Stat 100a: Introduction to Probability.
Outline for the day:

1. Midterm2 back.
2. CLT again.
3. Confidence intervals.
4. Sample size calculation.
5. Random walks.
6. Reflection principle.
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 \( \mu \) & SD \( \sigma \), then

\[
\left( \bar{X}_n - \mu \right) \div (\sigma/\sqrt{n}) \longrightarrow \text{Standard Normal. (mean 0, SD 1)}.
\]

In other words, \( \bar{X}_n \) has mean \( \mu \) and a standard deviation of \( \sigma/\sqrt{n} \).

Two interesting things about this:

(i) As \( n \to \infty \), \( \bar{X}_n \to \text{normal} \). 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). 

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

(ii) We can use this to find a range where \( \bar{X}_n \) 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, \( \left( \bar{X}_n - \mu \right) \div (\sigma/\sqrt{n}) \) is within -1.96 to +1.96. 

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

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

That is, 95\% of the time, \( \bar{X}_n \) is in the interval \( \mu \pm 1.96 (\sigma/\sqrt{n}) \).

\( = 5.88\% \pm 1.96(23.525\%/\sqrt{n}) \). 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 $X_1, X_2 \ldots, X_n$ are iid with mean $\mu$ & SD $\sigma$, then

\[
\left( \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \rightarrow \text{Standard Normal. (mean 0, SD 1).}
\]

In other words, $\bar{X}_n$ is like a draw from a normal distribution with mean $\mu$ and standard deviation of $\sigma/\sqrt{n}$.

That is, 95% of the time, $\bar{X}_n$ 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

$5 +/\ - 1.96 \times 60 / \sqrt{1600}$

$= 5 +/\ - 2.94$, or the range [$2.06, 7.94$].
3. Confidence Intervals (CIs) for $\mu$, ch 7.5.

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

$$\left( \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \right) \sim \text{Standard Normal. (mean 0, SD 1).}$$

So, 95% of the time, $\bar{X}_n$ is in the interval $\mu \pm 1.96 \left( \frac{\sigma}{\sqrt{n}} \right)$.

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

This range $\bar{X}_n \pm 1.96 \left( \frac{\sigma}{\sqrt{n}} \right)$ 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 $\left( \frac{\bar{X}_n - \mu}{s/\sqrt{n}} \right)$ 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 $\overline{X}_n \pm 1.96 (\sigma/\sqrt{n})$ is a 95% confidence interval for $\mu$.

(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 $\sim$ $10,000.

95% CI for $\mu$ 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)/($51)]^2 \sim 148,000$, so about 109,000 more hands.
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 $\frac{1}{2}$ 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

\[\text{i.e. } a-b = y - (-X_0) = y + X_0. \quad a+b=n, \text{ so } b = n-a, \quad 2a-n=y+X_0, \quad a=(n+y+X_0)/2\].
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

\[
P(A \text{ won more hands than } B \text{ throughout the telecast}) = \frac{a-b}{n}.
\]

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 \((n,x)\)

\[
= \binom{n-1}{a-1} - \binom{n-1}{a}
\]

\[
= \frac{(n-1)!}{[(a-1)! (n-a)!]} - \frac{(n-1)!}{[a! (n-a-1)!]}
\]

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

\[
= \binom{n}{a} \frac{a-b}{n}.
\]

And each path is equally likely, and has probability \( 1/\binom{n}{a} \).

So, \( P(\text{going from } (0,0) \text{ to } (n,a) \text{ without touching the x-axis}) = \frac{a-b}{n}. \)
8. Tournaments.