Covarianza

En Espacios muestrales bidimensionales es interesante analizar la relación entre ambas variables.

La COVARIANZA es una medida de asociación o dependencia entre 2 variables aleatorias. La covarianza puede ser positiva o negativa (la varianza siempre es positiva). Si valores grandes de *X* a menudo dan como resultado valores grandes de *Y*, o valores pequeños de *X*, dan como resultado valores pequeños de *Y*, la covarianza será positiva. Caso contrario será negativa.

Si *X* e *Y* son variables aleatorias con función de distribución conjunta (fdc), $f_{x,y}(x, y)$, para el caso discreto (VAD) o función de densidad de probabilidad conjunta (fdpc) para el caso continuo (VAC), entonces la covarianza de (x,y) ó $COV(x, y) = \sigma_{xy}$ es:

$$\boxed{\sigma_{xy} = E[(x - \mu_x) \, . \, (y - \mu_y)]}$$

$$\sigma_{xy} = E[(x-\mu_x)\,.\,(y-\mu_y)]$$

$$\sigma_{xy} = \sum_x \sum_y (x-\mu_x)\,.\,(y-\mu_y)\,.\,f_{x,y}(x,y)$$ VAD

$$\sigma_{xy} = \int_{y=-\infty}^{+\infty}\int_{x=-\infty}^{+\infty} (x-\mu_x)\,.\,(y-\mu_y)\,.\,f_{x,y}(x,y)\,dx\,dy$$ VAC

Una fórmula alternativa de cálculo de COV es:

$$\boxed{\sigma_{xy} = E(xy) - \mu_x . \mu_y}$$

Donde, por definición del operador Esperanza: $E[Y] = \int_{-\infty}^{+\infty} Y\,.\,f(x)\,dx$

$$E(xy) \Longrightarrow E(xy) = \int_{y=-\infty}^{+\infty}\int_{x=-\infty}^{+\infty} x\,.\,y\,.\,f_{x,y}(x,y)\,dx\,dy$$

Demostrar la fórmula alternativa de la *COV:* $\sigma_{xy} = E(xy) - \mu_x . \mu_y$

para el caso continuo:

Partamos de la definición de *COV:* $\sigma_{xy} = E[(x - \mu_x) . (y - \mu_y)]$ que para una VAC:

$$\sigma_{xy} = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} (x - \mu_x) . (y - \mu_y) . f_{x,y}(x, y)\, dx\, dy$$

¿Qué hacemos ahora?

Apliquemos propiedad distributiva

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} [\underline{x. y. f_{x,y}(x, y)} - \mu_x . y . f_{x,y}(x, y) - x . \mu_y . f_{x,y}(x, y) + \mu_x . \mu_y . f_{x,y}(x, y)]\, dx\, dy$$

Demostrar la fórmula alternativa de la *COV:* $\sigma_{xy} = E(xy) - \mu_x . \mu_y$

para el caso continuo:

Partamos de la definición de *COV:* $\sigma_{xy} = E[(x - \mu_x) \, . \, (y - \mu_y)]$ que para una VAC:

$$\sigma_{xy} = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} (x - \mu_x) \, . \, (y - \mu_y) \, . \, f_{x,y}(x,y) \, dx \, dy$$

¿Qué hacemos ahora?  Apliquemos propiedad distributiva

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} [\underline{x . y . f_{x,y}(x,y)} - \underline{\mu_x \, . \, y \, . \, f_{x,y}(x,y)} - x \, . \, \mu_y \, . \, f_{x,y}(x,y) + \mu_x \, . \, \mu_y \, . \, f_{x,y}(x,y)] \, dx \, dy$$

Demostrar la fórmula alternativa de la *COV:* $\sigma_{xy} = E(xy) - \mu_x . \mu_y$

para el caso continuo:

Partamos de la definición de *COV:* $\sigma_{xy} = E[(x - \mu_x) . (y - \mu_y)]$ que para una VAC:

$$\sigma_{xy} = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} (x - \mu_x) . (y - \mu_y) . f_{x,y}(x, y)\, dx\, dy$$

¿Qué hacemos ahora?

Apliquemos propiedad distributiva

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} [\underline{x. y. f_{x,y}(x, y)} - \underline{\mu_x . y . f_{x,y}(x, y)} - \underline{x . \mu_y . f_{x,y}(x, y)} + \mu_x . \mu_y . f_{x,y}(x, y)]\, dx\, dy$$

Demostrar la fórmula alternativa de la *COV:* $\sigma_{xy} = E(xy) - \mu_x . \mu_y$

para el caso continuo:

Partamos de la definición de *COV:* $\sigma_{xy} = E[(x - \mu_x) . (y - \mu_y)]$ que para una VAC:

$$\sigma_{xy} = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} (x - \mu_x) . (y - \mu_y) . f_{x,y}(x, y)\, dx\, dy$$

¿Qué hacemos ahora?

Apliquemos propiedad distributiva

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} [\underline{x . y . f_{x,y}(x, y)} - \underline{\mu_x . y . f_{x,y}(x, y)} - \underline{x . \mu_y . f_{x,y}(x, y)} + \underline{\mu_x . \mu_y . f_{x,y}(x, y)}]\, dx\, dy$$

¿Y ahora? Separemos en varias integrales distribuyendo en los términos del integrando

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} [x.y.f_{x,y}(x,y) - \mu_x .y . f_{x,y}(x,y) - x .\mu_y .f_{x,y}(x,y) + \mu_x .\mu_y . f_{x,y}(x,y)]\, dx\, dy$$

sabiendo que μ_x y μ_y son constantes ya que son los promedios poblacionales:

$$\sigma_{xy} = \underbrace{\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} x .y .f_{x,y}(x,y)\, dx\, dy}_{1} - \mu_x \underbrace{\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} y .f_{x,y}(x,y)\, dx\, dy}_{2}$$

$$- \mu_y \underbrace{\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} x . f_{x,y}(x,y)\, dx\, dy}_{3} + \mu_x .\mu_y \underbrace{\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} f_{x,y}(x,y)\, dx\, dy}_{4}$$

1 $$E(xy) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} x .y .f_{x,y}(x,y)\, dx\, dy$$

2 Como: $$\int_{x=-\infty}^{x=\infty} f_{x,y}(x,y)\, dx = h(y) \rightarrow \mathit{Función\ marginal\ de\ Y}$$

$$\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} y . f_{x,y}(x,y)\, dx\, dy = \int_{y=-\infty}^{y=\infty} y .h(y)\, dy = E(y) = \mu_y$$

$$\sigma_{xy} = \int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} [x.y.f_{x,y}(x,y) \; - \mu_x \, . \, y \, . \, f_{x,y}(x,y) \; - x \, . \, \mu_y \, . \, f_{x,y}(x,y) \; + \mu_x \, . \, \mu_y \, . \, f_{x,y}(x,y)] \, dx \, dy$$

sabiendo que μ_x y μ_y son constantes ya que son los promedios poblacionales:

$$\sigma_{xy} = \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} x \, . \, y \, . \, f_{x,y}(x,y) \, dx \, dy}_{1} \; - \mu_x \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} y \, . \, f_{x,y}(x,y) \, dx \, dy}_{2}$$

$$- \mu_y \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} x \, . \, f_{x,y}(x,y) \, dx \, dy}_{3} + \mu_x \, . \, \mu_y \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} f_{x,y}(x,y) \, dx \, dy}_{4}$$

3 De igual forma que en el caso anterior: $\int_{x=-\infty}^{x=\infty}\int_{y=-\infty}^{y=\infty} x \, . \, f_{x,y}(x,y) \, dy \, dx \; = E(x) = \mu_x$

4 Y la integral doble: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{x,y}(x,y) \, dx \, dy \; = \; 1$

Reemplazando estos valores en la función anterior nos queda:

$$\sigma_{xy} = \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} x\,.\,y\,.\,f_{x,y}(x,y)\;dx\;dy}_{1} \;-\; \mu_x \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} y\,.\,f_{x,y}(x,y)\;dx\;dy}_{2}$$

$$- \mu_y \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} x\,.\;f_{x,y}(x,y)\;dx\;dy}_{3} + \mu_x\,.\,\mu_y \underbrace{\int_{y=-\infty}^{y=\infty}\int_{x=-\infty}^{x=\infty} f_{x,y}(x,y)\;dx\;dy}_{4\;=\;1}$$

$$\sigma_{xy} = E(xy) - \mu_x.\mu_y \cancel{- \mu_y.\mu_x} \cancel{+ \mu_x.\mu_y}$$

$$\boxed{\sigma_{xy} = E(xy) - \mu_x.\mu_y}$$

El signo de la covarianza indica la relación directa (+) o inversa (-) entre *X* e *Y*. Si son estadísticamente independientes, la covarianza dará *0*. Lo opuesto no es necesariamente cierto, es decir si la Covarianza da *0* no necesariamente son estadísticamente independientes. La covarianza sólo describe la relación lineal entre dos variables aleatorias, por consiguiente, si una covarianza entre *X* e *Y* es cero pero *X* e *Y* no tienen una relación lineal, no significa que son independientes. Serán independientes en el caso que la fdp o fdpc sean de carácter lineal.

Ejemplo: Sea X el tiempo de reacción, en segundos, de dos componentes químicos e Y la temperatura (en °C) a la cual inicia la reacción. Suponga que las dos variables aleatorias, X e Y, tienen la densidad conjunta:

$$f(x,y) = \begin{cases} 4xy & para \quad 0 \leq x \leq 1;\ 0 \leq y \leq 1 \\ 0 & para \qquad otro\ caso \end{cases}$$

Determine si esta fdpc es válida. En caso afirmativo, calcule la *COV* entre *X* e *Y* para determinar si existe relación entre la velocidad de reacción y la temperatura y si la misma es directa o inversa.

Solución:

a) $f(x,y) \geq 0\ ;\ \forall(x,y)$

b) $\displaystyle\int_{y=-\infty}^{+\infty}\int_{x=-\infty}^{+\infty} f(x,y)dx\,dy =$

$$= 4\int_{y=0}^{1}\int_{x=0}^{1} xy\,dxdy = 4\int_{y=0}^{1} y\left.\frac{x^2}{2}\right|_0^1 dy = 2\int_{y=0}^{1} y\ dy = 2\left.\frac{y^2}{2}\right|_0^1 = 1$$

$$\sigma_{xy} = E(xy) - \mu_x . \mu_y$$

$$E(xy) = \int_{y=-\infty}^{+\infty} \int_{x=-\infty}^{+\infty} xy \, . f_{x,y}(x,y) \, dx \, dy = \int_{y=0}^{1} \int_{x=0}^{1} xy . 4xy \, dx \, dy$$

$$E(xy) = 4 \int_{y=0}^{1} \int_{x=0}^{1} x^2 y^2 \, dx \, dy = 4 \int_{y=0}^{1} y^2 \left. \frac{x^3}{3} \right|_0^1 dy = 4 \int_{y=0}^{1} y^2 \frac{1}{3} dy = \frac{4}{3} \left. \frac{y^3}{3} \right|_0^1 = \frac{4}{9}$$

$$\mu_x = E(x) = \int_{x=-\infty}^{+\infty} x \, g(x) \, dx \quad \text{con} \quad g(x) = \int_{y=-\infty}^{+\infty} f_{x,y}(x,y) dy = 4x \int_{y=0}^{1} y \, dy = 2x$$

$$\mu_x = E(x) = \int_{x=-\infty}^{+\infty} x \, g(x) \, dx = 2 \int_{x=0}^{1} x^2 dx = \frac{2}{3}$$

$$\mu_y = E(y) = \int_{y=-\infty}^{+\infty} y \, h(y) \, dy = 2 \int_{y=0}^{1} y^2 dy = \frac{2}{3}$$

$$\sigma_{xy} = E(xy) - \mu_x . \mu_y = \frac{4}{9} - \frac{2}{3} . \frac{2}{3} = 0$$

Como sabemos que la fdpc es de carácter lineal, $f(x, y) = 4xy$, entonces podemos decir que las variables *X* e *Y* son independientes.

Ejemplo: El número de repuestos de bolígrafos azules *X* y el número de repuestos de bolígrafos rojos *Y* se encuentran en una caja para la venta. Cuando se seleccionan dos repuestos para bolígrafo al azar se obtiene la distribución de probabilidad conjunta siguiente:

f(x,y)		*X* 0	1	2	*h(y)*
Y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
g(x)		5/14	15/28	3/28	1

Calcule la covarianza de *X* e *Y*

<u>Solución</u>: $\sigma_{xy} = E(xy) - \mu_x . \mu_y$

$$E(xy) = \sum_x \sum_y x.y.f(x,y) = \frac{3}{14}$$

$$\mu_x = E(x) = \sum_x x.g(x) = (0).\frac{5}{14} + (1).\frac{15}{28} + (2).\frac{3}{28} = \frac{21}{28} = \frac{3}{4}$$

$$\mu_y = E(y) = \sum_y y.h(y) = (0).\frac{15}{28} + (1).\frac{3}{7} + (2).\frac{1}{28} = \frac{14}{28} = \frac{1}{2}$$

$$\sigma_{xy} = \frac{3}{14} - \frac{3}{4}.\frac{1}{2} = \frac{3}{14} - \frac{3}{8} = \frac{12-21}{56} = -\frac{9}{56}$$

De manera tal que si en la extracción se observan más biromes azules se reducirá el número de biromes rojas.

$f(x,y)$		X			$h(y)$
		0	1	2	
Y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
$g(x)$		5/14	15/28	3/28	1

Coeficiente de Correlación

Aunque la covarianza indica el sentido de la dependencia entre *X* e *Y*, no indica nada al respecto de la fuerza de esta relación, ya que está ligado a las escalas de *X* e *Y*.

La versión de la covarianza libre de escala se llama "**Coeficiente de correlación**":

$$\boxed{\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_x \, . \, \sigma_y}}$$

Debería quedar claro para el lector que $\rho_{x,y}$ no tiene las unidades de *X* e *Y*. El coeficiente de correlación satisface la desigualdad $-1 \leq \rho_{x,y} \leq 1$. Toma un valor de cero cuando $\sigma_{xy} = 0$.

Ejemplo: Calcule $\rho_{x,y}$ del ejemplo de los repuestos de bolígrafos.

Solución: $$\rho_{x,y} = \frac{\sigma_{xy}}{\sigma_{x}\,.\,\sigma_{y}} \qquad \text{con} \qquad \sigma_{xy} = -\frac{9}{56}$$

$$V(x) = \sigma_x^2 = E(x^2) - E(x)^2 = \sum_x x^2\, g(x) - \left[\sum_x x\, g(x)\right]^2$$

$$\sigma_x^2 = (1)^2.\frac{15}{28} + (2)^2.\frac{3}{28} - \left(\frac{3}{4}\right)^2 = \frac{27}{28} - \frac{9}{16} = \frac{45}{112}$$

$$\sigma_y^2 = (1)^2.\frac{3}{7} + (2)^2.\frac{1}{28} - \left(\frac{1}{2}\right)^2 = \frac{16}{28} - \frac{1}{4} = \frac{16-7}{28} = \frac{9}{28}$$

$$\rho_{x,y} = -\frac{9}{56}.\frac{1}{\sqrt{\left(\frac{45}{112}\right)\cdot\left(\frac{9}{28}\right)}} = -\frac{1}{\sqrt{5}}$$

$$\rho_{x,y} = -0{,}4472$$

f(x,y)		*X*			*h(y)*
		0	1	2	
Y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
g(x)		5/14	15/28	3/28	1

Independencia estadística

Habíamos visto que si X e Y son variables aleatorias conjuntas con función de distribución de probabilidad $f_{x,y}(x,y)$ y si es posible que $f_{x,y}(x,y)$ puede factorizarse como un producto de las funciones de distribución de probabilidad marginales $g(x)$ y $h(y)$, entonces se dice que X e Y son variables aleatorias independientes $f_{x,y}(x,y) = g(x).h(y)$.

Podemos deducir entonces que en términos de esperanza X e Y podrían ser independientes cuando observamos que un valor aleatorio para X, podría no influir (o tiene la esperanza de no influir) en los valores aleatorios de Y si se verifica que:

$$\boxed{E(xy) = E(x) \, . \, E(y)}$$

Demostración de: $E(xy) = E(x).E(y)$ se cumple cuando $f(x,y) = g(x).h(y)$, es decir, cuando *X* e *Y* son independientes.

$$E(xy) = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=\infty} x \, . \, y \, . \, f_{x,y}(x,y) \, dy \, dx$$

Si *X* e *Y* son independientes, entonces sabemos que: $f_{x,y}(x,y) = g(x)\,.\,h(y)$

Por lo tanto reemplazando:

$$E(xy) = \int_{x=-\infty}^{x=\infty} \int_{y=-\infty}^{y=\infty} x \, y \, g(x) \, h(y) \, dy \, dx = \underbrace{\int_{x=-\infty}^{x=\infty} x \, g(x) \, dx}_{E(x)} \, . \underbrace{\int_{y=-\infty}^{y=\infty} y \, h(y) \, dy}_{E(y)}$$

Entonces: $E(xy) = E(x).E(y) = \mu_x.\mu_y$

Esta identidad, que es verificada cuando existe independencia, reafirma lo que se postuló respecto de la covarianza:

$$\boxed{\sigma_{xy} = E(xy) - \mu_x.\mu_y = 0}$$

ya que *X* e *Y* son independientes

Varianza de una combinación lineal

Habíamos visto que la $V(ax+b) = a^2V(x)$. En el caso de un par de variables aleatorias *(X, Y)* la varianza de una combinación lineal de estas variables, puede demostrarse que resulta:

$$\boxed{V(ax+by+c) = a^2V(x) + b^2V(y) + 2abCOV(x,y)}$$

Demostración para el caso de una VAC:

Por definición de Varianza: $V(x) = \sigma_x^2 = E(x^2) - E(x)^2$

$$V(ax+by+c) = \underbrace{E[(ax+by+c)^2]}_{(1)} - [\underbrace{E(ax+by+c)}_{(2)}]^2$$

Siendo (1): $E[(ax+by+c)^2] = E[a^2x^2 + b^2y^2 + c^2 + 2(axby + byc + axc)]$

$$= a^2E(x^2) + b^2E(y^2) + c^2 + 2abE(xy) + 2bcE(y) + 2acE(x)$$

y $(2)^2$: $[E(ax+by+c)]^2 = [aE(x) + bE(y) + c]^2$

$$= a^2\ [E(x)]^2 + b^2[E(y)]^2 + c^2 + 2abE(x)E(y) + 2bcE(y) + 2acE(x)$$

Por tanto: (1) - (2):

$$V(ax+by+c) = a^2E(x^2) + b^2E(y^2) + \cancel{c^2} + 2abE(xy) + 2\cancel{bcE(y)} + 2\cancel{acE(x)}$$
$$- a^2[E(x)]^2 - b^2[E(y)]^2 - \cancel{c^2} - 2abE(x)E(y) - \cancel{2bcE(y)} - \cancel{2acE(x)}$$

Agrupando términos

$$V(ax+by+c) = a^2\underbrace{[E(x^2) - E(x)^2]}_{V(x)} + b^2\underbrace{[E(y^2) - E(y)^2]}_{V(y)} + 2ab\underbrace{[E(xy) - E(x)E(y)]}_{COV(x,y)}$$

$$\boxed{V(ax+by+c) = a^2V(x) + b^2V(y) + 2abCOV(x,y)}$$