[[```Distance covariance’’’]] in [[statistics]] and in [[probability theory]] is a new measure of how much two variables change together. Distance covariance was introduced in 2005 by [[Gabor J. Szekely]] in several lectures to address the most important deficiency of Pearson’s classical [[covariance]], namely that Pearson’s [[covariance]] can easily be zero for dependent variables. Thus covariance = 0 (uncorrelatedness) does not imply independence while distance covariance = 0 does imply independence. The first results on distance covariance were published in 2007 and 2009.

```Definition’’’

If $(X_{k,}\ Y_k)$, k=1,2,…, n is a sample from two variables, X and Y, (they can be real valued or vector valued), then take all pairwise distances of observations: $a_{k,l} := |X_k – X_l|$ and $b_{k,l} := |Y_k – Y_l|$ for k,l=1,2,…,n. Then center these distances such that in the centered distance matrices $(A_{k,l})$ and $(B_{k,l})$ all row sums and all column sums equal zero. The centered distances are $A_{k,l} := a_{k,l} – a_{k.} – a_{.\,l} + a_{..}$ and $B_{k,l} := b_{k,l} – b_{k\,.} – b_{.\,l} + b_{..}$ where $a_{k.}$ is the arithmetic average of the numbers $a_{k,l}$ , l=1,2,…,n (the meaning of $a_{.\,l}$ is similar), a.. is the arithmetic average of all distances $a_{k,l}$ k,l=1,2,…,n, and we have the same notation for the b values. The empirical distance covariance is simply the arithmetic average of the products $A_{k,l}\ B_{k,l}$ that is

$$\mathrm{dcov}_n(X,Y) := (1/n^2)\ \Sigma_{k\ l} A_{k,l}\ B_{k,l}\ .$$

```Properties of $\mathrm{dcov}_n$’’

(i) $\mathrm{dcov}_n (X,Y) \geq 0$.

(ii) $\mathrm{dcov}_n = 0$ if and only if every observation is the same.

The most important effect of working with centered distances is that the population value of distance covariance is zero if and only if X and Y are independent. The population value of distance covariance is

$$\mathrm{dcov}(X,Y) := E|X-X’||Y-Y’| + E|X – X’|\ E|Y – Y’| - E|X – X’||Y – Y”| - E|X – X”||Y – Y’|$$

where E denotes expected value, X’ is an independent and identically distributed copy of X, Y’ is an independent and identically distributed copy of Y, finally X” (Y”) has the same distribution as X (Y) and independent not only of X (Y) but also of Y and Y’ (X and X’). One can show that dcov(X, Y) always exists if X and Y have finite expected values. Distance covariance can be expressed with Pearson’s covariance, cov, as follows: dcov(X,Y) = cov(|X-X’|, |Y-Y’|) – 2cov(|X-X’|,|Y-Y”|). This identity shows that the distance covariance is not the same as the covariance of distances, cov(|X-Y|, |Y-Y’|), which can be zero even if X and Y are not independent.

````Properties of dcov’’’

(i) $\mathrm{dcov}(a_1+b_1C_1X,a_2+b_2C_2Y) = |b_1b_2|\mathrm{dcor}(X,Y)$
for all constant vectors $a_1$, $a_2$ , scalars $b_1$, $b_2$, and orthonormal matrices $C_1$, $C_{2.}$

(i) If the random vectors $(X_1, Y_1)$ and $(X_2, Y_2)$ are independent then
$\mathrm{dcov}(X_1+X_2, Y_1+Y_2) \leq \mathrm{dcov}(X_1, Y_1) + \mathrm{dcov}\ (X_2, Y_2)$
with equality if and only if $X_1$ and $Y_1$ are both constants or $X_2$ and $Y_2$ are both constants or $X_1$, $X_2$, $Y_1$, $Y_2$ are mutually independent.

$\mathrm{dcov}_n(X,Y)$ is a biased estimate of dcov(X,Y) because $\mathrm{Edcov}_n(X,Y) = [(n-1)/n^2][(n-2)\mathrm{dcov}(X,Y)+E|X-X'|E|Y-Y'|]$. Thus the bias can easily be corrected.

It is interesting to note that if $a_{k,l}$ and $b_{k,l}$ were defined as squared distances then dcov would simply become the square of Pearson’s covariance.

``Distance variance’’ is a special case of distance covariance when the two variables are identical. The empirical distance variance,+69-

$\mathrm{dvar}_n(X):=\mathrm{dcov}_n(X,X) = (1/n^2)\ \Sigma_{\ k\ l}A_{k,l}^{\ 2}$
````

is a relative of [[Corrado Gini]]’s [[mean difference]] introduced in 1912 but Gini did not work with centered distances.

The population value of distance variance is

dvar(X):= $E|X – X'|^2 + E^2|X-X'| - 2E|X-X'||X-X''|$.

```Properties of dvar’’’

(i) dvar(X) = 0 if and only if X = E(X) almost surely.
(ii) dvar(a + bCX) = |b|dvar(X) for all constant vectors a , scalars b, and orthonormal matrices C.

The square root of the distance variance is the ``distance standard deviation’’. The (empirical) ``distance correlation’’ of two variables is obtained by dividing the (empirical) distance covariance of the two variables by the product of their (empirical) distance standard deviations. The (empirical) distance correlation is denoted by ($\mathrm{dcor}_n$(X,Y)) dcor(X,Y).

```Properties of $\mathrm{dcor}_n$ and dcor’’’

(i) 0 ≤ dcor_n(X,Y) ≤ 1 and 0 ≤ dcor(X,Y) ≤1.
(ii) dcor(X,Y) = 0 if and only if X and Y are independent.
(iii) dcor_n(X,Y) = 1 implies that dimensions of the linear spaces spanned by X and Y respectively are almost surely equal and if we assume that these spaces are equal, then in this subspace Y = a + b CX for some vector a, scalar b and orthonormal matrix C.

References

*Gini, C. (1912). *Variabilità e Mutabilità*. Bologna: Tipografia di Paolo Cuppini.
*Pearson, K. (1895). Royal Society Proceedings, 58, 241.

*Pearson, K. (1920). Notes on the history of correlation, Biometrika, 13, 25-45.

* Székely, G. J. Bakirov, N. K., and Rizzo, L. M. (2007). Measuring and testing independence by correlation of distances, The Annals of Statistics, 35, 2769-2794.

* Székely, G.J. and Rizzo, M.L. (2009). Brownian distance covariance, The Annals of Applied Statistics, 3/4, 1233-1308.