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

1. Return exams.
2. Harman/Negreanu and running it twice.
3. Continuous random variables and densities.
4. Uniform random variables.
5. Exponential random variables.
6. Normal random variables.
7. Pareto random variables.
8. Functions of independent random variables.
9. Moment generating functions of rvs.
10. Survivor functions.
11. Covariance and correlation.

- Next midterm is Nov19.
- Homework 2 is due Tue Nov12.
- Read through ch6.
- For gradegrubbing, see syllabus. If you would like a question reevaluated, submit your work and a WRITTEN explanation of why you think you deserve more points and how many more points you think you deserve to your TA. The TA will then give it to me, and I will consider it, and then give it back to the TA to give back to you.

http://www.stat.ucla.edu/~frederic/100a/F19 .

**2. Harman / Negreanu, and running it twice.**

Harman has 10♠ 7♠ . Negreanu has K♥ Q♥ . The flop is 10◆ 7♣ K◆ .

Harman's all-in. \$156,100 pot. P(Negreanu wins) = 28.69%. P(Harman wins) = 71.31%.

Let X = amount Harman has after the hand.

If they run it once, E(X) = \$0 x 29% + \$156,100 x 71.31% = **\$111,314.90.**

<u>If they run it twice, what is E(X)?</u>

There's some probability $p_1$ that Harman wins both times ==> X = \$156,100.

There's some probability $p_2$ that they each win one ==> X = \$78,050.

There's some probability $p_3$ that Negreanu wins both ==> X = \$0.

E(X) = \$156,100 x $p_1$ + \$78,050 x $p_2$ + \$0 x $p_3$.

If the different runs were *independent,* then $p_1$ = P(Harman wins 1st run & 2nd run) would = P(Harman wins 1st run) x P(Harman wins 2nd run) = 71.31% x 71.31% ~ 50.85%.

But, they're not quite independent! Very hard to compute $p_1$ and $p_2$.

*However, you don't need* $p_1$ *and* $p_2$ *!*

X = the amount Harman gets from the 1st run + amount she gets from 2nd run, so

E(X) = E(amount Harman gets from 1st run) + E(amount she gets from 2nd run)

= \$78,050 x P(Harman wins 1st run) + \$0 x P(Harman loses first run)

+ \$78,050 x P(Harman wins 2nd run) + \$0 x P(Harman loses 2nd run)

= \$78,050 x 71.31% + \$0 x 28.69% + \$78,050 x 71.31% + \$0 x 28.69% = **\$111,314.90.**

HAND RECAP Harman 10♠ 7♠ Negreanu K♥ Q♥ The flop is 10♦ 7♣ K♦ .

Harman's all-in. $156,100 pot.P(Negreanu wins) = 28.69%. P(Harman wins) = 71.31%.

---

The standard deviation (SD) changes a lot! **<u>Say they run it once</u>**. (see p127.)

$V(X) = E(X^2) - \mu^2$.

$\mu$ = $111,314.9, so $\mu^2$ ~ $12.3 billion.

$E(X^2)$ = ($156,100²)(71.31%) + (0²)(28.69%) = $17.3 billion.

V(X) = $17.3 billion - $12.3 bill. = $5.09 billion. SD σ = sqrt($5.09 billion)~$71,400.

So if they run it once, Harman expects to get back about $111,314.9 **+/- $71,400**.

**<u>If they run it twice?</u>** Hard to compute, but approximately, if each run were independent, then $V(X_1+X_2) = V(X_1) + V(X_2)$,

so if $X_1$ = amount she gets back on 1st run, and $X_2$ = amount she gets from 2nd run,

then $V(X_1+X_2) \sim V(X_1) + V(X_2)$ ~ $1.25 billion + $1.25 billion = $2.5 billion,

The standard deviation σ = sqrt($2.5 billion) ~ $50,000.

So if they run it twice, Harman expects to get back about $111,314.9 **+/- $50,000.**

## 3. Continuous random variables and their densities, ch6.1.

Density (or pdf = Probability Density Function) f(y):

$$\int_B f(y)\, dy = P(X \text{ in } B).$$

If F(c) is the cumulative distribution function, i.e. $F(c) = P(X \leq c)$,

then $f(c) = F'(c)$.

The survivor function is $S(c) = P(X > c) = 1 - F(c)$.

Expected value, $\mu = E(X) = \int y\, f(y)\, dy$. (= $\sum y\, P(y)$ for discrete X.)

For any function g, $E(g(X)) = \int g(y)\, f(y)\, dy$. For instance $E(X^2) = \int y^2 f(y) dy$.

Variance, $\sigma^2 = V(X) = Var(X) = E(X-\mu)^2 = E(X^2) - \mu^2$.

$SD(X) = \sqrt{V(X)}$.

For examples of pictures of pdfs, see p104, 106, and 107.

**4. Uniform example.**

Recall for a continuous random variable X,
the pdf f(y) is a function where $\int_a^b f(y)dy = P\{X \text{ is in } (a,b)\}$,
$E(X) = \mu = \int_{-\infty}^{\infty} y\, f(y)dy$,
and $\sigma^2 = Var(X) = E(X^2) - \mu^2$. $sd(X) = \sigma$.
If X is a continuous rv, then $P(X \le a) = P(X < a)$, because $P(X = a) = \int_a^a f(y)dy = 0$.
If X is uniform(a,b), then $f(y) = 1/(b-a)$ for y in (a,b), and y = 0 otherwise.

For example, suppose X and Y are independent uniform random variables on (0,1), and Z = *min*(X,Y). **a)** Find the pdf of Z. **b)** Find E(Z). **c)** Find SD(Z).

**a.** For c in (0,1), $P(Z > c) = P(X > c \;\&\; Y > c) = P(X > c)\, P(Y > c) = (1-c)^2 = 1 - 2c + c^2$.
So, $P(Z \le c) = 1 - (1 - 2c + c^2) = 2c - c^2$.
Thus, $\int_0^c f(c)dc = 2c - c^2$. So f(c) = the derivative of $2c - c^2 = 2 - 2c$, for c in (0,1).
Obviously, f(c) = 0 for all other c.
**b.** $E(Z) = \int_{-\infty}^{\infty} y\, f(y)dy = \int_0^1 c\,(2-2c)\, dc = \int_0^1 2c - 2c^2\, dc = c^2 - 2c^3/3]_{c=0}^{1}$
$= 1 - 2/3 - (0 - 0) = 1/3$.
**c.** $E(Z^2) = \int_{-\infty}^{\infty} y^2 f(y)dy = \int_0^1 c^2\,(2-2c)\, dc = \int_0^1 2c^2 - 2c^3\, dc = 2c^3/3 - 2c^4/4]_{c=0}^{1}$
$= 2/3 - 1/2 - (0 - 0) = 1/6$.
So, $\sigma^2 = Var(Z) = E(Z^2) - [E(Z)]^2 = 1/6 - (1/3)^2 = 1/18$.
$SD(Z) = \sigma = \sqrt{(1/18)} \sim 0.2357$.

**5. Exponential distribution, ch 6.4.**

Useful for modeling waiting times til something happens (like the geometric).

pdf of an exponential random variable is $f(y) = \lambda\ exp(-\lambda\ y)$, for $y \geq 0$, and $f(y) = 0$ otherwise.
The cdf is $F(y) = 1 - exp(-\lambda\ y)$, for $y \geq 0$.
If $X$ is exponential with parameter $\lambda$, then $E(X) = SD(X) = 1/\lambda$

If the total numbers of events in any disjoint time spans are independent, then these totals are Poisson random variables. If in addition the events are occurring at a constant rate $\lambda$, then the times between events, or *interevent times*, are exponential random variables with mean $1/\lambda$.

**Example.** Suppose you play 20 hands an hour, with each hand lasting exactly 3 minutes, and let $X$ be the time in hours until the end of the first hand in which you are dealt pocket aces. Use the exponential distribution to approximate $P(X \leq 2)$ and compare with the exact solution using the geometric distribution.

**Answer.** Each hand takes 1/20 hours, and the probability of being dealt pocket aces on a particular hand is 1/221, so the rate $\lambda$ = 1 in 221 hands = 1/(221/20) hours ~ 0.0905 per hour.
Using the exponential model, $P(X \leq 2 \text{ hours}) = 1 - exp(-2\lambda) \sim 16.556\%$.
This is an approximation, however, since by assumption *X* is not continuous but must be an integer multiple of 3 minutes.
Let *Y* = the number of hands you play until you are dealt pocket aces. Using the geometric distribution, $P(X \leq 2 \text{ hours}) = P(Y \leq 40 \text{ hands})$
$= 1 - (220/221)^{40} \sim 16.590\%$.

The survivor function for an exponential random variable is particularly simple: $P(X > c) = \int_c^\infty f(y)dy = \int_c^\infty \lambda\ exp(-\lambda\ y)dy = -exp(-\lambda\ y)]_c^\infty = exp(-\lambda\ c)$.

Like geometric random variables, exponential random variables have the *memorylessness* property: if *X* is exponential, then for any non-negative values *a* and *b*, $P(X > a+b \mid X > a) = P(X > b)$. *(See p115).*
Thus, with an exponential or geometric random variable, if after a certain time you still have not observed the event you are waiting for, then the distribution of the *future,* additional waiting time until you observe the event is the same as the distribution of the *unconditional* time to observe the event to begin with.

**6. Normal random variables.**

So far we have seen two continuous random variables, the uniform and the exponential.

Normal. pp 115-117. mean = $\mu$, SD = $\sigma$, $f(y) = 1/\sqrt{(2\pi\sigma^2)}\, e^{-(y-\mu)^2/2\sigma^2}$.

Symmetric around $\mu$,

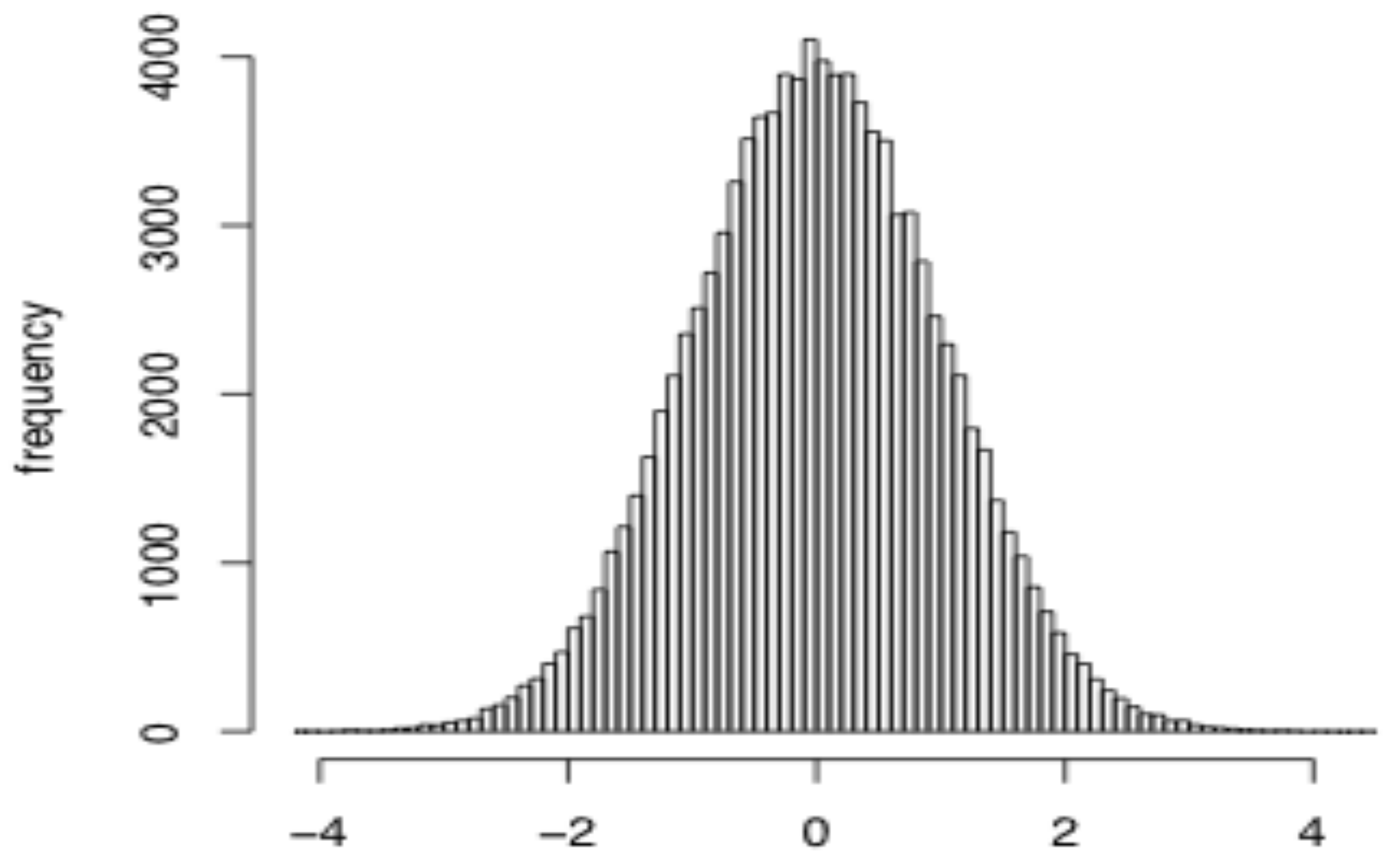
50% of the values are within 0.674 SDs of $\mu$,

68.27% of the values are within 1 SD of $\mu$, and

95% are within 1.96 SDs of $\mu$.

* Standard Normal. Normal with $\mu = 0$, $\sigma = 1$. See pp 117-118.

Standard normal density:
68.27% between -1.0 and 1.0
95% between -1.96 and 1.96



## 7. Pareto random variables, ch6.6

Pareto random variables are an example of *heavy-tailed* random variables, which means they have very, very large outliers much more frequently than other distributions like the normal or exponential.

For a Pareto random variable, the pdf is $f(y) = (b/a)\,(a/y)^{b+1}$, and the cdf is

$F(y) = 1 - (a/y)^b$,

for $y>a$, where $a>0$ is the *lower truncation point*, and $b>0$ is a parameter called the *fractal dimension*.

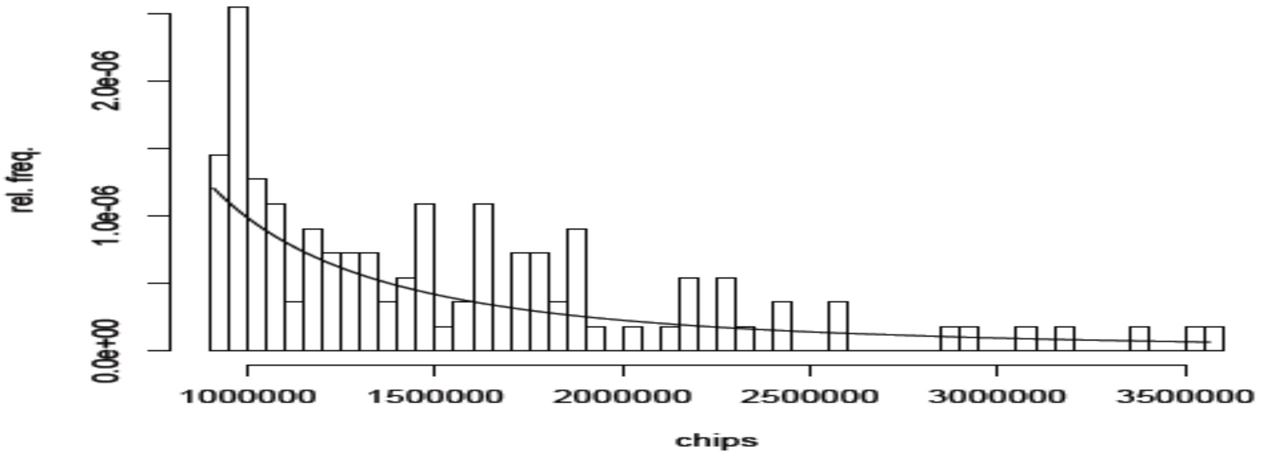



Figure 6.6.1: Relative frequency histogram of the chip counts of the leading 110 players in the 2010 WSOP main event after day 5. The curve is the Pareto density with $a = 900{,}000$ and $b = 1.11$.

**8. Functions of independent random variables.**

If X and Y are independent random variables, then

$E[f(X)\, g(Y)] = E[f(X)]\, E[g(Y)]$, for any functions f and g.

See Exercise 7.12. This is useful for problem 5.4 for instance.

**9. Moment generating functions of some random variables.**

Bernoulli(p). $\phi_X(t) = pe^t + q$.

Binomial(n,p). $\phi_X(t) = (pe^t + q)^n$.

Geometric(p). $\phi_X(t) = pe^t/(1 - qe^t)$.

Neg. binomial (r,p). $\phi_X(t) = [pe^t/(1 – qe^t)]^r$.

Poisson($\lambda$). $\phi_X(t) = e^{\{\lambda e^t - \lambda\}}$.

Uniform (a,b). $\phi_X(t) = (e^{tb} – e^{ta})/[t(b-a)]$.

Exponential ($\lambda$). $\phi_X(t) = \lambda/(\lambda-t)$.

Normal. $\phi_X(t) = e^{\{t\mu + t^2\sigma^2/2\}}$.

## 10. Survivor functions.

Recall the cdf $F(b) = P(X \leq b)$.
The survivor function is $S(b) = P(X > b) = 1 - F(b)$.
Some random variables have really simple survivor functions and it can be convenient to work with them.
If X is geometric, then $S(b) = P(X > b) = q^b$, for $b = 0,1,2,....$
For instance, let b=2. $X > 2$ means the 1st two were misses,
i.e. $P(X>2) = q^2$.
For exponential X, $F(b) = 1 - exp(-\lambda b)$, so $S(b) = exp(-\lambda b)$.

An interesting fact is that, if X takes only values in {0,1,2,3,...},
then $E(X) = S(0) + S(1) + S(2) + ....$
Proof.

$$
\begin{aligned}
S(0) &= P(X=1) + P(X=2) + P(X=3) + P(X=4) + ....\\
S(1) &= \phantom{P(X=1) + {}} P(X=2) + P(X=3) + P(X=4) + ....\\
S(2) &= \phantom{P(X=1) + P(X=2) + {}} P(X=3) + P(X=4) + ....\\
S(3) &= \phantom{P(X=1) + P(X=2) + P(X=3) + {}} P(X=4) + ....
\end{aligned}
$$

Add these up and you get
$0\,P(X=0) + 1P(X=1) + 2P(X=2) + 3P(X=3) + 4P(X=4) + ...$
$= \sum kP(X=k) = E(X)$.

## 11. Covariance and correlation.

For any random variables X and Y,

$$\text{var}(X+Y) = E[(X+Y)]^2 - [E(X) + E(Y)]^2$$
$$= E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 + 2E(XY) - 2E(X)E(Y)$$
$$= \text{var}(X) + \text{var}(Y) + 2[E(XY) - E(X)E(Y)].$$

$\text{cov}(X,Y) = E(XY) - E(X)E(Y)$ is called the *covariance* between X and Y,

$\text{cor}(X,Y) = \text{cov}(X,Y) / [SD(X)\ SD(Y)]$ is called the *correlation* bet. X and Y.

If X and Y are ind., then $E(XY) = E(X)E(Y)$,

so $\text{cov}(X,Y) = 0$, and $\text{var}(X+Y) = \text{var}(X) + \text{var}(Y)$.

Just as $E(aX + b) = aE(X) + b$, for any real numbers a and b,

$\text{cov}(aX + b,Y) = E[(aX+b)Y] - E(aX+b)E(Y)$

$= aE(XY) + bE(Y) - [aE(X)E(Y) + bE(Y)] = a\ \text{cov}(X,Y)$.

Ex. 7.1.3 is worth reading.

X = the # of 1st card, and Y = X if 2nd is red, -X if black.

$E(X)E(Y) = (8)(0)$.

$P(X = 2 \text{ and } Y = 2) = 1/13 * ½ = 1/26$, for instance, and same with any other combination,

so $E(XY) = 1/26\ [(2)(2)+(2)(-2)+(3)(3)+(3)(-3) + ... + (14)(14) + (14)(-14)]$

$= 0$.

So X and Y are *uncorrelated*, i.e. $\text{cor}(X,Y) = 0$.

But X and Y are not independent.

$P(X=2 \text{ and } Y=14) = 0$, but $P(X=2)P(Y=14) = (1/13)(1/26)$.