Lecture 15

Topics:

I) correlation

II) Bivariate Normal

III) prediction with Bivariate Normal.

Recall that the covariance btwn two R.V.s $cov(X,Y)$ is the number which is the missing link for computing the variance of $aX+bY$ when X & Y are dependent.

In some sense $cov(X,Y)$ measures "linear dependence". However, it is somewhat hard to interpret since it changes when the units of measurement are changed.

e.g. Randomly select a from california.

Let X_g = their ave sodium intake per day (in grams)

Y = their systolic blood pressure in mm Hg.

Converting grams to ounces.

X_o = their ave sodium intake per day (in ounces)

$$= 0.035274\, X_g.$$

Notice that a quantification of dependence should be the same for (X_g, Y) as for (X_o, Y).

Unfortunately

$$cov(X_g, Y) \neq \underbrace{cov(X_o, Y)}_{\substack{\text{this is actually} \\ 0.035274\, cov(X_g, Y).}}$$

So we need a measure of "linear dependence" that doesn't change when changing units (i.e. that is unit free).

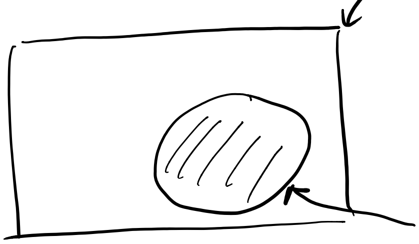
Definition: The correlation between two R.V.s (X,Y), denoted ρ or $\rho_{X,Y}$, is defined to be

$$\rho = \frac{cov(X,Y)}{sd(X)\, sd(Y)}$$

Fact 1 ρ is unit free, i.e. ρ stays the same when X or Y (or both) are converted to different units.

Fact 2

- $-1 \leq \rho \leq 1$ always
- if $\rho < 0$ then $cov(X,Y) < 0$
- if $\rho > 0$ then $cov(X,Y) > 0$.
- If X & Y are independent then $\rho = 0$ (but not the other way around). Picture:



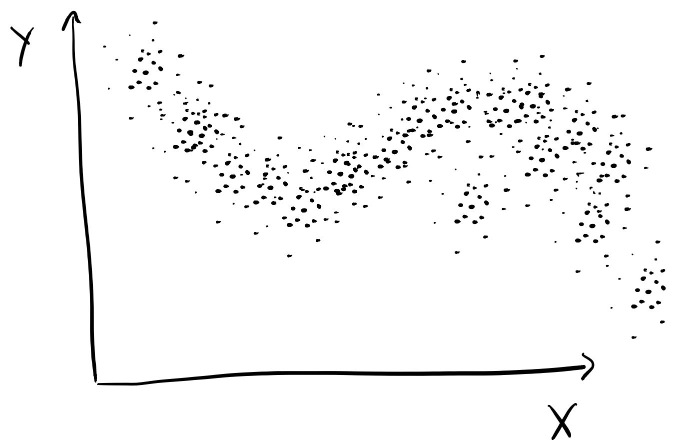
Fact 3

If $\rho = 1$, (X,Y) will always be on a fixed (non-random) line with positive slope

If $\rho = -1$, (X,Y) will always be on a line with negative slope.

Visualizing the joint PMF of a pair of R.V.s (X,Y)

To visualize the dependence between X & Y one can make a scatter plot of all possible pairs (X,Y) with repeats in proportion to probabilities



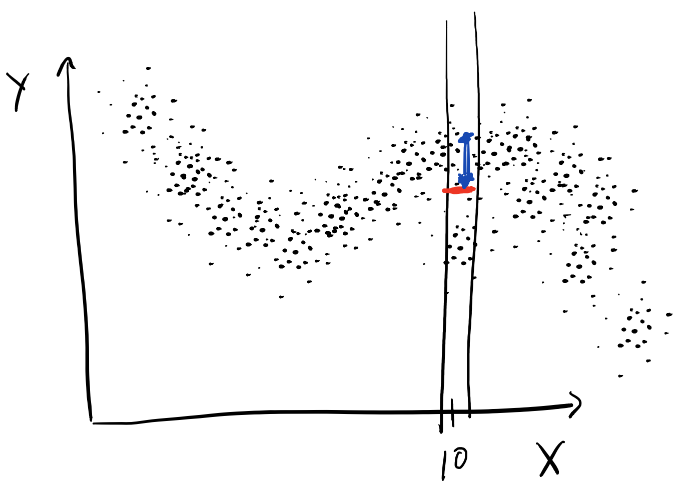
You can imagine that a random sample (X,Y) corresponds to the coordinates of a random dot drawn from this scatter plot.

Equivalently the coordinates of each dot represents the pair of numbers on the tickets in a box model for (X,Y).

With a box model $\boxed{\begin{matrix} X \\ Y \end{matrix}}$ for (X,Y)

Recall that $E(Y|X=x)$ & $sd(Y|X=x)$ are basically the ave & standard deviation of the Y value's on all tickets of the form $\boxed{\begin{matrix} x \\ Y \end{matrix}}$.

$\therefore$ $E(Y|X=x)$ & $sd(Y|X=x)$ can be visualized as follows.

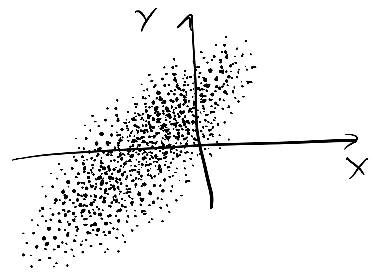


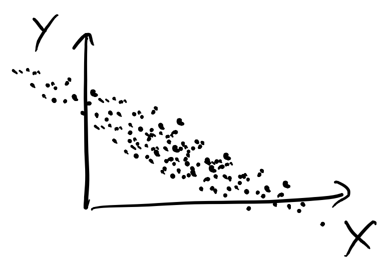
—: $E(Y|X=10) \approx$ the ave Y value in the column above 10

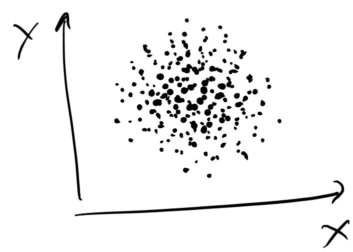
I: $sd(Y|X=10) \approx$ the s.d. of the Y values in the column above 10.

Bivariate Normal (X,Y)

Bivariate Normal (X,Y) is a special type of (continuous) pair of R.V.s which have scatter plots that look like this:







Def: (X,Y) is Bivariate normal if

1. X is normal
2. Y is Normal
3. For any x the conditional PDF of Y given $X=x$ is also Normal

As a consequense of the definition its also true that X given $Y=y$ is also Normal.

Predicting Y given X for Bivariate Normals

It is typically computationally difficult to compute the best prediction & s.d. of Y given $X=x$.

For Bivariate Normals it is extremely easy & all you need to compute these is the following 5 numbers (sometimes called parameters):

$E(X)$, $sd(X)$

$E(Y)$, $sd(Y)$

ρ ← The correlation btwn X, Y.

This gives us our next "master formulas".

Suppose (X,Y) are bivariate Normal with correlation ρ.

Then

Master Formula 3:

$E(Y|X=x)$ is computed using

$$\left(\frac{E(Y|X=x)-E(Y)}{sd(Y)}\right)=\rho\left(\frac{x-E(X)}{sd(X)}\right)$$

Master Formula 4:

$$var(Y|X=x)=(1-\rho^2)\,var(Y)$$

e.g. Suppose X & Y are BN with

$E(X)=10$ $sd(X)=1$

$E(Y)=1$ $sd(Y)=5$

$\rho=0.6$

Suppose you observe $X=11.5$. Predict Y & find the typical error of this prediction.

9

$$\frac{x - E(X)}{sd(X)} = \frac{11.5 - E(X)}{sd(X)} = \frac{11.5 - 10}{1} = 1.5$$

$\therefore$ By MF3:

$$\frac{E(Y|X=11.5) - E(Y)}{sd(Y)} = \frac{E(Y|X=11.5) - 1}{5}$$

$$\overset{MF3}{=} \rho\left(\frac{x - E(X)}{sd(X)}\right)$$

$$= 0.6 \cdot (1.5)$$

$$= 0.9$$

$$\therefore \quad \frac{E(Y|X=11.5) - 1}{5} = 0.9$$

$$\therefore \quad E(Y|X=11.5) = 5(0.9) + 1$$

$$= \underbrace{5.5}_{\text{best prediction of } Y \text{ given } X=11.5}$$

Also By MF4

$$var(Y|X=11.5) = (1 - \rho^2)\, var(Y)$$

$$= (1 - 0.6^2) \cdot 5^2$$

$$= 16.0$$

$\therefore$ the typical prediction error using 5.5 to predict Y when X is obs to be 11.5 is

$$sd(Y|X=11.5) = \sqrt{16} = 4$$

10