

Distribuciones muestrales y técnicas de evaluación de modelos

**Análisis de Datos con
Python**

L. en C.C. Manuel Soto Romero



- 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.



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

Pero antes...



Revisaremos el Mini-Proyecto de la Sesión Anterior.

¿Dudas?

¿Problemas?

¿Traumas?



Esta sesión puede ponerse muy filosófica





Es un conjunto de **todos los individuos** que porten 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.



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.



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?



- 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 injusto!

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?

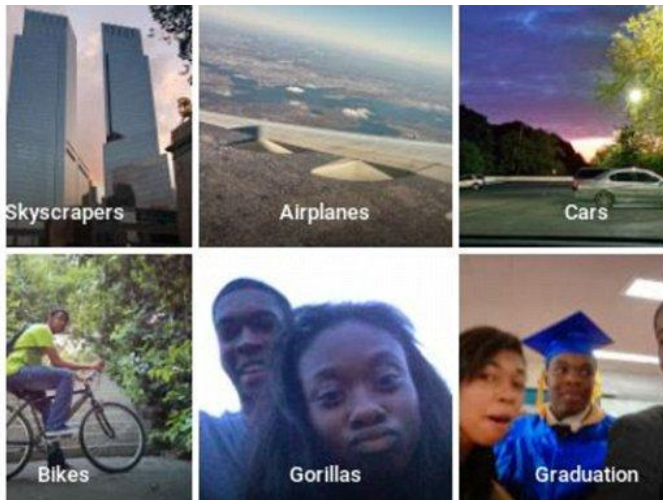


1. ¿De dónde surgen los sesgos en nuestros datos? ¿Cómo llegan ahí?
2. ¿A qué problemas podemos enfrentarnos cuando hay sesgos en nuestros datos?
3. ¿Qué podemos hacer para evitar sesgos en nuestros datos?
4. ¿De qué manera dañan a la sociedad estos sesgos?
5. ¿Qué papel juegan los científicos de datos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



Sofia Trejo

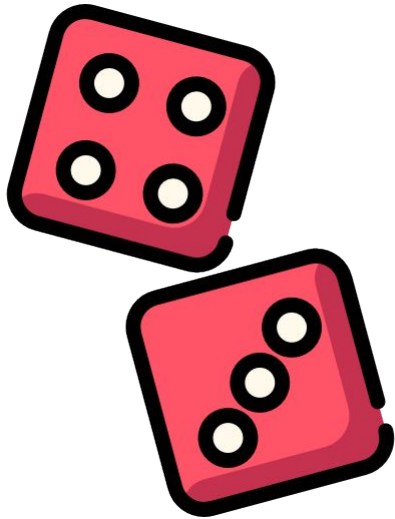
¿Qué hacemos entonces?



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.

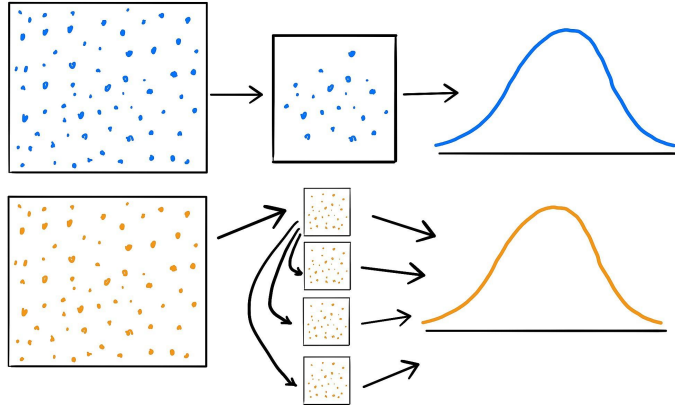
Tendremos que **"fingir"**



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.

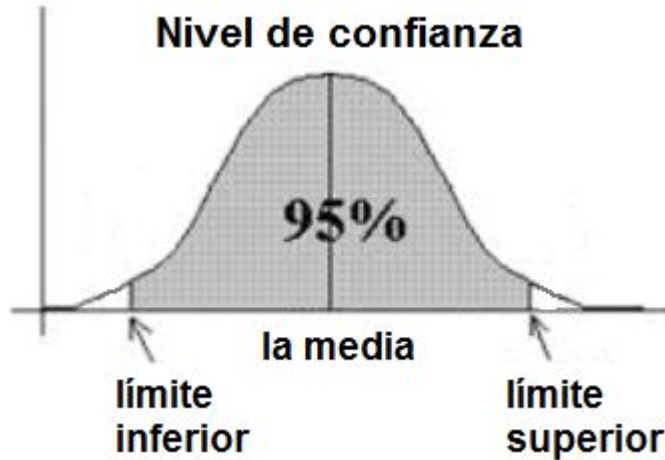
Se recomienda que esto se haga de forma automatizada.



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 **BEU**



Vayamos al Repo

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

Los intervalos de confianza son una manera de mostrar la incertidumbre de una manera muy fácil de comprender, suelen utilizarse en ciencia de datos para establecer qué tan precisa es nuestra medida estadística.



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](#)



NO OLVIDES REVISAR TU
POSTWORK Y TU PREWORK



Preguntas

