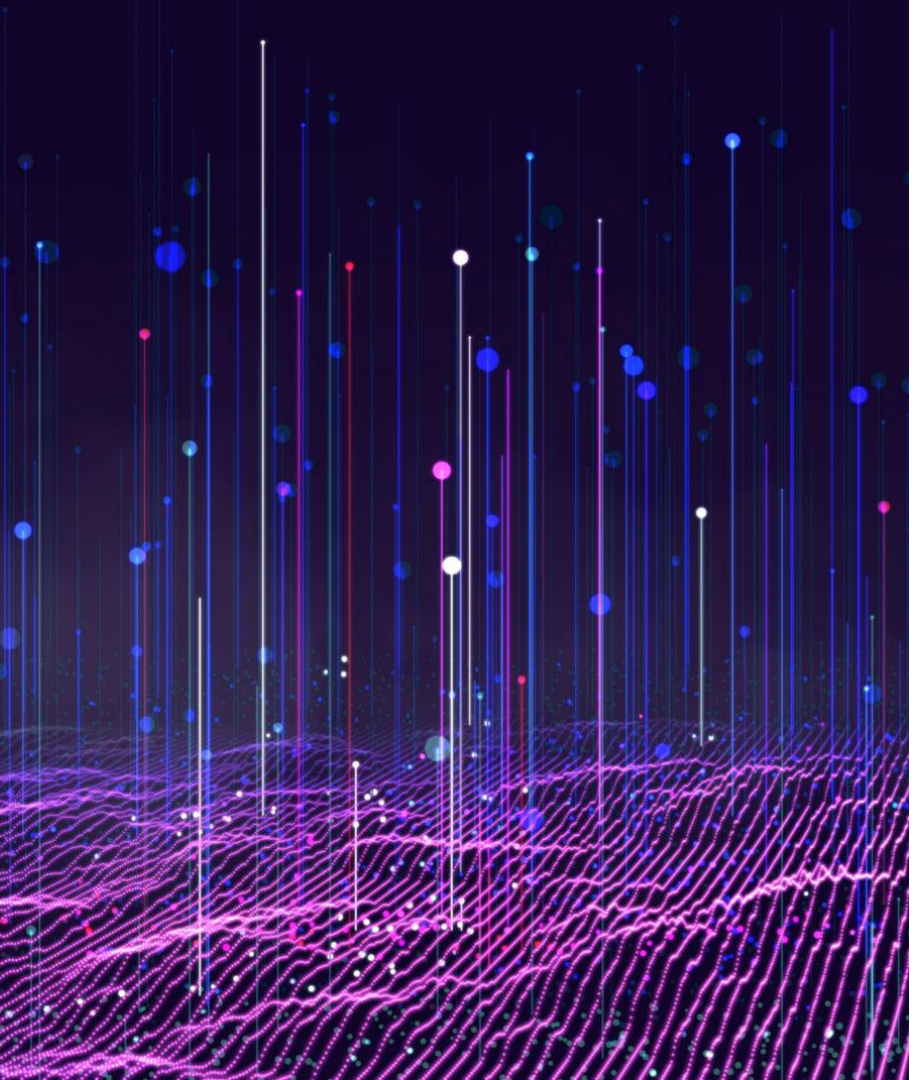


# Distribuciones muestrales y técnicas de evaluación de modelos

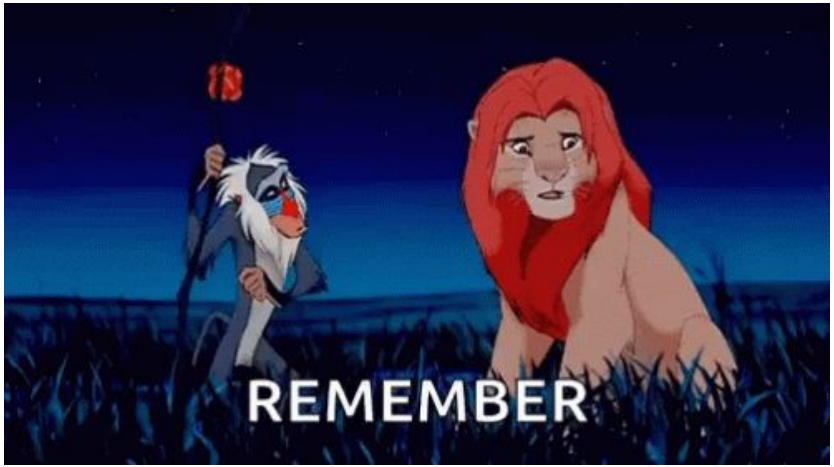
## Análisis de Datos con Python

Eduardo Selim Martínez Mayorga

# Objetivos de la sesión

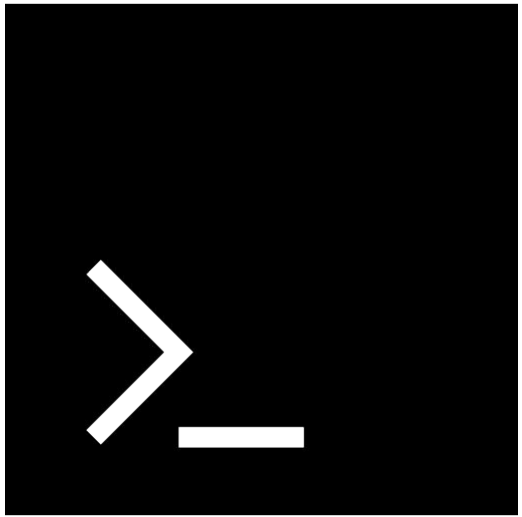
- Distinguir la diferencia entre población y muestra.
- Entender el concepto de 'sesgos' y por qué es tan importante estar conscientes de ellos.
- Aprender el concepto de muestreo aleatorio y cómo puede protegernos parcialmente de los sesgos.
- Utilizar la técnica 'bootstrap' como medio para explorar la distribución muestral de una estadística.
- Crear y utilizar histogramas, errores estándar e intervalos de confianza para evaluar la incertidumbre de una medida estadística.
- Utilizar técnicas para evitar sesgos en el entrenamiento de modelos, como la división de datasets y la validación cruzada.

# Recuerda...




- Correlaciones
- Coeficiente de Pearson
- Matriz de correlación
- Mapas de calor
- Gráficas de dispersión
- Gráficas de pares
- Regresión lineal

# Antes de empezar



**¡No olvides hacer pull del repo!**

**El material de la sesión se encuentra ahí.**

```
git pull origin master
```

# Prework

1. ¿Qué es una población en análisis estadístico?
    a. Es el grupo al cual tenemos acceso a la hora de realizar un experimento
    b. Es el subconjunto de datos que hemos recibido para realizar nuestros análisis
    c. Es el grupo completo acerca del cual se pretende obtener cierta información
    d. Es la cantidad de gente que vive en una región geográfica
    e. Un conjunto de muestras

# Prework

2. ¿Cómo podemos obtener una muestra que no esté sesgada (o que esté sesgada lo menos posible)?
    a. Tomando elementos de una población donde no existan los sesgos
    b. Tomando elementos de nuestra población de manera aleatoria
    c. Dividiendo nuestra población en partes iguales y tomando una de las partes como muestra
    d. Tomando muestras una y otra vez hasta que obtengamos una que no esté sesgada
    e. Seleccionando cuidadosamente los elementos que tomamos para la muestra

# Prework

3. ¿Qué es el error estándar?
    a. Es una medida de la variabilidad de la estadística que estamos analizando
    b. Es el error esperado a la hora de realizar un Boostrap
    c. Es el error típico en las medidas de variabilidad estadística
    d. Es la diferencia entre el tamaño de una población y el tamaño de una muestra
    e. Es la diferencia entre cada elemento del dataset y el promedio

# Prework

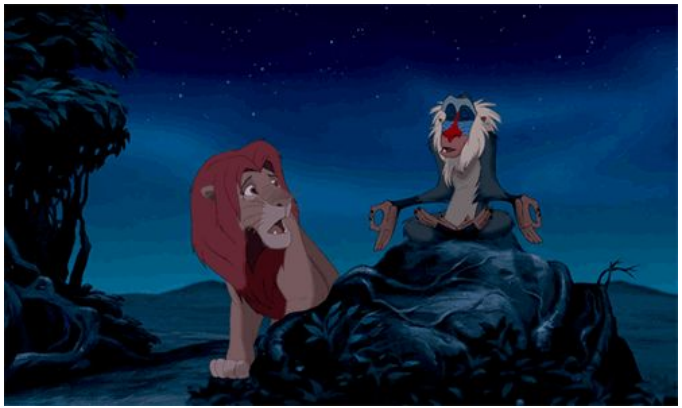
4. ¿Cómo podemos interpretar el valor del error estándar?
   a. Entre mayor el error estándar (y por lo tanto mayor curtosis), menos variabilidad e incertidumbre en nuestra medida estadística
   b. Entre mayor el error estándar (y por lo tanto menor curtosis), menos variabilidad e incertidumbre en nuestra medida estadística
   c. Entre menor el error estándar (y por lo tanto menor curtosis), más variabilidad e incertidumbre en nuestra medida estadística
   d. Entre mayor el error estándar (y por lo tanto mayor curtosis), más variabilidad e incertidumbre en nuestra medida estadística
   e. Entre menor el error estándar (y por lo tanto mayor curtosis), más variabilidad e incertidumbre en nuestra medida estadística

# Prework

5. A la hora de entrenar un modelo de Regresión Lineal, ¿de qué nos sirve dividir nuestro dataset en entrenamiento y prueba?
    a. Es parte del proceso organizado de un científico de datos
    b. Nos ayuda a disminuir el error estándar
    c. Nos permite tener cierta seguridad de que nuestro modelo tenga el mismo desempeño en el mundo real
    d. El modelo de Regresión Lineal sólo puede ser entrenado con el dataset de entrenamiento
    e. Facilita la obtención de una hipótesis válida

# Advertencia

**Esta sesión puede ponerse muy filosófica**



# Población

Es un conjunto de **todos los individuos** que poseen información sobre el fenómeno que se estudia.

**Ejemplo:**

Si se estudia el precio de la vivienda en una ciudad, la población será el total de las viviendas de dicha ciudad.

# Muestra

Es un subconjunto que es seleccionado de una población.

**Ejemplo:**

Si se estudia el precio de la vivienda de una ciudad, lo normal será no recoger información sobre todas las viviendas de la ciudad sino que se suele seleccionar un subgrupo (muestra) que se entienda que es suficientemente representativo.

# Veamos...

1. ¿Cómo definir poblaciones?
2. ¿A qué retos nos enfrentamos cuando definimos poblaciones?
3. ¿Cuáles pueden ser las consecuencias de definir incorrectamente a una población?
4. ¿Cómo generalizamos información de una muestra a una población?
5. ¿Qué problemas pueden surgir en este proceso de generalización?
6. ¿Cómo podemos evitar algunos de estos problemas o aunque sea minimizarlos?

# Sesgos

- Cuando escogemos una muestra, debemos asegurarnos de que nuestra muestra incluye datos representativos.

**Ejemplo:** Estudiar a una población de Científicos en México y tomar 90% hombres y 10% mujeres.

¡Esto es equivocado!

# Otro ejemplo

- El gobierno de Boston hizo una aplicación para detectar baches en las calles.
- Si varios conductores se detenían en una avenida, la App reportaba un bache en esta zona.

**¿Qué sesgo detectas aquí?**

# Veamos...

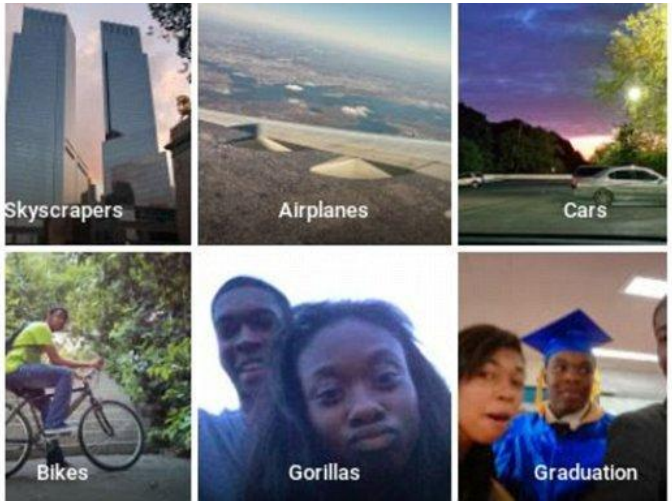
1. ¿De dónde surgen los sesgos en nuestros datos? ¿Cómo llegan ahí?
2. ¿A qué problemas podemos enfrentarnos cuando hay sesgos en nuestros datos?
3. ¿Qué podemos hacer para evitar sesgos en nuestros datos?
4. ¿De qué manera dañan a la sociedad estos sesgos?
5. ¿Qué papel juegan los científicos de datos en la eliminación de estos sesgos?

# Para saber más


**Sesgos en Inteligencia Artificial**

Sofia Trejo

# ¿Qué hacemos entonces?

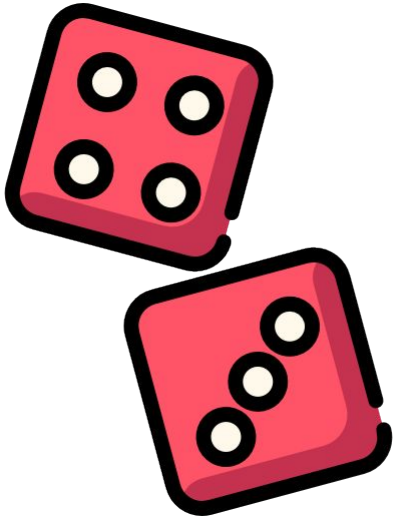



Cuando tenemos un conjunto en nuestras manos, lo más probable es que los datos incluidos en ese conjunto sean una **muestra** de toda la población existente.

En este caso debemos de preguntarnos si las medidas estadísticas que obtengamos son realmente **representativas** de la población. Normalmente es muy difícil (o imposible) regresar al origen de nuestra muestra para extraer más muestras y poder compararlas unas con otras.

Tendremos que “**fingir**”

# Muestreo aleatorio

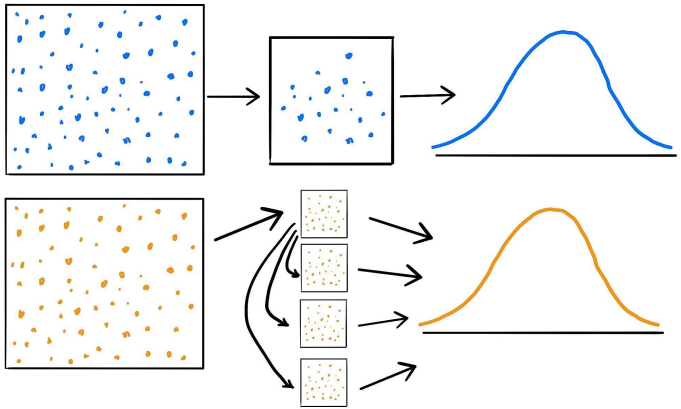


Es un procedimiento que selecciona una muestra cumpliendo dos propiedades fundamentales:

- Todos los individuos de la población tienen la misma probabilidad de ser elegidos.
- Todas las muestras del mismo tamaño son igualmente probables.

**Se recomienda que esto se haga de forma automatizada.**

# Boostrapping

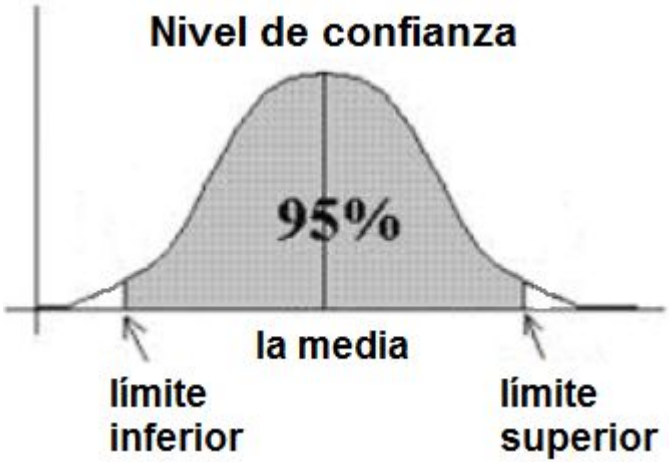


Vayamos al Repo

- Se utiliza para obtener distintas muestras a partir de la muestra que tenemos en nuestras manos y obtener un valor que nos indique qué tanta incertidumbre hay en la medida estadística que hemos realizado.

- Toma un elemento de tu conjunto de datos de manera aleatoria con reposición.
- Repite el paso 1 n veces (entre más cerca esté n a la longitud total de tu muestra, mejor).
- Toma la medida estadística que te interese de tus valores remuestreados.
- Repite los pasos 1 a 3 R veces (entre mayor sea R, mejor).
- Utiliza las medidas obtenidas para: a) Generar un histograma o boxplot b) Calcular el error estándar c) Calcular un intervalo de confianza

# Error estándar / Intervalos de Confianza




Vayamos al Repo

El error estándar nos dice qué tan dispersas están nuestras medidas estadísticas. Esta es una de las maneras de cuantificar incertidumbre. Usa la desviación estándar.

Los intervalos de confianza describen la variabilidad entre la medida obtenida en un estudio y la medida real de la población. Corresponde a un rango de valores, en el cual se encuentra, con alta probabilidad, el valor real de una determinada variable. Esta «alta probabilidad» se ha establecido por consenso en 95%. Así, un intervalo de confianza de 95% nos indica que dentro del rango dado se encuentra el valor real de un parámetro con 95% de certeza.

# Evaluación de modelos

1. ¿De dónde pueden provenir estos sesgos? ¿Cómo llegan a nuestros datos?
2. ¿Qué problemas pueden ocasionar? ¿Qué ejemplos tenemos de esto?
3. ¿Cómo podemos protegernos de este tipo de errores? ¿Es posible eliminar por completo los sesgos en nuestros datos?

**Vayamos al Repo**

# Postwork / Prework

**NO OLVIDES REVISAR TU POSTWORK Y TU PREWORK**

# Preguntas