Ciencia de Datos
Ingeniería Industrial
UTN FRBA
curso I5521

clase_01

Análisis exploratorio de datos. Descripción estadística.

UTN.BA INGENIERÍA INDUSTRIAL

UTN.BA UNIVERSIDAD TECNOLÓGICA NACIONAL FACULTAD REGIONAL BUENOS AIRES

AI & Art

Borges and AI

Léon Bottou [†] and Bernhard Schölkopf [‡]

† FAIR, Meta, New York, NY, USA

‡ Max Planck Institute for Intelligent Systems, Tübingen, Germany

Abstract

Many believe that Large Language Models (LLMs) open the era of Artificial Intelligence (AI). Some see opportunities while others see dangers. Yet both proponents and opponents grasp AI through the imagery popularised by science fiction. Will the machine become sentient and rebel against its creators? Will we experience a paperclip apocalypse? Before answering such questions, we should first ask whether this mental imagery provides a good description of the phenomenon at hand. Understanding weather patterns through the moods of the gods only goes so far. The present paper instead advocates understanding LLMs and their connection to AI through the imagery of Jorge Luis Borges, a master of 20th century literature, forerunner of magical realism, and precursor to postmodern literature. This exercise leads to a new perspective that illuminates the relation between language modelling and artificial intelligence.

agenda_clase_01

EDA

- Boxplot
- Outliers utilizando quantiles
- Correlaciòn Lineal (Pearson)

Pandas

Lab

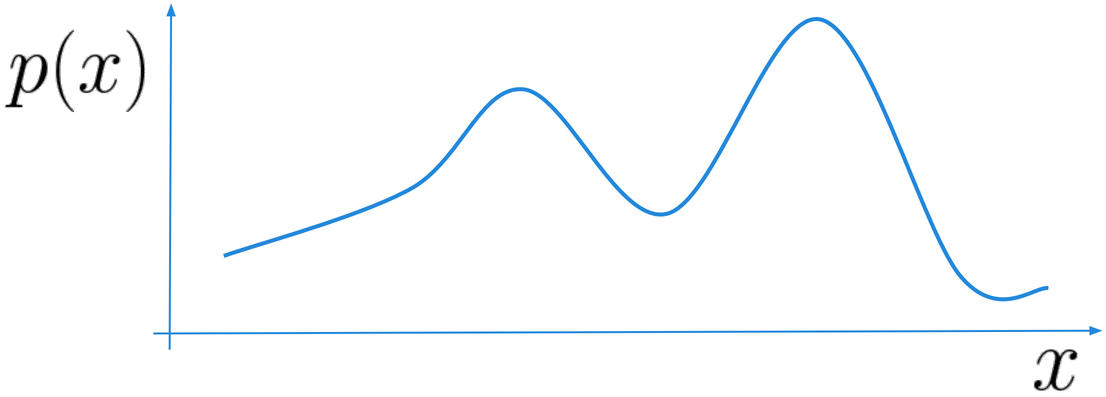
- EDA Subtes
- EDA GooglePlay

Histograma de frecuencias

Herramienta para estimar densidad empírica

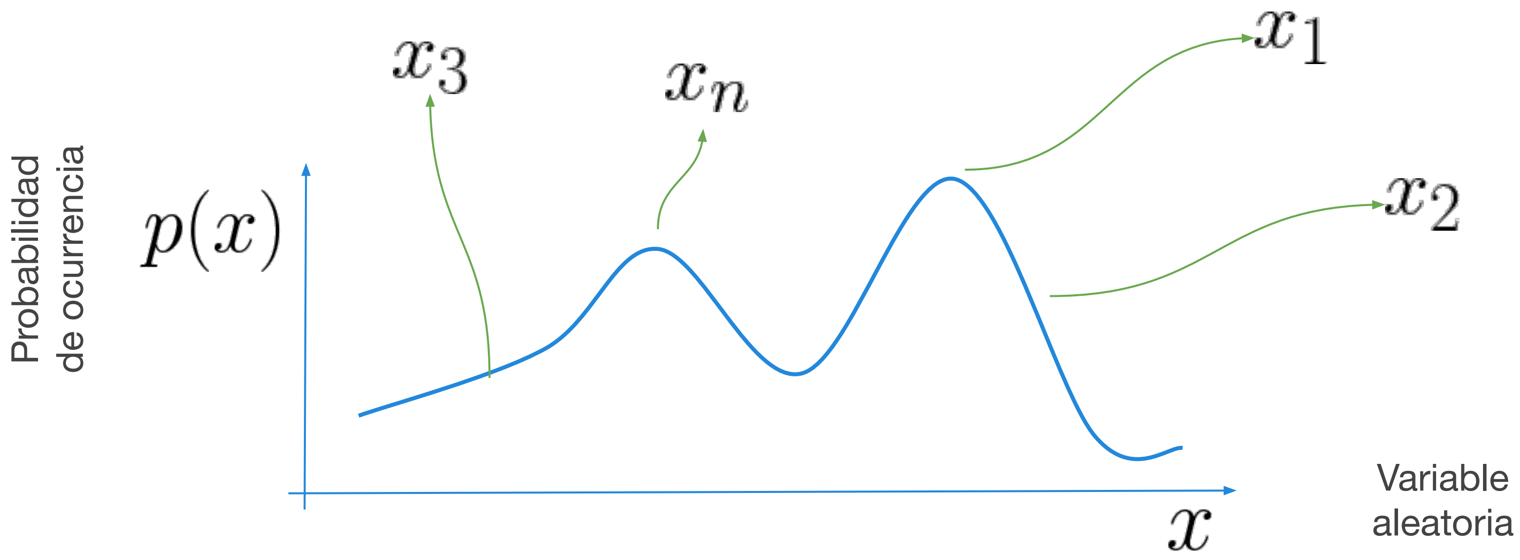
Distribución de probabilidad

La distribución de probabilidad es la **función** que asigna probabilidades de ocurrencia a distintos estados posibles de un experimento [1]. Es la **descripción** de un fenómeno **aleatorio** en términos de un espacio de muestreos y probabilidades de eventos.



[1] Ash, Robert B. (2008). *Basic probability theory* (Dover ed.). Mineola, N.Y.: Dover Publications

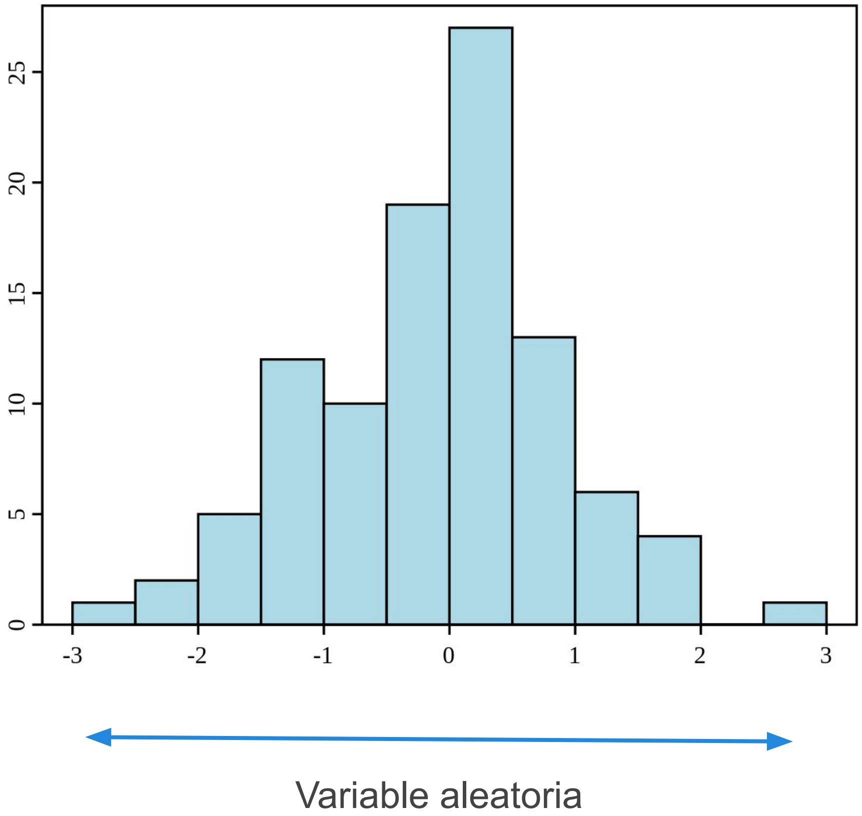
Muestreo desde una función de densidad de probabilidad



Suponiendo que **conocemos** la función de densidad de probabilidad de una variable aleatoria, vamos a **muestrear** multiples veces dicha funcion y obtener distintos valores de la variable aleatoria a simular. En el caso contrario si solo tenemos los datos y no conocemos la funcion de densidad que los genero se abordaran estrategias de maxima verosimilitud o metodos de estimacion no parametrica de la densidad [2]

[1] https://en.wikipedia.org/wiki/Maximum_likelihood_estimation
[2] https://en.wikipedia.org/wiki/Kernel_density_estimation .

Histograma de frecuencias



El histograma representa la frecuencia relativa de aparición de un valor de la variable aleatoria mediante la altura de las barras.

En el eje X tendremos los distintos valores que puede tomar una variable aleatoria a observar. En vez de contar valores únicos contamos todos los valores que caigan en un rango, es decir, la primer barra por ejemplo cuenta la cantidad de veces que la VA tomo los valores entre 100 y 125. La segunda barra cuenta la cantidad de veces que la VA tomo valores entre 125 y 150, etc.

Entonces al tomar muchas muestras (muestrear, samplear) una variable aleatoria podemos empíricamente entender cómo se distribuye los valores que la VA puede tomar. Entonces podemos decir que con un histograma podemos aproximar empíricamente la distribución de probabilidad.

Histograma de frecuencias

Cantidad de muestras por bin/caja

$$n_k = \sum \delta(x_{(kj)})$$

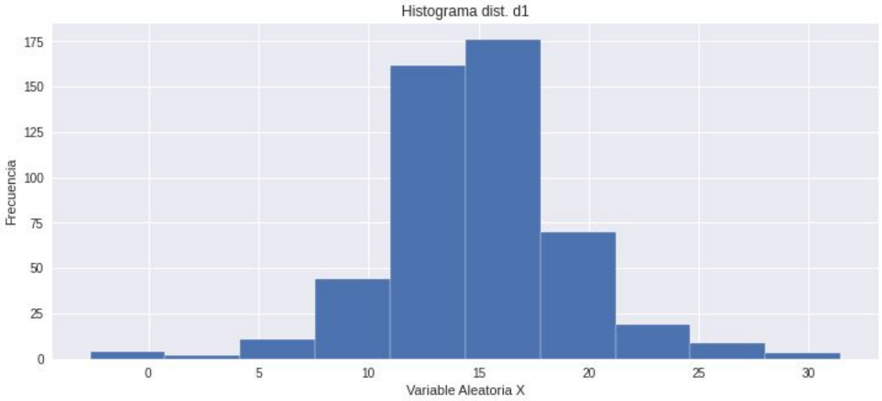
Funcion delta (contador)

$$\delta(x_{(ij)}) = 1$$

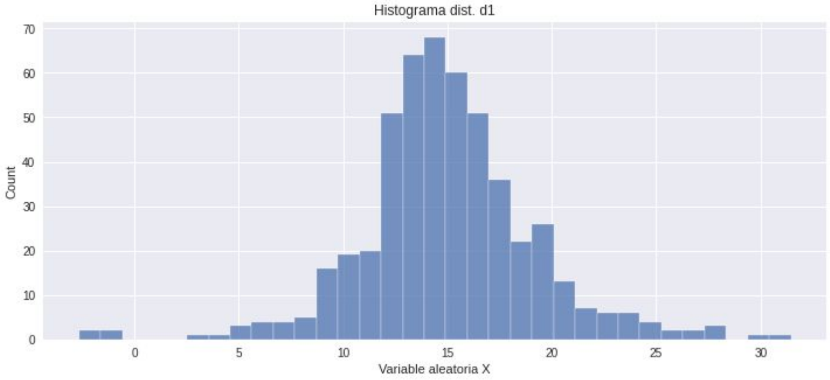
Muestras totales en los K bins

$$n = \sum_{i=1}^{k} \sum_{j=0}^{n_k} \delta(x_{(kj)})$$

Histograma de frecuencias

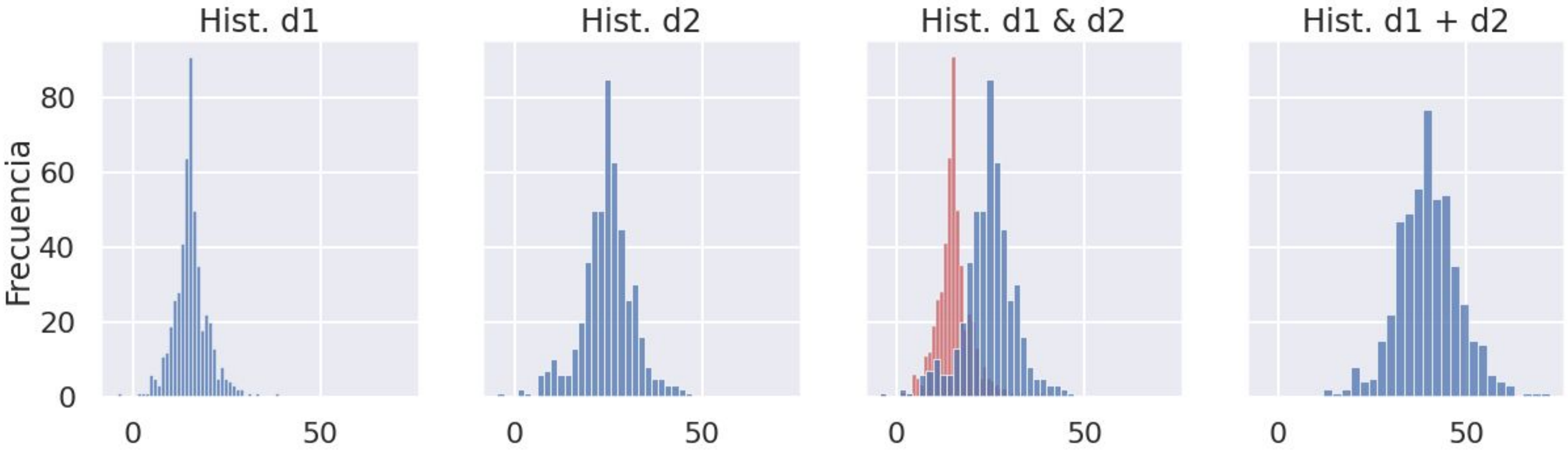


Histograma con bins = 10



Histograma con bins = 40

Histogramas



Histograma sobre 500 muestras de una distribución d1 normal mu = 15, sigma = 3

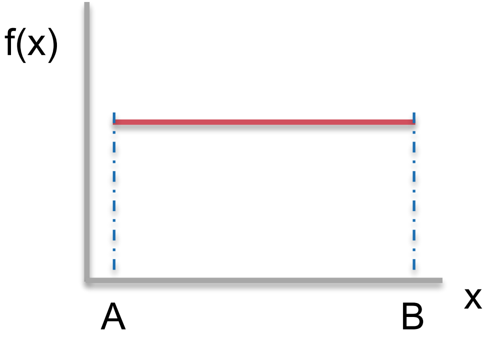
Histograma sobre 500 muestras de una distribución d2 normal mu = 25, sigma = 5

Los dos histogramas en simultáneo.

Histograma realizado sobre la suma de las 2 muestras obtenidas de las distribuciones d1 y d2.

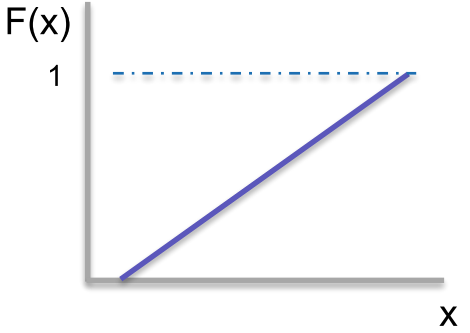
Distribución uniforme

Distribución de densidad de probabilidad.



$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if} \quad a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

Distribución de probabilidad acumulada.



Rango de valores posibles.

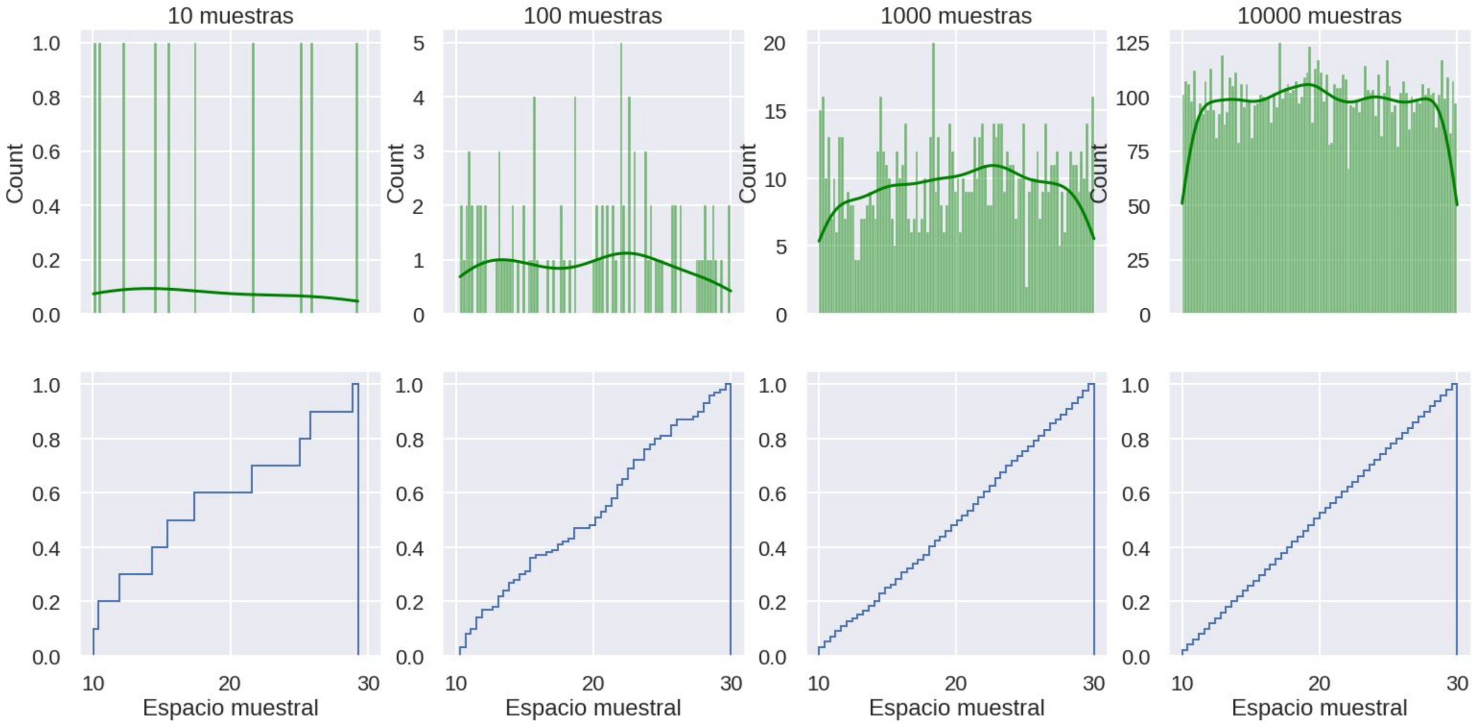
$$F(x) = \begin{cases} 0 & \text{if} \quad x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if} \quad x > b \end{cases}$$

La distribución de probabilidad uniforme asigna la misma probabilidad de ocurrencia a cada valor dentro del rango que puede generar una variable aleatoria.

Muestreo desde distribución uniforme

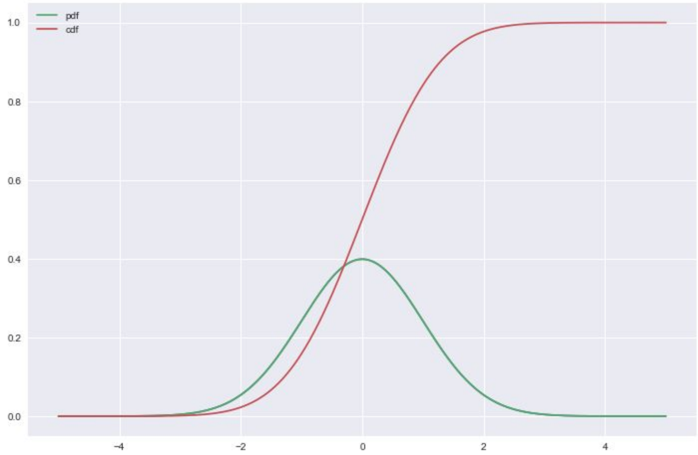
Probabilidad
Empirica
De ocurrencia

Probabilidad
Acumulada de
Ocurrencia



Distribución gaussiana - normal

Distribución de densidad (verde) y acumulada (roja) de probabilidad.



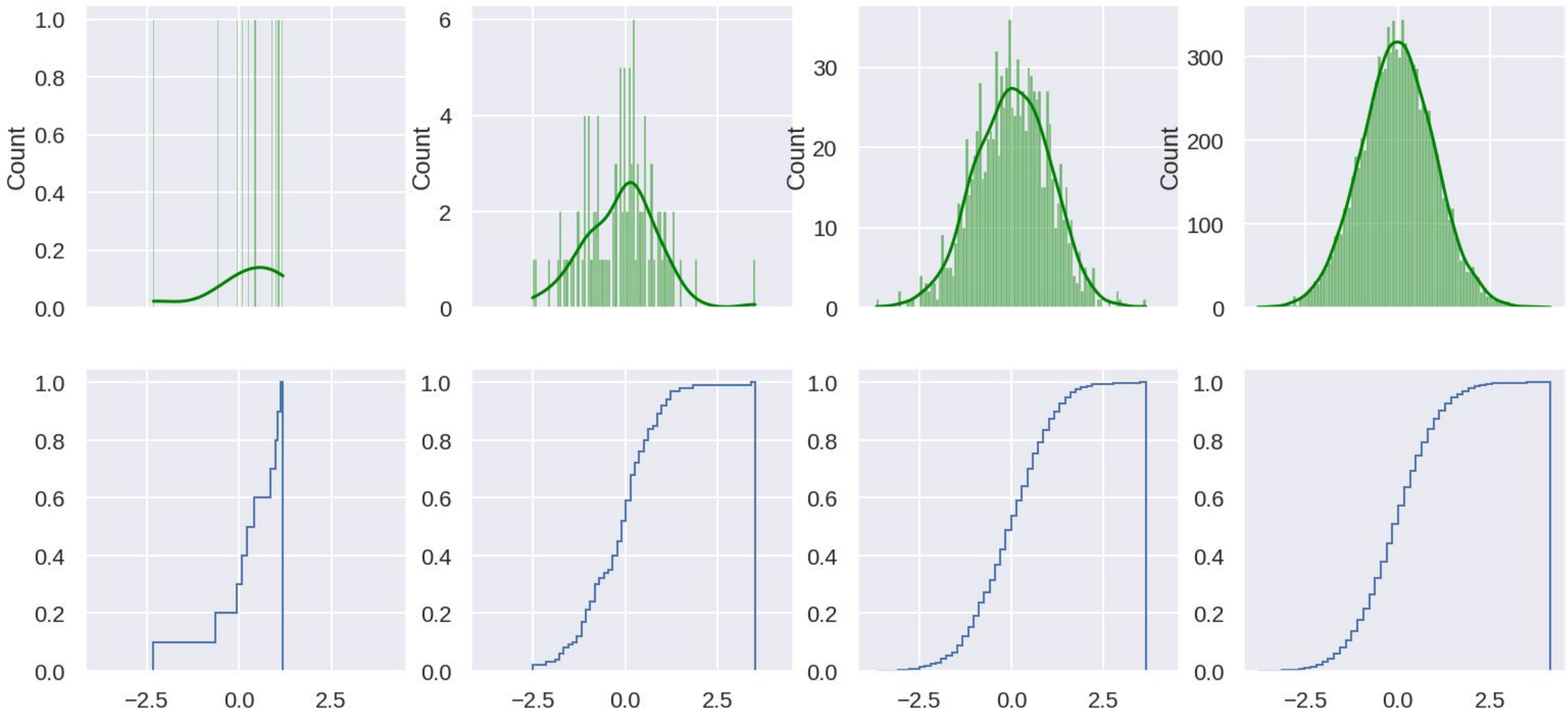
Rango de valores posibles de la VA.

Distribución simétrica de VA continua. El parámetro mu define la esperanza y el sigma el desvío standard.

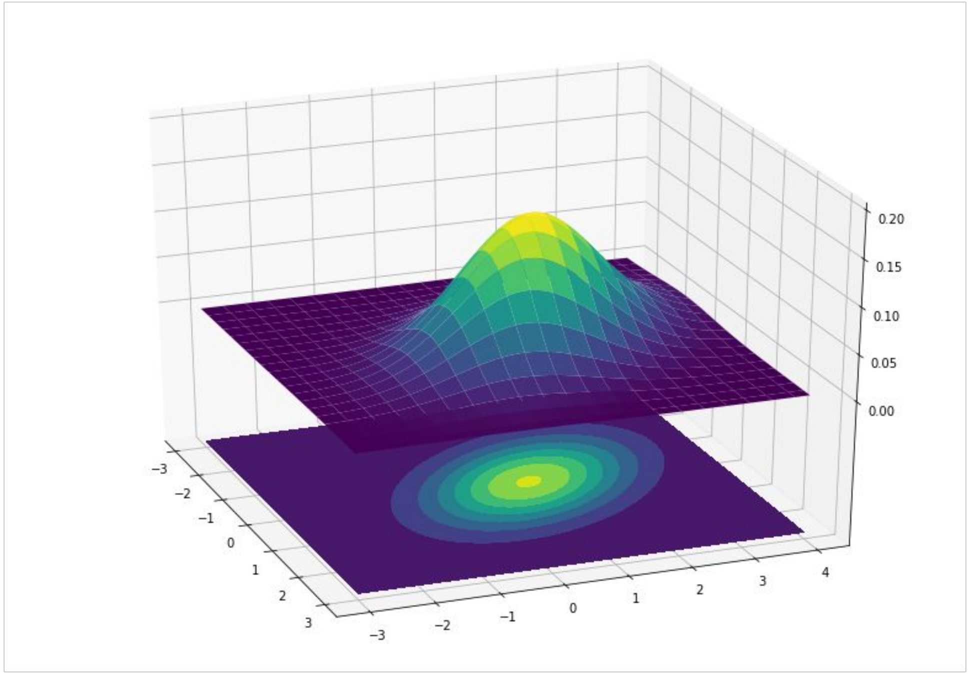
Suele utilizarse para modelar procesos reales en ciencias naturales, sociales, etc.

$$p(x) \sim \frac{1}{\sqrt{2\pi\sigma^2}}\mathbf{exp}\left[\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Muestreo de una distribución normal



Histograma 2D para gaussiana bivariada



$$p(X) \sim (2\pi)^{-d/2}|\Sigma^{-1/2}|\mathbf{exp}\left[-\frac{1}{2}(X-M)^t\,\Sigma^{-1}\,(X-M)\right]$$

Boxplot

Herramienta para estimar densidad empírica

Cuantiles

Los cuantiles suelen usarse como límites entre los grupos que dividen la distribución de **una** variable aleatoria en partes iguales; entendidas estas como intervalos que comprenden la misma proporción de valores.

Los mas populares son:

- Cuartiles, dividen la distribuciòn en 4 partes iguales (0.25, 0.5, 0.75)
- Quintiles, dividen la dist. en 5 partes iguales (0.2, 0.4, 0.6, 0.8)
- Deciles, dividen la dist. en 10 partes iguales (0.1,0.2…...0.9)
- Percentiles, dividen la dist. en 100 partes iguales (0.01…….0.99)

Cuantil & Cuartil

Datos Originales de una variable aleatoria

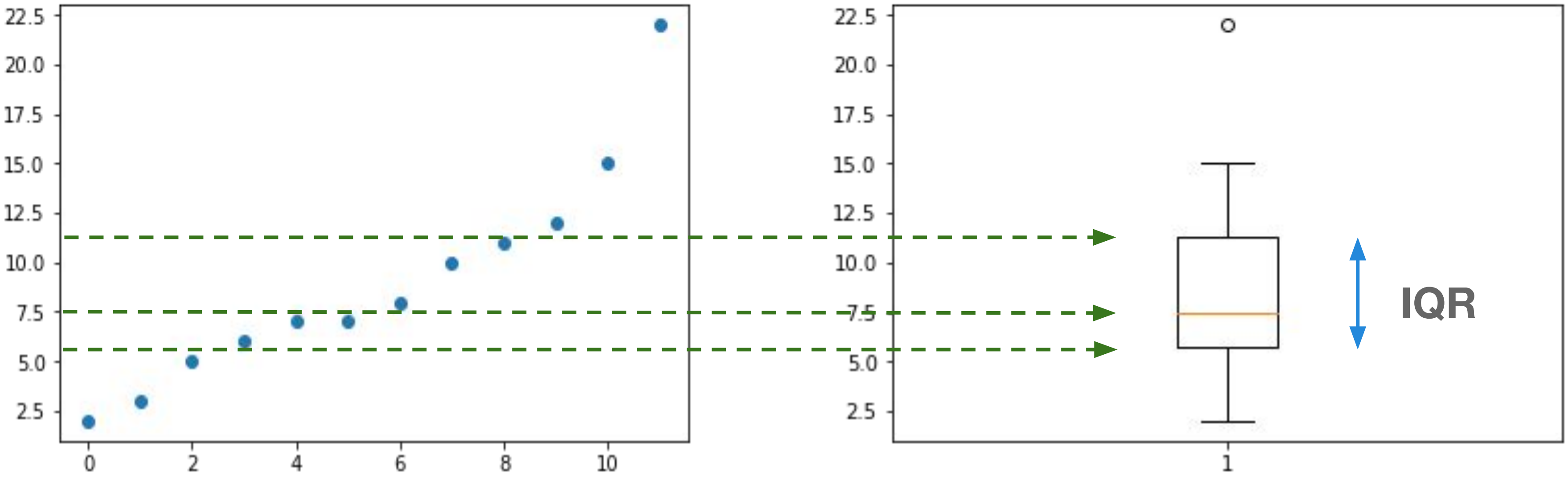
X = [15, 7, 3, 22, 10, 8, 6, 7, 2, 11, 5, 12]

Datos ordenados

[2, 3, 5, 6, 7, 7, 8, 10, 11, 12, 15, 22]

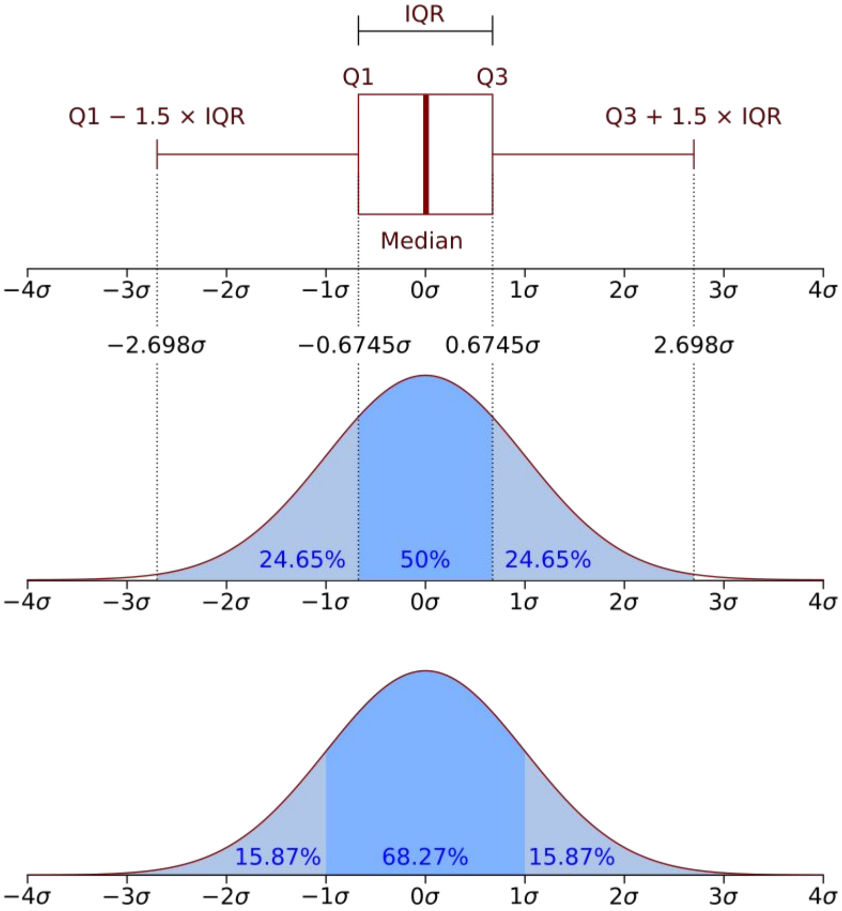
Min 1Q 2Q 3Q Max

Boxplot



En este caso por ejemplo tenemos una variable/feature que se mide en un lapso de 11 segundos. Queremos entender cómo se distribuyen los valores de la variable en cuestión.

Boxplot


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ
15.87%
68.27%
15.87%
−4σ −3σ −2σ −1σ 0σ 1σ 2σ 3σ 4σ

Cuartiles y Boxplots

Si ordenamos los datos de menor a mayor:

- El 25% de los datos será menor al 1er cuartil
- El 50% de los datos serà menor al 2do cuartil (mediana)
- El 75% de los datos serà menor al 3er cuartil
- Los valores que estén sobre el percentil 0.01 y 0.99 podrian considerarse outliers.

Mean, Median & Outliers

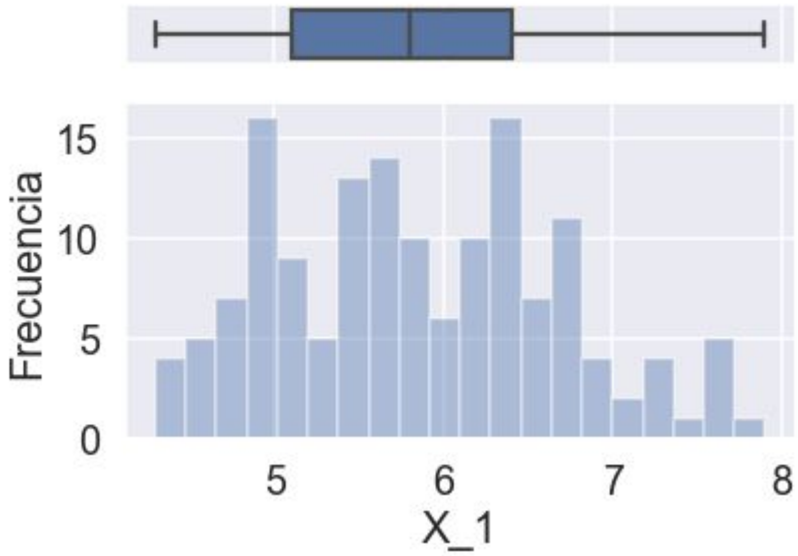
Filtrar por Cuantiles

Muchas veces, con el fin de quitar outliers de la distribuciòn de datos que deseamos analizar, lo que podemos realizar es:

- Quitar todos los datos que estén por encima del Percentil 99
- Quitar todos los datos que estén por debajo del Percentil 1
- Quitar todos los datos que estén por fuera del 1.5 * IQR (Inter Quartile Range).

Cuidado! Quitar datos del dataset dependerá de cada caso, es importante entender las consecuencias de quitar instancias consideradas anomalías.

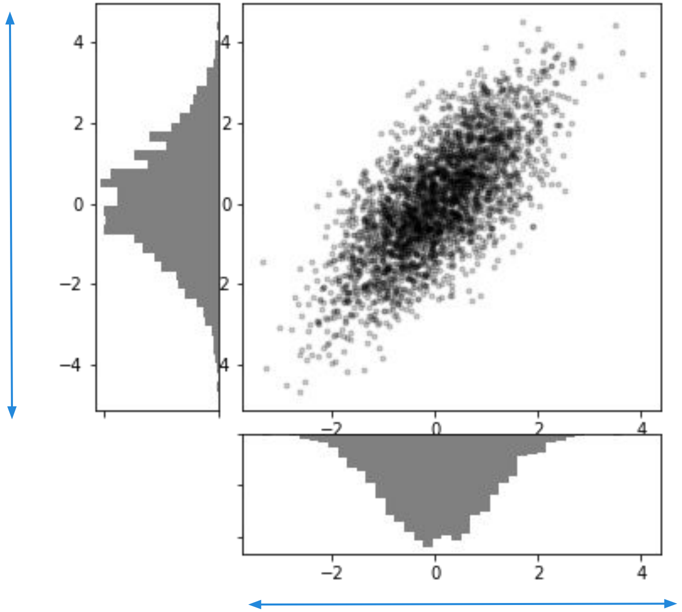
Boxplot & histograma



Un boxplot y un histograma en 1D son sinónimos y complementos para visualizar la densidad de probabilidad empírica de una variable.

Scatterplot + Histograma

x_1



$x \in \mathbb{R}^2$

x_2

Es es posible visualizar muestras en dos dimensiones con un scatterplot y simultáneamente histogramas en cada una de las variables que caracterizan a cada muestra. De igual manera en lugar de los histogramas podría haber un boxplot.

Correlación Lineal

Medida que expresa si dos variables aleatorias co-varían linealmente.

Correlación lineal (Pearson)

Es una forma de medir cuán cercanas están dos variables x e y (features) a tener una relación lineal entre ellas.

$$r = \frac{\sum_i^n (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{[\sum_i^n (x_{1i} - \bar{x_1})^2 (x_{2i} - \bar{x_2})^2]^{1/2}}$$

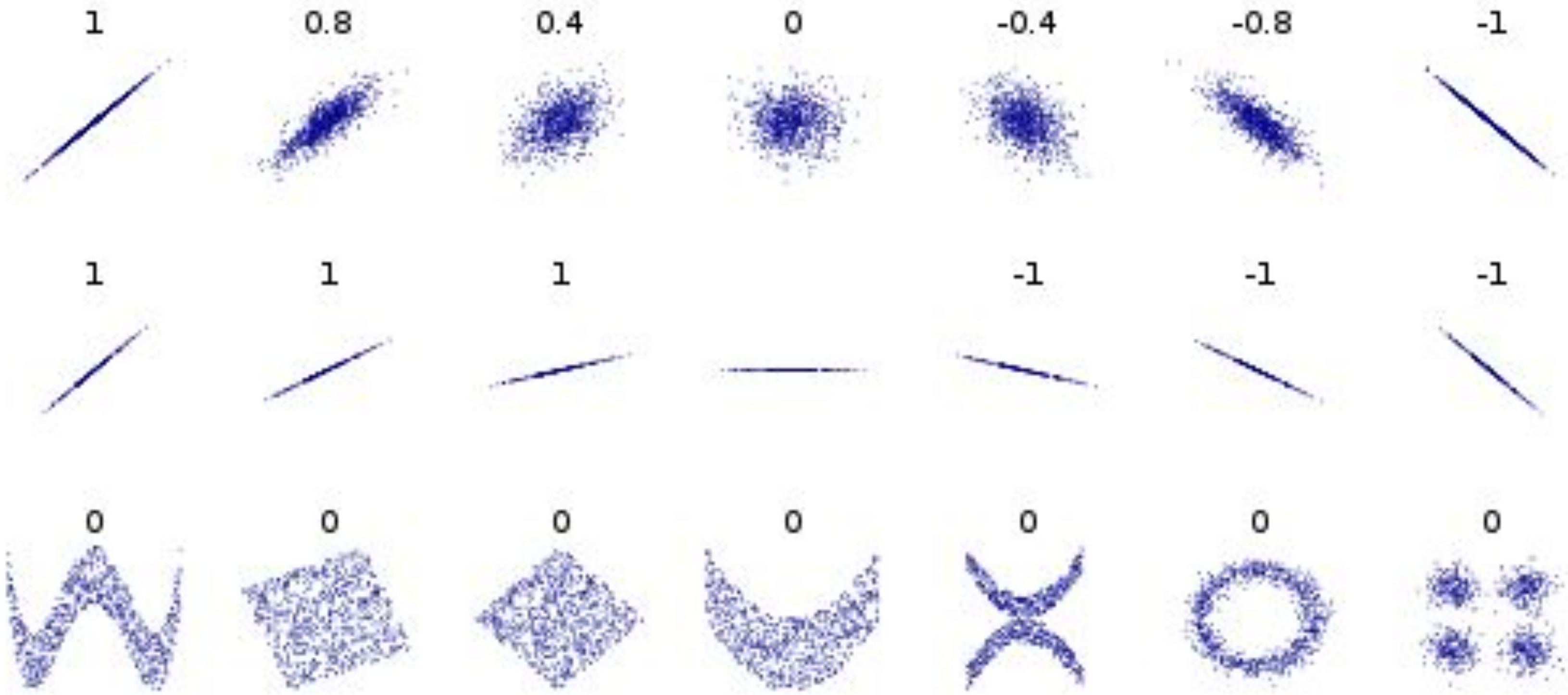
Matriz de correlación

$$R = \begin{array}{c} d \\ \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & r_{32} & r_{33} & r_{34} & r_{35} \\ r_{41} & r_{42} & r_{43} & r_{44} & r_{45} \\ r_{51} & r_{52} & r_{53} & r_{54} & r_{55} \end{bmatrix} \end{array} \; d$$

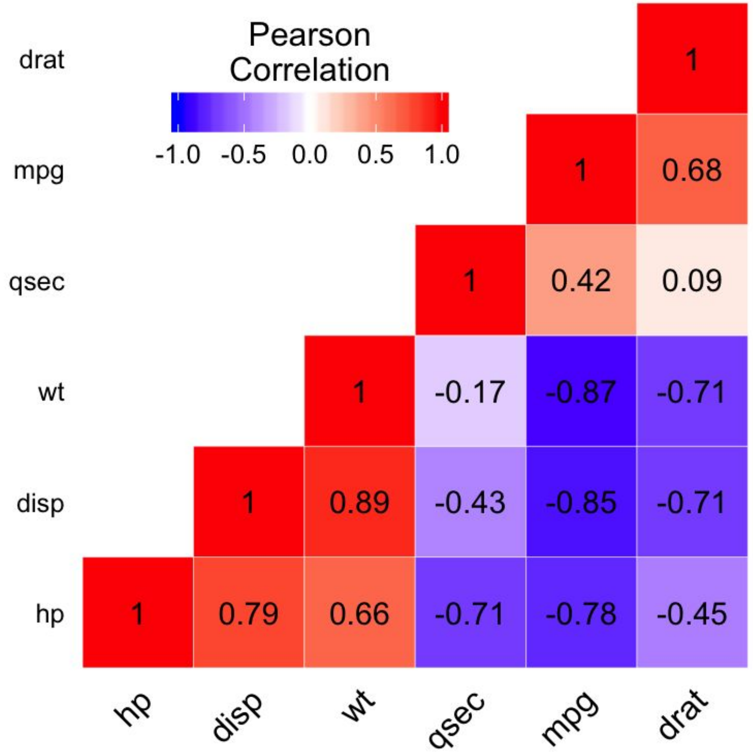
$$r = \frac{\sum_i^n (x_{1i} - \bar{x_1})(x_{2i} - \bar{x_2})}{[\sum_i^n (x_{1i} - \bar{x_1})^2 (x_{2i} - \bar{x_2})^2]^{1/2}}$$

$$\mathcal{X} \in \mathbb{R}^d$$

Correlaciòn lineal (Pearson)



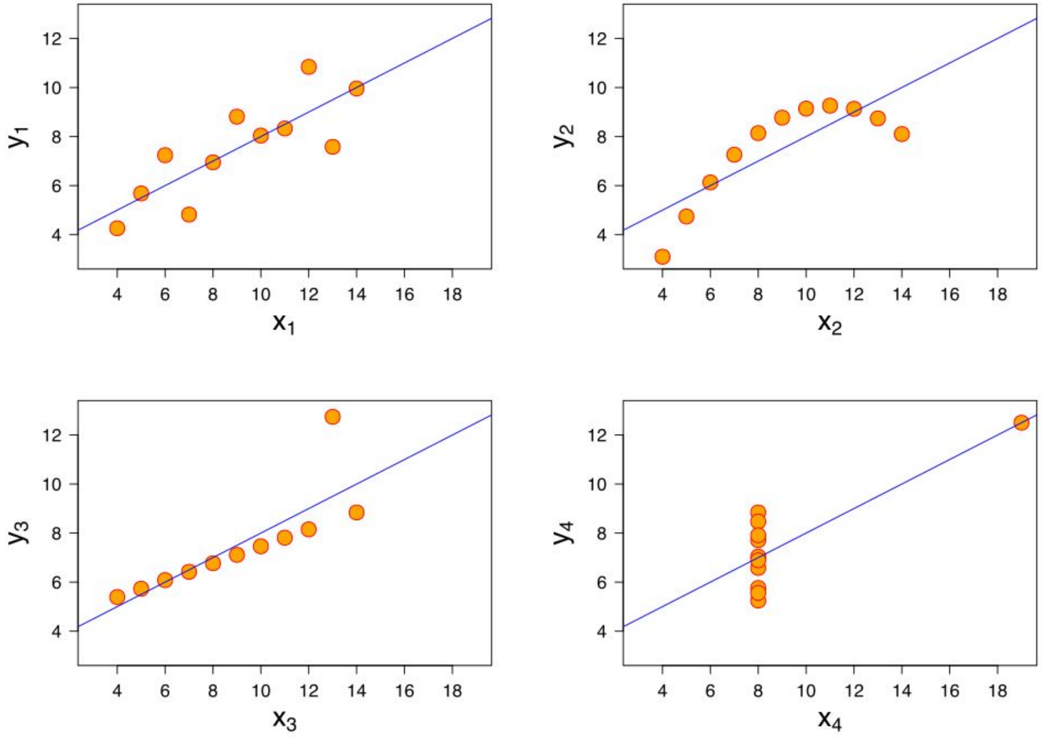
Correlación pairwise entre variables



En el ejemplo tenemos 6 variables/features. Podemos calcular la correlación lineal de Pearson par-a-par y visualizarla con un heatmap.

Atención: la correlación de Pearson **sólo** mide relación lineal entre variables. Que no exista correlación lineal no quiere decir que no exista relación alguna. Puede existir relación no lineal.

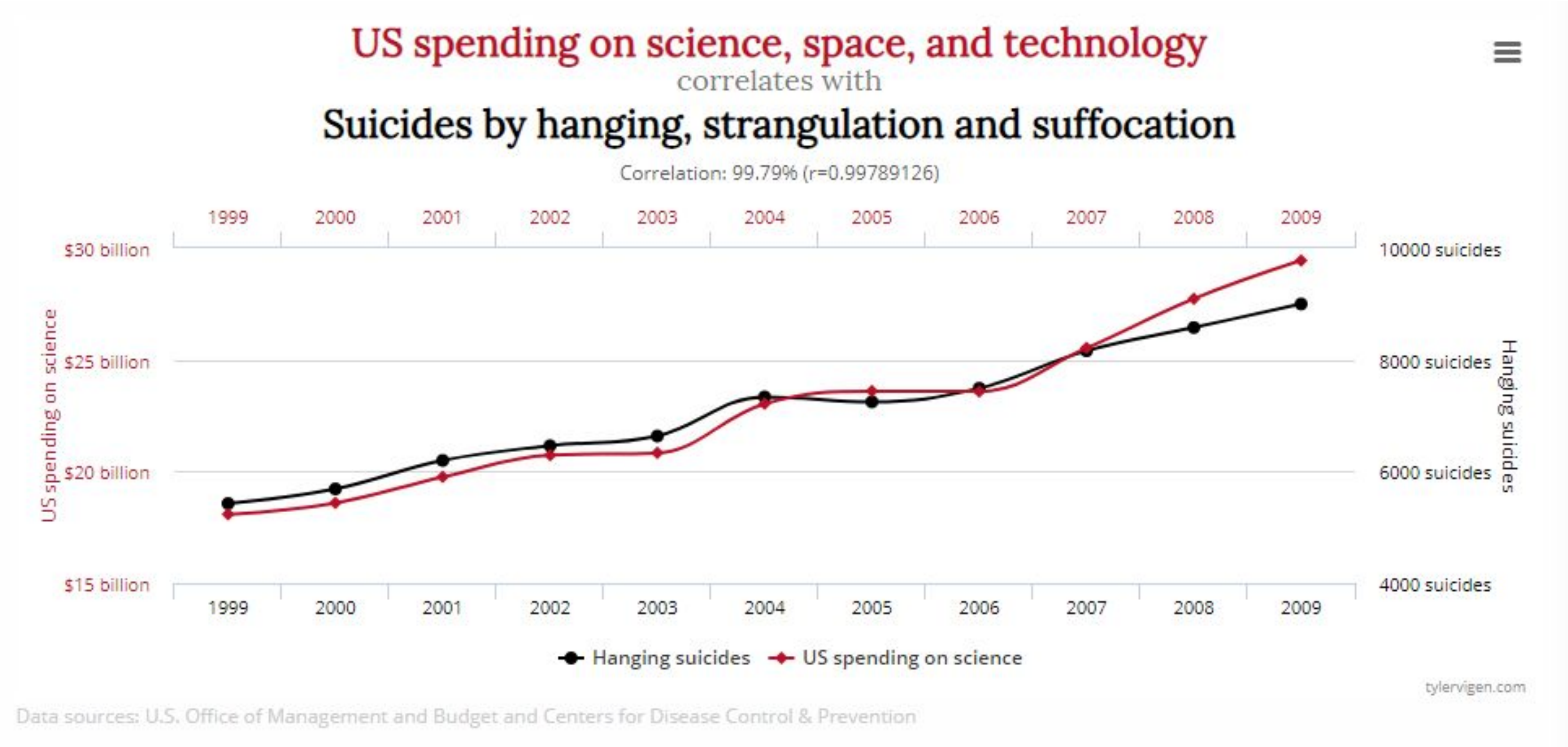
Correlaciòn lineal: trampas



Los 4 datasets tienen las mismas estadísticas descriptivas, sin embargo se ven muy distintos cuando se visualizan:

Media X = 9
Rxy = 0.81

Correlation is not causation



Pearson correlation in Python

```
x = np.array([[2,2],[3,3]])

R1 = np.corrcoef(x)
```

Pandas: Concat, Join, Merge

Pandas nos da la opción de poder combinar dataframes de distintas formas:

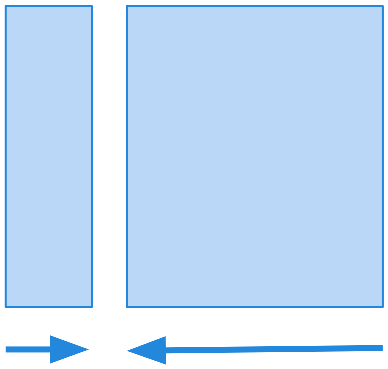
- **Concat**, unir dos dataframes por columnas o filas
- **Join & Merge** (vlookup)

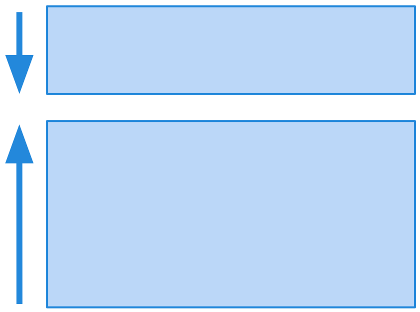
Pandas: Concat

Podremos concatenar dos dataframes por columnas o por filas. Esto quiere decir que si concatenamos por:

- Columnas: la cantidad de filas de ambos tiene que ser igual
- Filas: la cantidad de columnas de ambos tiene que ser igual

Concatenar por columnas



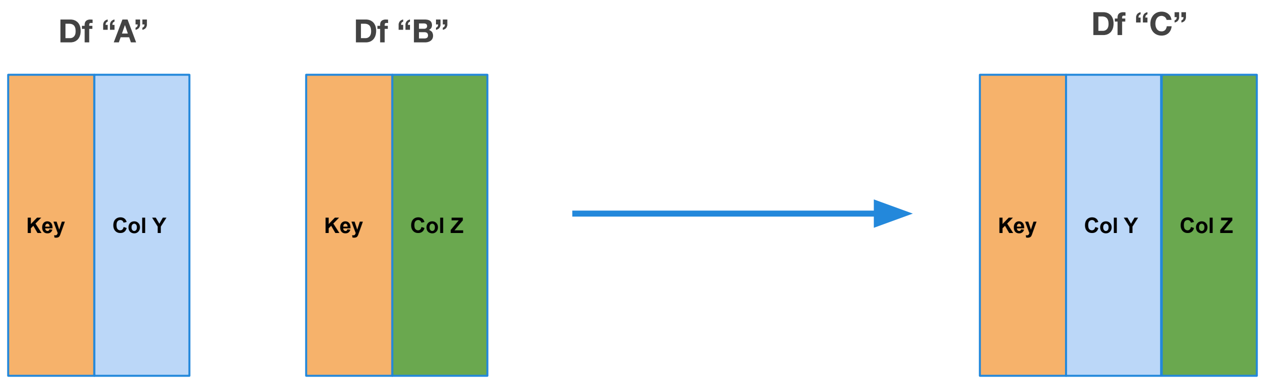


Concatenar por filas

Si la cantidad de filas o columnas no son iguales dependiendo el caso pandas generará nuevas columnas y filas para satisfacer la desigualdad y estas estarán llenas con NaNs.

Pandas: Join & Merge

Es lo más cercano al “vlookup” en excel. Esto permite poder tener una columna “key” de referencia en dos tablas (A y B). Permite llevar los datos de B asociados a “key” a la tabla A asociándolos a “key” también.



Pandas: Merge

```
In [39]: left = pd.DataFrame({'key': ['K0', 'K1', 'K2', 'K3'],
   ....:                      'A': ['A0', 'A1', 'A2', 'A3'],
   ....:                      'B': ['B0', 'B1', 'B2', 'B3']})
   ....:

In [40]: right = pd.DataFrame({'key': ['K0', 'K1', 'K2', 'K3'],
   ....:                       'C': ['C0', 'C1', 'C2', 'C3'],
   ....:                       'D': ['D0', 'D1', 'D2', 'D3']})
   ....:

In [41]: result = pd.merge(left, right, on='key')
```

left

	key	A	B
0	K0	A0	B0
1	K1	A1	B1
2	K2	A2	B2
3	K3	A3	B3

right

	key	C	D
0	K0	C0	D0
1	K1	C1	D1
2	K2	C2	D2
3	K3	C3	D3

Result

	key	A	B	C	D
0	K0	A0	B0	C0	D0
1	K1	A1	B1	C1	D1
2	K2	A2	B2	C2	D2
3	K3	A3	B3	C3	D3

https://pandas.pydata.org/docs/user_guide/merging.html

Pandas: pivot_table

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	A	2	y
2	one	B	3	z
3	two	A	4	q
4	two	B	5	w
5	two	B	6	t

```
df.pivot_table (index='foo',
                columns='bar',
                values='baz',
                aggfunc='sum')
```

bar	A	B
foo		
one	1 + 2	3
two	4	5 + 6

https://pandas.pydata.org/docs/user_guide/reshaping.html

A agarrar la PyLA