



# BOURBAKI

COLEGIO DE MATEMÁTICAS

# Índice

01. Introducción \_\_\_\_\_ pág. 02

02. Lectura de referencia: la teoría de los  
valores extremos y su versión del teo-  
rema límite central \_\_\_\_\_ pág. 04

03. Distribuciones infinitas \_\_\_\_\_ pág. 05

04. Teorema del límite central e intervalos  
de confianza \_\_\_\_\_ pág. 08

    01. Operaciones con variables aleatorias

        pág. 08

    02. Teorema del límite central \_\_\_\_\_ pág. 09

    03. Intervalos de confianza en sondeos \_\_\_\_\_

        pág. 10

# 01 Introducción

Bienvenidos a nuestro curso de Matemáticas Avanzadas para la Ciencia de Datos, nuestro curso tiene tres módulos dedicados a estudiar las ideas matemáticas más útiles para comprender los algoritmos y modelos matemáticos más comunes en Ciencia de Datos. Los tres módulos son los siguientes

- Probabilidad y Estadística
- Álgebra Lineal
- Optimización y cálculo diferencial

Todos los módulos tienen una duración de 8 semanas. El curso está acompañado de ejercicios y tareas en Python para practicar y reforzar los conocimientos aprendidos así como las implementaciones en bases de datos de los algoritmos estudiados. Pueden consultar el repositorio de esta semana en [este link](#).

La estructura de cada una de las semanas es la siguiente:

1. Treinta minutos dedicados a estudiar un artículo de referencia que motivará los conceptos matemáticos de esta semana.
2. Dos horas dedicadas a estudiar el tema de la semana y algunos ejercicios.

3. Una hora y media dedicada a practicar lo aprendido utilizando Python.

El primer módulo de probabilidad consta de los siguientes temas:

1. Kolmogorov, independencia y condicionamiento
2. Variables aleatorias discretas y sus momentos
3. Ley de los grandes números y máxima verosimilitud
4. El teorema límite central y los intervalos de confianza
5. Tests estadísticos
6. Inferencia bayesiana
7. Cadenas de markov y muestreros de gibbs
8. Redes bayesianas

## 02 Lectura de referencia: la teoría de los valores extremos y su versión del teorema límite central

Esta semana nos dedicaremos a estudiar los detalles y las aplicaciones del teorema límite central. La versión moderna de este teorema viene de principios del siglo XX, al lector interesado le recomendamos leer más sobre la historia de este teorema en [1]. En particular, uno de los textos que a pesar de su edad se conservan con excelente frescura es el orginal de P. Lévy [2] quien es uno de los matemáticos a los que le debemos este descubrimiento junto a Lindenberg y Lyapunov.

Ya hemos hablado de las distribucioens que contienen más valores extremos de lo habitual como por ejemplo la distribución de Laplace. Como lectura de referencia le recomendamos a los estudiantes leer sobre la ver-sión para valores extremos del teorema límite central que estudiaremos esta semana. El nombre de este teorema es en honor a los matemáticos Fisher-Tippett-Gnedenko y en lugar de hablar sobre la distribución normal lo hará sobre as distribuciones de Weibull, Gumbel y Fréchet. En particular recomendamos a los estudiantes leer una tesis de maestría so-bre algunas aplicaciones al problema de las inundaciones en el país vasco [3].

## 03 Distribuciones infinitas

Hasta ahora solo hemos hecho énfasis de espacios de probabilidad finitos sin embargo las variables aleatorias son mucho más expresivas para espacios de probabilidad infinitos, en esta sección nos dedicaremos a estudiar variables aleatorias un poco más complicadas.

**Example 00.1.** Si  $E$  es la experiencia aleatoria de lanzar un dado justo hasta obtener el número seis, definimos la siguiente variable aleatoria:  $X(\omega_1, \omega_2, \dots) =$

$$\inf_{j \geq 1} \{j : \omega_j = 6\}$$

**Example 00.2.** Sea  $\lambda > 0$ , definamos  $\mathbb{P}_{Poisson, \lambda}(i) = e^{-\lambda} \frac{\lambda^i}{i!}$ , ese es un ejemplo de una experiencia aleatoria numerable, llamada ley de Poisson.

**Remark 00.3.** La ley anterior corresponde a la probabilidad de que un evento raro ocurra después de muchas repeticiones. La justificación matemática de esta intuición es la siguiente: si  $\lim_{n \rightarrow \infty} n \cdot p_n = \lambda$  entonces  $\lim_{n \rightarrow \infty} \mathbb{P}_{Bin, p_n, n}(i) = \mathbb{P}_{Poisson, \lambda}(i)$ .

**Exercise 00.4.** Comparemos la distribución de Poisson con la ley binomial que calculamos en el ejemplo de las olimpiadas en las primeras notas, en este caso nuestro parámetro  $\lambda$  será igual a la esperanza de la variable aleatoria de Bernouilli, lo cuál corresponde con 1.535 gracias al cálculo de las primeras notas, de esa forma  $\mathbb{P}_{Poisson, 1.535}(0) = 0.215$ ,  $\mathbb{P}_{Poisson, 1.535}(1) = 0.33$ ,  $\mathbb{P}_{Poisson, 1.535}(2) = 0.253$ , y  $\mathbb{P}_{Poisson, 1.535}(3) = 0.129$ .

Notemos que si en lugar de considerar la variable aleatoria  $S_n$  consideramos la variable aleatoria  $n - S_n$  (o equivalentemente la variable aleatoria de Poisson con  $\lambda = n(1-p)$ ) es posible aproximar de la misma manera eventos altamente probables, una pregunta inmediata es qué pasa si deseamos calcular eventos cuya probabilidad no se ni muy pequeña ni muy alta, para ello es necesario utilizar el célebre Teorema Límite Central de Lévy.

## 00.1 Leyes de probabilidad continuas

---

Las leyes de probabilidad continuas (es decir definidas sobre el conjunto total de los números reales) son más complicadas de definir porque en ese caso las funciones de probabilidad no actúan sobre la familia total de subconjuntos, si lo hicieran esto generaría algunos problemas matemáticos los cuales trascienden el objetivo de este curso. En general solo hablaremos de las llamadas leyes continuas con densidad.

**Definition 00.1.** La ley de probabilidad uniforme sobre el conjunto  $[-1, 1]$  es la ley de probabilidad tal que  $\mathbb{P}_{unif}((\epsilon_1, \epsilon_2)) = \frac{1}{|\epsilon_1 - \epsilon_2|}$

**Definition 00.2.** La ley de probabilidad Gaussiana o normal con parámetros  $(\mu, \sigma^2)$  se define para los intervalos  $(-\infty, x]$  de la siguiente manera:

$$\mathbb{P}_{Gauss, \mu, \sigma^2}(-\infty, x] = \frac{1}{\sigma \cdot \sqrt{2\pi}} \int_{-\infty}^x \left( \exp \left( -\frac{(t - \mu)^2}{2 \cdot \sigma^2} \right) \right) dt$$

En estas notas no hemos definido la esperanza ni la covarianza para leyes de probabilidad no numerables, sin embargo es posible hacerlo:

**Proposition 00.5.** *Si  $X$  es una variable aleatoria tal que  $\mathbb{P}_X = \mathbb{P}_{Gauss,\mu,\sigma}$  entonces  $\mathbb{E}(X) = \mu$  y  $Var(X) = \sigma^2$ .*

**Exercise 00.6.** *(Distribución gaussiana multi-variada) Sea  $\mu \in \mathbb{R}^d$  y  $S \in \mathbb{R}^{d \times d}$  tal que  $S = S^T$  y  $xSx^T > 0$  para cualquier  $x \in \mathbb{R}^d \setminus \{\vec{0}\}$ . Definimos la ley de probabilidad gaussiana  $d$  dimensional con parámetros  $\mu, S$  de la siguiente forma:*

$$\mathbb{P}_{Gauss,\mu,S}((-\infty, x_1] \times \dots \times (-\infty, x_d]) = \frac{1}{(2\pi)^{d/2} |S|^{d/2}} \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_d} e^{-\frac{1}{2}(t-\mu)S^{-1}(t-\mu)^T} dt$$

# 04 Teorema del límite central e intervalos de confianza

## Operaciones con variables aleatorias

---

Una pregunta natural es en qué medida estas operaciones respetan propiedades más complicadas tales como la suma, la multiplicación o la conjunción. En esta sección enunciaremos algunas propiedades y contra-ejemplos sin demostración sobre las variables aleatorias.

**Proposition 01.1.** *Sean  $X, Y$  dos variables aleatorias independientes, sea  $Z = X + Y$  y  $W = (X, Y)$*

- *Si  $X, Y$  son gaussianas con distribuciones*

$$\mathbf{N}(\mu, \sigma^2), \mathbf{N}(\mu', \sigma'^2)$$

*Entonces  $Z$  tiene una ley de probabilidad  $\mathbf{N}(\mu + \mu', \sigma^2 + \sigma'^2)$ . También  $W$  sigue una distribución gaussiana.*

- *Si  $X, Y$  siguen una distribución de Poisson con parámetros  $\lambda, \lambda'$  entonces  $Z$  sigue una distribución de Poisson con parámetro  $\lambda + \lambda'$*

- *Si  $X, Y$  siguen distribuciones Binomiales  $\mathbb{P}_{Bin,n,p}, \mathbb{P}_{Bin,m,p}$ , entonces  $Z$  sigue una distribución binomial  $\mathbb{P}_{Bin,n+m,p}$*

**Example 01.2.** Sea  $X$  una variable aleatoria gaussiana, digamos  $\mathbf{N}(0, 1)$

- Si  $X'$  es una variable aleatoria tal que  $\mathbb{P}(X' = 1) = \frac{1}{2}, \mathbb{P}(X' = -1) = \frac{1}{2}$  y

definimos  $Y = X \cdot X'$ , entonces  $X + X'$  no es gaussiana.

- Si  $X'$  es una variable aleatoria definida igual a  $X$  cuando  $|X| \leq 1$  e igual a  $-X$  en el caso opuesto. Entonces  $(X, X')$  no sigue una distribución gaussiana multi-variada.

## Teorema del límite central

---

En el ejemplo 00.2 pudimos calcular el límite de una ley binomial cuando  $p_n$  se acercaba a cero a medida que  $n$  crece, a continuación vamos a enunciar el Teorema del límite central el cuál es una generalización de tal observación para eventos que no necesariamente son raros.

Una manera de motivarlo es recordando la ley de los grandes números. Si tenemos  $X_1, X_2, \dots$  variables aleatorias i.i.d, sabemos que  $\frac{X_1 + \dots + X_n}{n}$  y  $\mathbb{E}(X_1)$  se encuentran cerca. En matemáticas es importante conocer la velocidad con la que dos funciones se acercan entre sí. La manera en la que esto se hace es buscar convergencia cuando la multiplicamos por otra cantidad que nosotros sabemos si converge rápidamente o no, en este caso nos gustaría conocer si la siguiente familia de variables aleatorias converge para algún valor de  $\alpha$ :

$$n^\alpha \left( \frac{X_1 + \dots + X_n}{n} - \mathbb{E}(X_1) \right)$$

Eso significaría que  $\alpha$  será la velocidad de la convergencia del promedio usual a la esperanza. Desafortunadamente esta familia de variables aleatorias no converge hacia ninguna otra variable aleatoria para ningún valor de  $\alpha$ . Por eso es necesario introducir una nueva noción de convergencia más débil, la cual no significa que las variables aleatorias estén cerca sino que sus leyes de probabilidad lo están. El  $\alpha$  adecuado es  $\alpha = \frac{1}{2}$ .

**Theorem 02.1.** *Sean  $X_n$  es una familia de variables aleatorias independientes*

*e idénticamente distribuidas tales que  $\mathbb{E}(X_i) = \mu$ ,  $\text{Var}(X_i) = \sigma^2$  para todo  $i$ .*

*Definamos  $M_n = \frac{\sqrt{n} \left( \frac{(X_1 + X_2 + \dots + X_n)}{n} - \mu \right)}{\sigma}$  entonces para todos  $a < b$*

$$\lim_{n \rightarrow \infty} \mathbb{P}_{M_n}(a, b) = \mathbb{P}_{\text{Gauss}, 0, 1}(a, b)$$

## Intervalos de confianza en sondeos

---

Supongamos que tenemos estamos estudiando los resultados de una elección presidencial donde hay dos candidatos:  $A, B$ . Supongamos que un sondeo le pregunta a 2500 individuos sus intenciones de voto. Supondremos que evitamos preguntar a la misma persona dos veces, que el

resultado de una respuesta no influye en el resto y además que cualquier persona tiene la misma probabilidad de ser encuestado. Eso significa que estamos definiendo una familia de 2500 variables aleatorias independientes e idénticamente distribuidas  $X_i : \{A, B\} \rightarrow \{0, 1\}$  de la siguiente manera  $X(A) = 0, X(B) = 1$ . Notemos que como tenemos dos candidatos, en realidad estamos suponiendo una Ley de Bernoulli usual i.e.  $\mathbb{P}_{X_i}(1) = p, \mathbb{P}_{X_i}(0) = 1 - p$ . El objetivo de un tal sondeo es conocer el valor  $p$ .

**Exercise 03.1.** En la situación anterior demostrar que si llamamos

$$\bar{X}_n = \left( \frac{(X_1 + X_2 + \dots + X_n)}{n} \right)$$

entonces  $\mathbb{E}(\bar{X}_n) = p$ .

Gracias al ejercicio anterior vamos a utilizar  $\bar{X}_n$  y a aplicar tanto la ley de los grandes números como el teorema del límite central para calcular  $p$ .

Definimos  $\bar{X}_n(B) := p_n$ . Por ejemplo si 1300 individuos votaron por  $B$  entonces  $p_n$  será igual a .52. La pregunta que nos planteamos a continuación es: una vez hecha esta predicción cómo podemos saber si es buena? Recorriendo que el teorema del límite central nos dice la velocidad con la que una aproximación se acerca a la esperanza (en este caso a  $p$ ) es posible saber si es una mala aproximación.

**Exercise 03.2.** Convencerte con algunos casos particulares de que  $|a - b| < c$

es equivalente a  $(-c + b) < a < (c + b)$  y también a  $a - c < b < a + c$ .

Buscamos calcular lo siguiente:

$$\mathbb{P}(|\bar{X}_n - p| < \epsilon)$$

En lo particular nos gustaría que esta probabilidad sea alta: digamos .95 por ejemplo. Equivalentemente al cálculo anterior estamos interesados por la siguiente cantidad:

$$\mathbb{P}(\sqrt{n}|\bar{X}_n - p| < \sqrt{n}\epsilon) = \mathbb{P}\left(\frac{\sqrt{n}|\bar{X}_n - p|}{\sqrt{p(1-p)}} < \frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}\right)$$

El teorema central límite dice que

$$\mathbb{P}(|\bar{X}_n - p| < \epsilon) = \int_{-\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}}^{\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}} \exp\left(-\frac{x^2}{2}\right) dx$$

Ahora es necesario encontrar un valor de  $\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}}$  tal que la cantidad anterior sea igual a .95. Ello se puede hacer con una tabla de valores para la ley Gaussiana por ejemplo.

Si queremos que  $\mathbb{P}(|\bar{X}_n - p| < \epsilon) = .95$  entonces debemos de igualar  $\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}} = \frac{1.96 \cdot \sqrt{p(1-p)}}{\sqrt{n}}$

Desafortunadamente la cantidad anterior depende de  $p$ , sin embargo no-

temos que  $\sqrt{p(1-p)} \leq \frac{1}{2}$ . Por tanto es necesario que  $\epsilon \leq \frac{1.96 \cdot \sqrt{p(1-p)}}{\sqrt{n}} \leq$

$\frac{1.96}{2\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ . Resumimos el razonamiento anterior en el siguiente teorema:

**Theorem 03.3.** *Bajo las condiciones del sondeo descritas anteriormente, supongamos que  $n$  es tan grande que  $n \geq 50$ ,  $np \geq 5$ ,  $n(1-p) \geq 5$ . Podemos estar seguros con probabilidad del 95 porciento que la cantidad  $p$  está en  $[p_n - \frac{1}{n}, p_n + \frac{1}{n}]$ .*

**Exercise 03.4.** *Encontrar el intervalo de confianza correcto en el caso del sondeo de 2500 personas planteado en este párrafo.*

B O U R B A K I

COLEGIO DE MATEMÁTICAS

# Bibliografía

- [1] Hans Fischer. A History of the Central Limit Theorem. A History of the Central Limit Theorem, 2011.
- [2] Paul Levy. Théorie des erreurs. la loi de gauss et les lois exceptionnelles. Bulletin de la Société Mathématique de France, 52:49–85, 1924.
- [3] PabloSeñas Peón. Statistical Extreme Value Theory. Application to basins of the Basque Country (Teoría estadística de valores extremos. Aplicación a cuencas fluviales del País Vasco). PhD thesis, Universidad de Cantabria, 2020.



[escuela-bourbaki.com](http://escuela-bourbaki.com)