Stat 134 lec 38

last time

Sec 6.4 Correlation

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{SD(X)\,SD(Y)} = E(X^* Y^*)$$

(X^*: X in std units)

$-1 \le \text{Corr}(X,Y) \le 1$

$$\text{Corr}(aX+b, cY+d) = \pm \text{Corr}(X,Y) \quad a,b,c,d \in \mathbb{R},\ a,c \neq 0.$$

($+$ if $ac > 0$)

$N_1 + \cdots + N_k = C$ a sum of exchangeable RVs

$$\text{Corr}(N_1, N_2) = -\frac{1}{k-1}$$

Today

1. review student comments from last time.
2. sec 6.5 the bivariate normal distribution

(1) Review of student comments

Stat 134

Friday April 26 2019

1. An urn contains 90 marbles, of which there are 20 green, 25 black, and 45 red marbles. Alice randomly picks 10 marbles without replacement, and Bob randomly picks another 10 marbles without replacement. Let X_1 be the number of green marbles that Alice has and X_2 the number of green marbles that Bob has.
   To find $Corr(X_1, X_2)$ is

$$X_1 + X_2 + \cdots + X_9 = 20$$

   a true identity that is useful? Explain.

   a yes

   b no

   c not enough info to decide

Discuss with your neighbor for 1 minute how you did this.

b	The xi are not identically distributed as the distribution of xi depends on the values of xj j<i

Not true.
Imagine example of deck of 52 cards.
What is the chance there are 4 hearts in Cards 14-26?
answ $\frac{\binom{13}{4}\binom{39}{9}}{\binom{52}{13}}$

a	The Xns are identical, each being hypergeometric(90, 10, 20).

a	Assign the marbles in lots of 10, this makes X1 through X9 identically distributed (but not remotely independent), so we can use the trick in the guided example to find the correlation

Pull marbles M11 – M20 out without looking at other marbles.
$X_2 \sim$ Hyper(90, 10, 20)

a	After 9 turns, 90 marbles have been picked and those 90 marbles must contain all of the 20 green marbles since the marbles are picked without replacement

$X_1, \ldots, X_9$ i.d.
and $\sum$ equal to 20
so
$\text{Corr}(X_1, X_2) = -\frac{1}{9-1}$
$= -\frac{1}{8}$

② sec 6.5 Bivariate normal.

- A generalization of joint distribution of 2 iid normal to 2 correlated normals.
- An introduction to linear regression. We can learn how to predict final exam score from midterm score.

let X, Z iid $N(0,1)$, $-1 \le \rho \le 1$

$$Y = \rho X + \sqrt{1-\rho^2}\, Z$$

Y is normal being a linear combination of independent normals.

$$E(Y) = \rho \underbrace{E(X)}_{0} + \sqrt{1-\rho^2}\, \underbrace{E(Z)}_{0} = 0$$

$$\mathrm{Var}(Y) = \rho^2 \underbrace{\mathrm{Var}(X)}_{1} + (1-\rho^2) \underbrace{\mathrm{Var}(Z)}_{1} = 1$$

$$\Rightarrow Y \sim N(0,1)$$

Find $\mathrm{Corr}(X,Y)$

$$= \frac{\mathrm{Cov}(X,Y)}{SD(X)\,SD(Y)}$$

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, \rho X + \sqrt{1-\rho^2}\, Z)$$

$$= \mathrm{Cov}(X, \rho X) + \mathrm{Cov}(X, \sqrt{1-\rho^2}\, Z)$$

$$\underbrace{\mathrm{Cov}(X, \rho X)}_{\rho \mathrm{Var}(X)} = \rho \underbrace{\mathrm{Var}(X)}_{1} \qquad \underbrace{\mathrm{Cov}(X, \sqrt{1-\rho^2}\, Z)}_{0}$$

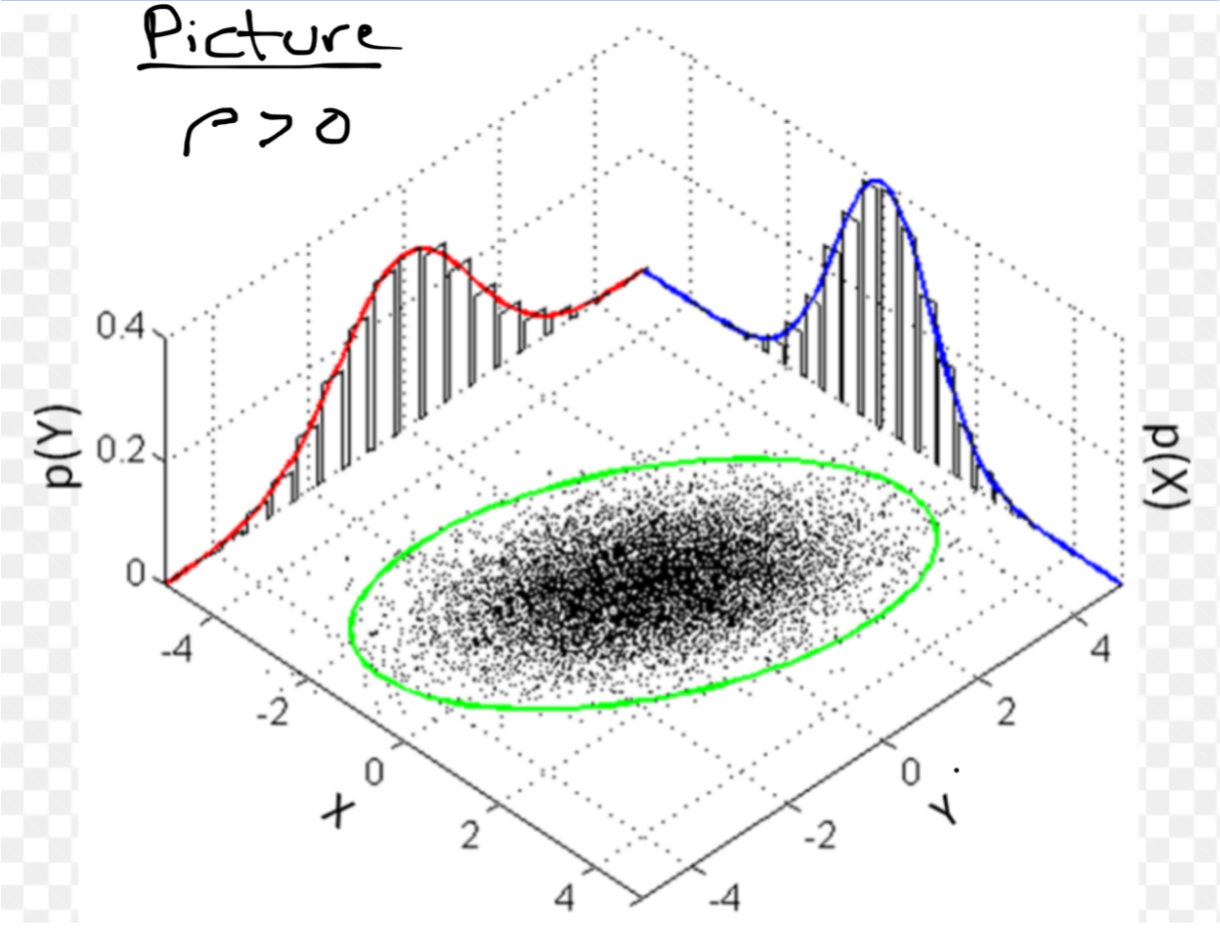
$$\mathrm{Corr}(X,Y) = \frac{\rho \mathrm{Var}(X)}{\underbrace{SD(X)}_{1}\,\underbrace{SD(Y)}_{1}} = \boxed{\rho}$$

Def^n (Standard Bivariate normal distribution)

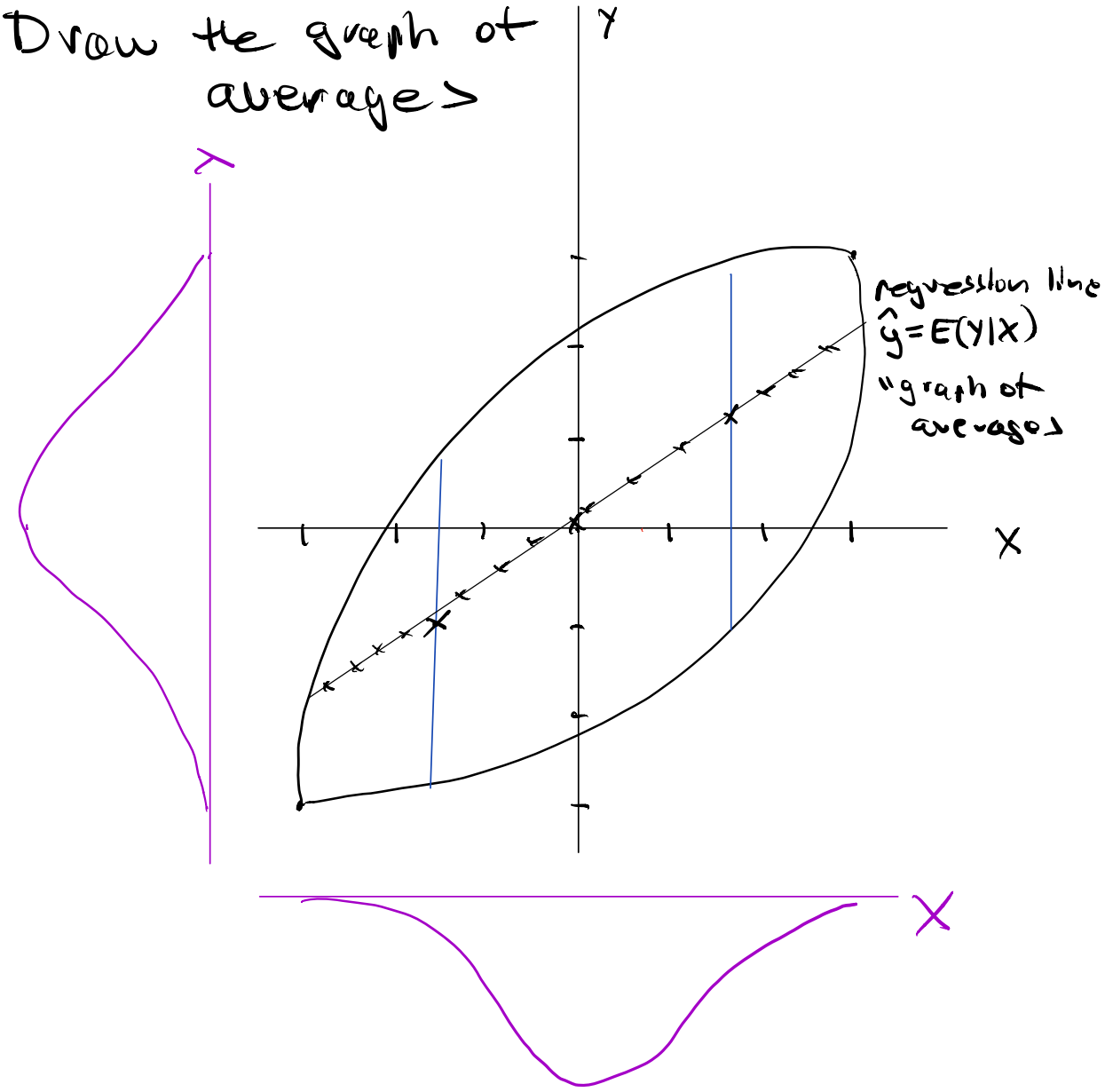
Let X, Z iid $N(0,1)$, $-1 \le \rho \le 1$

$$Y = \rho X + \sqrt{1-\rho^2}\, Z$$

We call the joint distribution (X, Y) the Standard bivariate normal with $\text{corr}(X, Y) = \rho$



Draw the graph of averages



$$y = \rho x + \sqrt{1-\rho^2}\, z$$

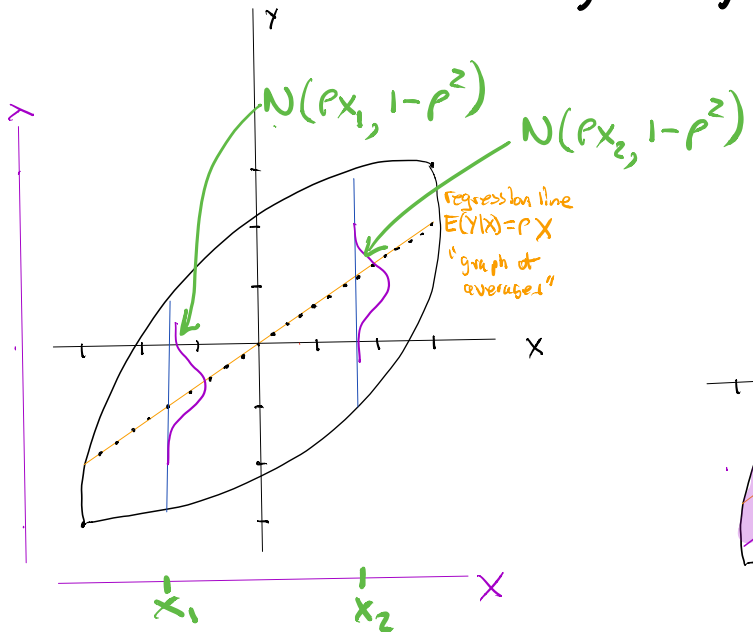
Find the "regression line" $E(y|x)$

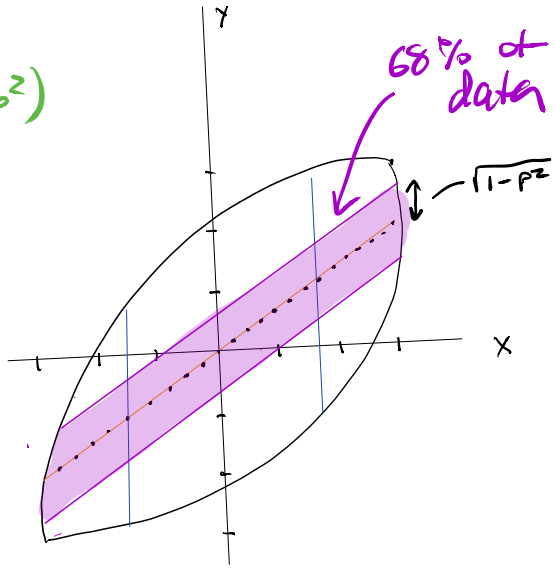
$$\begin{aligned} E(y|x) &= E\left(\rho x + \sqrt{1-\rho^2}\, z \,|\, x\right) \\ &= E(\rho x|x) + E\left(\sqrt{1-\rho^2}\, z|x\right) \\ &= \rho x + \sqrt{1-\rho^2}\, \underbrace{E(z)}_{0} \\ &= \rho x \end{aligned}$$

Find the conditional variance $\mathrm{Var}(y|x)$

$$\begin{aligned} \mathrm{Var}(y|x) &= \mathrm{Var}\left(\rho x + \sqrt{1-\rho^2}\, z \,|\, x\right) \\ &= \mathrm{Var}(\rho x|x) + \mathrm{Var}\left(\sqrt{1-\rho^2}\, z|x\right) \quad \text{since } x, z \text{ indep.} \\ &= \rho^2 \underbrace{\mathrm{Var}(x|x)}_{0} + (1-\rho^2)\underbrace{\mathrm{Var}(z|x)}_{\mathrm{Var}(z)=1} \quad \text{"} = \text{" since } x, z \text{ indep} \\ &= 1-\rho^2 \end{aligned}$$

Hence $\boxed{y|x \sim N(\rho x, 1-\rho^2)}$





Def^n (Bivariate Normal Distribution)

Random variables U and V have bivariate normal distribution with parameters $\mu_U, \mu_V, \sigma_U^2, \sigma_V^2, \rho$ iff the standardized variables

$$X = \frac{U-\mu_U}{\sigma_U}$$

$$Y = \frac{V-\mu_V}{\sigma_V}$$

have std. bivariate normal distribution with corr ρ.

Then $\rho = \text{Corr}(X,Y) = \text{Corr}(U,V)$.

regression line of bivariate normal distribution

$$\frac{V-\mu_V}{\sigma_V} = \rho \frac{U-\mu_U}{\sigma_U}$$ (y, x) — regression line in S.U.

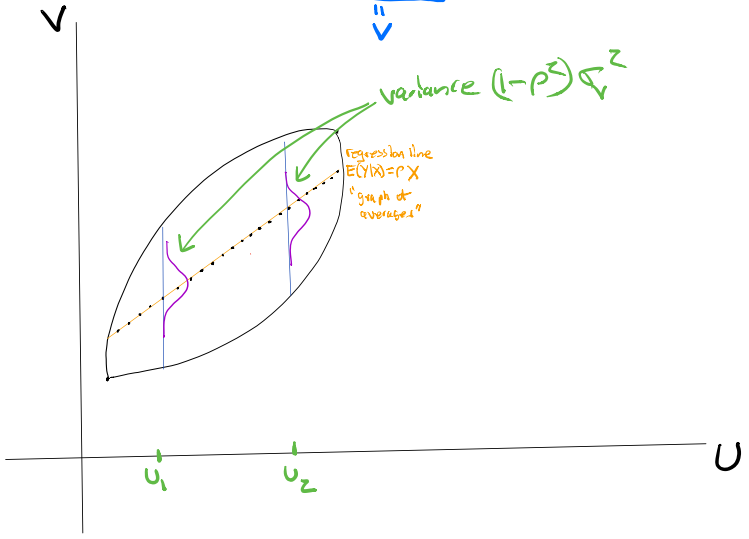
$$\Leftrightarrow \quad V-\mu_V = \frac{\sigma_V}{\sigma_U}\rho(U-\mu_U)$$

$$\Leftrightarrow \quad \hat{V} = \underbrace{\left(\frac{\sigma_V}{\sigma_U}\rho\right)}_{m} U + \underbrace{\mu_V - \frac{\sigma_V}{\sigma_U}\rho\mu_U}_{b}$$ regression line.

Facts

$\hat{V} = E(V|U) = \frac{\sigma_V}{\sigma_U}\rho U + \mu_V - \frac{\sigma_V}{\sigma_U}\rho\mu_U$ ← regression line

and $\text{var}(V|U) = \text{var}\left(\underbrace{\sigma_V y + \mu_V}_{=V} | X\right) = \sigma_V^2 \text{var}(y|x) = (1-\rho^2)\sigma_V^2$



Test 1 is $\mu_U = 60$, $\sigma_U = 20$

Test 2 is $\mu_V = 60$, $\sigma_V = 20$

$\rho = .6$

a) Find the regression line

$$\hat{V} = E(V|U) = \frac{\sigma_V}{\sigma_U}\rho U + \mu_V - \frac{\sigma_V}{\sigma_U}\rho\mu_U \quad \leftarrow \text{regression line}$$

$$= \frac{20}{20}(.6)U + 60 - \frac{20}{20}(.6)(60)$$

$$\boxed{\hat{V} = .6U + 24}$$

b) If you get a 70 on Test 1 what score do you predict to get on test 2?

$$\hat{V} = .6(70) + 24 = \boxed{66}$$

Notice you did relatively worse. Your test 1 score was $\frac{10}{20}$ sd above average but your test 2 score was only $\frac{6}{20}$ sd above average.

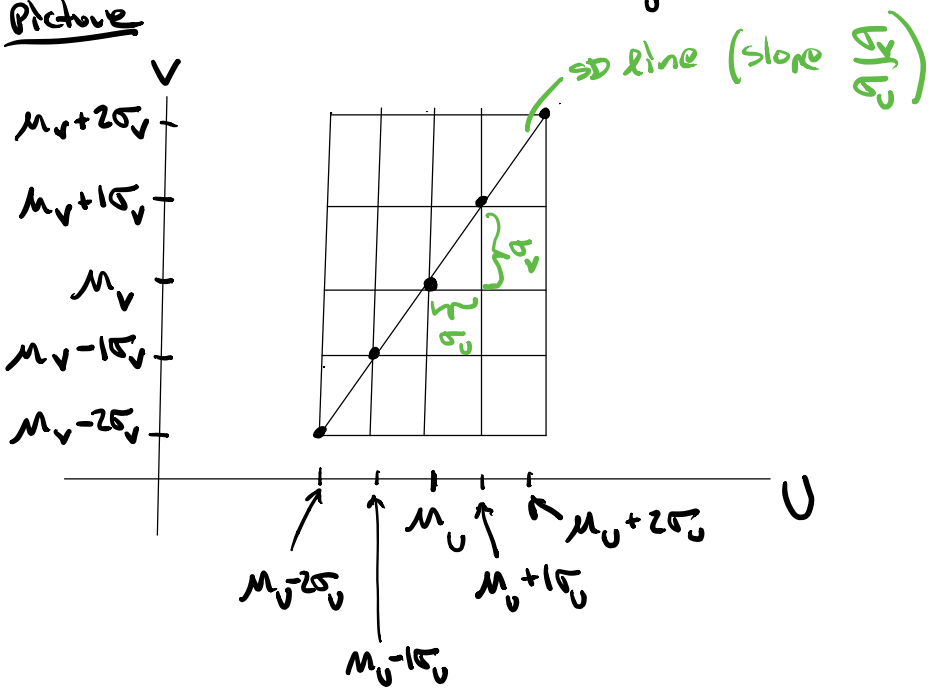
This is the "regression effect".

(3)

Regression line vs. SD line and regression effect

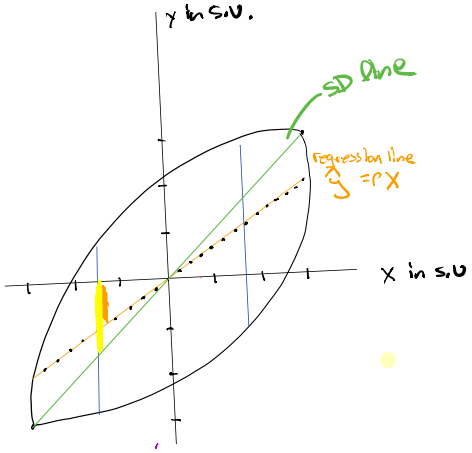
Defn The SD line is $V - \mu_V = \frac{\sigma_V}{\sigma_U}(U - \mu_U)$.

Picture



For U, V in s.u. the SD line is $y = 1 \cdot x$

V^* U^*



↑ steeper than regression line which has slope ρ.

Regression effect

The regression line has slope $-1 \le \rho \le 1$ compared with the SD line which has slope 1. For a fixed x you predict $y = \rho x$ which will be less than x if $x > 0$ and greater than x if $x < 0$. This means that if you do really well on a midterm (at least greater than 50th percentile — so $x > 0$ in s.u.) then you won't do as well on the final relative to the class (i.e you won't go all the way up to the SD line. The opposite is true however if you do poorly on the midterm.

Picture