# BOURBAKI

COLEGIO DE MATEMÁTICAS

BOURBAKI
COLEGIO DE MATEMÁTICAS

# Índice

# ÍNDICE

# 01 Introducción

El siguiente texto es la bitácora de un curso que impartimos en alianza con el Instituto Humai. El curso es una introducción a la probabilidad y estadística utilizada en machine learning. El contenido de este curso fue acompañado por ejercicios pertinentes en Python y enseñados por Ana Isabel Ascencio.

# 02 La teoría de la probabilidad

La teoría de la probabilidad es una de las áreas de las matemáticas que más tiempo tardó en formalizarse. A pesar de ser estudiada desde hace varios siglos no fue axiomatizada hasta el siglo XX, esto dio lugar por ejemplo al nacimiento de la estadística y se convirtieron así en áreas fundamentales para las matemáticas.

## Lenguaje

---

**Definition 01.1.** *Una experiencia aleatoria (o estocástica) es una experiencia $E$ tal que reproducida bajo condiciones idénticas puede conducir a resultados distintos. Lo contrario a un evento aleatorio es un evento determinista.*

**Exercise 01.2.** *De los siguientes eventos decidir cuáles de ellos son aleatorios y cuáles de ellos son deterministas:*

1. *Lanzar un dado tres veces*
2. *Lanzar un dado hasta que obtengamos el número seis*
3. *Sumar dos números enteros*
4. *Una base de datos con una sola columna binaria*

5. *Romeo le promete a Julieta llegar a las 21 horas, que Romeo cumpla su promesa*
6. *El precio de un activo financiero*
7. *El color de los ojos de una persona*
8. *El resultado de una regresión lineal*
9. *La hora a la que pasa un autobús*
10. *Elegir una palabra al abrir un libro*

**Definition 01.3.** *El espacio de una experiencia aleatoria $E$ es el conjunto de valores que puede tomar aquella experiencia como resultado, se denotará a lo largo del curso por $\Omega$.*

**Example 01.4.**

- *El espacio de la experiencia aleatoria 'lanzar un dado' es el conjunto que tiene los números del 1 al 6 (las caras del dado).*
- *El espacio aleatorio de la experiencia 'lanzar un dado dos veces' es el conjunto de parejas de números entre el 1 y el 6:* $\{(1,1),(1,2),(2,1),\ldots,(5,6),(6,5),(6,6)\}$.
- *El espacio asociado a elegir una palabra al abrir un libro es el conjunto de todas las palabras que aparecen en ese libro.*

**Exercise 01.5.** *De las experiencias aleatorias identificadas en el ejercicio 01.2, determinar el espacio de cada una.*

**Definition 01.6.** *Fija una experiencia aleatoria $E$ junto a su espacio $\Omega$, llamaremos un evento aleatorio a un subconjunto $A \subseteq \Omega$.*

**Example 01.7.** *Si nuestra experiencia aleatoria es lanzar dos dados, un ejemplo de un posible evento aleatorio es: la suma de los dos resultados después de lanzar un dado es menor o igual a 4, este evento visto como un subconjunto es* $A = \{(1,1),(1,2),(2,1),(1,3),(3,1),(2,2)\}$.

**Exercise 01.8.** *Enunciar por lo menos dos ejemplos de eventos aleatorios asociados a las experiencias descritas en el ejercicio 01.2 y describir sus elementos explicitamente.*

## Los axiomas de Kolmogorov

Hasta el momento solo hemos hablado de eventos aleatorios sin discribirlos numéricamente, es decir sin introducir el concepto fundamental de medida de probabilidad. Una medida de probabilidad (o distribución o función de probabilidad) es justamente una manera de cuantificar la incertidumbre. Aunque muchas ideas fueron introducidas antes, A. Kolmogorov fue la primera persona en introducir un formalismo para cuantificar la incertidumbre utilizando matemáticas rigurosas.

**Definition 02.1.** *Si* $\Omega$ *es un conjunto y* $A, B \subseteq \Omega$ *son dos subconjuntos, denotaremos por:*

1. $A \cup B$ *a la unión entre* $A$ *y* $B$.
2. $A \cap B$ *a la intersección entre* $A$ *y* $B$.

3. *$A^c$ al complemento de $A$.*

**Exercise 02.2.** *Describir con palabras aquellos eventos aleatorios que corresponden a cada uno de los descritos por $A \cup B$, $A \cap B$ y $A^c$ donde $A, B$ son los eventos elegidos en el ejercicio 01.5.*

**Definition 02.3.** *Fijemos un conjunto finito $\Omega$, una ley de probabilidad es una asignación numérica $\mathbb{P}$ a los eventos aleatorios tal que si $A, B \subseteq \Omega$:*

1. $0 \leq \mathbb{P}(A) \leq 1 = \mathbb{P}(\Omega)$
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, *si $A \cap B = \emptyset$.*

*A la pareja $(\Omega, \mathbb{P})$ se le conoce como un espacio de probabilidad.*

**Example 02.4.** *(Ley uniforme) Supongamos que $\Omega$ es un conjunto de tamaño $N$. Definamos la siguiente ley de probabilidad $\mathbb{P}_{unif}(A) = \frac{i}{N}$ donde $i$ es la cantidad de elementos en $A$.*

**Example 02.5.** *(Ley de Bernoulli) Sea $\Omega = \{\omega_0, \omega_1\}$ y $0 \leq p \leq 1$, definimos la ley de probabilidad de Bernoulli de la siguiente manera $\mathbb{P}(\omega_1) = p$ y $\mathbb{P}(\omega_0) = 1 - p$.*

**Example 02.6.** *(Producto de Bernoulli) Sea $\Omega = \{\omega_0, \omega_1\}^n$ y $0 \leq p \leq 1$, definimos la ley de probabilidad producto de Bernoulli de la siguiente manera $\mathbb{P}((v_1, \ldots, v_n)) = p^i (1-p)^{n-i}$ cuando el elemento $(v_1, \ldots, v_n)$ tiene $i \leq n$ entradas iguales a $\omega_1$ y $n - i$ entradas iguales a $\omega_0$.*

**Example 02.7.** *(Ley binomial) Sea* $\Omega$ *un conjunto de tamaño* $n$*, digamos* $\Omega = \{1,2,\ldots,n\}$ *y sea* $0 \le p \le 1$ *un número arbitrario. Definimos la ley de probabilidad siguiente:*

$$\mathbb{P}_{Bin,n,p}(i) = \left(\frac{n!}{(n-i)!i!}\right) p^i (1-p)^{n-i}$$

.

*Esta ley de probabilidad es llamada binomial con parámetros* $n, p$.

**Example 02.8.** *Supongamos que la probabilidad* $p$ *de ganar una medalla de oro en unas olimpiadas es el tamaño de su población dividida sobre la población mundial. Por ejemplo durante las olimpiadas del 2016 la población en Argentina era alrededor de 43 millones de personas mientras que la población mundial era de 7400 millones de personas, entonces* $p = 0.005$*, claramente una ley de Bernoulli es suficiente modelar este fenómeno. La ley Binomial nos es útil para calcular la probabilidad de que Argentina gane* $i$ *medallas de oro si suponemos que los 307 eventos son todos independientes entre ellos (más adelante hablaremos sobre este importante concepto). Entonces* $\mathbb{P}_{Bin,307,0.005}(0) = \binom{307}{0}(.005)^0(.995)^{307} = 0.214$*,*$\mathbb{P}_{Bin,307,0.005}(1) = \binom{307}{1}(.005)^1(.995)^{306} = 0.331$*,* $\mathbb{P}_{Bin,307,0.005}(2) = \binom{307}{2}(.005)^2(.995)^{305} = 0.254$ *,*$\mathbb{P}_{Bin,307,0.005}(3) = \binom{307}{3}(.005)^3(.995)^{304} = 0.013$ *y* $\mathbb{P}_{Bin,307,0.005}(i \le 3) = 0.214 + 0.331 + 0.254 + 0.013 = 0.812$.

El siguiente resultado simplifica la comprensión del concepto de ley de probabilidad:

**Proposition 02.9.** *Sea* $(\Omega, \mathbb{P})$ *un espacio de probabilidad, digamos* $\Omega = \{\omega_1, \omega_2, \ldots, \omega_m\}$*, la ley de probabilidad* $\mathbb{P}$ *es equivalente a una familia de* $m$ *números no nega-*

*tivos* $p_1, \ldots, p_m$ *tales que* $\sum_{i \le m} p_i = 1$.

La proposición anterior nos permite imaginar una representación gráfica de los espacios de probabilidad:

**Definition 02.1.** Sea $(\Omega, \mathbb{P},)$ un espacio de probabilidad, al conjunto $\{(\omega_i, \mathbb{P}(\omega_i)) \in \Omega \times [0,1] : \omega_i \in \Omega\}$ se le llamará la función de masa de probabilidad o gráfica de la medida de probabilidad.

**Remark 02.10.** *Normalmente estamos acostumbrados a visualizar las gráficas de las funciones en el plano cartesiano, sin embargo en algunos casos será complicado hacerlo pues nuestros espacios* $\Omega$ *no son naturalmente "unodimensionales" e.g. el espacio de probabilidad asociado al lanzamiento de dos dados.*

**Exercise 02.11.** *Para cada una de las experiencias aleatorias del ejercicio 01.2, determinar el espacio de probabilidad adecuado.*

**Exercise 02.12.** *Supongamos que tenemos una urna con* 30 *bolas de tres colores diferentes: blanco (10), rojo (8) y negro (12). Consideremos el evento aleatorio: elegir 5 bolas al azar. Determinar* $\Omega$ *y alguna ley de probabilidad para esta experiencia. £Existe alguna otra?*

**Exercise 02.13.** *Supongamos que tomamos 10 post-its y en cada uno escribimos alguna de las letras de la palebra MATEMÁTICO. Si doblamos cada uno de los post-its y los metemos en una urna, £cuál es el espacio de probabilidad asociado a esta experiencia? £qué probabilidad hay de obtener la palabra matemático al sacar los diez post-its de manera aleatoria?*

La probabilidad es la ciencia que después de fijar a una ley de probabilidad estudia las propiedades límites de la experiencia aleatoria asociada, por el contrario la estadística es el proceso inverso.

**Definition 02.2.** Sea $(\Omega, \mathbb{P})$ un espacio de probabilidad y $S \in \Omega^n$, una estadística se define como un espacio de probabilidad $(\Omega_S, \mathbb{P}_S)$ obtenido a partir de $S$ que pretende aproximar al espacio original.

**Example 02.14.** *Supongamos que una compañía quiere modelar la satisfacción de un usuario utilizando la siguiente asignación, si el usuario está satisfecho se le asignará el valor uno, de otro modo se le asignará el valor cero.*

*Notemos que es posible modelar la satisfacción promedio de todos los usuarios mediante un solo valor $\beta^* \in [0,1]$, sin embargo algunas veces podría ser complicado conocer la información de todos los usuarios y por tanto se considerará solo una muestra de tamaño $N$, digamos $S = \{x_1, \ldots, x_N\}$. En ese caso es posible definir una estadística mediante el promedio empírico $\frac{1}{N} \sum\limits_{i \leq N} x_i$.*

**Exercise 02.15.** *Determinar el espacio de probabilidad del ejemplo anterior así como la estadística que genera el promedio empírico:*

- *El espacio de probabilidad asociado es un espacio de bernoulli:*

$$(\{\omega_{sat}, \omega_{insat}\}, \mathbb{P})$$

*tal que $\mathbb{P}(\omega_{sat}) = \frac{n_{sat}}{n}, \mathbb{P}(\omega_{insat}) = \frac{n_{insat}}{n}$ donde $n_{sat}, n_{insat}, n$ son las cantidades de clientes satisfechos, insatisfechos y totales respectivamente.*

- *La estadística se define como*

$$\left(\{\omega'_{sat}, \omega'_{insat}\}, \mathbb{P}'\right)$$

*tal que* $\mathbb{P}'\left(\omega'_{sat}\right) = \frac{N_{sat}}{N}, \mathbb{P}'\left(\omega_{insat}\right) = \frac{N_{insat}}{N}$ *donde* $N_{sat}, N_{insat}$ *son las cantidades de clientes satisfechos e insatisfechos.*

# 03 Independencia y condicionamiento

**Definition 00.1.** Sea $(\Omega, \mathbb{P})$ un espacio de probabilidad. Si $A, B \subseteq \Omega$ son dos eventos aleatorios, diremos que ellos son independientes si $\mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A \cap B)$.

**Exercise 00.1.** *En la experiencia aleatoria de tirar dos dados, demostrar lo siguiente:*

1. *Obtener un uno en el primer lanzamiento y un dos en el segundo son eventos aleatorios independientes: digamos que el evento $A$ es "obtener un uno en el primer lanzamiento" y $B$ es el evento "obtener un dos en el segundo", entonces* $A = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$, $B = \{(1,2),(2,2),(3,2),(4,2),(5,2),(6,2)\}$ *y* $A \cap B = \{(1,2)\}$. *Calculando las probabilidades obtenemos:* $\mathbb{P}(A) = \frac{6}{36}, \mathbb{P}(B) = \frac{6}{36}, \mathbb{P}(A \cap B) = \frac{1}{36}$. *Esto significa que* $\mathbb{P}(B) \cdot \mathbb{P}(B) = \mathbb{P}(A \cap B)$ *y por tanto concluimos que ambos eventos son independientes. Intuitivamente esto es claro pues ambos ninguno de los eventos tiene injerencia en el otro.*

2. *Obtener un número uno en el primer lanzamiento y que la suma de los dos dados sea mayor a ocho son eventos aleatorios dependientes.*

3. *Obtener un número primo repetido y obtener un número par repetido son eventos aleatorios dependientes.*

**Exercise 00.2.** *Si $A, B$ son dos eventos independientes, £cómo son los eventos*

*$A^c, B^c$, independientes o dependientes?*

**Definition 00.3.** *Sea $(\Omega, \mathbb{P})$ una ley de probabilidad. Sean $A, B \subseteq \Omega$ son dos eventos aleatorios tales que $\mathbb{P}(B) > 0$. Definimos una nueva ley de probabilidad $\mathbb{P}(A|B) = \mathbb{P}\frac{(A\cap B)}{\mathbb{P}(B)}$ lo cual se lee como la ley de probabilidad condicionada al evento $B$.*

**Exercise 00.4.** *Notemos que dos eventos $A, B$ con probabilidad no cero son independientes si y solo si $\mathbb{P}(A|B) = \mathbb{P}(A)$.*

**Exercise 00.5.** *Consideremos los tres casos posibles del ejercicio 00.1, en todos ellos calcular $\mathbb{P}(A|B)$ y $\mathbb{P}(B|A)$.*

1. *Recordemos que para el primero de los casos tenemos*

   $A = \{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)\}$,

   $B = \{(1,2), (2,2), (3,2), (4,2), (5,2), (6,2)\}$ *y* $A \cap B = \{(1,2)\}$

   *Comencemos a calcular* $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)} = \frac{1/36}{6/36} = 1/6$, *por otro lado* $\mathbb{P}(B|A) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)} = \frac{1/36}{6/36} = 1/6$. *En este caso las probabilidades que acabamos de calcular son las probabilidades de los siguientes eventos: "si yo he obtenido un uno en el primero de mis lanzamientos, la probabilidad de obtener un dos en el segundo" y "si yo obtuve un dos en el segundo lanzamiento, la probabilidad de haber obtenido un uno en el primer lanzamiento".*

2.

3.

**Exercise 00.6.** *Demostrar que la función* $\mathbb{P}(-|B)$ *es una ley de probabilidad en el conjunto* $\Omega$.

**Exercise 00.7.** *Si* $\mathbb{P}$ *es una ley de probabilidad y* $A, B$ *son dos eventos aleatorios, £qué número podría ser más pequeño?* $\mathbb{P}(A|B)$ *o* $\mathbb{P}(A)$ *? £cuándo serán iguales?*

**Exercise 00.8.** *Si* $\Omega = \{1, 2, \ldots, 36\}$ *y* $\mathbb{P}_{Unif}$ *es la ley uniforme sobre un conjunto de tamaño* 36 *(recuerden que esto corresponde a la experiencia aleatoria de tirar 2 dados), si* $B = \{1, 2, 3, 4, 5, 6\}$*, determinar los valores de* $\mathbb{P}(-|B)$. *£A qué evento aleatorio podría corresponder la ley de probabilidad* $\mathbb{P}(-|B)$*?*

**Theorem 00.9.** *(Fórmula de Bayes) Supongamos que* $(\Omega, \mathbb{P})$ *una ley de probabilidad sobre un conjunto finito y* $A, B$ *dos eventos tales que* $\mathbb{P}(B), \mathbb{P}(A) > 0$. *Entonces* $\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)}$.

**Exercise 00.10.** *Deducir la fórmula de Bayes.*

**Exercise 00.11.** *Clasifiquemos los administradores de portafolios financieros en dos grupos, los bien informados y los mal informados. Supongamos que sabemos que con probabilidad igual a .8, si un administrador bien informado elige un activo financiero para su cliente, entonces el activo subirá de precio. Por el contrario la probabilidad de que eso ocurra si el administrador está mal informado es de .6. Supongamos que el 10 porciento de los administradores financieros están bien informados. La pregunta es la siguiente: un cliente elige al azar a un administrador y le pide que compre un activo financiero, suponiendo que el activo financiero subió, cuál es la probabilidad de que el administrador esté mal informado?*

*Demostración.* Si $M$ es el evento: el valor del activo sube e $I$ es el evento: el administrador está bien informado. Lo siguiente es cierto y no es difícil convencernos:

$$\mathbb{P}(M) = \mathbb{P}_I(M) \cdot \mathbb{P}(I) + \mathbb{P}_{I^c}(M) \cdot \mathbb{P}\left(I^c\right).$$

Utilizemos la fórmula de Bayes para concluir. □

# 04 Variables aleatorias discretas

La teoría moderna de probabilidades se basa en el manejo de las variables aleatorias, trataremos de convencer al estudiante de sus ventajas con algunos ejemplos.

**Definition 00.1.** Sea $(\Omega, \mathbb{P})$ una ley de probabilidad y $\mathbb{R}$ el conjunto de los números. Una variable aleatoria es una función $X : \Omega \to \mathbb{R}$.

**Exercise 00.1.** *Si $X : \Omega \to \mathbb{R}$ es una variable aleatoria, demostrar que es posible definir una ley de probabilidad sobre $\mathbb{R}$ de la siguiente forma: para cada $A \subseteq \mathbb{R}$ definimos $\mathbb{P}_X(A) = \mathbb{P}\left(X^{-1}(A)\right)$.*

**Exercise 00.2.** *Si $\Omega = \{1, \ldots, 6\}^2$ es el espacio de probabilidad correspondiente a la experiencia aleatoria de lanzar dos dados de manera independiente, definimos la siguiente variable aleatoria: $X_{sum}\left(i, j\right) = i + j$. Graficar la función $F : \mathbb{R} \to \mathbb{R}$ definida por $F(x) = \mathbb{P}_{X_{sum}}(\{x\})$.*

**Exercise 00.3.** *Sea $\left(\{0, 1\}^n, \mathbb{P}_{Bernoulli,n,p}\right)$ el producto de las leyes de probabilidad de Bernoulli. Para un $i \leq n$ y cada $\omega = (\omega_1, \ldots, \omega_n) \in \{0, 1\}^n$ definimos $X_i(\omega) = \omega_i$. Calcular $\mathbb{P}_{X_k}$.*

**Proposition 00.4.** *Sea $\left(\{0, 1\}^n, \mathbb{P}_{Bernoulli,n,p}\right)$ y para cada $i \leq n$ $X_i$ es la variable aleatoria definida en el ejercicio anterior, definimos la variable aleatoria $S_n = \sum_{i \leq n} X_i$. Entonces $\mathbb{P}_{S_n} = \mathbb{P}_{Bin,n,p}$.*

A continuación introducimos la definición de independencia para variables aleatorias.

**Definition 00.5.** *Sean* $X, Y : \Omega \to \mathbb{R}$ *dos variables aleatorias*

- *Definimos la variable aleatoria* $Z = (X, Y)$ *como la función* $Z : \Omega \to \mathbb{R}^2$ $Z(\omega) = (X(\omega), Y(\omega))$.
- *Decimos que las variables aleatorias* $X, Y$ *son independientes si para todo* $A \subseteq Im(X), B \subseteq Im(Y), \mathbb{P}_Z(A \times B) = \mathbb{P}_X(A) \cdot \mathbb{P}_Y(B)$.

**Example 00.6.** *Sea* $p \in [0,1]$ *tal que* $p \neq \frac{1}{2}$, *definimos el siguiente expacio de probabilidad* $\Omega = \{a, b, c, d\}$ *y* $\mathbb{P}$ *tal que* $\mathbb{P}(a) = \frac{p}{2}, \mathbb{P}(b) = \frac{p}{2}, \mathbb{P}(c) = \frac{1-p}{2}, \mathbb{P}(d) = \frac{1-p}{2}$. *También definimos las variables aleatorias:* $X : \Omega \to \{0,1\}$ *e* $Y : \Omega \to \{0,1\}$ *de la siguiente manera:* $X(a) = X(c) = 0, X(b) = X(d) = 1, Y(a) = Y(d) = 0, Y(b) = Y(c) = 1$. *Notemos que* $Z = (X, Y)(a) = (0,0), Z(b) = (1,1), Z(c) = (0,1), Z(d) = (1,0)$. *Estas variables aleatorias no son independientes pues* $\mathbb{P}_Z((0,0)) = \frac{p}{2}$ *sin embargo* $\mathbb{P}_X \otimes \mathbb{P}_Y(0,0) = \mathbb{P}\left(X^{-1}(0)\right) \cdot \mathbb{P}\left(Y^{-1}(0)\right) = \mathbb{P}(a, c) \cdot \mathbb{P}(a, d) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

## Esperanza, varianza y otros momentos de las variables aleatorias

---

**Definition 01.1.** *Si* $X : \Omega \to \mathbb{R}$ *es una variable aleatoria con y* $\mathbb{P}$ *una ley de probabilidad sobre el conjunto* $\Omega$ *determinada por* $p_i = \mathbb{P}(\omega_i)$, *definimos la* *esperanza de $X$ de la siguiente forma:*

$$\mathbb{E}[X] = \sum_{\omega_i \in \Omega} p_i \cdot X(\omega_i)$$

**Remark 01.2.** *Si $X$ es una variable aleatoria denotaremos por $Im(X)$ al conjunto de valores $x \in \mathbb{R}$ tales que existe algún $\omega \in \Omega$ con $X(\omega) = x$ y por $p_x = \mathbb{P}_X(\{x\})$.*

**Proposition 01.3.** *Si $X$ es una variable aleatoria, entonces*

$$\mathbb{E}(X) = \sum_{x \in Im(X)} x \cdot p_x.$$

*Demostración.*

$$\mathbb{E}(X) = \sum_{\omega_i \in \Omega} p_i \cdot X(\omega_i) = \sum_{x \in Im(X)} \left( \sum_{\omega_i : X(\omega_i) = x} p_i \cdot x \right) = \sum_{x \in Im(X)} \mathbb{P}(\{\omega : X(\omega_i) = x\}) \sum_{x \in Im(X)} x \cdot p_x$$

□

**Example 01.4.** *Sea $X : \{0,1\} \to \mathbb{R}$ la función identidad vista como una variable aleatoria sobre el espacio de probabilidad de Bernoulli $(\{0,1\}, \mathbb{P})$, entonces $\mathbb{E}[X] = \left(0 \cdot (1-p)\right) + (1 \cdot p) = p$.*

**Exercise 01.5.** *Definir dos variables aleatorias que corresponda a las 10 calificaciones durante un semestre de dos alumnos, todas ellas entre 5 y 10 y calcular sus esperanzas.*

**Exercise 01.6.** *Calcular la esperanza de $X_{sum}$ del ejercicio 00.2.*

Una propiedad muy importante de la esperanza es que es una funcional lineal en el siguiente sentido:

**Proposition 01.7.** *Si $X, Y$ son dos variables aleatorias y $a, b \in \mathbb{R}$, entonces $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$.*

**Proposition 01.8.** *Sea $\left(\{0,1\}^n, \mathbb{P}_{Bernoulli,n,p}\right)$ y $S_n$ la variable aleatoria definida en la proposición 00.4. Entonces $\mathbb{E}[S_n] = np$.*

**Example 01.9.** *En el ejemplo 02.8, la esperanza es* $307 \times 0.005 = 1.535$ *que es lo que se esperaría en promedio de medallas ganadas por Argentina de acuerdo al tamaño de su población.*

**Definition 01.1.** Sean $X, Y : \Omega \to \mathbb{R}$ dos variables aleatorias

- Definimos la varianza de $X$ como

$$Var(X) = \sum_{x \in Im(X)} (x - \mathbb{E}(X))^2 \cdot p_x$$

- Definimos la desviación estándar de una variable aleatoria como la raíz positiva de $Var(X)$, se denotará por $\sigma_X$.

- Definimos la covarianza entre $X$ y $Y$ como

$$Cov(X, Y) = \mathbb{E}(X \cdot Y) - \mathbb{E}(X) \cdot \mathbb{E}(Y)$$

- Si $X, Y$ son dos variables aleatorias, definimos su correlación como $Corr(X, Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y}$

**Remark 01.10.** $Cov(X,X) = Var(X)$.

**Example 01.11.** *Sea* $X : \{0,1\} \to \mathbb{R}$ *la variable aleatoria definida en el ejemplo 01.4 entonces* $Var(X) = \left(0-p\right)^2 \cdot \left(1-p\right) + \left(1-p\right)^2 \cdot p = p \cdot \left(1-p\right)$.

**Exercise 01.12.**

- *Calcular la varianza de la variable aleatoria definida en el ejercicio 01.5.*
- *Calcular la varianza de la variable* $X_{sum}$ *definida en el ejercicio 00.2.*

**Exercise 01.13.** *Si* $X, Y$ *son variables aleatorias independientes entonces* $Cov(X,Y) = 0$.

# 05 Ley de los grandes números

**Theorem 00.1.** *(Ley fuerte de los grandes números) Sean* $(\Omega, F, \mathbb{P})$ *un espacio de probabilidad y* $X_n : \Omega \to \mathbb{R}$ *una familia de variables aleatorias independientes e identicamente distribuidas tales que* $\mathbb{E}[|X_i|] < \infty$ *entonces existe un subconjunto* $N \in F$ *tal que* $\mathbb{P}_X(N) = 0$ *y además:*

$$\lim_{n\to\infty}\left(\frac{X_1(m) + \ldots + X_n(m)}{n}\right) = \mathbb{E}[X_1], \forall m \notin N$$

Una de las aplicaciones más importantes de la ley de los grandes números es el llamado método de Monte Carlo.

**Definition 00.1.** (Método Monte Carlo) Supongamos que queremos calcular cierto parámetro $\mu$. El método de Monte Carlo es una aplicación inmediata del resultado anterior, consiste en los siguientes pasos:

1. Definimos variables aleatorias i.d.d. cuya esperanza coincida con $\mu$.

2. Construimos un muestreo de tamaño arbitrario de nuestras variables aleatorias.

3. Gracias a la Ley de los grandes números el promedio empírico de este muestreo cada vez se acercará más a $\mu$

# Estadística uniforme y equidistribuida

**Definition 01.1.** Sea $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$, definimos:

1. El promedio empírico de $S$ como $\mu_S = \frac{1}{N}(x_1 + \ldots x_N)$.
2. El promedio empírico de cada una de las $j \leq d$ variables como $\mu_{S,j} = \frac{1}{N}\left(x_{1,j} + \ldots x_{N,j}\right)$
3. La varianza empírica de $S$ se define como $Var(S) = \frac{1}{N}\sum_{i \leq N}\left(x_i - \mu_S\right)^2$
4. La covarianza empírica de $S$ se define como una matriz simétrica de $d \times d$ entradas $(Cov(S))_{j,i} = (Cov(S))_{i,j} = \frac{1}{N-1}\sum_{k \leq N}\left(x_{k,i} - \mu_{S,i}\right)\left(x_{k,j} - \mu_{S,j}\right)$
5. La correlación empírica de $S$ se define como una matriz simétrica de $d \times d$ entradas $(Corr(S))_{j,i} = (Corr(S))_{i,j} = \frac{(Cov(S))_{j,i}}{\sqrt{(Cov(S))_{j,j}}\sqrt{(Cov(S))_{i,i}}}$

Gracias a la Ley Fuerte de los Grandes Números (y quizás algún argumento de continuidad) podemos concluir lo siguiente:

**Proposition 01.1.** *Supongamos que $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$ está generado por $N$ variables aleatorias $X_1, \ldots, X_N$ independientes e idénticamente distribuidas tales que $\mathbb{E}[|X_i|] < \infty$, entonces existe un subconjunto medible $N$ tal que $\mathbb{P}_X(N) = 0$ y además:*

1. $\lim_{n \to \infty} \mu_S(x) = \mathbb{E}[X_1], \forall x \notin N$
2. $\lim_{n \to \infty} (Var(S)(x)) = Var(X_1), \forall x \notin N$

3. $\lim_{n \to \infty} \left( Cov\,(S)_{i,j} \right) = Cov\left( X_i, X_j \right)$

4. $\lim_{n \to \infty} \left( Corr\,(S)_{i,j} \right) = Corr\left( X_i, X_j \right)$

# 06 Extensión de variables aleatorias y el Teorema Límite Central

Hasta ahora solo hemos hablado 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.

## Distribuciones infinitas

---

**Example 01.1.** *Si $E$ es la experiencia aleatoria de lanzar un dado justo hasta obtener el número seis, definimos la siguiente variable aleatoria:* $X(\omega_1, \omega_2, \ldots) = \underset{j\geq 1}{inf}\{j : \omega_j = 6\}$

**Example 01.2.** *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 01.3.** *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 $\underset{n\to\infty}{lim}\, n \cdot p_n = \lambda$ entonces $\underset{n\to\infty}{lim}\mathbb{P}_{Bin,p_n,n}(i) = \mathbb{P}_{Poisson,\lambda}(i)$.*

**Exercise 01.4.** *Comparemos la distribución de Poisson con la ley binomial que calculamos en el ejemplo 02.8, 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 en 01.9, de esa forma* $\mathbb{P}_{Poisson,1.535}(0) = 0.215$, $\mathbb{P}_{Poisson,1.535}(0) = 0.33$, $\mathbb{P}_{Poisson,1.535}(0) = 0.253$, *y* $\mathbb{P}_{Poisson,1.535}(0) = 0.129$.

Notemos que si en lugar de considerar la varibale 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.

## 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 sub-conjuntos, 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 02.1.**

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 02.2.** La ley de probabilidad Gaussiana o normal con parámetros $\left(\mu,\sigma^2\right)$ 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 02.1.** *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 02.2.** *(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 $x \in \mathbb{R}^d \setminus \{\bar{0}\}$. Definimos la ley de probabilidad gaussiana $d$ dimensional con parámetros $\mu, S$ de la siguiente forma:*

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

**Exercise 02.3.** *(Laplace) La distribución de Laplace con media $\mu$ y varianza $2b^2$ se define como*

$$\mathbb{P}_{Laplace}(-\infty,x) = \frac{1}{2b}\int_{-\infty}^{x} e^{-\frac{|x-\mu|}{b}}$$

# Operaciones con variables aleatorias

---

Ya hemos visto en la proposición 01.7 que algunas propiedades de las variables aleatorias se comportan bien bajo ciertas operaciones. 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 03.1.** *Sean $X, Y$ dos variables aleatorias independientes, sea $Z = X + Y$ y $W = (X, Y)$*

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

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

  *Entonces $Z$ tiene una ley de probabilidad $\mathbf{N}\left(\mu + \mu', \sigma^2 + \sigma'^2\right)$. 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 03.2.** *Sea $X$ una variable aleatoria gaussiana, digamos $\mathbf{N}(0, 1)$*

- *Si $X'$ es una variable aleatoria tal que $\mathbb{P}\left(X'=1\right)=\frac{1}{2}, \mathbb{P}\left(X'=-1\right)=\frac{1}{2}$ 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 $\left(X, X'\right)$ no sigue una distribución gaussiana multi-variada.*

## Teorema del límite central

En el ejemplo 01.2 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 $X_1, X_2, \ldots$ variables aleatorias i.i.d, sabemos que $\frac{X_1+\ldots+X_n}{n}$ 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 $\alpha$:

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

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 ningúna 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 04.1.** *Sean $X_n$ es una familia de variables aleatorias independientes e identicamente distribuidas tales que $\mathbb{E}(X_i) = \mu, Var(X_i) = \sigma^2$ para todo $i$. Definamos $M_n = \sqrt{n}\left(\frac{(X_1+X_2+\ldots X_n)}{n} - \mu\right)$ entonces para todos $a < b$*

$$\lim_{n\to\infty} \mathbb{P}_{M_n}(a,b) = \mathbb{P}_{Gauss,0,\sigma^2}(a,b)$$

.

# 07 Tests estadísticos

## 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. Definimos la varibale aleatoria $X : \{A, B\} \to \{0, 1\}$ por $X(A) = 0, X(B) = 1$. Supondremos que evitamos preguntar a la misma persona dos veces y también que el resultado de una respuesta no influye en el resto. Eso significa que estamos definiendo una familia de 2500 variables aleatorias $X_i$ i.i.d. Notemos que como tenemos dos candidatos, en realidad estamos suponiendo una Ley de Bernouilli 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 01.1.** *En la situación anterior demostrar que si llamamos* $\bar{X}_n = \left(\frac{(X_1+X_2+\ldots X_n)}{n} - \mathbb{E}(X\right.$ *entonces* $\mathbb{E}\left(\bar{X}_n\right) = p$.

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 voraron 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? Recordando 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.

Buscamos calcular lo siguiente:

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

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

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

.

El teorema central límite dice que

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

Dejamos como ejercicio demostrar que lo anterior es equivalente a:

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

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 $\mathbb{P}\left(|\bar{X}_n - p| < \epsilon\right) = .95$ entonces debemos de igualar $\frac{\sqrt{n}\epsilon}{\sqrt{p(1-p)}} = 1.96$ aproximadamente. Eso implica que es necesario que $|p_n - p| < \epsilon \leq \frac{1.96\cdot\sqrt{p(1-p)}}{\sqrt{n}}$

Desafortunadamente la cantidad anterior depende de $p$, sin embargo notemos que $\sqrt{p\left(1-p\right)} \leq \frac{1}{2}$. Por tanto es necesario que $|p_n - p| < \epsilon \leq \frac{1.96\cdot\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1.96}{2\cdot\sqrt{n}} \leq \frac{1}{\sqrt{n}}$. Resumimos el razoniamiento anterior en el siguiente teorema:

**Theorem 01.2.** *Bajo las condiciones del sondeo descritas anteriormente, si queremos estar con probabilidad del* 95 *porciento en una predicción acertada de los votantes del candidato $B$. Si $n$ es suficientemente grande, debemos estar en una predicción en el intervalo* $\left[p_n - \frac{1}{\sqrt{n}}, p_n + \frac{1}{\sqrt{n}}\right]$

## Test de independencia

**Exercise 02.1.** *Sean $X_1, \ldots, X_p$ un variables aleatorias independientes e idénticamente distribuidas cuya ley de probabilidad es* $\mathbf{N}(0,1)$*, definimos la variable aleatoria*

$$X_1^2 + \ldots + X_p^2$$

*Llamaremos a su ley de probabilidad asociadada $\chi^2_p$ la ley de pearson de $q$ grados de libertad.*

**Exercise 02.2.** *Sean $X, Y$ dos leyes de probabilidad que tienen asociadas funciones de probabilidad* $\mathbf{N}(0,1)$ *y* $\chi^2_p$ *respectivamente, definimos la variable aleatoria*

$$Z = \frac{X}{\sqrt{\frac{Y}{q}}}$$

*Su ley de probabilidad asociada la llamaremos la ley de probabilidad $t$ del estudiante $t_q$ con $q$ grados de libertad.*

**Definition 02.1.** Sea $X$ una variable aleatoria y $\mathbb{P}_X$ ley de probabilidad asociada. Para cada valor $p \in (0,1)$ definimos el quantil de orden $p$ de la variabble aleatoria $X$ de la siguiente manera:

$$q\left(p\right)(\mathbb{P}_X) = inf\{x \in \mathbb{R} : p \leq \mathbb{P}\left(X \leq x\right)\}$$

## Test de normalidad

# 08 Regresiones lineales: caso inyectivo

**Remark 00.1.** *Recordemos que una matriz real $X \in \cdot \mathbb{N}$ induce una función $X : \mathbb{R}^d \to \mathbb{R}^N$, decimos que esta matriz es inyectiva cuando la función que induce es inyectiva i.e. si $X(v) = 0 \in \mathbb{R}^N$ entonces $v = 0 \in \mathbb{R}^d$. La propiedad pincipal que utilizaremos sobre estas matrices es que si $X$ es inyectiva entonces $X^T \cdot X$ es invertible i.e. existe una matriz $\left(X^T \cdot X\right)^{-1}$ tal que $\left(X^T \cdot X\right)\left(X^T \cdot X\right)^{-1} = Id$.*

**Definition 00.1.** Fijemos dos enteros positivos $d, N$. Definimos una base de datos lineal e inyectiva como un conjunto $S = \{(x_1, y_1), \ldots (x_N, y_N)\}$ tal que $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, $x_{i,d} = 1$ para todo $i$ y existe algún $m^* = \left(m_1^*, m_2^*, \ldots m_d^*\right)$ que satisface $y_i = m^* x_i + \epsilon_i$ donde $\epsilon_i$ son variablse aleatorias que satisfacen:

- $\mathbb{E}(\epsilon) = 0$,
- Las variables aleatorias $\epsilon_i$ son independientes e identicamente distribuidas,
- La matriz de $N \times d$, $X = (x_i)_{i \leq N}$ es inyectiva.

Con las suposiciones anteriores si denotamos por $Y$ al vector $(y_i)_{i \leq N}$ y $\epsilon$ al vector $(\epsilon_i)_{i \leq N}$, una base de datos lineal e inyectiva $S$ se puede escribir como $Y = Xm^* + \epsilon$.

**Definition 00.2.** Dada una base de datos lineal e inyectiva $S$ y un vector $m \in \mathbb{R}^d$, definimos el error de $m$ como la variable aleatoria

$$err_S(m) = \frac{1}{N} \cdot \sum_{i \leq N} \left( \langle x_i, m \rangle - y_i \right)^2$$

El estimador óptimo de mínimos cuadrados se define de la siguiente manera:

$$m^* = \underset{m' \in \mathbb{R}^d}{argmin} \left( err_S(m') \right)$$

**Theorem 00.2.** *Sea $S$ una base de datos lineal e inyectiva, entonces el estimador óptimo de mínimos cuadrados existe y es único.*

*Demostración.* Vamos a demostrar que $m^* = \left(X^T X\right)^{-1} X^T Y$ satisface la condición pero comenzamos explicando cómo encontrarlo:

Si consideramos a $err_S(m')$ como una función de $m'$, sus mínimos (de existir) estarán en aquellos puntos donde el gradiente es cero.

Si calculamos el gradiente obtenemos lo siguiente:

$$\frac{2}{N} \sum_{i \leq N} \left( \langle x_i, m' \rangle - y_i \right) x_i$$

Es fácil deducir de lo anterior que $m^*$ debe de ser la solución a la igualdad con cero.

Para asegurarnos que es en verdad un mínimo es necesario calcular el Hessiano de $X$ y verificar que es definido positivo.

□

Sea $S$ una base de datos lineal e inyectiva y además supongamos que la matrix $X$ es diagonal (es decir que fuera de su diagonal todos los números son cero).

**Exercise 00.3.** *Bajo las hipótesis del párrafo anterior, demostrar que todas las entradas de la diagonal son distintas de cero.*

**Exercise 00.4.** *Dar un ejemplo cuando $N > d$ pero la matriz sigue siendo inyectiva.*

## Cualidades estadísticas del caso inyectivo

---

Notemos que en la suposición del modelo lineal inyectivo y determinista la aleatoriedad solo depende de $\epsilon$ y por tanto ambas matrices $X, Y$ están determinada. También es posible hacer un análisis cuando $X$ e $Y$ son puramente aleatorias.

**Proposition 01.1.** *Si $S$ es una base de datos lineal e inyectiva tal que $Var(\epsilon) = \sigma^2 Id_d$ entonces $\mathbb{E}(m_{LSS}) = m^*$ y $Var(m_{LSS}) = \sigma^2 \left(XX^T\right)$.*

*Demostración.* □

**Corollary 01.2.** *Para todo $x \in \mathbb{R}^d$, si definimos la variable aleatoria $y := m_{LSS} \cdot x$, entonces:*

$$\mathbb{E}\left[y\right] = m^* \cdot x, Var\left(y\right) = \sigma^2 x^T \left(XX^T\right) x$$

La proposición anterior solo es útil cuando conocemos la varianza de $\epsilon$, sin embargo en general eso no es posible. Una manera de calcular esa varianza una vez hecha la predicción es la siguiente:

**Definition 01.1.** Fija una base de datos lineal e inyectiva definimos la varianza natural del estimador óptimo de mínimos cuadrados $m^*$ de la siguiente forma:

$$\sigma_{nat}^2 = \frac{1}{N-d-1} \cdot \sum_{i \leq N} \left( \langle x_i, m_{LSS} \rangle - y_i \right)^2$$

**Proposition 01.3.** $\mathbb{E}\left(\sigma_{nat}^2\right) = \sigma^2$.

## Hipótesis Gaussianas y tests estadísticos

Hasta ahora no hemos hecho ninguna hipótesis sobre la naturaleza de las distribuciones $\epsilon_i$, si por ejemplo suponemos que esos errores son Gaussianos entonces es posible hacer un análisis más preciso:

**Proposition 02.1.** *Supongamos que* $\mathbb{P}(\epsilon_i) = \mathbf{N}\left(0, \sigma^2\right)$*, entonces:*

- $\mathbb{P}_{m_{LSS}} = \mathbf{N}\left(m^*, \sigma^2 \left(XX^T\right)^{-1}\right)$
- $(N-d-1)\,\sigma_{nat}^2 = \sigma^2 \chi_{N-p-1}^2$
- $\sigma_{nat}^2$ *y* $m_{LSS}$ *son variables aleatorias independientes.*

**Corollary 02.2.** *Para todo* $x \in \mathbb{R}^d$*, si definimos la variable aleatoria* $y := m_{LSS} \cdot x$*, entonces:*

$$\mathbb{P}_{m_{LSS}} = \mathbf{N}\left(m^* x, \sigma_{nat}^2 x \left(XX^T\right)^{-1} x\right)$$

Gracias a los resultados anteriores es posible formular algunos tests estadísticos útiles en la práctica:

**Corollary 02.3.** *(Intervalos de confianza) Sea $t_{N-d}$ la ley $t$ de probabilidad del estudiante con $N-d$ grados de libertad, para cualquier $\epsilon > 0$ y cualquier $x \in \mathbb{R}^d$*

$$\mathbb{P}\left(\left(m^* \cdot x\right) \in \left[g, g'\right]\right) = 1 - \epsilon$$

*donde*

$$g = m_{LSS} \cdot x - \sqrt{\sigma_{nat}^2 x \left(XX^T\right)^{-1} x} - q_{1-\frac{\epsilon}{2}}\left(t_{N-d}\right)$$

$$g' = m_{LSS} \cdot x + \sqrt{\sigma_{nat}^2 x \left(XX^T\right)^{-1} x} - q_{1-\frac{\epsilon}{2}}\left(t_{N-d}\right)$$

# 09 Máxima verosimilitud y Análisis discriminante lineal

## Máxima verosimilitud

---

Un acercamiento probabilista a machine learning es vía la máxima verosimilutud, en esta sección ejemplificaremos este punto de vista con un ejemplo.

En la última línea no es claro por qué es necesario calcular la esperanza de la estimación $\frac{1}{N}\sum_{i\leq N} x_i$. Esta última cantidad es enrealidad una variable aleatoria. Ahora vamos a explicar qué significa esta cantidad en términos de verosimilitud. Notemos que si queremos calcular la probabilidad de obtener el muestreo $S$ tenemos la siguiente fórmula:

$$\mathbb{P}(S) = \prod_{i\leq N} \left(\beta^*\right)^{x_i} \left(1-\beta^*\right)^{1-x_i} = \left(\beta^*\right)^{\sum_{i\leq N} x_i} \left(1-\beta^*\right)^{\sum_{i\leq N}(1-x_i)}$$

Eso implica que $log(\mathbb{P}(S)) = \left(\sum_{i\leq N} x_i\right) log\left(\beta^*\right) + \left(\sum_{i\leq N}(1-x_i)\right) log\left(1-\beta^*\right)$

Como $log$ es una función creciente, definimos el maximizador de la verosimilitud de la siguiente manera:

**Definition 01.1.**
- Dado un parámetro $\beta \in [0,1]$, definimos la cantidad $L(S,\beta) = \left(\sum_{i\leq N} x_i\right) log(\beta) + \left(\sum_{i\leq N}(1-x_i)\right) log(1-\beta)$ como la verosimilitud logarítmica del parámetro $\beta$ dada la base de datos $S$.

- Dada una base de datos $S$, definimos el parámetro que maximiza la verosimilitud de la siguiente manera:

$$\beta_{MLE} = \max_{\beta \in [0,1]} L(S, \beta)$$

**Proposition 01.1.** *Bajo las hipótesis de esta sección se tiene que* $\beta_{MLE} = \frac{1}{N} \sum_{i \leq N} x_i$

## Análisis discriminante lineal

BOURBAKI

COLEGIO DE MATEMÁTICAS

escuela-bourbaki.com