**Covariance of random variables**

Suppose there are two random variables, X and Y. It is often of interest to determine if they are independent. If they are dependent, then we would like quantify the amount and strength of dependence. Also, we would be interested in the type (positive or negative) of dependence.

Positive dependence means that “large” values of X tend to occur with “large” values of Y. “Small” values of X tend to occur with “small” values of Y. If we could plot all values in a population, the dependence would look like this:



Negative dependence means that “large” values of X tend to occur with “small” values of Y. “Small” values of X tend to occur with “large” values of Y. If we could plot all values in a population, the dependence would look like this:



Thus, positive dependence means as values of X increase, Y tends to increase as well (they move in the same direction). Negative dependence means as values of X increase, Y tends to decrease (they move in an opposite direction).

What would no dependence (independence) look like in the above plot?

Example: High school and college GPA

Suppose I had a joint PDF which quantified the possible values for which high school and college GPAs can take on. Let X = the student’s high school GPA and Y = the student’s college GPA.

Questions:

* Would you expect there to be a relationship between X and Y? In other words, are X and Y independent or dependent?
* If they are dependent, would you expect there to be a strong or weak amount of dependence?
* If they are dependent, would you expect a positive or negative dependence? What would positive and negative dependence mean in terms of the problem?

The numerical measure of the dependence between two random variables is called the “covariance”. It is denoted symbolically by σxy when X and Y denote the two random variables. Below are a few notes about it:

* σxy = 0 when there is independence.
* σxy > 0 when there is positive dependence
* σxy < 0 when there is negative dependence
* The further away σxy is from 0, the stronger the dependence.

Let X and Y be random variables with joint PDF f(x,y). Suppose E(X) = μx and E(Y) = μy. The covariance of X and Y is

 

when X and Y are discrete, and



when X and Y are continuous.

Common notation that is often used for the covariance is Cov(X,Y) = σxy.

Question: What is σxx?

A simplifying formula to find the covariance (like what we saw for variance) is

σxy = E(XY) – μxμy = E(XY) – E(X)×E(Y)

Proof:

E[(X-μx)(Y-μy)] = E[XY - Yμx - Xμy + μxμy]

= E[XY] - μxE[Y] - μyE[X] + E[μxμy]

= E[XY] - μxμy - μyμx + μxμy

= E[XY] - μxμy

Note that E(XY) ≠ μxμy except under a particular condition to be discussed later.

There is one problem with the covariance:

The measure of strength of dependence (how far it is from 0) is not necessarily bounded above or below.

The correlation coefficient, denoted by ρxy, fixes this problem. It is the covariance divided by the standard deviations of X and Y in order to provide a numerical value that is always between -1 and 1. Below are a few notes about it:

* -1 ≤ ρxy ≤ 1
* ρxy = 0 when there is independence.
* ρxy > 0 when there is positive dependence
* ρxy < 0 when there is negative dependence
* The closer to 1 that ρxy is, the stronger the positive dependence.
* The closer to -1 that ρxy is, the stronger the negative dependence.
* When X = Y, ρxy = 1. More generally, when X = a + bY for constants a and b > 0, ρxy = 1.
* When X = -Y, ρxy = -1. More generally, when X = a + bY for constants a and b < 0, ρxy = -1.

Let X and Y be random variables with covariance σxy and standard deviations σx and σy, respectively. The correlation coefficient for X and Y is



Sometimes one will see this denoted as Corr(X,Y).

Example: Grades for two courses (Cov.ipnyb)

Let X be a random variable denoting grade in a math course and Y be a random variable denoting grade in a statistics course. Suppose the joint PDF is



To find the covariance, I am going to use the σxy = E(XY) - μXμY expression.

Find μx:

 







One could also find this expected value another way by using the marginal PDF for X found in a previous section: g(x) = x2 + 2/3. Thus,

 





Find μy:

 





Find E(XY):

Since both X and Y are involved in the expectation, the joint PDF must be used.









Then σxy = E(XY) - μxμy = 3/8 – (7/12)×(2/3)

= 3/8 – 7/18 = 27/72 – 28/72 = -1/72 = -0.0139

Sage:





To find the correlation coefficient, I need to find the variances of X and Y in addition to the covariance between X and Y. To do this, I am going to use the shortcut formulas of Var(X) =  = E(X2) -  and Var(Y) =  = E(Y2) - . Because the individual means have already been found, I just need to find E(X2) and E(Y2).

Find E(X2):

 





Find Var(X):

Var(X) = E(X2) -  = 19/45 – (7/12)2 = 0.0819

Find E(Y2):







Find Var(Y):

Var(Y) = E(Y2) -  = 23/45 – (2/3)2 = 3/45 = 1/15 = 0.0667

Then 

Sage :





Describe the relationship between math course grade (X) and stat course grade (Y):

* Are math and stat course grades independent or dependent? Explain.
* If they are dependent, is there a strong or weak amount of dependence?
* If they are dependent, is there a positive or negative relationship between math and stat course grades?

On an exam, I may just ask you to describe the relationship between two random variables instead of prompting you with the above questions. In your explanation, you should still address these types of questions!

What we have developed is a way to understand the relationship between two different random variables. Where would this be useful? Suppose you want to study the relationships between:

* Humidity and temperature
* ACT and SAT score
* White and red blood cell counts
* Winning percentage and the number of yards per game on offense for NFL teams
* …