Co-variance and correlation

X, Y are random variable

with $\mathbb{E}[X] = \mu_x$ $\mathbb{E}[Y] = \mu_y$

$\operatorname{var}(X) = \sigma_x^2$, $\operatorname{var}(Y) = \sigma_y^2$

$0 < \sigma_x^2 < \infty$, $0 < \sigma_y^2 < \infty$

Definition 4.5.1:- The co-variance of X and Y is the number defined by

$$\operatorname{Cov}(X,Y) = \mathbb{E}\left[(X-\mu_x)(Y-\mu_y)\right]$$

Definition 4.5.2:- The correlation of X and Y is the number defined by

$$\rho_{xy} = \frac{\operatorname{Cov}(X,Y)}{\sigma_x \sigma_y}$$

ρ_{xy} = correlation coefficient.

Theorem 4.5.3: For any r.v X, and Y

$$\text{cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y$$

Proof:

$$\text{cov}(X,Y) = \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$$

$$= \mathbb{E}\left[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y\right]$$

$$= \mathbb{E}[XY] - \mathbb{E}[X]\mu_Y - \mathbb{E}[Y]\mu_X + \mu_X\mu_Y$$

$$= \mathbb{E}[XY] - 2\mu_X\mu_Y + \mu_X\mu_Y$$

$$\boxed{\text{cov}(X,Y) = \mathbb{E}[XY] - \mu_X\mu_Y}$$

Example 4.5.4: Correlation-I

Let the joint pdf of (X,Y) be

$$f(x,y) = 1, \quad \begin{matrix} 0 < x < 1 \\ x < y < x+1 \end{matrix}$$

marginal distribution of $X \sim \text{unif}(0,1)$

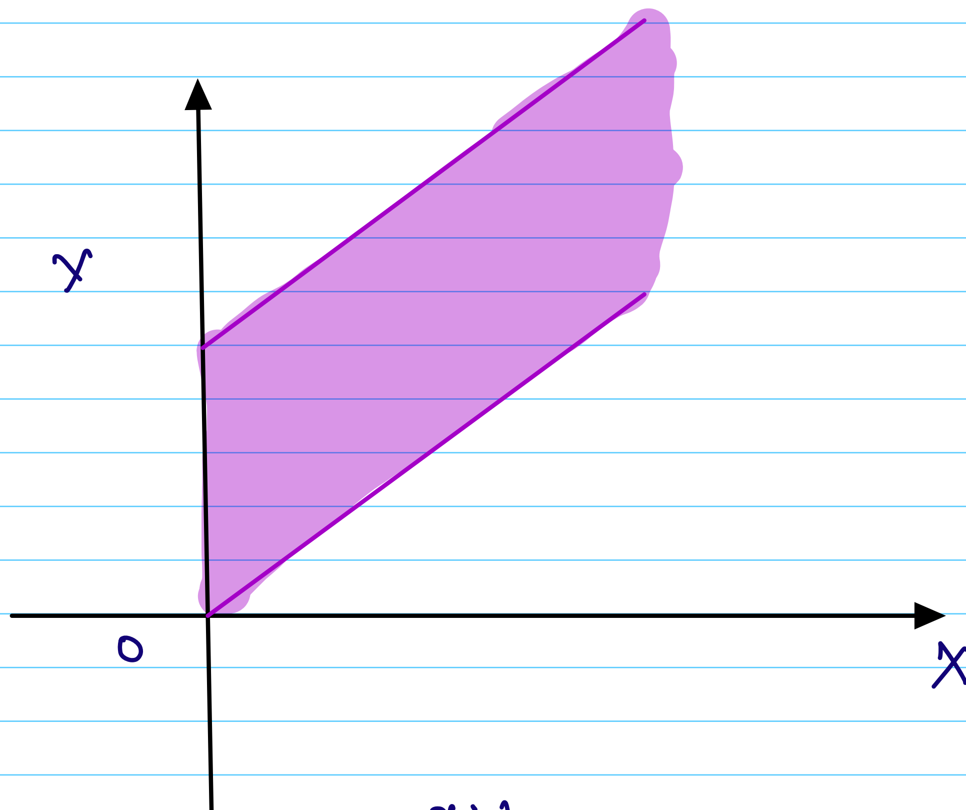
so $\mu_X = \frac{1}{2}$ $\quad \sigma_X^2 = \frac{1}{12}$

The marginal Pdf of Y is $f_Y(y) = y$, $0 < y < 1$ and $f_Y(y) = 2-y$, $1 \le y < 2$

$\mu_Y = 1$, $\sigma_Y^2 = \frac{1}{6}$.

Solution:

$$f_{XY}(x,y) = 1 \qquad 0 < x < 1$$
$$x < y < x+1$$



$$f_X(x) = \int_x^{x+1} f_{XY}(x,y)\, dy$$

$$= \int_x^{x+1} 1\, dy = 1 \qquad 0 < x < 1$$

$X \sim \text{unif}(0,1)$ $E[X] = \frac{1}{2}$

$\sigma_X^2 = \frac{1}{12}$

$$f_Y(y) = \int_x f(x,y)\,dx$$

$$f_Y(y) = \int_0^y 1\,dx \qquad 0 < y < 1$$

$$= \int_{y-1}^1 1\,dx \qquad 1 < y < 2$$

$$f_Y(y) = \begin{cases} y & 0 < y < 1 \\ 2-y & 1 < y < 2 \end{cases}$$

$$E[Y] = \int_0^1 y \cdot y\,dy + \int_1^2 (2-y)\,y\,dy$$

$$= \frac{1}{2} + \left. 2y - \frac{y^2}{2} \right]_1^2 = \frac{1}{2} + 2 - \frac{3}{2}$$

$$= 1$$

$\text{var}(Y) = \frac{1}{6}$

$$\text{Cov}(X,Y) = \mathbb{E}(XY) - \mu_X \mu_Y$$

$$= \int_{x=0}^{x=1}\int_{y=x}^{y=x+1} xy \, dy \, dx - \frac{1}{2} = \int_{x=0}^{x=1} x \cdot \frac{y^2}{2}\Big]_{y=x}^{y=x+1} dx - \frac{1}{2}$$

$$= \int_{x=0}^{x=1} x\left[\frac{(x+1)^2 - x^2}{2}\right] dx - \frac{1}{2}$$

$$= \int_{x=0}^{x=1} x\left(\frac{1}{2} + x\right) dx - \frac{1}{2}$$

$$= \frac{x^2}{4} + \frac{x^3}{3}\Big]_0^1 - \frac{1}{2} = \frac{7}{12} - \frac{1}{2}$$

$$= \frac{1}{12}$$

$$\text{Cov}(X,Y) = \frac{1}{12}$$

correlation coefficient $\rho = \dfrac{\text{Cov}(X,Y)}{\sigma_x \sigma_y}$

$$\Rightarrow \rho = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{6}}} = \frac{1}{\sqrt{2}}$$

Fundamental Properties of Co-variance and correlation.

Theorem 4.5.5:- if X and Y are independent random variable's, then $\text{cov}(X,Y) = 0$ and $\rho_{XY} = 0$

Proof:

$$\text{cov}(X,Y) = \mathbb{E}[XY] - \mu_X \mu_Y$$

if X, Y are independent then

$$\mathbb{E}[g(X)\, h(Y)] = \mathbb{E}[g(X)] \cdot \mathbb{E}[h(Y)]$$

$$\Rightarrow \quad \text{cov}(X,Y) = \mathbb{E}[X \cdot Y] - \mu_X \mu_Y$$

$$= \mathbb{E}[X]\, \mathbb{E}[Y] - \mu_X \mu_Y$$

$$= \mu_X \mu_Y - \mu_X \mu_Y$$

$$\boxed{\text{cov}(X,Y) = 0}$$

$$\boxed{\rho_{XY} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = 0}$$

It is important to Note that, however, that theorem 4.5.5 does not say that if $\mathrm{cov}(X, Y) = 0$, then X, Y are independent.

Ex:

$$X \sim f(x - \theta)$$

symmetric around θ, with $E[X] = \theta$

& $$Y = I(|X - \theta| < 2)$$

$\Rightarrow$ X & Y are not independent.

However

$$\mathrm{cov}(X, Y) = \int_{-\infty}^{\infty} x\, I(|x - \theta| < 2)\, f(x - \theta)\, dx$$

$\boxed{x - \theta = t}$

$$= \int_{-2}^{2} x \cdot f(x - \theta)\, dx = \int_{-2}^{2} (t + \theta) f(t)\, dt$$

$$= \theta \int_{-2}^{2} f(t)\, dt + \int_{-2}^{2} t f(t)\, dt = 0$$

$\Rightarrow$ symmetric around $0 \Rightarrow E[T] = 0$

$$E[XY] = 0 \int_{-2}^{2} f(t)\,dt = E[X]\,E[Y]$$

$$\Rightarrow \quad cov(X, Y) = 0$$

$\Rightarrow$ X, Y are uncorrelated, dependent r.v's.

Co-variance and correlation measure only a particular kind of linear relationship, that will be described further in Theorem 4.5.7.

$$X, Y \text{ independent} \Longrightarrow cov(X,Y) = 0$$

$$X, Y \text{ independent.} \;\cancel{\Longleftarrow}\; cov(X,Y) = 0$$

Theorem 4.5.6:

if X and Y are any two random variables and a, and b are any two constants, then

$$var(aX + aY) = a^2 var(X) + b^2 var(Y) + 2ab\, cov(X,Y)$$

if X, Y are independent r.v then

$$var(aX+bY) = a^2 var(X) + b^2 var(Y)$$

Proof:

$aX+bY \sim$ is a r.v

$$\Rightarrow \quad \mathbb{E}[aX+bY] = a\mu_x + b\mu_y$$

$$var(aX+bY) = \mathbb{E}\left[\left(aX+bY-(a\mu_x+b\mu_y)\right)^2\right]$$

$$= \mathbb{E}\left[\left(a(X-\mu_x) + b(Y-\mu_y)\right)^2\right]$$

$$= \mathbb{E}\left[a^2(X-\mu_x)^2\right] + \mathbb{E}\left[b^2(Y-\mu_y)^2\right]$$

$$+ \mathbb{E}\left[2ab(X-\mu_x)(Y-\mu_y)\right]$$

$$\boxed{\Rightarrow \quad a^2 var(X) + b^2 var(Y) + 2ab \, cov(X,Y)}$$

$\Rightarrow$ $Var(aX+bY) = a^2 Var(X) + b^2 Var(Y) + 2ab\,Cov(X,Y)$

if X, Y are positively co-rrelated

$\Rightarrow$ $Cov(X,Y) > 0$

$\Rightarrow$ $Var(X+Y)$ will be greater.
(greater than sum)

if X, Y are negatively co-rrelated

$\Rightarrow$ $Cov(X,Y) < 0$

$\Rightarrow$ $Var(X+Y)$ will be smaller
(less than sum)

Theorem 4.5.7: for any random variable's X and Y,

a. $-1 \leq \rho_{XY} \leq 1$

b. $|\rho_{XY}| = 1 \iff \exists\, a \neq 0$ & b such that $\mathbb{P}(Y = aX+b) = 1$. if $\rho_{XY} = 1$, then $a > 0$ and if $\rho_{XY} = -1$, then $a < 0$.

Proof: Consider a function $h(t)$

$$h(t) = \mathbb{E}\left[\left((X-\mu_X)t + (Y-\mu_Y)\right)^2\right]$$

$$\Rightarrow h(t) = \mathbb{E}\left[t^2 (X-\mu_x)^2\right] + \mathbb{E}\left[(Y-\mu_y)^2\right] + 2t\,\mathbb{E}\left[(X-\mu_x)(Y-\mu_y)\right]$$

$$\Rightarrow h(t) = t^2 \operatorname{var}(X) + \operatorname{var}(Y) + 2t \operatorname{cov}(X, Y)$$

$$\Rightarrow h(t) = \sigma_x^2 \cdot t^2 + 2\operatorname{cov}(X, Y) \cdot t + \sigma_y^2$$

roots are $= \dfrac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$h(t) \geq 0 \quad \forall t$

will have at most one real root.

$$\Rightarrow b^2 - 4ac \leq 0$$

$$\Rightarrow 4\operatorname{cov}(X, Y)^2 - 4\sigma_x^2 \sigma_y^2 \leq 0$$

$$\Rightarrow \operatorname{cov}(X, Y)^2 \leq \sigma_x^2 \sigma_y^2$$

$$\Rightarrow \quad -\sigma_x\sigma_y \leq \mathrm{cov}(X,Y) \leq \sigma_x\sigma_y$$

$$\Rightarrow \quad -1 \leq \frac{\mathrm{cov}(X,Y)}{\sigma_x\sigma_y} \leq 1$$

$$\Rightarrow \quad -1 \leq \rho \leq 1$$

also if it has a double real root

$\Rightarrow \quad b^2 - 4ac = 0 \Rightarrow |\rho_{xy}| = 1$

$\Rightarrow \quad h(t) = 0$ (for the double root, $t = \frac{-b}{2a}$)

$$\Rightarrow \quad \mathbb{E}\left[\left[(X-\mu_x)t + (Y-\mu_Y)\right]^2\right] = 0$$

$\Rightarrow$ expectation of a non-negative random variable $= 0$

$\Rightarrow$ all the probability density lie at 0

for the r.v

$$\left[(X-\mu_X)t + (Y-\mu_Y)\right]^2 \sim I_0$$

$$\Downarrow$$

$$\Rightarrow \mathbb{P}\left[\left[(X-\mu_X)t + (Y-\mu_Y)\right]^2 = 0\right] = 1$$

$$\Rightarrow \mathbb{P}\left[(X-\mu_X)t + (Y-\mu_Y) = 0\right] = 1$$

$$\Rightarrow \mathbb{P}\left[Xt + Y = \mu_X t + \mu_Y\right] = 1$$

$$\Rightarrow \mathbb{P}\left[Y = -tX + \mu_X t + \mu_Y\right] = 1$$

t will be the double root

$$t = \frac{-b}{2a}$$

$$\Rightarrow t = \frac{-2\,\mathrm{cov}(X,Y)}{2\sigma_X^2} = -\frac{\mathrm{cov}(X,Y)}{\sigma_X^2}$$

Example 4.5.8 (correlation II)

$X \sim \text{unif}(0,1)$

$Z \sim \text{unif}(0, \frac{1}{10})$

X and Z are independent

$Y = X + Z$ random variable

(X, Y) random vector.

$0 < x < 1, \quad 0 < z < \frac{1}{10}$

$$\left.\begin{aligned} U &= X \\ V &= X + Z \end{aligned}\right\} \quad \begin{aligned} X &= U \\ Z &= V - U \end{aligned} \qquad J = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix}$$

$J = 1$

$\Rightarrow 0 < u < 1, \quad 0 < v < \frac{11}{10}$

$f_{X,Z}(x,z) = 10 \qquad \begin{matrix} 0 < x < 1 \\ 0 < z < \frac{1}{10} \end{matrix}$

$$f_{U,V}(u,v) = 10 * 1 \qquad 0 < u < 1$$
$$u < v < u + \frac{1}{10}$$

$$\boxed{f_{X,Y}(x,y) = 10 \qquad 0 < x < 1, \quad x < y < x + \frac{1}{10}}$$

Another method:

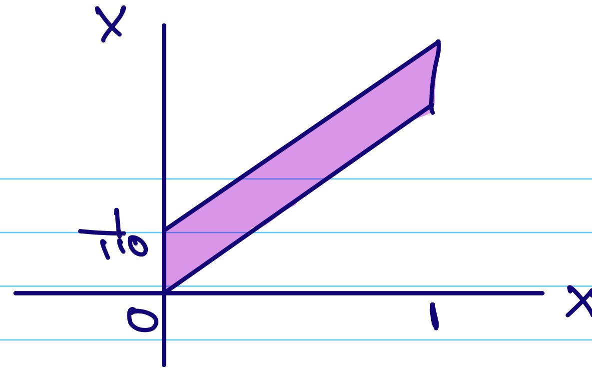
Given $X = x \qquad Y = x + Z$

$$\Rightarrow \quad Y|X \sim \text{unif}\left(x, x + \frac{1}{10}\right)$$
$$X \sim \text{unif}(0,1)$$

$$\Rightarrow \mathbb{P}(Y=y) = \int_x \mathbb{P}(X,Y)$$

$$\Rightarrow \mathbb{P}(Y=y) = \int_x \mathbb{P}(Y=y \mid X=x)\,\mathbb{P}(x)$$

$$\Rightarrow \mathbb{P}(Y=y) = \int_{x=x} \frac{1}{10}\, 1 .\, dx$$



$$\Rightarrow \mathbb{P}(Y=y) = \int_{x=0}^{x=y} \frac{1}{10}\,dx \quad \text{for } y<\frac{1}{10}$$

$$= \frac{y}{10} \quad \text{for } y<\frac{1}{10}$$

$$\Rightarrow \mathbb{P}(y=y) = \int_{x=y-1}^{x=y} \frac{1}{10}\,dx \quad \text{for } \frac{1}{10}\le y<1$$

$$= \frac{1}{10}(y-y+1) = \frac{1}{10}$$

$$\text{for } \frac{1}{10}\le y<1$$

$$\Rightarrow \mathbb{P}(y=y) = \int_{x=y-1}^{1} \frac{1}{10}\,dx \quad \text{for } 1\le y<\frac{11}{10}$$

$$= \frac{1}{10}(2-y)$$

$$\Rightarrow \mathbb{P}(y=y) = \begin{cases} \frac{y}{10} & 0 \le y < \frac{1}{10} \\ \frac{1}{10} & \frac{1}{10} \le y < 1 \\ \frac{2-y}{10} & 1 \le y < \frac{11}{10} \end{cases}$$

The representation of $Y = X + Z$ makes the computation of the co-variance and co-relation easy.

$$\mathbb{E}[X] = \frac{1}{2} \qquad \mathbb{E}[Y] = \mathbb{E}[X+Z]$$

$$\Rightarrow \mathbb{E}[Y] = \frac{1}{2} + \frac{1}{20} = \frac{11}{20}$$

$$\Rightarrow Cov(X,Y) = \mathbb{E}[XY] - E[X]E[Y]$$

$$= \mathbb{E}[X(X+Z)] - E[X]\,E[X+Z]$$

$$= \mathbb{E}[X^2] + E[\cancel{XZ}] - E[X]^2 - E[X]\cancel{E[Z]}$$

$$= \mathbb{E}[X^2] - E[X]^2$$

$$= var(X) = \frac{1}{12}$$

$$\mathrm{var}(Y) = \mathrm{var}(X+Z)$$
$$= \mathrm{var}(X) + \mathrm{var}(Z)$$
$$= \frac{1}{12} + \frac{1}{1200}$$

$$\Rightarrow \mathrm{Cov}(X,Y) = \frac{\frac{1}{12}}{\sqrt{\frac{1}{12}}\sqrt{\frac{1}{12}+\frac{1}{1200}}} = \sqrt{\frac{100}{101}}$$

there may be a strong relationship between X and Y, but if the relation is not linear, the correlation may be small.

<u>Example 4.5.9</u> (Correlation - III)

$$X \sim \mathrm{Unif}(-1,1) \qquad -1<x<1$$

$$f_X(x) = \frac{1}{2} \qquad -1<x<1$$

$Z \sim \text{Unif}(0, \frac{1}{10})$ $\qquad 0 < z < \frac{1}{10}$

$f_Z(z) = 10$

Now $\quad Y = X^2 + Z$

Method 1

$f_{X,Z} = 5$ $\qquad -1 < x < 1$

$\qquad 0 < z < \frac{1}{10}$

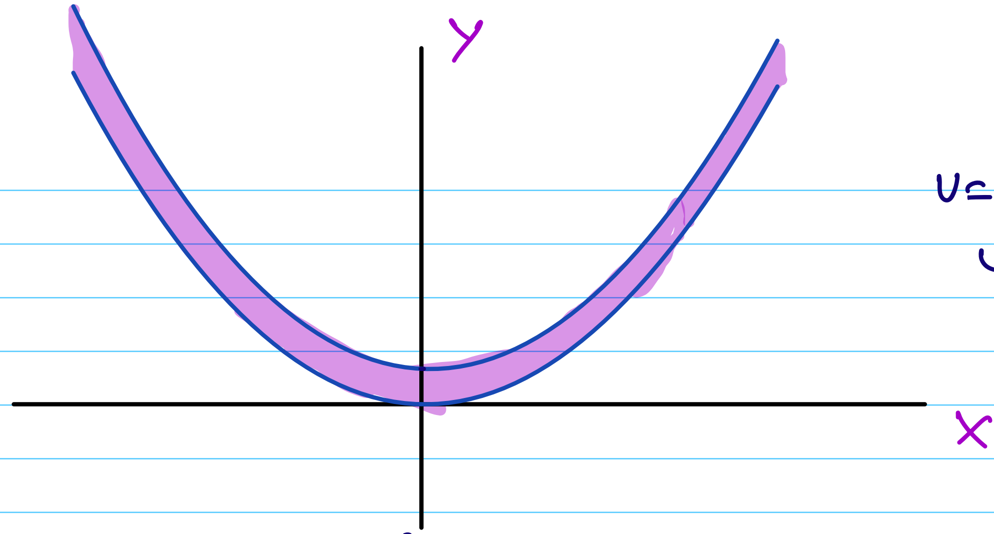
$$U = X \qquad\qquad X = U$$
$$V = X^2 + Z \implies Z = V - U^2 \qquad |J| = \begin{vmatrix} 1 & 0 \\ -2u & 1 \end{vmatrix}$$

$= 1$

$$-1 < u < 1$$
$$u^2 < v < u^2 + \frac{1}{10}$$

$f_{U,V}(u,v) = 5$ $\qquad -1 < u < 1$

$\qquad u^2 < v < u^2 + \frac{1}{10}$



$v = u^2$

$u = \sqrt{v}$

$v = u^2 + \frac{1}{10}$

$\sqrt{v - \frac{1}{10}}$

$$f_V(v) = \int_{\sqrt{v}}^{\sqrt{v-\frac{1}{10}}} 5 \, du$$

$\vdots$

Calculate $\mathrm{Cov}(X,Y)$, ρ_{XY}

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X(X^2+Z)] - E[X]E[X^2+Z]$$

$$= \mathbb{E}[X^3] + \mathbb{E}[XZ] - 0\,\mathbb{E}[X^2+Z]$$

$$= 0 + 0 - 0$$

$$\rho_{XY} = 0$$

$\Rightarrow$ even though X, Y are highly correlated but $\rho_{XY} = 0$ because it's not linear relation.

Definition 4.5.10 :-

The Bivariate normal Pdf with

$-\infty < \mu_x < \infty$, $-\infty < \mu_y < \infty$,

$0 < \sigma_x^2 < \infty$, $0 < \sigma_y^2 < \infty$,

$-1 < \rho_{xy} < 1$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2\right]\right]$$

$-\infty < x < \infty$

$-\infty < y < \infty$

Properties

(a) $X \sim N(\mu_x, \sigma_x^2)$

(b) $Y \sim N(\mu_y, \sigma_y^2)$

(c) $\rho_{xy} = \rho$

(d) a, b are constants

$$aX + bY \sim N(a\mu_x + b\mu_y,\ a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y)$$

Prove c from a,b

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_x \sigma_y}$$

$$= \frac{\mathbb{E}[(X-\mu_x)(Y-\mu_y)]}{\sigma_x \sigma_y}$$

$$= \mathbb{E}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_y}{\sigma_y}\right)\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) f_{X,Y}(x,y)\, dx\, dy$$

$$S = \left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)$$

$$t = \frac{x-\mu_x}{\sigma_x}$$

$$\Rightarrow \quad x = \sigma_x t + \mu_x$$

$$y = \sigma_y \frac{S}{t} + \mu_y$$

$$|J| = \begin{vmatrix} 0 & \sigma_x \\ \frac{\sigma_y}{t} & -\frac{\sigma_y s}{t^2} \end{vmatrix} = \frac{\sigma_x \sigma_y}{t}$$

$$\rho_{xy} = \mathbb{E}\left[\left(\frac{X-\mu_x}{\sigma_x}\right)\left(\frac{Y-\mu_Y}{\sigma_y}\right)\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} s \cdot \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(\frac{-1}{2(1-\rho^2)}\left(t^2 - 2\rho s + \frac{s^2}{t^2}\right)\right) \left|\frac{\sigma_x\sigma_y}{t}\right| ds\, dt$$

$$\vdots$$

$$\boxed{\rho_{xy} = \rho}$$

Note one Important fact:

All of the normal marginal and conditional pdf's are derived from the starting point of bivarite normality. The derivation does not go in the opposite direction. That is, marginal normality does not imply joint probability.