Covariance and Correlation

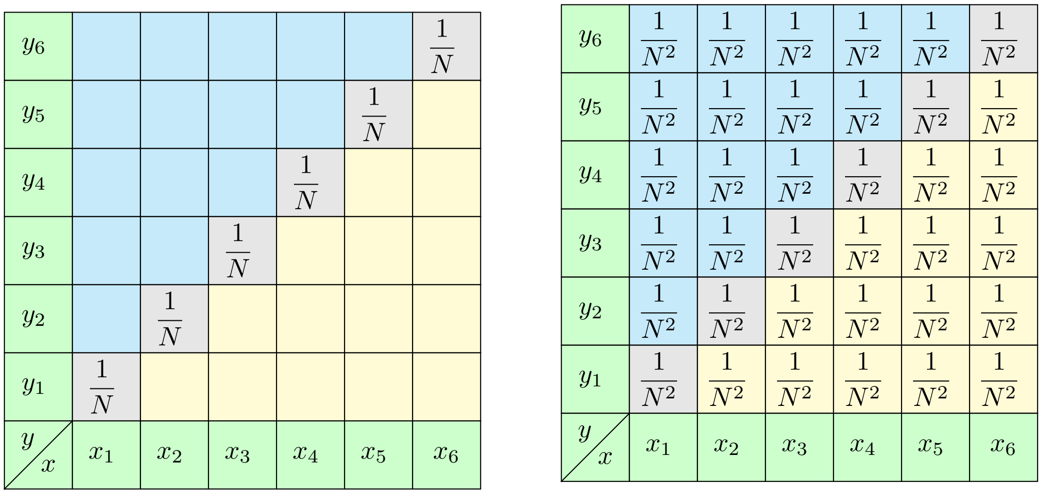
Parthiban Rajendran

November 7, 2018

1 Generalization

So far we have seen Covariance for discrete X, Y random variables. This could easily be transferred to continuous variables as well. However before generalization of the formula, we need to generalize the way the sample set is provided as well.

Suppose the sample set is given as $(X, Y) = (x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_N, y_N)$ then, if we say equi probable, then $p(X, Y)$ could be simply tabulated in different ways depending on the function $h(X, Y)$ that is, if we take the deformed or standard formula. This is illustrated in figure 1.



y_6						$\frac{1}{N}$
y_5					$\frac{1}{N}$	
y_4				$\frac{1}{N}$		
y_3			$\frac{1}{N}$			
y_2		$\frac{1}{N}$				
y_1	$\frac{1}{N}$					
y / x	x_1	x_2	x_3	x_4	x_5	x_6

Plot A
$p(X, Y)$ for standard formula
$h(X, Y) = (X - \overline{X})(Y - \overline{Y})$

y_6	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_5	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_4	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_3	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_2	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y_1	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$	$\frac{1}{N^2}$
y / x	x_1	x_2	x_3	x_4	x_5	x_6

Plot B
$p(X, Y)$ for deformed formula
$h(X_i, Y_i, X_j, Y_j) =$
$(X_i - X_j)(Y_i - Y_j)$

$p(X, Y)$ depending on $h(X, Y)$

This was simply because, of the way we indexed the sample points. In Plot A, we do not have a (x_2, y_1), because we just numbered as $(x_1, y_1), (x_2, y_2), (x_3, y_3) \cdots (x_N, y_N)$, and it worked because standard formula needed only one time indexing via i. But in Plot B, we had double indexing via i, j, this is why the probability at each *cell* also became $1/N^2$. Most often we do not use the deformed formula and stick to standard formula. Further, often the given probability density function (if given), would be something like this.

		y		
$\rho(x,y)$		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

Here our indexing style has to differ. Now we have $(x_1, x_2) = (100, 250)$ and $(y_1, y_2, y_3) = (0, 100, 200)$. If we line up these sample pairs, we get

$$(x_1, y_1), (x_1, y_2), (x_1, y_3), (x_2, y_1), (x_2, y_2), (x_2, y_3)$$

Thus even with standard formula due to data being in a different format, we would need to use double summation in order to vary i and j to different limits separately. Thus naturally our standard formula would become

$$\mathrm{Cov}(X, Y) = \sum_{i=1}^{2} \sum_{j=1}^{3} (x_i - \overline{x})(y_j - \overline{y}) p(x_i, y_i)$$

Generalizing the standard formula, and also extending to continuous X and Y, we could say,

Generalized Standard Covariance Formula

The **covariance** between two rv's X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_Y)]$$

$$= \begin{cases} \sum_x \sum_y (x - \mu_x)(y - \mu_y) p(x, y) & \text{X,Y discrete} \\ \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f(x, y) dx dy & \text{X,Y continuous} \end{cases} \tag{1}$$

Depending on samples are from population or we deal with entire population, either $\overline{x}$ or μ_X could be used respectively.

2 Example

We have already explained the concept with an example, so here will see a different approach.

Suppose joint and marginal pmf's for X = automobile policy deductible amount and Y = homeowner policy deductible amount are as below. Find the covariance.

		y		
$\rho(x,y)$		0	100	200
x	100	.20	.10	.20
	250	.05	.15	.30

x	100	250
$\rho_X(x)$	.5	.5

y	0	100	200
$\rho_Y(y)$	.25	.25	.5

This example was taken from devore2011 Since we need the means in the equation, let us calculate them first.

$$\mu_x = \sum_{i=1}^{2} x_i p_X(x_i) = 100(0.5) + 250(0.5) = 175\mu_y = \sum_{i=1}^{2} y_i p_Y(y_i) = 0(0.25) + 100(0.25) + 200(0.5) = 125$$

Coming to Covariance,

$$\text{Cov}(X,Y) = \sum_{i=1}^{2}\sum_{j=1}^{3}(x_i - \mu_x)(y_j - \mu_y)p(x_i, y_j)$$

$$= (x_1 - 175)(y_1 - 125)p(x_1, y_1) + (x_1 - 175)(y_2 - 125)p(x_1, y_2) + (x_1 - 175)(y_3 - 125)p(x_1, y_3)$$
$$+(x_2 - 175)(y_1 - 125)p(x_2, y_1) + (x_2 - 175)(y_2 - 125)p(x_2, y_2) + (x_2 - 175)(y_3 - 125)p(x_2, y_3)$$

$$= (100 - 175)(0 - 125)p(100, 0) + (100 - 175)(100 - 125)p(100, 100) + (100 - 175)(200 - 125)p(100, 200)$$
$$+(250 - 175)(0 - 125)p(250, 0) + (250 - 175)(100 - 125)p(250, 100) + (250 - 175)(200 - 125)p(250, 200)$$

$$= (100 - 175)(0 - 125)0.20 + (100 - 175)(100 - 125)0.10 + (100 - 175)(200 - 125)0.20$$
$$+(250 - 175)(0 - 125)0.05 + (250 - 175)(100 - 125)0.15 + (250 - 175)(200 - 125)0.30$$

$$= 1875$$

What just happpened? How come we took all possible pairs of (x, y) given in joing pmf as *samples*? Earlier, when we visualized TIA for random samples, we assumed that $h(X, Y)$ had equal probability for all of its values, thus resulting in a constant probability for entire summation. So it was enough if we look at it from the sky or top or whatever. If the probability density in the summation is a variable, then just by looking at 2D, we are missing the *contribution* of pmf to the summation. Now that we have varying pmf for different pairs of x, y, we need to account for that, because pairs having higher probability will attract more samples than those that would not, thus potentially forming a relationship between X and Y. This is evident the moment we visualize in 3D as shown in figure **??**. In 3D, it is evident now, the green has more volume, than red, so we could expect higher samples in these region than the red, thus suggesting in fact a *positive* correlation. Thus, yeah it is no more just a TIA ,but **total interested volume, TIV**. Also, a pmf resembles all possible values of (x, y), so could imagine, sample set of all possible values in any multiples (1 occurance per pair, or 10 occurance per pair, etc).

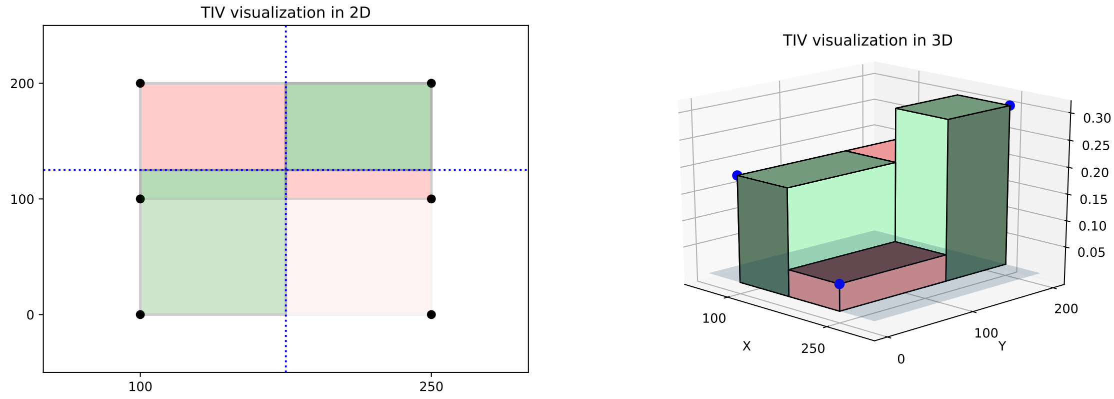


Figure 6: The Visualization of standard formula in 2D and 3D

Generalized Standard Covariance Visualization

The better generalized visualization of standard covariance formula is in volume, if underlying joint probability density function is not a constant.

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \sum_x \sum_y (x_i - \overline{x})(y_i - \overline{y})p(x_i, y_i) \\ &= (x_1 - \overline{x})(y_1 - \overline{y})p(x_1, y_1) + \cdots + (x_i - \overline{x})(y_i - \overline{y})p(x_i, y_i) + \cdots \\ &= V_{11} + \cdots + V_{ij} + \cdots\end{aligned} \tag{2}$$