**Clase 4: Estadística Descriptiva**

La Estadística Descriptiva sirve para describir, resumir y entender los datos.

**Medidas de Tendencia Central**:

* **Media**. Promedio. ; np.mean(nombre\_array)
* **Mediana** Valor en el cual hay la misma cantidad de valores a izquierda y derecha en una distribución; np.median(nombre\_array)
* **Moda**: Valor que aparece con mayor frecuencia en una distribución.

**Asimetría a la izquierda:** Media < Mediana < Moda.

**Asimetría a la Derecha:** Moda < Mediana < Media.

**Distribución Simétrica:** Moda = Mediana = Media.

**Medidas de Variabilidad**:

* **Rango**: Valor máximo – valor mínimo en una distribución. Rango = np.max(array) – np.min(array)
* **Rango Intercuartil**: Diferencia entre el 3er cuartil y el 1er cuartil de una distribución. A diferencia del rango, no se ve afectado fuertemente por *outliers*. q3, q1 = np.percentile(array, [75,25]); rango intercuartil = q3 - q1
* **Varianza**: Ayuda a entender cuán lejos están los números de una distribución de su media. ; variance = np.var(array)
* **Desvío Estándar**. Raíz cuadrada de la varianza. Es otra medida de dispersión de los datos pero **no** es la desviación promedio con respecto de la media porque los desvíos fueron elevados al cuadrado antes de sacar la raíz cuadrada. Una ventaja de la desviación estándar es que está medida en las mismas unidades que la media.

S = np.std(array).

En una distribución normal:

µ +- abarca el 68% de los valores;

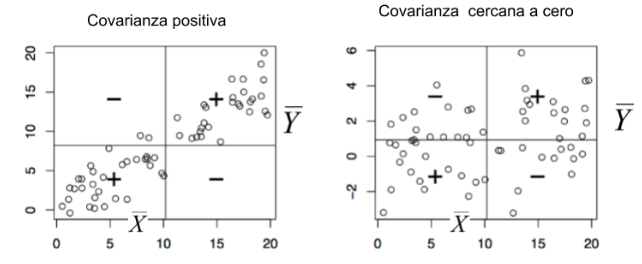
µ +- abarca el 95% de los valores;

µ +- abarca el 99,7% de los valores

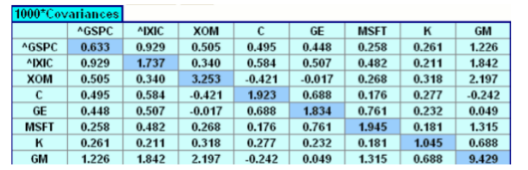
* **Coeficiente de Variación**: Desviación estándar dividido la media. Permite comparar la dispersión de diferentes variables. No tiene unidades.

**Medidas de Asociación Lineal entre Variables**:

**Covarianza:** 2 variables x e y tienen Covarianza positiva/negativa cuando tienden a encontrarse por encima de su media (cov positiva) o al estar una por encima de su media, la otra tender a estar por debajo de su respectiva media (negativa). Covarianza cercana a cero: Una variable puede encontrarse por encima o debajo de su media independientemente de lo que haga la otra:



Si se tiene un conjunto de datos con p variables, se puede representar una matriz de p x p llamada **matriz de varianzas y covarianzas**:



Dado un array con 2 arrays con los valores de cada variable como elementos,

Covarianza = np.cov(array)

**Correlación:** Es una versión estandarizada de la covarianza. Se divide por la desviación estándar de las variables:

Puede oscilar entre -1 y 1. Si es positiva, indica que la covarianza es positiva y viceversa. No tiene unidades físicas, con lo cual puede servir para comparar variables de diferentes unidades.

Dado un array con 2 arrays con los valores de cada variable como elementos,

Correlación = np.corrcoef(array)