Introducción Estadística y Numpy

OBJETIVOS DE LA CLASE

1. **Repasar las medidas de tendencia central (media, mediana y moda)**
2. **Repasar cómo la media, mediana y moda son afectadas por la asimetría**
3. **Estudiar las medidas de correlación y covarianza.**
4. **Utilizar la librería numpy y scipy**
5. **Realizar un laboratorio integrador de lo que hemos estado haciendo hasta ahora**

- Existen dos campos en la estadística:
  - **Descriptiva**
  - **Inferencial**

- **Foco en estadística descriptiva**: describir, sumarizar y comprender los datos.

- **Medidas de Tendencia Central**: proveer información descriptiva sobre el valor numérico que es considerado el más usual para una variable cuantitativa:
  - **Media**
  - **Mediana**
  - **Moda**

- **Asimetría** en la distribución de datos. Efecto de la media, mediana y moda.

- **Medidas de Variabilidad:**
  - **Rango**
  - **Varianza**
  - **Desvío Estándar**
  - **Coeficiente de Variación**
- NumPy tiene funciones para calcular todas estas medidas, pero antes vamos a ir a los conceptos fundamentales.

La **media** se define de la siguiente manera:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Por ejemplo, para la muestra 8, 5 y -1, su media es:

$$\bar{x} = \frac{8 + 5 + (-1)}{3} = 4$$

La **mediana** puede pensarse de manera simple como el valor del ”medio” de una lista ordenada de datos (o el valor que separa la primera mitad y la segunda mitad de una distribución).
Para una lista ordenada la mediana es calculada de diferente manera dependiendo de la cantidad de elementos de la misma:

- **Impar:**

    [1, 2, 3, 5, **7**, 8, 9, 10, 15]
    #elementos: 9
    La mediana es el valor de la posición 5 (la posición del “medio”)
    Mediana = 7

- **Par:**

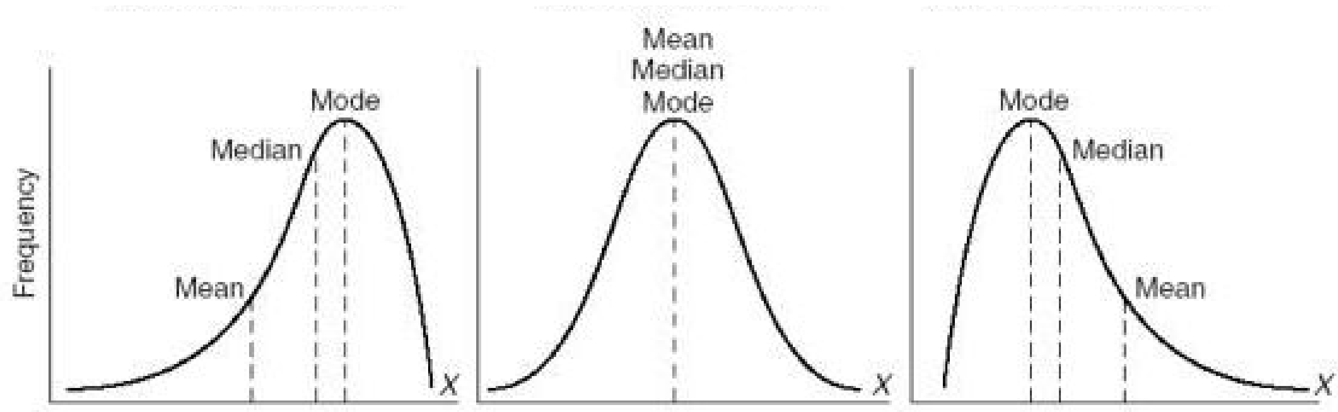
    [-5, -1, 0, **1**, **2**, 3, 8, 20]
    #elementos: 8
    La mediana es la media de los valores en las dos posiciones centrales
    Mediana = (1+2)/2 = 1.5

La **moda** es el valor que aparece con mayor frecuencia o más veces en la distribución.

Por ejemplo, la moda de [0,1,1,2,2,2,2,3,3,4,4,4,5] es 2.

La moda no es necesariamente única. Puede ocurrir que haya dos valores diferentes que sean los más frecuentes. Por ejemplo, para [10, 13, 13, 20, 20], tanto 13 como 20 son la moda.

Nos referimos a la **asimetría** en cuanto a la distribución de los datos[1]:

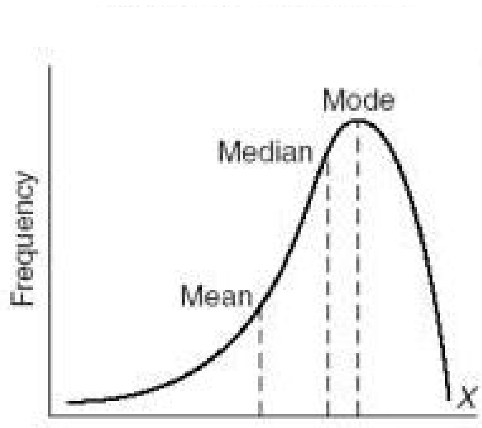


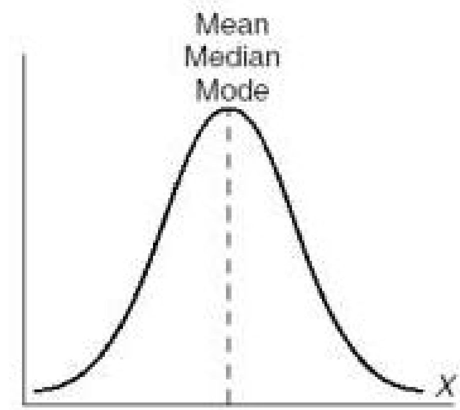
- Una distribución con **asimetría a derecha** significa que la cola del lado derecho es más larga que la de la izquierda (gráfico a la derecha)
- De la misma manera, una distribución con **asimetría a izquierda**, significa que la cola de la izquierda es más larga que la de la derecha (gráfico a izquierda).
- Por último, una **distribución simétrica** no presenta este fenómeno dado que sus colas son de igual longitud al ser simétrica.

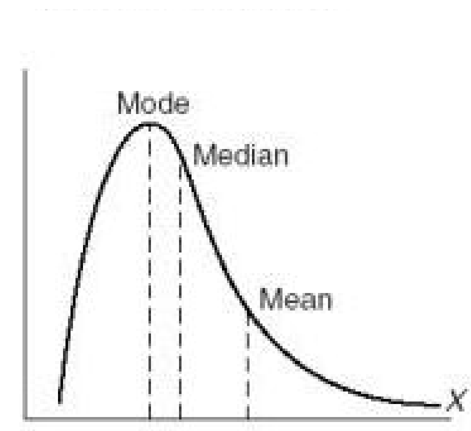
1: Estaremos hablando de asimetría en el contexto de distribuciones unimodales

ASIMETRÍA Y LAS MEDIDAS DE TENDENCIA CENTRAL

La media, mediana y moda son afectadas por la asimetría:







Asimetría a izquierda

Media < Mediana < Moda

Simetría

Media = Mediana = Moda

Asimetría a derecha

Moda < Mediana < Media

Las medidas de variabilidad indican cómo los datos están esparcidos. Nos vamos a focalizar en:

- **Rango**
- **Varianza**
- **Desvío estándar**
- **Coeficiente de variación**

Estas medidas proveen información complementaria (¡y no menos importante!) a las medidas de tendencia central (media, mediana y moda).

El **rango** es la diferencia entre el valor más bajo y más alto de la distribución.

En Python utilizamos la función numpy.ptp:

```
In [2]: import numpy as np
        a = np.array([2,3,6,7,8,11,4,6,17])
        np.ptp(a)

Out[2]: 15
```

La **varianza** es un valor numérico utilizado para describir cuánto varían los números de una distribución respecto a su media.

La varianza puede ser calculada como:

$$s^2 = \frac{\sum (X - \overline{X})^2}{N}$$

Esto es el **promedio de la diferencia elevada al cuadrado entre cada valor y la media**.

En Python utilizamos **numpy.var()** para calcular la varianza.

El **desvío estándar** es la raíz cuadrada de la varianza.

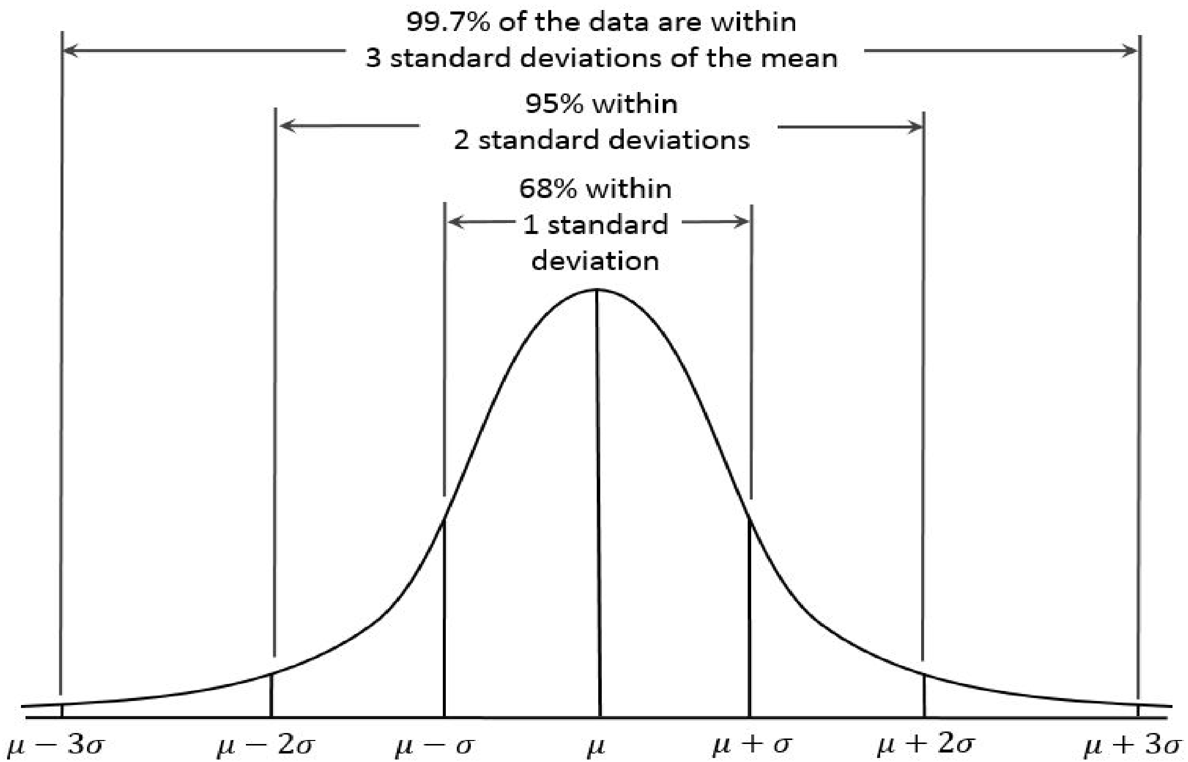
$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{N}}$$

- El desvío es una medida de la dispersión de los datos
- NO ES la desviación promedio con respecto de la media. Como los desvíos están elevados al cuadrado los desvíos muy grandes cuentan más que proporcionalmente.

En Numpy:

```
std = np.std(n)
```

Una ventaja del desvío estándar es que está **expresada en las mismas unidades** que la distribución.
(En cambio, la varianza tiene otras unidades ya que está elevada al cuadrado.)



La **desviación absoluta media** es una medida de dispersión más robusta que la varianza o el desvío estándar. Como se vé en su definición, pondera en igual medida a los valores próximos y a los valores lejanos.

$$\mathrm{MAD} = \mathrm{median}(\ |X_i - \mathrm{median}(X)|\ )$$

- Es la **mediana del valor absoluto de la diferencia a la mediana**.

El **coeficiente de variación** es el desvío estándar dividido por la media

$$CV = \left(\frac{S}{\overline{X}} \right) \cdot 100\%$$

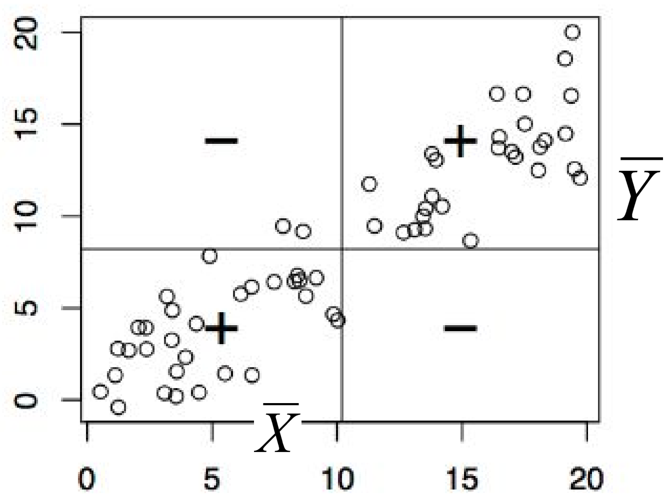
- El coeficiente de variación permite **comparar la dispersión de variables diferentes**.
    - Sirve si las variables tienen medias distintas.
    - También si las variables están expresadas en unidades distintas.

- El coeficiente de variación **no tiene unidades**.

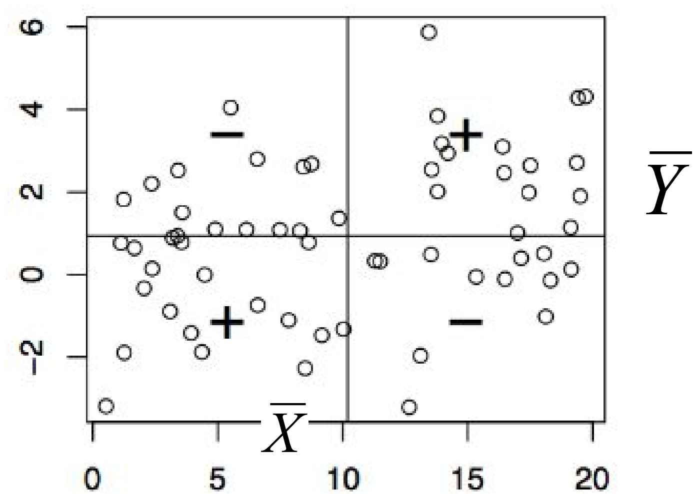
- **Covarianza**
- **Correlación**

- Decimos que dos variables X e Y, tienen covarianza positiva cuando tienden a encontrarse por encima de su media al mismo tiempo y tienen covarianza negativa cuando al mismo tiempo, tienden a encontrarse una por debajo y otra por encima.
- En cambio X e Y tienen covarianza cercana a cero cuando las variables pueden encontrarse por encima o por debajo de su media <u>*independientemente*</u> de lo que haga la otra.
- La covarianza mide la relación lineal entre ambas variables, es decir, qué tanto se asemeja la relación con una función lineal.

Covarianza positiva



Covarianza cercana a cero



La covarianza se mide como:

$$COV_{XY} = \frac{1}{N} \sum_{i=1}^{N} (X - \overline{X})(Y - \overline{Y})$$

La covarianza de un conjunto de datos con p variables se puede representar con una matriz de p x p llamada **matriz de varianzas y covarianzas**:

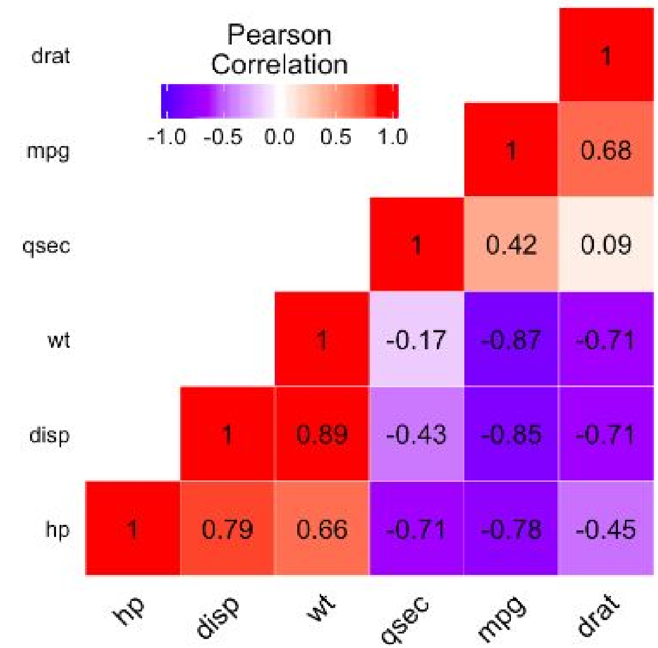
1000*Covariances	^GSPC	^IXIC	XOM	C	GE	MSFT	K	GM
^GSPC	0.633	0.929	0.505	0.495	0.448	0.258	0.261	1.226
^IXIC	0.929	1.737	0.340	0.584	0.507	0.482	0.211	1.842
XOM	0.505	0.340	3.253	-0.421	-0.017	0.268	0.318	2.197
C	0.495	0.584	-0.421	1.923	0.688	0.176	0.277	-0.242
GE	0.448	0.507	-0.017	0.688	1.834	0.761	0.232	0.049
MSFT	0.258	0.482	0.268	0.176	0.761	1.945	0.181	1.315
K	0.261	0.211	0.318	0.277	0.232	0.181	1.045	0.688
GM	1.226	1.842	2.197	-0.242	0.049	1.315	0.688	9.429

* En la diagonal se encuentra la varianza de cada acción

* En el resto de la matriz se encuentran las covarianzas

— La correlación es una **versión estandarizada** (dividida por los desvíos estándar) de la covarianza:

$$r_{xy} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_i - \overline{X}}{S_x}\right)\left(\frac{Y_i - \overline{Y}}{S_y}\right)$$



* La correlación está acotada entre 1 y -1.

* Siempre que la covarianza es positiva, la correlación es positiva y viceversa.

* Mientras que la correlación no tiene unidades físicas, la covarianza sí.

1. **Práctica guiada**

2. Práctica independiente

3. Laboratorio

1. Práctica guiada
2. **Práctica independiente**
3. Laboratorio

1. Práctica guiada
2. Práctica independiente
3. **Laboratorio**

FIN
Introducción Estadística y Numpy