Lecture 31

Agenda

1. Covariance and correlation

2. For Data points....

We will study about covariance and correlation between two random variables in this lecture. Parts of this lecture are similar to lecture 16, but there we did things for discrete random variables, now we will study them for general random variables.

Covariance

We know that if X and Y are two independent random variables then for any two functions $f()$ and $g()$,

$$E[f(X)g(Y)] = E[f(X)] \times E[g(Y)].$$

In particular $E(XY) = E(X)E(Y)$, i.e. $E(XY) - E(X)E(Y) = 0$. So in general,

Definition 1. *Let X and Y be two random variables, we define*

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

Just as $E(X)$, $V(X)$ helped us understand the probabilistic behaviour of X in specific ways, $Cov(X,Y)$ helps us understand the joint behaviour of X and Y as follows :

- If Y tend to increase when X increases and decrease when X decreases then $Cov(X,Y)$ is positive.

- If Y tend to increase when X decreases and increase when X decreases then $Cov(X,Y)$ is negative.

The following properties are useful,

1. $Cov(X,Y) = E[(X - E(X)) * (Y - E(Y))]$

2. $Cov(X_1 + X_2, X_3) = Cov(X_1, X_3) + Cov(X_2, X_3)$

3. $Cov(bX, Y) = b * Cov(X,Y)$ for some constant b

4. $Cov(X, X) = V(X)$

We now define correlation,

Definition 2. *The correlation coefficient between two random variables X and Y is defined as,*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

Without any intermediate discussion let's look at the properties of ρ_{XY}, because these explain the use of ρ_{XY}.

Properties of ρ_{XY}

1. ρ_{XY} is unitless.
2. $-1 \leq \rho_{XY} \leq 1$
3. If $\rho_{XY} = 1$ then $Y = a + bX$ for some $b > 0$.
4. If $\rho_{XY} = -1$ then $Y = a + bX$ for some $b < 0$.
5. ρ_{XY} measures the linear relationship between X and Y. As ρ_{XY} becomes closer to 0, X and Y become less and less linearly related. Thus the sign of ρ_{XY} gives the direction of the relationship and it's absolute value gives the strength.

Lemma 1. *If X and Y are independent, then*

$$\rho_{XY} = Cov(X,Y) = 0$$

Proof. Easy. □

But the converse of this result is not true.

Counterexample 1

(X, Y) have the following joint distribution

$$P(X = x, Y = y) = \frac{1}{8}$$

if $x \in \{-1, 0, 1\}$ and $y \in \{-1, 0, 1\}$ but $(x, y) \neq (0, 0)$. Thus we can check the following,

$$E(XY) = \sum_{x\in\{-1,0,1\}} \sum_{y\in\{-1,0,1\}} x * y * P(X = x, Y = y) = 0$$

$$E(X) = \sum_{x \in \{-1,0,1\}} x * P(X = x) = 0$$

$$E(Y) = \sum_{y \in \{-1,0,1\}} y * P(Y = y) = 0$$

Hence $Cov(X, Y) = 0$. But

$$P(X = 1)P(Y = 1) = \frac{3}{8} * \frac{3}{8} \neq \frac{1}{8} = P(X = 1, Y = 1)$$

which proves that X and Y are not independent.

Counterexample 2

Let the density of X be symmetric, i.e.

$$f_X(-x) = f_X(x)$$

for all $x \in \mathbb{R}$. Define $Y = X^2$; surely X and Y are not independent, but we can check that $Cov(X, Y) = 0$. `(HW)`

The properties of covariance that we just listed down for 2 or 3 random variables can be generalized as follows......

Lemma 2. *Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ be random variables. Then for constants $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_m$*

1. $Cov(a_1X_1 + a_2X_2 + \ldots + a_nX_n, b_1Y_1 + b_2Y_2 + \ldots + b_mY_m) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_ib_jCov(X_i, Y_j)$
2. $V(a_1X_1 + a_2X_2 + \ldots + a_nX_n) = \sum_{i=1}^{n} a_i^2 V(X_i) + 2\sum_{1 \le i < j} a_ia_jCov(X_i, X_j)$

For Data points.........

For a given random variable X we have defined $E(X)$ and $V(X)$. But these concepts carry over when we have n data points, $\{X_1, X_2, \ldots, X_n\}$. In that case we define the sample mean as

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

Similarly we define the sample variance as,

$$s_X^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

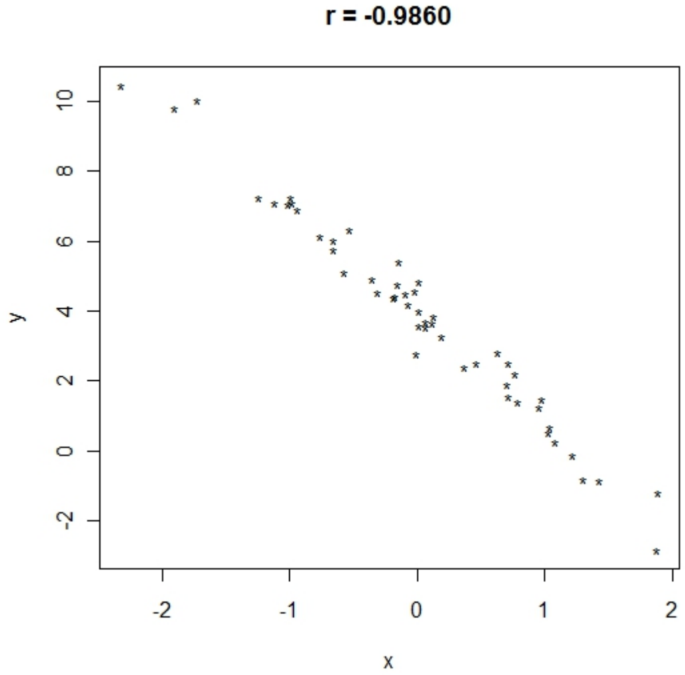

Figure 1:

You might have seen the use of $\frac{1}{n-1}$ instead of $\frac{1}{n}$ for the sample variance in some cases and there is a perfectly valid reason for that. But for now let's use the above formula.
When we have bivariate data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ we define the sample covariance as

$$covariance = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

and the sample correlation coefficient as

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{s_X s_Y}$$

Observe the similarity between these definitions for random variables and data points.
In figure 1 linear relationship is stronger than that in figure 2; which matches with the absolute value of r. In figure 1 r is negative and in figure 2 r is positive and that fact reflects in the direction of the plots.

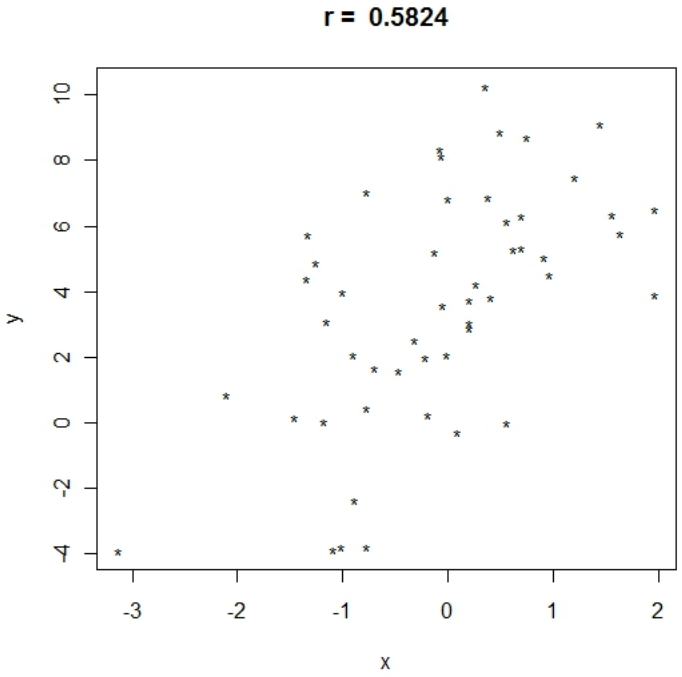

Figure 2: