

ACÀMICA

TEMA DEL DÍA

Estadística y Pandas

Hoy repasaremos algunos conceptos estadísticos, en particular Estadística Descriptiva. Luego, veremos cómo hacemos en Python para trabajar con conjuntos de datos usando Pandas.



Agenda

Daily

Explicación: Máscara, Pandas.

Break.

Hands-on training: Pandas

Cierre.



Daily



Daily



Sincronizando...

Toolbox



¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

Challenge



¿Cómo te ha ido?
¿Obstáculos?
¿Cómo seguimos?

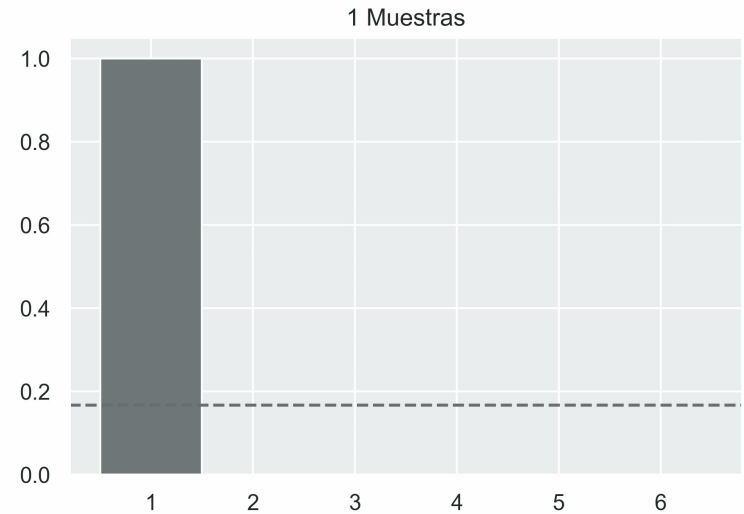
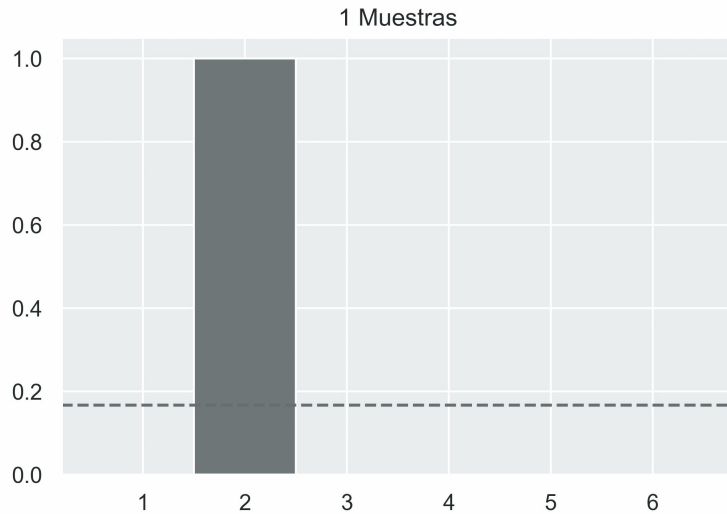
Probabilidad y estadística



Tenemos dos dados. Suponemos que uno está cargado. ¿Cómo descubrimos cuál es?



Tenemos dos dados. Suponemos que uno está cargado.
¿Cómo nos damos cuenta cuál?



Tipos de valores estadísticos:

Media: es el valor promedio estándar (lo que siempre conocimos por promedio).

Mediana: es el valor medio exacto en un conjunto de datos ordenados. Es decir, el 50% de los valores son menores que la media y el 50% son mayores.

Moda: el valor con mayor frecuencia en un conjunto de datos.

Ejemplo

Muestra: {5, 6, 7, 6, 7, 8, 6, 5, 6}

- **Media = 6.22**
- **Mediana = 6**
5, 5, 6, 6, 6, 6, 7, 7, 8
- **Moda = 6**
5, 5, 6, 6, 6, 6, 7, 7, 8

Varianza

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).

$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Varianza

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).

$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Promedio de la
muestra

Varianza

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).

Elementos de la muestra

$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Promedio de la muestra

Varianza

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).

El símbolo de sumatoria nos indica que debemos sumar sobre todos los valores del conjunto

Elementos de la muestra

$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

Promedio de la muestra

Varianza

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).

El símbolo de sumatoria nos indica que debemos sumar sobre todos los valores del conjunto


The diagram shows the formula for sample variance:
$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$
 Four red arrows point from text labels to parts of the formula: 1. From 'Elementos de la muestra' to the x_i term. 2. From 'Promedio de la muestra' to the \bar{x} term. 3. From 'Cantidad de elementos en la muestra' to the n in the denominator. 4. From 'El símbolo de sumatoria nos indica que debemos sumar sobre todos los valores del conjunto' to the summation symbol \sum .

Elementos de la muestra


Promedio de la muestra

Cantidad de elementos en la muestra

Veamos un ejemplo:

- Muestra: {5, 10, 8, 20} 
- N es 4
- El promedio, \bar{X} es 10,75

$$Var = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$


$$Var = \frac{(5-10,75)^2 + (10-10,75)^2 + (8-10,75)^2 + (20-10,75)^2}{4-1}$$

Máscaras



Máscaras - Filtros Booleanos

```
[66]: arreglo2d = np.arange(30).reshape(6,5)  
arreglo2d
```

```
[66]: array([[ 0,  1,  2,  3,  4],  
            [ 5,  6,  7,  8,  9],  
            [10, 11, 12, 13, 14],  
            [15, 16, 17, 18, 19],  
            [20, 21, 22, 23, 24],  
            [25, 26, 27, 28, 29]])
```

Máscaras - Filtros Booleanos

```
[66]: arreglo2d = np.arange(30).reshape(6,5)  
arreglo2d
```

```
[66]: array([[ 0,  1,  2,  3,  4],  
          [ 5,  6,  7,  8,  9],  
          [10, 11, 12, 13, 14],  
          [15, 16, 17, 18, 19],  
          [20, 21, 22, 23, 24],  
          [25, 26, 27, 28, 29]])
```

→
Creamos la
máscara

```
[67]: mask = arreglo2d < 20  
mask
```

```
[67]: array([[ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [False, False, False, False, False],  
          [False, False, False, False, False]])
```

Máscaras - Filtros Booleanos

```
[66]: arreglo2d = np.arange(30).reshape(6,5)  
arreglo2d
```

```
[66]: array([[ 0,  1,  2,  3,  4],  
          [ 5,  6,  7,  8,  9],  
          [10, 11, 12, 13, 14],  
          [15, 16, 17, 18, 19],  
          [20, 21, 22, 23, 24],  
          [25, 26, 27, 28, 29]])
```

→
Creamos la
máscara

```
[67]: mask = arreglo2d < 20  
mask
```

```
[67]: array([[ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [ True,  True,  True,  True,  True],  
          [False, False, False, False, False],  
          [False, False, False, False, False]])
```

```
[68]: arreglo2d[mask]
```

```
[68]: array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,  
          17, 18, 19])
```

←
Y seleccionamos aquellos
elementos que cumplen la
condición que representa
la máscara

Pandas



DATASET

Es el conjunto de datos que utilizaremos en el workflow de data science. Los podemos generar, obtener de terceros o simular.

datasets
estructurados

similar a planilla de cálculo. Información pre-procesada. Suelen venir en .txt, .csv, .xlsx, .json, etc.

datasets
no estructurados

audio, imágenes, texto en crudo
humanos / redes neuronales



DATASET

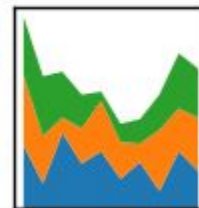
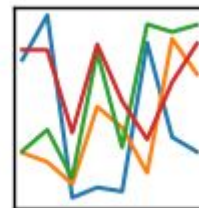
datasets
estructurados

similar a planilla de cálculo. Información pre-procesada. Suelen venir en .txt, .csv, .xlsx, .json, etc.

Para trabajar con datasets estructurados (y bueno, más), la librería estándar de Python es:

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



ARGENTINA DATASET

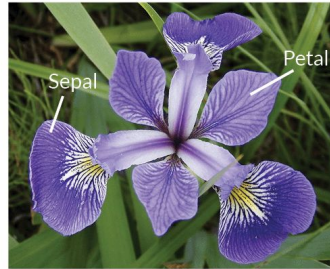


División Política,
Superficie y
Población



IRIS DATASET

Famoso dataset introducido por Ronald Fisher (padre de la estadística) en 1936.



Iris Versicolor



Iris Setosa



Iris Virginica



A close-up photograph of a white ceramic cup filled with a latte. The surface of the milk is decorated with intricate latte art, featuring a central heart shape surrounded by concentric, wavy lines. The cup is placed on a matching white saucer. In the background, a white napkin and a silver fork are visible, though they are out of focus. The overall lighting is soft and even.

¡BREAK!

Pandas: Instalación

1. Activar el ambiente: *"conda activate datascience"*
2. Instalar Pandas: *"conda install pandas"*

Hands-on training





Trabajamos en el Notebook que descargaste en la Toolbox 04, Sección 2: Pandas

Buenas prácticas de un data scientist



Buenas prácticas de ~~un data scientist~~ programador



PEP-20: The Zen of Python

Beautiful is better than ugly.
Explicit is better than implicit.
Simple is better than complex.
Complex is better than complicated.
Flat is better than nested.
Sparse is better than dense.
Readability counts.
Special cases aren't special enough to break the rules.
Although practicality beats purity.
Errors should never pass silently.
Unless explicitly silenced.
In the face of ambiguity, refuse the temptation to guess.
There should be one -and preferably only one- obvious way to do it.
Although that may not be obvious at first unless you're Dutch.
Now is better than never.
Although never is often better than **right** now.
If the implementation is hard to explain, it's a bad idea.
If the implementation is easy to explain, it may be a good idea.
Namespaces are one honking idea --let's do more of those!



Recursos



Probabilidad y Estadística

- <https://seeing-theory.brown.edu/basic-probability/index.html> - Este recurso no aparece en la Toolbox, pero pueden mirarlo si tienen tiempo.

Pandas

- [Python Data Science Handbook](#) - Capítulo 3, “Data Manipulation With Pandas”.



Para la próxima

- Termina el notebook de hoy
- Lee la Toolbox 05
- Resuelve el Challenge.

ACÀMICA