1 Formalization of Sample and Population Correlation

The *standardized covariance* with its unique characteristic is thus called **Pearson's Correlation Coefficient**, **r** as it was formalized by Pearson. It is not required to standardize the sample set everytime, and calculate the standardized covariance as slope of the resultant regression line. We could calculate directly from the given sample set as below.

$$r = \operatorname{cov}(X_s, Y_s) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{is} - \overline{x_s})(y_{is} - \overline{y_s})$$

Since standardized,

$$\overline{x_s} = \overline{y_s} = 0$$

$$x_{is} = \frac{x_i - \overline{x}}{s_X} \quad , \quad y_{is} = \frac{y_i - \overline{y}}{s_Y}$$

$$\begin{aligned}
\therefore r = \frac{1}{N-1}\sum_{i=1}^{N}(x_{is})(y_{is}) &= \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right) \\
&= \frac{1}{s_X s_Y}\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) \\
&= \frac{\operatorname{cov}(X,Y)}{s_X s_Y}
\end{aligned}$$

Thus the *sample correlation coefficient* **r** of a given sample set (X, Y) is given by

$$r = \frac{\operatorname{cov}(X,Y)}{s_X s_Y}$$

By analogy, a *population correlation coefficient* could also be derived. If (X, Y) are two discrete RVs, with $X = x_1, x_2, \cdots, x_N$, and $Y = y_1, y_2, \cdots, y_M$, and if $p(X,Y), p(X), p(Y)$ are their joint and marginal *pmf*s respectively, then a population correlation coefficient ρ could be defined as,

$$\rho = \frac{\sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(X,Y)}{\sqrt{\sum_x (x - \mu_X)^2 p(X) \sum_y (y - \mu_Y)^2 p(Y)}} \tag{1}$$

where, $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ are respective population parameters of X and Y. Recalling Covariance and Variance formula for population as below,

$$\begin{aligned}
\operatorname{Cov}(X,Y) = \sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(X,Y) \\
\sigma_X^2 = \sum_x (x - \mu_X)^2 p(X) \\
\sigma_Y^2 = \sum_y (y - \mu_Y)^2 p(Y)
\end{aligned}$$

and using that, one could rewrite ρ as

$$\rho = \frac{\sum_x \sum_y (x - \mu_X)(y - \mu_Y)p(X,Y)}{\sqrt{\sum_x (x - \mu_X)^2 p(X) \sum_y (y - \mu_Y)^2 p(Y)}} = \frac{\operatorname{Cov}(X,Y)}{\sigma_X \sigma_Y} \tag{2}$$

Sample and Population Correlation

The *sample correlation coefficient*, r of any given sample set (X, Y) is given by

$$r = \frac{\text{cov}(X, Y)}{s_X s_Y} \tag{3}$$

The *population correlation coefficient*, ρ of any given discrete RVs (X, Y) is given by

$$\rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4}$$

Similar ρ applicable to continuous RVs also, with integration suitably placed in place of summation.

2 Cosine Similarity

Interestingly correlation factor could be visualized to an extent in vector form or at least provides us easier computational method of calculation via matrices. Suppose there is a sample set (X, Y) of size 3. That is, if $(X, Y) = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\}$, we could represent them in a 3D vector form as below

$$\vec{x} = x_1\hat{i} + x_2\hat{j} + x_3\hat{k}$$
$$\vec{y} = y_1\hat{i} + y_2\hat{j} + y_3\hat{k}$$

In simpler matrix notation,

$$\vec{x} = [x_1, x_2, x_3]$$
$$\vec{y} = [y_1, y_2, y_3]^T$$

Using law of cosines, the angle θ between vectors $\vec{x}, \vec{y}$ can be calculated as

$$\text{cos}\theta = \frac{\vec{x} \bullet \vec{y}}{\|x\|\|y\|}$$

where

$$\vec{x} \bullet \vec{y} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \bullet \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1y_1 + x_2y_2 + x_3y_3 = \sum_i^3 x_iy_i$$

and

$$\|x\| = \sqrt{x_1^2 + x_2^2 + x_3^2} = \sqrt{\sum_i x_i^2}$$
$$\|y\| = \sqrt{y_1^2 + y_2^2 + y_3^2} = \sqrt{\sum_i y_i^2}$$

Readers are strongly advised to go through appendix **??** where the concept is explained in detail and also concluded that the above relation is applicable to any higher dimensional vector. Thus, recalling equation **??** from appendix, if the sample set size is N, then we could represent in matrix form and extend the cosine relationship as follows.

Let

$$\vec{x} = [x_1, x_2, x_3, \cdots, x_N]$$
$$\vec{y} = [y_1, y_2, y_3, \cdots, y_N]^T$$

then,

$$\vec{x} \bullet \vec{y} = \begin{matrix} \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} & \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \end{matrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \bullet \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = x_1y_1 + x_2y_2 + \cdots + x_Ny_N = \sum_i^N x_iy_i$$

and

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_N^2} = \sqrt{\sum_i^N x_i^2}$$
$$\|y\| = \sqrt{y_1^2 + y_2^2 + \cdots + y_N^2} = \sqrt{\sum_i^N y_i^2}$$

so,

$$\cos\theta = \frac{\vec{x} \bullet \vec{y}}{\|x\|\|y\|} = \frac{\sum_i^N x_iy_i}{\sqrt{\sum_i^N x_i^2}\sqrt{\sum_i^N y_i^2}}$$

If we subtract the mean of the RVs, from each of the elements as below, setting up **centered** vectors,

$$\vec{x_c} = [x_1 - \overline{x}, x_2 - \overline{x}, x_3 - \overline{x}, \cdots, x_N - \overline{x}]$$
$$\vec{y_c} = [y_1 - \overline{y}, y_2 - \overline{y}, y_3 - \overline{y}, \cdots, y_N - \overline{y}]^T$$

this similarly leads to

$$\cos\theta = \frac{\vec{x_c} \bullet \vec{y_c}}{\|x_c\|\|y_c\|} = \frac{\sum_i^N (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^N (x_i - \overline{x})^2}\sqrt{\sum_i^N (y_i - \overline{y})^2}}$$

which is same as *sample correlation coefficient*, r. Note that, the value of cosine ranges between ±1. So when both vectors are in same direction, the θ is 0, thus $\cos\theta = 1$, maximum value indicating perfect linearity. Similarly when both vectors are in opposite direction, $\theta = 180°$, implying $\cos\theta$ = -1. When the vectors are perpendicular to each other, $\theta = 90°$ implying $\cos\theta = 0$, thus zero correlation.

For those, who find it difficult to comprehend higher dimensional vector, remember that in any higher dimensional vector, the angle between the resultant two vectors is always on a plane (2D), thus the law of cosine still applies. This is also explained in appendix **??**

Cosine Similarity

The *sample correlation coefficient*, r of any given sample set (X, Y) can also be expressed in vector matrix form, giving a cosine relationship as

$$r = \cos\theta = \frac{\vec{x_c} \bullet \vec{y_c}}{\|x_c\|\|y_c\|} = \frac{\text{cov}(X, Y)}{s_X s_Y} \tag{5}$$

where, $\vec{x_c}$ and $\vec{y_c}$ indicate *centered* dataset