



Artículo redactado por [Carlos M. Mazzaroli](#) por diversión y demostrar los conocimientos adquiridos en estadística probabilística

# Matemáticas para Ciencias de Datos: Estadística Probabilística

En este artículo buscamos entender por qué la probabilidad es tan importante en Ciencias de datos y el Machine Learning en general.

El orden del mismo artículo y material de estudio fue dado por [Platzi](#) en un curso impartido por el profesor [Francisco Camacho](#). También se enlazaron materiales de apoyo desde YouTube u otras páginas webs.

¡Espero que disfruten de este artículo para aprender juntos!

## Tema 1: Incertidumbre y probabilidad

### ¿Qué es la probabilidad?

En que situaciones necesitamos usar la probabilidad? para eso iremos al concepto básico que nos induce a esto. Por ello nos haremos esta pregunta:

**Que es la probabilidad?**

**La probabilidad es la herramienta a la que recurrimos cuando hay incertidumbre.**

La incertidumbre surge a la hora de tomar decisiones cuando tenemos información incompleta. Los juegos de azar son el ejemplo perfecto, ya que no podemos predecir el resultado en un juego de cartas o dados en un casino.

Esto se debe a situaciones que tienen un grado de complejidad donde no es posible tener todas las variables suficiente, con todos los datos para predecir de que si lanzas los dados de cierta manera o juegas las cartas de cierta manera se terminara por dar un resultado.

Podemos resumir lo anterior como **la incertidumbre es la toma de decisiones con información incompleta**.

"El azar no es más que la medida de nuestra ignorancia. Los fenómenos fortuitos son, por definición, aquellos cuyas leyes o causas simplemente ignoramos"

**Henri Poincaré**

Por el hecho de que vivimos en una realidad donde la gran mayoría de nuestras decisiones las tomamos con información incompleta, los matemáticos desarrollaron un esquema para cuantificar esta incertidumbre, dando así el área de la Probabilidad en estadística.

**Probabilidad**

En matemáticos decimos que la probabilidad es el lenguaje y conjunto de herramientas matemáticas que nos permite cuantificar la incertidumbre.

Así entendemos la propia palabra *probabilidad* como el área de investigación, el concepto matemático puntual de la probabilidad tiene algunas sutilezas que conducen a confusiones.

### Axiomas de la probabilidad

Entender los elementos esenciales de la probabilidad

Todo conjunto lógico tiene que estar basado a un conjunto de axiomas, que significa que es un conjunto de sentencias o afirmaciones, que no son derivables de algo más fundamental, es decir que no requiere demostración, por lo que lo asumimos como verdad.

**Sucesos**

Para comenzar definiremos los sucesos, ya que en cualquier libro de referencia, cualquier curso, o idioma donde estudies las matemáticas de la probabilidad, encontrarás que la definición más simple es que la probabilidad ( $P$ ), es la división de dos cantidades, números de sucesos exitosos sobre número de sucesos totales.

$$P = \frac{N^{\circ} \text{ sucesos exitosos}}{N^{\circ} \text{ sucesos totales}} \text{ - creencia del total}$$

Ejemplo, cuando lanzas un dado, el resultado son 6 posibilidades, cada una de esas posibilidades, es un suceso. Los sucesos totales son 6, pero cuando queremos saber cuáles son las probabilidades de que el dado caiga en 2, como el 2 es un suceso de los seis, decimos que la probabilidad es de un sexto ( $\frac{1}{6}$ ).

Pero esto tiene una sutileza que nos conduce a dos escuelas de pensamiento en estadística, la escuela Frecuentista y escuela Bayesiana.

### Por qué se divide el pensamiento estadístico?

Cuando definimos que un sexto es la probabilidad de que un dado caiga en 2, aunque no parezca evidente, estamos asumiendo que todas las caras son igualmente probables. Por lo tanto asumimos que todas las caras son igualmente probables.

Lo mismo podemos definir con una moneda, donde tenemos dos posibles sucesos (cara o cruz), donde asumimos que la probabilidad de que caiga cara es de un medio ( $\frac{1}{2}$ ) y que caiga cruz también es de un medio.

Pero que tan cierto es esto?

Dependiendo de como interpretemos que es un suceso, o el número de sucesos exitosos, podríamos hacer un ejercicio donde tiramos una moneda al aire 10 veces, y si la probabilidad es como la entendemos hasta ahora, de las 10 probabilidades, 5 deberían caer en cara y 5 en cruz.

Haz el ejercicio y anota si realmente de las 10 veces te cayó 5 veces en cara o cruz.

Probablemente no fue así verdad, no?

Aca es cuando diferenciamos las escuelas Frecuentistas y Bayesianas.

## Escuela Frecuentista

Este pensamiento de probabilidad, nos explica que estos números que llamamos probabilidad (en esta ocasión puede ser la cara y la cruz), son números que solo se alcanzan a la medida que haces infinitos lanzamientos de la moneda o dado, la proporción del número de lanzamientos exitosos y el número de lanzamientos totales tiende a un medio, se acerca cada vez más a 0.5. Como no hay forma de demostrar esto, por eso denominamos esto como un axioma.

Las probabilidades sobre estos posibles sucesos que llamamos elementales, es porque son las ocurrencias más básicas sobre un suceso probabilístico, donde una moneda solo tendría dos opciones, un dado seis opciones. Por esto debemos diferenciar sobre lo que es un suceso elemental y un suceso

### Suceso elemental:

Podemos entender a un suceso elemental como:

"El resultado de lanzar un dado es 4"

Donde decimos que es elemental porque el resultado 4 solo se puede dar de una manera, que es que el dado caiga en 4 y nada más.

Suceso

Un suceso en general se percibe como:

"El suceso de lanzar un dado es par"

No es elemental porque es la unión de varios sucesos elementales, donde el resultado puede ser 2, 4 o 6.

El uso de sucesos y sucesos elementales, se encuentran en el espacio muestral ( $EM$ ), que es el conjunto de todos los posibles resultados de un evento aleatorio. El dado tendría un espacio de seis, ya que tiene seis caras en la que puede caer, a cada uno de estos elementos del espacio muestral es lo que llamamos los sucesos elementales.

En probabilidad entendemos que todo evento aleatorio, viene escrito en un espacio muestral donde cada elemento son todas las posibles ocurrencias de ese evento aleatorio probabilístico, donde cada uno de los elementos le asignamos un número que es una propiedad intrínseca. En el ejemplo de los dados le darianos el elemento  $1/6$ , ya que cada cara es igualmente probable y a esto lo asumimos como un axioma.



Este tipo de probabilidad fundamental, sobre cada suceso elemental, de un espacio muestral, definimos a un sexto como la probabilidad de que caiga cada cara, a esto lo definimos como un axioma.

En la vida real tendríamos que hacer infinitos intentos de esta situación en particular, al ser un escenario abstracto, lo asumimos cierto dentro de un esquema axiomático, es decir, la probabilidad hace parte de los axiomas y de esas propiedades intrínsecas de un problema aleatorio.

La probabilidad que se le asigna a cada posible ocurrencia de un sistema aleatorio, posee varias propiedades que deben cumplirse para que el esquema axiomático tenga sentido:

- $0 < P < 1$  Tienen que ser números que vayan del 0% al 100%, donde 0 es igual a 0% y 1 es igual a 100%.
- *certeza*  $\rightarrow P = 1$  Un elemento totalmente cierto lo llamamos con un 1.
- *imposibilidad*  $\rightarrow p = 0$  Un elemento imposiblemente cierto lo llamamos con un 0.
- *disjuntos*  $\rightarrow P(A \cup B) = P(A) + P(B)$  La probabilidad de dos elementos disjuntos sucedan, es la suma de las probabilidades de cada uno de estos.

Ejemplo:

La probabilidad de que al arrojar un dado caiga en 2 y en 4 es la suma de la probabilidad de ambos sucesos, ya que no puede caer en 2 y en 4 al mismo tiempo

Por lo que decimos que la probabilidad de que el dado caiga en 2 o en 4 es de dos sextos ( $\frac{2}{6}$ ), ya que son dos eventos posibles dentro de los seis eventos posibles.

## Que es realmente la probabilidad?

Para concluir debemos determinar que es realmente la probabilidad.

Muchos dicen que es la creencia que tenemos sobre sucesos elementales. Ya que desde la perspectiva frecuentista, no tenemos forma de determinarlos realmente, así que la probabilidad se incluye como un axioma en un conjunto de reglas que nos permite cuantificar la incertidumbre.

# Probabilidad en Machine Learning

En este módulo entenderemos como el Machine Learning y Ciencias de dato en general, como, el concepto o herramientas de probabilidad

Recordemos que la probabilidad es un conjunto de herramientas y lenguaje matemático que nos permite cuantificar la incertidumbre.

Pero... donde se encuentra la incertidumbre en el machine learning?

La fuente de la incertidumbre se encuentra en:

- Los datos.
- Atributos del modelo.
- Arquitectura del modelo.

Recuerda que en la vida real, recolectar y hacer la medición de los datos es un proceso imperfecto, ya que todos los instrumentos de medición tienen un margen de error, que ya nos introduce parte de incertidumbre en los datos.

En Machine learning hablamos que un modelo se alimenta de atributos, o variables predictoras, estas variables con frecuencia son subconjuntos reducidos del problema total y real, lo que hace que esta reducción de variables ya es otra capa de incertidumbre.

En matemáticas un modelo se entiende como una representación simplificada de la realidad, y al ser una representación simplificada, ya induce otra capa más de incertidumbre.

Estas tres son las principales fuentes de incertidumbre dentro del ML (Machine Learning), y por supuesto, estas fuentes de incertidumbre, son cuantificables con probabilidad

## Modelo de Clasificación

Ejemplo, un clasificador de documento de texto, donde tenemos un conjunto de documentos de texto de distintas categorías, y nuestro modelo tiene que leer e identificar cual es el tema de conversación y ubicarlos en una categoría correspondiente de cada documento.



Entonces, el modelo asignara cierta probabilidad a cada documento y así de determinara la clasificación de los documentos.

A este modelo es lo que llamamos un clasificador probabilístico, y por definición del uso del propio modelo, ya damos uso a la probabilidad para identificar cual es la categoría más probable

### Funcionamiento interno del modelo de Clasificador Probabilístico



En nuestro caso, lo que hace el modelo es que tiene documentos, donde cada uno tiene etiquetas, la etiqueta sería la categoría o tema al que se refiere.

### Fase de Entrenamiento



Los documentos tienen una función que extra los atributos, es decir, simplifica los elementos fundamentales del documento que me ayudaran a hacer la extracción de la categoría, a esto es lo que llamamos el **Extractor de Atributos**.

Una vez realizada la extracción, el documento fue reducido a un conjunto de atributos (Vector), que se pasara como input al algoritmo de Machine Learning de clasificación. El cual sería un algoritmo de Machine Learning supervisado, porque le doy las etiquetas, donde el algoritmo leería los atributos y en base a esto, asignaría una etiqueta.

Así es como entrenaríamos a un algoritmo de ML supervisado

### fase de Predicción

Luego de la fase de entrenamiento, viene la fase de Predicción



En el proceso donde el algoritmo aprendió a unir los atributos con las etiquetas correctas, ya podemos agarrar el modelo, entregarles documentos, en el cual usando el mismo proceso de extracción de atributos, para luego entrar en el modelo de clasificación y predecir la etiqueta.

Entonces podríamos decir que tengo una tarea donde tengo que clasificar muchos documentos, pero como físicamente no se puedo por el tamaño y número de documentos, usaría mi algoritmo que está bien entrenado y el modelo me diría si el texto está hablando sobre política, deportes, etc

*La mayoría de modelos de clasificación funcionan con este esquema en general*

## Todas las etapas del modelo probabilístico



Todas las etapas de un modelo, en ciertos aspectos involucra probabilidad.

¿Pero como?

## Entrenamiento

En la parte de entrenamiento, antes de pasar los documentos para extraer atributos e identificar la arquitectura, debemos escoger el modelo a usar

- Diseño

El modelo a usar es lo que definimos por diseño, donde escogeremos si el modelo usaría probabilidad o no (No todos los modelos se apoyaran de la matemática probabilística, ya que no todos los modelos son probabilísticos)

En este caso usaremos de ejemplo el modelo de Naive Bayes, ya que es un modelo probabilístico, ya que el clasificador Naïve-Bayes aprende de los datos de entrenamiento y luego predice la clase de la instancia de prueba con la mayor probabilidad posterior.

- Entrenamiento

Una vez elegido el diseño, tenemos que definir el entrenamiento.

El entrenamiento básicamente es que el modelo aprenda algo que sabrá mas adelante, que es el concepto de distribución de probabilidad.

Esto es una manera de saber que probabilidad asignarle a cada una de las posibles ocurrencias de los datos, donde nos encontramos con el esquema MLE (Maximum Likelihood Estimation)

### ¿Qué es un parámetro de un modelo?

En los modelos de aprendizaje automático, los parámetros son las variables que se estiman durante el proceso de entrenamiento con los conjuntos de datos. Por lo que sus valores no los indica manualmente el científico de datos, sino que son obtenidos.

### MLE

En estadística, la estimación de máxima verosimilitud (MLE) es un método para estimar los parámetros de una distribución de probabilidad supuesta, dados algunos datos observados. Esto se logra maximizando una función de verosimilitud para que, bajo el modelo estadístico asumido, los datos observados sean los más probables.

- Calibración

Luego llegaría la calibración, que es el ajuste del modelo a través de los hiperparámetros, donde iremos calibrando el modelo para que el error del modelo sea cada vez más pequeño.

Los hiperparámetros se encuentran por fuera del esquema de optimización, donde a veces se lo denomina como tuneo o calibración de hiperparámetros, donde hay veces que se usan algoritmos de optimización bayesiana

### ¿Qué es un hiper parámetro?

Son valores que generalmente no puedo configurar con la optimización del modelo, por lo que suelen ser indicados por el científico de datos. El valor óptimo de un hiper parámetro no se puede conocer a priori para un problema dado. Por lo que se tiene que utilizar valores genéricos, reglas genéricas, los valores que han funcionado anteriormente en problemas similares o buscar la mejor opción mediante prueba y error. Siendo una buena opción buscar los hiper parámetros la validación cruzada.

## Predicción

Luego del proceso de entrenamiento, viene el proceso de predicción, donde nos encontramos con la fase de interpretación del modelo

- interpretación de la predicción

Independientemente de si el modelo es probabilístico o no, para la correcta interpretación del modelo, la persona requiere de ciertos conceptos de probabilidad, ya que entender cómo funciona el cálculo de probabilidad del modelo, nos permite tener una interpretación correcta del mismo

# Tema 2: Fundamentos de probabilidad

## Tipos de probabilidad

Profundizaremos sobre el concepto de probabilidad mismo, hablando de los distintos tipos de probabilidad

Distintos tipos de situaciones tienen que cuantificar con conceptos adicionales sobre el concepto de probabilidad básico

### Tipos de probabilidades

- Conjunta (Joint)
- Marginal
- Condicional  $P(A|B)$

Para explicar estos conceptos que suelen darse de manera muy abstracta, serán explicados con un juego de dos dados, estudiando el espacio muestral, donde la malla será la matriz por la que las filas y columnas representan el estado del primer dado y del segundo, teniendo un espacio muestral de 36

combinaciones



Identifiquemos las siguientes probabilidades y sus interpretaciones formulando la siguiente pregunta

## Probabilidad Conjunta

Formula General:

$$P(A \cap B)$$

¿Cuál es la probabilidad de que ambos dados caigan en número par?

Esta pregunta se puede responder de forma sencilla, considerando, que el espacio muestral de los 2 dados y sus 36 posibilidades solo tenemos 9 sucesos exitosos que cumplan con el.



Entonces decimos que los estados o sucesos exitosos son 9 posibilidades de 36, por lo tanto la probabilidad quedaría como  $\frac{9}{36}$  y simplificado  $\frac{1}{4}$ .

- $P(\text{par}, \text{par}) = \frac{9}{36} = \frac{1}{4}$

Esta probabilidad que corresponde a un suceso como tal, en realidad sería la unión de dos sucesos, es decir que el dado A haya caído en par y el dado B también haya caído en par, por lo tanto corresponde a dos sucesos separados, que es lo que llamamos una probabilidad conjunta (joint)

- $\underbrace{P(\text{par} * A, \text{par} * B)}_{\text{Conjunta (joint)}} = \frac{9}{36} = \frac{1}{4}$

Una probabilidad conjunta es una probabilidad de dos o más sucesos, que calculamos haciendo un conteo directo al espacio muestral

## Probabilidad Condicional

$$P(A|B)$$

¿Cuál es la probabilidad es de que un dado caiga en par, sabiendo que el dado B ya cayó en par?

Como ves, esta pregunta es ligeramente distinta a la pregunta anterior, ya que supone una condición previa que restringe el uso enteró del espacio muestral.

Ahora solo tenemos que considerar las situaciones donde ya sabemos que B es par



La parte de la pregunta que nos dice "el dado B ya cayó en par", lo que hace es restringir el espacio muestral, teniendo antes 36 posibilidades a tener ahora solo 18 posibilidades.

Entonces ahora que reducimos el espacio muestral, decimos ¿cuál es la probabilidad de que a caiga en par, sabiendo que b ya es par?.

$$\underbrace{P(A = \text{par}|B = \text{par})}_{\text{condicional}}$$

Como esto impone una condición previa a este tipo de probabilidades con la barrita vertical ("!" o "tal que"), la definimos como Probabilidad Condicional.

Habiendo ya restringido el espacio muestral, volveremos a realizar un conteo pero solo teniendo en cuenta el espacio restringido



El número de sucesos exitosos no cambio, sino solo el número de sucesos posibles por lo que el resultado con el espacio muestral reducido quedará tal que

$$\underbrace{P(A = \text{par}|B = \text{par})}_{\text{condicional}} = \frac{9}{18}$$

Así vemos que tenemos dos probabilidades distintas, para dos preguntas distintas. Pero entonces nos preguntamos...

¿Cómo están relacionadas estas probabilidades?

Resulta que podemos formular la pregunta de cuál es la probabilidad de que B caiga en par, esto sería una probabilidad tradicional, ya que no ponemos ninguna condición, donde disponemos del espacio muestral completo teniendo las 36 opciones disponibles, pero... ¿cuántos de éstos sucesos corresponden al estado par?



Entonces de la misma manera 36 opciones, donde tenemos 3 columnas con 6 sucesos que cumplen la premisa de la pregunta, dándonos 18 sobre 36

$$\underbrace{P(B = \text{par})}_{\text{EM completo}} = \frac{18}{36}$$

Ahora vemos que tenemos 3 probabilidades que fueron los resultados de 3 preguntas diferentes, pero podemos decir que **la probabilidad conjunta del suceso A y B, es igual a la probabilidad condicional de A dado B, por, la probabilidad de B**

$$\underbrace{P(A, B)}_{\text{conjunta}} = \underbrace{P(A|B)}_{\text{condicional}} * \underbrace{P(B)}_{\text{prob de B}}$$

Esto no es un caso particular del ejemplo dado, sino, que es una fórmula general, la **Regla del Producto**.

$$\underbrace{P(A, B)}_{\text{Regla del producto}} = P(A|B) * P(B)$$

## Probabilidad Marginal

Es cuando se obtiene una probabilidad sencilla a partir de una probabilidad conjunta. Es decir cuando se tiene las probabilidades conjuntas de 2 sucesos y se quiere saber solo la probabilidad de que suceda el primer suceso independiente de lo que pasa con el otro, así eso se define como **la suma de todas la probabilidades conjuntas sobre los demás estados que no está considerando A**.

$$\underbrace{P(A)}_{\text{Marginal}} = \sum p(A, B)$$

## Conclusión

Por medio del juego de los dados logramos definir de forma natural tres tipos de probabilidades

Primero definimos la **Probabilidad Conjunta** que como vimos, es una probabilidad que considera la ocurrencia de diferentes, pero simultáneos eventos aleatorios.

Luego esta se relaciona con la **Probabilidad Condicional** por medio de la Regla del Producto. Es importante aclarar que la probabilidad condicional **NO IMPLICA CAUSALIDAD**, es decir, que la probabilidad de que suceda A dado que sucedió B, no quiere decir que B sea la causa de A. Puede que en situaciones particulares ocurra, pero son dos conceptos diferentes.

Con esto terminamos por decir que la **Probabilidad Marginal**, se obtienen haciendo sumas sobre ciertas variables aleatorias o ciertos sucesos sobre ciertas variables aleatorias dentro de la probabilidad conjunta. Siempre que hagamos sumas de probabilidades conjuntas y dejemos libre una de las variables, decimos que estamos obteniendo la probabilidad marginal de dicho evento aleatorio.

## Ejemplos de cálculo de probabilidad

### Correlación de eventos

Con un par de ejemplos sencillos, buscaremos ganar más intuición sobre el uso de la probabilidad, como calcularla y cómo debemos interpretarlas. Aprenderemos cómo la correlación de eventos nos permite descubrir y aplicar asociaciones lógicas entre distintos eventos.

#### Ejemplo 1: Juego de los Dados

Para este ejemplo, tendremos en cuenta 3 eventos aleatorios:

- $A = \{\text{El resultado de lanzar un dado es } 4\}$
- $B = \{\text{El resultado de lanzar un dado es par}\}$
- $C = \{\text{El resultado de lanzar un dado es impar}\}$

Podríamos preguntarnos sobre la probabilidades de cada uno de estos eventos, sin condicionarlos a la ocurrencia previa de otro evento.

Veamos la diferencia entre la probabilidad condicionada entre uno u otros sucesos, y las probabilidades sin condicionar para ver qué conceptos surgen.

---

Dados nuestros tres sucesos, consideraremos nuestras probabilidades de una manera sencilla.

Veamos en primer lugar una probabilidad tradicional.

Dónde queremos saber cual es la probabilidad de que suceda A, sin ninguna condición adicional.

$$A = \{\text{El resultado de lanzar un dado es } 4\}$$

Al ser un dado, sabemos que las posibilidades son 6, y que caiga en 4 es solo una de ellas, por lo tanto decimos que la probabilidad de A, es un sexto:

$$P(A) = \frac{1}{6} \rightarrow 16.6\%$$

#### Correlación Positiva

Que sucede, si ahora nos preguntamos, ? cual es la probabilidad de que suceda A, sabiendo que ya ha sucedido B?

$$A = \{\text{El resultado de lanzar un dado es } 4\}$$

$$B = \{\text{El resultado de lanzar un dado es par}\}$$

$$P(A|B) = ?$$

Gramaticalmente esto lo traduciríamos a la vida real como el hecho de que lance una vez el dado, y cayó en un número par, provocando que nuestro espacio muestral se haya reducido, así que el número de posibilidades ya no es 6, sino 3, y de esas 3, solo una posibilidad corresponde al evento exitoso.

$$P(A|B) = \frac{1}{3} \rightarrow 33.3\%$$

Por lo tanto decimos esta probabilidad condicional de que suceda A dado B, es mayor que la probabilidad tradicional de que suceda A, nos dice, que el hecho de que haya ocurrido B, aumenta la probabilidad de que ocurra A. Entonces es cuando decimos que los eventos A y B están **positivamente correlacionados**.

Recordemos el concepto de Correlación de manera sencilla. La correlación es lo que nos dice como dos variables, eventos, sucesos o activos se relacionan el uno al otro. ([Profundizar conceptos de Correlación. Tema 3, Capítulo VI](#))

### Correlación Negativa

Por otro lado también podemos preguntarnos sobre cuál es la probabilidad de que suceda A sabiendo que ya sucedió C.

$$A = \{\text{El resultado de lanzar un dado es } 4\}$$

$$C = \{\text{El resultado de lanzar un dado es impar}\}$$

$$P(A|C) = ?$$

Entonces, si ya sabemos que C sucedió, y el resultado es algún número impar, vemos que el espacio muestral se vio limitado a estas opciones: {1, 3, 5}, y resulta que A solo puede ser el número {4}.

Como no hay una intersección entre el {1, 3, 5} y el elemento {4}, esto representa sucesos excluyentes, por lo tanto quedamos con un intersección o conjunto vacío.

$$\{1, 3, 5\} \cap \{4\} = \emptyset$$

Por lo tanto decimos que la condición C reduce el espacio muestral, diciéndonos que los sucesos posibles son 3, pero los sucesos exitosos son 0, dándonos una probabilidad de 0. Entonces decimos que la ocurrencia de C, redujo la ocurrencia de A, por lo tanto decimos que los eventos A y C están **negativamente correlacionados**

Que la probabilidad nos de 0, osea que los elementos sean Excluyente, no quiere decir que los elementos sean independientes, sino, son altamente dependientes

*Excluyente ≠ Independiente*

### Conclusión del Juego de los dados

Con este sencillo ejercicio evidenciamos como cuando dos eventos pueden estar tanto positivamente y negativamente correlacionados. Y concluimos en:

- Correlación Positiva es cuando la ocurrencia de un evento, **aumenta** la probabilidad de suceso de otro evento correlacionados
- Correlación negativa es cuando la ocurrencia de un suceso **disminuye** la probabilidad de ocurrencia del otro.
- Dos elementos excluyentes, no son independientes, por lo contrario, tienen correlación negativa.

### Ejemplo 2: Juego de ruleta

Para este ejemplo, usaremos el conocido juego de la ruleta que se pueden ver en muchos casinos, donde tendremos a dos jugadores, que cada uno apostará a cuatro números dentro de las ocho posibilidades dentro del rango del espacio muestral.



El conjunto del *jugador, 1* lo llamaremos como conjunto A y al del *jugador, 2* conjunto B, donde diferenciaremos los conjuntos de números que apostaron por el color rojo y azul.



Para facilitar el entendimiento del ejercicio, pasaremos a llamar los conjuntos de los jugadores a conjunto A y B.

$$\text{Jugador, 1} \longrightarrow A = 1, 2, 3, 4$$

$$\text{Jugador, 2} \longrightarrow B = 5, 6, 7, 8$$

Aclarado esto, comenzamos el ejercicio preguntándonos:

**¿Cuál es la probabilidad de que gane el jugador 1?**

En este caso, tanto los sucesos del *jugador, 1* y el *jugador, 2* son excluyentes, ya que sus apuestas son diferentes, por lo tanto no tienen un conjunto de intersecciones.

Al analizar la probabilidad de que gane el *jugador, 1*, está dada por las 8 casillas de la ruleta y de esas 8 posibilidades, solo 4 harán que gane el *jugador, 1*.

Quedando tal que:

KaTeX parse error: Expected 'EOF', got '%' at position 22: ... 4/8 = 1/2 = 50%

Pero... ¿Qué pasaría si ahora agregamos una condición?

**¿Cuál es la probabilidad de que gane el jugador 1, sabiendo que el resultado de la ruleta se encuentra dentro del conjunto B?**

$$P(A|B) = ?$$

Vemos que la condición de restringir el espacio muestral completo, al conjunto B, redujo el número de eventos exitosos, teniendo como el número de eventos exitosos igual a 0, ya que al no existir un punto de intersección entre el conjunto A y B, no queda más que un conjunto vacío.

$$P(A|B) = \frac{0}{4} = 0$$

Entonces evidenciamos un ejercicio de eventos excluyentes.

#### ¿Qué sucedería si presentamos una situación ligeramente diferente?

Donde el *Jugador*, 2 cambie su apuesta al conjunto de números 4,5,6,7 y el *Jugador*, 1 mantiene su conjunto del principio.

Quedando los conjuntos tal que:

$$\text{Jugador, 1} = 1, 2, 3, 4 \longrightarrow A = 1, 2, 3, 4$$

$$\text{Jugador, 2} = 4, 5, 6, 7 \longrightarrow B = 4, 5, 6, 7$$

Recuerda que en general lo que elija el *Jugador*, 1 y el *Jugador*, 2, no tienen porque tener relación alguna entre sus apuestas, ya que cada uno está apostando a las posibilidades, donde cada uno eligió números a su propio criterio.

En este caso logramos ver que si ocurre una intersección entre el conjunto del *Jugador*, 1 y el nuevo conjunto del *Jugador*, 2.



Entonces nos volvemos a preguntar...

#### ¿Cuál es la probabilidad de que gane el jugador 1, sabiendo que el resultado de la ruleta se encuentra dentro del conjunto B?

Al igual que la vez anterior, **la condición B restringe el espacio muestral** a 4 de las 8 posibilidades, y la intersección entre el conjunto A y B solo ocurre en una posibilidad, por lo tanto la probabilidad de A dado B quedaría tal que:

$$\text{KaTeX parse error: Expected 'EOF', got '%' at position 18: } \dots A|B) = 1/4 = 25\%$$

#### ¿Qué buscamos demostrar con este ejemplo?

Lo que queremos decir, es que gracias al conocimiento previo de saber que el *Jugador*, 2 haya ganado, la probabilidad de que el *Jugador*, 1 también ganará, es del 25%.

A diferencia de antes, cuando las probabilidades de que el *Jugador*, 1 ganó, sabiendo que el *Jugador*, 2 ganó, eran del 0%.

#### Conclusiones del Juego de la Ruleta

- La ocurrencia del conocimiento previo de que el jugador 2 haya ganado, lo que provocó fue que se reduzca la probabilidad de que el jugador 1 haya ganado. Por lo tanto sabemos que la coincidencia de los eventos A y B representan eventos que se encuentran negativamente correlacionados.

#### Reto para practicar

Sabemos que el jugador 1 mantiene los números que eligió al principio y jugador 2 cambio los suyos por el 2, 3, 6 y 7.

#### ¿Cuál es la probabilidad de que gane el jugador 1, sabiendo que el jugador 2 gano?

$$\text{jugador 1} = 1, 2, 3, 4$$

$$\text{jugador 2} = 2, 3, 6, 7$$

## Ejemplos avanzados con probabilidad

Continuaremos desarrollando ejemplos para

### Paradoja ¿niño o niña?

1. Una mujer tiene dos bebés donde el mayor es un varón.
2. Una mujer tiene dos bebés donde uno de ellos es varón.

Parecen parecidos pero no, en probabilidades cambia

Tablero formulamos la siguiente pregunta

Cual es la probabilidad de esta mujer tenga dos hijos varones.

El fin del ejercicio es darse cuenta que la información de cada enunciado es diferente.

Y para el cálculo de probabilidades de un ejerc q parece tan sencillo, primero deberemos calcular el espacio muestral, donde dibujaremos una matriz, donde en un eje se encontrarán los posibles géneros de un hijo, y en el otro eje los géneros del otro hijo.

$M$	$FM$	$MM$
$F$	$FF$	$MF$
$F$		$M$
$\underbrace{\hspace{1cm}}$ Espacio Muestral		

Para dar un ejemplo, daremos que sin conocimiento previo nos preguntamos:

### ¿Cuál es la posibilidad de que una mujer tenga 2 hijos varones?

Tal cual lo presentamos sin ninguna condición, planteamos una probabilidad tradicional, y consideramos el espacio muestral completo, donde el número de eventos posibles es 4, y el número de eventos exitosos es 1.

$M$	$ $	$FM$	$(MM)$
$F$	$ $	$FF$	$MF$
$F$	$ $	$M$	

Y la probabilidad quedaría tal que:

$$P(MM) = \frac{1}{4} = 25\%$$

Por lo tanto, la probabilidad es solamente de  $\frac{1}{4}$  sin ninguna condición previa, donde esta probabilidad no representa ni al caso 1 ni 2 mencionados anteriormente, sino simplemente a una situación general donde no tenemos información previa al género de los hijos de dicha mujer.

### Situación 1

Pero qué sucedería si ahora imponemos la situación donde tenemos la información previa de que el mayor de los hijos es un varón. Entonces replantearíamos el ejercicio ahora de esta forma:

#### ¿Cuál es la probabilidad de que ambos hijos sean varones, sabiendo que el mayor es varón?

$$P(MM | \text{Mayor Varón}) = ?$$

Con la información que tenemos podemos restringir el espacio muestral sabiendo que uno de los ejes (hijos) es el mayor, restringiendo el espacio muestral a solamente dos estados.

$M$	$ $	$FM$	$ $	$MM$
$F$	$ $	$FF$	$ $	$MF$
$F$	$ $	$M$		

Y de esta manera apreciamos que entre estos dos estados, solo uno satisface el enunciado. Quedando la probabilidad tal que:

$$P(MM | \text{Mayor M}) = \frac{1}{2} = 50\%$$

Este resultado funciona bien con la situación 1, pero ¿Qué diferencia existe entre la situación 1 y 2?

### Situación 2

Comparemos la sutileza gramatical entre la situación 1 y 2

1. Una mujer tiene dos bebés donde el mayor es un varón.
2. Una mujer tiene dos bebés donde uno de ellos es varón.

La diferencia de decir entre el hijo mayor es varón, y uno de ellos es varón, implica que en realidad no sabemos cuál de ellos es varón.

Aunque escape de la intuición, este pequeño cambio genera una diferencia en el espacio muestral, y lo demostraremos en la matriz para que se pueda apreciar el cambio del enunciado 2.

$M$	$[FM]$	$[MM]$
$F$	$FF$	$MF$
$F$	$M$	

El hecho de decir que, uno de ellos es el varón, representan 3 posibles estados en el espacio muestral, porque en cada uno de los ejes, al menos uno de los hijos es varón.

Cuando escribimos esto en la probabilidad, decimos, que la probabilidad de que ambos hijos sean varones, sabiendo que alguno de ellos es varón es de un estado exitoso sobre tres posibles estados:

$$P(MM | \text{alguno M}) = \frac{1}{3} = 33.\bar{3}\%$$

Así podemos demostrar que las posibilidades de la situación 1 y 2 son diferentes, aunque parezcan igual y se debe a que la cantidad de información que contiene cada frase debido a esa sutileza gramatical es diferente y esto determina distintos resultados en probabilidad.

Donde en la situación 1 la probabilidad de que los dos hijos sean varones es mayor, debido a la mayor cantidad de información dada en el enunciado.

## El problema de Monty Hall



En nuestra segunda paradoja trataremos el caso del programa de televisión *Let's make a deal*, dado por el conductor Monty Hall, al que de se debe su nombre en probabilidad a esta paradoja, cómo, **El problema de Monty Hall**

El ejercicio consistía en que el presentador le presentaba a un participante tres puertas, donde el participante tenía que elegir una entre las tres posibles puertas, donde detrás de dos puertas no había nada y solo en una había un premio.

En una situación tradicional, el presentador le preguntaría al participante que elija una puerta, entonces nos preguntaríamos

#### ¿Cuál es la probabilidad de que el participante elija la puerta correcta?

Siendo una probabilidad tradicional, dibujaremos el espacio muestral y veríamos todas las opciones, donde podremos ver que las tres opciones inicialmente, todas son igualmente probables, por lo tanto el participante piense que la probabilidad de elegir la puerta correcta sea de un tercio por la naturaleza de la situación.

$P_1$	$P_2$	$P_3$
0	0	1
0	1	0
1	0	0

$\longrightarrow 1/3 \rightarrow 33.3\%$

El truco del show era de que una vez que el participante eligiera una puerta, Monty Hall abre una de las puertas (obviamente el presentador abriría una puerta sin recompensa, este era el punto de información adicional), luego de el presentador abriera la puerta sin recompensa, Monty le preguntaría al participante, que ahora cuenta de información adicional (la puerta que abrió Monty):

Ahora que sabe que esta puerta no tenía recompensa, ¿mantiene la puerta que eligió, o prefiere cambiar de puerta?

Entonces, la intuición a primera vista nos hace decir:

¿Cuál es la probabilidad de que cambie de puerta y gane? o ¿Cuál es la probabilidad de que mantenga la puerta y gane?

Dibujemos el espacio muestral para representar esta idea de forma más visual:

Este sería muestral antes de elegir una puerta

$P_1$	$P_2$	$P_3$
0	0	1
0	1	0
1	0	0

Entonces digamos que elegimos la puerta 1, y el presentador abrió la puerta 3 ya que esta no tiene premio. El espacio muestral se nos restringiría a las situaciones donde la puerta 3 no tiene premios, quedando tal que

$P_1$	$P_2$	$P_3$
0	1	0
1	0	0

$\longrightarrow 1/2 = 50\%$

Donde el número de estados posibles ya no es 3, sino 2, quedando la probabilidad de éxito de un medio.

Entonces el razonamiento es que al tener 2 opciones donde solo 1 es de éxito y ambas tienen 50% de posibilidades de ser la correcta, el participante asume que da igual si mantiene la puerta o no, ya que en ambas tiene 50% de probabilidad de ganar. Este es el primer razonamiento que se podría hacer con la intuición natural de las probabilidades, pero la paradoja es que esto es **falso**.

¿Por qué cambian las probabilidades?

Pero... ¿en realidad hay más probabilidades de ganar si cambio la puerta una vez que el presentador haya descartado una?

La respuesta es sí! Porque resultan en situaciones distintas, al igual que en el ejemplo anterior, las probabilidades pueden verse modificadas cuando hay un cambio en la cantidad de información disponible a la hora de tomar una decisión, y este cambio de información fue el hecho de que el presentador haya abierto la puerta, modificando la probabilidad de éxito.

Cambiemos el esquema mediante el cual ahora calcularemos nuestras nuevas probabilidades.



En este nuevo diagrama consideraremos dos columnas nuevas, donde representará el caso de si mantenemos la misma puerta que elegimos, o el caso en que cambiemos la puerta sabiendo la información adicional que nos dé el presentador.

## Situación 1

Entonces supongamos que elegimos la puerta 1 y el presentador abriera la puerta 2, ya que la 3 tiene el premio, por lo tanto no la abriría.



¿Qué sucedería en esta situación?

En esta situación, si me mantuviera en la puerta 1, que es la que elegimos al principio, no ganaríamos el premio, pero si la cambiamos al a puerta 3, entonces sí lo ganaríamos. Quedando el diagrama tal que:



## Situación 2

Sigamos usando el mismo razonamiento con el siguientes caso, donde abrimos la puerta 1, y el presentador tendrá que abrir la puerta 3, ya que en la puerta 2 se encuentra el premio.



Donde si me quedo con la puerta que elegí perdería, pero si la cambiara, ganaría.

## Situación 3

Y en la ultima situación, seria que abriera la puerta 1 que contiene el premio, y daria igual la puerta que abriera el presentador, ya que ni la puerta 2 y 3 tienen premio, siendo el único caso donde si mantengo la puerta ganaría.



## Resolución

Estas son las situaciones que tendríamos que tener en cuenta para saber ¿Cual es la probabilidad de ganar si me quedo con la misma puerta?

Para saber esto veamos la matriz completa de todas las situaciones que vimos en el ejercicio.



Si nos hubiéramos quedado con la misma puerta que elegimos al principio, solo en un caso hubiéramos tenido la probabilidad de ganar de los tres casos, quedando tal que:



Pero a diferencia de si hubiéramos cambiado de puerta luego de la información adicional dada por el presentador, tendríamos dos casos de eventos exitosos contra los tres que teníamos.



## Conclusión

Con estos dos ejercicios demostramos que el cálculo de probabilidades, no siempre es intuitivo y que hay tener cuidado al entender cual es el espacio muestral sobre cual estamos trabajando, dado que tengamos información adicional, o no, sobre cierta situación a la cual realizaremos el cálculo de probabilidades.

Con estos dos ejercicios dimos el inicio para desarrollar nuestra intuición probabilística!

[Video explicativo Monty Hall - Javie Santaolla](#)

# Tema 3: Distribuciones de probabilidad

## ¿Qué es una distribución?

¿Qué es una distribución de probabilidad?

Es una función, en el sentido matemático del cálculo, donde a cada uno de los posibles estados de una variable aleatoria dentro del espacio muestral, se le asigna una probabilidad.

Como ejemplo tenemos el ejercicio del dado que tiene un espacio muestral de 6 posibles estados, y cada uno de esos estados tiene una probabilidad de  $\frac{1}{6}$ , en este caso esta distribución sería una función constante, donde a cada estado se le asigna un valor, siendo la misma, una función discreta.

En general mencionaremos a la  $X$  mayúscula como una variable aleatoria, donde  $P$ , será la función, que a cada una de las ocurrencias o valores posibles de esta variable aleatoria, se le asignará un número que denominaremos la probabilidad.

$$X \text{ aleatoria} \longrightarrow \underbrace{P(X = x)}_{\text{probabilidad de ocurrencia}}$$

De esta manera comprendemos que  $P$  es función de la variable aleatoria

$$P = f(X)$$

Una convención en probabilidad, es que, las **letras mayúsculas denotan las variables**, mientras que las **letras minúsculas denotar los posibles valores que estas variables aleatorias pueden tomar**.

$X \rightarrow$  variable aleatoria

$x \rightarrow$  valores posibles en el espacio muestral

Al igual que sucede en el cálculo, las funciones poseen un Dominio.

El dominio viene por **todos los valores posibles de la variable aleatoria por la cual la función puede ser calculada**, donde estos dominios podrán así dividirse en conjuntos discretos o continuos, donde tendremos tanto **funciones discretas o funciones continuas**.

$$Dom(X) = \begin{cases} \text{Discreto}, & \{1, 2, 3, 4, 5, 6\} \\ \text{Continuo}, & [0, \infty] \end{cases}$$

#### [Artículo dedicado a Funciones Matemáticas para Ciencias de Datos](#)

Un ejemplo de una distribución discreta podría usarse de ejemplo el juego de los dados, porque los valores tienen un número finito de estados, donde tenemos a las 6 caras del dado, y la variable aleatoria sería la cara que me daría el dado como resultado.

A diferencia de las variables aleatorias que pueden ser continua, por ejemplo, la temperatura, ya que puede considerarse como una variable aleatoria y es continua, porque no precisa necesariamente tener que ser un número entero, sino, que puede ser un valor decimal cualquiera dentro de un rango definido.

Profundizaremos sobre los aspectos matemáticos de estas funciones particulares, que llamamos distribuciones de probabilidad:

## Distribución de Probabilidad

Coincidiremos a  $X$  mayúscula como una variable aleatoria donde  $P$  de  $X$ , será una función de distribución de probabilidad, o también conocida como **densidad de probabilidad**

$X \rightarrow P(X) \rightarrow$  densidad de probabilidad

Donde  $P(X)$  puede tener un carácter **discreto** o un carácter **continuo**, que estarán determinados por los valores posibles de la variable aleatoria  $X$

$$P(X) \begin{cases} \text{Discreto} \\ \text{Continuo} \end{cases}$$

Como toda función, se puede graficar, así que cuando una distribución de probabilidad es continua, podemos describirlo como una distribución gaussiana en un plano de ejes cartesianos, donde dado un punto  $x$  ( $x$  minúscula = ocurrencia de un valor específico dentro del conjunto de variables), la imagen, dada la función, es la probabilidad de que ocurra ese valor particular



Y como toda función en cálculo, la podemos derivar o integrar.

## Integral de una distribución

¿Qué significa la integral de una distribución?

Al igual que podemos preguntarnos ¿Cuál es la probabilidad de que la variable tenga un valor en particular? que sería haciendo esto con la función de densidad de probabilidad.

$$P(X = x) = ?$$

También podemos preguntarnos

¿Cuál es la probabilidad de que mi variable aleatoria tenga valores menores o iguales que un valor específico dado?

$$P(X \leq x) = ?$$

Para calcular esto, debemos recordar los conceptos de cálculo integral, donde serían todos los valores que se encuentre por detrás del valor umbral, y esto es lo que llamamos en cálculo, un área bajo la curva



Así es como sabemos que este tipo de probabilidades ( $P(X \leq x)$ ) están dadas por una integral en función de la distribución.

Donde decimos que la probabilidad de que mi variable aleatoria tome valores menores o iguales que un cierto valor específico, está dado por la integral de mi

distribución de probabilidad  $P$  de  $X$ , respecto a la variable de  $X$ , integrando sobre todos los posibles valores que sean menores o igual que  $x$ .

$$P(X \leq x) = \int_{-\infty}^x P(X) dX$$

Esto es una integral, es el área debajo de la curva y también representa una probabilidad.

En general decimos **cuando  $x$  minúsculo no es un valor numérico**, sino, un valor cualquiera dado que puede considerarse como un parámetro, esto determinará una nueva función que llamamos **la Distribución Acumulada  $C(X)$**

$$P(X \leq \underbrace{x}_{* \text{parametro}}) = \int *X \leq x P(X) dX$$

### Función de distribución acumulada

Entonces decimos que **la distribución acumulada**, representa la probabilidad de que mi variable aleatoria tome valores menores o iguales que esta  $x$  dada, lo que llamamos una función de probabilidad acumulada.

$$P(X \leq \underbrace{x}_{* \text{parametro}}) = \int *X \leq x P(X) dX = C(x) \leftarrow \text{Función, Probabilidad, Acumulada}$$

**La interpretación de la función de probabilidad acumulada, es la integral de la función de densidad de probabilidad.**

Y sirve para responder el tipo de preguntas que surgen cuando no nos preguntamos por un valor en particular de la probabilidad, sino, **cuando nos referimos a un rango dentro de la probabilidad**, por ejemplo:

¿Cuál es la probabilidad de que al tirar un dado el resultado sea un número menor o igual a  $x$ ?

Por ejemplo, cuál es la probabilidad de que al tirar un dado el resultado sea un número menor o igual al 2, *donde este ejemplo específicamente se trata de una función discreta*.

**La función de probabilidad acumulada también se utiliza para funciones discretas**, solo que en este caso, la gráfica en el plano cartesiano ya no se vería como una curva suave, sino, como un histograma.

Donde cada una de sus caras tendrá una probabilidad, que sería la frecuencia con la que ocurriría cada uno de estos eventos.



Cuando queremos calcular la probabilidad acumulada de una función discreta, porque queremos responder la misma pregunta sobre ¿Cuál es la probabilidad de que mi variable aleatoria tomará valores menores o iguales de cierto valor? ya no se usarán integrales, sino, sumas discretas, donde sumó todas las probabilidades en las que mi variable aleatoria tenga los valores menores o iguales al parámetro de referencia.

$$P(X \leq x) = \sum_{-\infty}^x P(X) \rightarrow \text{Función, Probabilidad, Acumulada}$$

Esta también forma parte de la definición de **Probabilidad Acumulada**, pero es la función para los casos de **funciones discretas**.

### Reto

Desarrolla una expresión matemática para el siguiente caso:

¿Cuál es la probabilidad de que mi variable aleatoria tome valores entre dos umbrales?

$$P(a \leq X \leq b) = ?$$

Pista:

Consideramos este caso como una variable aleatoria continua, donde tengo el umbral entre a y b



### Solución

$$P(a \leq X \leq b) = \int_a^b P(X) dx = P(b) - P(a) = C(X)$$

### Conclusión

La probabilidad es un campo que depende mucho de los elementos del cálculo, porque esas funciones que nos permiten determinar probabilidades sobre los diferentes estados de una variable aleatoria, son específicamente las que definimos en el cálculo sobre un punto de vista matemático, tales como sus propiedades matemáticas de derivación e integración que se aplican sobre funciones normales en cálculo, pueden ser aplicadas en probabilidad.

En el caso particular de la distribución de distribución acumulada, que es la integral de la función de la densidad de probabilidad.

Pero no te asustes! porque estas son las bases para que entiendas los mecanismos detrás del cálculo de probabilidad, pero en la práctica, pasaremos al código con Python para desarrollar esto como lo haría todo un científico de datos!

## Distribuciones discretas

Profundizaremos en como trabajar con distribuciones discretas, tales como el lanzamiento de monedas y dados, donde para este tipo de ejercicios, surge de manera natural la **distribución de Bernoulli**.

## Distribución de Bernoulli

Una distribución de Bernoulli es una función que asigna a la variable binaria dos valores, cuando  $X$  sea igual a 1 (éxito) ocurre con la probabilidad  $p$ , donde  $p$  valdría 0,5 dado un caso de probabilidad equilibrada, y cuando  $X$  sea igual a 0 (fracaso) se representa con la probabilidad de  $1 - p$ , porque la suma de las probabilidades tiene que dar el 100%.

Se dice que la variable aleatoria  $X$ , se distribuye como una Bernoulli de parámetro  $p$  con  $0 < p < 1$

### Fórmula de Bernoulli

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$

$$0 < p < 1$$

Desde la definición de esta función, podemos empezar a considerar situaciones más complejas con base al ejercicio de lanzar monedas, pero... no te preguntas sobre ¿cómo podemos hacer más complejo lanzar una moneda?, si solo hay 2 posibilidades del 50%, la respuesta es lanzar una  $n$  cantidad de monedas, 2, 3, 4 o las que yo quiera.

Por lo tanto, decimos que cuando tenemos secuencias repetitivas de eventos binarios (eventos tipo Bernoulli), es cuando tenemos que hablar de la famosa **Distribución binomial**, entonces en este punto es cuando empezaremos a desarrollar sus pasos precedentes, hasta poder entender de forma natural y sencilla sobre como surgen de manera fundamental las distribuciones binomiales.

### Ejemplo Distribución Bernoulli

#### Caso 1

Podemos decir que tenemos tres monedas que lanzaremos 1 vez cada una o que lanzaremos la misma moneda tres veces, donde la probabilidad de que obtengamos cara y cruz en cada lanzamiento es igualmente probable en ambos casos.

Entonces nos preguntamos:

**¿Cuál es la probabilidad de 3 lanzamientos de una monedas, en 2 de esos 3 lanzamientos obtengamos cara?**

Al graficar y contar las combinaciones, vemos que existen 8 posibles escenarios al lanzar las monedas:



Así sabemos que trabajamos con un espacio muestral de 8 posibilidades, y de esas 8 solo en 3 casos obtendríamos 2 caras.



Por lo tanto, decimos que la probabilidad de lanzar 3 veces una moneda y obtener 2 caras es de:

$$P(2, Caras | 3, Lanzamientos) = \frac{3}{8}$$

Asumimos en esta situación que la probabilidad de obtener cara o cruz, es igualmente probable en cada lanzamiento, porque esta es la hipótesis con la que hemos trabajado desde el principio del artículo, donde decimos que hay probabilidades fundamentales que son axiomáticas y, por lo tanto, asumimos que son igualmente probable.

Pero en caso de la distribución de Bernoulli, cuando esto no sucede, definimos el número  $p$  minúscula, que asigna la probabilidad de uno u otro suceso.

*En la vida real este parámetro se ajusta acorde los datos que obtengamos en práctica de X experimento.*

#### Caso 2

Si complicamos el caso de la distribución binomial, donde ya no nos preguntándonos por 3 lanzamientos de monedas, sino, por  $n$  lanzamientos, donde  $n$  podría ser un número muy grande, y de esos  $n$  lanzamientos, podemos tener  $k$  caras, que es la variable donde definiremos cuantas caras queremos tener según el número  $n$  de lanzamientos.

$$P(K = Caras | n = Lanzamientos) = ?$$

En este punto vemos que la formula se complica, ya que mientras más lanzamientos haya, el espacio muestral ( $EM$ ) crece de manera mayor que la exponencial.

[Calcular espacio muestral para una n cantidad de probabilidades](#)

Entonces en este punto nos preguntamos:

**¿Existe alguna fórmula general para contar todos estos posibles estados y sobre ellos hacer el conteo de probabilidades?**

La respuesta es claro que si, de esto es lo que se trata específicamente [la función de distribución binomial](#).

## Introducción a la distribución binomial

Volviendo a nuestro problema de los 3 lanzamientos de una moneda, usaremos la letra  $k$  para definir el número de caras o sucesos exitosos, que queremos obtener a partir de los  $n$  lanzamientos.

Veamos como sería la distribución binomial de este problema, donde ya sabemos que es binomial, pero...

**¿Qué quiere decir que una distribución sea binomial exactamente?**

Gráficamente sabemos que cuando una distribución es discreta, la gráfica tendrá forma de un diagrama de barras o histograma, donde cada barra representa la frecuencia relativa de un evento posible en el eje X.

Entonces decimos que los eventos posibles son el resultado que podemos obtener de cada lanzamiento: **0 caras, 1 cara, 2 caras y como máximo 3 caras**. Estas son las 4 posibilidades sobre las que podemos calcular las frecuencias relativas.



De todos los eventos, sabemos que tenemos 8 posibilidades, que sería nuestro espacio muestral.



Entonces... **¿Cuántas opciones del EM resultan en 0 caras, en 1 cara, 2 caras y 3 cars?**

Como vemos en el gráfico del espacio muestral, solo en un caso tenemos 0 caras, en 3 casos tenemos 1 y 2 caras, y en solo un caso tenemos 3 caras. Quedando la probabilidad quedaría tal que:

Donde:

$$P(k = cara, ; n = lanzamientos) = \frac{k}{EM}$$

$$P(0, 3) = \frac{1}{8} \quad P(1, 3) = \frac{3}{8} \quad P(2, 3) = \frac{3}{8} \quad P(3, 3) = \frac{1}{8}$$



De esta forma podemos reflejar el concepto de como se vería la distribución binomial para este caso en particular.

## Combinatorio o Coeficiente binomial

Volviendo a la pregunta inicial.

**¿Existe alguna fórmula general para contar todos estos posibles estados y sobre ellos hacer el conteo de probabilidades?**

¡Y como ya mencionamos antes, si disponemos una fórmula general!

En matemáticas contamos con un elemento que son los coeficientes binomiales, números combinatorios o combinaciones son números estudiados en matemáticas combinatoria que corresponden al número de formas en que se puede extraer subconjuntos a partir de un conjunto dado.

El número combinatorio  $\binom{n}{k}$  es el número de subconjuntos  $k$  elementos que satisfacen algún requisito de un conjunto con  $n$  elementos, y el subconjunto  $k$  tiene que ser menor que el conjunto  $n$ .

**Fórmula del Combinatorio:**

$$C_n^k = \binom{n}{k} = \frac{n!}{k! \cdot (n - k)!}$$

donde:

$n$  = número de intentos

$k$  = número de aciertos

$r \leq n$

## Ejemplo Combinatorio o Coeficiente binomial

Ahora que sabemos que existe una manera de contar de forma general todos los posibles estados dentro de un espacio muestral, pasaremos con un ejemplo para ver como nos ayuda a conocer la probabilidad de éxito donde:

**Queremos obtener la probabilidad de obtener  $k$  veces cara, dado  $n$  lanzamientos.**

donde:

$n = 3$

$k = 1$

Donde la probabilidad quedaría tal que:

$$P(1, 3) = ?$$

Usando el Combinatorio pasaremos a contar los posibles estados de éxito de obtener 1 cara dado 3 lanzamientos de una moneda quedando tal que:

$$P(k, n) \rightarrow C_k^n$$

Donde sabemos que  $n$  es el número de lanzamientos que son 3 y  $k$  sería el número de éxitos dado  $n$  que es 1:

$$P(1, 3) \rightarrow C_1^3 = \binom{3}{1} = \frac{3!}{1! \cdot (3 - 1)!} = \frac{1 * 2 * 3}{1 * 1 * 2} = \frac{\cancel{1} * \cancel{2} * 3}{\cancel{1} * \cancel{2}} = 3$$

Así es como sabemos las posibles maneras de obtener este resultado, que es 3.

Por último decimos que la probabilidad está dada por el resultado del combinatorio entre el espacio muestral total, que es 8, ya que son todos los estados posibles de lanzar 3 veces una moneda.

$$P(k, n) = \frac{C}{EM} \rightarrow P(3, 1) = \frac{3}{8}$$

Entonces vemos que a través de la fórmula combinatoria somos capaces de contar los estados dado dicho evento y lo podemos demostrar, comparando el resultado que obtuvimos con el gráfico de la distribución binomial que realizamos anteriormente, donde la barra que reflejaba la probabilidad de obtener 1 cara de 3 lanzamientos de una moneda era de  $\frac{3}{8}$ .

Así a través del símbolo combinatorio que nos permite contar los estados posibles, podemos desarrollar el cálculo de conteo de probabilidades, de esta manera dando paso a la introducción de la fórmula general de Distribución Binomial.

## Distribución Binomial

Definimos a la distribución binomial o distribución binómica como una distribución de probabilidad discreta que cuenta el número de éxitos en una secuencia de  $n$  ensayos de Bernoulli independientes entre sí, con una probabilidad fija  $p$  de ocurrencia de éxito entre los ensayos.

Decimos que la probabilidad de un suceso particular, es igual al número de estados que conducen a ese suceso, multiplicado por la probabilidad de cada estado individual:

$$p \rightarrow \text{suceso} \rightarrow p * \text{estados suceso}$$

Usando el ejemplo de las monedas:

Donde sabemos que el número 3 son los estados exitosos que obtuvimos a partir del combinatorio  $C_k^n$  y la probabilidad de individual de cada uno de los estados es de  $\frac{1}{8}$ .

Quedándonos el resultado tal que:

$$p = \frac{1}{8} * 3 = \frac{3}{8}$$

Estas probabilidades estarían dadas por la fórmula general que encontramos en la literatura como la Distribución Binomial.

## Desarrollo fórmula Distribución Binomial

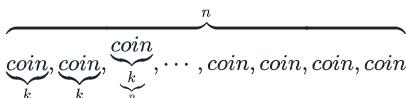
Cuando tengamos una probabilidad que dados de  $n$  intentos, y queramos obtener  $k$  resultados, el número de estados exitosos es igual a  $n$  combinado  $k$ .

$$P(k, n) = \binom{n}{k}$$

Esto lo que quiere decir es que de una  $n$  cantidad de estados, solo habrá una cantidad  $k$  de estados que satisfagan los resultados deseados y cada estado  $k$  tendrá una probabilidad  $p$

Ejemplo:

De lanzar una  $n$  cantidad de monedas, solo habrá una  $k$  cantidad que caiga en cara, donde cada una de esas  $k$  monedas tendrá una probabilidad  $p$ .



Asumimos que  $p$  para el caso ideal tendría una probabilidad de  $1/2$ , pero puede que esto no sea así, denotamos la letra  $p$  por si las probabilidades no se encuentran balanceadas y multiplicamos la probabilidad de cada uno de estos eventos, que sería la probabilidad de la primera moneda, por la probabilidad de la segunda, la tercera y así considerando a todas las  $k$ , que sería equivalente a elevar la probabilidad  $p$  a la  $k$ .

$$p(k_1) * p(k_2) * \dots * p(k_n) = p^k$$

Quedando la formula general tal que:

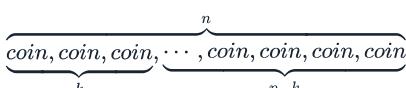
$$P(k, n) = \binom{n}{k} p^k$$

Si la probabilidad de que el evento individual éxito es  $p$ , decimos que la probabilidad del evento fallido es  $q = 1 - p$  tal cual vimos en la distribución de Bernoulli.

$$P(k, n) = \binom{n}{k} p^k q^{n-k}$$

De misma manera debemos descubrir la probabilidad de cada evento no exitoso, lo cual resolvemos multiplicando la probabilidad de cada uno de estos e igualmente sería equivalente a elevar  $q$  por el número de todos los eventos no exitosos.

Para saber cuantos eventos fallidos quedaron, simplemente tenemos que restarle a la totalidad de eventos  $n$  realizados, la cantidad de eventos exitosos  $k$ .



Por lo tanto, la probabilidad de fracaso quedaría elevado a la  $n-k$

Quedándonos la formula tal que:

### Fórmula de Distribución Binomial

$$P(X) = \binom{n}{k} p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Y así es como damos a la fórmula de la Distribución Binomial

donde:

$n$  = número de intentos

$k$  = número de aciertos

$p$  = probabilidad de éxito en un intento

$q = (1 - p)$  probabilidad de fracaso en un intento

### Conclusión

Así es como definimos a la distribución binomial, como, una distribución o función de densidad de probabilidad, donde podemos calcular de una secuencia de eventos de tipo Bernoulli cuantos éxitos puedo tener de variables binarias.

- $P(k \text{ caras} | n \text{ lanzamientos})$

$$\bullet P(k; n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Esta no es la única distribución que podemos trabajar con variables aleatorias binarias, ya que existen varias.

Donde también existen casos de variables aleatorias discretas no binarias, como por ejemplo la distribución multinomial, que es la generalización natural de la binomial.

Formula general de Distribución multinomial

$$P(X_1, \dots, X_n) = \frac{n!}{k_1! \cdot \dots \cdot k_n!} P_1^{k_1} \cdot \dots \cdot P_n^{k_n}$$

No profundizaremos en la distribución multinomial, porque lo importante es entender que existen otras distribuciones para variables discretas, con nombres interesantes que podrás impactar y asustar a tus amigos con simplemente nombrarlas.

Otras distribuciones

- [Poisson](#)
- [Geométrica](#)
- [Hipergeométrica](#)
- [Binomial negativa](#)
- [t de Student](#)

Entonces nos surge la duda de ¿Cómo sabremos cuando usar cada distribución habiendo tantas?, la verdad es que existen ciertas experiencias, investigaciones y experimentos aleatorios donde cada una de estas distribuciones se aplican de manera óptima.

En el siguiente capítulo veremos que hay casos donde tenemos un conjunto de datos particular y no sabemos al comienzo su distribución, aprenderemos que existen técnicas para ajustar la mejor distribución de probabilidad al conjunto de datos que tengamos.

¡Ya que en la vida real no sabemos exactamente las distribuciones en probabilidad y en conjuntos de datos, sino que tendremos que aprenderlas y tendremos ayuda de algoritmos que aprenden la distribución a partir de los datos, dando así el inicio al Machine Learning probabilístico!

## Usando la distribución binomial

```
# Dependencias

import numpy as np
from numpy.random import binomial
from scipy.stats import binom
import scipy.stats
from math import factorial
import matplotlib.pyplot as plt
```

### Función de la distribución binomial con Python

Para nuestro primer ejercicio, representaremos esta función en Python.

$$P(k, n; p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

$n = \text{intentos}$  $k = \text{exitos}$  $p = \text{probabilidad de que ocurra un evento } k$ 

```
def my_binomial(k,n,p):
    return factorial(n)/(factorial(k)*factorial(n-k))*pow(p,k)*pow(1-p,n-k)

my_binomial(2,3,0.5)
```

0.375

## Metodo de Scipy funcion binomial

[Documentacion](#)

## PMF Probability Mass Function

```
# scipy.stats.binom(numero de intentos, probabilidad).pmf(numero de exitos)
dist = binom(3, 0.5)

# pmf = probability mass function = funcion de densidad de probabilidad
pmf = dist.pmf(2)
```

## Función de distribución acumulada con Python

$$P(k \leq 2, n = 3; p = \frac{1}{2}) = \sum_{k=0}^2 \left[ \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \right] = \frac{7}{8}$$

Puedes intentar validar el resultado de este ejercicio en papel antes de pasar al código

```
probabilidades_individuales = scipy.stats.binom(3,0.5).pmf(range(0,3))
probabilidades_sumadas = round(np.sum(probabilidades_individuales),3)

print(f"""
Probabilidades Individuales: {probabilidades_individuales}
Probabilidades Sumadas: {round(probabilidades_sumadas,3)}
""")
```

```
Probabilidades Individuales: [0.125 0.375 0.375]
Probabilidades Sumadas: 0.875
```

## CDF Cumulative Distribution Function

```
# scipy.stats.binom(numero de intentos = 3 , probabilidad de eventos = 0.5).cdf(casos exitoso = 2)

# Cumulative distribution function. = funcion de distribucion acumulada
dist.cdf(2)
```

0.875

# Comprobamos que los resultados sean correctos

```
print(f"""
{probabilidades_sumadas}
{dist.cdf(2)}
{7/8}
""")
```

0.875  
0.875  
0.875

## Simulaciones de secuencias con generadores aleatorios

- Los generadores aleatorios tienen como propósito simular muestras de datos que resultarían de muestreos en la vida real de procesos aleatorios como lanzar una moneda o un dado.

```
# simulacion con 100 lanzamientos de moneda equilibrada
# (ejecuta esta celda varias veces para observar la variacion de los resultados)
```

```
p = 0.5
n = 3
```

```
binomial(n,p) #numpy.random.binomial
```

```
2
```

### Distribución simulada

- La probabilidad experimental o simulada es el resultado de un experimento aleatorio.

```
# Resultados de probabilidades teóricas, mientras más intentos hagamos, más se acerca a las probabilidades de la escuela frecuentista
```

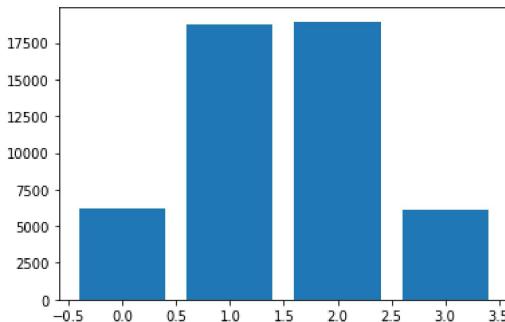
```
arr = []
```

```
def simulada(intentos, exito, probabilidad):
    for i in range(intentos):
        arr.append(binomial(exito, probabilidad))
x,y = np.unique(arr, return_counts=True)

plt.bar(x,y)
plt.show()

print(np.unique(arr, return_counts=True)[1]/len(arr)) # imprimimos las probabilidades
```

```
simulada(50000,3,.5) # Si subimos el numero de intentos, las probabilidades quedaran tal como la distribucion acumulada de probabilidad
```



```
[0.1233 0.37452 0.37902 0.12316]
```

```
values=[0,1,2,3]
[binom(3,.5).pmf(k) for k in values]
```

```
[0.125, 0.3750000000000001, 0.3750000000000001, 0.125]
```

### Distribución teórica

- La probabilidad teórica se basa en el modelo matemático desarrollado en la teoría de la probabilidad.

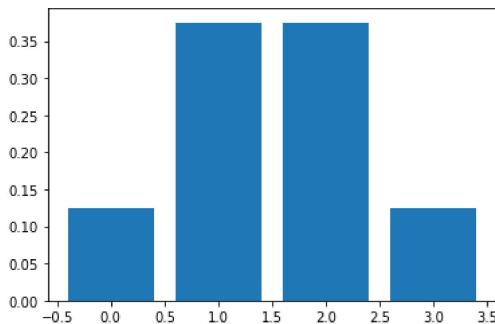
```
arr = []
```

```
def simulada(intentos, exito, probabilidad):
    for _ in range(intentos):
        arr.append(binomial(exito, probabilidad))

x = np.copy(values)
y = [binom(3,.5).pmf(k) for k in values]
plt.bar(x,y)
plt.show()

print(y) # imprimimos las probabilidades
```

```
simulada(10000, 3, .5)
```



```
[0.125, 0.3750000000000001, 0.3750000000000001, 0.125]
```

### Comparación de distribución teórica y simulada

La idea es poder visualizar como el acercamiento de una probabilidad simulada, se va acercando más a su probabilidad teórica en cuanto al crecimiento de sus experimentos.

```
def plot_hist(num_trials):
    values = [0,1,2,3]
    arr = []

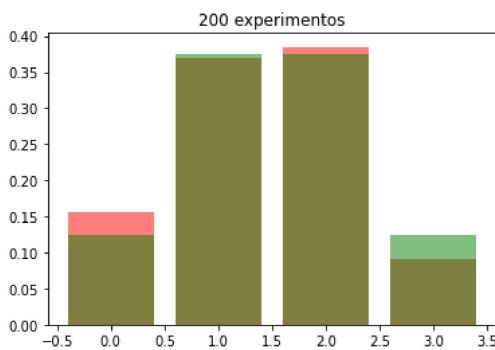
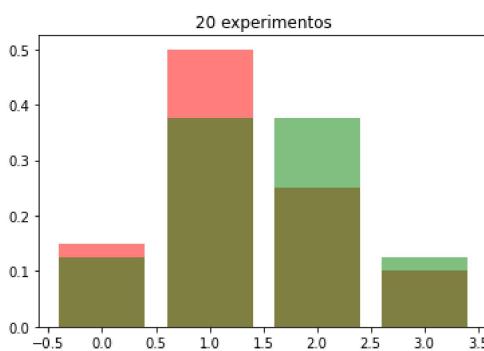
    for _ in range(num_trials):
        arr.append(binomial(3,0.5))

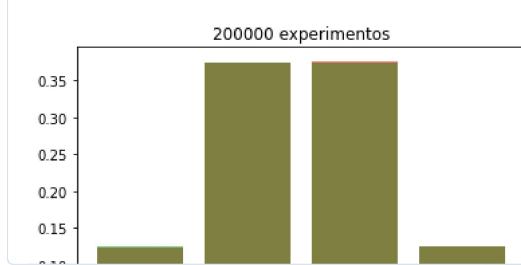
    distribucion_simulada = np.unique(arr, return_counts=True)[1]/len(arr)
    distribucion_teorica = [binom(3,0.5).pmf(k) for k in values]

    plt.bar(values, distribucion_simulada, alpha=0.5, color = 'red')
    plt.bar(values, distribucion_teorica, alpha=0.5,color='green')

    plt.title('{} experimentos'.format(num_trials))
    plt.show()

plot_hist(20)
plot_hist(200)
plot_hist(200000)
```





## Conclusión

Con esto vemos como nos ayudamos de Python para hacer simulaciones de eventos aleatorios que siguen una ley de distribución binomial.

De esta manera evidenciamos el pensamiento de la escuela frequentista, donde las probabilidades teóricas se cumplen en el momento que el número de elementos es muy grande, de esta manera consiguiendo una validación del experimento

Este esquema, muy usado en la vida del científico de datos, nos puede servir cuando tengamos algún experimento que sea muy costoso llevarlo a cabo, pero si podemos remplazar dichos experimentos por su versión simulada, podríamos lograr hacer relativamente más económico hacer experimentos que validen nuestros ejercicios como científicos de datos.

## Distribuciones continuas

Anteriormente, trabajamos las distribuciones discretas, en particular la binomial. Ahora nos toca ver y trabajar las distribuciones continuas son aquellas que toman valores que no son necesariamente un número entero, sino que ya nos encontramos con variables que se encuentran dentro del los [números reales](#).

Para adentrarnos al tema, empezaremos exemplificando con famosa [distribución normal \(gaussiana\)](#), que es una distribución de variable continua que con más frecuencia aparecen en estadística y teoría de probabilidades

Nos ayudaremos de un del siguiente dataset para poder hacer un procesamiento y análisis de los datos para entender como a partir de estos datos nace nuestra distribución normal.

## Distribución normal teórica

Función de densidad:

$$P(X) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

$\mu$  = media de la distribución

$\sigma$  = desviación estándar de la distribución

```
# Dependencias

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm
```

Función normal en Python

```
def gaussiana(x, median, std):
    return 1/(std*np.sqrt(2*np.pi))*np.exp(-0.5*(x-median)/std**2)
```

Graficamos la función

```
x = np.arange(-4, 4, 0.1)
y = gaussiana(x, 0.0, 1.0)

plt.plot(x,y)

[<matplotlib.lines.Line2D at 0x7f42f804e250>]
```



### Desplazamos la media de la distribución

Veremos que la distribución es la misma, pero se desplaza a donde su media esté ubicada

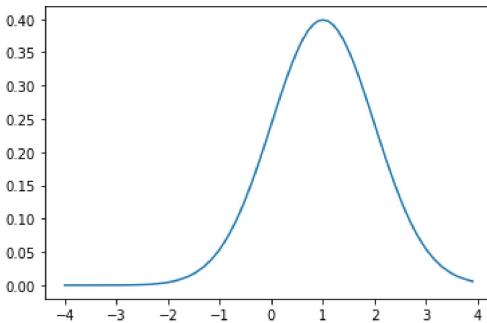
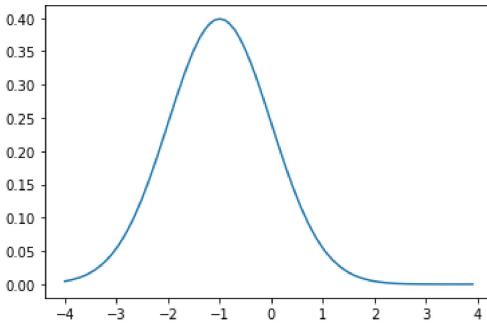
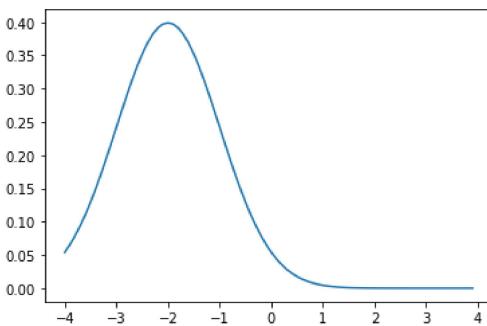
```
x = np.arange(-4, 4, 0.1)
y1 = gaussiana(x, -2.0, 1.0) # media de -2
y2 = gaussiana(x, -1.0, 1.0) # media de -1
y3 = gaussiana(x, 1.0, 1.0) # media de 1
y4 = gaussiana(x, 2.0, 1.0) # media de 2
```

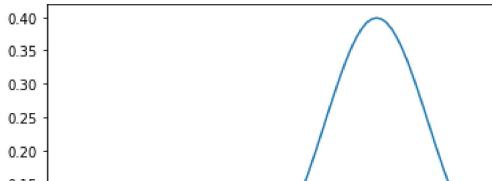
```
plt.plot(x,y1)
plt.show()
```

```
plt.plot(x,y2)
plt.show()
```

```
plt.plot(x,y3)
plt.show()
```

```
plt.plot(x,y4)
plt.show()
```





### Desplazamos la desviación estándar de la distribución

Al modificar la desviación estándar veremos como la distancia entre cada variable se reduce, o se aumenta.

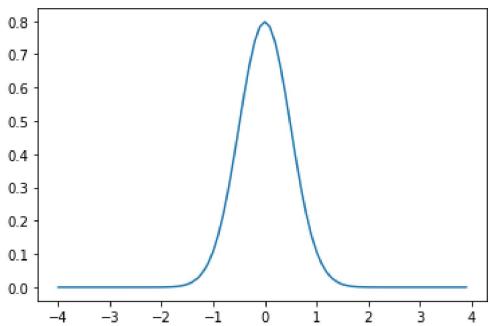
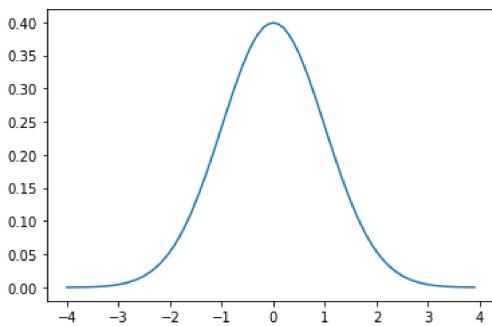
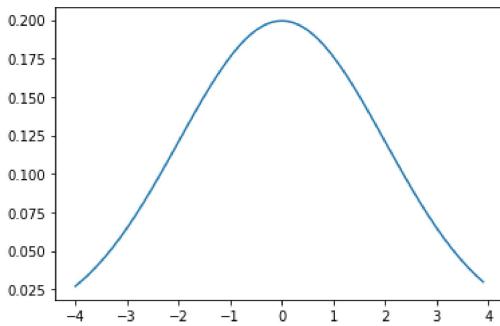
```
x = np.arange(-4, 4, 0.1)
y1 = gaussiana(x, 0, 2.0) # Desviacion estandar de 2
y2 = gaussiana(x, 0, 1.0) # Desviacion estandar de -2
y3 = gaussiana(x, 0, 0.5) # Desviacion estandar de -1
y4 = gaussiana(x, 0, 0.1) # Desviacion estandar de 1

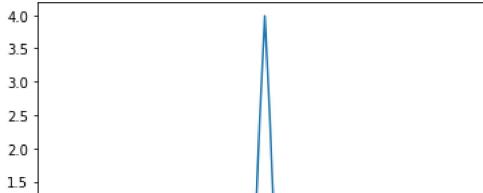
plt.plot(x,y1)
plt.show()

plt.plot(x,y2)
plt.show()

plt.plot(x,y3)
plt.show()

plt.plot(x,y4)
plt.show()
```





### Distribución normal con SciPy

Una vez que entendemos el mecanismo de una función podemos ayudarnos de librerías como SciPy para facilitar y optimizar el código.

[Scipy.norm\(\).pdf\(\)](#)

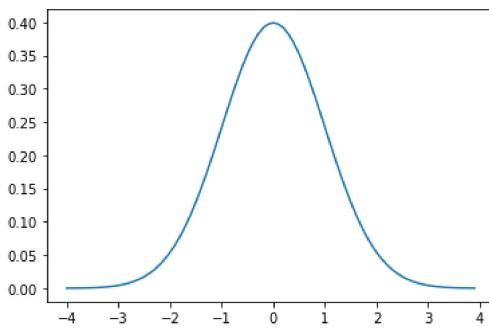
```
# scipy.stats.norm(mean, std)
# pdf = Probability density function.

dist = norm(0,1)

x = np.arange(-4,4, 0.1)
y = [dist.pdf(value) for value in x]

plt.plot(x,y)
```

[<matplotlib.lines.Line2D at 0x7f42f7c60df0>]



### Función de probabilidad acumulada con Python

Es conveniente recordar que toda distribución sea discreta o continua, tiene su distribución acumulada de probabilidad, que su fórmula la vimos al comienzo del capítulo 3, que su fórmula es la integral de esta función

$$P(X \leq \underbrace{x}_{\text{parámetro}}) = \int_{X \leq x} P(X)dX = C(x)$$

Sin embargo, cuando hablamos toquemos en otro artículo el cálculo matemático, verás que no es fácil de integrar una distribución normal y mucho menos nos pondremos a hacer los cálculos a mano, porque afortunadamente constamos con SciPy que tiene métodos para calcular esto de una forma numérica, ósea una aproximación de su integral

```
# cdf = Cumulative distribution function.

dist = norm(0,1)

x = np.arange(-4,4,0.1)
y = [dist.cdf(value) for value in x]

plt.plot(x,y)
```

[<matplotlib.lines.Line2D at 0x7f42f7c8ac40>]

10



## Distribución Normal en Python

Pasaremos a analizar el siguiente dataset sobre el tamaño de las alas de moscas domésticas y nos sirve para ejemplificar una distribución normal desde el campo de la biometría.

Descargaremos el archivo Excel y lo cargaremos con Pandas

### Dataset

```
# Importamos la base de datos
```

```
df = pd.read_excel('./data/s057.xls')
df.head()
```

Normally Distributed Housefly Wing Lengths				
	Unnamed: 1	Unnamed: 2	Unnamed: 3	
0	Sokal, R.R., and P.E.Hunter. 1955.	NaN	NaN	NaN
1	NaN	NaN	NaN	NaN
2	length (x.1mm)	NaN	NaN	NaN
3	36	NaN	Bin	Frequency
4	37	NaN	36-38	2

```
# Guardamos solo la columna de la longitud de alas que tiene todos los datos, omitiendo los 3 primeros valores que no nos sirven para la
```

```
arr = df['Normally Distributed Housefly Wing Lengths'].values[3:]
print(arr)
```

```
[36 37 38 38 39 39 40 40 40 41 41 41 41 41 41 41 42 42 42 42 42 42 42 43
 43 43 43 43 43 43 44 44 44 44 44 44 44 44 45 45 45 45 45 45 45 45 45
 45 45 46 46 46 46 46 46 46 46 46 46 47 47 47 47 47 47 47 47 48 48 48
 48 48 48 48 49 49 49 49 49 49 49 49 50 50 50 50 50 50 51 51 51 52 52
 53 53 54 55]
```

```
# Separamos los valores únicos del array con NumPy y activamos que cuente cuantas veces se repite cada valor único
```

```
values, dist = np.unique(arr, return_counts=True)

print(f"""
{values} Lista de valores únicos
{dist} Lista de frecuencia de valores únicos
""")
```

```
[36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55] Lista de valores únicos
[ 1  1  2  2  4  6  7  8  9 10 10  9  8  7  6  4  2  2  1  1] Lista de frecuencia de valores únicos
```

```
# Graficamos con matplotlib
```

```
plt.bar(values,dist)
plt.xlabel('Largo de alas en milímetros')
plt.ylabel('frecuencia')

Text(0, 0.5, 'frecuencia')
```

### PDF Probability Density Function

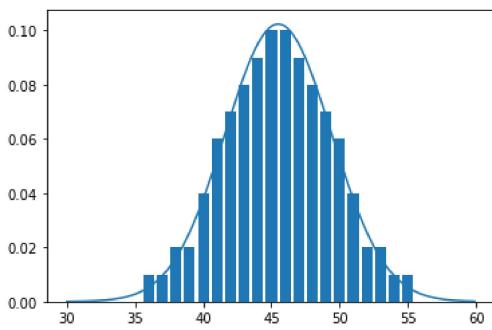
```
# El gráfico que obtuvimos a simple vista luce como una campana de gauss, pero para asegurarnos de que esto sea así, lo verificaremos haciendo lo siguiente

mean = arr.mean()                                     # Calculamos el promedio del array de datos
std = arr.std()                                       # Calculamos la desviación estándar de los datos
x = np.arange(30, 60, 0.1)                            # Creamos una lista de valores que se encuentre un poco mas alejado de los valores límites de la muestra
dist = norm(mean, std)                                # Con scipy definimos la distribución normal con el promedio y desviación estandar de los datos
y = [dist.pdf(value) for value in x]                  # Calculamos Y con la densidad de probabilidad en cuanto a los datos de dist

plt.plot(x,y)                                         # Graficamos la estimación de la distribución

values, dist = np.unique(arr, return_counts=True)
plt.bar(values,dist/len(arr))                         # Graficamos la distribución del ejercicio anterior, y normalizamos la lista dist con el total de datos

plt.show()
```



### Conclusión

Este procedimiento fue para verificar que nuestros datos se asemejan a una distribución normal, forzando los parámetros de la distribución gaussiana con el promedio y desviación estándar de los propios datos.

Así comprendemos que la distribución normal aparece de forma natural en situaciones cotidianas, donde los datos reflejan cosas de la naturaleza misma, sin embargo, no es la única distribución que existe, podemos encontrarnos con otras distribuciones como la exponencial, gama, Pareto que se usarán en situaciones específicas.

## ¿Cómo estimar una distribución?

En el capítulo anterior entendimos como una distribución gaussiana o normal es el patrón natural de distribuciones de probabilidades de un conjunto de datos reales y esto nos llevó a tener que ajustar una función de probabilidad a un conjunto de datos, como mencionamos anteriormente, esto se debe a los principios del Machine Learning, que trata de ajustar una distribución a un conjunto de datos, para con dicha distribución hacer predicciones.

¡Pasemos a aprender mejor sobre como hacer la **estimación paramétrica** con la práctica!

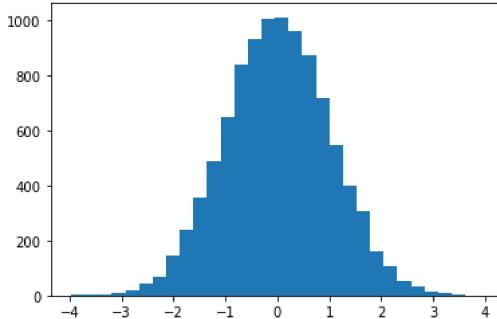
```
# Dependencias

import numpy as np
from numpy.random import normal
from scipy.stats import norm
import matplotlib.pyplot as plt
```

Creamos un conjunto de datos aleatorios con NumPy para poder simular los datos de una distribución normal.

Este paso lo realizamos, ya que no estamos tomando datos reales de alguna investigación, sino que los generamos de forma artificial para usarlos como base para el ejercicio de estimación paramétrica y no paramétrica. Ambas estimaciones son tipos de estimaciones de densidades, pero existen pequeñas diferencias

```
sample = normal(size = 10000) # generador aleatorio
plt.hist(sample, bins = 30);
```



## Estimación Paramétrica

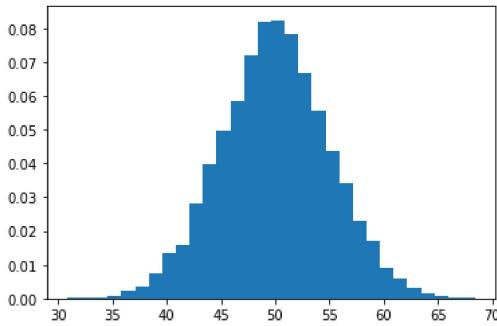
La estimación paramétrica consiste en suponer una función para la distribución y ajustar o forzar los parámetros de los datos a dicha distribución.

Generamos los datos aleatorios con su media en 50, desviación estándar de 5 y un espacio muestral de 10.000 datos aleatorios.

```
sample = normal(loc=50, scale=5, size=10000) # mean=50, std=5, size=10000
```

Si graficamos el resultado, veremos algo como lo expuesto en la siguiente imagen, la cual tiene una forma de distribución normal:

```
plt.hist(sample,bins=30, density=True);
```



El siguiente paso es calcular una función teórica que se ajuste al conjunto de datos.

En este caso se conoce la media y el desvío estándar porque los colocamos nosotros para realizar la simulación, pero en una situación real deberemos de calcular la media y el desvío estándar de los datos

```
mean = sample.mean()
std = sample.std()
```

Con estos valores, se crea una instancia de un objeto cuyos parámetros son precisamente mean y std.

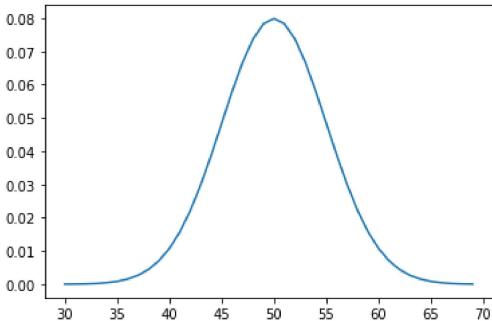
Es decir, tenemos la base para aplicar la fórmula de la función de distribución Gausiana, pero nos faltan los valores sobre los cuales vamos a calcular esas probabilidades.

Entonces primero se produce el objeto **dist** y luego se genera un array **values** cuyo rango va a variar entre los extremos de los datos reales y calculamos las probabilidades.

```
dist = norm(mean,std)
values = [value for value in range(30,70)]
probabilidades = [dist.pdf(value) for value in values]
```

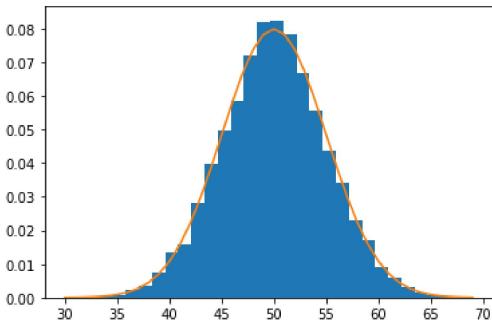
Si graficamos ahora solo la función teórica nos queda lo siguiente:

```
plt.plot(values, probabilidades)
plt.show()
```



Finalmente, graficamos los datos y la curva teórica calculada y observamos que se asemejan.

```
plt.hist(sample, bins=30, density=True)
plt.plot(values, probabilidades)
plt.show()
```

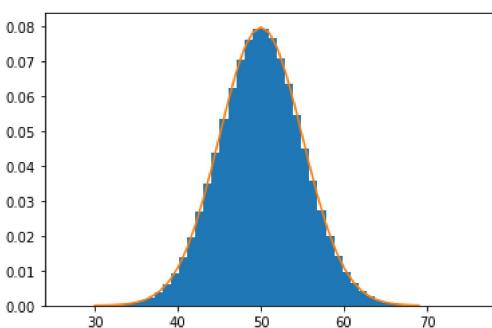


# Código completo

```
sample = normal(loc=50, scale=5, size=500000) # mu = 50, sigma = 5
mu = sample.mean()
sigma = sample.std()

dist = norm(mu, sigma)
values = [value for value in range(30, 70)]
probabilidades = [dist.pdf(value) for value in values]

plt.hist(sample, bins=50, density=True)
plt.plot(values, probabilidades)
plt.show();
```



## Estimación No Paramétrica

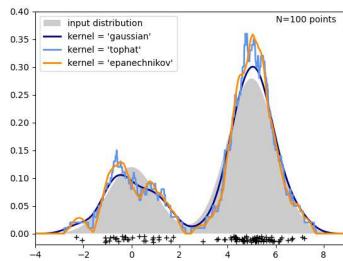
la estimación no paramétrica se aplica cuando los datos no se ajustan a ninguna distribución conocida, por lo tanto no se ajustan los parámetros de dicha distribución, sino que se trata de una combinación de varias distribuciones.

Para este tipo de ejercicios nos ayudaremos de un método que ya viene incluido dentro de la librería de Scikit-learn, el cual se llama Kernel Density Estimation

### Kernel Density estimation

- parámetro de suavizado: smoothing parameter
- función base: basis function

La idea de este método, es que cuando nos encontramos con distribuciones no gaussianas no podríamos lograr una estimación de la distribución, porque no se podrían ajustar los datos a la misma tal como en el caso de una distribución bimodal.



¡Para entenderlos mejor pasemos a la práctica!

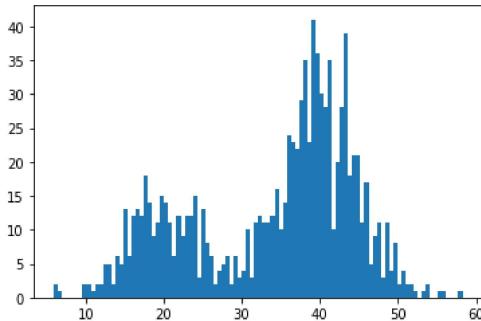
Primero simularemos dos distribuciones normales como ya lo hicimos anteriormente y las juntaremos en una sola distribución bimodal a través del método `hstack()` de NumPy.

```
# Dependencias

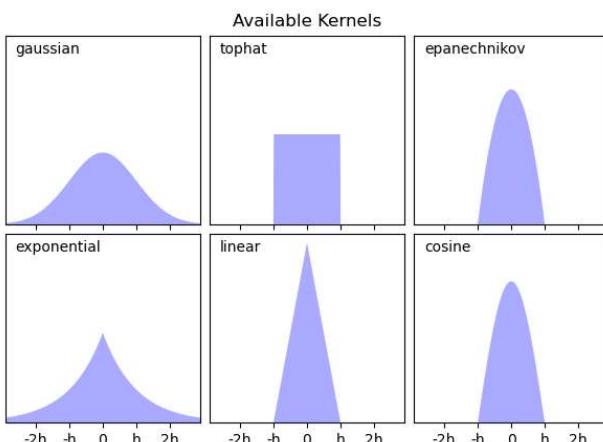
import matplotlib.pyplot as plt
from numpy import hstack
from sklearn.neighbors import KernelDensity

# construccion de una distribucion bimodal
sample1 = normal(loc=20, scale=5, size=300)
sample2 = normal(loc=40, scale=5, size=700)
sample = hstack((sample1, sample2))

# graficamos
plt.hist(sample, bins=100);
```



Una vez que tenemos los datos simulados, comenzamos el proceso de estimación, para ello se crea un objeto `model`, el cual de instancia a través de los parámetros `bandwidth` (parámetro de suavizado) y `kernel` (función base que se adapta según la forma de la distribución).



Esto es equivalente al caso anterior de estimación paramétrica, donde teníamos la función normal teórica, y luego calculábamos las probabilidades, solo que ahora no tenemos solo una función de densidad de probabilidad, sino un conjunto de distribuciones.

Una vez creado el objeto, se ajustan los datos a las necesidades del objeto, para esto se utiliza el método `reshape`, el cual los ordena en una matriz de  $n$  filas y 1 columna y luego se ajusta el modelo a estos datos.

```

model = KernelDensity(bandwidth=2, kernel='gaussian') # (parametro de suavizado, funcion base)
# print(sample)
sample = sample.reshape((len(sample), 1))
# print(sample)
model.fit(sample)

KernelDensity(bandwidth=2)

```

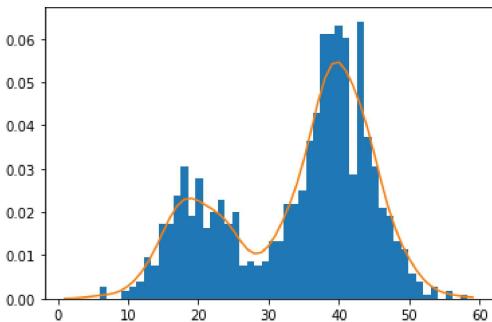
Estamos ajustando el modelo a los datos, es como si estuviéramos calculando la media y el desvío estándar de una distribución normal, pero nos faltan calcular los promedios para darle forma a la función teórica.

Esto es lo que hacemos a continuación, se crea un array en el rango de los datos reales sobre cuáles queremos estimar la función, y luego calculamos las probabilidades.

```

values = np.asarray([value for value in range(1, 60)])
values = values.reshape((len(values),1))
probabilities = model.score_samples(values) # probabilidad logarítmica para optimizar calculos computacionales
probabilities = np.exp(probabilities) # invertimos los resultados de las probabilidades logarítmicas con su exponencial para tener la es
plt.hist(sample, bins=50, density=True)
plt.plot(values, probabilities)
plt.show()

```



Con este ejercicio vimos dos maneras de ajustar probabilidades teóricas a un conjunto de datos real.

Este tipo de problema es fundamental porque como ya mencionamos en Machine Learning por lo general y en todo modelo que se trabaje sobre un conjunto de datos, consiste en que siempre tendremos que ajustar una densidad de probabilidad a un conjunto de datos real, por esta razón este tema es tan importante.

En el siguiente tema veremos un framework muy popular en el momento de ajustar densidades de probabilidad a datos reales

## Mini Proyecto:

Consumiremos los datos de una API de la NASA para poder hacer una estimación de velocidad de asteroides que pasan cerca de la tierra en tiempo real, usaremos lo aprendido para hacer una estimación de la misma analizando si para este caso se ajusta mejor una estimación paramétrica o no paramétrica según la forma de los datos.

Hay que tener en cuenta que la API solo nos deja ver los datos desde la fecha de inicio a 7 días en adelante por ejemplo

- inicio: 1 de febrero del 2022
- fin: 7 de febrero del 2022

```

# Dependencias
from scipy.stats import norm
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import KernelDensity
from datetime import date, timedelta
import requests
import json

# Valores de entrada
start_date = date.today() - timedelta(days=7)
end_date = date.today()
api_key = 'JksQ8eXx4qD051uW0X4Q23ptCbGzr0B7e5DpcDKp'
api_url = 'https://api.nasa.gov/neo/rest/v1/feed?start_date={start_date}&end_date={end_date}&api_key={api_key}'.format(start_date = star

```

```
# Consumo de API

def obtain_asteroid_speed():
    try:
        request = requests.get(api_url)
        data = json.loads(request.text)
    except:
        print('xD')

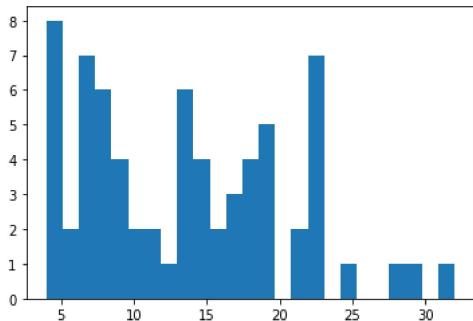
    speeds = []

    for day in data['near_earth_objects']:
        for obj in data['near_earth_objects'][day]:
            speeds.append(round(float(obj['close_approach_data'][0]['relative_velocity']['kilometers_per_second'])))
    speeds = np.array(speeds)

    return speeds

arr = obtain_asteroid_speed()
```

```
# Forma de los datos
plt.hist(arr, bins=25);
```



## Resolucion

Intenta resolverlo solo, podras ver como quedarian los resultados y sacar tus propias conclusiones!

```
def parametric_estimation_asteroid_speed():

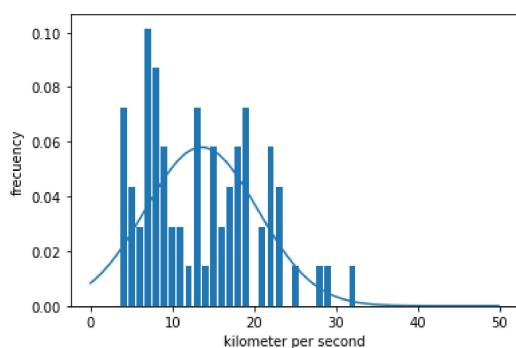
    mean = arr.mean()
    std = arr.std()
    x = np.arange(0, 50, 0.1)
    dist = norm(mean, std)
    y = [dist.pdf(value) for value in x]

    values, dist = np.unique(arr, return_counts=True)

    plt.bar(values,dist/len(arr))
    plt.plot(x,y)

    plt.xlabel('kilometer per second')
    plt.ylabel('frequency')
    plt.show()

parametric_estimation_asteroid_speed()
```



```

arr = arr.reshape((len(arr),1))

def kernel_density_estimation():
    model = KernelDensity(bandwidth=2, kernel='epanechnikov')
    model.fit(arr)

    values = np.asarray([value for value in range(1,35)])
    values = values.reshape((len(values),1))
    probabilities = model.score_samples(values)
    probabilities = np.exp(probabilities)

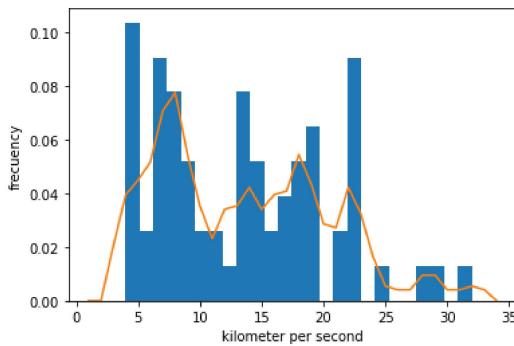
    plt.hist(arr, bins=25,density=True)
    plt.plot(values,probabilities)

    plt.xlabel('kilometer per second')
    plt.ylabel('frequency')

    plt.show()

kernel_density_estimation()

```



## Tema 4: MLE (Maximum Likelihood Estimation)

### ¿Qué es MLE?: Estimación de máxima verosimilitud

En el capítulo anterior aprendimos lo que es la estimación de densidad de probabilidad tanto con métodos paramétricos y no paramétricos, también vimos lo importante que es estimar la distribución de probabilidad de un conjunto de datos, por lo tanto, en este capítulo veremos uno de los esquemas más comunes al realizar este proceso.

#### MLE

MLE que son las siglas del inglés **Maximum Likelihood Estimation** que la conocemos como **estimación de máxima verosimilitud** o **EMV** por sus siglas en español, es una técnica que nos permite estimar densidades de probabilidad dentro de un esquema de trabajo muy general.

Podemos decir que el objetivo del MLE es encontrar la forma mas óptima de ajustar una distribución a los datos.

- **Probabilidad:** que tantas veces (observaciones) puede pasar algo en una cantidad de posibilidades.
- **Verosimilitud:** qué tan bien describe a los datos/observaciones un modelo estadístico.

#### Elementos del MLE

Los elementos esenciales de este esquema son:

- Escoger la distribución: Teniendo solo una muestra de los datos.
- Escoger los parámetros de la distribución: Que nos permitan ajustar mejor o peor la distribución a los datos.

Ejemplo:

En el ejercicio del tema anterior teníamos una distribución normal y calculamos directamente la media y desviación estándar de estos datos para ajustar la mejor campana de Gauss de los datos

El problema general de la EMV es que los datos obedecen a una distribución de probabilidad de una población, la cual no tendremos el conocimiento total sobre los datos de la misma, así que en general la distribución de probabilidad de la muestra de datos que tenemos es diferente a la distribución de probabilidad del problema poblacional donde podríamos entender su forma en el hipotético caso que tuviéramos el conocimiento de todos esos datos posibles

Viendo esto en el esquema frecuentista sabemos que esto es teóricamente posible, pero solo en nuestra imaginación, ya que en la vida real, nunca tendremos todos los datos existentes para nuestras investigaciones.

Por lo tanto, decimos que escoger la distribución de probabilidad sobre una muestra de datos es la primera restricción que debemos considerar.

Luego de tener la muestra de la población general y la distribución, ya posea unos parámetros que usaremos para ajustar la distribución a nuestros datos.

 esto también se refiere al Machine Learning, porque también aquí tendríamos unas variables que deberemos tunear o calibrar para poder ajustar un conjunto de datos

Así llegamos al punto donde decimos que el MLE es un problema de optimización

## ¿Un problema de optimización?

Decimos que el MLE es un problema de optimización porque formulamos el esquema de trabajo de la siguiente manera:

Tenemos un conjunto de datos  $X$  ( $X$  representa en general un dataset donde puede tener muchos datos o representar cualquier proceso de eventos aleatorios)

Por otro lado, tenemos los parámetros de la distribución que queremos ajustar

Entonces hay una probabilidad que tenemos de ajustar los datos a una distribución en concreto.

A esta distribución, que resulta de haber ajustado los parámetros para ciertos datos en general, se escribe con letra  $L$  ( $L$  representa a la sigla en inglés Likelihood que es verosimilitud en español)

- \$

$$P(X; \theta) = L(X; \theta)$$

¿Una vez que tenemos esta densidad de probabilidad de datos unos parámetros y dados un conjunto de un dataset, nos preguntamos ¿Y yo como resuelvo el problema de optimización   ?

Al haber tantos conjuntos de parámetros que nos permitan ajustar los datos con distintos grados de probabilidad, lo que haremos es que de todas esas posibles combinaciones, elegir aquella la cual su probabilidad es la máxima posible.

El máximo de la función  $L$  de  $X$  dado los parámetros  $\theta$  ( $\theta$ )

- \$

$$\max ; L(X; \theta)$$

Una hipótesis importante es que a veces esta distribución de probabilidad sobre el dataset se puede factorizar como el producto de varias probabilidades, donde cada probabilidad concierne a un data point del conjunto de datos, por lo tanto, estos data point son  $X_i$  en la expresión general.

- \$

$$\max ; L(X; \theta) \rightarrow \max ; \prod \limits_{i=1}^n P(X_i; \theta)$$

Entonces decimos que tenemos un producto de probabilidades.

Pero ante esto sucede un problema frecuente que se presenta cada vez que factorizamos distribuciones de probabilidad como el producto de varias probabilidades, ya que las probabilidades al ser números pequeños al ir multiplicándolas, estas van adquiriendo más decimales hacia la derecha, y esto computacionalmente hablando existe una precisión límite que por las cuales debajo de esta las computadoras ya no pueden computar, a esto se lo denomina como underflow.

Por lo tanto, cuando trabajamos con este tipo de problemas, es normal que se nos presente el underflow.

Para solucionar este tipo de problemas es aplicando el logaritmo de las probabilidades en vez de las probabilidades en sí mismas, ya que al realizar el logaritmo aplicamos una de sus propiedades que **el logaritmo de un producto es igual a la suma de los logaritmos**, que es una propiedad matemática de los logaritmos.

Por lo tanto, convertimos un problema de multiplicaciones a sumas, donde los logaritmos convierten números pequeños a números negativos relativamente más grande que facilitan la computación desde un punto numérico.

**Como resultado, decimos que en un problema de estimación de máxima verosimilitud lo que hacemos es** calcular el máximo del logaritmo de la verosimilitud en función de  $L$  que es igual a calcular el máximo de la sumatoria de los logaritmos de las probabilidades individuales, donde cada probabilidad corresponde a un data point dado un parámetro  $\theta$ .

$$\max \log L(X; \theta) \rightarrow \max \sum_i \log P(X_i; \theta)$$

Este es el problema general que se formula para hallar la densidad de probabilidad que mejor se ajusta a cierto conjunto de datos.

[Video complementario sobre estimación de máxima verosimilitud <https://seeing-theory.brown.edu/probability-distributions/es.html>](https://seeing-theory.brown.edu/probability-distributions/es.html)

## MLE en machine learning

Visto el esquema de estimación de máxima verosimilitud, concluimos que:

El objetivo del MLE es encontrar una forma óptima de ajustar la distribución de los datos para que podamos representarlos mejor y facilitar el proceso de trabajar con distribuciones más generales.

En este capítulo desarrollaremos este concepto teórico con un ejemplo en particular.

## ¿Cómo se usa el MLE en Machine Learning?

Decimos que en general el **machine learning** consiste en **ajustar densidades a datos**, desde un punto probabilístico todos los problemas se reducen a esta simple frase, ya sea que tengamos modelos de machine learning **supervisados** como la **clasificación y regresión** o modelos **no supervisados** como la **clusterización**.

Solamente por motivos de aclaración mencionamos que la regresión lineal con estimación de máxima verosimilitud se reduce a que la regresión lineal consiste en que tenemos un conjunto de datos y poseamos una intuición de que esos datos deberían seguir un modelo lineal, siendo de una variable una función tal que:

$$y = m \cdot x + b$$

m: pendiente de la recta

b: ordenada al origen

Lo interesante de esto es que cuando estudiemos la teoría en ciencias de datos, donde  $b_0$  antes era la pendiente que ahora es el peso (weight) y donde  $b_1$  antes era la ordenada al origen, pero ahora es él bias

$$y = b_0 \cdot x + b_1$$

$b_0$ : weight

$b_1$ : bias

Por lo tanto, decimos que estas convenciones que se dan entre la matemática tradicional y el machine learning representan lo mismo

$$y = \underbrace{m}_{\text{pendiente}} * x + \underbrace{b}_{\text{ordenada al origen}} = \underbrace{b_0}_{\text{weight}} * x + \underbrace{b_1}_{\text{bias}}$$

Dado el problema de que debemos encontrar el modelo lineal que mejor ajuste un conjunto de datos, entonces aplicando estimación de máxima verosimilitud se leería tal que:

La probabilidad de dado el conjunto de datos  $x$  obtenga  $y$ , que lo expresamos como el máximo de la suma de logaritmos de las probabilidades de cada pareja  $(x_i, y_i)$  de un conjunto de datos suponiendo el modelo  $h$ .

$$P(y|x) \rightarrow \max \sum_i \ln P(y_i|x_i; \underbrace{h}_{\text{modelo}})$$

$$h \rightarrow y = m \cdot x + b$$

💡 En general, la estimación de máxima verosimilitud como sirve para cualquier distribución, el modelo  $h$  no siempre será un modelo lineal, solamente que en este caso  $h$  es la hipótesis que utilizaremos para ajustar un modelo lineal.

## Práctica teórica

Pasos

- Escoger la distribución:
- Escoger los parámetros de la distribución:

Fórmulas:

$$\max \left\{ \sum_i \ln P(x_i | y_i; h) \right\}$$

$$h \rightarrow y = mx + b$$

$$P \rightarrow \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y-\mu}{\sigma} \right)^2}$$

### Demostración de estimación de máxima verosimilitud

Teniendo ya las fórmulas por usar presentadas, contamos con un conjunto de datos imaginarios al cual queremos ajustar un modelo lineal, el cual sería la recta de color rojo.



Por lo cual asumimos que la recta deberá tener una ecuación  $y = mx + b$  donde buscamos determinar los coeficientes  $m$  y  $b$  que mejor ajusten los datos.

$$h \rightarrow y = mx + b$$

Por lo tanto, formulamos el esquema de máxima verosimilitud, suponiendo que la desviación de los datos respecto a la recta principal (ruido o noise) tomaría una distribución gaussiana, por esto mismo nuestra hipótesis de modelamiento probabilístico es que la probabilidad tendría una forma gaussiana.

$$; ; ; P \longrightarrow \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2}$$



Así pues, determinado el primer punto del MLE donde ya declaramos que nuestra distribución  $f$  tendrá forma gaussiana.

$MLE \longrightarrow dist; f \longrightarrow Gaussiana$

Decimos en el esquema tradicional mediante el cual aprendemos la regresión lineal es con los mínimos cuadrados.

¿Pero qué quiere decir los mínimos cuadrados?

Cuando nos encontramos con un dato real  $(x_i, y_i)$  y a este valor se le aplica la función teórica del modelo donde el resultado que deberíamos obtener es  $y = mx + b$ , la  $y$  obtenida de esta fórmula es diferente a la  $y_i$  de los datos reales.

Al ser estos valores diferentes esto determinara un error para cada data point dentro del conjunto de datos.



Este error lo calculamos como la diferencia entre  $y_i$  de los datos reales, menos el  $y$  del modelo teórico:

$$y = (m, x_i + b)$$

$$y_i - y$$

$$y_i - (m, x_i + b)$$

Esta diferencia tiende a resultados positivos como negativos, por lo que a estos errores individuales los elevamos al cuadrado para luego hacer la sumatoria de todos los errores buscando minimizar esta suma

$$\sum_i (y_i - (m, x_i + b))^2$$

De esta manera lo que buscamos con el problema de regresión lineal es encontrar el mínimo de la suma de los errores cuadráticos, que también se lo conoce como el método de mínimos cuadrados, donde lo más usual es entender la regresión lineal desde este punto de vista.

$$\min \sum_i (y_i - (m, x_i + b))^2$$

Lo que buscamos demostrar es que la estimación de máxima verosimilitud es equivalente a la forma en la que aprendemos la regresión lineal para entender que de misma forma es un problema probabilístico con el mismo esquema

Pero llegamos a la pregunta

¿Cómo demostraríamos que la estimación de máxima verosimilitud a los mínimos cuadrados?

Empezamos por nuestra hipótesis de probabilidad donde decimos que el ruido de los datos seguirá una distribución gaussiana y la hipótesis de modelamiento donde decimos que el modelo tiene que ser un modelo lineal.

- Función lineal:  $; ; ; h \longrightarrow y = mx + b$

$$\bullet \text{ Función de campana de Gauss: } ; ; ; P \longrightarrow \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2}$$

¿Cómo combinamos estas dos fórmulas de nuestra hipótesis?

Combinaremos estas dos funciones dentro de la expresión que usamos para hacer la estimación de máxima verosimilitud

KaTeX parse error: Expected '}', got '\right' at position 99: ...; | ;x\_i ; ; h \right\}

Resulta que dentro de nuestra función de Gauss tenemos la variable del ruido ( $y$ ) y la media ( $\mu$ ) donde la tendencia media es la tendencia de la recta que buscamos calcular

$$P \longrightarrow \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{y - \mu}{\sigma} \right)^2}$$

Es decir, la resta entre estos dos términos de la campana de Gauss ( $y - \mu$ ) es totalmente equivalente entre los términos de los mínimos cuadrados ( $y_i - (m, x_i + b)$ )

Por lo tanto, matemáticamente resultaría que la función de estimación de máxima verosimilitud es igual al máximo de la suma de las probabilidades logarítmicas de la función gaussiana, donde cambiaremos el exponente ( $y - u$ ) por ( $y_i - (mx_i + b)$ ).

KaTeX parse error: Expected '}', got '\right' at position 71: ...; ;x\_i ; ; h \right\} = \max \left\{ \dots

Formulado el problema ya hasta este punto, parece que a simple vista no hay relación alguna entre nuestra nueva expresión del MLC y el método de mínimos cuadrados, salvo por la diferencia entre el ruido y la media que demostramos anteriormente.

$$\text{MLE} \rightarrow \max \sum_i -i \ln\left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_i-(mx_i+b)}{\sigma}\right)^2}\right)$$

$$\text{MSE} \rightarrow \min \sum_i -i(y_i - (m, x_i + b))^2$$

Sigamos desglosando a mayor profundidad nuestra expresión de máxima verosimilitud para demostrar que son el MLE y el MSE son problemas equivalentes

Si vemos ese término en específico del MLE

KaTeX parse error: Expected '}', got 'right' at position 179: ...t)^2}}\right) \underline{\right)}

Podemos ver que estamos calculando el logaritmo del producto de dos cosas, que como ya mencionamos anteriormente, el logaritmo de un producto es igual a la suma de dos logaritmos, sea el  $\ln$  de  $a$  más el  $\ln$  de  $b$

$$\ln \left( \underbrace{\left( \frac{1}{\sigma\sqrt{2\pi}} \right)}_{\text{constante}} * \underbrace{\left( e^{-\frac{1}{2}\left(\frac{y_i-(mx_i+b)}{\sigma}\right)^2} \right)}_{\text{logarítmico}} \right) = \ln(a) + \ln(b)$$

Esto es lo mismo que decir, el máximo de la suma del logaritmo constante de  $(\frac{1}{\sigma\sqrt{2\pi}})$  más la suma del término logarítmico  $(e^{\frac{1}{2}\left(\frac{y_i-(mx_i+b)}{\sigma}\right)^2})$

KaTeX parse error: Expected '}', got 'right' at position 283: ...2 } \right) \underline{\right)}

Donde despreciamos el término constante, ya que al tener todos sus elementos fijos no generaría ninguna diferencia en el problema de maximización que implica escoger un montón de probabilidades donde todas tendrán este mismo término constante

KaTeX parse error: Expected '}', got 'right' at position 296: ...2 } \right) \underline{\right)}

Quedando la expresión reducida tal que

KaTeX parse error: Expected '}', got 'right' at position 181: ...2 } \right) \underline{\right)}

La siguiente reducción la daremos con las funciones inversas del logaritmo natural y la exponencial, ya que al decir que de un valor  $x$  le aplicamos una función  $f$  y al resultado le aplicamos la función inversa  $f^{-1}$  el resultado final volvería a ser el valor  $x$ , que exactamente lo que está ocurriendo en nuestra

$$f^{-1}(f(x)) = x$$

Que es exactamente lo que sucede cuando calculamos el logaritmo natural de una exponencial, por lo tanto, ahora nuestra reducción quedaría tal que

KaTeX parse error: Expected '}', got 'right' at position 168: ...2 } \right) \underline{\right)}

Resulta que si despejamos las últimas constantes que nos quedan que serían  $-\frac{1}{2}$  y  $\sigma$ , el signo menos de la fracción cambiaria mi máximo al mínimo, quedando tal que:

KaTeX parse error: Expected '}', got 'right' at position 120: ...}^2 \right) \underline{\right)}

Así para finalizar demostramos que el problema de estimación de máxima verosimilitud también consiste en calcular el mínimo de la suma de los errores cuadráticos, así demostramos que el problema de mínimos cuadrados también es un problema de MLE

## Demostracion matematica



[Maximum Likelihood For the Normal Distribution, step-by-step!!!](#)

[Estimación puntual - Método de máxima verosimilitud \(+ 2 ejemplos\) - \[Clase 1/6\]](#)

## Regresión logistica

La Regresión logística es una tecnica que se puede usar tanto en la estadística tradicional, como en el Machine Learning.

La regresión logística es similar a la regresión lineal que vimos en el ejercicio anterior, pero esta nos ayuda a predecir cuando tenemos que clasificar datos de manera binaria, siendo algo **verdadero**, o siendo algo **falso**.



Otra de sus particularidades, es que a diferencia de la regresión lineal con la regresión logística es que ahora ajustamos los datos con una línea con forma de "S" que proviene de la [función sigmoid](#), donde los valores se ubicaran dentro de los rangos 0 y 1 (0: falso | 1: verdadero)



$$\text{función sigmoid: } y = \frac{1}{1 + e^{-x}}$$

¿Cómo acomodamos los datos para usar la regresión logística?

Primero deberemos comprender que nos encontramos ante un problema de clasificación con atributos, donde tenemos varias independientes que dependiendo de sus valores podremos predecir su clase

$$\underbrace{\{X_1, X_2, \dots, X_n\}}_{\text{atributos}} \longrightarrow y \xleftarrow[0,1]{} \text{clase}$$

Por ejemplo, predecir si una transacción bancaria fue fraudulenta o no, donde 1 podría ser *fraudulento* y 0 *no fraudulento* y las variables representarían datos como:

$X_1$  : Hora de transacción

$X_2$  : Monto de transacción

$X_3$  : Distancia entre la distancia del dueño de la tarjeta y la transacción

$X_n$  : etc.

Dentro de estos problemas combinar a las variables con un peso ( $\beta$  beta) o parámetros que nos permitan determinar descubrir cuáles variables son más relevantes o no

$$\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n = X$$

El resultado de la sumatoria de los pesos combinados a las variables nos dará por resultado una [combinación lineal](#) de todas mis variables resultando en  $X$ , que será el exponente dentro de nuestra función sigmoidal

$$P = \frac{1}{1 + e^{-X}}$$

## MLE para Regresión logística

[Video explicativo máxima verosimilitud para regresión logística](#)

Formularímos el problema de máxima verosimilitud misma manera que lo hicimos con la regresión lineal:

$$\max \left\{ \sum_i \ln P(y_i | x_i; h) \right\}$$

Tendremos el máximo de las sumas de los logaritmos naturales de las probabilidades de  $y$  dado  $x$  y una hipótesis de modelamiento  $h$

donde:

$y_i$  : clase o categoría de cada elemento y  $x_i$  : son los atributos de cada elemento, donde además cada elemento del dataset satisface una distribución de Bernoulli:

$$P = \begin{cases} p, & \text{si } y = 1, \\ 1 - p, & \text{si } y = 0. \end{cases}$$

En este caso la verosimilitud está dada por:

$$L = \hat{y}y + (1 - \hat{y})(1 - y)$$

Esta función da como resultado probabilidades altas cuando  $\hat{y} \sim y$ .

## Aplicación de regresión logística con Python

```
# Dependencias

from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
from matplotlib import cm
import numpy as np
import pandas as pd
```

$$L = \hat{y}y + (1 - \hat{y})(1 - y)$$

[numpy.meshgrid](#)

```
# Funcion de verosimilitud para reg. logistica
def likelihood(y, yp):
    return yp*y + (1-yp)*(1-y)
```

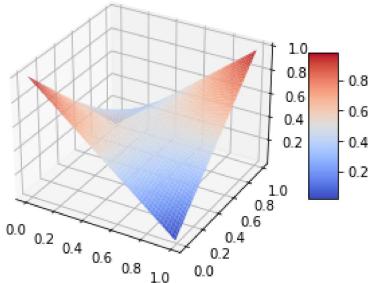
```
fig = plt.figure()
ax = plt.axes(projection='3d')

# generaremos X(vector), Y(vector) y Z(matriz)
Y = np.arange(0, 1, 0.01)
```

```
YP = np.arange(0,1, 0.01)
Y, YP = np.meshgrid(Y, YP)
Z = likelihood(Y, YP)

# Superficie
surf = ax.plot_surface(Y,YP,Z, cmap=cm.coolwarm)
fig.colorbar(surf, shrink=0.5, aspect=5)

plt.show()
```



Vemos que la verosimilitud entre la variable tiene máximos en 1 cuando las predicciones se cumplen con los datos reales, y en 0 cuando las predicciones son opuestas a los datos reales

Ahora veamos la verosimilitud con un gráfico interactivo para poder ver la comparación de las variables predichas con los datos reales.

```
# La libreria que nos ayudaremos es plotly  
#pip install plotly
```

```
Collecting plotly
  Downloading plotly-5.10.0-py2.py3-none-any.whl (15.2 MB)
|██████████| 15.2 MB 11.3 MB/s eta 0:00:01

Collecting tenacity>=6.2.0
  Downloading tenacity-8.0.1-py3-none-any.whl (24 kB)

Installing collected packages: tenacity, plotly
Successfully installed plotly-5.10.0 tenacity-8.0.1
Note: you may need to restart the kernel to use updated packages.
```

```
import plotly.graph_objects as go

# generamos X(vector), Y(vector) y Z(matriz)
y = np.arange(0, 1, 0.01)
yp = np.arange(0, 1, 0.01)
Y, YP = np.meshgrid(y, yp)
Z = likelihood(Y, YP)

# graficar
fig = go.Figure(data=[go.Surface(z=Z, x=y, y=yp)])

fig.update_traces(contours_z=dict(show=True, usecolormap=True,
                                    highlightcolor="limegreen", project_z=True))

fig.update_layout(title=' ', autosize=False,
                  width=500, height=500,
                  margin=dict(l=65, r=50, b=65, t=90))

fig.show()
```

Ahora, de manera más interactiva vemos cuando los datos reales y los datos predichos coinciden ( $x, y$ ) la verosimilitud ( $z$ ) entre las variables aumenta.

Considerando  $p \rightarrow \log(p)$ , y sumando la verosimilitud para todos los puntos del dataset obtenemos:

$$\begin{aligned} & \max \sum_i (y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \\ & = \min - \sum_i (y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \end{aligned}$$

Que es la conocida función de costo para clasificación conocida como [Cross-entropy](#).

## Regresión logística con Scikit-learn

Recordemos que:

$$\hat{y} = \frac{1}{1 + \exp(-\text{log-odds})}$$

Donde  $\text{log-odds} = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$  y los betas son los parámetros del modelo.

Aplicaremos un ejercicio de clasificación simple con el dataset Iris:

- [https://en.wikipedia.org/wiki/Iris\\_flower\\_data\\_set](https://en.wikipedia.org/wiki/Iris_flower_data_set)
- [https://scikit-learn.org/stable/auto\\_examples/datasets/plot\\_iris\\_dataset.html](https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html)

Aprovecharemos la clase [sklearn.linear\\_model.LogisticRegression](#) para poder crear el modelo de regresión logística donde ya implementa todas las funciones que necesitaremos

```
# Dependencias
from sklearn.datasets import load_iris
from sklearn.linear_model import LogisticRegression
```

```
# atributos del dataset iris
atrib_names = ['sepal_length', 'sepal_width', 'petal_length', 'petal_width']

# X: atributos de los objetos
# y: tipos de flores
X, y = load_iris(return_X_y=True)
```

```
# Miramos los primeros 2 atributos del datapoint
print(X[:4])
print(y[:4])
```

```
[[5.1 3.5 1.4 0.2]
 [4.9 3. 1.4 0.2]
 [4.7 3.2 1.3 0.2]
 [4.6 3.1 1.5 0.2]]
 [0 0 0 0]
```

Revisamos los parámetros resultantes  $\beta_i$ :

```
# Instanciamos el modelo de clasificación logístico para las dos primeras clases x: [0,1] y para los 100 primeros atributos de cada dato
clf = LogisticRegression(random_state=10, solver='liblinear').fit(X[:100], y[:100])
```

```
# Vemos la matriz de coeficientes
clf.coef_
```

```
array([[-0.40247392, -1.46382925,  2.23785648,  1.00009294]])
```

Estos 4 números que tenemos en nuestro array hacen referencia a los  $\beta_i$  de nuestras 4 variables, siendo los parámetros que mejor ajustan a las predicciones del modelo de clasificación a las categorías de los datos reales

Ahora veamos un sencillo ejemplo de como clasificar datos con la regresión logística

```
# Logistic Regresion  
# Preparamos los datos de entrada y salida  
#X: se mantiene como esta  
y = y.reshape((len(y), 1))
```

```
x[ :2 ]
```

```
array([[5.1, 3.5, 1.4, 0.2],  
       [4.9, 3. , 1.4, 0.2]])
```

```
y[ :2 ]
```

```
array([[0],  
      [0]])
```

## # Logistic Regressions

```
from sklearn.linear_model import LogisticRegression
```

```
# Ajustamos los datos al modelo de regresion logistica  
model = LogisticRegression().fit(X,y)
```

```
/home/mazzaroli/anaconda3/lib/python3.9/site-packages/sklearn/utils/validation.py:63: DataConversionWarning: A column-vector y was passed when a 1d array was
    return f(*args, **kwargs)
/home/mazzaroli/anaconda3/lib/python3.9/site-packages/sklearn/linear_model/_logistic.py:763: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. OF ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (`max_iter`) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

[https://scikit-learn.org/stable/modules/linear\\_model.html#logistic-regression](https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression)

```
n_iter i = check optimize result(
```

# Exatitud del modelo

*# Devuelve la precisión media en los datos de prueba y las etiquetas dadas.*

```
model.score(X, y)
```

0.9733333333333334

# Realizamos la predicción

```
expected = y
predicted = model.predict(X) # lo ideal seria tener nuevos inputs, pero para este caso de estudio usaremos el mismo
predicted # Resultado de las predicciones para las flores setosa, versicolor y virginica
```

Para comprender mejor este resultado haremos un reporte de clasificación y una matriz de confusión que nos darán información extra sobre esta predicción

```
from sklearn import metrics
```

Nos guiaremos dentro de la puntuación f1-score para guiarnos sobre la precisión de los datos, recordemos que el dataset iris tiene sus 3 flores 0: setosa, 1: versicolor y 2: virginica

```
print(metrics.classification_report(expected, predicted))
```

	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	50
1.0	0.98	0.94	0.96	50
2.0	0.94	0.98	0.96	50
accuracy			0.97	150
macro avg	0.97	0.97	0.97	150
weighted avg	0.97	0.97	0.97	150

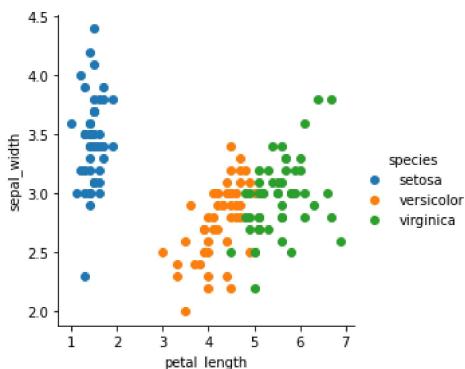
```
import seaborn as sns
iris = load_iris()

iris = pd.DataFrame(data = np.c_[iris['data'], iris['target']], columns=atrib_names + ['species'])

iris['species'] = iris['species'].map({ 0:'setosa', 1:'versicolor', 2:'virginica'})

sns.FacetGrid(data=iris, hue='species', height=4).map(plt.scatter, 'petal_length', 'sepal_width').add_legend()

<seaborn.axisgrid.FacetGrid at 0x7f2a0e7b5d90>
```



Como vemos para la especie setosa hay un 100% de precisión a la hora de hacer clasificaciones, y con la versicolor y virginica un 97% de precisión.

Dentro del caso práctico es un dataset bastante sencillo de clasificar, como vemos para las flores setosas se encuentran alejadas de la versicolor y virginica, pero existe un pequeño outliers dentro de la versicolor y virginica donde se presta para la confusión de la clasificación

```
print(metrics.confusion_matrix(expected, predicted))
```

```
[[50  0  0]
 [ 0 47  3]
 [ 0  1 49]]
```

Recordemos que el dataset iris consta de 150 datapoints donde se dividen en 50 datapoints para cada especie.

- Por lo tanto, vemos que en el primer valor se predijo 50/50 datapoints correctamente para las setosas
- Para la versicolor tuvimos un 47 datapoints correctamente clasificados, y 3 datapoints se clasificaron como de la virginica
- Y para finalizar tuvimos 49 datapoints correctamente clasificados para la virginica y 1 datapoint se clasificó como versicolor

## Tema 5: Inferencia bayesiana

### Teoremas de Bayes

Llegado a este punto, podemos abordar uno de los temas que se habló al principio de este artículo, donde nos elevaremos un escalón en la estadística probabilística, dejando la escuela frecuencia para adentrarnos a la **escuela bayesiana**.

Asumiendo que ya nos encontramos familiarizados con las probabilidades condicionales, daremos un pequeño repaso de las probabilidades condicionales.

Enunciado:

Fuimos de viaje a Baikalia y le preguntamos a 14 personas si les gustaban los ñoquis con queso o con estofado.

- 2 nos respondieron que les gustaban sus ñoquis con estofado y queso rallado.
- 5 personas nos dijeron que no les gustaba el estofado pero si con queso.
- 4 personas no le gustaban sus ñoquis con queso pero si con estofado.
- 3 personas les gustaban sus ñoquis sin estofado y sin queso.

	$ESTOFADO_{si}$	$ESTOFADO_{no}$	$P3$
$QUESO_{si}$	2	5	
$QUESO_{no}$	4	3	

Luego calculábamos las probabilidades para cada suceso dividiendo el número de resultados por el total de personas que entrevistamos

	$ESTOFADO_{si}$	$ESTOFADO_{no}$	$P3$
$QUESO_{si}$	$p = 2/14$	$p = 5/14$	
$QUESO_{no}$	$p = 4/14$	$p = 3/14$	

Ahora calcularemos las probabilidades para cada persona que le gustan sus ñoquis con y sin queso, como también para las personas que les gustan sus ñoquis con o sin estofado

	$ESTOFADO_{si}$	$ESTOFADO_{no}$	total queso
$QUESO_{si}$	$p = 2/14$	$p = 5/14$	$2 + 5 = 7$ $p = 7/14$
$QUESO_{no}$	$p = 4/14$	$p = 3/14$	$3 + 4 = 7$ $p = 7/14$
total estofado	$2 + 4 = 6$ $p = 6/14$	$5 + 3 = 8$ $p = 8/14$	

Entonces podemos calcular la **Probabilidad Condicional** de que alguien en Baikalia que **no le guste el estofado, pero si el queso con sus ñoquis, sabiendo que le gusta el queso**:

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si})$$

Para hacer esto dividímos la probabilidad de las personas que no le guste el estofado por la probabilidad de las personas que si le guste el queso

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si}) = \frac{5/14}{2+5/14} = 0.71$$

Vemos que el numerador es la **probabilidad incondicional** de que a alguien en Baikalia no le guste el estofado, pero si el queso en sus ñoquis

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{2+5/14} = 0.71$$

y el denominador es la **probabilidad incondicional** de que a alguien en Baikalia le guste el queso

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(QUESO_{si})} = 0.71$$

Por lo tanto, **la probabilidad de que a alguien no le guste el estofado, pero si el queso en sus ñoquis, dado que sabemos que le gusta el queso, es igual a la probabilidad de que a alguien no le gusta el estofado, pero si el queso sobre la probabilidad de que a alguien le guste el queso**.

Entonces lo que podemos decir acerca de la probabilidad condicional es que es la probabilidad de que un evento ocurra, teniendo en cuenta el conocimiento que tenemos acerca del evento

Ahora comparemos la probabilidad de que a **una persona no le guste la carne y si el queso, sabiendo de antemano que le gusta el queso**, con la probabilidad que calculamos anteriormente donde sabemos que a la persona no le gusta la carne.

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | ESTOFADO_{no}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(ESTOFADO_{no})} = \frac{5/14}{8/14} = 0.63$$

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(QUESO_{si})} = \frac{5/14}{7/14} = 0.71$$

En ambos casos queremos saber la probabilidad para el mismo evento, saber a quién no le gusta el estofado, pero si el queso, sin embargo, ya que tenemos distintos conocimientos previos para cada caso, tendremos probabilidades diferentes para cada evento.

## Derivando el teorema de Bayes

Ahora tratemos de resolver la **probabilidad condicional sin saber la probabilidad de que no te guste el estofado y si el queso**  
 $p(ESTOFADO_{no}, QUESO_{si})$

Incluso si no conocemos la probabilidad de que a alguien no le guste el estofado, pero si el queso

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | QUESO_{si}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(QUESO_{si})}$$

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | ESTOFADO_{no}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(ESTOFADO_{no})}$$

Podemos multiplicar ambos miembros de la primera ecuación por la probabilidad de que a alguien si le guste el queso

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si}) = \frac{P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})}{P(\text{QUESO}_{si})} * P(\text{QUESO}_{si})$$

Donde cancelaremos estos términos

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si}) = \frac{P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})}{\cancel{P(\text{QUESO}_{si})}} * \cancel{P(\text{QUESO}_{si})}$$

Y en la izquierda quedamos con que la probabilidad de que a alguien no le guste el estofado, pero si el queso es igual a la probabilidad de que a alguien no le guste el estofado, pero si el queso dado que sabemos que le gusta el queso multiplicado la probabilidad de que a alguien le guste el queso

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si}) = P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})$$

A mismo modo, podemos multiplicar ambos miembros de la segunda ecuación por la probabilidad de que a alguien no le guste el estofado.

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no}) = \frac{P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})}{P(\text{ESTOFADO}_{no})} * P(\text{ESTOFADO}_{no})$$

Donde estos dos términos a la izquierda quedan cancelados

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no}) = \frac{P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})}{\cancel{P(\text{ESTOFADO}_{no})}} * \cancel{P(\text{ESTOFADO}_{no})}$$

Igual que antes, el miembro de la derecha resulta en la probabilidad de que a alguien no le guste el estofado, pero si el queso

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no}) = P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})$$

Ahora tenemos los dos miembros del lado izquierdo de ambas ecuaciones, donde la probabilidad de que encontramos a alguien que no le guste el estofado, pero si el queso es igual.

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si}) = P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})$$

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no}) = P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})$$

El enunciado era resolver estos términos

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no})$$

Sin conocer el término

$$P(\text{ESTOFADO}_{no}, \text{QUESO}_{si})$$

Como ambas ecuaciones resultan en el mismo resultado, decimos que ambas ecuaciones son iguales una a la otra:

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si}) = P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no})$$

Recordemos que buscamos resolver este término  $P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si})$  y este otro término

$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no})$ , así que empezaremos por el término de la izquierda dividiendo ambos lados por la probabilidad de que a alguien le guste el queso

$$\frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})}{P(\text{QUESO}_{si})} = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no})}{P(\text{QUESO}_{si})}$$

La probabilidad de que a alguien le guste el queso se cancela por sí misma en el miembro de la izquierda

$$\frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * \cancel{P(\text{QUESO}_{si})}}{\cancel{P(\text{QUESO}_{si})}} = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no})}{P(\text{QUESO}_{si})}$$

y así resolvemos este miembro

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no})}{P(\text{QUESO}_{si})}$$

Ahora movemos el miembro que se encontraba en la derecha al miembro de la izquierda, y de mismo modo el miembro de la izquierda al de la derecha

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no}) = P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})$$

y dividimos ambos miembros por la probabilidad de que a alguien no le guste el estofado

$$\frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * P(\text{ESTOFADO}_{no})}{P(\text{ESTOFADO}_{no})} = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})}{P(\text{ESTOFADO}_{no})}$$

y la probabilidad de que a alguien no le guste el estofado, se cancela en el miembro izquierdo

$$\frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) * \cancel{P(\text{ESTOFADO}_{no})}}{\cancel{P(\text{ESTOFADO}_{no})}} = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})}{P(\text{ESTOFADO}_{no})}$$

¡logramos resolver el otro término!

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{ESTOFADO}_{no}) = \frac{P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})}{P(\text{ESTOFADO}_{no})}$$

De esta forma en ambos casos resolvemos el problema porque ya no necesitaremos de la probabilidad de que a alguien no le guste el estofado, pero si el queso!

¡Pero aún más importante derivamos el teorema de Bayes!

El Teorema de Bayes nos dice que esta probabilidad condicional que está basada en que sabemos que a las personas les gusta el queso  $P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \text{QUESO}_{si})$  puede ser derivada desde esta probabilidad condicional, donde se basa en que sabemos que no les gusta

el estofado  $P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no})$ .

Alternativamente, el Teorema de Bayes nos dice que esta probabilidad condicional que está basada en que sabemos que a las personas no les gusta el estofado  $P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no})$  se puede derivar desde la probabilidad condicional donde sabemos que a las personas les gusta el queso  $P(ESTOFADO_{no} \text{ y } QUESO_{si}|QUESO_{si})$

Generalmente, si le damos el valor  $A$  la probabilidad de que a las personas no le gusta el estofado y a  $B$  que les gusta el queso

$$A = ESTOFADO_{no} \quad B = QUESO_{si}$$

Podríamos reescribir cada ecuación con la fórmula general del Teorema de Bayes

$$P(A, B|B) = \frac{P(A, B|A) * P(A)}{P(B)}$$

$$P(A, B|A) = \frac{P(A, B|B) * P(B)}{P(A)}$$

En otras palabras, dada la probabilidad condicional donde sabemos una cosa acerca de un evento, puede ser derivada desde otra cosa que sabemos acerca de este mismo evento.

## Importancia del Teorema de Bayes

Ahora te preguntarás... **¿cuál es la utilidad del Teorema de Bayes después de realizar un poco de álgebra para derivar su fórmula?**

Pues cuando tenemos toda la información en algún gráfico bonito o de misma manera en una tabla de contingencia, entonces el Teorema de Bayes no es tan importante, pues en el caso en que tengamos todos los datos, ni siquiera tomamos en cuenta el Teorema de Bayes.

Pero normalmente en la mayoría de las veces **\*\*no tendremos toda la información. \*\***

En otras palabras podríamos enunciar:

- La probabilidad de que a alguien no le guste el estofado dado que le guste el queso es de 0.71  
→  $p(ESTOFADO_{no}, QUESO_{si}|QUESO_{si}) = 0.71$
- Pensamos que la probabilidad de que a alguien le guste el queso es **cercano a 0.6** →  $p(QUESO_{si}) \approx 0.6$
- La probabilidad de que a alguien no le guste el estofado es de 0.57 →  $p(ESTOFADO_{no}) = 0.57$

Entonces esta es toda la información que tenemos...

- $p(ESTOFADO_{no}, QUESO_{si}|QUESO_{si}) = 0.71$
- $p(QUESO_{si}) \approx 0.6$
- $p(ESTOFADO_{no}) = 0.57$

Luego cambiamos los valores en el Teorema de Bayes

$$P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no}) = \frac{P(ESTOFADO_{no} \text{ y } QUESO_{si}|QUESO_{si}) * (QUESO_{si})}{P(ESTOFADO_{no})}$$

$$P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no}) = \frac{0.71 * 0.6}{0.57} \approx 0.75$$

Y obtendríamos **aproximadamente 0.75**.

Lo que significa que dada esta información, que incluye una conjectura sobre la probabilidad de que a alguien le guste el queso... que la probabilidad de que alguien le guste el queso sabiendo que no les gusta el estofado, es aproximadamente 0.75

Habrás notado que **al calcular esta probabilidad condicional con el Teorema de Bayes, donde lo usamos porque no teníamos toda la información necesaria, el resultado que obtuvimos fue distinto a cuando calculamos la probabilidad de cuando sabíamos todos los resultados.**

- $P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no}) = \frac{0.71 * 0.6}{0.57} \approx 0.75$
- $P(ESTOFADO_{no} \text{ y } QUESO_{si}|ESTOFADO_{no}) = \frac{p(ESTOFADO_{no}, QUESO_{si})}{p(ESTOFADO_{no})} = \frac{\frac{5}{14}}{\frac{8}{14}} = 0.63$

**Esto se debe a que no sabíamos el valor exacto para las personas que les gustaba el queso  $p(QUESO_{si})$ ,** por lo cual tuvimos que adivinar. Y sabemos que puede sonar a una mala idea hacer una adivinanza para calcular una probabilidad... pero es la única opción que tenemos cuando tenemos un espacio muestral demasiado grande.

Por ejemplo, sería casi imposible preguntar a todas y cada una de las personas en Argentina que les gusta el queso... por lo tanto, muchas veces nos tocará hacer una estimación.

La Estadística Bayesiana trata de entender que implica hacer una adivinanza de este estilo y todo lo que conlleva.

## Notación general para Teorema de Bayes

En estos ejercicios usamos una nomenclatura para facilitar nuestro entendimiento en el aprendizaje.

En la mayoría de lugares donde busques acerca de la probabilidad condicional, sabiendo que a esta persona no le gusta el estofado

$$P(ESTOFADO_{no} \text{ y } QUESO_{si} | \underbrace{ESTOFADO_{no}}_{\text{información}})$$

No se la incluye al establecer la probabilidad

$$P(\text{ESTOFADO}_{no} \text{ y } \text{QUESO}_{si} | \underbrace{\text{ESTOFADO}_{no}}_{})$$

Ahora nuestra probabilidad condicional se lee tal que:

La probabilidad de que alguien le guste el queso, dado que no le guste el estofado

$$P(\text{QUESO}_{si} | \text{ESTOFADO}_{no})$$

A misma manera, como sabemos que a esta persona le gusta el queso, no se lo incluye al establecer la probabilidad

$$P(\text{ESTOFADO}_{no} \text{ y } \underbrace{\text{QUESO}_{si}}_{\text{QUESO}_{si}} | \text{QUESO}_{si})$$

$$P(\text{QUESO}_{si} | \text{ESTOFADO}_{no}) = \frac{P(\text{ESTOFADO}_{no} | \text{QUESO}_{si}) * P(\text{QUESO}_{si})}{P(\text{ESTOFADO}_{no})}$$

Ahora la probabilidad condicional se leería:

La probabilidad de que a alguien no le guste el estofado, dado que si le guste el queso

$$P(\text{ESTOFADO}_{no} | \text{QUESO}_{si})$$

Es importante recordar que en ambos casos, solo hay un evento, y ambas probabilidades condicionales hacen referencia a la misma probabilidad

$\text{QUESO}_{si}$	$\text{ESTOFADO}_{si}$ $p = 2/14$	$\text{ESTOFADO}_{no}$ $p = 5/14 \leftarrow$	total queso $2 + 5 = 7$ $p = 7/14$
$\text{QUESO}_{no}$	$p = 4/14$	$p = 3/14$	$3 + 4 = 7$ $p = 7/14$
total estofado	$2 + 4 = 6$ $p = 6/14$	$5 + 3 = 8$ $p = 8/14$	

La única diferencia real entre estas dos probabilidades condicionales es la información previa sobre el evento.

Por lo cual el beneficio de utilizar la nomenclatura trabajada en el desarrollo es para hacer más fácil la lectura y mostrar que en ambos casos estamos hablando de la misma cosa.

## Sigamos trabajando con el Teorema de Bayes

Visto el punto anterior, ahora realicemos otro ejercicio para entender mejor el Teorema de Bayes, teniendo en cuenta que ya repasamos las probabilidades condicionales.

El ejemplo que usaremos es el siguiente:

Imaginemos que tenemos en frente a una mujer de 40 años o más que debe hacerse un estudio de mamografía. El cual es un estudio que deben hacerse todas las mujeres a los 40 años para poder detectar síntomas del cáncer de seno.

El test tiene el siguiente parámetro.

La mujer, al llegar a la clínica, el técnico le comenta que el dispositivo con el que se usara para detectar si hay cáncer, tiene una sensibilidad del 80%.

¿Pero qué quiere decir esto?

Primero definamos las variables de probabilidad que usaremos, diremos que la variable  $x$  es igual a que el test de positivo, y la variable  $y$  es igual a que la paciente tenga cáncer de seno

$$\begin{array}{ll} x = 1 = \text{test positivo} & y = 1 = \text{paciente con cáncer} \\ x = 0 = \text{test negativo} & y = 0 = \text{paciente sin cáncer} \end{array}$$

Definido lo anterior, ahora formulemos lo que se refirió el técnico sobre la sensibilidad del 80%:

La probabilidad de que el test de cáncer de positivo, dado que sabemos de antemano un paciente tiene cáncer es del 80%

$$p(x = 1, y = 1 | y = 1) = 0.8 = 80\%$$

Cuando hablamos de sensibilidad, nos referimos a la precisión del dispositivo para detectar cáncer, lo cual es probable que se haya empleado el dispositivo en un grupo de personas que ya sabían de antemano que tenían cáncer y usaron el dispositivo para ver que tan efectivo era para detectar la enfermedad, donde en toda la población que sé probó el dispositivo, solo pudo detectar el cáncer un 80% de las veces.

Siendo así tenemos la probabilidad de que un paciente tenga cáncer es del 0.004%

$$p(y = 1) = 0.004 = 0.4\%$$

Y añadimos los falsos positivos que surgieron en los estudios, donde teníamos pacientes que sabíamos que no tenían cáncer, pero el dispositivo dio el test positivo a cáncer.

$$p(x = 1 | y = 0) = 0.1 = 10\%$$

Volviendo a la fórmula general del Teorema de Bayes, agreguemos unos componentes.

Formula general de Bayes:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

teniendo la formula a mano, démosle nombres a las probabilidades que tenemos dentro del teorema

$$\overbrace{P(A|B)}^{\text{posterior}} = \frac{\overbrace{P(B|A)}^{\text{verosimilitud}} * \overbrace{P(A)}^{\text{prior}}}{\underbrace{P(B)}_{\text{evidencia}}}$$

Ahora vemos que tenemos nuevos nombres para cada probabilidad, tanto como la **verosimilitud**, **prior** y la **evidencia**.

$P(B|A)$ : la **verosimilitud** es probabilidad condicional de  $B$  dado  $A$ , siendo la probabilidad de que el test de positivo dado que el paciente tenga cáncer.

$P(A)$ : las probabilidades a **priori** son la probabilidad de que un paciente cualquiera tenga cáncer, **independientemente, no es reflejo de sí la persona sometida al test tenga cáncer o no (antes de saber el estado del paciente)**.

$P(B)$ : la **evidencia**, es la probabilidad de que el test de positivo a cáncer, esta probabilidad actuaría como evidencia de que el paciente tiene cáncer de seno.

[Para profundizar más sobre este tema](#)

Entonces decimos que:

- $y = A$  = paciente con cáncer
- $x = B$  = test positivo cáncer

y nuestras probabilidades quedarían tal que

$$\text{verosimilitud} \rightarrow p(x = 1, y = 1|y = 1) = 0.8 = 80\%$$

$$\text{priori} \rightarrow p(y = 1) = 0.004$$

$$\text{evidencia} \rightarrow p(x = 1) = ?$$

Vemos que no tenemos la evidencia para la fórmula general, así que resolvamos este faltante de información escribiendo la fórmula de Bayes con los datos del ejercicio

$$P(y = 1|x = 1) = \frac{p(x = 1|y = 1) * p(y = 1)}{p(x = 1)}$$

Nosotros buscamos resolver la probabilidad de que el test de positivo dado que el paciente tenga cáncer, multiplicado por la probabilidad de que un paciente tenga cáncer, dividido la probabilidad de que el test de positivo, todo esto resulta en la probabilidad que nos ayudara a la correcta interpretación de si debemos preocuparnos o no por el control de la paciente, que sería cuál es la probabilidad de que la paciente tenga cáncer, sabiendo que el test dio positivo ( $P(y = 1|x = 1)$ ).

Hay que recordar que aunque el examen de positivo existe un factor de error, por lo cual hay que considerar ambas opciones.

El único dato que nos falta para poder resolver este ejercicio es la **evidencia**  $p(x)$ , que se puede obtener de calcular una probabilidad conjunta, ya que  $p(x)$  es una **Probabilidad Marginal** como vimos en el tema 2

$$\underbrace{P(A)}_{\text{Marginal}} = \Sigma p(A, B)$$

$$P(x) = \Sigma p(x, y)$$

y recordemos a misma manera que una probabilidad conjunta es igual al producto de la probabilidad de  $x$  dado  $y$  por la probabilidad de  $y$

$$\underbrace{P(A, B)}_{\text{Regla del producto}} = P(A|B) * P(B)$$

$$\Sigma P(x, y) = P(x|y) * P(y)$$

Donde en este caso, en particular lo representaríamos como la probabilidad de que el test de positivo es igual a la sumatoria de la probabilidad de que el test de positivo, dado que el paciente tenga o no cáncer multiplicado, la probabilidad de que el paciente tenga o no cáncer.

$$P(x = 1) = \Sigma P(x = 1|y) * P(y)$$

Si te resulta confuso de que  $y$  no tenga su signo igual a 1 tranquilo, es porque la posibilidad de que el paciente tenga cáncer son dos (0 y 1), por lo tanto, tendríamos que realizar el producto con ambos estados en  $y = 0$  e  $y = 1$  para luego realizar la sumatoria de las probabilidades

$$p(x = 1) = P(x = 1|y = 1) * P(y = 1) + P(x = 1|y = 0) * P(y = 0)$$

y recordemos por una vez más nuestros valores para así reescribirlos en la fórmula

$$P(x = 1|y = 1) = 0.8$$

$$P(x = 1|y = 0) = 0.1$$

$$P(y = 1) = 0.004$$

$$P(y = 0) = 0.996 = P(y = 1) - 1$$

Ahora transcribamos los valores para hallar  $p(x = 1)$

$$p(x=1) = P(x=1|y=1) * P(y=1) + P(x=1|y=0) * P(y=0)$$

$$p(x=1) = 0.8 * 0.004 + 0.1 * 0.996$$

$$p(x=1) = 0.0032 + 0.0996$$

$$p(x=1) = 0.1028$$

Hallada la evidencia, ya tenemos todos los valores para poder calcular la probabilidad de que una paciente tenga cáncer dado que el test haya dado positivo

$$P(y=1|x=1) = \frac{P(x=1|y=1) * P(y=1)}{P(x=1)}$$

$$P(y=1|x=1) = \frac{0.8 * 0.004}{0.1028}$$

$$P(y=1|x=1) = 0.031 = 3.1\%$$

Podemos pensar que al escuchar que el técnico dijo que la sensibilidad de la máquina era de un 80% pensemos que si el test salía positivo, tenemos un 80% de tener cáncer, pero si no era su precisión a la hora de detectar si era positivo o negativo, y tras usar el Teorema de Bayes vemos que la probabilidad posteriori de tener cáncer dado que el test dio positivo es del 3.1%

[Video complementario](#)

## Naive Bayes en machine learning

¿Qué es un clasificador de Naive Bayes en machine learning?

Básicamente, es un problema donde tenemos que realizar una clasificación de los datapoints que tenemos en un dataset dado

Dataset:

$$X = \{x_1, x_2, x_3, \dots, x_n\} \quad \{y\}$$

donde usaremos la base del **clasificador Naive Bayes** será el **Teorema de Bayes**

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

¿Ahora que sabemos la fórmula de probabilidad, te preguntaras sobre, como lo aplicamos a un problema de clasificación?

Para el dataset de muestra tendríamos unos variables independientes  $x$  que podían ser características de una persona como altura, peso, edad y buscaríamos calcular la variable dependiente  $y$  que podría ser si esta persona es obesa o no.

$$\underbrace{X}_{\text{persona}} = \{\underbrace{x_1}_{\text{peso}}, \underbrace{x_2}_{\text{altura}}, \underbrace{x_3}_{\text{edad}}, \dots, x_n\} \quad \{ \underbrace{y}_{\text{obeso|si|no}} \}$$

Teniendo en cuenta este ejemplo particular, veamos como modificaríamos el Teorema de Bayes para usarlo en machine learning.

Recordemos que el posteriori del Teorema de Bayes es **la probabilidad de  $A$  dado  $B$** , por lo tanto,  $B$  ya es información con la que constamos, y podemos modificar la fórmula del Teorema a lo siguiente:

Sabemos que  $B$  es el  $X$ , ya que es el dataset que tenemos dado, y  $A$  es  $y$  que es la probabilidad que buscamos calcular

$$P(\underbrace{A|B}_{A \text{ dado } B}) = \frac{\underbrace{P(B|A)}_{\text{verosimilitud}} * \underbrace{P(A)}_{\text{prior}}}{\underbrace{P(B)}_{\text{evidencia}}}$$

Para transformar la **verosimilitud** sabemos que  $B$  es el  $X$ , por lo tanto, podemos escribir la probabilidad de  $x_1$  dado  $y$ , por la probabilidad de  $x_2$  dado  $y$ , y así por todos los  $x_n$  dado  $y$  que haya en el datapoint:

$$P(y|x_1, x_2, \dots, x_n) = \frac{\underbrace{P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)}_{\text{verosimilitud}} * P(A)}{P(B)}$$

Luego transformaremos las probabilidades **priori**, donde sabemos que es  $y$ :

$$P(y|x_1, x_2, \dots, x_n) = \frac{\underbrace{P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)}_{\text{verosimilitud}} * \underbrace{P(y)}_{\text{prior}}}{P(B)}$$

Solo para recordar, marquemos en la probabilidad posteriori quien es  $A$  y quien es  $B$ :

$$P(\underbrace{y|}_{\text{A}} \underbrace{x_1, x_2, \dots, x_n}_{\text{B}}) = \frac{\underbrace{P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)}_{\text{verosimilitud}} * \underbrace{P(y)}_{\text{prior}}}{P(B)}$$

Ahora transformemos la **evidencia**, donde sabemos que  $B$  es todo el conjunto de  $xn$  o  $X$ :

$$P(\underbrace{y|}_{\text{A}} \underbrace{x_1, x_2, \dots, x_n}_{\text{B}}) = \frac{\underbrace{P(x_1|y) * P(x_2|y) * P(x_3|y) * \dots * P(x_n|y)}_{\text{verosimilitud}} * \underbrace{P(y)}_{\text{prior}}}{\underbrace{P(x_1) * P(x_2) * P(x_3) * \dots * P(x_n)}_{\text{evidencia}}}$$

Abreviamos la verosimilitud que es la multiplicación de todo  $x_n$  dado  $y$ , por el producto de  $x_i$  dado  $y$

$$P(\overbrace{y}^A | \overbrace{x_1, x_2, \dots, x_n}^B) = \frac{\overbrace{P(y)}^{prior} * \overbrace{\prod_{i=1}^n P(x_i|y)}^{verosimilitud}}{\underbrace{P(x_1) * P(x_2) * P(x_3) * \dots * P(x_n)}_{evidencia}}$$

La evidencia la consideraremos como una constante, ya que será igual en cada registro que tengamos del datapoint, por lo tanto, la omitimos y decimos que la probabilidad de  $y$  dado  $x$  es directamente proporcional a la ecuación que desarrollamos:

$$P(\overbrace{y}^A | \overbrace{x_1, x_2, \dots, x_n}^B) \propto \overbrace{P(y)}^{prior} * \overbrace{\prod_{i=1}^n P(x_i|y)}^{verosimilitud}$$

Ahora, partiendo desde el punto en que ambas ecuaciones son directamente proporcionales, para encontrar el output  $y$  de  $\{x_1, x_2, x_3, \dots, x_n\}$ , calcularemos el máximo valor que tendrán las probabilidades de lo que estemos computando con el clasificador Naive Bayes

$$y = argmax_y \overbrace{P(y)}^{prior} * \overbrace{\prod_{i=1}^n P(x_i|y)}^{verosimilitud}$$

lo cual suponiendo que hicimos la probabilidad de que una persona sea obesa o no, primero nos computaría las probabilidades de **si es obesa** y luego las probabilidades de **no es obesa**, y el valor que tenga su probabilidad más alta sería la el resultado que nos devolvió el modelo de clasificación

ej 1:

$$y_{obeso} = 0.8$$

$$y_{no\ obeso} = 0.3$$

ej 2:

$$y_{obeso} = 0.24$$

$$y_{no\ obeso} = 0.57$$

[Video con ejemplo práctico](#)

[Explicación en video + ejemplo práctico](#)

## Conclusiones

¡Hemos estudiado tanto las probabilidades frecuentistas como bayesianas y ha sido un largo artículo por el cual recorrimos, vimos algunos ejemplos en código y una gran cantidad de teoría!

Cabe recalcar que el esquema bayesiano es muy importante en ciencias de datos y machine learning en general, ya que buscamos construir la base de nuestro conocimiento para en próximos artículos entrar a máxima profundidad sobre algoritmos de regresión lineal, clasificación y mucho mas.