# Resumen del capítulo: Teoría de la probabilidad

## Experimentos, resultados elementales y eventos

Un **experimento** es una prueba repetible en la que se produce uno de varios **resultados**. Un resultado puede estar compuesto por varios **resultados elementales** que, por definición, no pueden desglosarse más.

En las situaciones más sencillas todos los resultados elementales son igualmente probables. Podemos llamar a tal experimento un "experimento justo". En un experimento justo con *n* resultados elementales, la probabilidad de cada resultado es 1/*n*.

El conjunto de todos los posibles resultados elementales de un experimento se llama **espacio muestral**. Puede seleccionar un subconjunto que contenga varios resultados elementales. Esto se llama un **evento**.

Un **evento imposible** es un evento que nunca puede suceder por lo que la probabilidad de que ocurra es 0. Un **evento seguro** es un evento que definitivamente ocurrirá por lo que su probabilidad es igual a 1. La probabilidad de otros eventos está entre 0 y 1.

Siempre que mantengas la condición de que todos los resultados elementales tengan la misma probabilidad, la **probabilidad del evento** será el número de resultados elementales en el evento dividido entre el número total de resultados (es decir, el tamaño del espacio de muestra). De manera más general (incluso cuando los resultados elementales no son igualmente probables), la probabilidad de un evento será equivalente a la suma de las probabilidades de sus resultados elementales constituyentes.

## La ley de los grandes números

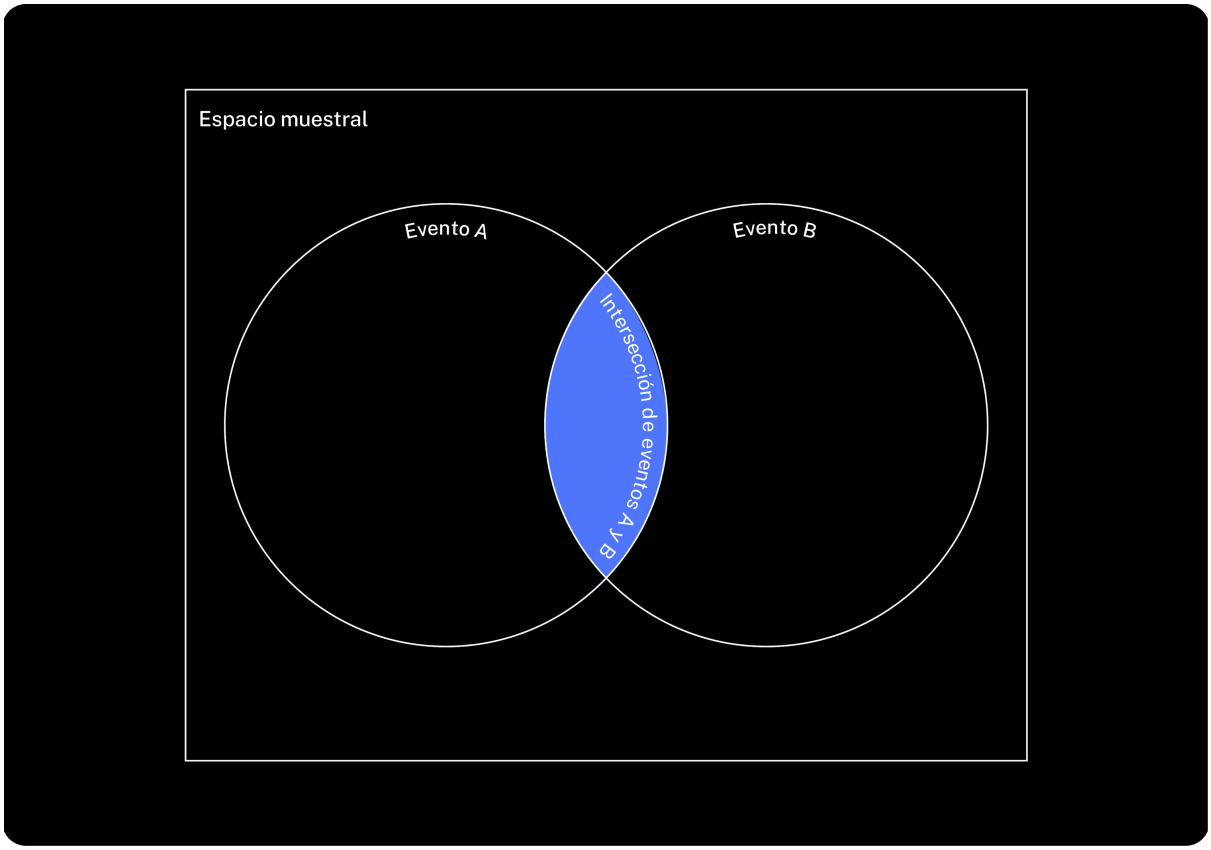
La **ley de los grandes números** dice que cuantas más veces repitas un experimento, más se acercará la frecuencia de un evento dado a su probabilidad.

Podemos usar esta regla a la inversa. Si no conocemos la probabilidad de un evento, pero podemos repetir el experimento muchas veces, podemos estimar su

probabilidad a partir de la frecuencia de los resultados.

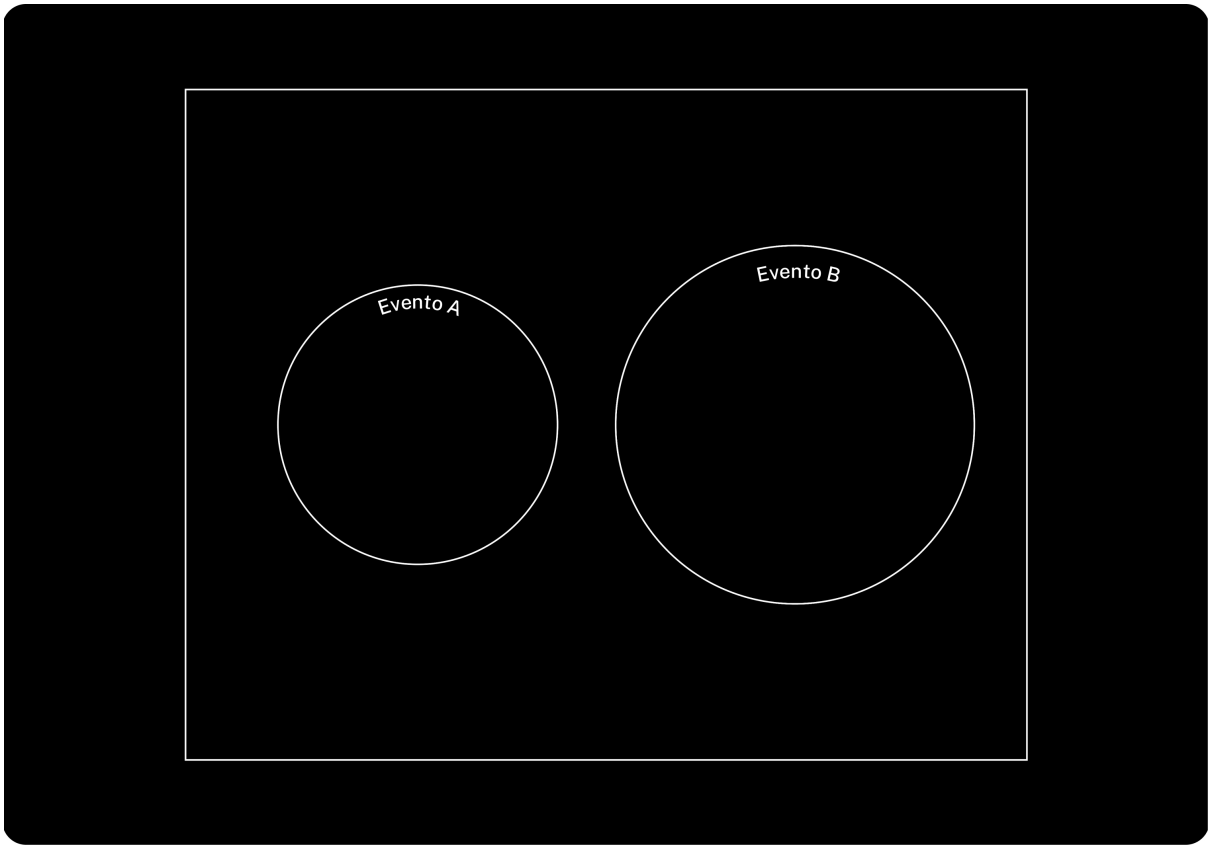
## Eventos mutuamente excluyentes e independientes; Multiplicar probabilidades

Para ilustrar la intersección entre eventos, puedes usar un **diagrama de Venn**:




Si los eventos A y B se cruzan, significa que hay resultados elementales que ocurren tanto en A como en B.

Los eventos que no pueden ocurrir simultáneamente en el mismo experimento se denominan **mutuamente excluyentes**: su diagrama de Venn no muestra ninguna intersección:




La probabilidad de que ambos eventos mutuamente excluyentes ocurran es cero.

Los eventos se denominan **independientes** si la ocurrencia de uno no afecta la probabilidad del otro. Si los eventos son independientes, entonces la probabilidad de su intersección es igual al producto de sus probabilidades de ocurrencia. Esta regla funciona en sentido inverso.

Si los eventos mutuamente excluyentes ocupan todo el espacio muestral, la suma de sus probabilidades será 1.

Puedes saber si los eventos son mutuamente excluyentes a partir de un diagrama de Venn. No es tan fácil ilustrar la independencia; necesitas comprobar si el producto de las probabilidades de los eventos es igual a la probabilidad de la intersección de estos eventos.

## Números aleatorios, distribuciones de probabilidad e intervalos de valores

Una **variable aleatoria** es una variable que toma un **valor aleatorio** que no se puede predecir antes de que se lleve a cabo el experimento. Los experimentos tienen resultados que pueden describirse tanto cuantitativa como cualitativamente. ¿Alguna vez la pitón logró salir del laberinto? ¿Llovió? ¿De qué color cayó la bola de la ruleta? Un número aleatorio se define numéricamente en función de estos resultados. Es una forma de proyectar los resultados del experimento, sin importar cómo se hayan definido, en la recta numérica.

Al igual que todas las variables cuantitativas, las variables aleatorias pueden ser **discretas** o **continuas**.

La **distribución de probabilidad** de una variable aleatoria se puede mostrar en una tabla que contenga todos los valores posibles de la variable y la probabilidad de que ocurran.

Utiliza el tipo de datos **matriz** de la librería NumPy para almacenar tablas numéricas:

```
table = np.array([[2,3,4,5,6,7],
[3,4,5,6,7,8],
...
[7,8,9,10,11,12]])
```

Si necesitas obtener una lista de todas las claves de la librería cuando trabajas con una librería, usa el método `keys()`. Obtén una lista de todos los valores usando el método `values()`:

```
dictionary = {...}
print(dictionary.keys())
print(dictionary.values())
```

## Valor esperado y varianza

Puedes definir una variable aleatoria para un experimento y encontrar un valor numérico hacia el cual tenderá en múltiples repeticiones del experimento. Este valor es conocido como el **valor esperado** de la variable aleatoria.

Si el experimento consiste en resultados elementales igualmente probables que se definen numéricamente, el valor esperado será igual al promedio de los valores

posibles.

El **valor esperado** de una variable aleatoria (*X*) es la suma de todos los valores de la variable aleatoria (representados con una *x* minúscula) multiplicada por sus probabilidades:

$$E(X) = \sum p(x_i)x_i$$

El valor esperado es como una medida de posición, pero para variables aleatorias en lugar de datasets. Te dice en qué valor se distribuye la variable aleatoria y, según la ley de los grandes números, hacia qué valor tenderá cuando se repita el experimento.

Dado que la variable aleatoria se distribuye en torno a esta "medida de posición", puedes determinar cuál es su varianza. Para hacerlo, necesitas encontrar el valor esperado del cuadrado de la variable aleatoria. No es difícil dado que los valores cambian, pero su probabilidad no.

Ya que conocemos el valor esperado de la variable aleatoria y su cuadrado, la **varianza** se puede encontrar usando la fórmula:

$$Var(X) = E(X^2) - (E(X))^2$$

## Probabilidad de éxito en experimentos binomiales

Los experimentos con dos resultados posibles se conocen como **experimentos binomiales**. Por lo general, aunque no siempre, uno de los resultados se llama "éxito" y el otro "fracaso". Si la probabilidad de éxito es *p*, la probabilidad de fracaso es 1 - *p*, ya que la suma de las probabilidades de los resultados debe ser igual a 1.

## La distribución binomial

El número de formas de obtener *k* éxitos a partir de *n* repeticiones de un experimento se puede encontrar utilizando la fórmula:

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

donde ! (leído como "factorial") es igual al producto de números naturales desde 1 hasta el número dado: n! = 1 · 2 · 3 · 4 · … · (n-1) · n.

Puedes calcular el factorial usando la librería de matemáticas y su método `factorial`:

```
from math import factorial
x = factorial(5)
```

Repasemos el experimento binomial, (donde un experimento con dos resultados se repite *n* veces). Si la probabilidad de éxito es *p* y la de fracaso es 1 - *p* y el experimento se repite *n* veces, entonces la probabilidad de obtener *k* éxitos después de *n* intentos es:

$$C_n^k p^k (1-p)^{n-k}$$

Aquí están las condiciones que nos permiten confirmar que la variable aleatoria se distribuye de forma binomial:

- Se realiza un número fijo y finito de intentos (*n*)
- Cada intento es un simple experimento binomial con exactamente dos resultados
- Los intentos son independientes entre sí
- La probabilidad de éxito (*p*) es la misma para todos los *n* intentos

## La distribución normal

El **teorema del límite central** es un teorema clave en estadística. En términos algo simplificados, establece que "Muchas variables aleatorias independientes, sumadas, dan una distribución normal".

La distribución normal describe valores continuos reales. Cuenta con dos parámetros, media y varianza:

$X \sim \mathbb{N}(\mu, \sigma 2)$

Esta notación se puede leer como: La variable *X* tiene una distribución normal con una media de *mu* (μ) y una varianza de *sigma al cuadrado* ($\sigma^2$) (que corresponde a una desviación estándar de *sigma*).

Para encontrar la probabilidad de que ocurra cualquier intervalo determinado a partir de parámetros de distribución conocidos, llamamos a dos métodos del paquete scipy.stats: **norm.ppf** y **norm.cdf.**

- `ppf`: percent point function (función de punto porcentual).
- `cdf`: cumulative distribution function (función de distribución acumulada).

Ambos funcionan con la distribución normal, dada una media particular (valor esperado) y una desviación estándar.

- `norm.ppf` proporciona el *valor* de una variable cuando se conoce la probabilidad del intervalo a la izquierda de ese valor.
- `norm.cdf`, por otro lado, proporciona la *probabilidad* del intervalo a la izquierda del valor cuando este valor es conocido.

Calcula la distribución normal usando el método `norm()` del paquete scipy.stats con dos argumentos: valor esperado y desviación estándar. Vamos a encontrar la probabilidad de obtener un valor particular, *x*:

```
from scipy import stats as st

# establece una distribución normal
distr = st.norm(1000, 100)

x = 1000

result = distr.cdf(x) # calcula la probabilidad de obtener el valor x
```

Utilizando la función `norm.cdf` podemos calcular la probabilidad de obtener un valor en el intervalo entre *x1* y *x2*:

```
from scipy import stats as st

# establece una distribución normal
distr = st.norm(1000, 100)

x1 = 900
x2 = 1100

result = distr.cdf(x2) - distr.cdf(x1)
# calcula la probabilidad de obtener un valor entre x1 y x2
```

Para encontrar un valor que tiene una cierta probabilidad, usamos el método `norm.ppf`:

```
from scipy import stats as st

# establece una distribución normal
distr = st.norm(1000, 100)

p1 = 0.841344746

result = distr.ppf(p1)
```

## Aproximación normal a la distribución binomial

Con un gran número de repeticiones de un experimento binomial, la distribución binomial se aproxima a la distribución normal.

Para una distribución binomial discreta, dado un número de intentos $n$ y una probabilidad de éxito $p$, el valor esperado es igual a $n \cdot p$ y la varianza es $n \cdot p \cdot (1-p)$ .

Si $n$ es superior a 50, estos parámetros de distribución binomial pueden tomarse como la media y la varianza de una distribución normal bastante cercana a la binomial. La distribución normal será la más cercana a la binomial cuando el valor esperado tenga $n \cdot p$ como el valor de la media y $n \cdot p \cdot (1-p)$ como la varianza.