



Ciencia de Datos y BigData

Análisis y Visualización de Datos - Varias Variables

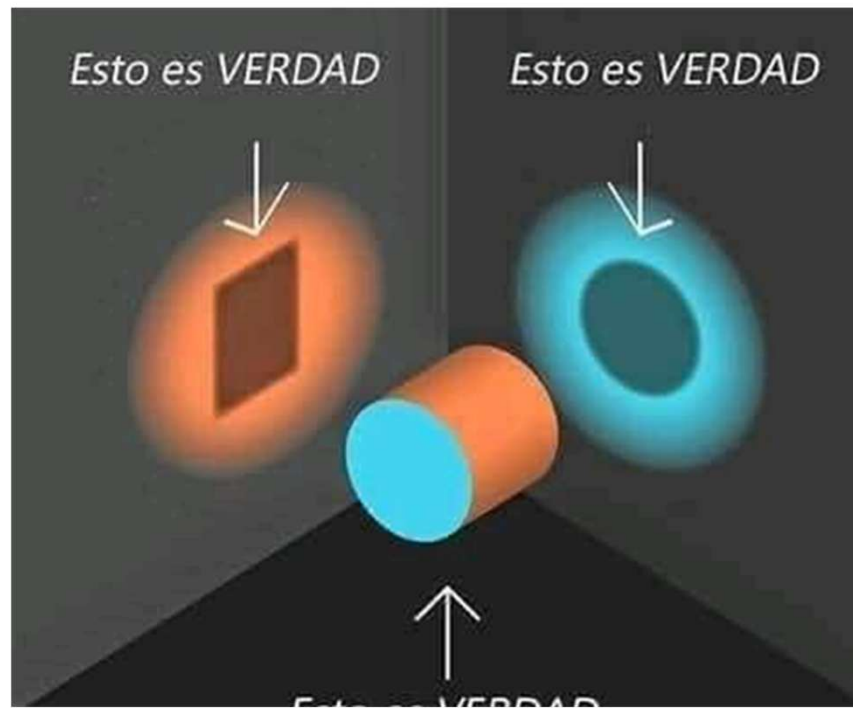
Dr. José Ramón Iglesias

DSP-ASIC BUILDER GROUP
Director Semillero TRIAC
Ingeniería Electronica
Universidad Popular del Cesar

Demo con Notebook

[EL460_Varias Variables.ipynb](#)

Para pensar...



Varias Variables

En un mismo experimento o análisis podemos tomar en cuenta varios aspectos o medidas relevantes a la vez, así como combinación de situaciones, etc.

Sean $\mathbf{X}: \Omega \rightarrow \mathbb{R}$ e $\mathbf{Y}: \Omega \rightarrow \mathbb{R}$ **variables aleatorias**.

tendrán asociadas una función de **densidad/probabilidad conjunta**:

- $f(x, y)$ p/ X e Y v.a. continuas, densidad, y
- $f(x, y) = P(X = x, Y = y)$ p/discretas, prob. o densidad puntual

Notación: $P(X=x, Y=y) = P((X=x) \cap (Y=y))$. la coma significa intersección

Varias Variables

X: $\Omega \rightarrow \mathbb{R}$ **e** **Y**: $\Omega \rightarrow \mathbb{R}$ **variables aleatorias** con función de densidad o probabilidad conjunta **$f(x,y)$** .

Se cumple que las **densidades marginales** son:

$$f_Y(y) = \sum_x f(x, y)$$

o categórica)

p/ X discreta (numérica

$$f_X(x) = \sum_y f(x, y)$$

o categórica)

p/ Y discreta (numérica

Varias Variables

X: $\Omega \rightarrow \mathbb{R}$ **e** **Y**: $\Omega \rightarrow \mathbb{R}$ **variables aleatorias** con función de densidad o probabilidad conjunta **$f(x,y)$** .

Se definen las **densidades condicionales como**:

$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}, \text{ si } f_Y(y) > 0$$

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)}, \text{ si } f_X(x) > 0$$

Percentiles

Los percentiles de crecimiento son percentiles de distribuciones condicionadas por edad. Ver: el siguiente [link curvas de crecimiento](#)

Varias Variables: propiedades

$X: \Omega \rightarrow \mathbb{R}$ e $Y: \Omega \rightarrow \mathbb{R}$ variables aleatorias

Se pueden combinar v.a. numéricas.

Por ejemplo, se puede definir la v.a. suma:

$X+Y: \Omega \rightarrow \mathbb{R}$ es una nueva v.a. $(X+Y)(\omega) = X(\omega) + Y(\omega)$ p/c/ $\omega \in \Omega$.

Y así con cualquier combinación de dos o más variables.

$X-Y$, X/Y^2 , $\log(X.Y)$, etc.

se cumple que:

- $E(X+Y) = E(X) + E(Y)$, $E(X-Y) = E(X) - E(Y)$, (media de la suma..., media de la resta..)
- $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) - \text{Cov}(X, Y)$.

Varias Variables, numéricas

X: $\Omega \rightarrow \mathbb{R}$ e **Y**: $\Omega \rightarrow \mathbb{R}$ **variables aleatorias**

Se define la **Covarianza** y el **Coeficiente de Correlación** entre ellas como:

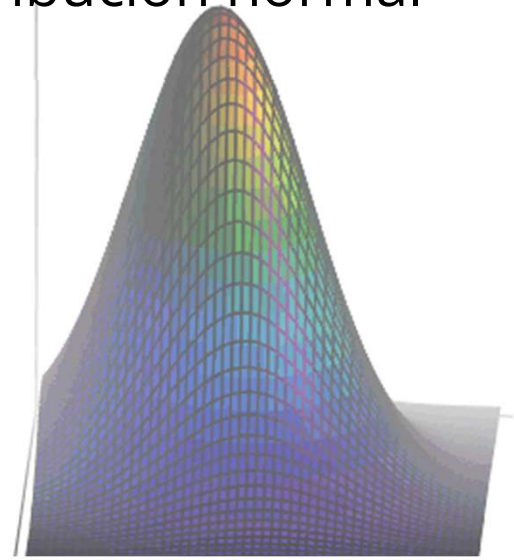
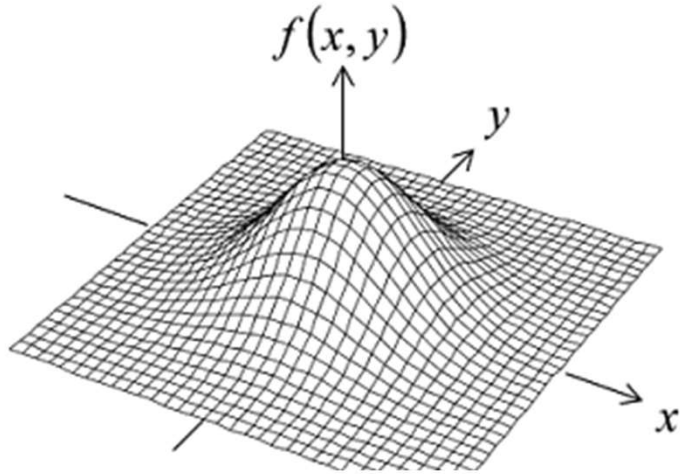
$\text{Cov}(X, Y) = E\{(X - \mu_X)(Y - \mu_Y)\}$, para $\mu_X = E(X)$ y $\mu_Y = E(Y)$.

, p/ $\sigma_1^2 = \text{Var}(X)$ y $\sigma_2^2 = \text{Var}(Y)$

$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1 \sigma_2}$$

Ejemplo: distribución Normal bivariada

Diremos que el par (X, Y) de v.a. tiene distribución normal bivariada si función de densidad conjunta es:



$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x-\mu_1)^2 + \sigma_1^2(y-\mu_2)^2 - 2\sigma_1\sigma_2\rho(x-\mu_1)(y-\mu_2))\right\}$$

$$\mu_1 = E(X), \quad \mu_2 = E(Y), \quad \sigma_1^2 = \text{Var}(X), \quad \sigma_2^2 = \text{Var}(Y), \quad \rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1\sigma_2}$$

Varias Variables, categóricas

```
In [9]: # Tabla de contingencia class / survived  
pd.crosstab(index=titanic['survived'],  
             columns=titanic['class'], margins=True)
```

Out[9]:

class	1st class	2nd class	3rd class	All
survived				
no	122	167	528	817
yes	203	118	178	499
All	325	285	706	1316

Notebook (ejercicio)

[CIED2 Varias Variables.ipynb](#)

Independencia entre Variables

X e Y v.a. se dicen **independientes** si $f(x, y) = f_X(x)f_Y(y)$

Para más variables:

Sean **X_1, X_2, \dots, X_n** , variables aleatorias se dicen mutuamente **independientes** si

$$f(x_1, x_2, \dots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\dots f_{X_n}(x_n) = \prod_{i=1}^n f_{X_i}(x_i)$$

Varias Variables: independencia

X: $\Omega \rightarrow \mathbb{R}$ e **Y**: $\Omega \rightarrow \mathbb{R}$ **variables aleatorias**

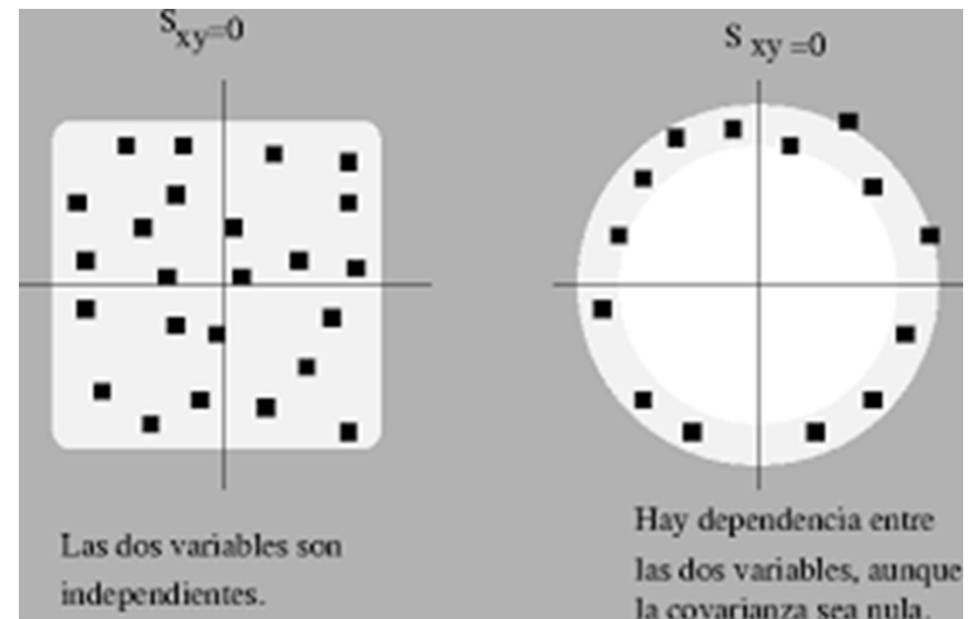
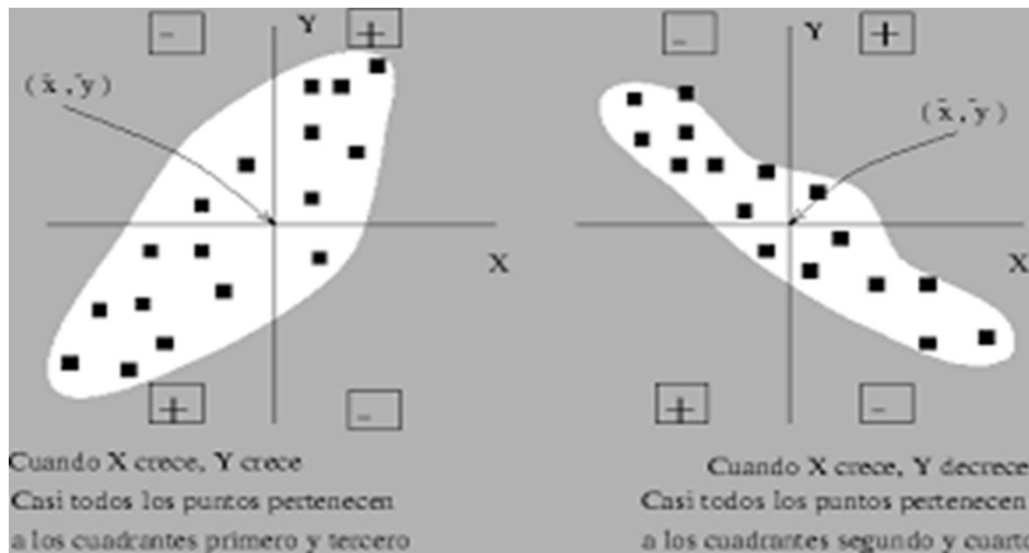
Si X e Y son independientes se cumple:

- $E(X.Y)=E(X).E(Y)$,
- $\text{Cov}(X,Y)=0$
- $\rho=\text{Corr}(X,Y)=0$
- $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$.
- $P(X=x, Y=y)=P(X=x). P(Y=y)$,
- $P(X=x | Y=y)=P(X=x)$,

$$f(x, y) = f_X(x) f_Y(y)$$

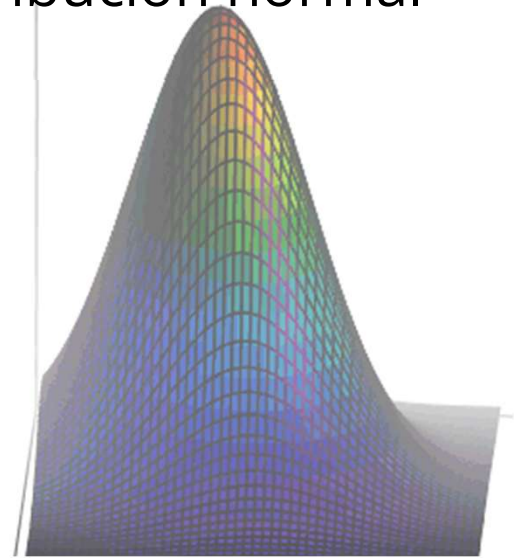
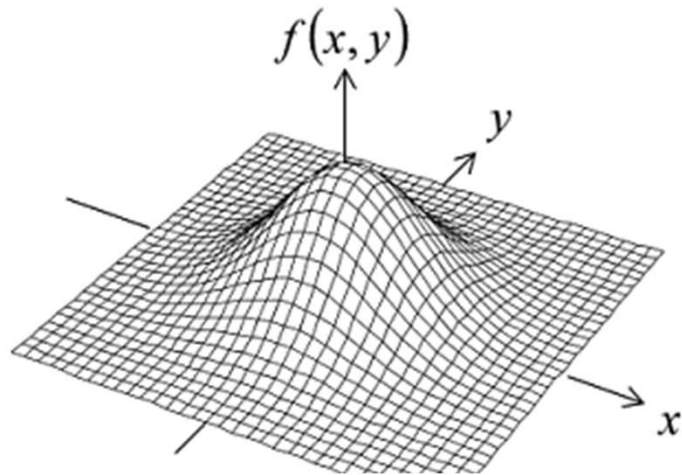
$$f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)} = \frac{f_X(x) \cancel{f_Y(y)}}{\cancel{f_Y(y)}} = f_X(x)$$

Covarianza en gráficos



Ejemplo: distribución Normal bivariada

Diremos que el par (X, Y) de v.a. tiene distribución normal bivariada si función de densidad conjunta es:



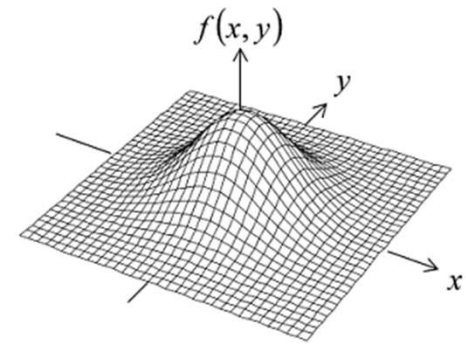
$$f(x,y)= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x-\mu_1)^2 + \sigma_1^2(y-\mu_2)^2 - 2\sigma_1\sigma_2\rho(x-\mu_1)(y-\mu_2))\right\}$$

$$\mu_1 = E(X), \quad \mu_2 = E(Y), \quad \sigma_1^2 = \text{Var}(X), \quad \sigma_2^2 = \text{Var}(Y), \quad \rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1\sigma_2}$$

Ejemplo: distribución Normal bivariada

X e Y v.a. con distribución normal bivariada son independientes si y sólo si $\rho=0$ si y sólo si $\text{Cov}(X,Y)=0$

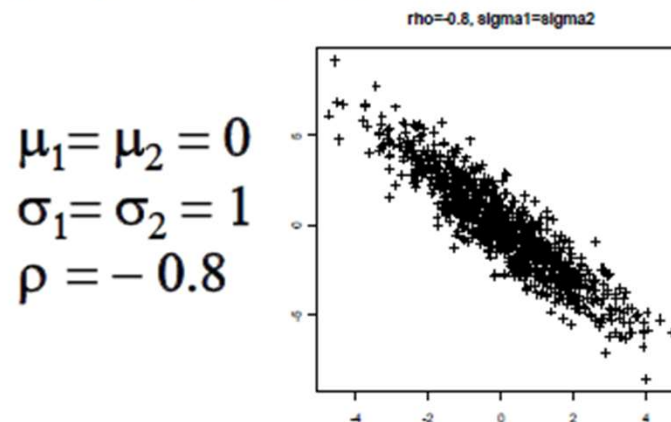
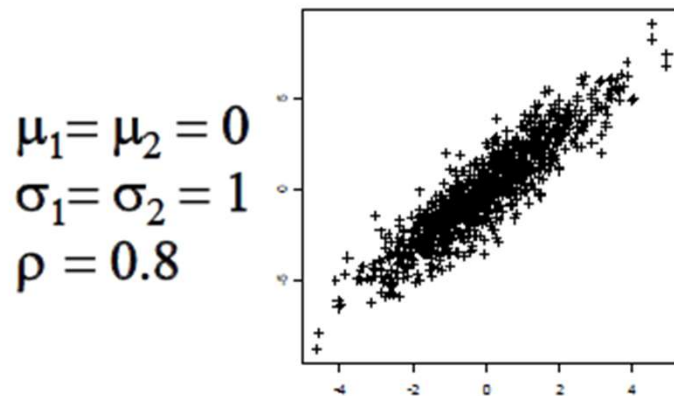
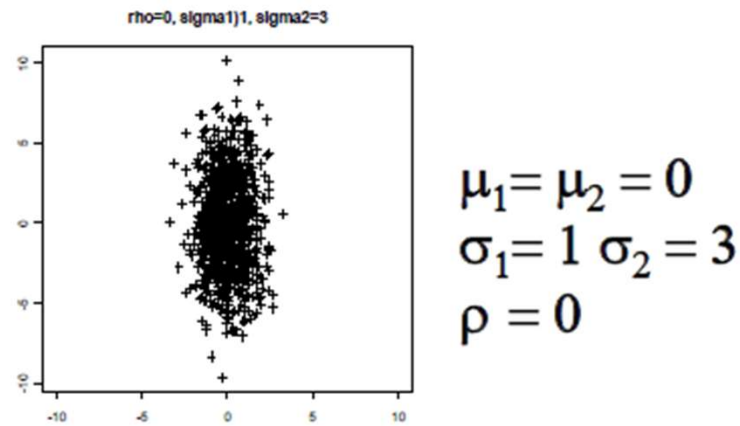
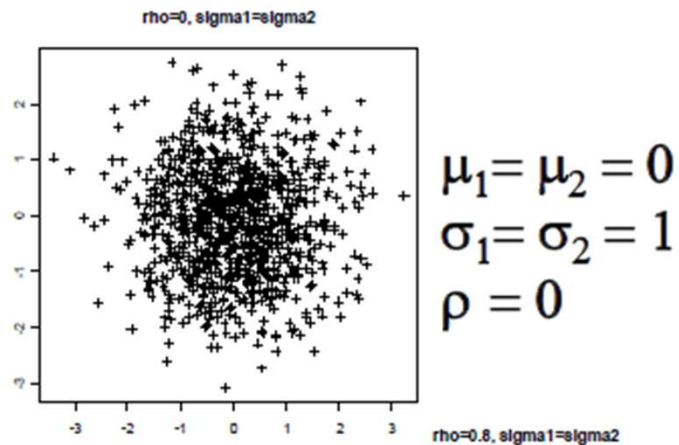
$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_1 \sigma_2} = 0$$



$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2\sigma_1^2\sigma_2^2(1-\rho^2)}(\sigma_2^2(x-\mu_1)^2 + \sigma_1^2(y-\mu_2)^2 - 2\sigma_1\sigma_2\rho(x-\mu_1)(y-\mu_2))\right\}$$

$$f(x,y) = \left[\frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right) \right] \left[\frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(y-\mu_2)^2}{2\sigma_2^2}\right) \right] = f_X(x)f_Y(y)$$

Datos con distribución Normal bivariada



Ejemplo: Distribución Normal multivariada

Llamamos **vector aleatorio** a $\underline{Y}=(Y_1, Y_2, \dots, Y_n)$ o a $\underline{Y}=[Y_1 \ Y_2 \ \dots \ Y_n]^t$ a un **arreglo de v.a. 's**

Diremos que el vector \underline{Y} tiene distribución normal multivariada si la función de densidad conjunta del arreglo es:

$$f(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) \right)$$

donde $\mathbf{x} = [x_1 \ \dots \ x_n]^t$ y $\mu = [\mu_1 \ \dots \ \mu_n]^t$
, vector de medias.

Σ se dice **Matriz de Var-Covarianza**, matriz nxn, simétrica no singular y $|\Sigma|$ denota el determinante de Σ .

Asociación de Variables

- Para variables Categóricas: **tablas de contingencia**

	Diestro	Zurdo	TOTAL
Hombre	43	9	52
Mujer	44	4	48
TOTAL	87	13	100

- Para variables Numéricas: Covarianza y Correlación
-

Covarianza y Correlación

$$Cov(X, Y) = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

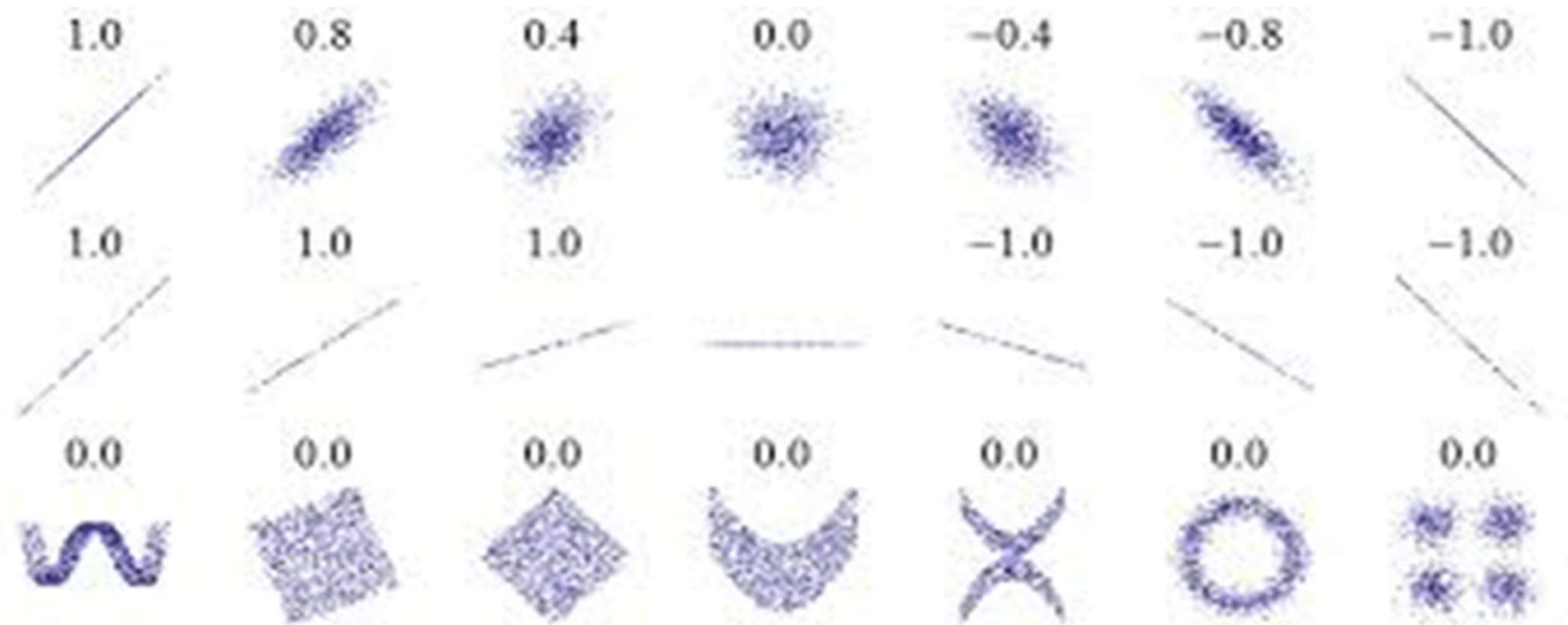
Si $Cov_{xy} > 0$, la correlación (“alineación”) es directa.

• Si $Cov_{xy} < 0$, la correlación (“alineación”) es inversa

$$\rho = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_1 \sigma_2}$$

El valor del índice de correlación $r = \rho$ varía en el intervalo $[-1, 1]$,

Correlación en gráficos



Interpretación

- Si $r = 1$, existe una correlación positiva perfecta. El índice indica una dependencia total entre las dos variables denominada *relación directa*: cuando una de ellas aumenta, la otra también lo hace en proporción constante.
 - Si $0 < r < 1$, existe una correlación positiva.
 - Si $r = 0$, no existe relación lineal. Pero esto no necesariamente implica que las variables son independientes: pueden existir todavía relaciones no lineales entre las dos variables.
 - Si $-1 < r < 0$, existe una correlación negativa.
 - Si $r = -1$, existe una correlación negativa perfecta. El índice indica una dependencia total entre las dos variables llamada *relación inversa*: cuando una de ellas aumenta, la otra disminuye en proporción constante.
-

Coeficiente de correlación lineal. Spearman

O Coeficiente de Spearman

Para (X, Y) par de v.a. Si no sabemos si su distribución conj. es Normal o tenemos poco datos. O si la/s variable/s son del tipo ordinal.

- analíticamente tiene un cálculo tedioso
-

Coeficiente de correlación lineal. Tau de Kendall

O Coeficiente de Correlación por Rangos de Kendall

Medida de asociación no paramétrica utilizada para variables cualitativas ordinales o de razón (numéricas). Estas variables son distribuidas en categorías con varios niveles que cumplen un orden, por ejemplo, muy bajo, bajo, medio, alto y muy alto.

- Sólo se puede aplicar a partir de tablas cuadradas.
 - Las variables utilizadas deben ser de nivel ordinal, intervalo o razón
 - Su resultado debe encontrarse en el rango de -1 a 1.
 - Tiene sentido su aplicación, si las variables objeto de estudio no poseen una distribución poblacional conjunta normal
-

Cuidado!

Recuerden que **la correlación no implica causalidad**. Por ejemplo, si las ventas de helados están correlacionadas positivamente con los ataques de los tiburones a los nadadores, eso no significa que el consumo de helados de alguna manera hace que los tiburones ataquen. Otra variable, como el clima cálido, puede provocar un aumento tanto en las ventas de helados como en las visitas a las playas.

