Stat 134 Lec 38

Last time

Sec 6.5 Bivariate Normal

Let $X, Z \overset{iid}{\sim} N(0,1)$

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

defn (X,Y) is std bivariate normal

Properties

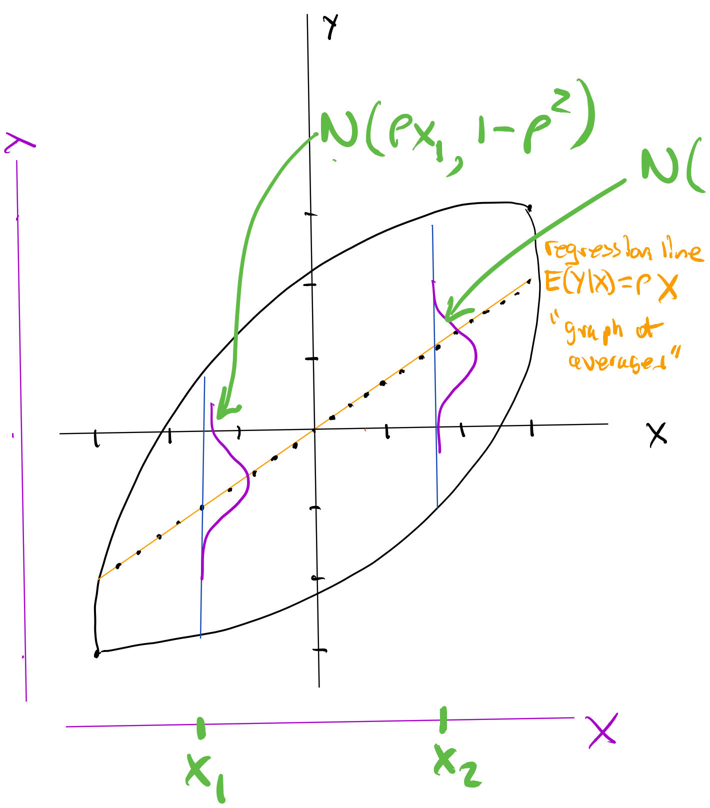
① $Y \sim N(0,1)$

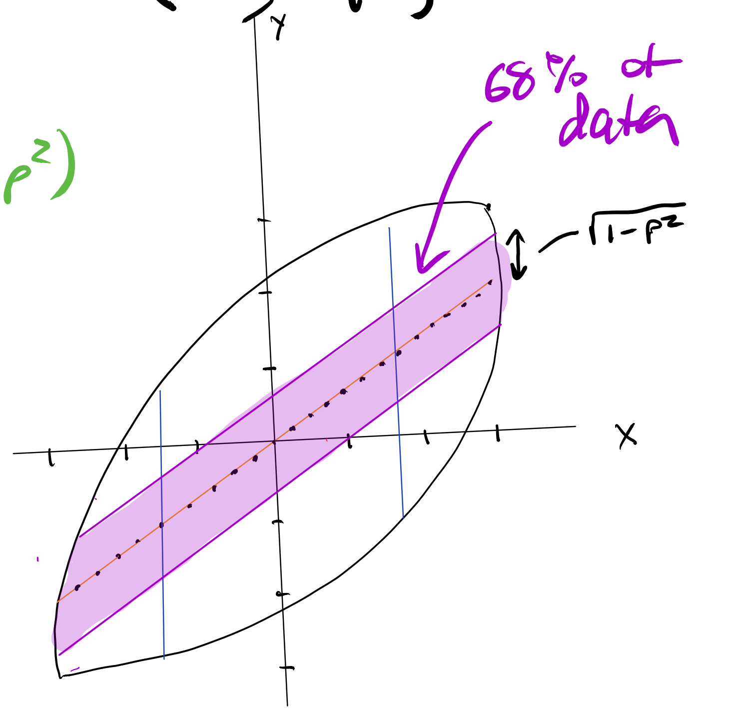
② $Corr(X,Y) = \rho$

③ $Y \mid X=x \; = \rho x + \sqrt{1-\rho^2}\, Z \sim N(\rho x, 1-\rho^2)$

$E(Y \mid X=x) = \rho x$

$Var(Y \mid X=x) = 1-\rho^2$





$E(Y \mid X=x)$

Regression line in s.u. is $\hat{y} = \rho x$

Today Sec 6.5

① Student explanation concept test last time

② Practice with bivariate normal

③ Regression effect.

④ More practice.

Next time Finish sec 6.5 and start review.

1

Stat 134

Wednesday Nov 28 2018

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

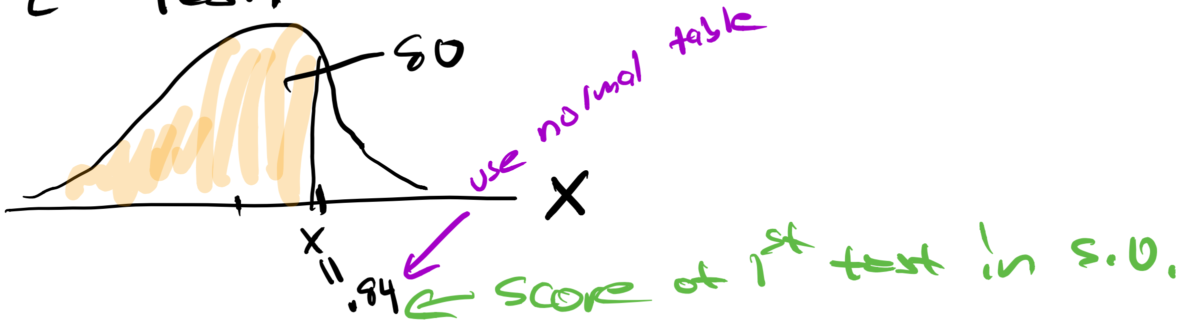
   c not enough info to decide

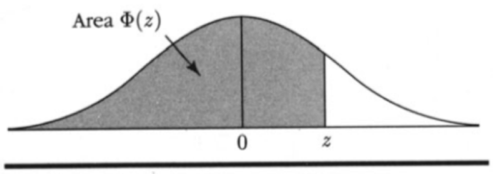
Discuss with neighbor how you did this.

a	-1/(9-1) = -1/8	
b	The variables are not interchangeable. That is, that the second and third distributions depend on the previous ones	
a		
a	The Xs are exchangeable and 9 draws will form a census of the population. So the variance of the sum will be 0 and solving for the corr will be easier.	
a		
a		
a	Because they draw without replacement, the total number of all green marbles that can be drawn is 20. Since all 90 marbles are drawn in X1...X9, we know all the green marbles are drawn in these 9 draws, and draws are exchangeable with considered unconditionally. Thus the correlation is -(1/8). (The derivation of the correlation without the identity uses the method of indicators.)	
a		
a	It's a sum of exchangeable random variables, because you don't condition on first draw. Sum has to be 20.	
a		
b		

2 Practice with Bivariate normal

ex Two test scores are bivariate normal $\rho = .6$. Given that the 1st test is at the 80th percentile, predict the percentile of the 2nd test.





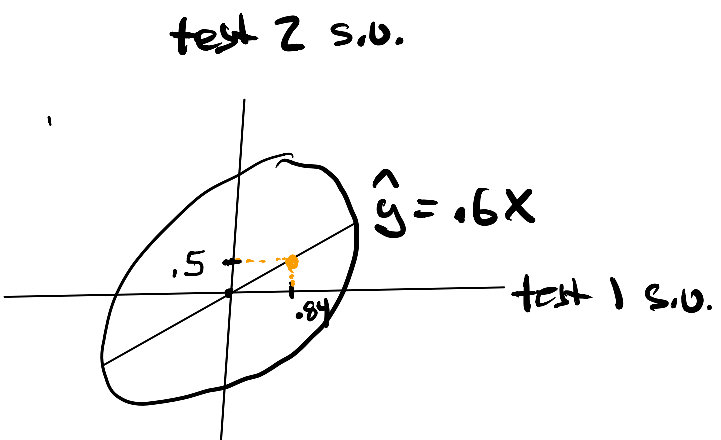
Appendix 5
Normal Table

Table shows values of $\Phi(z)$ for z from 0 to 3.59 by steps of .01. Example: to find $\Phi(1.23)$, look in row 1.2 and column .03 to find $\Phi(1.2 + .03) = \Phi(1.23) = .8907$. Use $\Phi(z) = 1 - \Phi(-z)$ for negative z.

	.0	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7703	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545

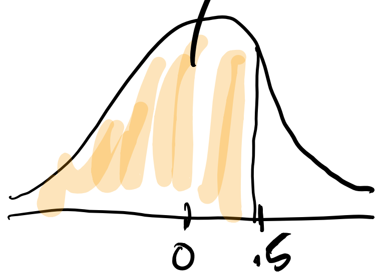
$\Phi^{-1}(.80) = .84$

in R:
qnorm(.80)



$\Rightarrow \hat{y} = .6(.84) = .5$

$\Phi(.5) = .69$



$\Rightarrow$ A score of $.5$ (in s.u.) for test 2 is only the 69^{th} percentile.

Regression line

In applications you often aren't given x and y in s.u.

$$\frac{y-\bar{y}}{SD(y)} = \rho \frac{x-\bar{x}}{SD(x)}$$ is regression line in s.u.

$$\Leftrightarrow y-\bar{y} = \frac{SD(y)}{SD(x)}\rho(x-\bar{x})$$

$$\Leftrightarrow \hat{y} = \underbrace{\left(\frac{SD(y)}{SD(x)}\rho\right)}_{m} x + \underbrace{\bar{y} - \frac{SD(y)}{SD(x)}\rho\bar{x}}_{b}$$

is regression line

Fun fact:

$$m = \frac{SD(y)}{SD(x)}\rho = \frac{SD(y)}{SD(x)}\operatorname{corr}(x,y) = \frac{\cancel{SD(y)}}{SD(x)} \frac{\operatorname{cov}(x,y)}{SD(x)\cancel{SD(y)}}$$

$$= \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)}$$

ex Test 1 is $\bar{x} = 60$

$SD(x) = 20$

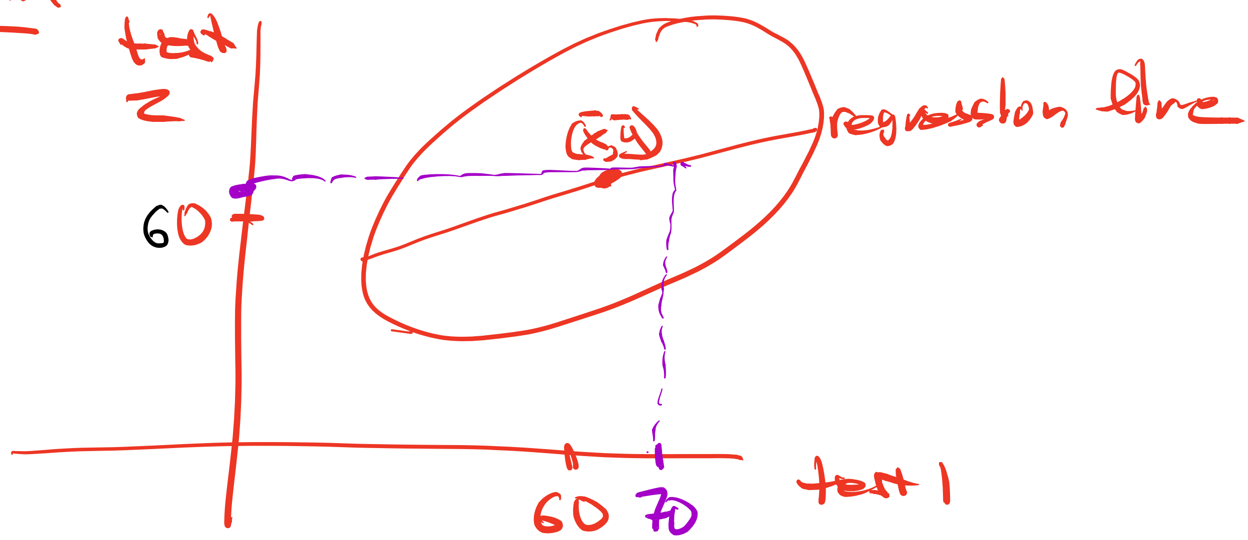
Test 2 is $\bar{y} = 60$

$SD(y) = 20$

$\rho = .6$

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

soln



soln

$$\hat{y} = \frac{SD(y)}{SD(x)} \rho x + \left(\bar{y} - \frac{SD(y)}{SD(x)} \rho \bar{x}\right)$$

$$= 1(.6)(70) + 60 - 1(.6)(60)$$

$$= \boxed{66}$$

Notice you did relatively worse

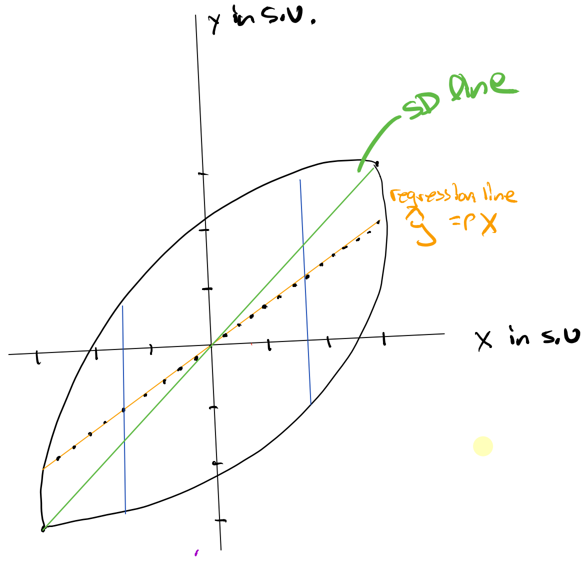
This is the "regression effect".

(3)

Regression line vs. SD line and regression effect

Def^n The SD line is $y-\bar{y} = \frac{SD(y)}{SD(x)}(x-\bar{x})$.

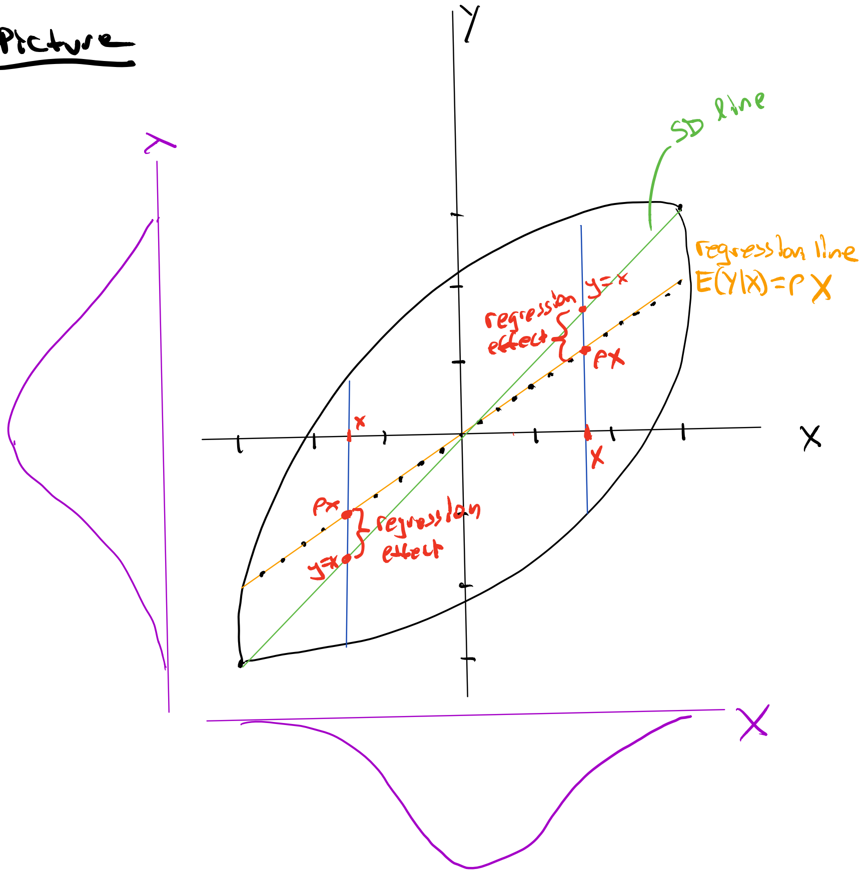
For X, Y in s.u. the SD line is $y = 1.X$ steeper than regression line.



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 50^{th} 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



Y
X
SD line
regression line
E(Y|X) = ρ X
regression effect
y=x
ρx
x
regression effect

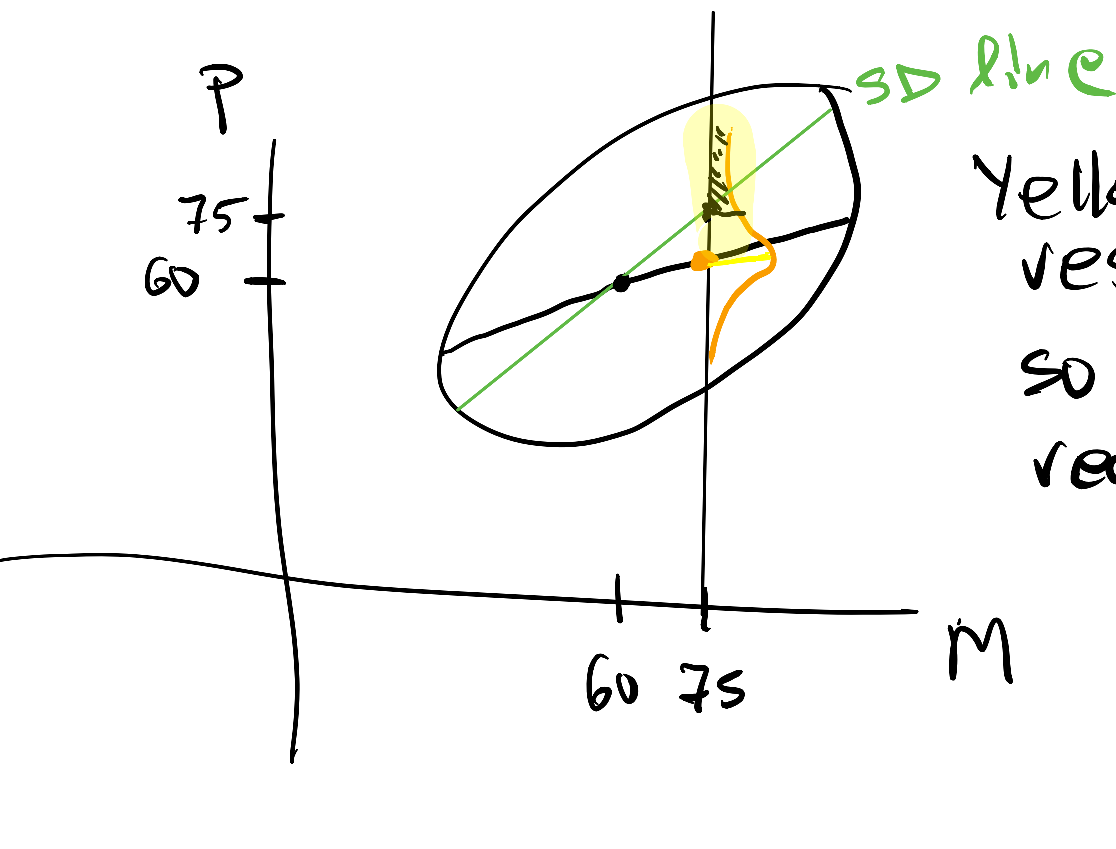
Concept test

A test score in Math and Physics is bivariate normal, $\rho > 0$. The average is 60 on both tests and the SDs are the same. Of students scoring 75 on the Math test:

a about half scored over 75 on Physics

b more than half scored over 75 on Physics

c less than half scored over 75 on Physics



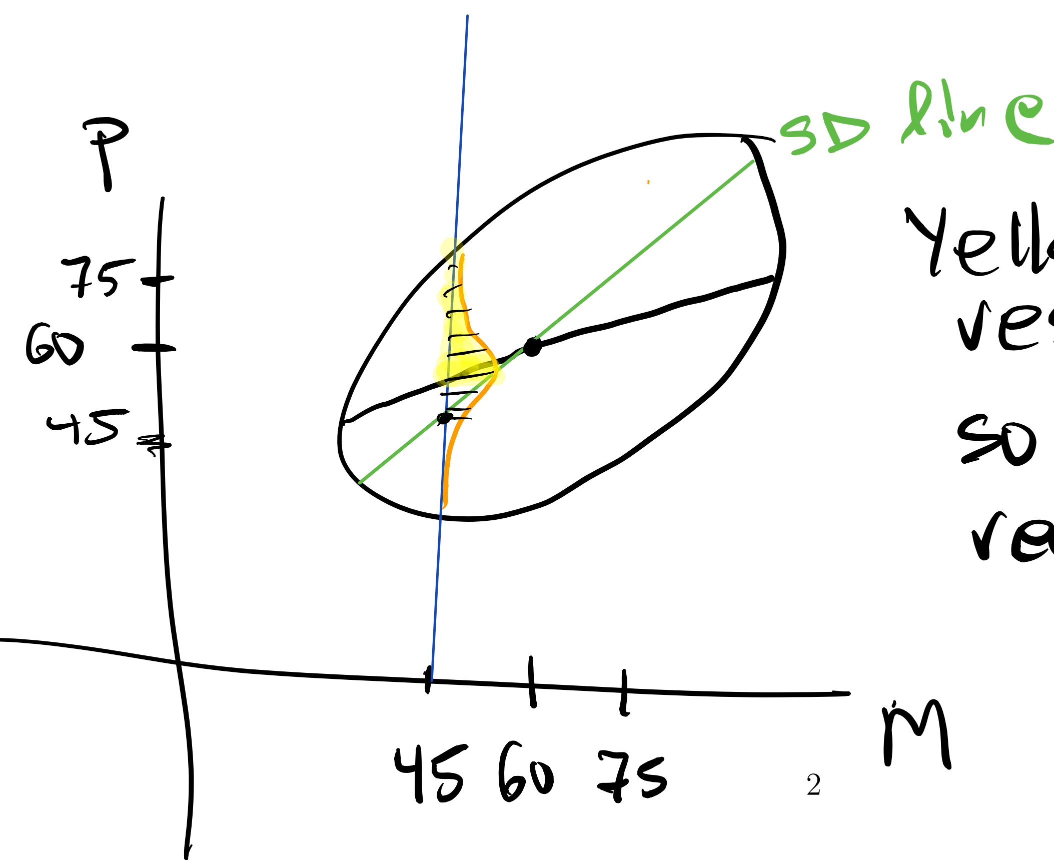
Yellow shaded region is 50%
so black shaded region < 50%

2. A test score in Math and Physics is bivariate normal, $\rho > 0$. The average is 60 on both tests and the SDs are the same. Of students scoring 45 on the Math test:

a about half scored over 45 on Physics

b more than half scored over 45 on Physics

c less than half scored over 45 on Physics



Yellow shaded region is 50%
so black shaded region > 50%

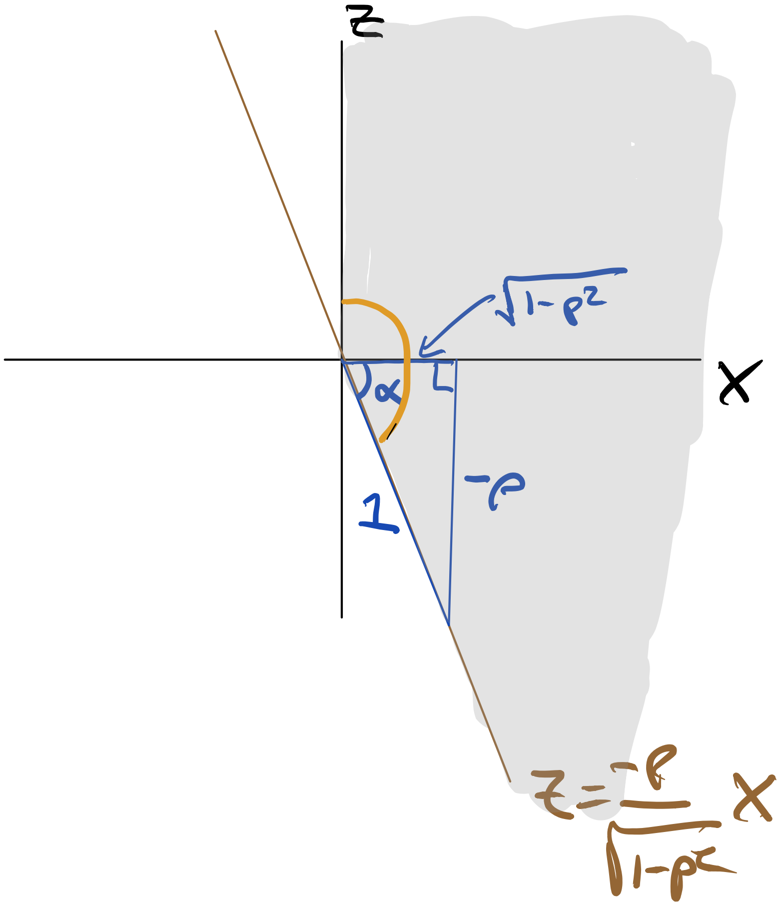
(4) More Practice.

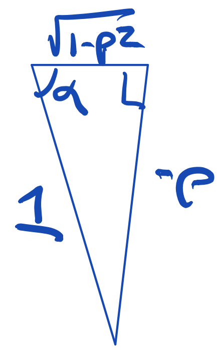
eg: X, Y std bivariate normal, $\rho > 0$

Find $P(X>0, Y>0)$

$$P(X>0, Y>0) = P(X>0, \rho X + \sqrt{1-\rho^2}\, Z > 0)$$

$$= P\left(X>0, Z > \frac{-\rho}{\sqrt{1-\rho^2}} X\right)$$





$$\tan\alpha = \frac{-\rho}{\sqrt{1-\rho^2}}$$

$$\alpha = \tan^{-1}\left(\frac{-\rho}{\sqrt{1-\rho^2}}\right)$$

$$P\left(X>0, Z > \frac{-\rho}{\sqrt{1-\rho^2}} X\right) = \frac{90 + |\alpha|}{360} = \boxed{\frac{90 + \left|\tan^{-1}\left(\frac{-\rho}{\sqrt{1-\rho^2}}\right)\right|}{360}}$$