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

Outline for the day

1. Return midterm2.
2. Bivariate and marginal density.
3. CLT.
4. CIs.
5. Sample size calculations.

Please keep silent until I am finished returning Midtem 2.

Everyone's score is boosted by 1 point out of 14.
So if it says 9 in red, circled, on the last problem of your exam, then you effectively got a score of 10/14.

HW3 is due Wed Mar2, 2pm by email .

**Bivariate and marginal density.**

Suppose X and Y are random variables.

If X and Y are discrete, we can define the joint pmf $f(x,y) = P(X = x \text{ and } Y = y)$.

Suppose X and Y are continuous for the rest of this page.

Define the bivariate or joint pdf $f(x,y)$ as a function with the properties that $f(x,y) \geq 0$, and for any a,b,c,d,

$$P(a \leq X \leq b \text{ and } c \leq Y \leq d) = \int_a^b \int_c^d f(x,y)\, dy\, dx.$$

The integral $\int_{-\infty}^{\infty} f(x,y)\, dy = f(x)$, the pdf of X, and this function $f(x)$ is sometimes called the *marginal* density of X. Similarly $\int_{-\infty}^{\infty} f(x,y)\, dx$ is the marginal pdf of Y.

$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx = \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f(x,y)\, dy\right] dx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\, f(x,y)\, dy\, dx.$

Just as $P(A|B) = P(AB)/P(B)$, $f(x|y) = f(x,y)/f(y)$.

X and Y are independent iff. $f(x,y) = f_x(x)f_y(y)$.

Now $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy\, f(x,y)\, dy\, dx$. This can be useful to find $\text{cov}(X,Y) = E(XY) - E(X)E(Y)$.

What is $E(X^2Y+e^Y)$? It $= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x^2y+e^y)\, f(x,y)\, dy\, dx$.

Bivariate and marginal density.

Suppose the joint density of X and Y is $f(x,y) = a \exp(x+y)$, for X and Y in $(0,1) \times (0,1)$. What is a? What is the marginal density of Y? What type of distribution does X have conditional on Y? What is $E(X|Y)$? What is the mean of X when Y = .2? Are X and Y independent?

$\iint a \exp(x+y)\, dxdy = 1 = a\iint \exp(x)\exp(y)dxdy = a\int_0^1 \exp(x)\, dx \int_0^1 \exp(y)\, dy = a(e-1)^2$, so $a = (e-1)^{-2}$.

The marginal density of Y is $f(y) = \int_0^1 a \exp(x+y)\, dx = a \exp(y) \int_0^1 \exp(x)dx = a \exp(y)(e-1) = \exp(y)/(e-1)$.

Conditional on Y, the density of X is $f(x|y) = f(x,y)/f(y) = a \exp(x+y)(e-1)/\exp(y) = \exp(x)/(e-1)$. So X|Y is like an exponential(1) random variable restricted to (0,1).

$E(X|Y) = \int_0^1 x \exp(x)/(e-1)\, dx = 1/(e-1)\, [x \exp(x) - \int \exp(x)dx] = 1/(e-1)\, [x \exp(x) - \exp(x)]_0^1 = 1/(e-1)\, [e - e - 0 + 1] = 1/(e-1)$.
When Y = .2, $E(X|Y) = 1/(e-1)$.

$f(y) = \exp(y)/(e-1)$ and similarly $f(x) = \exp(x)/(e-1)$,
so $f(x)f(y) = \exp(x+y)/(e-1)^2 = f(x,y)$. Therefore, X and Y are independent.

# **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):

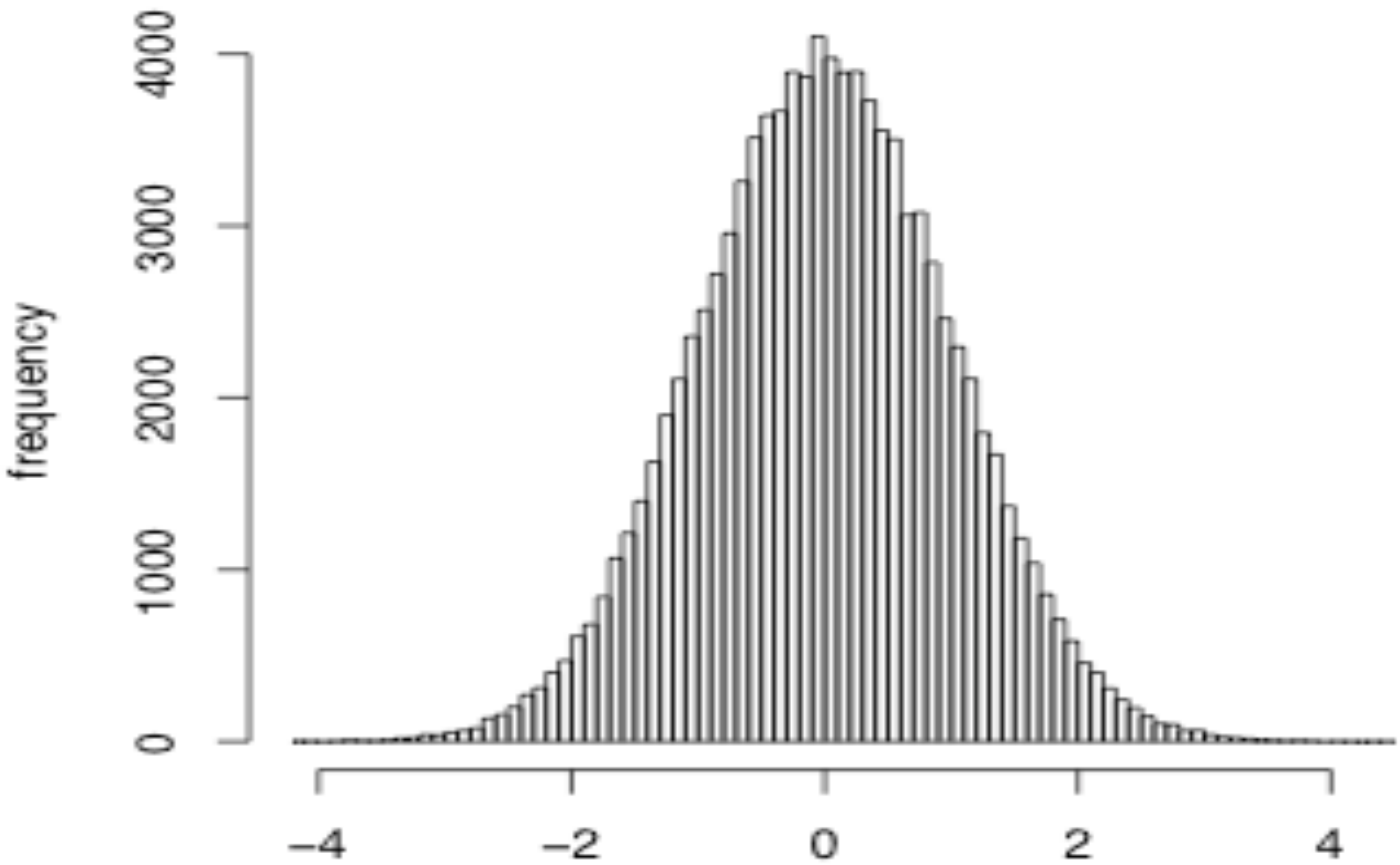
$\overline{X}_n$ ---> $\mu$ .

CENTRAL LIMIT THEOREM (CLT):

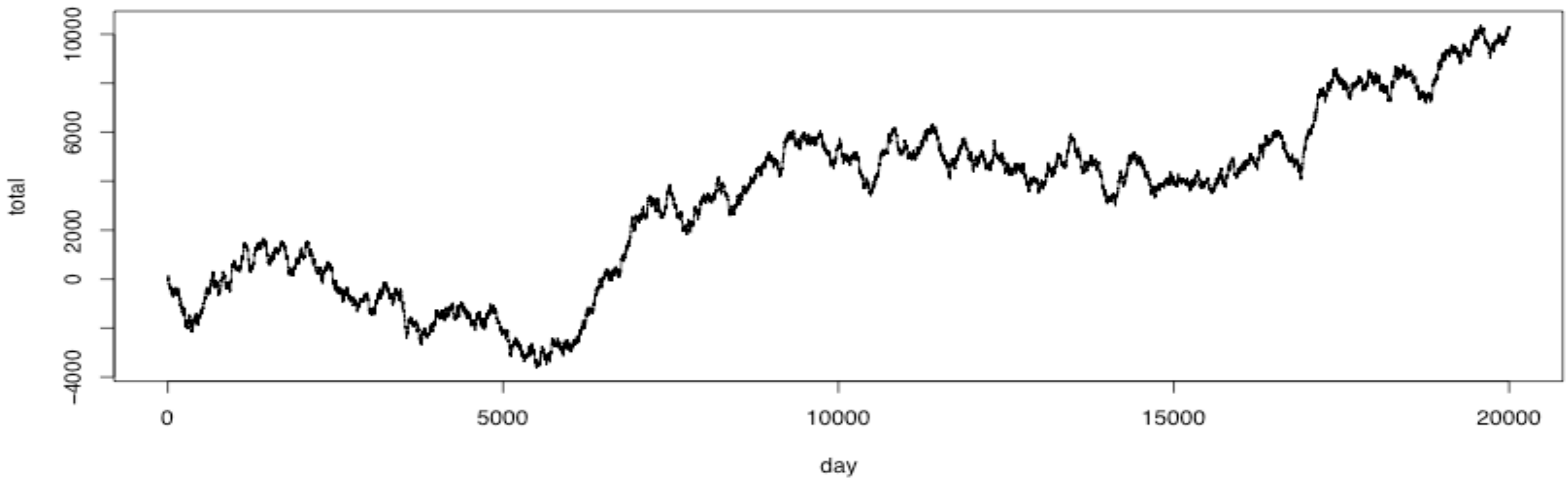
$(\overline{X}_n - \mu) \div (\sigma/\sqrt{n})$ ---> Standard Normal.

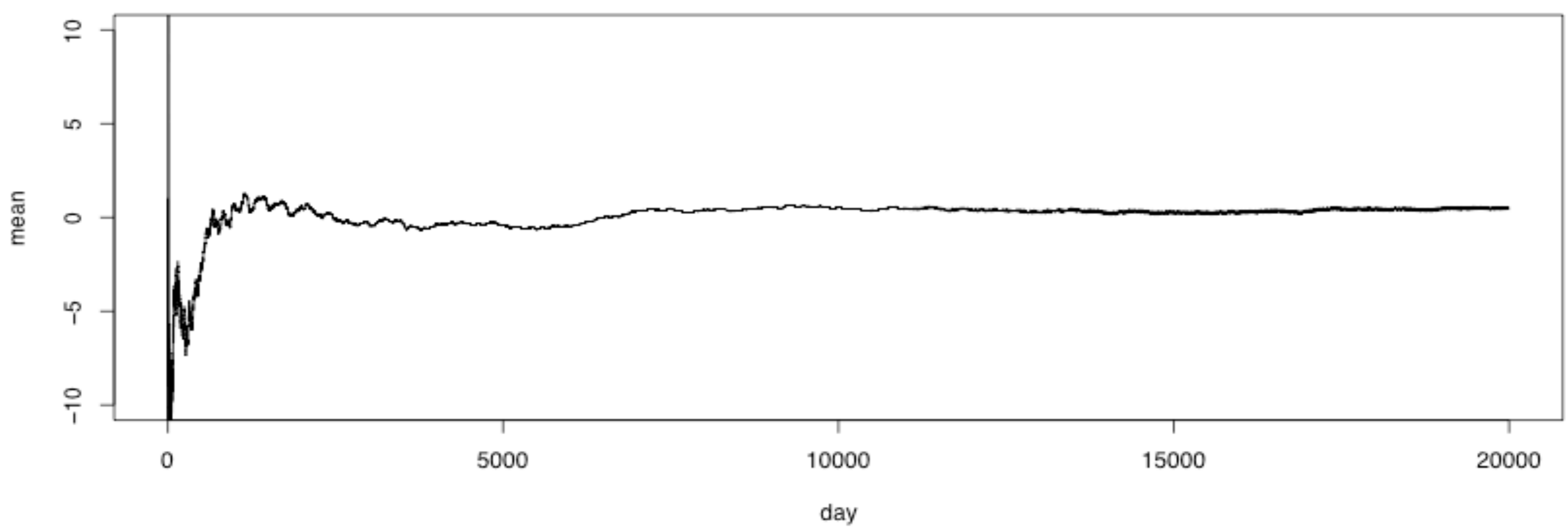
Useful for tracking results.

# 95% between -1.96 and 1.96

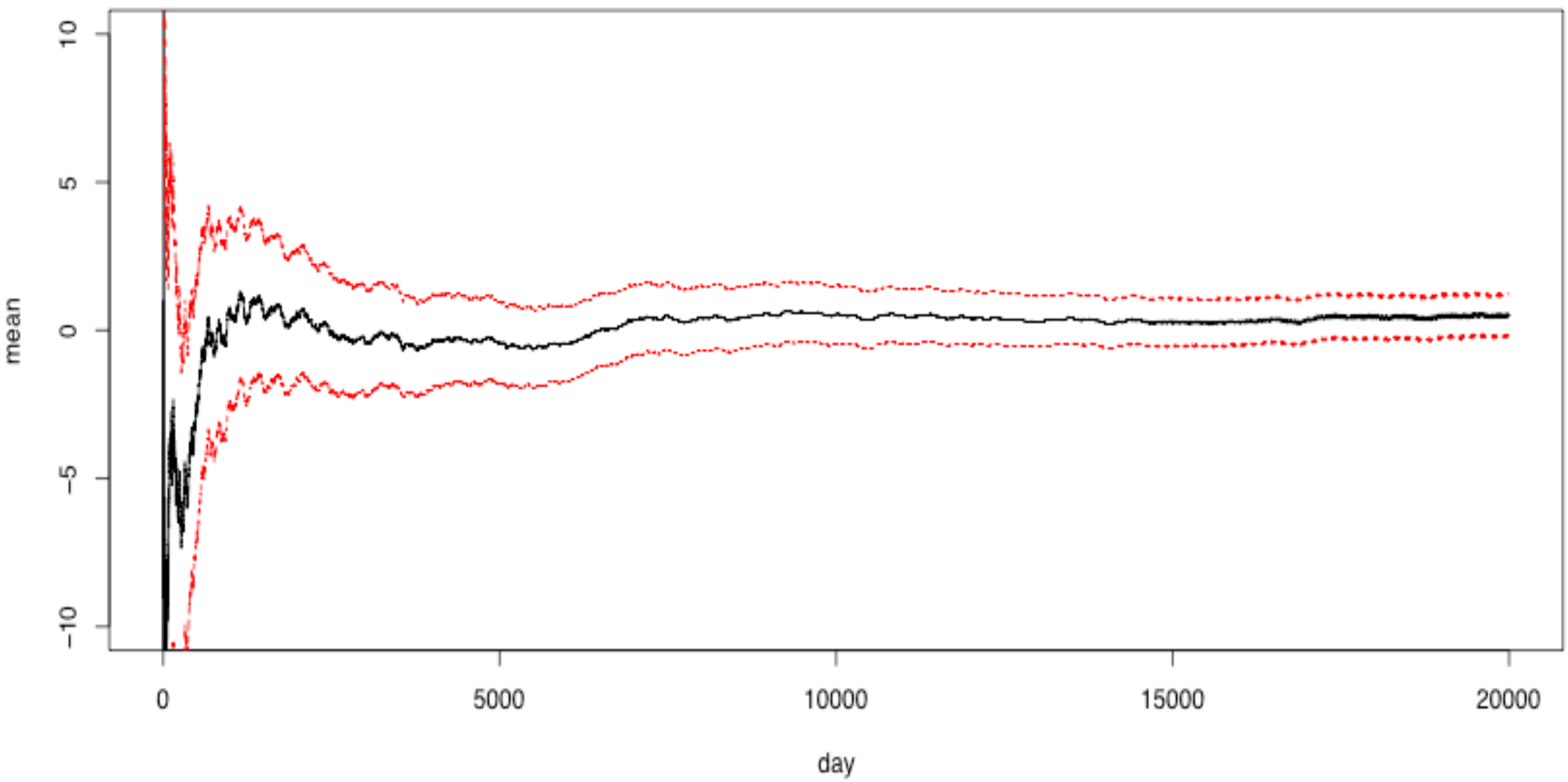


# Truth: -49 to 51, exp. value $\mu = 1.0$



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



<u>Central Limit Theorem (CLT):</u> if $X_1$ , $X_2$ …, $X_n$ are iid with mean $\mu$ & 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 $\mu$ 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%. σ = √(pq) = √(5.88% x 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 (σ/√n).**

**= 5.88% +/- 1.96(23.525%/√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

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

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].

**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

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

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 $\overline{X}_n$ +/- $1.96\,(\sigma/\sqrt{n})$.

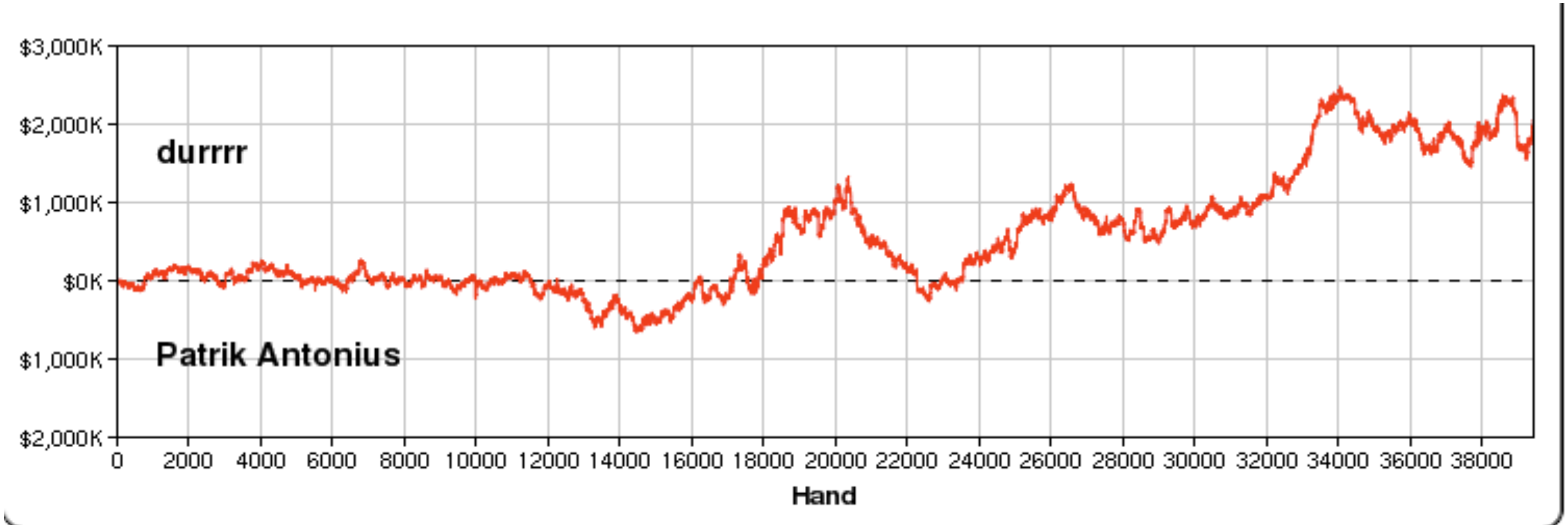
This range $\overline{X}_n$ +/- $1.96\,(\sigma/\sqrt{n})$ is called a 95% confidence interval (CI) for $\mu$.

[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!

**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.