Stat 100a, Introduction to Probability.

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

1. Hand in hw2.
2. Pareto distribution.
3. Functions of independent rvs.
4. Moment generating functions of rvs.
5. Survivor functions.
6. Covariance and correlation.
7. Bivariate normal.
8. Example problems.

Bring a PEN or PENCIL and CALCULATOR and any books or notes you want to the exams.
The midterm is Tue Nov 19.
HW3 is due Tue Dec3 and is on the course website.
http://www.stat.ucla.edu/~frederic/100a/F19 .

1. Hand in hw2.

2. 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*.

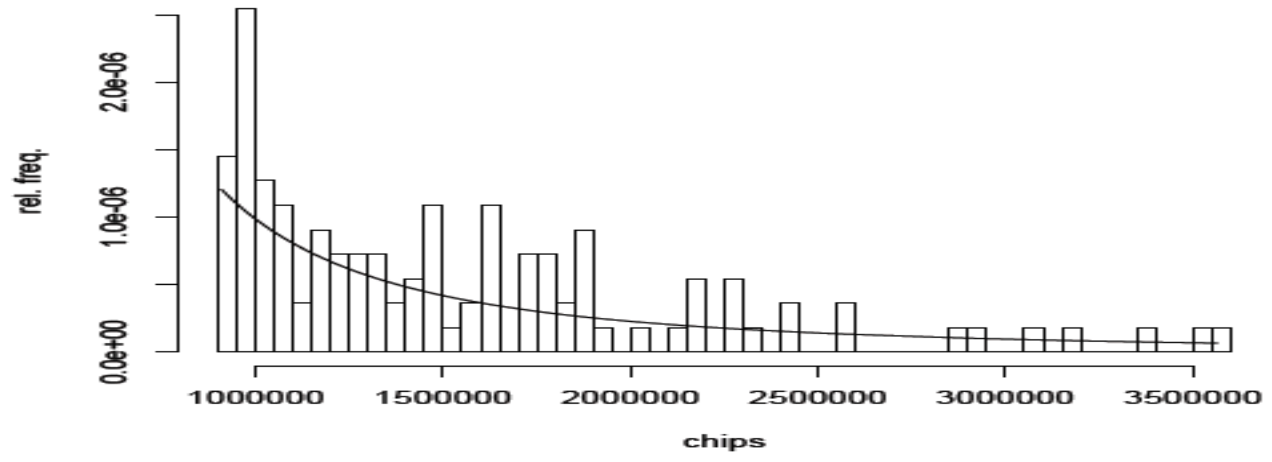


Figure 6.6.1: Relative frequency histogram of the chip counts of the leading 110 players in the 2010 WSOP main event after day 5. The curve is the Pareto density with $a = 900{,}000$ and $b = 1.11$.

3. 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.

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

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

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

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

5. 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 $b = 0,1,2,\ldots$.

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 $\{0,1,2,3,\ldots\}$,

then $E(X) = S(0) + S(1) + S(2) + \ldots$.

Proof.

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

Add these up and you get

$0\,P(X=0) + 1P(X=1) + 2P(X=2) + 3P(X=3) + 4P(X=4) + \ldots$

$= \sum kP(X=k) = E(X)$.

6. Covariance and correlation.

For any random variables X and Y,

$var(X+Y) = E[(X+Y)^2] – [E(X) + E(Y)]^2$

$= E(X^2) – [E(X)]^2 + E(Y^2) – [E(Y)]^2 + 2E(XY) – 2E(X)E(Y)$

$= var(X) + var(Y) + 2[E(XY) – E(X)E(Y)].$

cov(X,Y) = E(XY) – E(X)E(Y) is called the *covariance* between X and Y.

$cov(X,X) = E(X^2) – [E(X)]^2 = var(X).$

cor(X,Y) = 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 cov(X,Y) = 0, and in this circumstance var(X+Y) = var(X) + var(Y).

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

cov(aX + b,Y) = E[(aX+b)Y] – E(aX+b)E(Y)

= aE(XY) + bE(Y) – [aE(X)E(Y) + bE(Y)] = a cov(X,Y).

For rvs W,X,Y, and Z, cov(W+X, Y+Z) = cov(W,Y) + cov(W,Z) + cov(X,Y) + cov(X,Z).

Why? cov(W+X,Y+Z) = E(WY+WZ+XY+XZ) – E(W+X)E(Y+Z)

= E(WY+WZ+XY+XZ) – (E(W)+E(X))(E(Y)+E(Z))

= E(WY) + E(WZ) + E(XY) – E(XZ) – E(W)E(Y) – E(W)E(Z) – E(X)E(Y)-E(X)E(Z).

Note cov(X,Y) = cov(Y,X) and same for correlation.

Covariance and correlation.

Ex. 7.1.3 is worth reading.

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

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

P(X = 2 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. cor(X,Y) = 0.

But X and Y are not independent.

P(X=2 and Y=14) = 0, but P(X=2)P(Y=14) = (1/13)(1/26).

7. Bivariate normal.

$X \sim N(0,1)$ means X is normal with mean 0 and variance 1.

If $X \sim N(0,1)$ and $Y = a + bX$, then Y is normal with mean a and variance b^2.

Suppose X is normal, and $Y|X$ is normal. Then (X,Y) are *bivariate normal*.

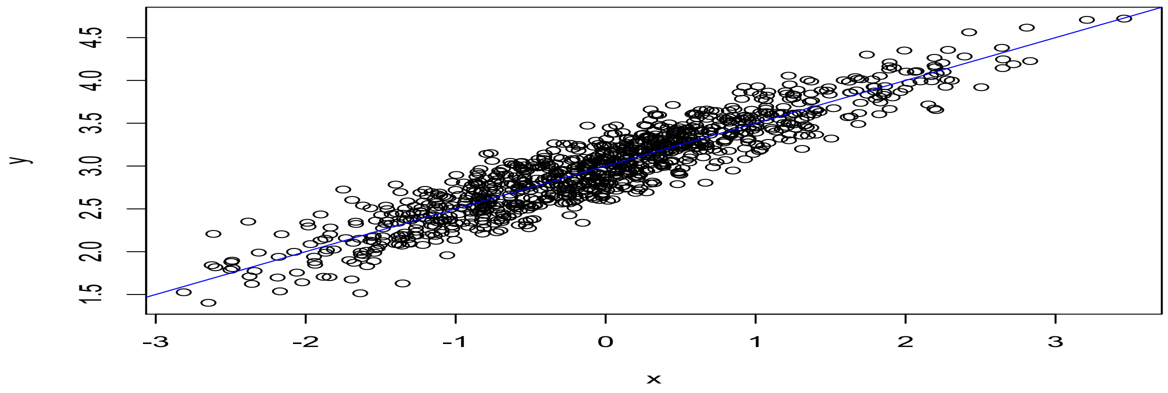
For example, let $X = N(0,1)$. Let $\varepsilon = N(0, 0.2^2)$, ε independent of X.

Let $Y = 3 + 0.5 X + \varepsilon$.

Then (X,Y) are bivariate normal.

$Y|X = (3+0.5X) + \varepsilon$ which is normal since ε is normal.

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



Bivariate normal.

For example, let $X = N(0,1)$. Let $\varepsilon = N(0, 0.2^2)$ and independent of X. Let $Y = 3 + 0.5\, X + \varepsilon$.

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

$E(X) = 0$.

$E(Y) = E(3 + 0.5X + \varepsilon) = 3 + 0.5\, E(X) + E(\varepsilon) = 3$.

Given X, $E(Y|X) = E(3 + 0.5X + \varepsilon \mid X) = 3 + 0.5\, X$. We will discuss this more later.

$\mathrm{var}(X) = 1$.

$\mathrm{var}(Y) = \mathrm{var}(3 + 0.5\, X + \varepsilon) = \mathrm{var}(0.5X + \varepsilon) = 0.5^2\, \mathrm{var}(X) + \mathrm{var}(\varepsilon) = 0.5^2 + 0.2^2 = 0.29$.

$\mathrm{cov}(X,Y) = \mathrm{cov}(X, 3 + 0.5X + \varepsilon) = 0.5\, \mathrm{var}(X) + \mathrm{cov}(X, \varepsilon) = 0.5 + 0 = 0.5$.

$\rho = \mathrm{cov}(X,Y)/(\mathrm{sd}(X)\, \mathrm{sd}(Y)) = 0.5 / (1 \text{ x } \sqrt{.29}) = 0.928$.

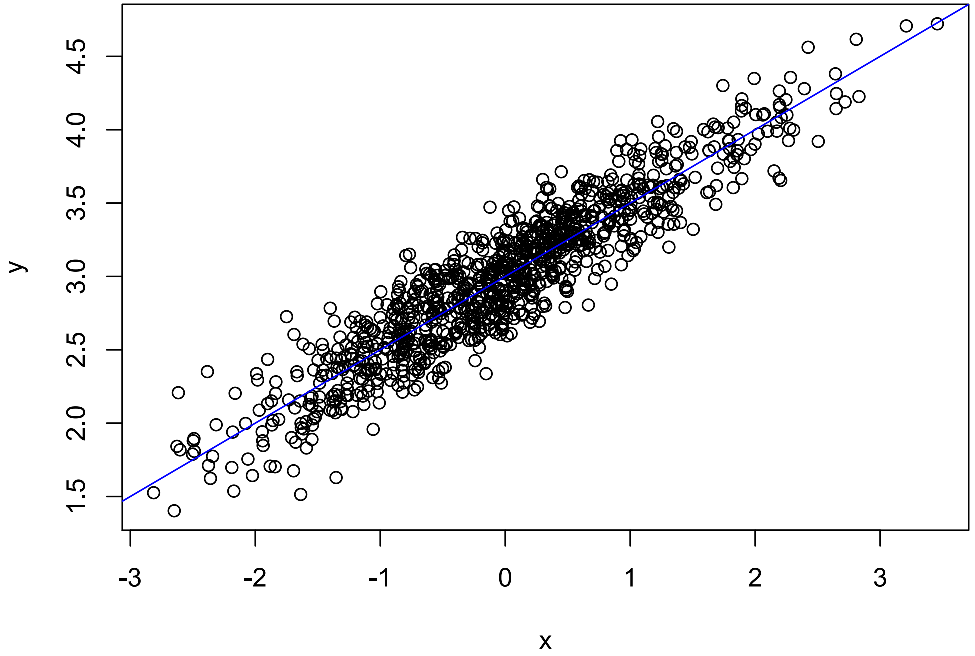
In general, if (X,Y) are bivariate normal, can write $Y = \beta_1 + \beta_2 X + \varepsilon$, where $E(\varepsilon) = 0$, and ε is normal and ind. of X. Following the same logic, $\rho = \mathrm{cov}(X,Y)/(\sigma_x\sigma_y) = \beta_2\, \mathrm{var}(X)/(\sigma_x\sigma_y) = \beta_2\, \sigma_x/\sigma_y$, so $\rho = \beta_2\, \sigma_x/\sigma_y$, and $\beta_2 = \rho\, \sigma_y/\sigma_x$.

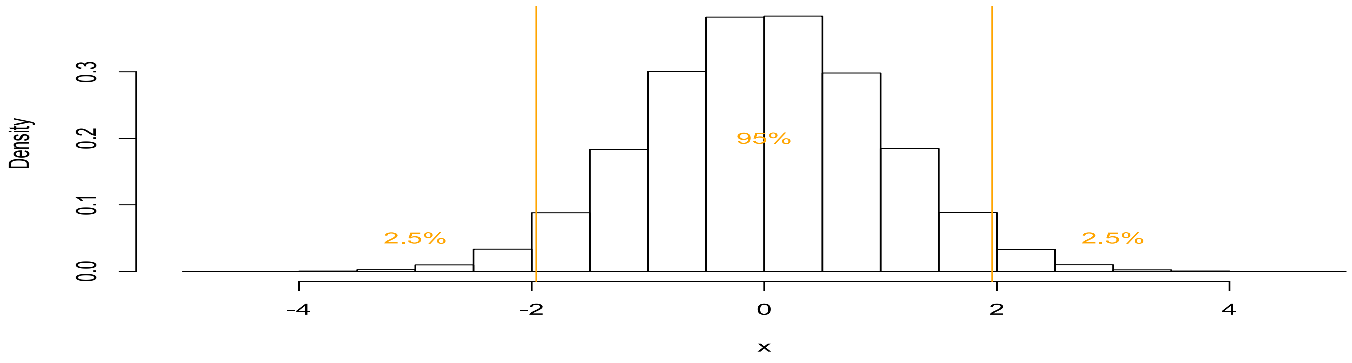
Bivariate normal.

For example, let X = N(0,1). Let ε = N(0, 0.2^2) and independent of X. Let Y = 3 + 0.5 X + ε.

In R,

```
x = rnorm(1000,mean=0,sd=1)
eps = rnorm(1000,mean=0,sd=.2)
y = 3 + .5*x+eps
plot(x,y)
cor(x,y) # 0.9282692.
```





Bivariate normal.

If (X,Y) are bivariate normal with $E(X) = 100$, $var(X) = 25$, $E(Y) = 200$, $var(Y) = 49$, $\rho = 0.8$,

What is the distribution of Y given X = 105? What is $P(Y > 213.83 \mid X = 105)$?

Given X = 105, Y is normal. Write $Y = \beta_1 + \beta_2 X + \varepsilon$ where ε is normal with mean 0, ind. of X.

Recall $\beta_2 = \rho\, \sigma_y/\sigma_x = 0.8 \times 7/5 = 1.12$.

So $Y = \beta_1 + 1.12\, X + \varepsilon$.

To get β_1, note $200 = E(Y) = \beta_1 + 1.12\, E(X) + E(\varepsilon) = \beta_1 + 1.12\, (100)$. So $200 = \beta_1 + 112$. $\beta_1 = 88$.

So $Y = 88 + 1.12\, X + \varepsilon$, where ε is normal with mean 0 and ind. of X.

What is $var(\varepsilon)$?

$49 = var(Y) = var(88 + 1.12\, X + \varepsilon) = 1.12^2\, var(X) + var(\varepsilon) + 2(1.12)\, cov(X,\varepsilon)$

$= 1.12^2\, (25) + var(\varepsilon) + 0$. So $var(\varepsilon) = 49 - 1.12^2\, (25) = 17.64$ and $sd(\varepsilon) = \sqrt{17.64} = 4.2$.

So $Y = 88 + 1.12\, X + \varepsilon$, where ε is $N(0, 4.2^2)$ and ind. of X.

Given X = 105, $Y = 88 + 1.12(105) + \varepsilon = 205.6 + \varepsilon$, so $Y|X=105 \sim N(205.6, 4.2^2)$.

Now how many sds above the mean is 213.83? $(213.83 – 205.6)/4.2 = 1.96$,

so $P(Y>213.83 \mid X=105)$ = P(normal is > 1.96 sds above its mean) = 2.5%.

8. Example problems.

X is a continuous random variable with cdf $F(y) = 1 - y^{-1}$, for y in $(1,\infty)$, and $F(y) = 0$ otherwise.
a. What is the pdf of X?
b. What is f(1)?
c. What is E(X)?

a. $f(y) = F'(y) = d/dy\ (1 - y^{-1}) = y^{-2}$, for y in $(1,\infty)$, and $f(y) = 0$ otherwise.

To be a pdf, f(y) must be nonnegative for all y and integrate to 1.
$f(y) \geq 0$ for all y, and $\int_{-\infty}^{\infty} f(y)dy = \int_1^{\infty} y^{-2}\, dy = -y^{-1}]_1^{\infty} = 0 + 1 = 1$. So, f is indeed a pdf.

b. $f(1) = 1^{-2} = 1$.

c. $E(X) = \int_{-\infty}^{\infty} y\, f(y)dy = \int_1^{\infty} y\, y^{-2}\, dy = \int_1^{\infty} y^{-1}\, dy = \ln(\infty) - \ln(1) = \infty$.

X is a continuous random variable with cdf $F(y) = 1 - y^{-2}$, for y in $(1,\infty)$, and $F(y) = 0$ otherwise.
a. What is the pdf of X?
b. What is f(1)? Is this a problem?
c. What is E(X)?
d. What is $P(2 \leq X \leq 3)$?
e. What is $P(2 < X < 3)$?

a. $f(y) = F'(y) = d/dy\ (1 - y^{-2}) = 2y^{-3}$, for y in $(1,\infty)$, and $f(y) = 0$ otherwise.

To be a pdf, f(y) must be nonnegative for all y and integrate to 1.
$f(y) \geq 0$ for all y, and $\int_{-\infty}^{\infty} f(y)dy = \int_1^{\infty} 2y^{-3}\, dy = -y^{-2}]_1^{\infty} = 0 + 1 = 1$. So, f is indeed a pdf.

b. $f(1) = 2$. This does not mean $P(X=1)$ is 2. It is not a problem.

c. $E(X) = \int_{-\infty}^{\infty} y\, f(y)dy = 2\int_1^{\infty} y\, y^{-3}\, dy = 2\int_1^{\infty} y^{-2}\, dy = -2y^{-1}]_1^{\infty} = 0 + 2 = 2.$

d. $P(2 \leq X \leq 3) = \int_2^3 f(y)dy = 2\int_2^3 y^{-3}\, dy = -y^{-2}\,]_2^3 = -1/9 + 1/4 \sim 0.139.$

Alternatively, $P(2 \leq X \leq 3) = F(3) - F(2) = 1 - 1/9 - 1 + 1/4 \sim 0.139.$
e. Same thing.

Suppose X is uniform(0,1), Y is exponential with E(Y)=2, and X and Y are independent. What is cov(3X+Y, 4X-Y)?

cov(3X+Y, 4X-Y) = 12 cov(X,X) – 3cov(X,Y) + 4cov(Y,X) – cov(Y,Y)
= 12 var(X) – 0 + 0 – var(Y).

For exponential, E(Y) = $1/\lambda$ and var(Y) = $1/\lambda^2$, so $\lambda=1/2$ and var(Y) = 4.
What about var(X)?
$E(X^2) = \int y^2 f(y)dy$
$= \int_0^1 y^2 dy$ because f(y) = 1 for uniform(0,1) for y in (0,1),
$= y^3/3 \,]_0^1$
$= 1/3.$
var(X) = $E(X^2) - \mu^2 = 1/3 - ¼ = 1/12$.

cov(3X+Y, 4X-Y) = 12 (1/12) – 4
= -3.