Stat 100a: Introduction to Probability.

<u>Outline for the day:</u>

1. Midterm2 back.
2. CLT again.
3. Confidence intervals.
4. Sample size calculation.
5. Random walks.
6. Reflection principle.
7. Ballot theorem.
8. Tournaments.

HW3 is due Wed.

1. Midterm 2 back.

Please be completely silent until I am done passing back the exams.

2. Central Limit Theorem (CLT): if $X_1, X_2 \ldots, X_n$ are iid with mean μ & SD σ, then

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

In other words, $\overline{X}_n$ has mean μ 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).

μ = 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 μ +/- 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 μ & SD σ, 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 μ and standard deviation of $\sigma \div \sqrt{n}$.

That is, 95% of the time, $\overline{X_n}$ is in the interval μ +/- 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 μ +/- 1.96 $(\sigma/\sqrt{n})$, where μ = \$5, σ = \$60, and n=1600. So the answer is

\$5 +/- 1.96 x \$60 / √(1600)

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

3. Confidence Intervals (CIs) for μ, ch 7.5.

Central Limit Theorem (CLT): if $X_1, X_2 \ldots, X_n$ are iid with mean μ & SD σ, 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 μ +/- $1.96\ (\sigma/\sqrt{n})$.

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

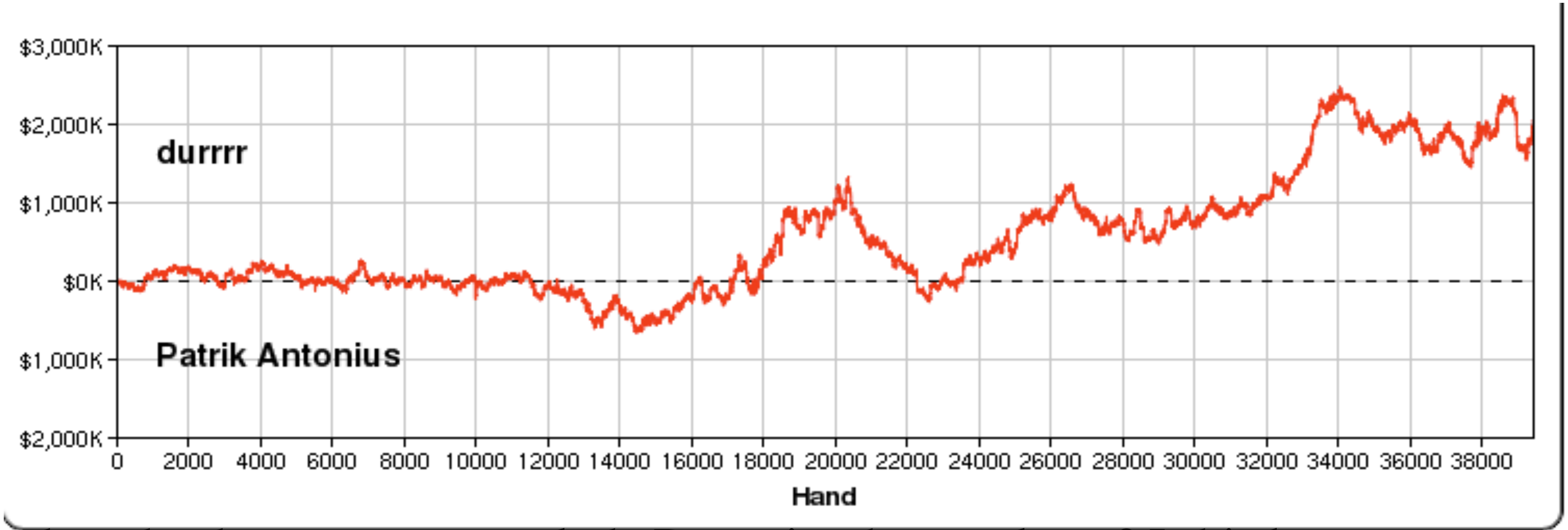
This range $\overline{X}_n$ +/- $1.96\ (\sigma/\sqrt{n})$ is called a 95% confidence interval (CI) for μ.

[Usually you don't know σ 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 μ. 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 μ is \$51 +/- 1.96 (\$10,000 / $\sqrt{39{,}000}$) = \$51 +/- \$99 = (-\$48, \$150).

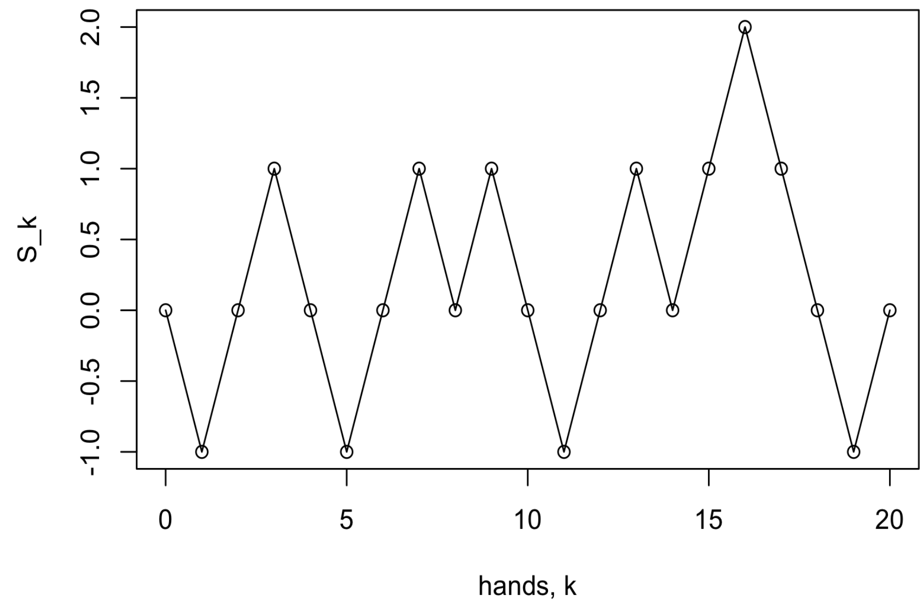
Results are inconclusive, even after 39,000 hands!

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

5. 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.

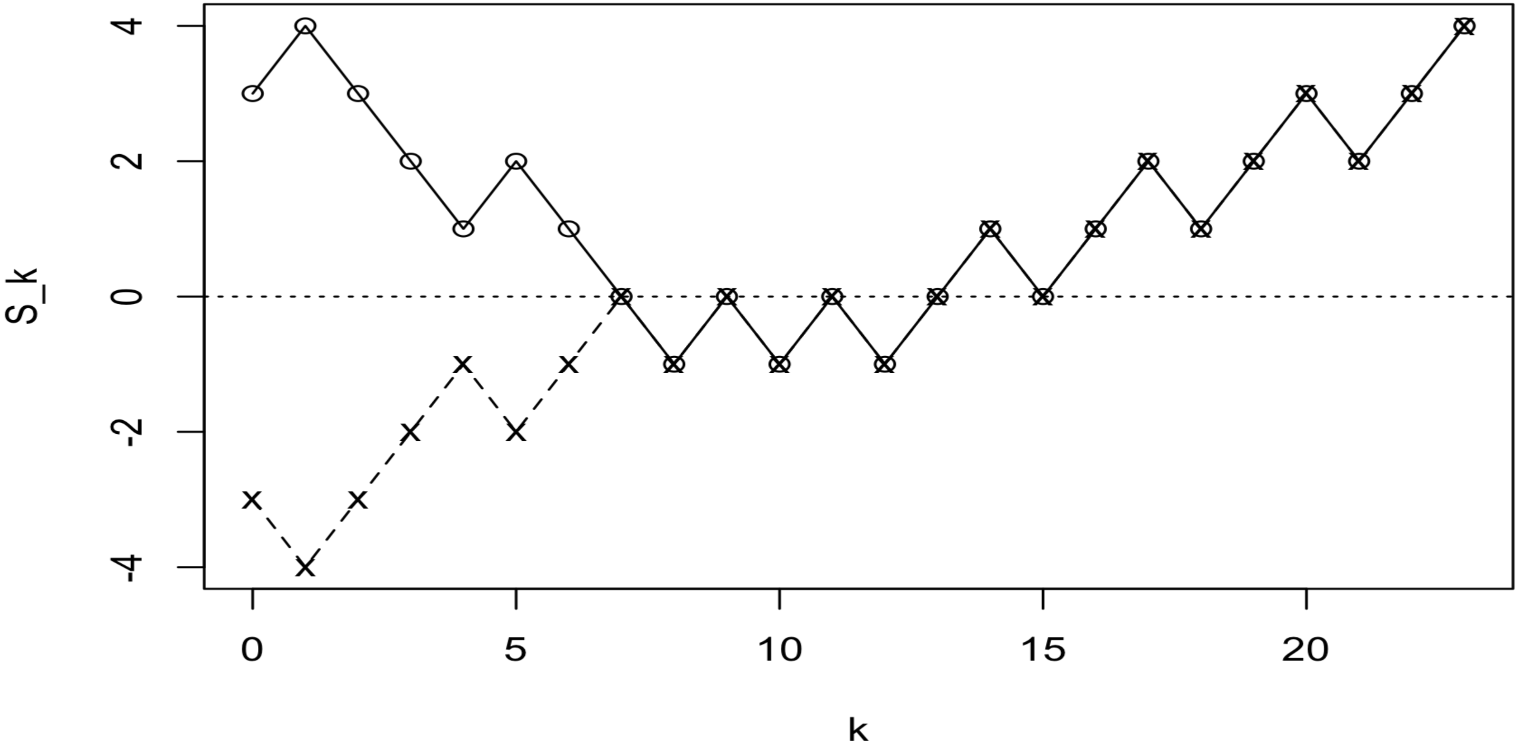
* *Reflection principle:* The number of paths from $(0,X_0)$ to (n,y) that touch the x-axis = the number of paths from $(0,-X_0)$ to (n,y), for any n,y, and $X_0 > 0$.

* *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.

6. 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

[i.e. $a-b = y - (-X_0) = y + X_0$. $a+b=n$, so $b = n-a$, $2a-n=y+X_0$, $a=(n+y+X_0)/2$].

7. 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).

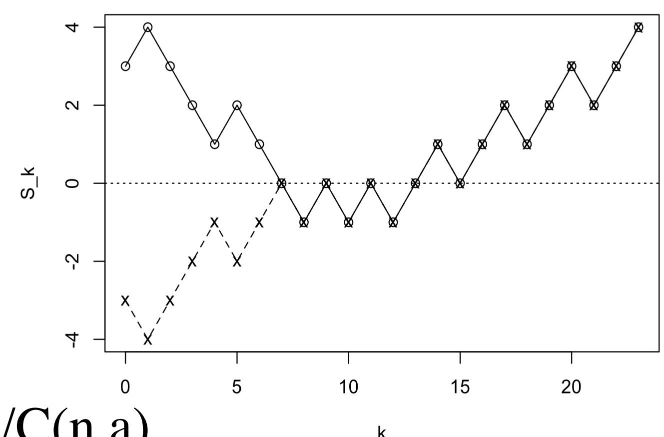
So the number of paths from (1,1) to (n,x) that **do not** touch the x-axis equals the number of paths from (1,1) to (n,x) minus the number of paths from (1,-1) to (n,x)

= 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,a) without touching the x-axis = (a-b)/n.

8. Tournaments.