Ciencia de Datos

- **Módulo 2**

Estadística
Álgebra relacional

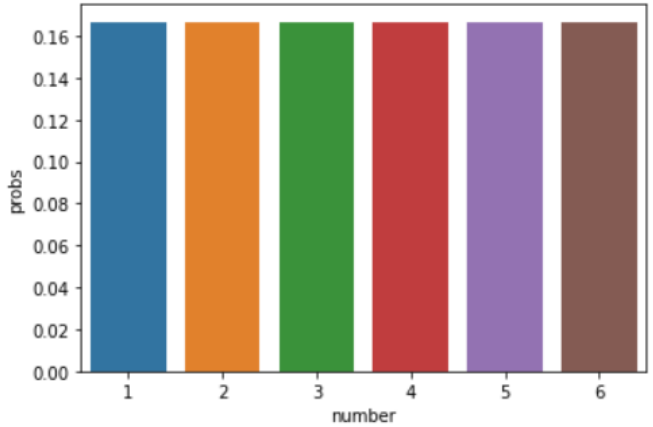
UTNFra
GyEE Gestión y Evaluación Educativa
aprendé programando Virtual
SECUNDARIA del futuro

Estadística

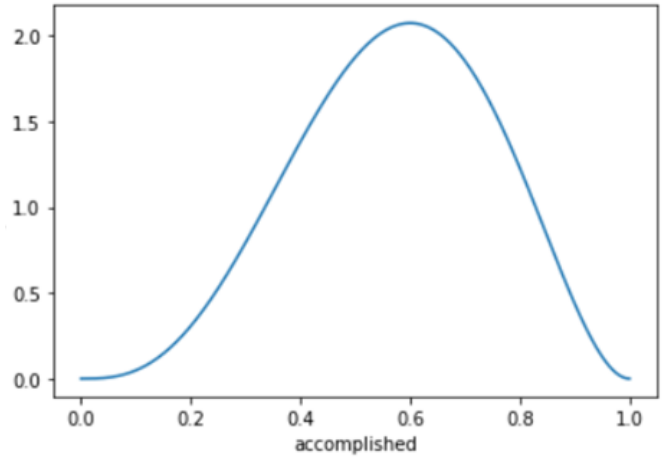
La estadística es la ciencia que se encarga de inferir valores o generar predicciones sobre algún evento del cual tenemos información previa.
Básicamente sirve para obtener información general con solo una muestra de las observaciones, permitiéndonos calcular
la confianza que podemos tener sobre nuestras predicciones.

Estadística - Distribuciones

En estadística todas las variables tienen lo que se llama función de distribución, esto es una función que representa la probabilidad de cada evento posible de la variable, considerando todos los eventos. A estas distribuciones las llamamos poblacionales.



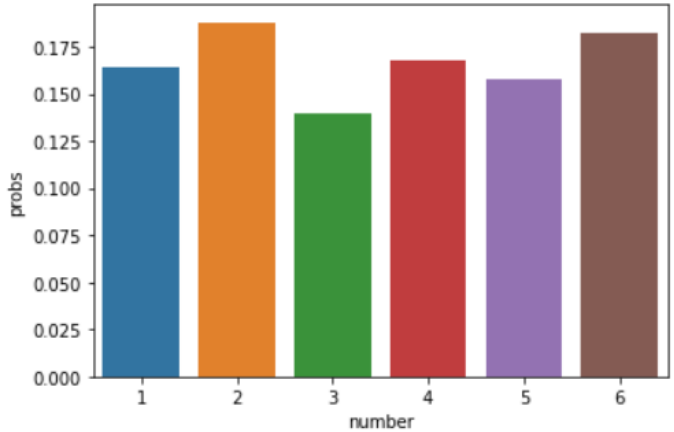
Lanzamientos de un dado de 6 caras



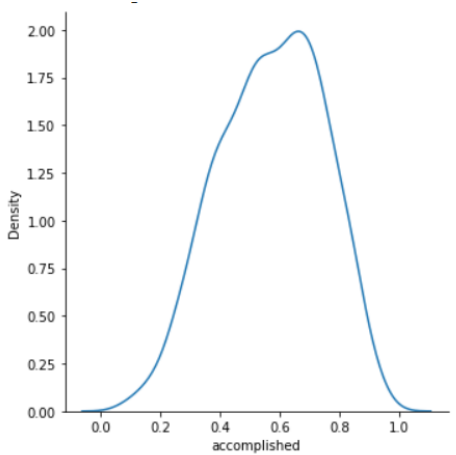
Porcentaje de cumplimiento máximo de un maratón.

Estadística - Distribuciones

Al analizar los datos, dado que trabajamos con un número limitado de eventos, nos encontramos con que las distribuciones no son exactamente iguales a las funciones de distribución. A estas distribuciones las llamamos empíricas. La estadística nos ayuda a descubrir qué tipo de distribución tienen nuestros datos.



Lanzamientos de un dado de 6 caras



Porcentaje de cumplimiento máximo de una maratón en particular.

Estadística - Tipos

Existen dos tipos de estadística:

- Descriptiva: Busca describir los datos que se tienen, analizar distribuciones y correlaciones, etc. Tiene una relación directa con el análisis exploratorio de los datos.
- Inferencial: Conociendo el comportamiento de los datos recolectados, busca obtener información más general sobre el total de los datos (no solo los recolectados). De alguna forma relacionado a la parte del modelado en la ciencia de datos actual.

Estadística - Estadísticos

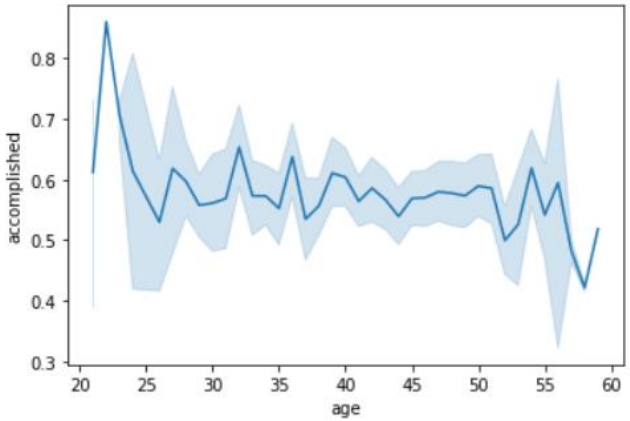
Para realizar inferencias se usan estadísticos, estas son funciones que aproximan parámetros o valores de la distribución poblacional de una variable, por ejemplo, el promedio de los datos es un estadístico para la esperanza, o valor esperado, de la distribución poblacional (El valor esperado de una distribución se puede ver como el promedio de infinitas observaciones).

	Dados	Maratón
Promedio Muestra	3.514	57.23%
Valor esperado	3.5	57.14%

Estadística - Modelos

Los primeros modelos de aprendizaje automático, fueron modelos estadísticos, como lo son las regresiones lineales, logística etc. Estos son modelos que no solo nos ayudan a predecir valores futuros sino que a su vez, los parámetros aprendidos nos dan información sobre nuestras variables.

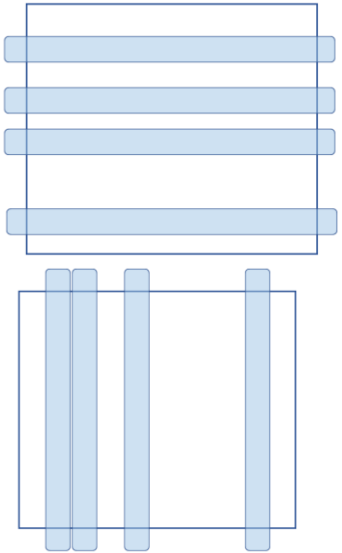
Por ejemplo, si en el estudio de cumplimiento del maratón agregamos la edad y entrenamos una regresión lineal obtenemos cuánto porcentaje más, o menos, cumple un corredor por cada año de vida más que tenga.



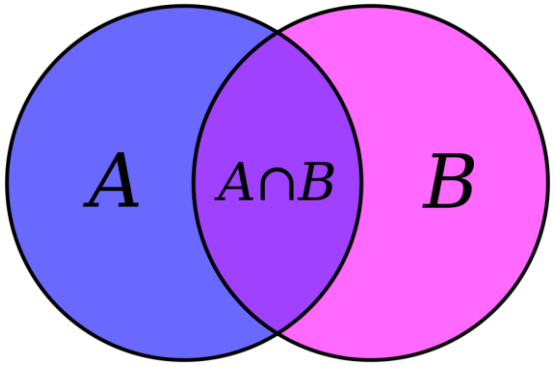
Álgebra Relacional

Las operaciones que hacemos con conjuntos o relaciones es lo que llamamos álgebra relacional, algunas de las operaciones sobre un mismo conjunto son:

- Restricción: Genera un subconjunto del conjunto principal, el cual contiene solo los elementos con alguna propiedad específica, por ejemplo, queremos los pasajeros del titanic que tengan menos de 20 años de edad.
- Proyección: Genera un subconjunto del conjunto principal, con todas las observaciones pero solo un grupo de variables, por ejemplo, solo queremos la información de nombre y edad de los pasajeros del titanic.



Álgebra Relacional



Otras operaciones entre pares de conjuntos son:

- Unión (A∪B): Es el conjunto que posee todos los elementos de los conjuntos A y B.
- Intersección (A∩B): Es el conjunto que posee sólo los elementos de A que también están en B.
- Diferencia (A - B): Es el conjunto de todos los datos de A que no están en B.
- Producto cartesiano (A⨯B): Es un nuevo conjunto el cual combina, todos los elementos de A con cada uno de los elementos de B.