Stat 100a: Introduction to Probability.

<u>Outline for the day</u>

1. CIs and sample size calculations.
2. Review.

Exam 2 is Tue Nov 20. Bring a calculator.

It will be on everything up to and including CIs and sample size calculations.

HW3 is due Tue Dec 4.

Confidence Intervals (CIs) for μ, ch 7.5.

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

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

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.

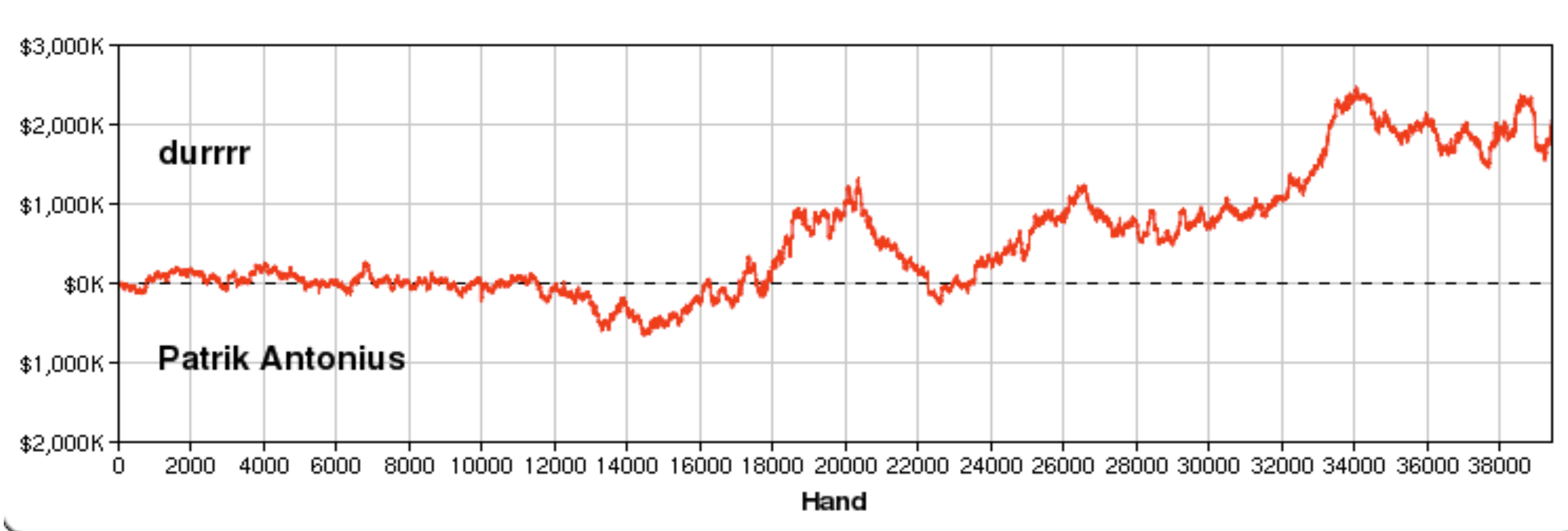
[Usually you don't know σ and have to estimate it using the sample std deviation, s, of your data, and $(\overline{X}_n - \mu) \div (s/\sqrt{n})$ has a t_{n-1} distribution if the X_i are normal.

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 (σ/√n) is a 95% confidence interval for μ. 1.96 (σ/√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 / √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 (σ/√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.

Review problems for Exam 2.

What is the probability you will have a full house on the turn?

For example, 3♥ 3♠ 3♣ 7♣ 7♦ K♥.

There are 13 possibilities for the number on the triplet, and for each such choice,

there are C(4,3) choices for the suits on the triplet, and for each such choice,

there are 12 possibilities for the number on the pair, and for each such choice,

there are C(4,2) possibilities for the suits of the pair, and for each such choice,

there are 44 possibilities left for the last card to go with the full house.

It is tempting to answer 13 * C(4,3) * 12 * C(4,2) * 44.

But this would ignore 3♥ 3♠ 3♣ 7♣ 7♦ 7♥. So add in C(13,2)*C(4,3)*C(4,3).

The answer is (13 * C(4,3) * 12 * C(4,2) * 44 + C(13,2)*C(4,3)*C(4,3))/C(52,6)

~ 1 in 123.

♠ ♣ ♥ ♦

Review problems for Exam 2.

What is the probability you will have a straight on the turn? Include the case where you have a straight and you also have a flush, or where you have a straightflush.

For example, 4♥ 5♠ 6♣ 7♣ 8♦ Q♥ or 4♥ 5♥ 6♥ 7♣ 8♥ Q♥.

First, note there are 10 straights. A2345, 23456, ..., 10JQKA.

It is tempting to say $10 * 4^5 * 32$.

10 choices for the numbers on the straight, 4*4*4*4*4 choices for the suits on the straight, and 52-4*5 = 32 possibilities for the other card.

But there are two problems here. We have not counted 4♥ 5♠ 6♣ 7♣ 8♦ 5♦.

And, we have doublecounted 4♥ 5♠ 6♣ 7♣ 8♦ 9♠.

One way to do this is as follows. First, consider cases with no pairs, and count A2345 separately. For any other straight, there are 28 cards to go with it so it is not a straight already counted. For 45678, for instance, the last card must not be 34567 or 8, so there are 28 such cards. For A2345, there are 32 cards to go with it. So we have $9*4^5*28+4^5*32$ combinations with no pairs, plus $10*5*C(4,2)*4^4$ combinations with a pair like 4♥ 5♠ 6♣ 7♣ 8♦ 5♦, and answer is $(9*4^5*28+4^5*32 + 10*5*C(4,2)*4^4) / C(52,6) \sim 1$ in 55.4.

Review problems for Exam 2.

Why $10*5*C(4,2)*4^4$ combinations with a pair like 4♥ 5♠ 6♣ 7♣ 8♦ 5♦?

There are 10 choices for the numbers on the straight, like 45678.

For each such choice, there are 5 choices for the number to be paired up, like 5.

For each such choice, there are C(4,2) possibilities for the suits on the pair.

For each such choice, there are 4 possibilities for the suits on the other 4 cards.

Review problems for Exam 2.

Suppose you have 10 players at the table. What is the expected number of players who have 2 face cards?

P(2 face cards) = C(12,2)/C(52,2) = 4.98%.

Let $X1 = 1$ if player 1 has 2 face cards, and $X1 = 0$ otherwise.

$X2 = 1$ if player 2 has 2 face cards, and $X2 = 0$ otherwise. etc.

$X = \sum Xi$ = total number of players with 2 face cards.

$E(X) = \sum E(Xi) = 10 \times 4.98\% = 0.498$.

Let $X = N(0, 0.8^2)$ and $\varepsilon = N(0, 0.1^2)$ and ε is independent of X. Let $Y = 7 + 0.2 X + \varepsilon$.

Find E(X), E(Y), E(Y|X), var(X), var(Y), cov(X,Y), and ρ = cor(X,Y).

$E(X) = 0$.

$E(Y) = E(7 + 0.2X + \varepsilon) = 7 + 0.2\, E(X) + E(\varepsilon) = 7$.

$E(Y|X) = E(7 + 0.2X + \varepsilon \mid X) = 7 + 0.2X + E(\varepsilon \mid X) = 7 + 0.2X$ since ε and X are ind.

$\text{var}(X) = 0.64$.

$\text{var}(Y) = \text{var}(7 + 0.2\, X + \varepsilon) = \text{var}(0.2X + \varepsilon) = 0.2^2\, \text{var}(X) + \text{var}(\varepsilon) + 2{*}0.2\, \text{cov}(X,\varepsilon)$

$= 0.2^2(.64) + 0.1^2 + 0 = 0.0356$.

$\text{cov}(X,Y) = \text{cov}(X, 7 + 0.2X + \varepsilon) = 0.2\, \text{var}(X) + \text{cov}(X, \varepsilon) = 0.2(0.64) + 0 = 0.128$.

$\rho = \text{cov}(X,Y)/(\text{sd}(X)\, \text{sd}(Y)) = 0.128 / (0.8 \text{ x } \sqrt{.0356}) = 0.848$.

Suppose (X,Y) are bivariate normal with $E(X) = 10$, $var(X) = 9$, $E(Y) = 30$, $var(Y) = 4$, $\rho = 0.3$,

What is the distribution of Y given $X = 7$?

Given $X = 7$, Y is normal. Write $Y = \beta_1 + \beta_2 X + \varepsilon$ where ε is normal with mean 0, ind. of X.

Recall $\beta_2 = \rho\ \sigma_y/\sigma_x = 0.3 \text{ x } 2/3 = 0.2$.

So $Y = \beta_1 + 0.2\ X + \varepsilon$.

To get β_1, note $30 = E(Y) = \beta_1 + 0.2\ E(X) + E(\varepsilon) = \beta_1 + 0.2\ (10) + 0$. So $30 = \beta_1 + 2$. $\beta_1 = 28$.

So $Y = 28 + 0.2\ X + \varepsilon$, where ε is normal with mean 0 and ind. of X.

What is $var(\varepsilon)$?

$4 = var(Y) = var(28 + 0.2\ X + \varepsilon) = 0.2^2\ var(X) + var(\varepsilon) + 2(0.2)\ cov(X,\varepsilon)$
$= 0.2^2\ (9) + var(\varepsilon) + 0$. So $var(\varepsilon) = 4 - 0.2^2(9) = 3.64$ and $sd(\varepsilon) = \sqrt{3.64} = 1.91$.

So $Y = 28 + 0.2\ X + \varepsilon$, where ε is $N(0, 1.91^2)$ and ind. of X.

Given $X = 7$, $Y = 28 + 0.2(7) + \varepsilon = 29.4 + \varepsilon$, so $Y|X{=}7 \sim N(29.4, 1.91^2)$.

P(straight flush or royal flush on the turn)?

The trick is not to double count cases like 345678 all of clubs. The way to do it is to count the highest one, AKQJ10, separately. There are 9 other straight flushes, and for each choice, like say 45678, there are 46 cards left for the other card to go with it so our choice of straight flush remains the same. For instance, with 45678 of clubs, we want to exclude the 9 of clubs as a possibility for the other card, because otherwise our straight flush would be 56789 of clubs. But with AKQJ10 of clubs, now there are 47 possibilities for the other card to go with it.

There are 4 suits, and for each suit, there are 9 straight flushes with 46 cards to go with them plus one with 47 cards to go with it, so the answer is $(4*9*46+4*47)/C(52,6) = 0.0000906$.

P(straight flush on the turn but not royal flush) = $(4*9*46)/C(52,6) = 0.0000813$.

Review list.

1) Basic principles of counting.
2) Axioms of probability, and addition rule.
3) Permutations & combinations.
4) Conditional probability.
5) Independence.
6) Multiplication rules.
7) Odds ratios.
8) Discrete RVs and probability mass function.
9) Expected value.
10) Pot odds calculations.
11) Luck and skill.
12) Variance and SD.
13) $E(aX+b)$ and $E(X+Y)$.
14) Bayes's rule.
15) Markov and Chebyshev inequalities.
16) Bernoulli, binomial, geometric, Poisson, and negative binomial rvs.
17) Moment generating functions.
18) pdf, cdf, uniform, exponential, normal and Pareto rvs. $F'(y) = f(y)$.
19) Survivor functions.
20) Covariance and correlation.
21) Bivariate normal.
22) Conditional expectation.
23) LLN, CLT, CIs, and sample size calculations.
24) Bivariate density and marginal density.

What integrals do you need to know?

You need to know $\int e^{ax}\,dx$, $\int ax^k\,dx$ for any k, and $\int \log(x)\,dx$,

and basic stuff like $\int [af(x) + g(x)]\,dx = a\int f(x)dx + \int g(x)dx$,

and you need to understand that $\int\int f(x,y)\,dy\,dx = \int[\int f(x,y)\,dy]\,dx$.

Bivariate and marginal density example.

Suppose the joint density of X and Y is $f(x,y) = a(xy + x + y)$, for X and Y in $(0,2) \times (0,2)$. What is a? What is the marginal density of Y? What is the density of X conditional on Y? What is $E(X|Y)$? Are X and Y independent?

$\iint a(xy + x + y)\, dy\, dx = 1 = a\int [xy^2/2 + xy + y^2/2\,]_{y=0}^{2}\, dx = a\int [2x + 2x + 2 - 0 - 0 - 0]dx$
$= a(x^2 + x^2 + 2x)\,]_{x=0}^{2} = a(4 + 4 + 4 - 0 - 0 - 0) = 12a$, so $a = 1/12$.
The marginal density of Y is $f(y) = \int_0^2 a(xy+x+y)\, dx$
$= ay \int_0^2 x dx + a\int_0^2 x dx + ay \int_0^2 dx$
$= y/12\, (x^2/2)\,]_{x=0}^{2} + 1/12\, (x^2/2)]_{x=0}^{2} + y/12\; x]_{x=0}^{2}$
$= 2y/12 + 2/12 + 2y/12$
$= y/3 + 1/6$.
Check that this is a density. $\int_0^2 (y/3 + 1/6)\, dy = (y^2/6 + y/6)\,]_{y=0}^{2} = 4/6 + 2/6 - 0 - 0 = 1$.

Conditional on Y, the density of X is $f(x|y) = f(x,y)/f(y) = (xy+x+y) / [12(y/3+1/6)]$
$= (xy+x+y)/(4y+2)$.
$E(X|Y) = \int_0^2 x(xy+x+y)/(4y+2)\, dx = (x^3y/3 + x^3/3 + x^2y/2)/(4y+2)]_{x=0}^{2}$
$= (8y/3+8/3+2y-0-0-0)/(4y+2) = 14y/3 + 8/3$.
$f(y) = y/3 + 1/6$ and similarly $f(x) = x/3 + 1/6$,
so $f(x)f(y) = xy/9 + x/18 + y/18 + 1/36 \neq f(x,y)$. So, X and Y are not independent.

Suppose X = the number of hands until you get dealt at least one black card. After this, you play 100 more hands and Y = the number of hands where you get dealt pocket aces out of these next 100 hands.

Let $Z = 4X + 7Y$. What is the SD of X? What is SD(Y)? What is E(Z)? What is SD(Z)?

X is geometric(p), where $p = 1 - P(\text{both red}) = 1 - C(26,2)/C(52,2) \sim 75.5\%$. SD(Y) = $\sqrt{q}/p = 0.656$.

Y is binomial(n,p), n = 100 and $p = C(4,2)/C(52,2) \sim 0.452\%$. SD(X) = $\sqrt{(npq)} = 0.671$.

$E(Z) = 4E(X) + 7E(Y) = 4(1/.755) + 7(100)(.00452) = 8.46$.

X and Y are independent so $Var(Z) = Var(4X) + Var(7Y) = 16Var(X) + 49Var(Y) = 16(.656^2) + 49(.671^2) = 28.9$. So $SD(Z) = \sqrt{28.9} = 5.38$.

Let X = 1 if you are dealt pocket aces and 0 otherwise. Let Y = 1 if you are dealt two black cards and 0 otherwise. What is cov(3X, 7Y)?

cov(3X, 7Y) = 21cov(X,Y).
cov(X,Y) = E(XY) – E(X)E(Y).
E(X) = 1 P(pocket aces) + 0 P(not pocket aces) = C(4,2)/C(52,2) = 0.452%.
E(Y) = 1 P(2 black cards) + 0 P(not 2 black cards) = C(26,2)/C(52,2) = 24.5%.
Here XY = 1 if X and Y are both 1, and XY = 0 otherwise.
So E(XY) = 1 P(X and Y = 1) + 0 P(X or Y does not equal 1)
= P(2 black aces) + 0
= 1 / C(52,2) = 0.0754%.
cov(X,Y) = .000754 - .00452(.245) = -.0003534.
cov(3X, 7Y) = 21 (-.0003534) = -.00742.