DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440

Clase 5 - Agenda

ESTADÍSTICA Y PANDAS

- Numpy
- Estadística: conceptos
- Pandas

¿ Cómo les fue con numpy ?

¿Dudas ?

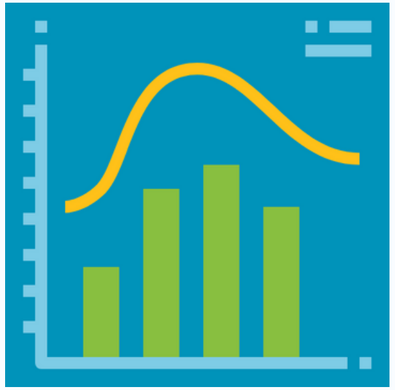
Abrimos notebook "clase 5 - ejercicios numpy.ipynb"

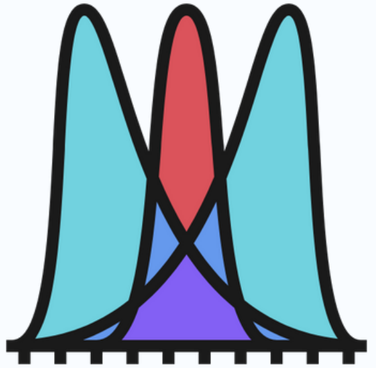
PROBABILIDAD Y ESTADÍSTICA

Probabilidad -> Que espero ver
Estadística -> Lo que vi

ICARO 2021

Variables aleatorias





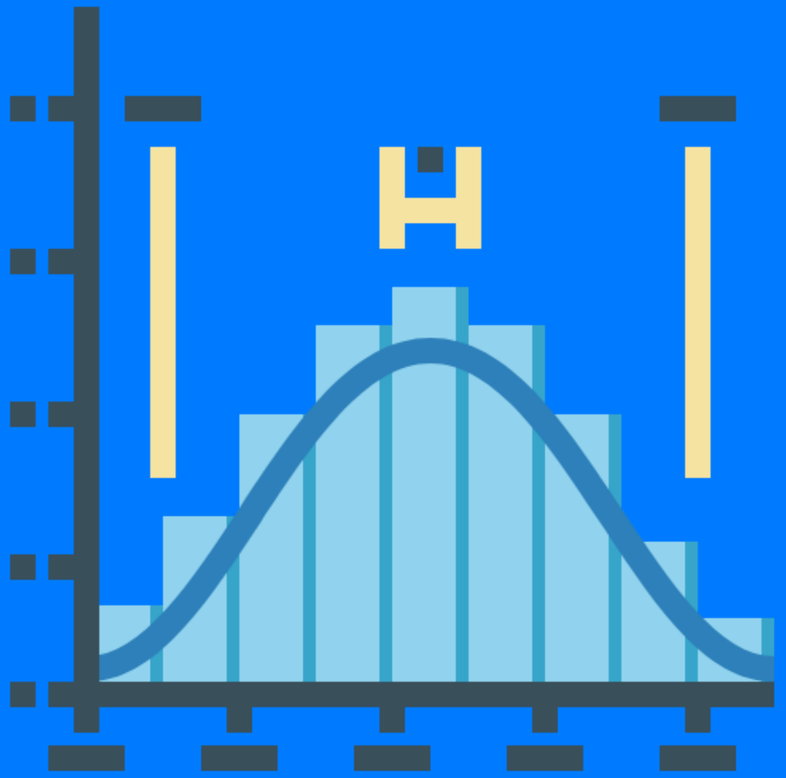
Discretas

No puede tomar ningún valor entre dos consecutivos

Continuas

Puede tomar cualquier valor dentro de un intervalo.

Ejemplos?



Variables discretas

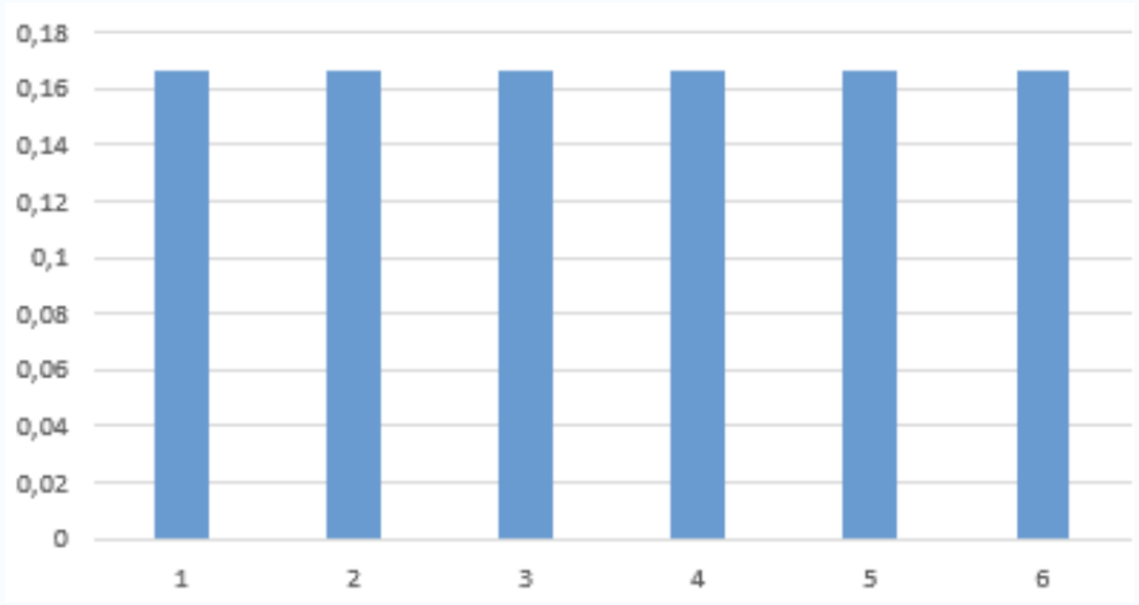
Distribución

La distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra.

Distribución uniforme

La distribución de probabilidad uniforme asigna la misma probabilidad para todo un rango de valores.

Ejemplos?



Xdado: {1,2,3,4,5,6}

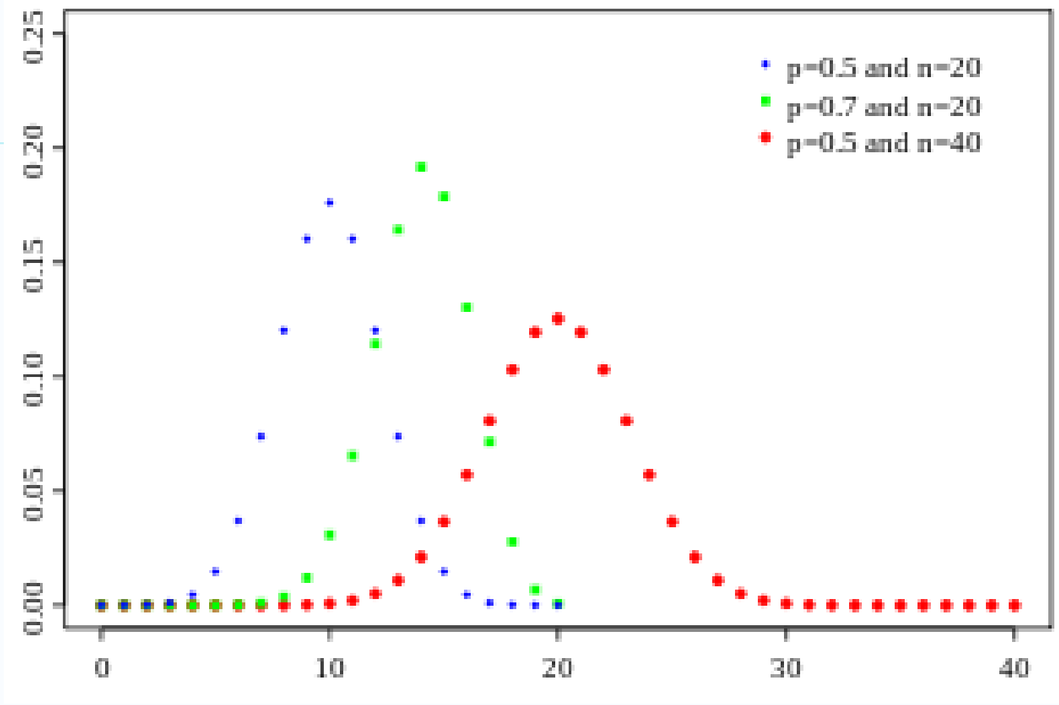
P(X = 1,2,3,4,5,6) = 1/6

Distribución binomial

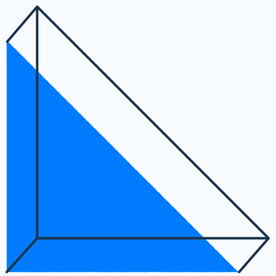
Una distribución binomial es una distribución de probabilidad discreta que describe el número de éxitos al realizar n experimentos independientes entre sí, acerca de una variable aleatoria.

Sirve para responder preguntas del tipo:

Si tiro n veces una moneda con probabilidad p de sacar cara (o ceca), ¿cuál es la probabilidad de sacar x caras (o cecas)?



Distribución Binomial



Resultados posibles

En cada ensayo, experimento o prueba solo son posibles dos resultados (éxito o fracaso).

Probabilidad constante

La probabilidad del éxito ha de ser constante. Esta se representa mediante la letra p. La probabilidad de que salga cara al lanzar una moneda es 0,5 y esta es constante.
La probabilidad de fracaso ha de ser también constate. Esta se representa mediante la letra q = 1-p.

Experimentos independientes

El resultado obtenido en cada experimento es independiente del anterior. Por lo tanto, lo que ocurra en cada experimento no afecta a los siguientes.

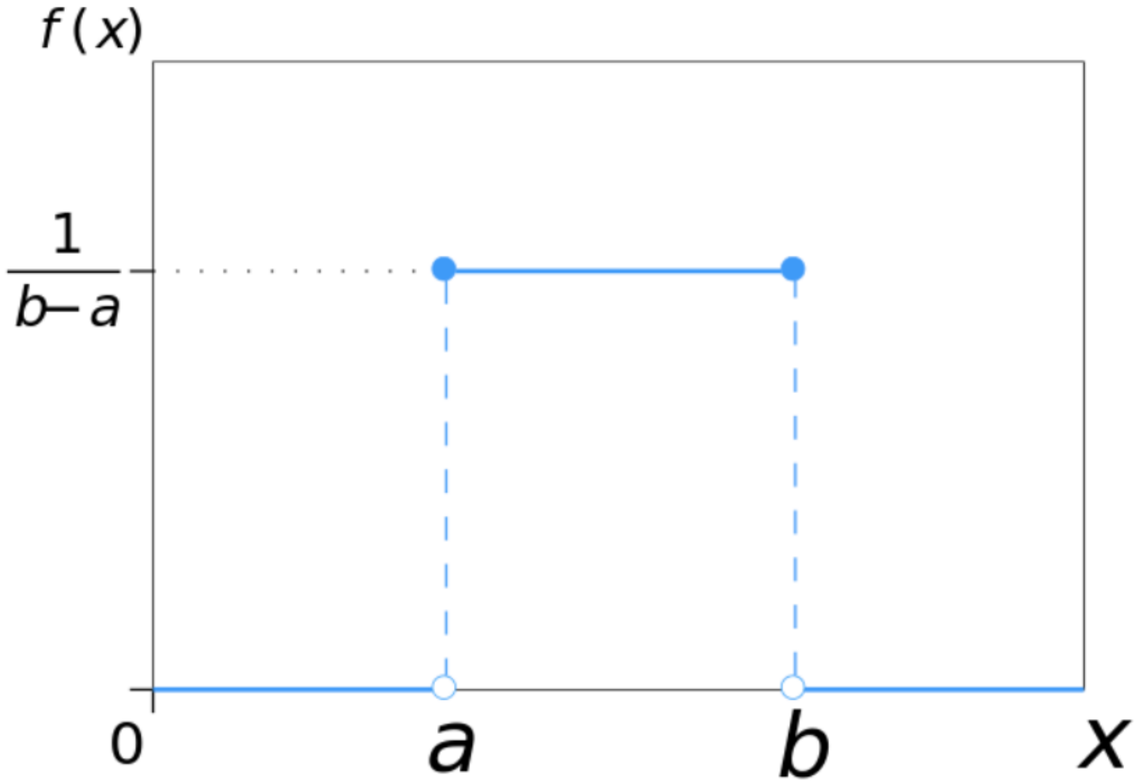
Experimentos mutuamente excluyentes

Los sucesos son mutuamente excluyentes, es decir, no pueden ocurrir los 2 al mismo tiempo.

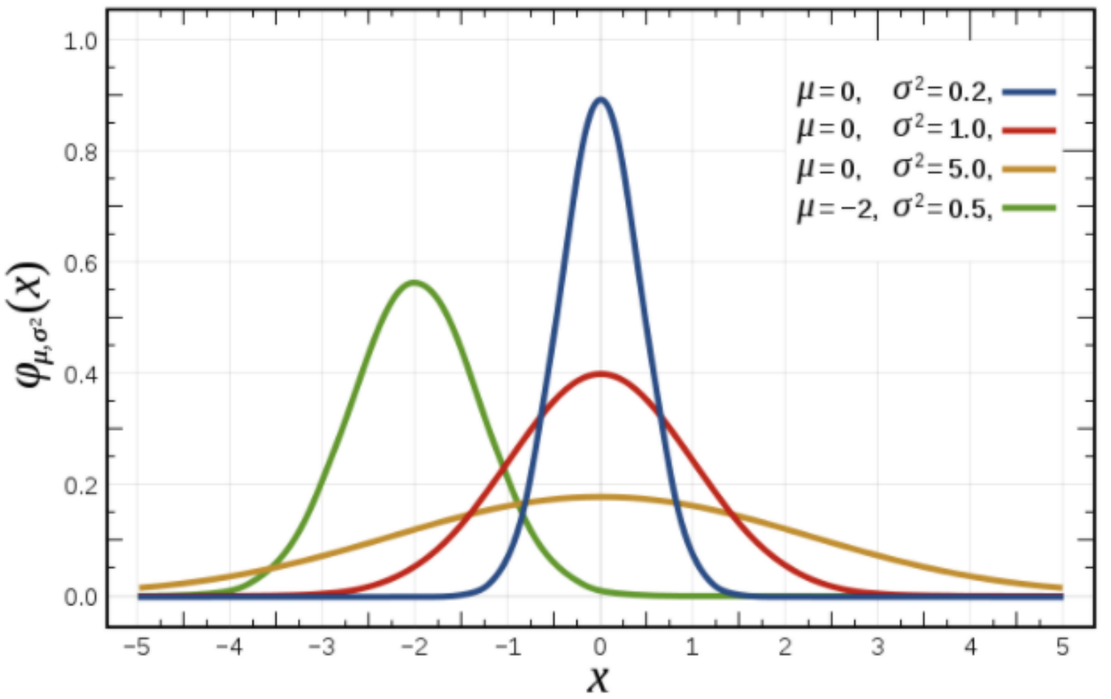
Variables continuas

Para variables continuas utilizamos el concepto de **densidad de probabilidad**

Densidad uniforme



Densidad gaussiana o normal



La función de densidad de una distribución normal tiene forma de campana. Es simétrica en torno a la media.
El área total bajo la curva es 1 (como corresponde a una función de densidad).

Medidas de resumen

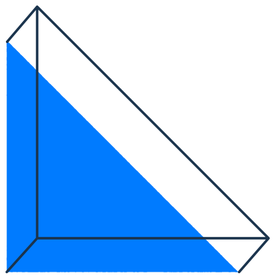
Por lo general en estadística se trabaja con grandes volúmenes de datos y para poder entender el comportamiento de ellos y encontrar patrones es necesario sintetizar esta información.

Las medidas de resumen, resumen en una sola cifra toda la información contenida en una variable.

Se dividen entre grupos:

- Medidas de tendencia central
- Medidas de dispersión
- Medidas de posición

Medidas de tendencia central



Buscan representar donde se encuentran centrados los datos. Son:

- La media (μ): Sumamos todos los datos y luego dividimos por la cantidad de observaciones.
- La mediana: Es un valor tal que bajo ella se encuentran el 50% de las observaciones.
- La moda: es la observación que más se repite.

El principal problema es:
¿ Cuándo conviene usar cada una ?

La media es sensible a valores extremos.

Cuando hay presencia de datos extremos se recomienda utilizar la mediana como medida de tendencia central pues esta será más representativa que la media.

Medidas de dispersión

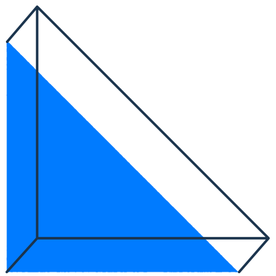
Las medias más usadas de dispersión son:

- El rango: Diferencia entre el valor máximo y mínimo.
- La varianza (σ^2): Mide que tan dispersos están los datos en torno a la media.

$$\sigma^2 = \frac{\sum_1^N (x_i - \bar{X})^2}{N}$$

- La desviación estándar (σ): Es la raíz cuadrada de la varianza.

En general se trabaja más con la desviación estándar que con la varianza.
La desviación estándar se hace para poder trabajar en las unidades de medida iniciales.

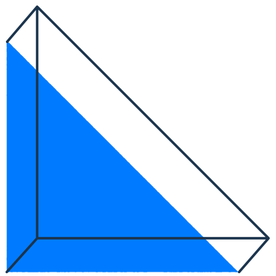


Medidas de dispersión

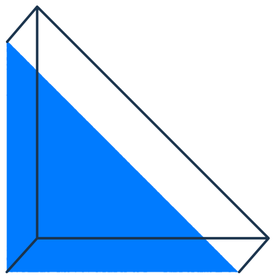
La regla empírica (dist normal)

Existe una regla llamada Regla Empírica que señala:

- Aproximadamente el 68% de las datos estén entre la media y una desviación estándar
- Aproximadamente el 95% de las datos estén entre la media y dos desviaciones estándar
- Aproximadamente el 100% de las datos estén entre la media y tres desviaciones estándar



Medidas de posición



Percentiles

Los percentiles dividen un conjunto de observaciones ordenadas de menor a mayor en 100 partes iguales.

Así hasta el primer percentil, P10, hay un 10% de las observaciones, hasta el segundo percentil, P20, hay un 20% de las observaciones ya así sucesivamente.

Otra manera que se usa con frecuencia para dividir un conjunto de observaciones son los cuartiles: Se dividen las observaciones en cuatro grupos tal que cada uno de ellos contenga el 25% de las observaciones.

Abrimos notebook "clase 5-pandas.ipynb"

Preguntas?

Para la próxima clase:

- Terminar notebook de pandas si no lo hicieron