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

1. Uniform random variables, continued.
2. Covariance and correlation.
3. Bivariate normal.

Homework 2 is due today, Mon Feb14, 2pm. Email to STAT100AW22@stat.ucla.edu.
No class or OH Feb 21 President's Day.
Exam 2 will be on Wed Feb23, 2pm-3:15pm.
The computer project is due on Sat Mar5, 8:00pm.
Read through chapter 7.2. ♠ ♣ ♥ ♦

Leave all answers as decimals, not fractions, for all homeworks.
http://www.stat.ucla.edu/~frederic/100A/W22

If X is a uniform (0,1) random variable, a) why is $P(X > c) = 1-c$, for $0 \le c \le 1$? b) What is $E(X)$? c) What is $V(X)$?

Remember, for uniform(a,b), the pdf is $f(y) = 1/(b-a)$ for $a<y<b$ and the pdf is 0 otherwise. Here $a = 0$, $b = 1$, so $f(y) = 1$ for $0<y<1$.

a. $P(X > c) = P\{X \text{ is in } (c,\infty)\} = \int_c^\infty f(y)dy = \int_c^1 1\, dy = 1-c.$

b. $E(X) = \int_{-\infty}^\infty y\, f(y)dy = \int_0^1 y\, f(y)dy = y^2/2\,]_0^1 = ½ - 0 = ½.$

c. $E(X^2) = \int_{-\infty}^\infty y^2\, f(y)dy = \int_0^1 y^2\, f(y)dy = y^3/3\,]_0^1 = 1/3.$
$V(X) = E(X^2) - \mu^2 = 1/3 – (1/2)^2 = 1/3 – 1/4 = 1/12.$

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,

$\text{so } E(XY) = 1/26\ [(2)(2)+(2)(-2)+(3)(3)+(3)(-3) + \ldots + (14)(14) + (14)(-14)]$

$= 0$.

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 $P(X=2)P(Y=14) = (1/13)(1/26)$.

Covariance and correlation.

For any random variables X and Y,

$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X,Y)$.

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

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

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

$\text{cov}(aX + b, Y) = a\, \text{cov}(X,Y)$.

$\text{var}(aX+b) = \text{cov}(aX+b, aX+b) = a^2\text{var}(X)$.

No such simple statement is true for correlation.

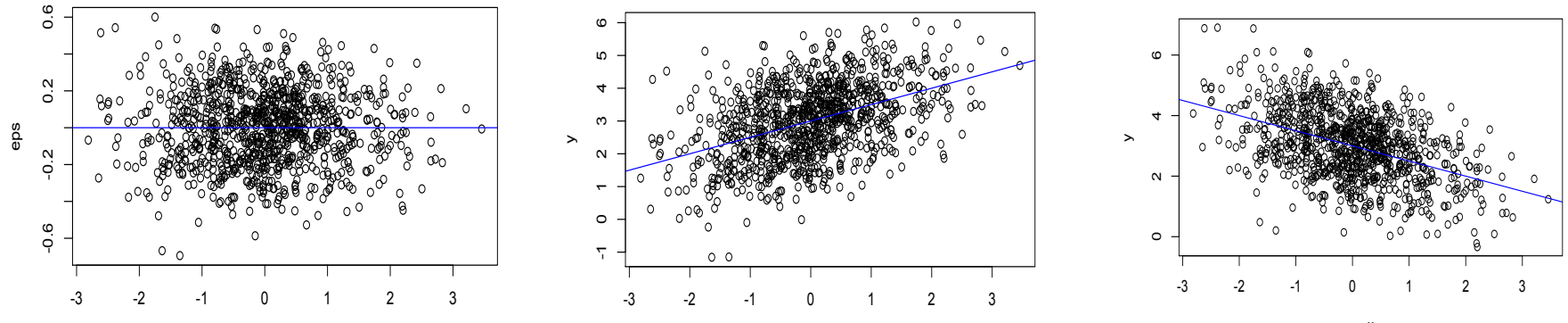
If $\rho = \text{cor}(X,Y)$, we always have $-1 \le \rho \le 1$.

$\rho = -1$ iff. the points (X,Y) all fall exactly on a line sloping downward, and

$\rho = 1$ iff. the points (X,Y) all fall exactly on a line sloping upward.

$\rho = 0$ means the best fitting line to (X,Y) is horizontal.

$\rho = 0$ $\rho = 0.44$ $\rho = -0.44$.



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*.

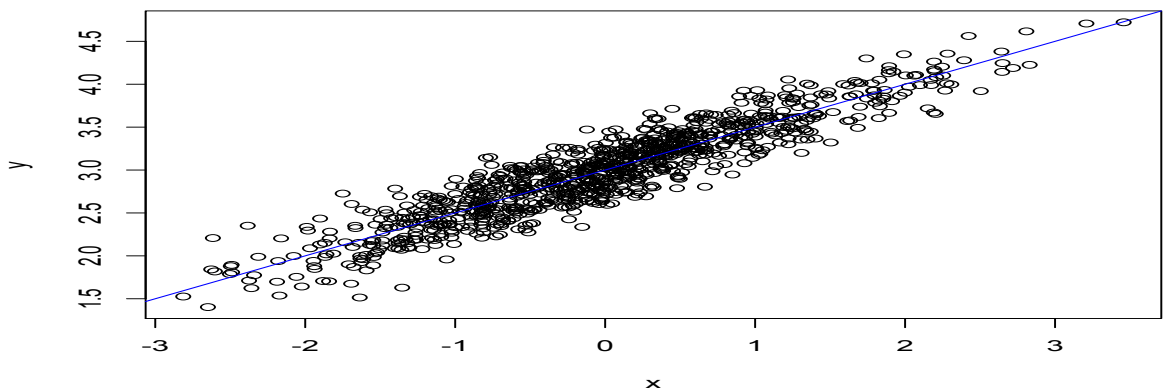
For example, let $X = N(0,1)$. Let $\varepsilon = N(0, 0.2^2)$, ε independent of X.

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

Then (X,Y) are bivariate normal.

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

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

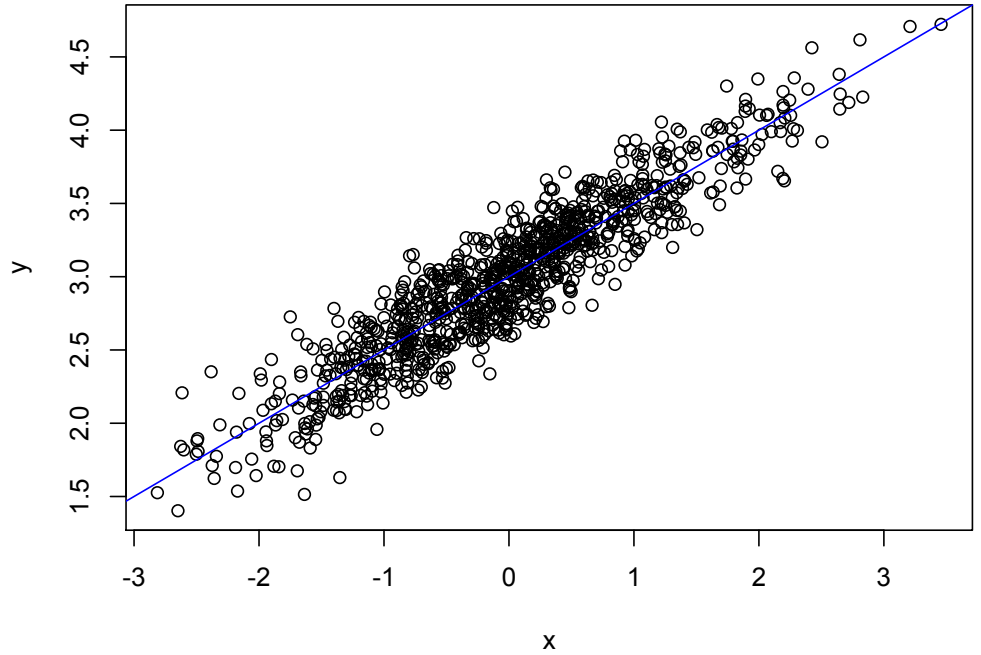


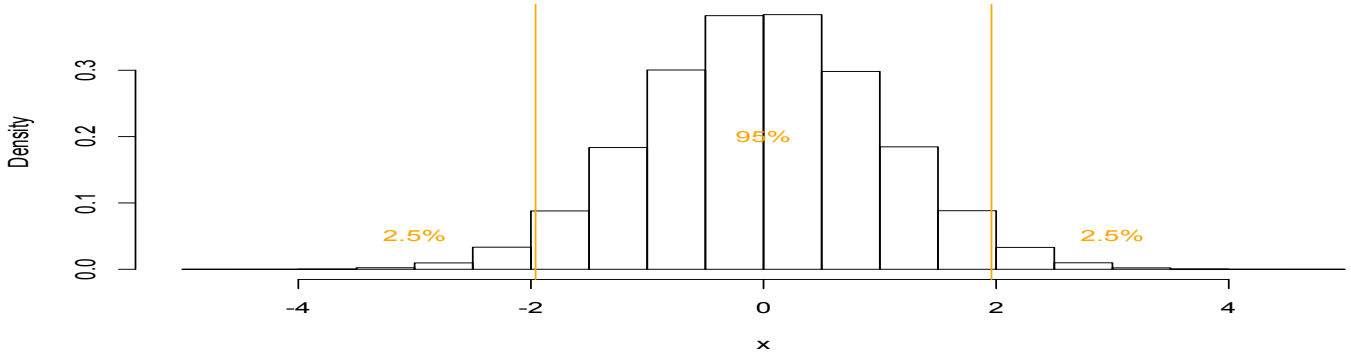
Bivariate normal.

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

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.

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 $\rho = cor(X,Y)$.

$E(X) = 0$.

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

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

$var(X) = 1$.

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

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

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

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, $\rho = cov(X,Y)/(\sigma_x\sigma_y) = \beta_2\ var(X)/(\sigma_x\sigma_y) = \beta_2\ \sigma_x/\sigma_y$, so $\rho = \beta_2\ \sigma_x/\sigma_y$, and $\beta_2 = \rho\ \sigma_y/\sigma_x$.

Bivariate normal.

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 | 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(ε)?

$49 = \text{var}(Y) = \text{var}(88 + 1.12\, X + \varepsilon) = 1.12^2\, \text{var}(X) + \text{var}(\varepsilon) + 2(1.12)\, \text{cov}(X,\varepsilon)$
$= 1.12^2\,(25) + \text{var}(\varepsilon) + 0$. So $\text{var}(\varepsilon) = 49 - 1.12^2\,(25) = 17.64$ and $\text{sd}(\varepsilon) = \sqrt{17.64} = 4.2$.

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 | X=105) = P(normal is > 1.96 sds above its mean) = 2.5%.

Bivariate normal.

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

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