Python - Estadística

Conceptos

Rogelio Ferreira Escutia

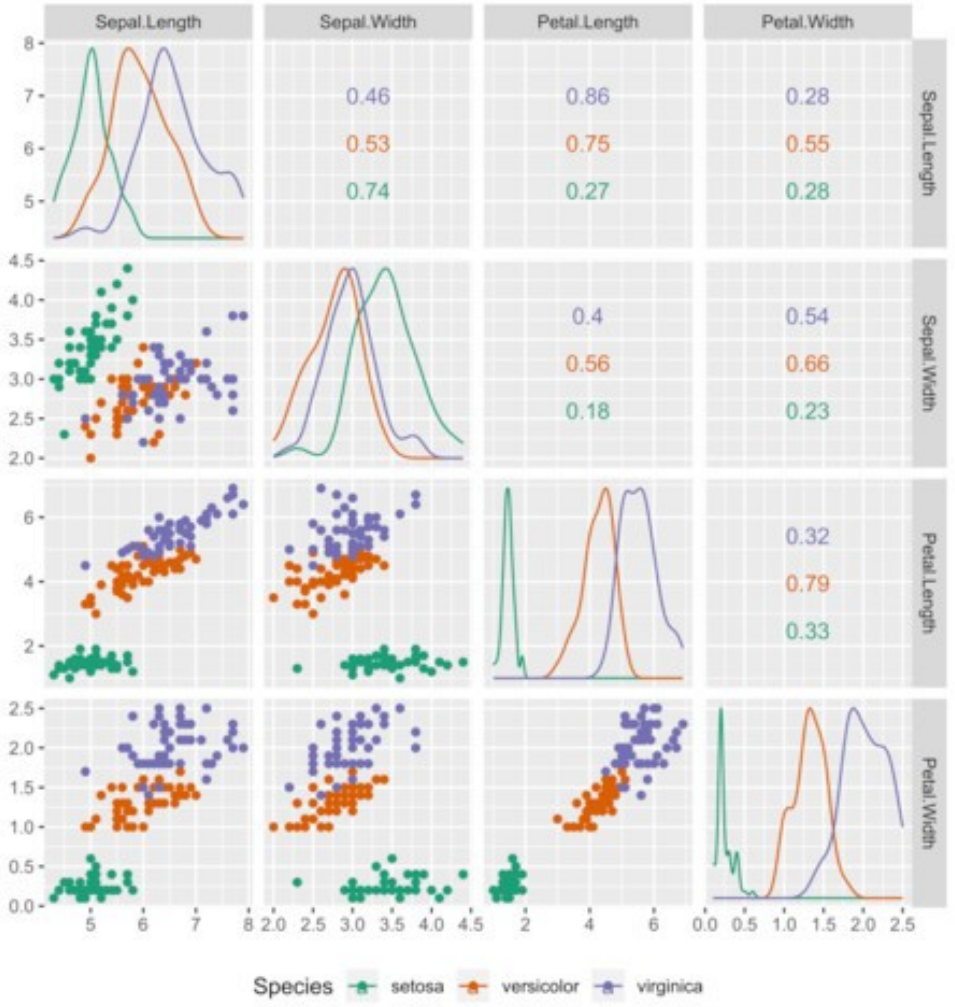
Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

xumarhu.net

Conceptos

Definición

- **“Es el área que se dedica a la colección, organización, análisis, interpretacion y representación de los datos”**



“Statistics”, https://en.wikipedia.org/wiki/Statistics, abril 2021

Datos

- **Tipos de datos según su contenido:**

Analytical Data Type	Computational Data Type	Typical Statistics	Typical Visualizations
Nominal, categorical	Character, string,	Counts, relative frequencies (%), modes	Bar chart, pie chart*
Ordinal	Character, string	Counts, relative frequencies (%), modes, percentiles, quartiles	Bar chart
Interval	Integer, floating point	Max, min, range, mean (arithmetic), median, mode, differences, std dev, addition, subtraction	Boxplot, histogram
Ratio	Integer, floating point	Max, min, range, median, mean (arithmetic and geometric), ratios, differences, std dev, addition, subtraction, multiplication, division	Boxplot, histogram

"Creating Good Data", Harry J. Foxwell, Apress Media, USA 2020

Datos

- **Categoría: Sexo, escuela, modelo de carro, etc.**
- **Numérica: calificación, cantidad de dinero**
- **Contínua: peso, tiempo, altura**
- **Cualitativa: bueno, malo**

"Statisticas", Derek Banas, https://www.youtube.com/watch?v=YCPYNXtwKAc, abril 2021

Bibliotecas utilizadas

Bibliotecas mas utilizadas

- **statistics: Es la biblioteca que viene incluida en Python y que está orientada a estadística.**
- **NumPy: Biblioteca orientada cómputo numérico.**
- **SciPy: Es para computación científica y está basado en NumPy (incluye scipy.stats la cual es orientada específica mente a estadística).**
- **Pandas: Basada en NumPy es utilzada para arreglos multidimensionales utilizando los dataframes.**
- **Matplotlib: Sirve para graficar y se combina muy buen con NumPy, SciPy y Pandas.**

Estadística Descriptiva

Población

- **Población es un conjunto de elementos similares a los que se les someterá a un análisis estadístico.**

- **Este conjunto puede estar integrado por objetos como aviones, mesas, personas, estrellas, etc..**

- **También pueden ser valores que nos representan variables físicas como temperatura, presión, humedad, etc..**

Muestra

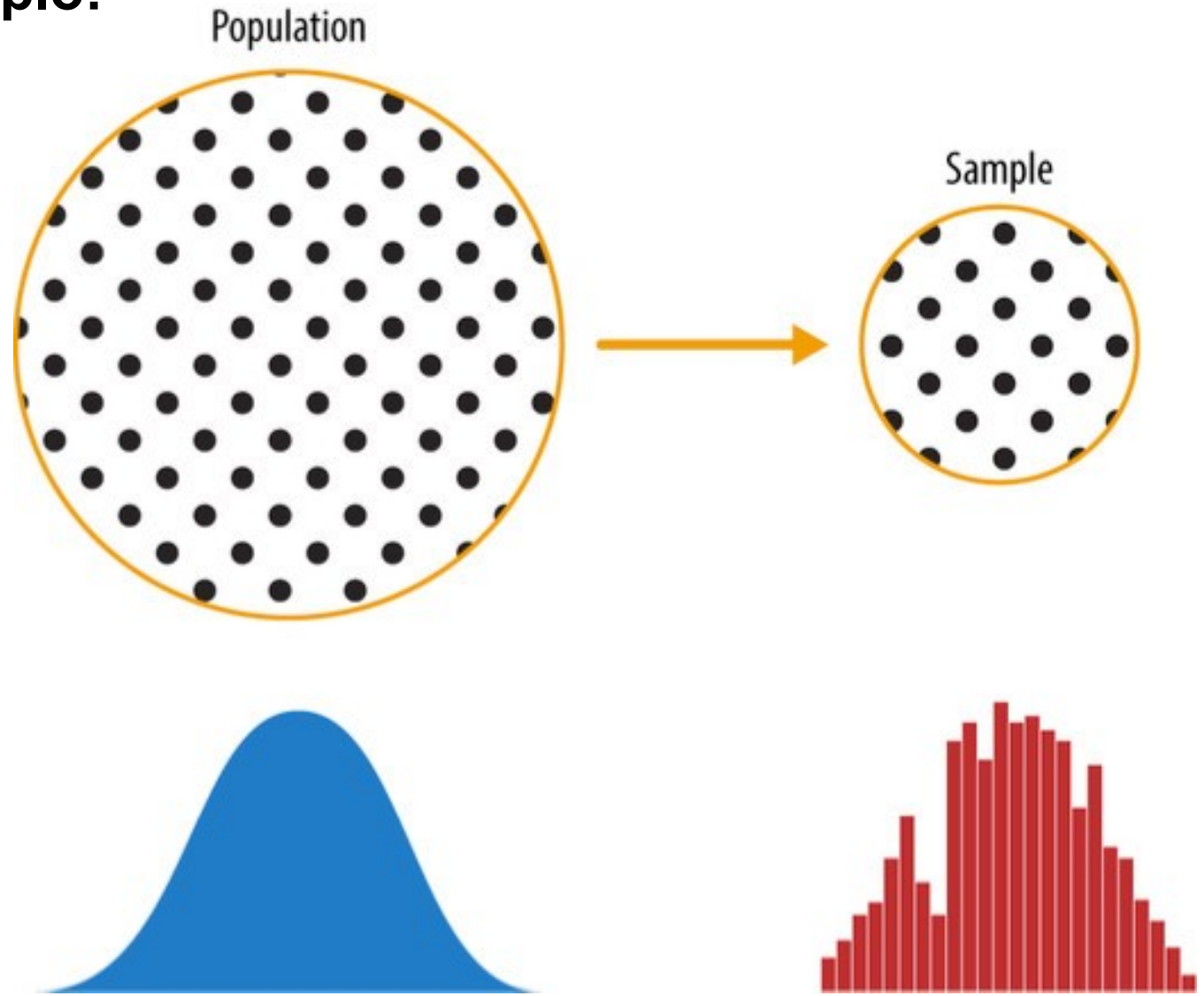
- **En muchas ocasiones la población a analizar puede ser muy grande o muy compleja y tardada de analizar, por lo cual se elige escoger una "muestra", la cual es un subconjunto de la población y que debe ser "representativa" de la población.**

- **La muestra puede ser pequeña y fácil de analizar, pero también, puede ser que no represente a la población de la cual se extrajo, por lo cual se dice que es una muestra "sesgada".**

- **El análisis de una muestra "no representativa" ó "sesgada", nos llevará a conclusiones y suposiciones erróneas.**

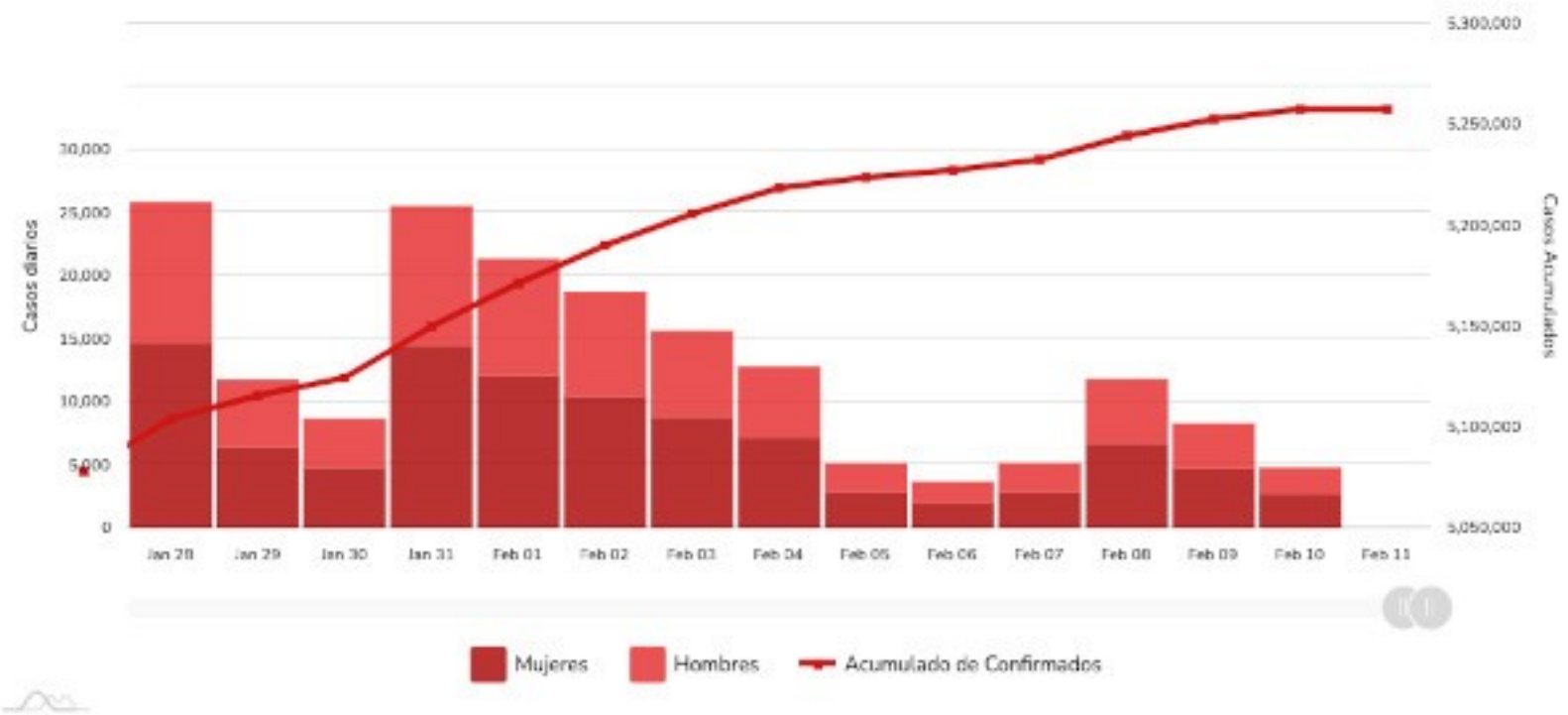
Población y Muestra

- **Ejemplo:**



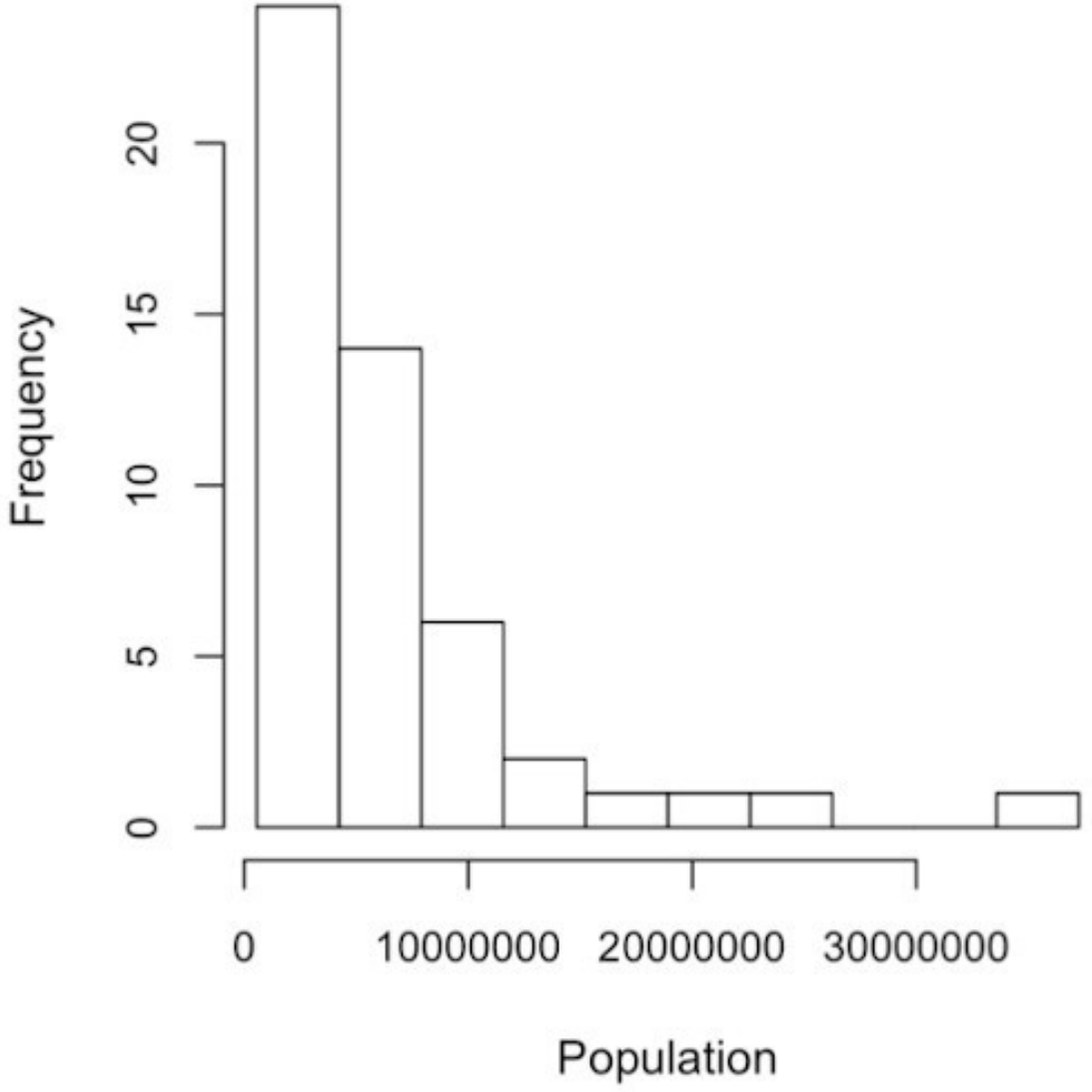
Histogramas

- **En el área de estadística se utiliza mucho el histograma, el cual es un gráfico donde se representa el conjunto de datos a analizar y se representa en forma de barras, donde el tamaño de las barras es proporcional a la cantidad de datos a representar.**



Histograma

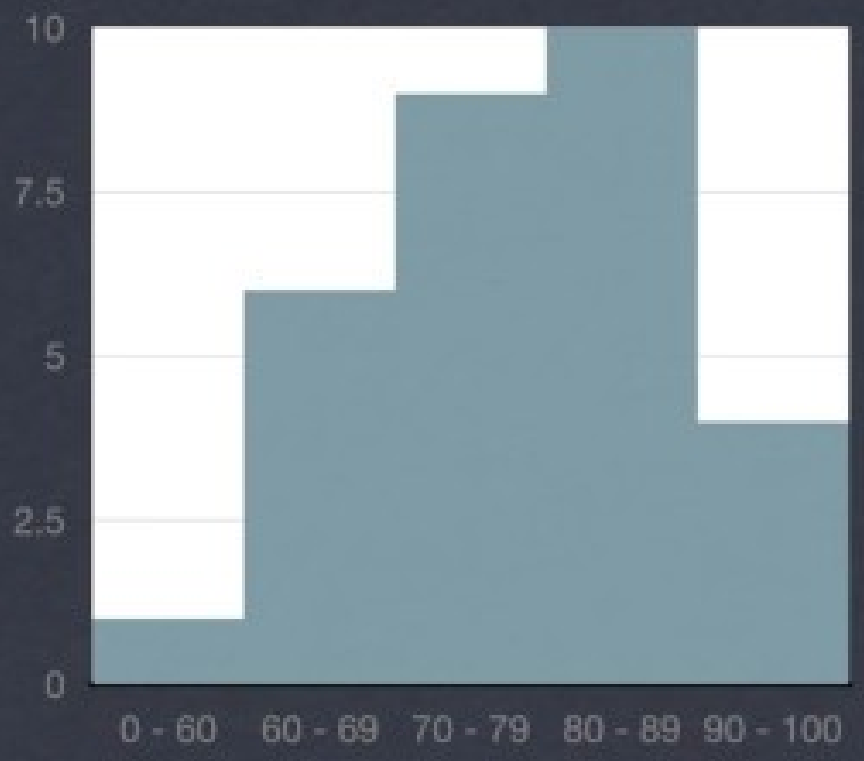
- **Gráfica de los datos que mas se repiten:**



"Practical Statistics for Data Scientist", Peter Bruce, Andrew Bruce & Peter Gedeck, O'Reilly, California USA 2020

Frecuencias e Histograma

TEST SCORE	# of STUDENTS
0 - 60	1
60 - 69	6
70 - 79	9
80 - 89	10
90 - 100	4
TOTAL	30

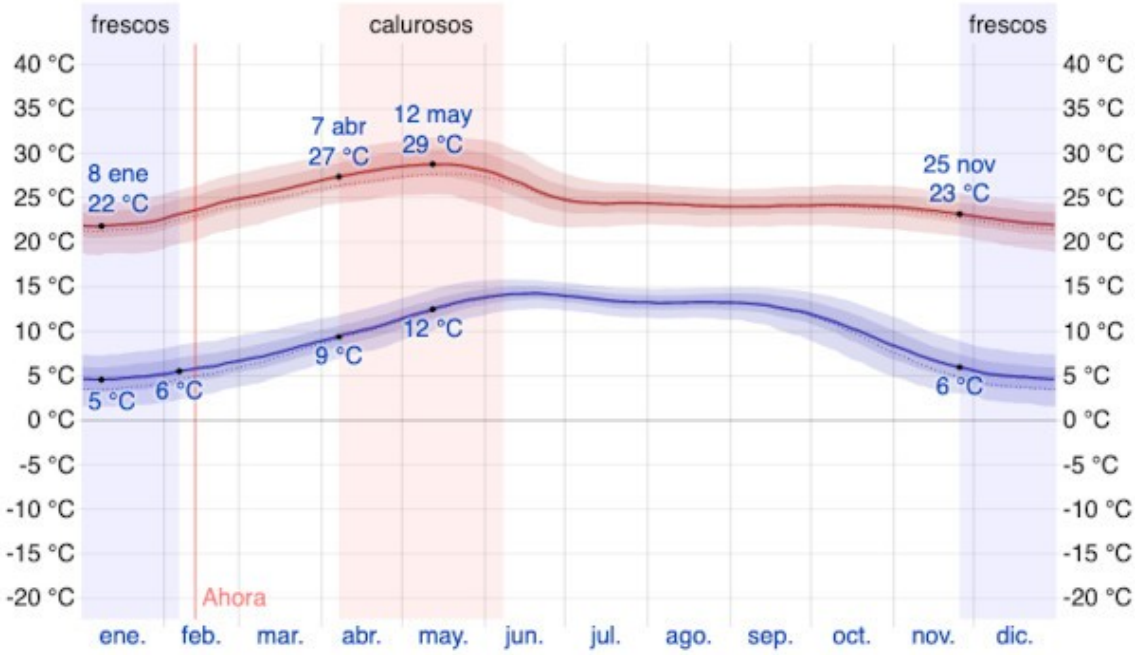


"Statisticas", Derek Banas, https://www.youtube.com/watch?v=YCPYNXtwKAc, abril 2021

Mínimo y Máximo

- **Dentro de un conjunto de datos, el “máximo” es el valor mas alto y el “mínimo” es el valor más bajo de todos. Estos puntos son importantes para delimitar el área donde se distribuyen los valores de nuestros datos.**

- **A continuación se muestra la gráfica de temperaturas máximas y mínimas que se registran para la ciudad de Morelia según el sitio “Weather Spark”:**



Media

- **La “Media Aritmética” (ó “Promedio” como se le dice comúnmente), es la suma de todos los componentes entre el número de componentes (media aritmética).**

$$\mathrm{AM} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

Promedio

```
datos = [40, 17, 35, 16]

def promedio(lista):
    suma = 0
    for i in lista:
        suma = suma + i
    promedio = suma / len(lista)
    return promedio

print("\nLista de datos: ",datos)
print("\nEl Promedio es: ", promedio(datos),"\n")
```

```
Lista de datos:  [40, 17, 35, 16]

El Promedio es:  27.0
```

Mediana

- **Dado un conjunto de datos ordenados, la “mediana” es el valor que se encuentra al centro de todos los valores.**

- **Si el número de datos es par, entonces la “mediana” es la media de los 2 valores centrales.**

Mediana

```
datos = [40, 17, 35, 16]        # Para probar con una lista de tamaño par
datos = [40, 17, 35, 16, 30]    # Para probar con una lista de tamaño impar

def mediana(lista):
    lista_ordenada = sorted(lista)
    print("Lista ordenada: ",lista_ordenada)
    if len(lista_ordenada)% 2 == 0:
        i = int((len(lista_ordenada)) / 2) - 1
        j = i + 1
        mediana = (lista_ordenada[i] + lista_ordenada[j]) / 2
    else:
        i = int((len(lista_ordenada)) / 2)
        mediana = lista_ordenada[i]
    return mediana

print("\nLista de datos: ",datos)
print("La Mediana es: ", mediana(datos),"\n")
```

```
Lista de datos:  [40, 17, 35, 16, 30]
Lista ordenada:  [16, 17, 30, 35, 40]
La Mediana es:  30
```

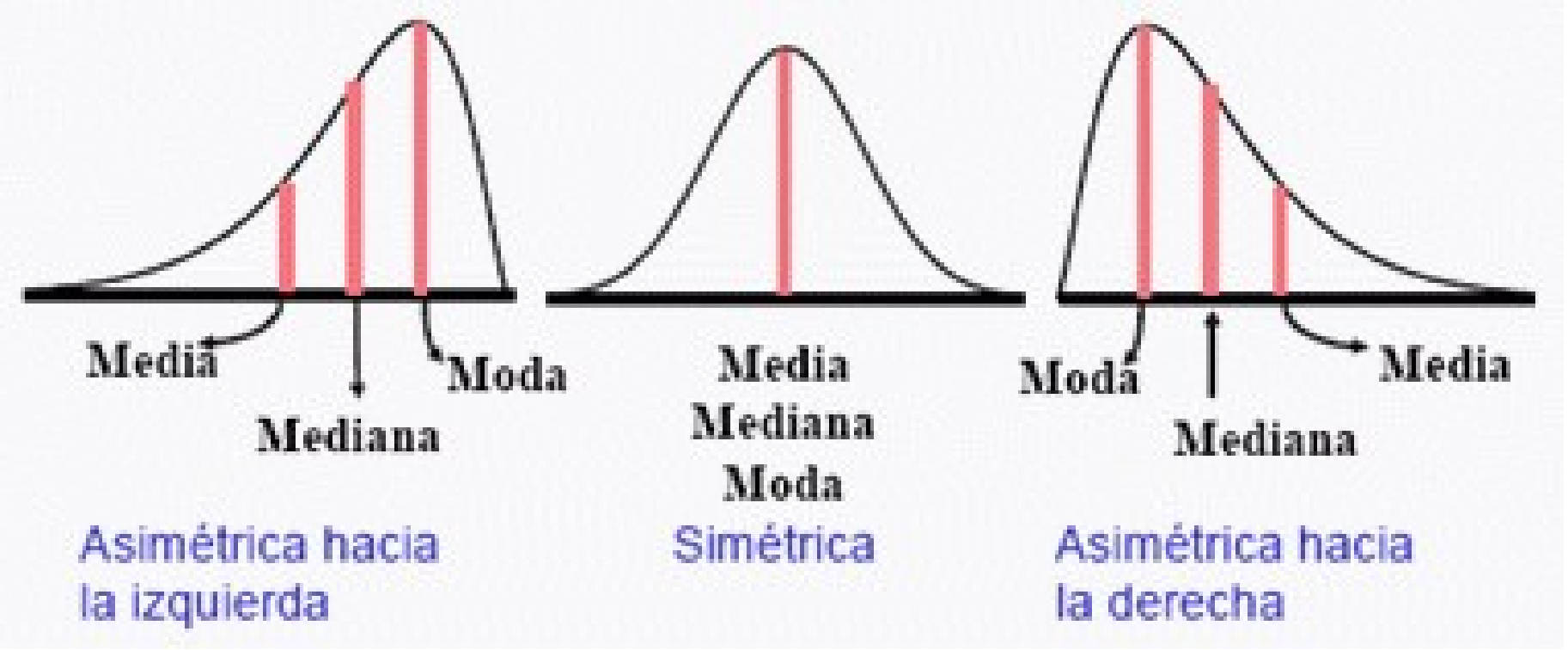
Moda (1)

- **“Moda” es el valor que mas veces se repite dentro de un conjunto de datos.**

Moda (2)

```
from collections import defaultdict

sample = [1, 3, 2, 5, 7, 0, 2, 3]

def mode(values):
    counts = defaultdict(lambda: 0)

    for s in values:
        counts[s] += 1

    max_count = max(counts.values())
    modes = [v for v in set(values) if counts[v] == max_count]
    return modes

print(mode(sample))
```

Media, Mediana y Moda



Medidas de Dispersión

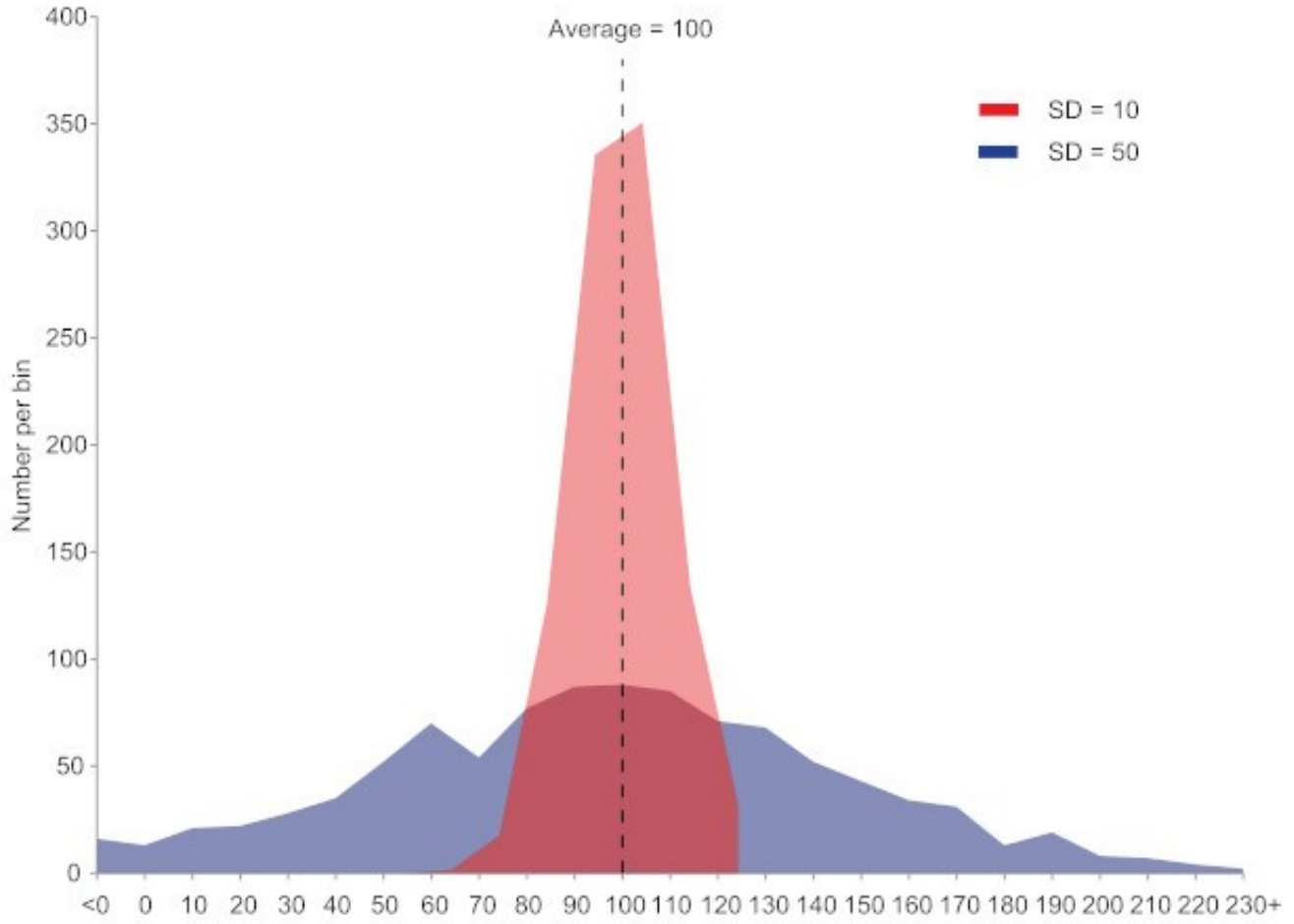
Dispersión

- **La “dispersión” es muy importante, ya que nos indica el grado de separación de nuestros datos.**

- **Otra manera de expresarlo, es que la dispersión es, qué tanto se estiran ó se comprimen.**

Dispersión

- **La “dispersión” se puede observar a través de la siguiente gráfica donde se observan 2 conjuntos de datos que tienen el mismo promedio, pero la muestra azul tiene mayor dispersión y la roja menos dispersión**



Varianza

- **La “varianza” (también denominada “variancia”) es una medida de dispersión, la cual nos sirve para representar la variabilidad de un conjunto de datos con respecto a la media aritmética de los mismos datos.**

- **La varianza siempre tiene valores iguales o mayores a 0.**

Varianza

VARIANCE

- How is Data Spread Around Mean?
- σ^2 : Variance of Population
- S^2 : Variance of Sample

SAMPLE [4, 3, 6, 5, 2]

$$\text{MEAN} = \frac{20}{5} = 4$$

$$S^2 = \frac{\Sigma (x_i - \bar{x})^2}{n - 1}$$

$$= \frac{(4 - 4)^2 + (3 - 4)^2 + (6 - 4)^2 + (5 - 4)^2 + (2 - 4)^2}{4} = \frac{10}{4} = 2.5$$

"Statisticas", Derek Banas, https://www.youtube.com/watch?v=YCPYNXtwKAc, abril 2021

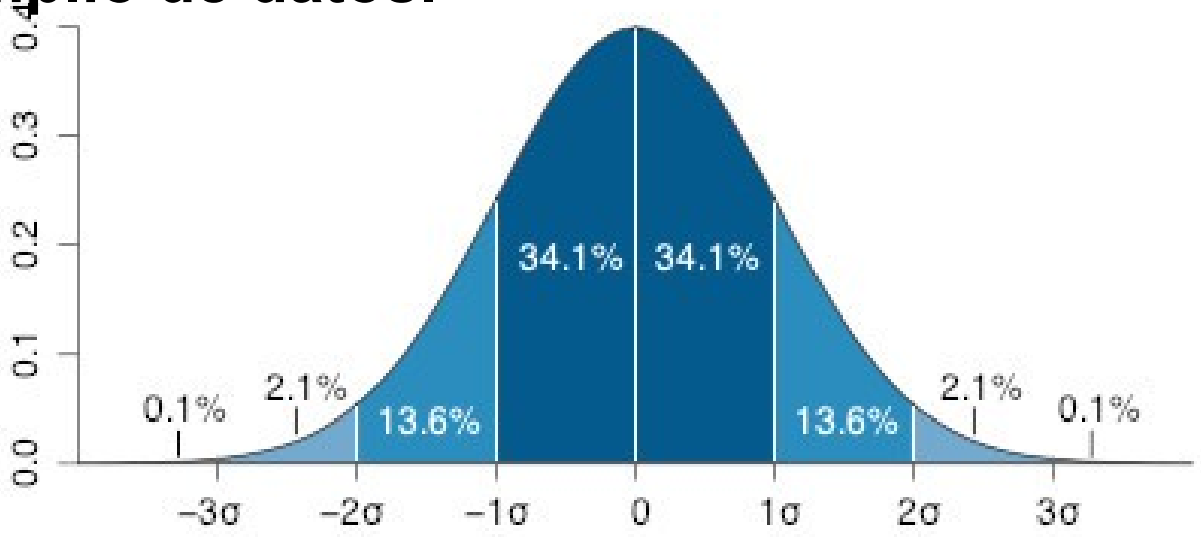
Varianza

```
sample = [4,3,6,5,2]

def variance(values):
    mean = sum(values) / len(values)
    var = sum((v - mean) ** 2 for v in values) / len(values)
    return var

print(variance(sample))  # prints 4.359375
```

Desviación Estándar

- **La “desviación estándar” (también conocida como “desviación típica”) es una medida para cuantificar la dispersión de un conjunto de datos y se calcula como la raíz cuadrada de la varianza.**

- **Si la desviación estándar es baja, entonces indica que los datos están agrupados cerca de la media del conjunto de datos. Y al contrario, entre más alta sea la desviación estándar, indica que los valores extienden sobre un rango mas amplio de datos.**

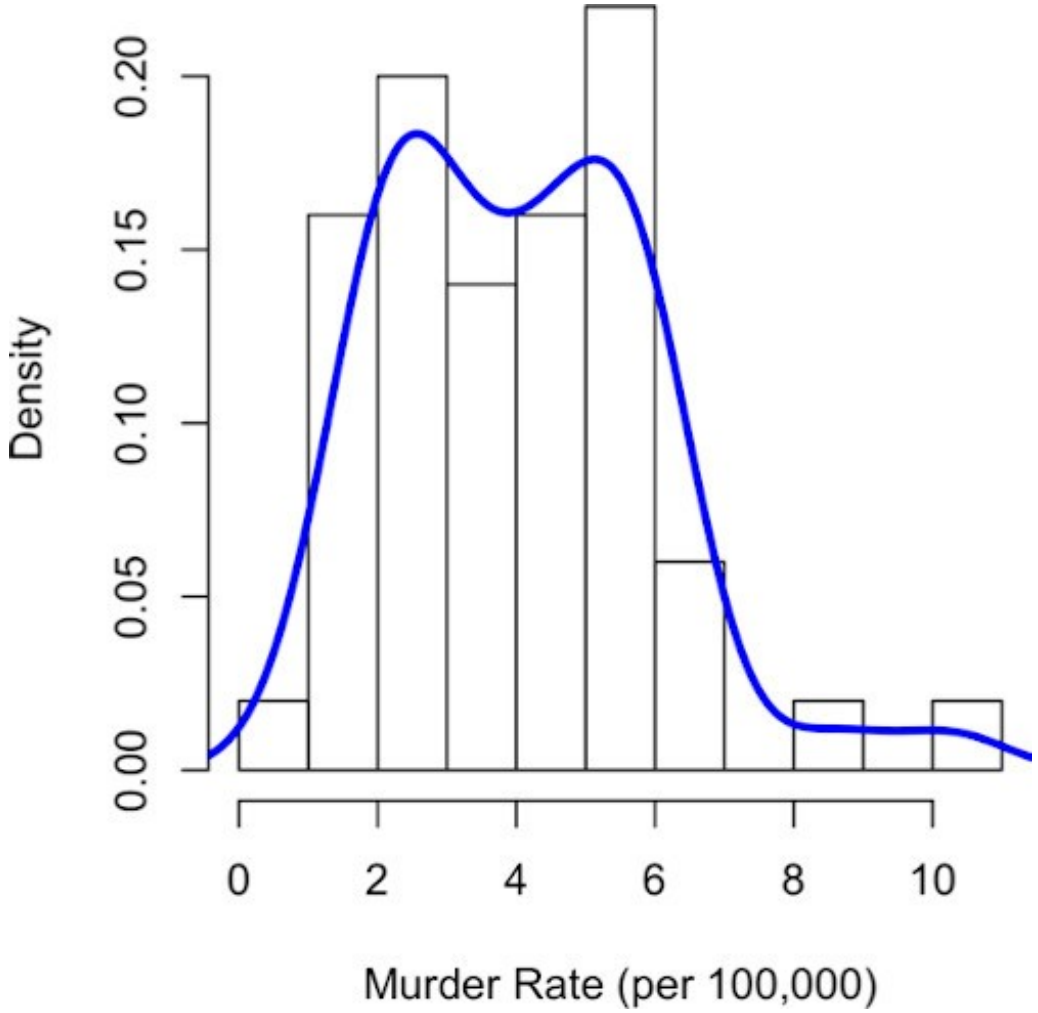


Desviación Estándar

```python
from math import sqrt

# Number of pets each person owns
sample = [1, 3, 2, 5, 7, 0, 2, 3]

def variance(values):
    mean = sum(values) / len(values)
    var = sum((v - mean) ** 2 for v in values) / (len(values) - 1)
    return var

def std_dev(values):
    return sqrt(variance(values))

print(std_dev(sample))  # prints 2.0879116360612584
```

Densidad

- **Gráfica de la concentración de los datos:**



"Practical Statistics for Data Scientist", Peter Bruce, Andrew Bruce & Peter Gedeck, O'Reilly, California USA 2020

Correlación

- **La “correlación” es un valor que nos indica la cantidad de proporcionalidad existente entre 2 variables, cuando una de ellas varía sistemáticamente con respecto a los valores homónimos de la otra.**

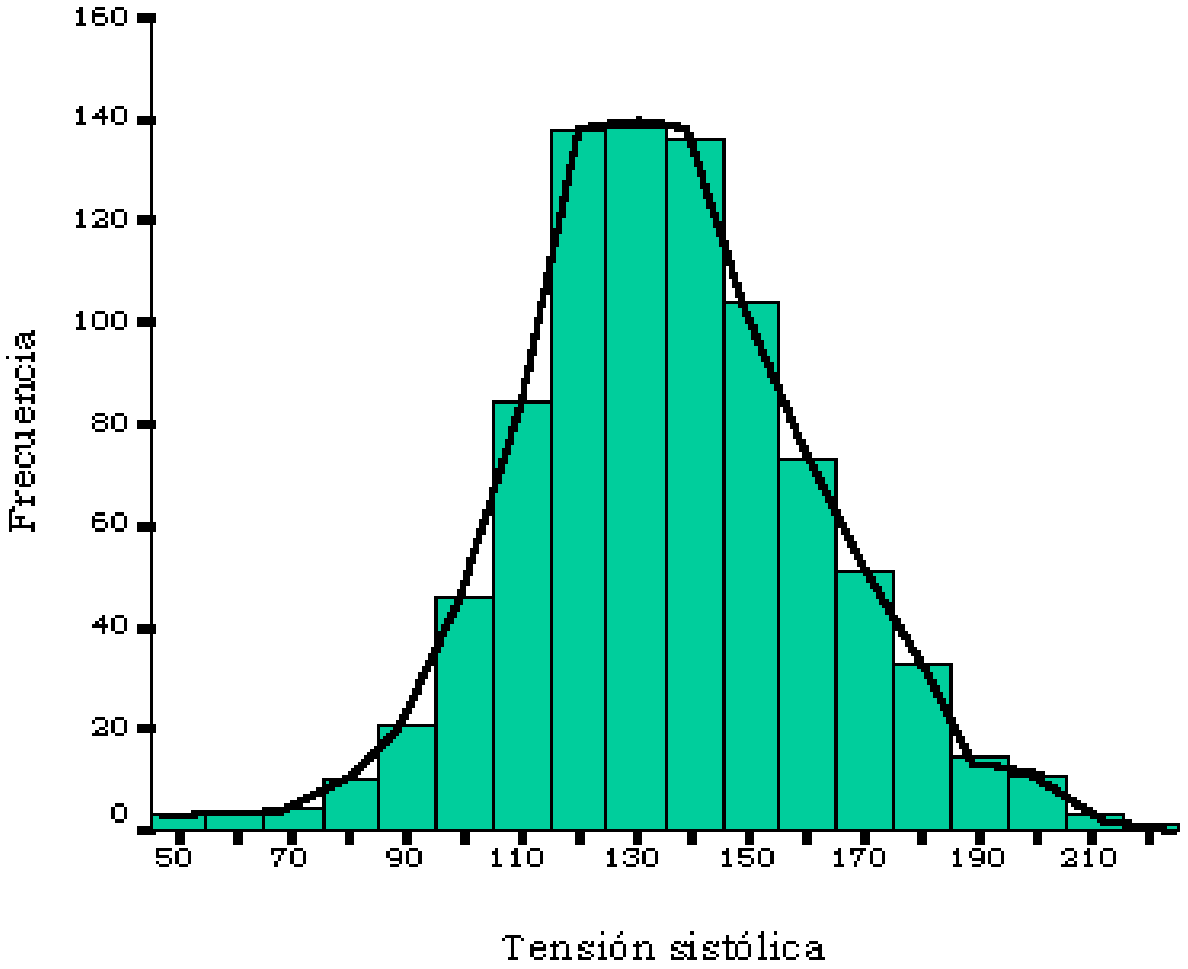
- **Si la correlación es alta entre las variables A y B, esto nos indica que si los valores de A aumentan o disminuyen, también lo hacen los valores de B.**

- **La correlación entre 2 variables no implica directamente que haya alguna relación de causalidad entra ambas.**

Distribuciones

Distribución Normal

- **Distribución de una muestra de la presión arterial de 5000 pacientes:**

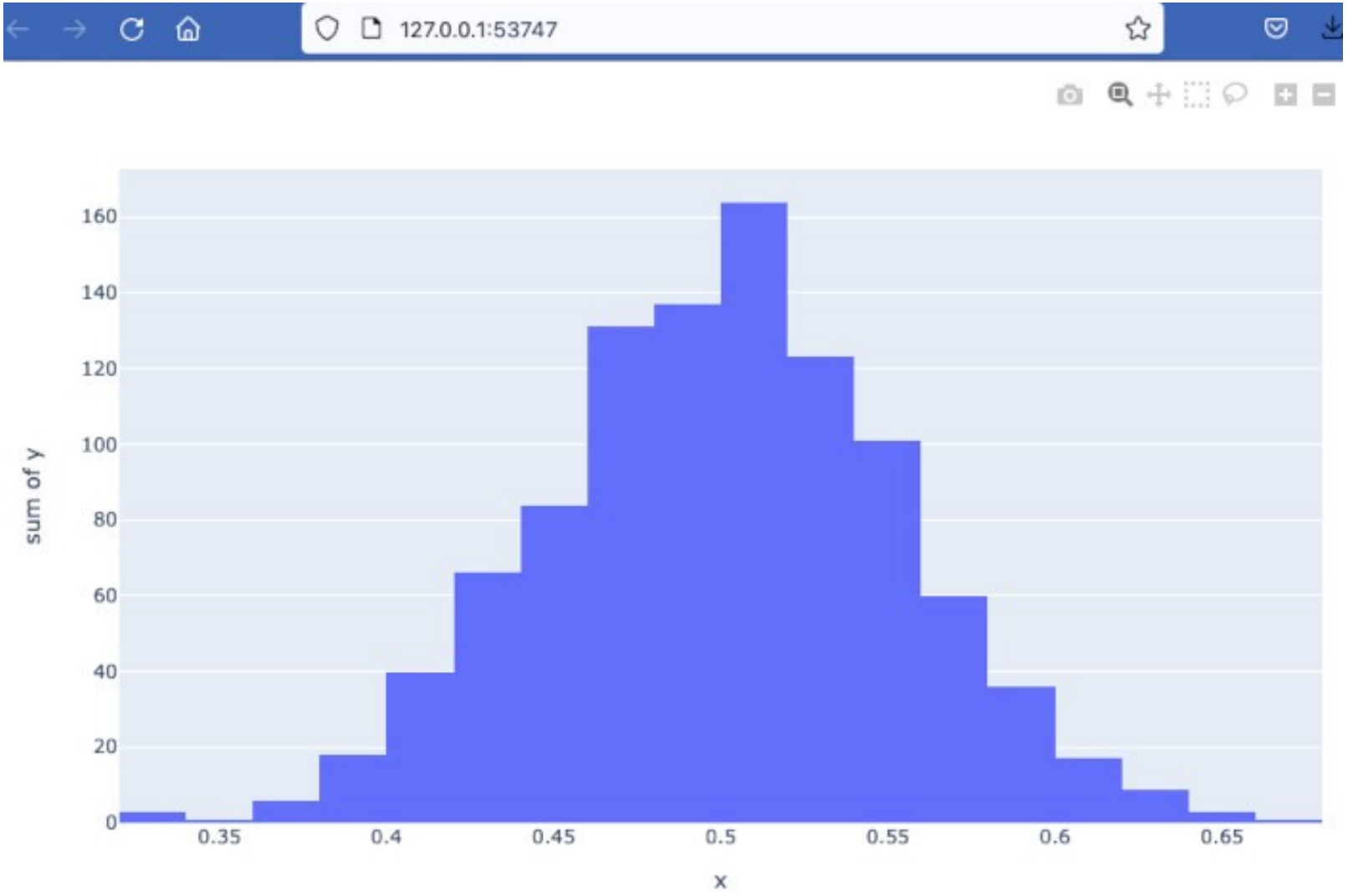


"Distribución normal", https://www.fisterra.com/mbe/investiga/distr_normal/distr_normal.asp, abril 2021

Distribución Normal (1)

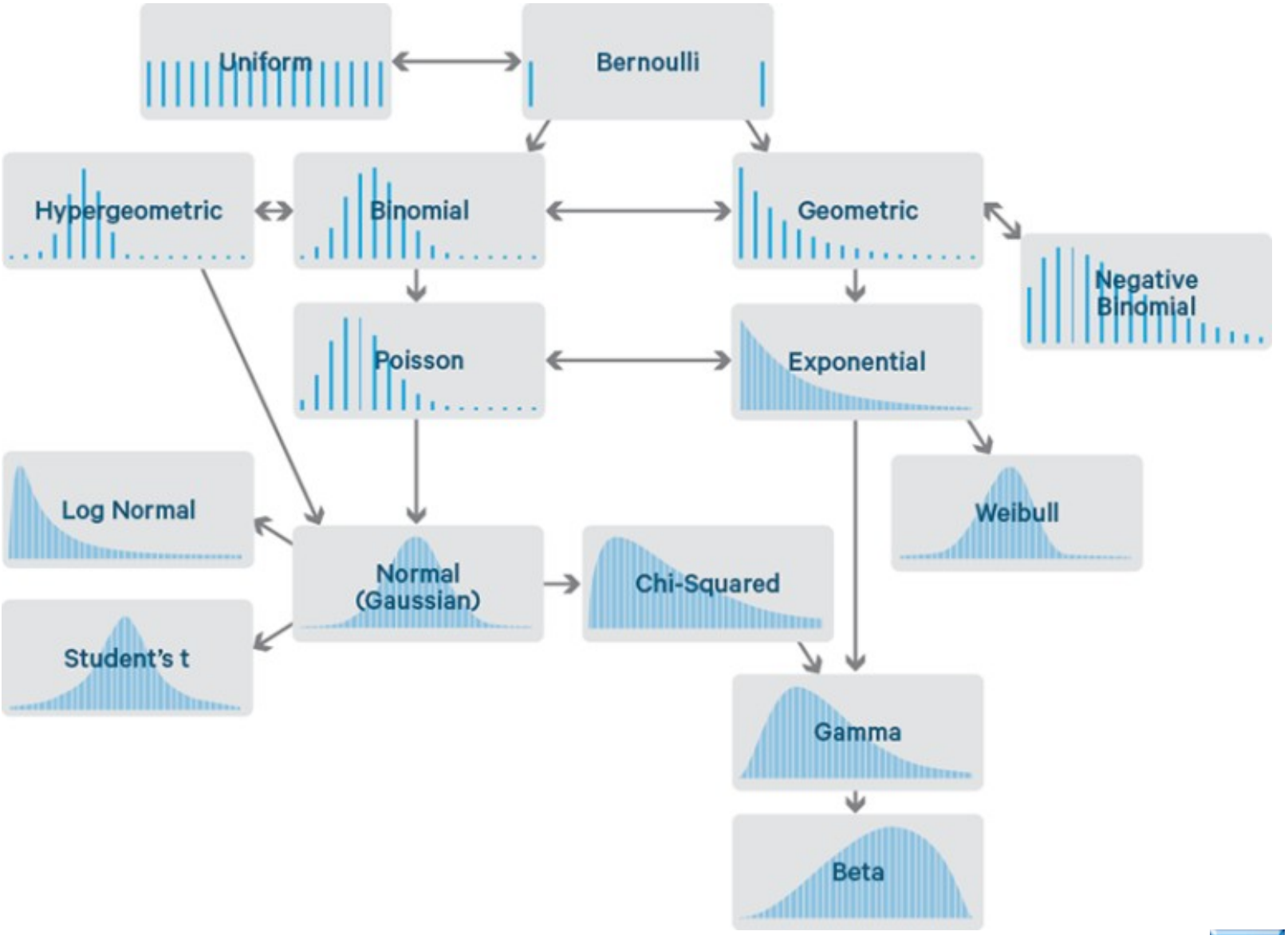
```
import random

import plotly.express as px

sample_size = 30
sample_count = 1000

# Central limit theorem, 1000 samples each with 30 random numbers between 0.0 and 1.
0
x_values = [(sum([random.uniform(0.0, 1.0) for i in range(sample_size)]) /
sample_size) for _ in range(sample_count)]
y_values = [1 for _ in range(sample_count)]

px.histogram(x=x_values, y = y_values, nbins=20).show()
```

Distribución Normal (2)



"Essential Math for Data Science", Thomas Nield, O'Reilly 2022

Distribuciones



"¿Cómo saber si una variable sigue una distribución normal en Python?", https://machinelearningparatodos.com/como-saber-si-una-variable-sigue-una-distribucion-normal-en-python/, abril 2021

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

rogelio.fe@morelia.tecnm.mx

rogeplus@gmail.com

xumarhu.net

@rogeplus

https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg

rogelioferreiraescutia