Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Covarianza y Correlación

1

Profesora: *Dra. Fabiola Ocampo Botello*

2

A menudo se desea analizar la relación existente entre variables, es decir, en qué medida la variabilidad de los valores de una variable x se relaciona con la variación de una variable *y*.

Es decir, cómo se mueven los valores de x y los valores de *y*.

¿Los valores más altos de x se relaciona con los valores más altos de *y*?, o
¿Los valores más altos de x se relaciona con los valores más bajos de *y*?, o
¿Los valores de x no se "mueven" con los valores de *y*?

Esto es, la relación entre ambas variables, también llamada correlación es positiva, negativa o nula. Esta relación se establece en términos de coeficientes de correlación.

3

Covarianza

Anderson, Sweeney & Williams (2008) establecen que la covariaza y la correlación son medidas descriptivas de la relación entre dos variables.

La fórmula de la covariaza es:

COVARIANZA MUESTRAL

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} \qquad (3.10)$$

Imagen tomada de Anderson, Sweeney & Williams (2008)

Para comprender la covarianza se considera el ejemplo presentado por Anderson, Sweeney & Williams (2008) de la página 110-112.
Cuyo enunciado es:
El administrador de la tienda desea determinar la relación entre el número de comerciales televisados en un fin de semana y las ventas de la tienda durante la semana siguiente.

4

Mason, Lind & Marshal (2000:432) establecen que el análisis de correlación es un conjunto de técnicas empleado para medir la intensidad de la asociación entre dos variables.
Es decir, qué tan intensa es la relación entre tales variables.

Los mismos autores presentan un ejemplo:

Suponga que el gerente de una tienda desea conocer si existe una relación entre el número de llamadas telefónicas de ventas realizadas en un mes y la cantidad de copiadoras vendidas en ese lapso de tiempo.

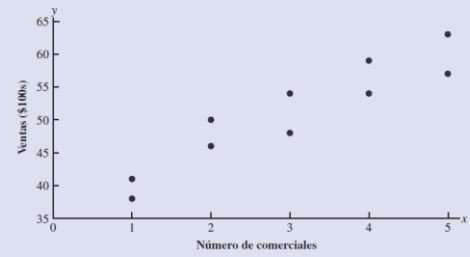
5

TABLA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO

Semana	Número de comerciales x	Volumen de ventas ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Imágenes tomadas de Anderson, Sweeney & Williams (2008)

FIGURA 3.7 DATOS MUESTRALES DE LA TIENDA DE EQUIPOS DE SONIDO



6

Al aplicar la fórmula 3.10.

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{9} = 11$$

TABLA 3.8 CÁLCULO DE LA COVARIANZA MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totales	30	510	0	0	99

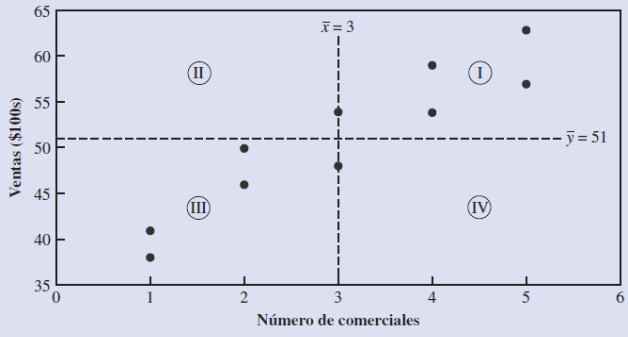
$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

Imagen tomada de Anderson, Sweeney & Williams (2008)

Interpretación de covarianza

7

FIGURA 3.8 DIAGRAMA DE DISPERSIÓN DIVIDIDO PARA LA TIENDA DE EQUIPOS DE SONIDO



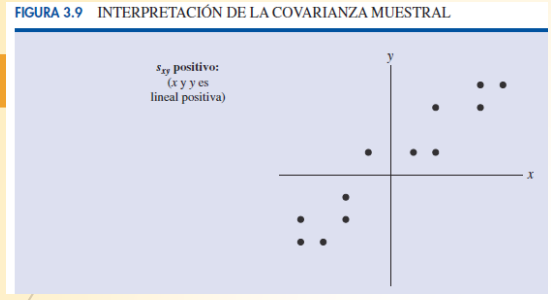
Una línea vertical punteada en $\bar{x} = 3$ y una línea horizontal punteada en $\bar{y} = 51$.

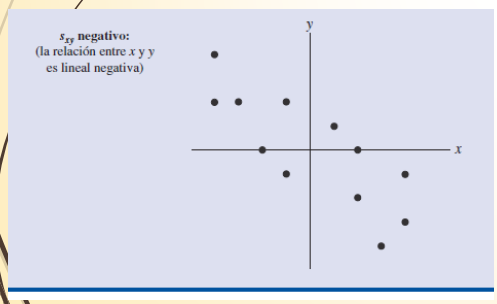
Estas líneas dividen a la gráfica en cuatro cuadrantes.
Si el valor de sxy es positivo, los puntos que más influyen sobre sxy deberán encontrarse en los cuadrantes I y III.

Imagen tomada de Anderson, Sweeney & Williams (2008)

8

FIGURA 3.9 INTERPRETACIÓN DE LA COVARIANZA MUESTRAL





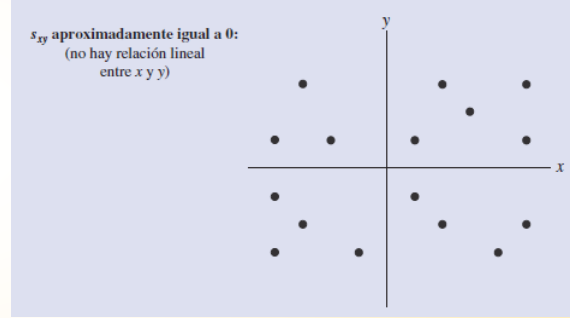


Imagen tomada de Anderson, Sweeney & Williams (2008)

9

Anderson, Sweeney & Williams (2008) establecen que un problema en el uso de la covarianza, como medida de la fuerza de la relación lineal, es que el valor de la covarianza depende de las unidades de medición empleadas para x y y. Una medida de la relación entre dos variables, a la cual no le afectan las unidades de medición empleadas para x y *y*, es el **coeficiente de correlación.**

COEFICIENTE DE CORRELACIÓN DEL PRODUCTO–MOMENTO DE PEARSON: DATOS MUESTRALES

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad \textbf{(3.12)}$$

donde

r_{xy} = coeficiente de correlación muestral
s_{xy} = covarianza muestral
s_x = desviación estándar muestral de x
s_y = desviación estándar muestral de y

Imagen tomada de Anderson, Sweeney & Williams (2008)

10

Considerando el ejemplo presentado anteriormente.

El coeficiente de correlación de los datos de la tienda de equipos para sonido. A partir de la tabla 3.8, se calcula la desviación estándar muestral de las dos variables.

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{9}} = 1.49$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{566}{9}} = 7.93$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{(1.49)(7.93)} = +0.93$$

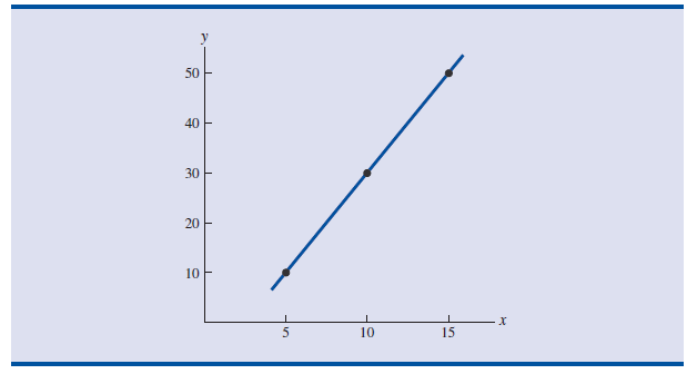
Imágenes tomadas de Anderson, Sweeney & Williams (2008)

11

Ejemplo del cálculo de la correlación presentada en las páginas 115-116 del libro de Anderson, Sweeney & Williams (2008).

x_i	y_i
5	10
10	30
15	50

FIGURA 3.10 DIAGRAMA DE DISPERSIÓN QUE REPRESENTA UNA RELACIÓN LINEAL POSITIVA PERFECTA



Imágenes tomadas de Anderson, Sweeney & Williams (2008)

12

TABLA 3.9 CÁLCULOS PARA OBTENER EL COEFICIENTE DE CORRELACIÓN MUESTRAL

	x_i	y_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	5	10	−5	25	−20	400	100
	10	30	0	0	0	0	0
	15	50	5	25	20	400	100
Totales	30	90	0	50	0	800	200

$\bar{x} = 10 \quad \bar{y} = 30$

Para emplear la ecuación (3.12) en el cálculo de la correlación muestral, es necesario calcular primero sxy, sx y sy. En la tabla 3.9 se muestran parte de los cálculos. Con los resultados de la tabla 3.9 se tiene

$$s_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{200}{2} = 100$$

$$s_x = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{50}{2}} = 5$$

$$s_y = \sqrt{\frac{\Sigma(y_i - \bar{y})^2}{n - 1}} = \sqrt{\frac{800}{2}} = 20$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{100}{5(20)} = 1$$

Imágenes tomadas de Anderson, Sweeney & Williams (2008)

Es importante establecer coeficientes de correlación, variaciones concomitantes que indiquen su covariación.

Kerliger y Lee (2002) establecen que los coeficientes de correlación están basado en la variación concomitante de los miembros de conjuntos de pares ordenados. Si covarían, varían altos valores con altos valores, valores medios con valores medios y valores bajos con valores bajos o valores bajos con valores altos, se dice que hay una relación positiva, negativa o cero.

13

TABLA 5.4 *Tres conjuntos de pares ordenados con diferentes direcciones y grados de correlación.*

(*I*) $r = 1.00$		(*II*) $r = -1.00$		(*III*) $r = 0$	
X	*Y*	*X*	*Y*	*X*	*Y*
1	1	1	5	1	2
2	2	2	4	2	5
3	3	3	3	3	3
4	4	4	2	4	1
5	5	5	1	5	4

Tabla tomada de Kerlinger y Lee (2002:85)

Ejercicios propuestos por Anderson, Sweeney & Williams (2008):

Ejercicios 48. En un estudio del departamento de transporte sobre la velocidad y el rendimiento de la gasolina en automóviles de tamaño mediano se obtuvieron los datos siguientes.

Velocidad	30 50 40 55 30 25 60 25 50 55
Rendimiento	28 25 25 23 30 32 21 35 26 25

Calcule e interprete el coeficiente de correlación muestral

Ejercicio 50: El Promedio Industrial Dow Jones (DJIA, por sus siglas en inglés) y el Standard & Poor's 500 Index (S&P 500) se usan para medir el mercado bursátil. El DJIA se basa en el precio de las acciones de 30 empresas grandes; el S&P 500 se basa en los precios de las acciones de 500 empresas. Si ambas miden el mercado bursátil, ¿cuál es la relación entre ellas? En los datos siguientes se muestra el aumento porcentual diario o la disminución porcentual diaria del DJIA y del S&P 500 en una muestra de nueve días durante tres meses (The Wall Street Journal, 15 de enero a 10 de marzo de 2006).

14

DJIA	0.20	0.82	−0.99	0.04	−0.24	1.01	0.30	0.55	−0.25
S&P 500	0.24	0.19	−0.91	0.08	−0.33	0.87	0.36	0.83	−0.16

Calcule e interprete el coeficiente de correlación muestral

15

Referencias bibliográficas

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.

Carollo Limeres, M. Carmen. (2012). Regresión lineal simple. Apuntes del departamento de estadística e investigación operativa . Disponible en: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf

Kerlinger, F. N. & Lee, H. B. (2002). Investigación del comportamiento. Métodos de investigación en ciencias sociales. 4ª ed. México: Mc. Graw Hill.

Mason, Lind & Marshal. (2000). Estadística para administración y economía. Alfaomega. 10ª edición.