Lecture 22

Continuing our discussion on jointly distributed random variables

Expected Value

Let X, Y be jointly distributed r.v and $h(x,y)$ be an arbitrary function of X, Y.

Then

$$E[h(X,Y)] = \sum_x \sum_y h(x,y)\, p(x,y) \quad (X, Y \text{ are discrete})$$

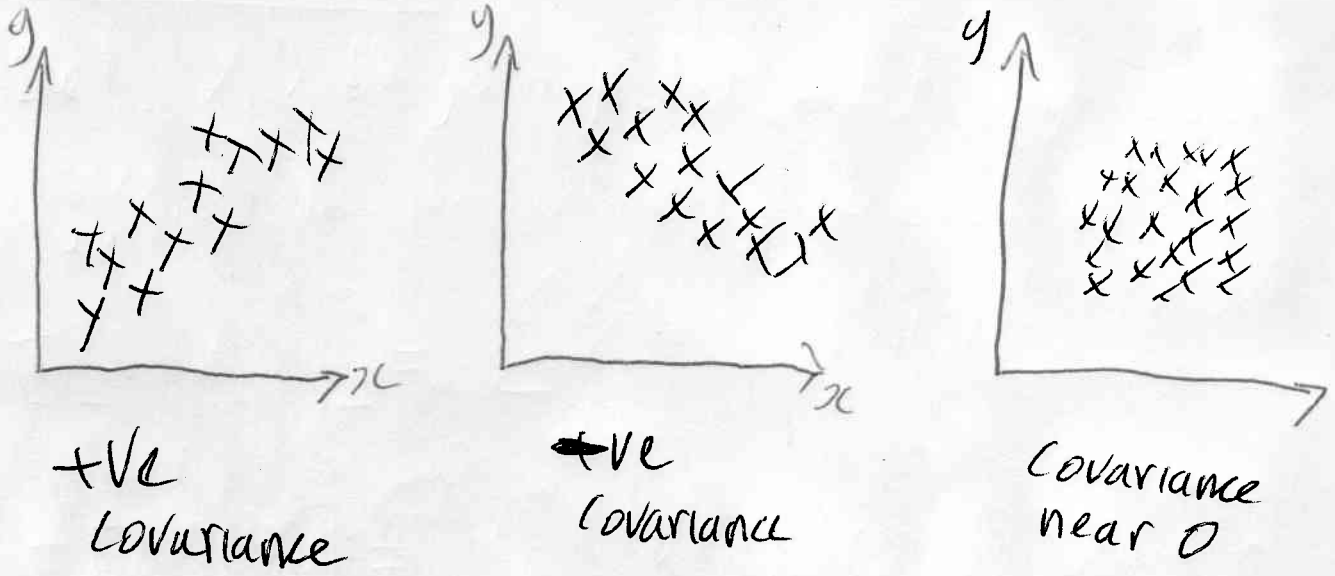
$$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y)\, f(x,y)\, dy\, dx \quad (X, Y \text{ are continuous})$$

Covariance is a measure of how strongly two variables are related to each other.

$$Cov(X,Y) = E((X - \mu_x)(Y - \mu_y)]$$

$$Cov(X,Y) = \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \qquad X, Y \text{ discrete}$$

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)f(x,y)\,dy\,dx \qquad X, Y \text{ continuous}$$



+Ve Covariance

-Ve Covariance

Covariance near 0

Note that for computational purposes it is easier to use

$$Cov(X,Y) = E(XY) - \mu_x\mu_y$$

Proof (for continuous case)

$$Cov(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_x)(y-\mu_y)f(x,y)\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy - \mu_x y - \mu_y x + \mu_x\mu_y)f(x,y)\,dy\,dx$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\,dy\,dx - \mu_x \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\,dy\,dx}_{=\mu_y}$$

$$- \mu_y \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\,dy\,dx}_{=\mu_x} + \mu_x\mu_y \underbrace{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dy\,dx}_{=1}$$

$$= E[XY] - 2\mu_x\mu_y + \mu_x\mu_y$$

$$= E[XY] - \mu_x\mu_y$$

Unfortunately covariance depends on the units of measurement. Therefore changing the units of measurements can have considerable effect on the magnitude of the covariance. To remedy this problem we use a rescaled version of the covariance.....

Correlation

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

(If we rescale by dividing by the two standard deviations)

Note that the following are properties of correlation

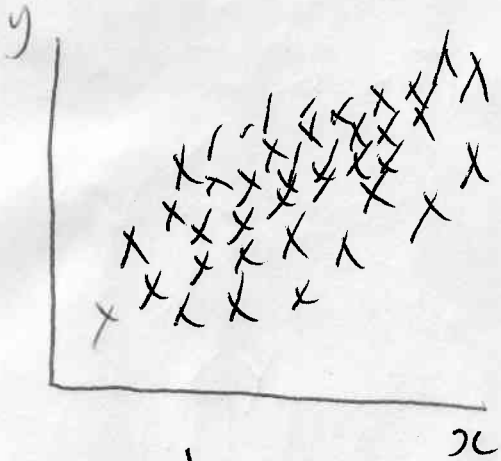
1. $-1 \leq \rho_{X,Y} \leq 1$ for any r.v X, Y

2. $\rho_{aX+b, cY+d} = \rho_{X,Y}$ (i.e rescaling has no effect on correlation)

3. If X, Y are independent $\rho_{X,Y} = 0$

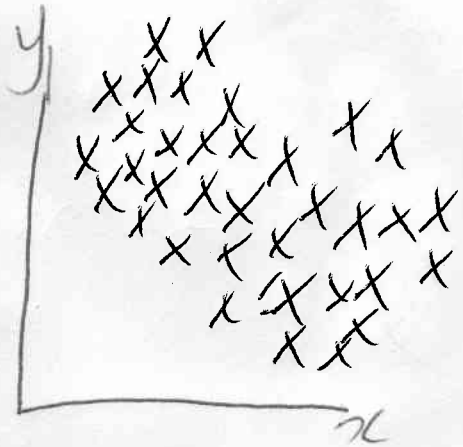
4. $\rho = 1$ or $-1 \iff Y = aX+b$ for some a, b $(a \neq 0)$

Interpreting correlation



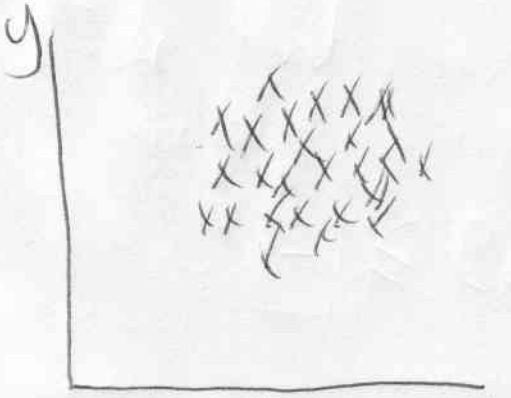
positive relationship

$\rho_{X,Y} > 0$

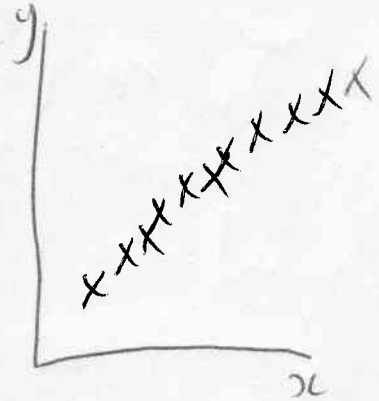


negative relationship

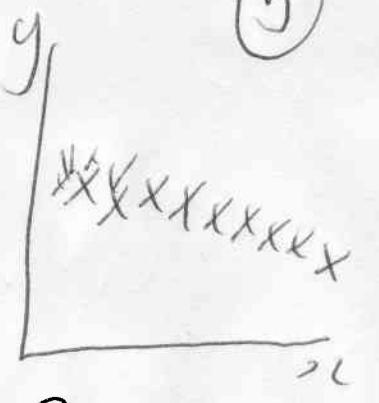
$\rho_{X,Y} < 0$



no "linear" relationship
$\rho_{x,y} \approx 0$



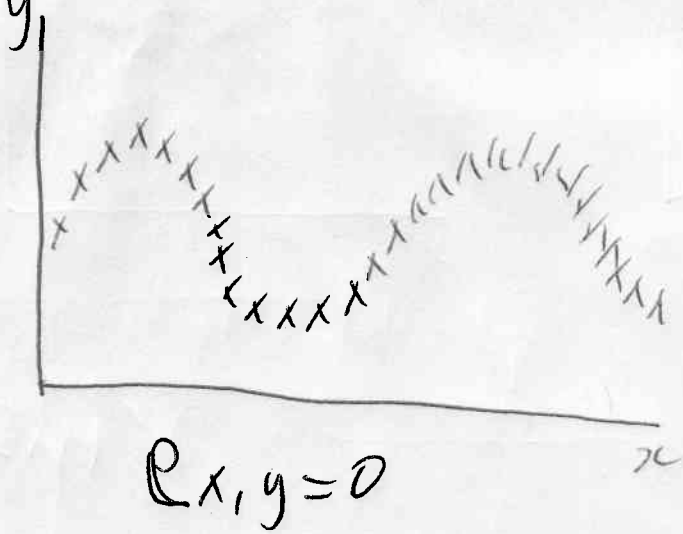
$\rho_{x,y}$ near +1
strong positive linear relationship



$\rho_{x,y}$ near -1
strong negative linear relationship

Note that correlation is only a measure of the strength of the linear relationship between x, y. It is possible to have a very strong non-linear relationship between x, y and have $\rho_{x,y} = 0$

eg



$\rho_{x,y} = 0$

Examples

1. Recall from last time. Joint distribution of X, Y

$$f(x,y) = k(x^2+y^2) \quad 20 \le x \le 30, \quad 20 \le y \le 30$$

$$= 0 \quad \text{o.w} \quad k = \frac{3}{380000}$$

what is $Cov(X,Y)$?

Recall that $Cov(X,Y) = E(XY) - E(X)E(Y)$

what is $E[X]$?

$$E[X] = \int_{20}^{30} x f_X(x)\,dx$$

$$= k\int_{20}^{30} x\left[10x^2 + \frac{30^3-20^3}{3}\right]dx$$

$$= k\left[\frac{10x^4}{4} + \frac{30^3-20^3}{6}x^2\right]\Bigg|_{20}^{30}$$

$$= k\left[\frac{10(30^4-20^4)}{4} + \frac{(30^3-20^3)(30^2-20^2)}{6}\right]$$

$$= 25.3289$$

Similarly it can be shown that

$E[Y] = 25.3289$

$$E[XY] = k\int_{20}^{30}\int_{20}^{30} xy\left[x^2+y^2\right] dx\,dy$$

$$= k\int_{20}^{30}\left[\frac{x^3y^2}{2}+\frac{xy^4}{4}\right]_{20}^{30} dx$$

$$= k\int_{20}^{30}\left[\frac{x^3(30^2-20^2)}{2}+x\frac{(30^4-20^4)}{4}\right] dx$$

$$= k\left[\frac{x^4(30^2-20^2)}{8}+\frac{x^2(30^4-20^4)}{8}\right]_{20}^{30}$$

$$= k\left[\frac{(30^4-20^4)(30^2-20^2)}{8}+\frac{(30^2-20^2)(30^4-20^4)}{8}\right]$$

$$= \frac{3}{380000}\left[81250000\right] = 641.4474$$

So $COV(X,Y) = 641.4474 - (25.3289)(25.3289)$

$= -.1078$ (4dp)

What is $Cor(X,Y)$?

$$Cor(X,Y) = \rho_{x,y} = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

$$Var(X) = E(X^2) - (E(X))^2 \qquad Var(Y) = E(Y^2) - (E(Y))^2$$

$$E(X^2) = \int_{20}^{30} x^2 f_X(x)\,dx$$

$$= K \int_{20}^{30} x^2 \left[10x^2 + \frac{30^3 - 20^3}{3}\right] dx$$

$$= K \left[\frac{10x^5}{5} + \frac{30^3 - 20^3}{9} x^3\right]_{20}^{30}$$

$$= \frac{3}{380000} \left[(30^5 - 20^5) + \frac{(30^3 - 20^3)^2}{9}\right]$$

$$= 649.8246$$

And Similarly it can be shown

$$E[Y^2] = 649.8246$$

So

$$Var(X) = 649.8246 - (25.3289)^2$$

$$= 8.2694$$

and $Var(Y) = 8.2694$

$$\Rightarrow \sigma_X = SD(X) = \sqrt{Var(X)} = \sqrt{8.2694}$$

$$= 2.8757$$

$$\sigma_Y = 2.8757$$

So $$Cor(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{-.1078}{(2.8757)(2.8757)}$$

$$= -.0130 \quad (4dp)$$

So there is almost no linear relationship between x and y. (Note that does not mean a strong non linear relationship might exist).