Stat 134 lec 39

Sec 6.5 Bivariate normal

Last time:

Let $X, Z \overset{iid}{\sim} N(0,1)$ and $Y = \rho X + \sqrt{1-\rho^2}\, Z$, $-1 \le \rho \le 1$

We call the joint distribution (X,Y) the standard bivariate normal w/ correlation ρ

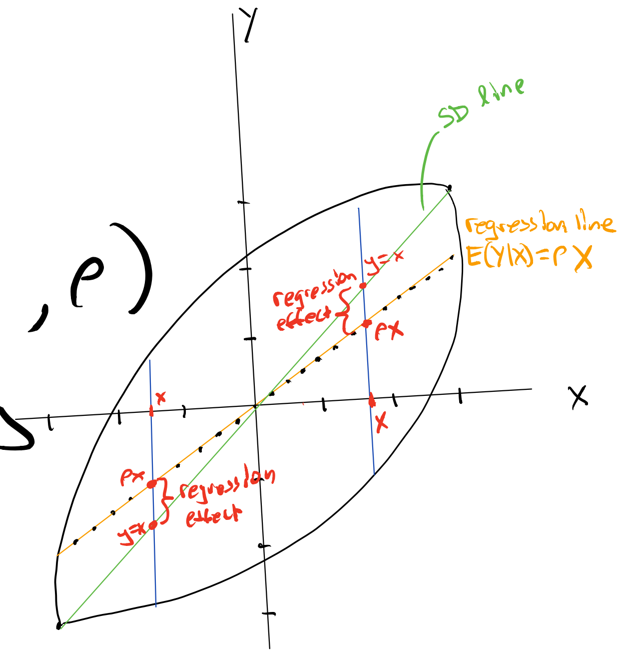
Properties

a) Regression line: $\hat{y} = \rho X$

$$\hat{y} = \underbrace{\rho \frac{\sigma_y}{\sigma_x}}_{m} X + \underbrace{\mu_y - \rho \frac{\sigma_y}{\sigma_x} \mu_x}_{b}$$

We say $(X,Y) \sim BV(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$ ← bivariate normal

b) The regression effect makes the predicted score regress towards the mean.



MGF

$M_Y(t) = E(e^{tY}) = E(e^{1 tY}) = M_{tY}(1)$

$M_{X,Y}(s,t) = E(e^{sX+tY}) = M_{sX+tY}(1)$ multivariate MGF

$Z \sim N(\mu, \sigma^2)$ iff $M_Z(a) = e^{\mu a} e^{\frac{\sigma^2 a^2}{2}}$

today ① Review student responses to Concept test.
② Properties of bivariate normal.
③ Practice with bivariate normal

Next time Review (see discussion board on b-courses)

1

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:

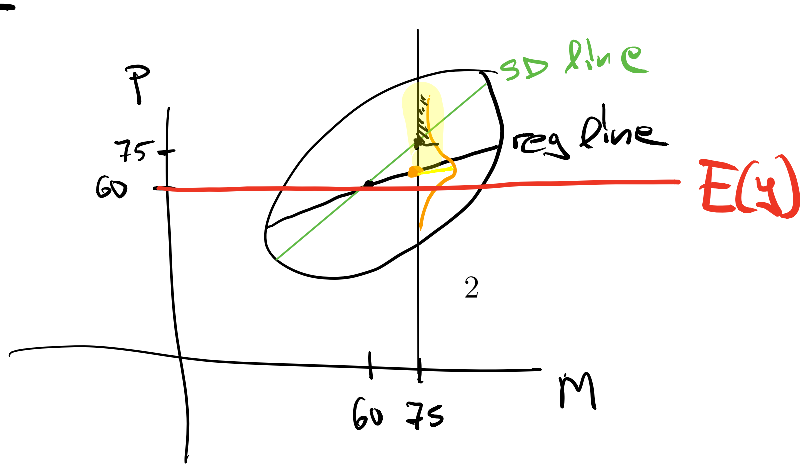
a about half scored over 75 on Physics

b more than half scored over 75 on Physics

c less than half scored over 75 on Physics

Review your thinking with a neighbor.

Picture



c	the second test			
c	Regression effect makes the score closer to the mean			
c	75 is above the average			
c				
c				
c				
b				
	There is a positive intersect			
b				
b				
b				
b				
b				
a				
b	You expect those above the mean to do worse, and those below the mean to do better			
b				
b				
b				
b	It's the regression effect we just talked about -- a student scoring below the mean of X should expect to score above the mean of Y, if X,Y are bivariate normal w/ positive correlation \ro.			
b				
b				
a				
b	In this case, the regression effect leads those students to performing better due to their previous score below the mean.			
c				
c				
c				
b	Assumption regression line.			
b				
c	Regression effect			
b				
c	The predicted score for the physics test using the regression line will yield a predicted score less than 75 for a math score of 75. Therefore a student would have had to score above 75 on the math test to get a 75 on the physics test, and fewer students did this because the test scores are distributed normally			
c	y=15p +60 < 75			
b	45 < 60, so predicted score will be higher on the second test			
b				
c				
c	The predicted score for the physics test using the regression line will yield a predicted score less than 75 for a math score of 75. Therefore a student would have had to score above 75 on the math test to get a 75 on the physics test, and fewer students did this because the test scores are distributed normally			
b				

② Properties of Bivariate normal.

Thm Let (X,Y) be standard bivariate normal.

The MGF of (X,Y) is

$$M_{(X,Y)}(s,t) = e^{\frac{s^2}{2}+\frac{t^2}{2}+st\rho}$$

Pf/ Recall that $X \sim N(0,1)$ so $M_X(s) = e^{\frac{s^2}{2}}$

$$M_{X,Y}(s,t) = E\left[e^{sX+tY}\right]$$

$$= E\left[e^{sX+t(\rho X+\sqrt{1-\rho^2}Z)}\right]$$

$$= E\left[e^{(s+t\rho)X} \cdot e^{t\sqrt{1-\rho^2}Z}\right]$$

independence

$$= E\left[e^{(s+t\rho)X}\right] E\left[e^{t\sqrt{1-\rho^2}Z}\right]$$

$$= M_X(s+t\rho) \cdot M_Z(t\sqrt{1-\rho^2})$$

$$= e^{\frac{(s+t\rho)^2}{2}} \cdot e^{\frac{(t\sqrt{1-\rho^2})^2}{2}}$$

$$= e^{\frac{s^2}{2}+st\rho+\frac{t^2}{2}\rho^2} \cdot e^{\frac{t^2}{2}-\frac{t^2\rho^2}{2}}$$

$$= \boxed{e^{s^2/2+t^2/2+st\rho}}$$

$\square$

Thm If (X,Y) is bivariate normal then

$\rho = \text{Corr}(X,Y) = 0$ iff X,Y are independent.

Pf/

$$M_{(X,Y)}(s,t) = M_{sX+tY}(1) = M_{sX}(1) \cdot M_{tY}(1) = M_X(s) M_Y(t)$$

iff X and Y are independent.

Since (X,Y) is bivariate normal,

$$M_{(X,Y)}(s,t) = e^{s^2/2 + t^2/2 + st\rho}$$

$$= M_X(s) M_Y(t) \quad \text{iff } \rho = 0$$

since $M_X(s) = e^{s^2/2}$

$M_Y(t) = e^{t^2/2}$

$\square$

Recall from lec 30 that the sum of independent normal random variables is normal.

What about dependent normal random variables?

Thm Let $X, Y \sim N(0,1)$ and $\operatorname{corr}(X,Y) = \rho$.

(X,Y) is std bivariate normal iff

$sX + tY$ is normal for all constants s, t.

Pf/

$\Rightarrow$: Suppose X, Y std bivariate normal.

(i.e. $Y = \rho X + \sqrt{1-\rho^2}\, Z$ for X, Z iid $N(0,1)$)

$$sX + tY = sX + t\left(\rho X + \sqrt{1-\rho^2}\, Z\right)$$

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

is normal since X, Z are independent normals.

$\Leftarrow$: Suppose $sX + tY$ is normal.

note $E(sX+tY) = sE(X) + tE(Y) = 0$

$$\operatorname{Var}(sX+tY) = \operatorname{Var}(sX) + \operatorname{Var}(tY) + \underbrace{2\operatorname{Cov}(sX, tY)}_{= 2\operatorname{Corr}(sX,tY)SD(X)SD(Y)}$$

$$= s^2 + t^2 + 2st\rho$$

$$\Rightarrow \quad sX + tY \sim N(0, s^2 + t^2 + 2st\rho)$$

Recall,

For $Z \sim N(\mu, \sigma^2) \Rightarrow M_Z(a) = e^{\mu a} e^{\sigma^2 \frac{a^2}{2}}$

σ^2

Then $M_{(X,Y)}(s,t) = M_{sX+tY}(1) = e^{(s^2+t^2+2st\rho)\frac{1^2}{2}}$

normal

MGF of (X,Y)

$$= e^{\frac{s^2}{2} + \frac{t^2}{2} + st\rho}$$

$\Rightarrow$ (X,Y) is bivariate normal $\square$

Corollary

Let $X \sim N(\mu_x, \sigma_x^2)$
$Y \sim N(\mu_y, \sigma_y^2)$
and $\mathrm{Corr}(X,Y) = \rho$

$(X,Y) \sim BV(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$ iff
$sX + tY$ is normal for all constants s, t.

Pf/ By definition we say
$(X,Y) \sim BV(\mu_x, \sigma_x, \mu_y, \sigma_y, \rho)$ iff
$(X^*, Y^*) \sim BV(0, 1, 0, 1, \rho)$

in S.U.

Note that $\mathrm{corr}(X,Y) = \mathrm{Corr}(X^*, Y^*)$.

Furthermore,
Note that $sX^* + tY^*$ is normal iff
$sX + tY$ is normal since

$$X^* = \frac{X - \mu_x}{\sigma_x} \quad \text{and} \quad Y^* = \frac{Y - \mu_y}{\sigma_y},$$ and

a linear transformation of a normal RV is normal.

$\square$

ex Let M = a students score on the midterm
F = a students score on the final

$$E(M) = 70 \quad E(F) = 65 \quad \rho = .6$$
$$SD(M) = 8 \quad SD(F) = 10$$

Find the chance that a student scores higher on the final than the midterm?

Soln

normal since $(M, F) \sim BV(70, 8, 65, 10, .6)$

$$P(F > M) = P(F - M > 0)$$

$$= P\left(\frac{F-M - E(F-M)}{SD(F-M)} > \frac{0 - E(F-M)}{SD(F-M)} \right)$$

$$E(F-M) = 65 - 70 = -5$$

$Corr(F,M)SD(F)SD(M)$

$$Var(F-M) = Var(F) + Var(M) - 2Cov(F,M)$$
$$= 10^2 + 8^2 - 2(.6)(8)(10)$$
$$= 68$$

$$SD(F-M) = \sqrt{68}$$

$$\Rightarrow \frac{0 - E(F-M)}{SD(F-M)} = \frac{0-(-5)}{\sqrt{68}} = .61$$

$$\Rightarrow P(F > M) = 1 - \Phi(.61) = .27$$

Stat 134

Monday December 3 2018

1. Let (X,Y) be bivariate normal. Then (2X+3Y+4, 6X-Y-4) is bivariate normal.

   a true

   b false

$a(2x+3y+4)+b(6x-y-4)$

$=(2a+6b)X+(3a-b)Y+4a-4b$

is normal since (X,Y) bivariate normal (using corollary above).

Review MGF

Main properties

1. $M_X(0) = 1$

2. $M_{aX}(t) = M_X(at)$

3. $M'_X(0) = E(X)$

   $M''_X(0) = E^2(X)$

   $\vdots$

   $M^{(k)}_X(0) = E^{(k)}(X)$

4. $M_X(t)$ is unique for t in a neighborhood of 0. So if $M_X(t) = e^{t^2/2}$ for all t then $X \sim N(0,1)$.

ex

CLT

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, mean μ, SD σ ← any distribution

$S_n = \sum_{i=1}^{n} X_i$

$S_n \longrightarrow N(n\mu, n\sigma^2)$ as $n \to \infty$

Pf/ We show that

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} \longrightarrow Z \sim N(0,1) \quad \text{as } n \to \infty$$

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \frac{1}{\sqrt{n}}\left(\frac{S_n - n\mu}{\sigma}\right)$$

$$= \frac{1}{\sqrt{n}}\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma}\right)$$

$$= \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n}\underbrace{\frac{X_i - \mu}{\sigma}}_{= Y_i}\right)$$

$$= \sum_{i=1}^{n}\frac{Y_i}{\sqrt{n}}$$

$$E(Y_i) = E\left(\frac{X_i - \mu}{\sigma}\right) = \frac{1}{\sigma}E(X_i - \mu) = 0$$

$$Var(Y_i) = \frac{1}{\sigma^2}Var(X_i - \mu) = \frac{1}{\sigma^2}\cdot\sigma^2 = 1$$

So $E^2(Y_i) = Var(Y_i) + E(Y_i)^2 = 1$

Make a Taylor series of $M_{Y_i}(t)$, around 0:

$$M_{\frac{Y_i}{\sqrt{n}}}(t) = M_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = M_{Y_i}(0) + M'_{Y_i}(0)\frac{t}{\sqrt{n}} + \frac{M''_{Y_i}(0)\frac{t^2}{n}}{2!} + \cdots$$

$$= 1 + \underbrace{E(Y_i)}_{= 0}\frac{t}{\sqrt{n}} + \frac{\overbrace{E^2(Y_i)}^{= 1}}{2!}\frac{t^2}{n} + \cdots$$

$$= 1 + \frac{1}{n}\left[\frac{t^2}{2} + \underbrace{\frac{t^3 M'''(0)}{3!\,n^{1/2}}}_{\to 0} + \cdots\right]$$

all terms $\to 0$ since have n

$\to 1 + \frac{1}{n}\frac{t^2}{2}$ as $n \to \infty$ in denom,

So

$$M_{\frac{S_n - n\mu}{\sqrt{n}\sigma}}(t) = M_{\frac{Y_1}{\sqrt{n}}}(t) \cdots M_{\frac{Y_n}{\sqrt{n}}}(t)$$

$$\to \left[1 + \frac{1}{n}\frac{t^2}{2}\right]^n \approx e^{\frac{t^2}{2}}$$

which is MGF of $N(0,1)$

Hence $\frac{S_n - n\mu}{\sqrt{n}\sigma} \to N(0,1)$

$\square$

Appendix

Example of $X, Y \sim N(0,1)$ whose joint <u>isn't</u> std bivariate normal.

<u>ex</u> $X \sim N(0,1)$, $W = \begin{cases} 1 & \text{w/ prob } 1/2 \\ -1 & \text{w/ prob } 1/2 \end{cases}$
indep RVs.

let $Y = WX = \begin{cases} X & \text{if } W = 1 \\ -X & \text{if } W = -1 \end{cases}$

Next we show that $Y \sim N(0,1)$.

Note by change of variable rule that $-X \sim N(0,1)$.

Then
$$\begin{aligned}
F_Y(y) &= P(Y \leq y) \\
&= P(Y \leq y \mid W=1)P(W=1) + P(Y \leq y \mid W=-1)P(W=-1) \\
&= \tfrac{1}{2}P(X \leq y \mid W=1) + \tfrac{1}{2}P(-X \leq y \mid W=-1) \\
&= \tfrac{1}{2}P(X \leq y) + \tfrac{1}{2}P(-X \leq y) \quad \text{since } X, W \text{ indep.} \\
&= \tfrac{1}{2}\Phi(y) + \tfrac{1}{2}\Phi(y) \quad \text{since } X, -X \sim N(0,1) \\
&= \Phi(y).
\end{aligned}$$

Thus $Y \sim N(0,1)$.

Then $X + Y = \begin{cases} 2X & \text{with prob } 1/2 \\ 0 & \text{with prob } 1/2 \end{cases}$

is a mixed distribution, <u>not</u> normal.

So X, Y normal but $X+Y$ not normal.

Furthermore $\mathrm{Cov}(X, Y) = 0$ since

$E(YX) = E(WX^2) = \underbrace{E(W)}_{=0}E(X^2)$ by indep of W and X.

but X, Y are clearly dependent contradicting property of bivariate normal.

Extra Problem

문) Let M = midterm score
F = final score

$E(M) = 66 \quad E(F) = 73 \quad \rho = 0.6$
$SD(M) = 10 \quad SD(F) = 8$

Let overall score be $S = .3M + .7F$

Find $Cov(F, S)$.

Soln $Cov(F,S) = Cov(F, .3M + .7F)$

$= .3\,Cov(F, M) + .7\,Var(F)$

$= .3\,Corr(F, M)SD(F)SD(M) + .7\,Var(F)$

$= (.3)(.6)\,8 \cdot 10 + .7(64)$

$= 59.2$

If asked to find $Corr(F,S)$

$$Corr(F,S) = \frac{Cov(F,S)}{SD(F)SD(S)}$$

$Var(S) = Var(S) = Var(.3M + .7F)$

$= Var(.3M) + Var(.7F) + 2Cov(.3M, .7F)$

$= .09\,Var(M) + .49\,Var(F) + 2(.3)(.7)\underbrace{Cov(M,F)}_{corr(M,F)SD(M)SD(F)}$

$= .09(100) + .49(64) + 2(.3)(.7)(.6)(10)(8) = 60.52$

so $Corr(F,S) = \dfrac{59.2}{10 \cdot \sqrt{60.52}} = .76$