ≥ 47 more hands)

= 1/2 P(starting with 1 chip, stay above 0 for at least 47 more hands).

So, multiplying both sides by 2,

P(starting with 1 chip, stay above 0 for at least 47 hands) = Choose(48,24)$(½)^{48}$

= 11.46%.