Statistica Matematica

Los Del DGIIM, losdeldgiim.github.io

Doble Grado en Ingeniería Informática y Matemáticas
Universidad de Granada

Statistica Matematica

Los Del DGIIM, losdeldgiim.github.io

Joaquín Avilés de la Fuente

Granada, 2025-2026

Índice general

1. Richiami di Probabilità **5**

1.1. Teoria della misura 7
1.2. Variabili aleatorie 8
1.2.1. Propietà funzione di ripartizione 9
1.3. Richiami di combinatoria 12
1.4. Distribuciones Discretas 12
1.4.1. Distribución uniforme discreta 12
1.4.2. Distribución binomial 13
1.4.3. Distribución binomial negativa 13
1.4.4. Distribución de Poisson 14
1.5. Distribuciones Continuas 15
1.5.1. Distribución Uniforme Continua 15
1.5.2. Distribución Normal 16
1.5.3. Distribución Exponencial 16
1.6. Distribuciones Reproductivas 17

2. Statistica **19**

2.1. Distribuzioni uniformi 19
2.2. Modelli a urne (con reinserimento) 19
2.2.1. Caso ordenado 20
2.2.2. Caso no ordenado 22
2.3. Modelli a urne (senza reinserimento) 24
2.3.1. Caso ordenado 24
2.3.2. Caso no ordenado 25
2.4. Distribuzione di Poisson 27
2.4.1. Processo di Poisson 28
2.5. Tempi di attesa 29
2.6. Distribuzione Normale 31
2.7. Probabilità condizionata 32
2.8. Multi-stage models 34
2.9. Indipendenza 38
2.10. Esperanza 40
2.11. Varianza 41
2.12. Leggi dei grandi numeri 43
2.13. Teorema Central del Límite 46

3. Applicazioni **49**
3.1. Regioni di confidenza 49
3.1.1. Metodo diretto per costruire regione di confidenza 50
3.1.2. Metodo alternativo per costruire regione di confidenza 53

4. Inferenza Statistica **57**
4.1. Valores muestrales 59
4.2. Regressione lineare semplice 64
4.3. Likelihood/Verosimiglianza 68
4.4. Entropia 73
4.5. Distribución Gaussiana multivariante 76
4.6. Distribuciones relacionadas con la normal 77
4.7. Teorema de estimación de Cramér-Rao-Fisher 80
4.8. Inferenza Bayesiana 83

5. Ejercicios **89**
5.1. Sección 1 89
5.2. Sección 2 89
5.3. Sección 3 89
5.4. Sección 4 91
5.5. Sección 5 91
5.6. Sección 8 91
5.7. Sección 9 92
5.8. PDF extra 92

1. Richiami di Probabilità

Antes de nada es importante notar la traducción de algunos términos clave entre inglés, italiano y español que son clave para entender la bibliografía y los apuntes del curso:

- **Probability density function** = Distribuzione/Funzione di densità di probabilità = Función de densidad de probabilidad
- **Probability mass function** = Distribuzione/Funzione di massa di probabilità = Función masa de probabilidad
- **Cumulative distribution function** = Distribuzione/Funzione di ripartizione: Función de distribución acumulativa

como podemos ver muchas veces en italiano se usa indistintamente el término *distribuzione* o *funzione* para referirse a las funciones de probabilidad. De hecho, a veces se usa el término **legge** (probabilidad) para referirse a la función de probabilidad, ya que en italiano *legge* significa ley, y en este caso se refiere a la ley que rige el comportamiento de la variable aleatoria.

Observación. Es importante destacar que en otros cursos de Probabilidad y Estadística se ha usado $\mathcal{A}$ para denotar la $\sigma-$álgebra, pero en este curso usaremos $\mathcal{F}$ para evitar confusiones con los eventos.

Observación. En muchas ocasiones, usaremos el término *función de densidad* para referirnos a la función masa de probabilidad en variables aleatorias discretas, ya que es un término más general que se aplica tanto a variables aleatorias discretas como continuas, por lo que cuando hablemos de función de densidad, nos referiremos a ambas, y se notará muchas veces también en ambos casos como $P(X)$.

Utilizamos una moneda trucada donde la probabilidad de que salga cara (fenómeno A) es $p \in (0,1)$, por lo que observamos el fenómeno y calcularemos p. Tenemos $X = \mathbb{1}_A$ donde

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{si } x \in A, \\ 0, & \text{si } x \notin A \end{cases}$$

con $P(A) = p$ y espacio muestral $\Omega = \{0,1\}$. Sabemos que una sola prueba no basta, por lo que haremos muchas (lanzamos n veces la moneda) y tenemos un **stimatore**

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Ai}$$

Recordemos la desigualdad de Chebyshev para mas tarde probar la **Legge dei numeri grandi (LGN)**.

Teorema 1 (Disuguaglianza di Markov)**.** *Sea X una variable aleatoria. Si $X \geqslant 0$ y $\exists E[X]$, entonces:*

$$P[X \geqslant \varepsilon] \leqslant \frac{E[X]}{\varepsilon} \qquad \forall \varepsilon \in \mathbb{R}^+$$

Proposición 2. *(Disuguaglianza Chebyshev) Si X es una variable aleatoria tal que $\exists E[X^2]$, se tiene*

$$P[|X - E[X]| \geqslant \delta] \leqslant \frac{\mathrm{Var}[X]}{\delta^2}, \qquad \forall \delta > 0$$

Demostración. Se basa en aplicar la desigualdad básica a la variable $(X - E[X])^2$. Para $\varepsilon = \delta^2$, se tiene que:

$$P[|X - E[X]| \geqslant \delta] = P[(X - E[X])^2 \geqslant \delta^2] \leqslant \frac{E[(X - E[X])^2]}{\delta^2} = \frac{\mathrm{Var}[X]}{\delta^2}$$

□

Teorema 3 (Legge dei numeri grandi (LGN))**.** *Sea $(X_n)_{n\geqslant 1}$ una sucesión de variables aleatorias tal que*

1) X_i independiente a X_j para $i \neq j$

2) $E[X_i] = \mu \ \forall i \in \mathbb{N}$

3) X_i integrables con varianza finita $\mathrm{Var}[X_i] = E[(X_i - E[X_i])^2] = \sigma^2 \ \forall i \in \mathbb{N}$

entonces

$$P\left[\left|\overline{X}_n - \mu\right| \geqslant \delta\right] \leqslant \frac{\sigma^2}{n\delta^2}, \qquad \forall \delta > 0$$

Demostración. Sea $Z = \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, entonces $E[Z] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu$. Por independencia en $(*)$, tenemos que

$$E[(X_i - E[X_i])(X_j - E[X_j])] \overset{(*)}{=} E[X_i - E[X_i]]E[X_j - E[X_j]] = 0, \qquad \forall i \neq j$$

Y finalmente desarrollando obtenemos

$$E(|Z - E[Z]|^2) = E\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]\right|^2\right] = E\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} E[X_i]\right|^2\right] =$$

$$= E\left[\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - E[X_i])\right|^2\right] = E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - E[X_i])(X_j - E[X_j])\right] =$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[(X_i - E[X_i])(X_j - E[X_j])] \overset{(*)}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

donde en $(*)$ sabemos que por independencia (demostrado antes) $E[(X_i - E[X_i])(X_j - E[X_j])] = 0$ para $i \neq j$, quedando solo la suma de varianzas. Tras aplicar Chebyshev para la variable Z, se tiene lo buscado

$$P\left[|Z - E[Z]| \geqslant \delta\right] \leqslant \frac{Var[Z]}{\delta^2} = \frac{\sigma^2}{n\delta^2}$$

□

1.1. Teoria della misura

Definición 1.1.1 (Spazio misurable)**.** Un **espacio medible** es un par ordenado $(\Omega, \mathcal{F})$ que consiste en

- Ω (Espacio Muestral): conjunto no vacío que contiene todos los posibles resultados de un experimento o fenómeno.
- $\mathcal{F}$ (σ-álgebra): familia (o colección) de subconjuntos de Ω que satisface tres axiomas específicos que garantizan que las operaciones lógicas entre sucesos sigan siendo sucesos.

Definición 1.1.2 ($\sigma-$algebra)**.** Sea $\mathcal{F}$ una colección de subconjuntos de un conjunto no vacío Ω, decimos que es una $\sigma-$álgebra si cumple

1. $\emptyset \in \mathcal{F}$
2. Si $A \in \mathcal{F} \implies A^c \in \mathcal{F}$
3. Si $(A_n)_{n\geqslant 1} \in \mathcal{F} \implies \bigcup_{n\geqslant 1} A_n \in \mathcal{F}$

Definición 1.1.3 (Spazio di probabilità)**.** Un **espacio de probabilidad** es una tripla $(\Omega, \mathcal{F}, P)$ donde

1) $(\Omega, \mathcal{F})$ es un espacio medible.

2) $P : \mathcal{F} \to [0, 1]$ es una función llamada **medida de probabilidad** que asigna a cada suceso un número real entre 0 y 1, cumpliendo los Axiomas de Kolmogorov

   - Normalización (N): la probabilidad del espacio total es 1, es decir, $P(\Omega) = 1$
   - σ-aditividad (A): para cualquier secuencia de sucesos $\{A_i\}_{i\geqslant 1}$ que sean disjuntos entre sí (es decir, $A_i \cap A_j = \emptyset$ para $i \neq j$), la probabilidad de su unión es la suma de sus probabilidades individuales

$$P\left(\bigcup_{n\geqslant 1} A_n\right) = \sum_{n\geqslant 1} P(A_n)$$

Quesito. Perché non posso sempre scegliere $\mathcal{F} = \mathcal{P}(\Omega)$? (l'insieme delle parti di Ω) Esto es posible siempre que Ω numerable.

Teorema 1.1.4 (Vitali 1905)**.** *The power set is too large. Let $\Omega = \{0, 1\}^{\mathbb{N}}$ be the sample space for infinite coin tossing. Then there is no mapping $P : \mathcal{P}(\Omega) \to [0, 1]$ with the properties*

- *(N) Normalisation. $P(\Omega) = 1$.*
- *(A) σ-Additivity. If $A_1, A_2, \ldots$ are pairwise disjoint, then*

$$P\left(\bigcup_{n\geqslant 1} A_n\right) = \sum_{n\geqslant 1} P(A_n).$$

(The probabilities of countably many incompatible events can be added up.)

- *(I) Flip invariance. For all $A \subseteq \Omega$ and $n \geqslant 1$ one has $P(T_n A) = P(A)$; here*

$$T_n : \Omega \to \Omega, \quad \omega = (\omega_1, \omega_2, \ldots) \mapsto (\omega_1, \ldots, \omega_{n-1}, 1 - \omega_n, \omega_{n+1}, \ldots)$$

is the mapping from Ω onto itself that inverts the result of the nth toss, and $T_n A = \{T_n(\omega) : \omega \in A\}$ is the image of A under T_n. (This expresses the fairness of the coin and the independence of the tosses.)

Es importante destacar que la propiedad de *Flip invariance* simplemente nos indica que la definición de la función de probabilidad P debe tener sentido, ya que puede existir P que cumpla (N) y (A) pero no (I), es decir, que no sea consistente con la naturaleza del experimento aleatorio que estamos modelando. En este caso, lo que se está expresando en concreto es que si en un conjunto A de secuencias de lanzamientos de moneda, invertimos el resultado de un lanzamiento específico (el n-ésimo), la probabilidad asignada a ese conjunto no debería cambiar, ya que hemos cambiado un resultado de manera simétrica y justa en todos los casos.

Proposición 1.1.5. *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, allora*

1) $(A_n)_{n\geqslant 1} \in \mathcal{F}$ *tal que* $A_n \subseteq A_{n+1}\ \forall n \implies P(\bigcup_{n\geqslant 1} A_n) = \lim_{n\to\infty} P(A_n)$

2) $(B_n)_{n\geqslant 1} \in \mathcal{F}$ *tal que* $B_{n+1} \subseteq B_n\ \forall n \implies P(\bigcap_{n\geqslant 1} B_n) = \lim_{n\to\infty} P(B_n)$

1.2. Variabili aleatorie

Observación. Trabajaremos siempre con un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ fijo.

Definición 1.2.4 (Variabile aleatoria)**.** Una **variable aleatoria** X es una función medible sobre un espacio de probabilidad. Es decir, una función:

$$X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})$$

tal que la imagen inversa de cualquier conjunto de Borel es medible. Esto se caracteriza por:

$$X^{-1}(B) \in \mathcal{F} \qquad \forall B \in \mathcal{B}(\mathbb{R})$$

Observación. Esta definición de variable aleatoria se puede aplicar a cualquier espacio medible de llegada, no solo a $\mathbb{R}$ con su $\sigma-$álgebra de Borel.

Definición 1.2.5 (Legge di probabilità)**.** Sea $X : \Omega \to \mathbb{R}$ una variable aleatoria, la **ley de probabilidad** de X es la medida de probabilidad P_X definida en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ como la siguiente función

$$\begin{array}{rccc} P_X : & (\mathbb{R}, \mathcal{B}) & \longrightarrow & [0,1] \\ & B & \longmapsto & P(X^{-1}(B)) \end{array}$$

también se nota como $\mu_X(B) = P_X(B) = P[X \in B]$.

Definición 1.2.6 (Funzione di distribuzione (cumulativa)/funzione di ripartizione)**.** Se define la **función de distribución (acumulativa)** de una variable aleatoria como:

$$\begin{array}{rccc} F_X: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & P_X\left(]-\infty, x]\right) \end{array}$$

también se nota como $F_X(x) = P(X \leqslant x) = \mu_X(]-\infty, x])$.

Corolario 1.2.5.1. *Sea X una variable aleatoria, y sea $(z_n)_{n\geqslant 1}$ una sucesión de números reales tal que $z_n \geqslant z_{n-1} \geqslant \ldots \geqslant z_1 = z$ y $\lim_{n\to\infty} z_n = z$, entonces*

$$\lim_{n\to\infty} F_X(z_n) = F_X(z)$$

Teorema 1.2.6 (Probability rules)**.** *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, se cumplen las siguientes propiedades para eventos arbitrarios $A, B, A_1, A_2, \cdots \in \mathcal{F}$.*

(a) $P(\emptyset) = 0$

(b) ***Aditividad finita****: $P(A \cup B) + P(A \cap B) = +P(A) + P(B)$, y en particular $P(A) + P(A^c) = 1$*

(c) ***Monotonía****: si $A \subseteq B$ entonces $P(A) \leqslant P(B)$*

(d) $\sigma-$***subaditividad****: $P\left(\bigcup_{i=1}^{\infty} A_i\right) \leqslant \sum_{i=1}^{\infty} P(A_i)$*

(e) $\sigma-$***continuidad****: si $A_n \nearrow A$ o $A_n \searrow A$ (es decir, los A_n son crecientes con unión A, o decrecientes con intersección A), entonces $P(A_n) \to P(A)$ cuando $n \to \infty$*

1.2.1. Propietà funzione di ripartizione

Proposición 1.2.7. *Sea X una variable aleatoria, entonces su función de distribución F_X cumple*

1) F_X es monótona no decreciente

2) $\lim_{x\to-\infty} F_X(x) = 0$ y $\lim_{x\to+\infty} F_X(x) = 1$

3) si $z_n \searrow z$, entonces $\lim_{n\to\infty} F_X(z_n) = F_X(z)$

4) se $z_n \nearrow z$, entonces $\lim_{n\to\infty} F_X(z_n) = F_X(z) \iff P(X = z) = 0$

Definición 1.2.7 (Variabile aleatoria discreta)**.** Sea una variable aleatoria $X : \Omega \to \mathbb{R}$ con $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, se dice que es **discreta** si

$$X(\Omega) = \{X(w) : w \in \Omega\}$$

es a lo máximo numerable.

Definición 1.2.8 (Variabile aleatoria continua)**.** Una variable aleatoria X se dice **continua** si

$$P(X = x) = 0, \qquad \forall x \in \mathbb{R}$$

o equivalentemente, si su función de distribución F_X es continua en todo punto de $\mathbb{R}$.

Definición 1.2.9 (Misura)**.** Una **medida** es cualquier función que asigna un "tamaño" o "peso" a los conjuntos de una σ-álgebra. Para que sea una medida, solo debe cumplir tres reglas básicas

1) **No negatividad**: el tamaño no puede ser negativo ($\mu(A) \geqslant 0$).

2) **Vacío nulo**: el conjunto vacío mide cero ($\mu(\emptyset) = 0$).

3) **σ-aditividad**: el tamaño de la unión de piezas disjuntas es la suma de los tamaños de las piezas
$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$$

Definición 1.2.10 (Misura assolutamente continua)**.** Sea un espacio medible $(\Omega, \mathcal{F})$ y dos medidas μ y ν definidas en él. Decimos que μ es **absolutamente continua** respecto a ν (y lo denotamos como $\mu \ll \nu$) si se cumple que

$$\text{si } \nu(A) = 0 \implies \mu(A) = 0 \quad \forall A \in \mathcal{F}$$

Teorema 1.2.8 (Teorema de Radon-Nikodym)**.** *Sea $(\Omega, \mathcal{F})$ un espacio medible y μ y ν dos medidas definidas en él. Si μ es absolutamente continua respecto a ν (es decir, $\mu \ll \nu$), entonces existe una función integrable $f : \Omega \to [0, +\infty]$ ($f \in L^1_\mu$) tal que*

$$\mu(A) = \int_A f(x)\, d\nu(x)$$

Proposición 1.2.9. *Sea X una v.a. ass. continua con función de densidad f_X, entonces se tiene*

$$E[g(X)] = \int_{\mathbb{R}} g(z) f_X(z) dz$$

donde $g : \mathbb{R} \to \mathbb{R}$ positiva.

Demostración. **Paso 1**: Funciones Indicadoras (El átomo). Supongamos primero que g es una función indicadora sobre un conjunto de Borel A, es decir

$$g(x) = \mathbb{1}_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

- Lado izquierdo: $E[g(X)] = E[\mathbb{1}_A(X)]$. Por definición, la esperanza de una indicadora es la probabilidad del suceso: $P(X \in A)$.
- Lado derecho: $\int_{\mathbb{R}} \mathbb{1}_A(z) f_X(z) dz = \int_A f_X(z) dz$. Por definición de variable absolutamente continua, esta integral es exactamente $\mu_X(A)$, que es $P(X \in A)$.

Conclusión: la igualdad se cumple para funciones indicadoras.
Paso 2: Funciones Simples (Linealidad). Una función simple es una combinación lineal de indicadoras $g(x) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(x)$.Debido a que tanto la Esperanza (E) como la Integral $(\int)$ son operadores lineales, podemos sacar las constantes y separar las sumas

$$E\left[\sum a_i \mathbb{1}_{A_i}(X)\right] = \sum a_i E[\mathbb{1}_{A_i}(X)] = \sum a_i \int_{A_i} f_X(z)dz =$$
$$= \int_{\mathbb{R}} \left(\sum a_i \mathbb{1}_{A_i}(z)\right) f_X(z)dz = \int_{\mathbb{R}} g(z) f_X(z)dz$$

Paso 3: Funciones No Negativas (Límite Monótono). Aquí es donde ocurre la "magia"del análisis. Si g es una función medible y positiva $(g \geqslant 0)$, existe una sucesión de funciones simples $(g_n)_{n\geqslant 1}$ que crece hacia g punto a punto $(g_n \nearrow g)$.
Aplicamos el Teorema de la Convergencia Monótona (TCM): como $g_n \nearrow g$, entonces $E[g_n(X)] \to E[g(X)]$. Como ya demostramos en el Paso 2 que la igualdad vale para cada g_n tenemos lo siguiente

$$E[g_n(X)] = \int_{\mathbb{R}} g_n(z) f_X(z)dz$$

Al tomar el límite en ambos lados, el TCM nos permite "meter.[el] límite dentro de la integral

$$\lim_{n\to\infty} \int_{\mathbb{R}} g_n(z) f_X(z)dz = \int_{\mathbb{R}} g(z) f_X(z)dz$$

□

Observación. Si $f \in L^1_\mu$ se debe cumplir que la integral del valor absoluto de la función sea finita, es decir

$$\int_\Omega |f(\omega)|\, d\mu(\omega) < \infty$$

Si esta condición se cumple, decimos que f es integrable.

Observación. El tipo de integrales con las que estamos trabajando son integrales de Lebesgue, que generalizan las integrales de Riemann y permiten integrar funciones más generales en espacios medibles.Si la medida ν tiene una función de densidad $\rho(x)$ (como la densidad de una normal o una binomial), entonces se cumple que

$$d\nu(x) = \rho(x)\, dx$$

En este caso, la integral .[ex]traña"se convierte en una integral común

$$\int_A f(x)\, d\nu(x) = \int_A f(x)\rho(x)\, dx$$

Definición 1.2.11 (Variabile aleatoria assolutamente continua)**.** Una variable aleatoria X se dice **absolutamente continua** si

$$\mu_X(A) = \int_A f_X(x)\, dx, \qquad \forall A \in \mathcal{B}(\mathbb{R})$$

para una función $f_X : \mathbb{R} \to [0, +\infty]$ positiva e integrable sobre $\mathbb{R}$. Modo alternativo $\mu_x \ll Leb$.

Observación. 1) $\int_{\mathbb{R}} f(x)dx = \mu_X(\mathbb{R}) = P(X \in \mathbb{R}) = P(\Omega) = 1$ donde f es la función de densidad de X.

2) si X es absolutamente continua $\implies$ X es continua.

1.3. Richiami di combinatoria

Estudiaremos las diferentes formas de contar arreglos, selecciones o agrupaciones de elementos de un conjunto finito. Consideremos un conjunto Ω con n elementos distintos, es decir, $\Omega = \{1, 2, \ldots, n\}$, con por ejemplo n bolas enumeradas.

1) **Disposizioni di classe k di n elementi con ripetizioni con ordine**: las *disposiciones de clase k con repeticiones y con orden* son los vectores $(x_1, x_2, \ldots, x_k)$ con $x_i \in \Omega$. Forman el siguiente conjunto

$$D_n^k = \{(x_1, \ldots, x_k) : x_i \in \Omega\} \text{ y su cardinal es } |D_n^k| = n^k$$

2) **Disposizioni di classe k di n elementi senza ripetizioni con ordine**: las *disposiciones de clase k sin repeticiones y con orden* son los vectores $(x_1, x_2, \ldots, x_k)$ con $x_i \in \Omega$ y $x_i \neq x_j$ para $i \neq j$. Forman el siguiente conjunto

$$D_{sr,n}^k = \{(x_1, \ldots, x_k) : x_i \in \Omega, x_i \neq x_j \text{ per } i \neq j\}, \text{ y su cardinal es } |D_{sr,n}^k| = \frac{n!}{(n-k)!}$$

3) **Disposizioni di classe k di n elementi senza ripetizioni senza ordine (=combinazioni)**: las *combinaciones de clase k sin repeticiones y sin orden* son las anteriores pero sin importar el orden, es decir, los subconjuntos $\{x_1, x_2, \ldots, x_k\} \sim (y_1, y_2, \ldots, y_k)$ se diferencian en una permutación, por lo que forman el cardinal es

$$C_n^k = \frac{D_{sr,n}^k}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

1.4. Distribuciones Discretas

En esta sección, recordaremos algunas de las distribuciones de probabilidad discretas más importantes que se utilizan en estadística y probabilidad.

1.4.1. Distribución uniforme discreta

Definición 1.4.12. Sea X una variable aleatoria. Se dice que la variable aleatoria X se distribuye uniformemente alrededor de los puntos $x_1, \ldots, x_n$ si dichos valores son equiprobables. Se notará como sigue:

$$X \rightsquigarrow U(x_1, \ldots, x_n)$$

Proposición 1.4.10. *Sea X una variable aleatoria tal que $X \rightsquigarrow U(x_1, \ldots, x_n)$. Entonces, su función masa de probabilidad es:*

$$f(x) = P[X = x] = \begin{cases} 1/n & x = x_i, i = 1, \ldots, n \\ 0 & \textit{en caso contrario} \end{cases}$$

Cuando $Re_X = \{1, 2, \ldots, n\}$, veamos algunos casos particulares de series:

$$E[X] = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2}$$
$$E[X^2] = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{1}{n}\frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$
$$Var(X) = \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$$

1.4.2. Distribución binomial

Definición 1.4.13 (Distribución binomial)**.** Se dice que una variable aleatoria X sigue una distribución binomial de parámetros n y p, $n \in \mathbb{N}$, $p \in]0, 1[$ si modela el número de éxitos en n repeticiones independientes de un ensayo de Bernouilli con probabilidad p de éxito, manteniéndose esta constante en las n repeticiones. Se notará como

$$X \rightsquigarrow B(n, p)$$

Observación. Es fácil ver que la distribución de Bernouilli es un caso particular de una distribución binomial, considerando $n = 1$.

Proposición 1.4.11. *Sea $X \rightsquigarrow B(n, p)$. Entonces, la* ***función masa de probabilidad*** *de la distribución binomial es:*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x \in \{0, \ldots, n\}$$

Proposición 1.4.12. *Sea $X \rightsquigarrow B(n, p)$. Entonces, tenemos que:*

$$E[X] = np$$

Proposición 1.4.13. *Sea $X \rightsquigarrow B(n, p)$. Entonces, tenemos que:*

$$\mathrm{Var}[X] = np(1-p)$$

1.4.3. Distribución binomial negativa

Definición 1.4.14 (Distribución binomial negativa)**.** Se dice que una variable aleatoria X sigue una distribución binomial negativa de parámetros r y p, $r \in \mathbb{N}$, $p \in]0, 1[$ si modela el número de fracasos antes de llegar al $r-$ésimo éxito en un ensayo de Bernouilli con probabilidad p de éxito, manteniéndose esta constante en todas las repeticiones. Se notará como

$$X \rightsquigarrow BN(r, p)$$

Observación. Es fácil ver que la Distribución Geométrica es un caso particular de una distribución binomial negativa, considerando $r = 1$.

Razonemos la función masa de probabilidad. Como la variable X modela el número de fracasos antes de llegar al $r-$ésimo éxito, tenemos que en total se producen x fracasos y r fracasos. Como la probabilidad se mantiene constante, tenemos que la probabilidad para cierta ordenación de los sucesos es $p^r(1-p)^x$. No obstante, las distintas formas de reorganizar los sucesos son combinaciones de x fracasos y $r-1$ éxitos[1]; es decir, combinaciones sin repetición. Por tanto, tenemos que en total hay $\binom{x+r-1}{x}$ formas distintas. Por tanto, a priori deducimos que la función masa de probabilidad de la distribución binomial negativa es:

$$f(x) = \binom{x+r-1}{x}(1-p)^x p^r \qquad x \in \mathbb{N} \cup \{0\}$$

No obstante, hay una expresión alternativa de la función masa de probabilidad muy útil para las demostraciones. Para ello, se introduce la siguiente definición:

Definición 1.4.15. Dado $r \in \mathbb{R}, x \in \mathbb{N}$, se define[2] el siguiente número combinatorio:

$$\binom{-r}{x} = \frac{(-r)(-r-1)\cdots(-r-x+1)}{x!};\ \binom{-\alpha}{0} = 1, \qquad \forall r \in \mathbb{R}, x \in \mathbb{N}$$

Proposición 1.4.14. *Sea $X \rightsquigarrow BN(r,p)$. Entonces, la **función masa de probabilidad** de la distribución binomial negativa es:*

$$f(x) = \binom{x+r-1}{x}(1-p)^x p^r = \binom{-r}{x}(p-1)^x p^r \qquad x \in \mathbb{N} \cup \{0\}$$

Proposición 1.4.15. *Sea $X \rightsquigarrow BN(r,p)$. Entonces, tenemos que:*

$$E[X] = \frac{r(1-p)}{p}$$

Lema 1.4.16. *Sea $X \rightsquigarrow BN(r,p)$. Entonces, tenemos que:*

$$E[X^2] = \frac{r(r+1)(1-p)^2}{p^2} + \frac{r(1-p)}{p}$$

Corolario 1.4.16.1. *Sea $X \rightsquigarrow BN(r,p)$. Entonces, tenemos que:*

$$\mathrm{Var}[X] = \frac{r(1-p)}{p^2}$$

1.4.4. Distribución de Poisson

Definición 1.4.16 (Distribución de Poisson)**.** Sea X una variable aleatoria. Se dice que X sigue una distribución de Poisson si representa el número de de ocurrencias de un determinado suceso durante un periodo de tiempo fijo o una región fija del espacio.

Ha de cumplir las siguientes condiciones:

[1] El éxito $r-$ésimo no se añade porque su posición está fijada, ya que debe ser el último suceso.

[2] El número combinatorio está ya definido, pero con más restricciones. Esta definición concuerda con la anterior en los supuestos anteriores.

1. El número de ocurrencias en un intervalo o región específicas ha de ser independiente de las ocurrencias en otras zonas o intervalos de tiempo.
2. Si se considera un intervalo de tiempo muy pequeño (o una región muy pequeña):
   - La probabilidad de una ocurrencia es proporcional a la longitud del intervalo (el volumen de la región).
   - La probabilidad de dos o más ocurrencias es despreciable.

Dicho de otra manera, si de media se tiene que en determinado intervalo se producen λ ocurrencias, tenemos que sigue una distribución de Poisson de parámetro λ, notado por:

$$X \rightsquigarrow \mathcal{P}(\lambda)$$

Proposición 1.4.17. *Sea $X \rightsquigarrow \mathcal{P}(\lambda)$. Entonces, su* ***función masa de probabilidad*** *es:*

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Proposición 1.4.18. *Sea $X \rightsquigarrow \mathcal{P}(\lambda)$. Entonces, tenemos que:*

$$E[X] = \lambda$$

Lema 1.4.19. *Sea $X \rightsquigarrow \mathcal{P}(\lambda)$. Entonces, tenemos que:*

$$E[X^2] = \lambda^2 + \lambda$$

Corolario 1.4.19.1. *Sea $X \rightsquigarrow \mathcal{P}(\lambda)$. Entonces, tenemos que:*

$$\mathrm{Var}[X] = \lambda$$

1.5. Distribuciones Continuas

En esta sección, recordaremos algunas de las distribuciones de probabilidad continuas más importantes que se utilizan en estadística y probabilidad.

1.5.1. Distribución Uniforme Continua

Definición 1.5.17 (Distribución Uniforme Continua)**.** Una variable aleatoria continua X sigue una distribución uniforme en el intervalo $[a, b]$, con $a, b \in \mathbb{R}$, $a < b$, si su función de densidad toma un valor constante en dicho intervalo, siendo nula fuera de él. Lo denotaremos como $X \sim \mathcal{U}(a, b)$.

Proposición 1.5.20. *Sea $X \sim \mathcal{U}(a, b)$. Entonces, se tiene que:*

$$\begin{array}{rccl} f: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & \begin{cases} \dfrac{1}{b-a} & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases} \end{array}$$

Proposición 1.5.21. *Sea $X \sim \mathcal{U}(a,b)$, entonces su función de distribución es:*

$$\begin{array}{rccl} F_X: & \mathbb{R} & \longrightarrow & [0,1] \\ & x & \longmapsto & \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & x \in [a,b] \\ 1 & x > b \end{cases} \end{array}$$

Calculemos ahora los momentos de una variable aleatoria X tal que $X \sim \mathcal{U}(a,b)$.

Proposición 1.5.22. *Sea X una variable aleatoria tal que $X \sim \mathcal{U}(a,b)$. Entonces, su momento no centrado de orden $k \in \mathbb{N} \cup \{0\}$ es:*

$$m_k = E[X^k] = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}$$

Como consecuencia, tenemos que:

$$m_1 = E[X] = \frac{b^2 - a^2}{2(b-a)} = \frac{b+a}{2}$$

1.5.2. Distribución Normal

Esta es la distribución de probabilidad más importante en la Teoría de la Probabilidad y la Estadística Matemática.

Definición 1.5.18 (Distribución Normal)**.** Una variable aleatoria continua X sigue una distribución normal con parámetros $\mu \in \mathbb{R}$ y $\sigma^2 \in \mathbb{R}^+$, si su función de densidad es:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\dfrac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}$$

donde $\sigma = +\sqrt{\sigma^2}$. Lo denotaremos como $X \sim \mathcal{N}(\mu, \sigma^2) = \mathcal{N}_{\mu,\sigma^2}$.

A pesar de darse como definición, hemos de demostrar que efectivamente es una función de densidad. Para ello, incluimos el siguiente Lema, cuya demostración no se incluye por su complejidad, al requerir de integración en varias variables con cambio a coordenadas polares.

Lema 1.5.23 (Integral de Gauss)**.** *Sea $a \in \mathbb{R}^+$, $b \in \mathbb{R}$. Entonces:*

$$\int_{-\infty}^{\infty} e^{-a(x+b)^2}\, dx = \sqrt{\frac{\pi}{a}}$$

1.5.3. Distribución Exponencial

Definición 1.5.19 (Distribución Exponencial)**.** Una variable aleatoria continua X sigue una distribución exponencial con parámetro $\lambda \in \mathbb{R}^+$, si su función de densidad es:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

Lo denotaremos como $X \sim \exp(\lambda)$.

Veamos algunas aplicaciones de esta distribución:

- La distribución exponencial se utiliza para modelar el tiempo aleatorio entre dos fallos consecutivos en *fiabilidad* o entre dos llegadas consecutivas en *teoría de colas*.
- También se aplica en la modelización de tiempos aleatorios de supervivencia (*Análisis de Supervivencia*).
- En general, X suele representar un tiempo aleatorio transcurrido entre dos sucesos, que se producen de forma aleatoria y consecutiva en el tiempo. Dichos sucesos se contabilizan mediante un proceso de Poisson homogéneo. El parámetro λ representa la razón de ocurrencia de dichos sucesos, que en este caso es constante.

Proposición 1.5.24. *Sea X una variable aleatoria tal que $X \sim \exp(\lambda)$. Entonces, su función de distribución es:*

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

Observación. Dada una v.a. $X \sim \exp(\lambda)$, se tiene que

$$F_X(t) = P(X < t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda t}, \qquad t \geqslant 0$$

Proposición 1.5.25. *Sea X una variable aleatoria de forma que $X \sim \exp(\lambda)$. Entonces, sus momentos no centrados de orden $k \in \mathbb{N}$ son:*

$$E[X^k] = \frac{k!}{\lambda^k}$$

Como consecuencia, tenemos que:

$$E[X] = \frac{1}{\lambda} \qquad \text{Var}[X] = E[X^2] - E[X]^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

1.6. Distribuciones Reproductivas

Definición 1.6.20 (Reproductividad de distribuciones)**.** Una distribución de variables aleatorias se dice reproductiva si, dadas X_1, X_2, ..., X_n variables aleatorias que sigan una distribución de dicho tipo (aunque la distribución de cada variable tenga distintos parámetros), entonces la variable aleatoria dada por:

$$X = \sum_{i=1}^{n} X_i$$

sigue una distribución del mismo tipo.

Proposición 1.6.26 (Distribución Reproductiva - Binomial)**.** *Sean $X_1, \ldots, X_n$ variables aleatorias independientes y $X_i \sim B(k_i, p)$ para $i = 1, \ldots, n$. Entonces:*

$$\sum_{i=1}^{n} X_i \sim B\left(\sum_{i=1}^{n} k_i, p\right)$$

Proposición 1.6.27 (Distribución Reproductiva - Poisson)**.** *Sean $X_1, \ldots, X_n$ variables aleatorias independientes y $X_i \sim \mathcal{P}(\lambda_i)$ para $i = 1, \ldots, n$. Entonces:*

$$\sum_{i=1}^{n} X_i \sim \mathcal{P}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Proposición 1.6.28 (Distribución Reproductiva - Binomial Negativa)**.** *Sean $X_1, \ldots, X_n$ variables aleatorias independientes y $X_i \sim BN(k_i, p)$ para $i = 1, \ldots, n$. Entonces:*

$$\sum_{i=1}^{n} X_i \sim BN\left(\sum_{i=1}^{n} k_i, p\right)$$

Proposición 1.6.29. *Sean $X_1, \ldots, X_n$ variables aleatorias independientes y $X_i \sim G(p)$ para $i = 1, \ldots, n$. Entonces:*

$$\sum_{i=1}^{n} X_i \sim BN(n, p)$$

Proposición 1.6.30 (Distribución Reproductiva - Normal)**.** *Sean $X_1, \ldots, X_n$ variables aleatorias independientes y $X_i \sim N(\mu_i, \sigma_i^2)$ para $i = 1, \ldots, n$. Entonces:*

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

2. Statistica

Notación. Cuando hablamos de espacios muestrales discretos, la $\sigma-$álgebra la notamos como $\mathcal{F} = 2^{\Omega}$.

2.1. Distribuzioni uniformi

Definición 2.1.1. (Distribución uniforme discreta) Sea el espacio muestral $\Omega = \{x_1, x_2, \ldots, x_n\}$ con n resultados posibles. Una variable aleatoria X que toma cada valor x_i con probabilidad $\frac{1}{n}$ se dice que sigue una **distribución uniforme discreta** en $\{x_1, x_2, \ldots, x_n\}$, de manera que

$$\mathcal{U}_{\Omega}(A) = P(A) = \frac{\#A}{\#\Omega} = \frac{\text{casi favorevoli}}{\text{casi possibili}} \quad \forall A \subseteq \Omega$$

donde tenemos $\mathcal{F} = \mathcal{P}(\Omega) = 2^{\Omega}$, y se denota como $X \sim U\{x_1, x_2, \ldots, x_n\}$.

Definición 2.1.2. Sea $\Omega \subseteq \mathbb{R}^n$ un conjunto de Borel con volumen n-dimensional $0 < \lambda^n(\Omega) < +\infty$. La medida de probabilidad $\mathcal{U}_{\Omega}$ en $(\Omega, \mathcal{B}(\Omega))$ con densidad de Lebesgue constante $f(x) = \frac{1}{\lambda^n(\Omega)}$, dada por

$$\mathcal{U}_{\Omega}(A) = \int_A f(x)\, d\lambda^n(x) = \int_A \frac{1}{\lambda^n(\Omega)}\, d\lambda^n(x) = \frac{\lambda^n(A)}{\lambda^n(\Omega)}, \quad A \in \mathcal{B}(\Omega)$$

se llama **distribución uniforme continua** en Ω

Teorema 2.1.1 (Teorema de unicidad, Dynkin)**.** *Sean P_1, P_2 dos medidas de probabilidad en el espacio de probabilidad $(\Omega, \mathcal{F})$. Si $\mathcal{D} \in \mathcal{F}$ cumple*

1) $\sigma(\mathcal{D}) = \mathcal{F}$, es decir, $\mathcal{D}$ genera $\mathcal{F}$

2) $A, B \in \mathcal{D} \implies A \cap B \in \mathcal{D}$

3) $P_{1|_D} = P_{2|_D}$, es decir, $P_1(A) = P_2(A) \quad \forall A \in \mathcal{D}$

entonces $P_1 = P_2$.

2.2. Modelli a urne (con reinserimento)

Tenemos una urna con N bolas, donde N_{a_i} indica el número de bolas de tipo a_i y el conjunto de distintos tipo de bolas es $E = \{a_1, a_2, \ldots, a_{|E|}\}$. Sabemos además que $N = \sum_{a \in E} N_a = N$.

Se quieren extraer n bolas de la urna, devolviendo cada bola extraída antes de proceder a la siguiente extracción (reinsserción), por lo que el espacio muestral es $\Omega = \{(\omega_1, \omega_2, \ldots, \omega_n) : \omega_i \in E\} = E^n$.

2.2.1. Caso ordenado

Suponiendo que cada bola tiene la misma probabilidad de ser extraída, tenemos que la probabilidad de obtener una secuencia concreta $(\omega_1, \omega_2, \ldots, \omega_n)$ de las especies de las bolas es

$$P(\{(\omega_1, \omega_2, \ldots, \omega_n)\}) = \frac{N_{\omega_1}}{N} \cdot \ldots \cdot \frac{N_{\omega_n}}{N} = \prod_{i=1}^{n} p(\omega_i)$$

donde $p : E \to [0, 1]$ es la función de probabilidad de extraer una bola de tipo a_i, es decir, $p(a_i) = \dfrac{N_{a_i}}{N}$. En este caso estamos trabajando en función de las especies extraídas y no respecto a las bolas concretas.

Probemos a intentar ver la situación desde otra perspectiva, considerando el espacio muestral como $\overline{\Omega} = \{1, \ldots, N\}^n$, es decir, cada bola corresponde a un número desde 1 hasta N (están enumeradas). Definiremos los conjuntos de cada tipo de bola como $C_a \subseteq \{1, \ldots, N\}$, es decir, C_a es el conjunto de números que corresponden a bolas de tipo a, y por tanto $|C_a| = N_a$.

En este sentido, tenemos que la probabilidad uniforme es la siguiente

$$\overline{P}(\{\bar{\omega}\}) = \frac{1}{N^n} \quad \forall \bar{\omega} \in \overline{\Omega}$$

ya que todas las bolas son diferentes, pues estan enumeradas (trabajamos ahora con la numeración de las bolas y no con el tipo). Definamos ahora las variables aleatorias siguientes

$$X_i : \overline{\Omega} \to \Omega, \quad X(\bar{\omega}_1, \ldots, \bar{\omega}_n) = a \text{ si } \bar{\omega}_i \in C_a$$

donde a es la especie de $\bar{\omega}_i$. De esta manera podemos definir la variable aleatoria conjunta

$$X = (X_1, \ldots, X_n) : \overline{\Omega} \to \Omega, \quad X(\bar{\omega}_1, \ldots, \bar{\omega}_n) = (X_1(\bar{\omega}_1, \ldots, \bar{\omega}_n), \ldots, X_n(\bar{\omega}_1, \ldots, \bar{\omega}_n))$$

que nos da la especie de cada bola extraída. Estamos en situación por tanto de definir la probabilidad uniforme para el espacio muestral Ω como $P = \overline{P} \circ X^{-1}$, es decir,

$$P(\{(\omega_1, \ldots, \omega_n)\}) = \overline{P}(X = (\omega_1, \ldots, \omega_n)) \quad \forall (\omega_1, \ldots, \omega_n) \in \Omega \quad \equiv \quad P(A) = \overline{P}(X^{-1}(A)) \quad \forall A \subseteq \Omega$$

donde si además tenemos en cuenta que $\{X = (\omega_1, \ldots, \omega_n)\} = C_{\omega_1} \times \ldots \times C_{\omega_n}$, concluimos que

$$P(\{(\omega_1, \ldots, \omega_n)\}) = \overline{P}(C_{\omega_1} \times \ldots \times C_{\omega_n}) = \frac{|C_{\omega_1}| \cdot \ldots \cdot |C_{\omega_n}|}{N^n} = \prod_{i=1}^{n} p(\omega_i)$$

donde $p(\omega_i) = \dfrac{N_{\omega_i}}{N}$, obteniendo finalmente la misma probabilidad que habíamos deducido antes, habiendo desarrollado desde otro punto de vista.

Es importante notar que la expresión $\{X = (\omega_1, \ldots, \omega_n)\} = C_{\omega_1} \times \ldots \times C_{\omega_n}$, viene de pensar que si tenemos una secuencia concreta de especies $(\omega_1, \ldots, \omega_n) \in \Omega$, para que la variable aleatoria conjunta X tome esos valores, la primera bola extraída debe pertenecer al conjunto C_{ω_1} (es decir, debe ser una bola de tipo ω_1), la segunda bola extraída debe pertenecer al conjunto C_{ω_2}, y así sucesivamente hasta la n-ésima bola. Por tanto, el conjunto de todas las secuencias de bolas numeradas que cumplen esta condición es precisamente el producto cartesiano $C_{\omega_1} \times C_{\omega_2} \times \ldots \times C_{\omega_n}$.

Debemos destacar que hemos podido demostrar la probabilidad de extraer una secuencia concreta de especies de bolas $(\omega_1, \omega_2, \ldots, \omega_n)$ de dos maneras distintas: una directamente pensando en las probabilidades de extraer cada especie en cada extracción, y otra pensando en la numeración de las bolas y usando variables aleatorias que nos permiten mapear la numeración a las especies. Ambas formas son válidas y nos llevan al mismo resultado final, gracias en que hay reinserción, de manera que la probabilida de sacar una bola de tipo a en cada extracción es siempre la misma, independientemente de las extracciones anteriores. Esto es un detalle que en el caso de NO reinserción tendremos que separar, es decir, el caso de bolas enumeradas y el caso de especies de bolas no serán equivalentes.

Definición 2.2.3 (Misura prodotto con densità discrete)**.** Sea $E = \{a_1, a_2, \ldots, a_{|E|}\}$ (un conjunto finito) y $p : E \to [0, 1]$ una función de probabilidad discreta en E, es decir, $\sum_{a \in E} p(a) = 1$. La medida de probabilidad P se dice **medida producto** en el espacio de probabilidad $(\Omega, \mathcal{F})$ y es de la forma

$$P(\{(\omega_1, \omega_2, \ldots, \omega_n)\}) = \prod_{i=1}^{n} p(\omega_i) \quad \forall (\omega_1, \omega_2, \ldots, \omega_n) \in \Omega = E^n$$

y su función de densidad es

$$p^{\otimes n}(\omega) = \prod_{i=1}^{n} p(\omega_i) \quad \forall \omega = (\omega_1, \omega_2, \ldots, \omega_n) \in E^n$$

llamada **n-ésima densidad de producto de** p.

Ejemplo. Sea $E = \{0, 1\}$ y $p(1) = p \in [0, 1]$, obtenemos la densidad producto tal que

$$p^{\otimes n}(\omega) = \prod_{i=1}^{n} p(\omega_i) = p^{\sum_{i=1}^{n} \omega_i} (1 - p)^{n - \sum_{i=1}^{n} \omega_i} \quad \forall \omega = (\omega_1, \omega_2, \ldots, \omega_n) \in \{0, 1\}^n$$

y por tanto $P(\omega) = p^{\otimes n}(\omega)\ \forall \omega \in \{0, 1\}^n$. Esta distribución de probabilidad se llama **distribución de Bernoulli de parámetro** p **con** n **lanzamientos** dada por $Ber(n, p)$, y modela la probabilidad de obtener x éxitos en un experimento de lanzar una moneda sesgada n veces, donde la probabilidad de obtener cara (éxito) es p y la de obtener cruz (fracaso) es $1 - p$.

2.2.2. Caso no ordenado

Vamos a estudiar ahora el caso en el que no nos importa el orden en el que se extraen las bolas, sino únicamente la cantidad de bolas de cada especie extraída. Para ellos, definiendo $\Omega = E^n$ como antes, tenemos la medida de probabilidad P.
Definamos ahora otro espacio muestral en el mismo sentido del problema

$$\hat{\Omega} = \{ \overrightarrow{k} = (k_a)_{a\in E} : k_a \geqslant 0, \sum_{a\in E} k_a = n\}$$

donde este espacio muestral no contiene secuencias concretas de bolas extraídas, sino que contiene vectores $\overrightarrow{k}$ que indican cuántas bolas de cada especie $a \in E$ han sido extraídas en total, sin importar el orden. Es claro que la restricción $\sum_{a\in E} k_a = n$ es necesaria, ya que el total de bolas extraídas es n, y debemos tener en cuenta de contar un total de n bolas entre todas las especies.
Definamos la variable aleatoria siguiente que nos permite "olvidarnos del orden"

$$S : \Omega \to \hat{\Omega}, \quad S(\omega_1, \omega_2, \ldots, \omega_n) = (S_a)_{a\in E}$$

dove

$$S_a(\omega) = \sum_{i=1}^{n} \mathbb{1}_{\{a\}}(\omega_i), \quad \forall \omega \in \Omega,\ a \in E$$

que calcula el número de bolas de cada especie a en la secuencia $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$, consiguiendo obtener el vector del que hablabamos antes en el espacio muestral $\hat{\Omega}$.
Como medida producto sobre $\hat{\Omega}$, definimos $\hat{P} = P \circ S^{-1}$, es decir,

$$\hat{P}(\{\overrightarrow{k}\}) = P(S^{-1}(\overrightarrow{k})) = P(S = \overrightarrow{k}) \quad \forall \overrightarrow{k} \in \hat{\Omega}$$

donde $\{S = \overrightarrow{k}\} = \bigcup_{\omega: S_a(\omega) = \overrightarrow{k}} \{\omega\}$, teniendo por tanto

$$\hat{P}(\{\overrightarrow{k}\}) = P(S = \overrightarrow{k}) = \sum_{\omega: S_a(\omega) = \overrightarrow{k}} P(\{\omega\})$$

Para calcular $P(S = \overrightarrow{k})$, debemos contar cuántas secuencias $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ en Ω cumplen que para cada especie $a \in E$, el número de veces que a aparece en la secuencia es k_a. Tenemos que estudiar por tanto dos aspectos importantes: cuántas secuencias ω cumplen $S_a(\omega) = \overrightarrow{k}$ (lo que decíamos anteriormente) y cuál es la probabilidad $P(\{\omega\})$ de cada una de estas secuencias, veámoslo

1) Para calcular cada una de las secuencias seguimos lo explicado en el *caso ordenado*, es decir, cada secuencia $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ tiene probabilidad

$$P(\{\omega\}) = \prod_{i=1}^{n} p(\omega_i) = \prod_{i=1}^{n} \frac{N_{\omega_i}}{N} = \prod_{a\in E} p(a)^{k_a}$$

ya que en la secuencia ω, la especie a aparece k_a veces para cada $\overrightarrow{k} \in \hat{\Omega}$, y $\frac{N_{\omega_i}}{N} = p(\omega_i)$ significa la probabilidad de extraer una bola de tipo ω_i.

2) Para calcular cuántas secuencias distintas ω producen el mismo vector de recuento $\overrightarrow{k}$, necesitamos trabajar con un problema clásico de combinatoria: número de secuencias con n elementos donde hay $k_{a_1}, k_{a_2}, \ldots$ repeticiones de cada tipo de bola. Este número es el coeficiente multinomial

$$\text{Número de secuencias} = \frac{n!}{\prod_{a\in E} k_a!}$$

donde $n = \sum k_a$.
Veámos ahora una explicación más detallada del problema de combinatoria estudiado. Imagina que tienes n huecos vacíos en una fila y tienes que decidir dónde poner cada color. Este enfoque es muy intuitivo

- **Paso 1**: de los n huecos, eliges k_1 para el primer color. ¿De cuántas formas? De $\binom{n}{k_1}$ formas.
- **Paso 2**: ahora te quedan $n - k_1$ huecos libres. De esos, eliges k_2 para el segundo color, teniendo $\binom{n-k_1}{k_2}$ formas.
- **Paso 3**: repites el proceso hasta que no queden huecos.

Si multiplicamos todas estas elecciones (porque son pasos sucesivos), la magia del álgebra hace que casi todo se cancele

$$\binom{n}{k_1} \cdot \binom{n-k_1}{k_2} \cdots = \frac{n!}{k_1!(n-k_1)!} \cdot \frac{(n-k_1)!}{k_2!(n-k_1-k_2)!} \cdots$$

donde se puede or cómo el denominador de una fracción cancela el numerador de la siguiente. Al final, solo sobrevive

$$\frac{n!}{k_1!k_2!\ldots k_m!} = \frac{n!}{\prod_{a\in E} k_a!}$$

Destacar que nuestro objetivo en general era poder obtener la *distribución multinomial*, y para ello hemos necesitado definir el *coeficiente multinomial*, a parte de hacer este desarrollo tanto en el caso ordenado como en el no ordenado.

Definición 2.2.4 (Coefficiente multinomiale)**.** Sean n y $\overrightarrow{k} = (k_a)_{a\in E}$ con $k_a \geqslant 0$ y $\sum\limits_{a\in E} k_a = n$. El **coeficiente multinomial** se define como

$$\binom{n}{\overrightarrow{k}} = \frac{n!}{\prod\limits_{a\in E} k_a!}$$

Teorema 2.2.2 (Teorema del coeficiente multinomial)**.** *Sea $n \in \mathbb{N}$ y sean $x_1, x_2, \ldots, x_m \in \mathbb{R}$. El teorema establece que:*

$$(x_1 + x_2 + \cdots + x_m)^n = \sum_{\substack{\overrightarrow{k}\, talque \\ k_1+k_2+\cdots+k_m=n}} \binom{n}{\overrightarrow{k}} \cdot x_1^{k_1} x_2^{k_2} \ldots x_m^{k_m}$$

donde $\overrightarrow{k} = (k_1, k_2, \ldots, k_m)$ tal que $k_i \geqslant 0$, y la suma se extiende a todos los vectores $\overrightarrow{k}$ con entradas no negativas que suman n.

Observación. Para el caso $|E| = 2$ con $E = \{a_1, a_2\}$, el coeficiente multinomial se reduce al coeficiente binomial clásico

$$\binom{n}{\overrightarrow{k}} = \frac{n!}{k_{a_1}!k_{a_2}!} = \frac{n!}{k_{a_1}!(n-k_{a_1})!} = \binom{n}{k_{a_1}}$$

Tenemos finalmente por tanto que $|\{\omega \in \Omega : S_a(\omega) = \overrightarrow{k}\}| = \binom{n}{\overrightarrow{k}}$, y por tanto

$$\hat{P}(\{\overrightarrow{k}\}) = P(S = \overrightarrow{k}) = \sum_{\omega : S_a(\omega) = \overrightarrow{k}} P(\{\omega\}) = \binom{n}{\overrightarrow{k}} \prod_{a \in E} p(a)^{k_a}$$

Definición 2.2.5 (Distribuzione multinomiale)**.** Sea E un conjunto finito con $|E| \geqslant 2$ y $\hat{\Omega} = \{\overrightarrow{k} = (k_a)_{a \in E} : k_a \geqslant 0, \sum_{a \in E} k_a = n\}$. Una medida de probabilidad $\hat{P}$ en $(\hat{\Omega}, 2^{\hat{\Omega}})$ se dice que sigue una **distribución multinomial** de parámetros n y $p : E \to [0, 1]$ (función de probabilidad discreta en E) si

$$\hat{P}(\{\overrightarrow{k}\}) = \binom{n}{\overrightarrow{k}} \prod_{a \in E} p(a)^{k_a} \quad \forall \overrightarrow{k} \in \hat{\Omega}$$

y se denota como $\hat{P} \sim \mathcal{M}(n, p)$. Alternativamente, dado $\Omega = E^n$ y $P = p^{\otimes n}$ es equivalente decir

$$\mu_S = P \circ S^{-1} \sim \mathcal{M}(n, p)$$

donde $S : \Omega \to \hat{\Omega}$ es la variable aleatoria que "olvida el orden y recuerda solo los tipos".

2.3. Modelli a urne (senza reinserimento)

Tenemos una urna con N bolas con todas distintas entre sí, y queremos extraer $n \leqslant N$ bolas sin reinsertarlas. Estudiaremos el caso ordenado y el caso no ordenado.

Además, después estudiaremos el caso en el que las bolas no son todas distintas, sino que hay distintas especies de bolas, de manera que el conjunto de distintos tipos de bolas es $E = \{a_1, a_2, \ldots, a_{|E|}\}$ con N_{a_i} bolas de tipo a_i y $N = \sum_{a \in E} N_a$, como hicimos anteriormente.

2.3.1. Caso ordenado

En este caso, el espacio muestral es

$$\overline{\Omega}_{\neq} = \{\bar{\omega} = (\bar{\omega}_1, \bar{\omega}_2, \ldots, \bar{\omega}_n) : \bar{\omega}_i \neq \bar{\omega}_j \text{ si } i \neq j\}$$

donde cada $\bar{\omega}_i$ es la numeración de una bola en la urna, suponiendo que todas están enumeradas.

Domanda.¿Cuál es la medida de probabilidad correcta?

Sabemos que la medida de probabilidad P es uniforme, por lo que $P(\{\bar{\omega}\}) = \dfrac{|\{\bar{\omega}\}|}{|\overline{\Omega}_{\neq}|} = \dfrac{1}{|\overline{\Omega}_{\neq}|} \quad \forall \bar{\omega} \in \overline{\Omega}_{\neq}$, y calculando el cardinal del espacio muestral tenemos la probabilidad buscada, veámoslo

$$|\overline{\Omega}_{\neq}| = N \cdot (N-1) \cdot \ldots \cdot (N-n+1) = \frac{N!}{(N-n)!}$$

2.3.2. Caso no ordenado

Bolas enumeradas

En este caso, definimos el espacio muestral como

$$\tilde{\Omega} = \{\tilde{\omega} \subseteq \{1, 2, \ldots, N\} : |\tilde{\omega}| = n\}$$

donde se supone que las bolas están enumeradas.

Observación. Es importante notar la pequeña diferencia con el espacio muestral $\overline{\Omega}_{\neq}$ del caso ordenado, donde las secuencias tenían en cuenta el orden de extracción, mientras que en $\tilde{\Omega}$ no importa el orden, sino únicamente qué bolas han sido extraídas, esto se almacena en el caso ordenado usando vectores $\bar{\omega}$, mientras aquí usamos simplemente $\tilde{\omega}$ como conjunto.

Es claro que todos los elementos $\tilde{\omega} \in \tilde{\Omega}$ tienen la misma probabilidad, es decir, $\tilde{P}$ es una medida de probabilida uniforme y viene dada por lo que

$$\tilde{P}(\{\tilde{\omega}\}) = \frac{|\{\tilde{\omega}\}|}{|\tilde{\Omega}|} = \frac{1}{\binom{N}{n}} \quad \forall \tilde{\omega} \in \tilde{\Omega}$$

donde para saber el cardinal del espacio muestral, sabemos que el espacio muestral tiene el número de combinaciones posibles de n elementos entre N sin contar ordenación, ya que tenemos elementos $\tilde{\omega}$ como conjuntos, lo que nos lleva a $|\tilde{\Omega}| = \dfrac{N!}{n!(N-n)!} = \binom{N}{n}$.

Veámos como podemos demostrarlo de otro modo, definiendo la variable aleatoria siguiente

$$Y : \overline{\Omega}_{\neq} \to \tilde{\Omega} \quad Y(\bar{\omega}_1, \ldots, \bar{\omega}_n) = \{\tilde{\omega_1}, \ldots, \tilde{\omega_n}\} \quad \text{donde } \tilde{\omega_i} = \bar{\omega}_i$$

Lo que queremos ahora demostrar es $\tilde{P} = P \circ Y^{-1}$, veámoslo

$$P(Y^{-1}(\{\tilde{\omega}\})) = \frac{|Y^{-1}(\{\tilde{\omega}\})|}{|\overline{\Omega}_{\neq}|} = \frac{|\{\bar{\omega} : Y(\bar{\omega}) = \tilde{\omega}\}|}{\dfrac{N!}{(N-n)!}} \overset{(*)}{=} \frac{\dfrac{n!}{N!}}{(N-n)!} = \frac{1}{\binom{N}{n}} = \tilde{P}(\{\tilde{\omega}\}) \quad \forall \tilde{\omega} \in \tilde{\Omega}$$

donde en $(*)$ hemos usado que el número de secuencias ordenadas $\bar{\omega}$ que producen el mismo conjunto $\tilde{\omega}$ es $n!$, ya que hay $n!$ maneras distintas de ordenar el conjunto $\tilde{\omega}$ en vectores de la forma $\bar{\omega}$.

Bolas de colores (especies)

Estudiaremos ahora el caso donde nos centramos en las especies de bolas, es decir, el conjunto de especies de bolas es $E = \{a_1, a_2, \dots, a_{|E|}\}$ con N_{a_i} bolas de tipo a_i y $N = \sum a \in E N_a$, y olvidamos la numeración de las bolas. Recordemos la definición de los conjuntos $C_a \subseteq \{1, \dots, N\}$ $\forall a \in E$, donde C_a es el conjunto de números que corresponden a bolas de tipo a, y por tanto $|C_a| = N_a$.
En este caso el espacio muestral es

$$\hat{\Omega} = \{\overrightarrow{k} = (k_a)_{a\in E} : k_a \geqslant 0, \sum_{a\in E} k_a = n\}$$

Domanda. ¿Cuál es la medida de probabilidad correcta?
Para responder a esta pregunta, definimos la variable aleatoria siguiente

$$T : \tilde{\Omega} \to \hat{\Omega}, \quad T(\tilde{\omega}) = T(\{\tilde{\omega_1}, \dots, \tilde{\omega_n}\}) = (|\tilde{\omega} \cap C_a|)_{a\in E} \quad \forall \tilde{\omega} \in \tilde{\Omega}$$

donde lo que intenta dicha v.a. es olvidar la numeración de las bolas y quedarse únicamente con el recuento de cuántas bolas de cada especie han sido extraídas. Definimos por tanto la medida de probabilidad $\hat{P} = \tilde{P} \circ T^{-1}$ en $(\hat{\Omega}, 2^{\hat{\Omega}})$, pues simplemente queremos trasladar la medida de probabilidad uniforme del espacio muestral $\tilde{\Omega}$ al espacio muestral $\hat{\Omega}$ usando la variable aleatoria T, que se encarga de mapear conjuntos de bolas extraídas a vectores de recuento de especies. Calculemos ahora dicha medida de probabilidad

$$\hat{P}(\{\overrightarrow{k}\}) = \tilde{P}(T^{-1}(\{\overrightarrow{k}\})) = \tilde{P}(T = \overrightarrow{k}) = \frac{|\{\tilde{\omega} : T(\tilde{\omega}) = \overrightarrow{k}\}|}{\binom{N}{n}} \quad \forall \overrightarrow{k} \in \hat{\Omega}$$

Falta ahora estudiar cuantos elementos (cardinal) hay en el conjunto $\{\tilde{\omega} : T(\tilde{\omega}) = \overrightarrow{k}\}$.

Lo primero que haremos será definir el problema de combinatoria en pequeños problemas más manejables, de manera que lo que intentamos es contar cuántos conjuntos $\tilde{\omega}$ de bolas extraídas cumplen que para cada especie $a \in E$, el número de bolas de tipo a en el conjunto es k_a (obtenido de $|\tilde{\omega} \cap C_a| = k_a$), así uniendo todos los tipos de bolas tenemos esta relación biunivocua, es decir, existe una biyección entre ambos conjuntos lo cual no demostraremos

$$\{\tilde{\omega} : T(\tilde{\omega}) = \overrightarrow{k}\} \equiv \prod_{a\in E} \{\tilde{\omega_a} \subseteq C_a : |\tilde{\omega_a}| = k_a\}$$

Ahora simplemente debemos calcular el cardinal de cada conjunto del producto cartesiano, y luego multiplicarlos (porque son elecciones sucesivas). Para el color $a \in E$, el número de maneras de elegir k_a bolas de tipo a entre las N_a disponibles es $\binom{N_a}{k_a}$, esto equivale a pensar que tenemos en el vector k_a huecos posibles para hacer combinaciones del total de bolas del tipo a (N_a). Por tanto, el cardinal del conjunto buscado es

$$|\{\tilde{\omega} : T(\tilde{\omega}) = \overrightarrow{k}\}| = \prod_{a\in E} \binom{N_a}{k_a}$$

donde si tenemos $k_a > N_a$ para algún $a \in E$, entonces $\binom{N_a}{k_a} = 0$. Finalmente tenemos que la medida de probabilidad buscada es

$$\hat{P}(\{\overrightarrow{k}\}) = \frac{\prod_{a\in E}\binom{N_a}{k_a}}{\binom{N}{n}} \quad \forall \overrightarrow{k} \in \hat{\Omega}$$

Definición 2.3.6 (Distribuzione ipergeometrica)**.** Sea E un conjunto finito con $|E| \geqslant 2$, $\overrightarrow{N_a} = (N_a)_{a\in E}$ con $N_a \in \mathbb{N}_0$ y $N = \sum_{a\in E} N_a$, y $n \leqslant N$. Una medida de probabilidad $\hat{P} = \mathcal{H}_{n,\overrightarrow{N}}$ en $(\hat{\Omega}, 2^{\hat{\Omega}})$ con función densidad (o función masa de probabilidad)

$$\mathcal{H}_{n,\overrightarrow{N}}(\{\overrightarrow{k}\}) = \frac{\prod_{a\in E}\binom{N_a}{k_a}}{\binom{N}{n}} \quad \forall \overrightarrow{k} \in \hat{\Omega}$$

se dice que sigue una **distribución hipergeométrica** con parámetros n y $\overrightarrow{N}$.

Ejemplo. Sea $E = \{0, 1\}$ con N_1 bolas de tipo 1 y N_0 bolas de tipo 0, y $N = N_0 + N_1$. En este caso, el espacio muestral es

$$\hat{\Omega} = \{(k_0, k_1) : k_0, k_1 \geqslant 0, k_0 + k_1 = n\}$$

y la distribución hipergeométrica es

$$\mathcal{H}_{n,(N_0,N_1)}(k_0, k_1) = \frac{\binom{N_0}{k_0}\binom{N_1}{k_1}}{\binom{N}{n}} \quad \forall (k_0, k_1) \in \hat{\Omega}$$

donde es común expresar la distribución en función de un solo parámetro, por ejemplo k_1, teniendo en cuenta que $k_0 = n - k_1$, obteniendo

$$\mathcal{H}_{n,(N_0,N_1)}(k_1) = \frac{\binom{N_0}{n-k_1}\binom{N_1}{k_1}}{\binom{N}{n}} \quad \forall k_1 = 0, 1, \ldots, \text{mín}\{n, N_1\}$$

esto sería la distribución hipergeométrica clásica de parámetros N, N_1, n.

2.4. Distribuzione di Poisson

Ejemplo. Vamos a trabjar sobre reclamaciones de seguro, un clásico ejemplo de distribución de Poisson. ¿Cuantás reclamaciones recibe una compañía de seguros en un intervalo de tiempo fijo $[0, t[]$, con $t > 0$? Obviamente, el número de reclamaciones es aleatorio, y el espacio muestral es $\Omega = \mathbb{N}_0$. Pero, ¿qué P describe la situación?

Es claro que si tomo un n grande y divido el intervalo $[0, t[$ en n subintervalos de longitud t/n, puedo asumir que en cada subintervalo la compañía recibe a lo sumo una reclamación. En principio la probabilidad de recibir una reclamación en un subintervalo depende de su longitud t/n, por lo que podemos definir un factor de probabilidad $\alpha > 0$ tal que la probabilidad de recibir una reclamación en un subintervalo es $p_n = \alpha\frac{t}{n}$.

Una buena primera aproximación sería la distribución de Bernoulli con parámetro $\alpha\frac{t}{n}$, donde hemos supuesto (como puede ser lógico) que cada subintervalo es independiente de los demás, es decir, que haya llamada en un intervalo no afecta a la probabilidad de que haya llamada en otro intervalo.

Ejemplo.

Observación. Distribución binomial de parámetros θ, n es $\mathcal{B}(\theta, n) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$

Tenemos por tanto que la probabilidad de recibir k reclamaciones en el intervalo $[0, t[$ es $P(k) = \lim_{n\to\infty} \mathcal{B}_{\alpha\frac{t}{n},n}(k)$, donde sabemos que existe dicho límite gracias al siguiente teorema.

Teorema 2.4.3 (Approssimazione di Poisson della distribuzione binomiale)**.** *Sea p_n una sucesión de números tal que $np_n \to \lambda \geqslant 0$, entonces*

$$\lim_{n\to\infty} B_{p_n,n}(k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad \forall k \in \mathbb{N}_0$$

Demostración. □

Teorema 2.4.4. *Para todo $n \in \mathbb{N}_0$ y $p \in (0,1)$, entonces*

$$||\mathcal{B}_{p,n} - \mathcal{P}_{np}|| := \sum_{k\geqslant 0} |\mathcal{B}_{p,n}(k) - \mathcal{P}_{np}(k)| \leqslant 2np^2$$

Definición 2.4.7 (Distribuzione di Poisson)**.** Sea $\Omega = \mathbb{N}_0$ con $\sigma-$algebra 2^Ω, entonces la medida de probabilidad $\mathcal{P}_\lambda$ en $(\Omega, 2^\Omega)$ dada por

$$\mathcal{P}_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda} \quad \forall k \in \mathbb{N}_0$$

se llama **distribución de Poisson** de parámetro $\lambda > 0$.

Ejemplo.

2.4.1. Processo di Poisson

Sea $\alpha > 0$, definiremos una secuencia de variables aleatorias i.i.d. (independientes e idénticamente distribuidas) $(L_i)_{i\geqslant 1}$ que siguen una distribución exponencial con parámetro α, es decir,

$$L_i : (\Omega, \mathcal{F}, P) \to \mathbb{R} \quad L_i \sim \mathcal{E}(\alpha) \quad \forall i \geqslant 1$$

de manera que su función de densidad viene dada por

$$f_{L_i}(x) = \begin{cases} \alpha e^{-\alpha x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

por lo que tenemos claramente, $P(L_i \leqslant t) = 1 - e^{-\alpha t} \quad \forall t \geqslant 0$, e interpretamos las variables aleatorias L_i como los tiempos entre la llegada $i-1$-ésima y la llegada i-ésima en un proceso de llegadas.

Definimos también ahora las siguientes variables aleatorias (más adelante estudiaremos en detalle las v.a. N_t)

$$T_k = \sum_{i=1}^{k} L_i : \Omega \longrightarrow \mathbb{R} \quad \forall k \geqslant 1 \quad \text{y} \quad N_t = \sum_{k \geqslant 1} \mathbb{1}_{(0,t]}(T_k) \quad \forall t > 0$$

donde T_k representa el tiempo hasta la k-ésima llegada (suma de los tiempos entre llegadas) y N_t representa el número de llegadas en el intervalo $]0, t]$.

Domanda. ¿Por qué eliminamos la llegada en $t = 0$?
Tenemos que

$$P(L_1 > 0) = P(\bigcup_{t>0} \{L_1 > t\}) = \lim_{t \to 0^+} P(L_1 > t) = \lim_{t \to 0^+} e^{-\alpha t} = 1$$

por lo que, $P(L_i = 0) = 0$, por lo que la probabilidad de que haya una llegada en $t = 0$ es nula. De hecho, sabemos que $P(L_i \leqslant 0) = 0$, por lo que como es imposible que ocurra algún suceso en $t \leqslant 0$, no tiene sentido contar la llegada en $t = 0$ (obtendríamos el mismo resultado contando o no la llegada en $t = 0$).

Teorema 2.4.5 (Construction of the Poisson process)**.** *Bajo las condiciones descritas anteriormente, tenemos que para cada $t > 0$*

- *N_t es una variable aleatoria que sigue una distribución de Poisson con parámetro αt, es decir, $N_t \sim \mathcal{P}_{\alpha t}$*
- *$N_s, N_t - N_s$ son independientes para todo $s \in]0, t[$.*

Demostración. Tenemos que demostrar que dado $t > 0$ la función $N_t : (\Omega, \mathcal{F}) \to (\mathbb{N}, \mathcal{P}(\mathbb{N}))$ es una variable aleatoria ya que necesitamos estar seguros que los elementos como $\{N_t = k\}$ están en $\mathcal{F}$, puesto que sino no podríamos calcular su probabilidad ($P(N_t = k)$) y por tanto no estaríamos trabajando con una v.a., por lo que veámos esta simple demostración □

Teorema 2.4.6 (Versione generale)**.** *Bajo las condiciones descritas anteriormente, para cualquier partición $0 < t_1 < t_2 < \ldots < t_n < +\infty$, las variables aleatorias $N_{t_i} - N_{t_{i-1}}$ $\forall 1 \leqslant i \leqslant n$ son independientes y siguen una distribución de Poisson con parámetro $\alpha(t_i - t_{i-1})$.*

Definición 2.4.8 (Poisson process)**.** Una familia de variables aleatorias $\{N_t\}_{t \geqslant 0}$ que cumple las propiedades del Teorema 2.4.5 se llama **proceso de Poisson** con intensidad $\alpha > 0$.

2.5. Tempi di attesa

Vamos a trabajar con un experimento bernouilliano con longitud infinita y parámetro $p \in]0, 1[$, es decir, un experimento que tiene dos posibles resultados:

éxito (con probabilidad p) y fracaso (con probabilidad $1-p$). Tenemos que el espacio muestral es $\Omega = \{0,1\}^{\mathbb{N}}$, donde 0 es fracaso y 1 éxito, y definiremos T_r = tiempo de espera hasta el r-ésimo$\in \mathbb{N}$ éxito, matemáticamente dado por

$$T_r : \{0,1\}^{\mathbb{N}} \to \mathbb{N} \quad T_r(\omega) = \text{mín}\{k \in \mathbb{N} : \sum_{i=1}^{k} \omega_i = r\} \quad \forall \omega = (\omega_1, \ldots, \omega_n, \ldots) \in \Omega$$

donde cada $\omega_i \in \{0,1\}$ representa el resultado del i-ésimo experimento bernoulliano.

Domanda. ¿Cuál es la distribución (legge) de T_r?

Definiremos la medida de probabilidad $\mu_{T_r} = P \circ T_r^{-1}$ en $(\mathbb{N}, 2^{\mathbb{N}})$, que es la distribución de T_r y para calcularla debemos estudiar $P(T_r = n)$ para cada $n \in \mathbb{N}$. Es claro que $T_r = n$ si en los primeros $n-1$ experimentos bernoullianos hay exactamente $r-1$ éxitos (y por tanto $n-r=k$ fracasos), y el n-ésimo experimento es un éxito. Por tanto, tenemos que el número de experimentos es $r+k$, y para calcular la probabilidad de $T_r = n$ usaremos la distribución binomial con $r+k-1$ experientos, pues el n-ésimo experimento es un éxito (probabilidad p), veámoslo

$$P(T_r = n) = \mathcal{B}_{r+k-1,p}(r-1) \cdot p = \binom{r+k-1}{r-1} p^{r-1}(1-p)^k \cdot p \overset{(*)}{=} \binom{-r}{k} p^r (p-1)^k \quad \forall n \geqslant r$$

donde en $(*)$ hemos usado que $(1-p)^k = (-1)^k(p-1)^k$ y la definición de coeficiente binomial para $-r$ (número negativo), que viene dado como

$$\binom{-r}{k} = \frac{(-r)\cdots(-r-k+1)}{k!} = (-1)^k \frac{r \cdots (r+k-1)}{k!} = (-1)^k \binom{r+k-1}{r-1} \quad \forall k, r \in \mathbb{N}_0$$

Observación. Recordemos que estamos trabajando con el espacio de probabilidad $(\Omega, \mathcal{F}, P)$ del experimento bernoulliano infinito, donde P mide probabilidades de secuencias de resultados donde $P(1) = p$ es el éxito y $P(0) = 1-p$ es el fracaso.

Es importante tener claro que cuando hablamos de una secuencia de variables aleatorias de Bernoulli, nos referimos a un experimento con infinitos ensayos independientes, cada uno con dos posibles resultados (éxito o fracaso) y una probabilidad fija de éxito $p \in]0,1[$, es decir, variables aleatorias $(X_n)_{n\geqslant 0}$ donde $X_i = \mathbb{1}_{A_i}$ son independientes y $P(A_i) = p$ para cada $i \geqslant 0$.

Es posible que las variables aleatorias de Bernoulli formen parte de un experimento más grande, de manera que una v.a. concreta X_k, tenga como dominio los infinitos experimentos, pero al ser independiente de los demás, su distribución es la misma que la de una v.a. de Bernoulli pura, fijándose solo en su resultado.

Un ejemplo claro de esto es el experimento de lanzar una moneda justa infinitas veces, donde cada lanzamiento es independiente de los demás y la probabilidad de cara es $p = 1/2$. Si definimos la variable aleatoria X_n como el resultado del n-ésimo lanzamiento (1 si es cara, 0 si es cruz), entonces

$$X_n : \Omega \to \{0,1\} \quad X_n(\omega) = \omega_n \quad \forall \omega = (\omega_1, \omega_2, \ldots) \in \Omega$$

donde $\Omega = \{0,1\}^{\mathbb{N}}$ es el espacio muestral del experimento de lanzar la moneda infinitas veces.

Definición 2.5.9 (Distribuzione binomiale negativa o di Pascal)**.** Sea $r > 0$ y $0 < p < 1$, la medida de probabilidad $\mathcal{B}_{r,p}$ en $(\mathbb{N}_0, 2^{\mathbb{N}_0})$ con función de densidad

$$\mathcal{B}_{r,p}(k) = \binom{-r}{k} p^r (p-1)^k \quad \forall k \in \mathbb{N}_0$$

se llama **distribución binomial negativa o de Pascal** con parámetros r y p.

Observación. Como bien hemos demostrado anteriormente, esta distribución cálcula la probabilidad de que el r-ésimo éxito ocurra en el experimento número $n = r + k$.

2.6. Distribuzione Normale

Domanda. ¿Existe una distribución uniforme en dimensión infinita?
Tomaremos el siguiente espacio muestral

$$\Omega_N = \{x \in \mathbb{R}^N : ||x||_2 \leqslant vN\} \text{ con } v > 0$$

que es la bola de radio $\sqrt{vN}$ en $\mathbb{R}^N$, y donde se aproximará la dimensión infinita cuando $N \to \infty$. Definimos la medida de probabilidad $P_N = \mathcal{U}_{\Omega_N}$ como la distribución uniforme (continua) en $(\Omega_N, \mathcal{B}(\Omega_N))$, es decir

$$P_N(A) = \frac{\lambda_N(A)}{\lambda_N(\Omega_N)} \quad \forall A \in \mathcal{B}(\Omega_N)$$

donde λ_N es la medida de Lebesgue en $\mathbb{R}^N$. También usaremos la siguiente variable aleatoria

$$X_N : \Omega_N \to \mathbb{R} \quad X_N(x) = x_1 \quad \forall x = (x_1, x_2, \ldots, x_N) \in \Omega_N$$

que selecciona la primera coordenada de cada vector $x \in \Omega_N$. Nuestro objetivo es estudiar la distribución de X_N cuando $N \to \infty$, es decir, estudiar la medida de probabilidad $\mu_{X_N} = P_N \circ X_N^{-1}$ en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Teorema 2.6.7 (Borel-Poincaré)**.** *Sea $v > 0$ y $a < b$ con $a, b \in \mathbb{R}$, entonces*

$$\lim_{N\to\infty} P_N(a \leqslant X_N \leqslant b) = \frac{1}{\sqrt{2\pi v}} \int_a^b e^{-\frac{x^2}{2v}} dx$$

Demostración. □

2.7. Probabilità condizionata

Supongamos que quiero estudiar el evento B y alguien me dice que el evento A ha ocurrido, ¿cómo cambia la probabilidad de B? Para ello dada la medida de probabilidad P (trabajamos en el espacio de probabilidad $(\Omega, \mathcal{F}, P)$) calcularemos una nueva medida de probabilidad P_A tomando en cuenta la información que tenemos, es decir, que A ha ocurrido, por lo que nos preguntamos, ¿que propiedades debe tener P_A?

1) $P_A(A) = 1$.

2) $\exists c_A$ tal che $\forall B \subseteq A$ con $B \in \mathcal{F}$, entonces $P_A(B) = c_A P(B)$.

Proposición 2.7.8. *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $A \in \mathcal{F}$ con $P(A) > 0$, entonces existe una única medida de probabilidad P_A en $(\Omega, \mathcal{F})$ que cumple las propiedades 1) y 2) anteriores, dada por*

$$P_A(B) = \frac{P(A \cap B)}{P(A)} \quad \forall B \in \mathcal{F}$$

Demostración. □

Observación. Recordemos dos propiedades importantes de la probabilidad condicionada P_A

- $P(A) > 0$
- $P(B \mid A) = \dfrac{P(A \cap B)}{P(A)} \quad \forall A, B \in \mathcal{F}$ con $P(A) > 0$

Definición 2.7.10. Sea $A \in \mathcal{F}$ con $P(A) > 0$, la **probabilidad condicionada a** A en $(\Omega, \mathcal{F})$ viene dada por

$$P_A(B) = P(B \mid A) = \frac{P(A \cap B)}{P(A)} \quad \forall B \in \mathcal{F}$$

Una interpretación intuitiva de la probabilidad condicionada $P(B \mid A)$ es que es la fracción de casos favorables a B dentro de los casos favorables a A, entre la probabilidad total de que ocurra A.

Teorema 2.7.9. *Sea $A \in \mathcal{F}$ con $P(A) > 0$, entonces*

1) $P(\cdot \mid A)$ es una medida de probabilidad en $(\Omega, \mathcal{F})$.

2) Sia $(B_i)_{i \in \mathbb{N}}$ una collezione di eventi disgiunti in $\mathcal{F}$ e $\bigcup_{i \in \mathbb{N}} B_i = \Omega$, allora

$$P(E) = \sum_{i \in \mathbb{N}} P(B_i) P(E \mid B_i) \quad \forall E \in \mathcal{F}$$

*3) **(Bayes)** Para todo $k \in \mathbb{N}$ fijado, se tiene*

$$P(B_k \mid A) = \frac{P(B_k) P(A \mid B_k)}{\sum_{i \in \mathbb{N}} P(B_i) P(A \mid B_i)}$$

donde $(B_i)_{i \in \mathbb{N}}$ sono come 2).

4) Sea una colección $(A_i)_{i=1,\dots,n} \subseteq \mathcal{F}$ *de eventos tales que* $P(A_1 \cap A_2 \cap \ldots \cap A_n) > 0$, *entonces*

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1) \cdot \ldots \cdot P(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

Demostración. □

Ejemplo. *Paradoja de Monty Hall.* Supongamos que estamos en un concurso televisivo donde hay tres puertas cerradas: detrás de una de ellas hay un coche (premio) y detrás de las otras dos hay cabras (sin premio). El concursante debe elegir una puerta, por ejemplo la puerta 1. Después, el presentador (que sabe dónde está el coche) abre una de las otras dos puertas, por ejemplo la puerta 3, que tiene una cabra detrás. Ahora el concursante tiene la opción de quedarse con su elección inicial (puerta 1) o cambiar a la otra puerta cerrada (puerta 2). ¿Qué debería hacer el concursante para maximizar sus probabilidades de ganar el coche? Definamos ahora el problema matemáticamente.
Sea $\Omega = \{1, 2, 3\}$ el conjunto de puertas, y notamos la 1 como la puerta ganadora. Definimos las siguientes variables aleatorias

$$C : \Omega \to \{1, 2, 3\} \quad \text{y} \quad H : \Omega \to \{1, 2, 3\}$$

donde C y H indican la elección del concursante y del presentador respectivamente. Sabemos que el evento $A = \{C \neq H\} \cap \{H \neq 1\}$ ha ocurrido, es decir, el concursante ha elegido una puerta distinta a la que ha abierto el presentador, y el presentador no ha abierto la puerta ganadora. Nuestro objetivo es calcular la probabilidad condicionada $P(C = 1 \mid A)$, es decir, la probabilidad de que el concursante haya elegido la puerta ganadora dado que el evento A ha ocurrido. Tenemos entonces

$$\left.\begin{array}{l} P(H = 1) = 0 \implies P(H \neq 1) = 1 \\ (\text{Asunción del presentador}) P(H \neq C) = 1 \end{array}\right\} \implies P(A) = 1 \implies$$
$$\implies P(C = 1 \mid A) = \frac{P(\{C = 1\} \cap A)}{P(A)} = P(\{C = 1\} \cap A) = P(\{C = 1\}) = \frac{1}{3}$$

donde hemos obtenido que la probabilidad de que el concursante haya elegido la puerta ganadora desde el principio es $1/3$. Para ver bien el evento A, veámoslo en detalle:

$$\{C \neq H\} = \{\omega \in \Omega \mid C(\omega) \neq H(\omega)\} \implies \{C = 1\} \subseteq \{C \neq H\}$$
$$\{H \neq 1\} = \{\omega \in \Omega \mid H(\omega) \neq 1\} \implies \{C = 1\} \subseteq \{H \neq 1\}$$

es decir, es claro que $\{C = 1\} = \{\omega \in \Omega \mid C(\omega) = 1\} \subseteq \{C \neq H\} \cap \{H \neq 1\}$. Por tanto, la probabilidad haya elegido una puerta sin premio es

$$P(C \neq 1 \mid A) = 1 - P(C = 1 \mid A) = \frac{2}{3}$$

por lo que, el concursante debería cambiar su elección inicial para maximizar sus probabilidades de ganar el coche.

2.8. Multi-stage models

En esta sección estudiaremos modelos de probabilidad que se desarrollan en múltiples etapas (multi-stage), donde cada etapa puede depender de las anteriores. Queremos contruir un espacio de probabilidad $(\Omega, \mathcal{F}, P)$ para el experimento completo, así como variables aleatorias $(X_i)_{i=1,\dots,n}$ en Ω que describan los resultados de los subexperimentos individuales. Supongamos que tenemos lo siguiente

1) Conocemos la distribución (legge) de X_1.

2) Conocemos la distribución de X_k dado $X_1, \dots, X_{k-1}$ para cada $k \geqslant 2$.

es decir, nosotros sabemos cómo se comporta cada subexperimento individualmente, y como se comporta cada subexperimento condicionado a los anteriores. Véamos ahora el siguiente teorema que nos permitirá construir el espacio de probabilidad completo.

Teorema 2.8.10 (Construcción de medidas de probabilidad mediante probabilidades condicionadas)**.** *Sean $\Omega_1, \dots, \Omega_n$ espacios muestrales discretos no vacíos, y dado*

1. *$p_1 : \Omega_1 \to [0,1]$ función de densidad (densità di probabilità), con $\sum_{\omega_1 \in \Omega_1} p_1(\omega_1) = 1$.*
2. *Para cada $k \geqslant 2$, y para cada $\omega_i \in \Omega_i$ con $i < k$, tenemos*

$$p_{k|\omega_1,\dots,\omega_{k-1}} : \Omega_k \to [0,1] \quad \text{función de densidad}$$

3. *Definimos el espacio muestral completo como $\Omega = \prod_{i=1}^n \Omega_i$.*

Entonces, $\exists!$ P medida de probabilidad en $(\Omega, 2^\Omega)$ tal que

a) $P(X_1 = \omega_1) = p(\omega_1)$ dove $X_1 : \Omega \to \Omega_1$ es la v.a. proyección en la primera coordenada.

b) Para cada $k = 2, \dots, n$ se tiene

$$P(X_k = \omega_k \mid X_1 = \omega_1, \dots, X_{k-1} = \omega_{k-1}) = p_{k|\omega_1,\dots,\omega_{k-1}}(\omega_k)$$

donde $X_k : \Omega \to \Omega_k$ es la v.a. proyección en la k-ésima coordenada.

Por tanto, viene dada por

$$P(\{\omega\}) = p_1(\omega_1) \cdot p_{2|\omega_1}(\omega_2) \cdots p_{n|\omega_1,\dots,\omega_{n-1}}(\omega_n) \quad \forall \omega = (\omega_1, \dots, \omega_n) \in \Omega$$

Demostración. □

Ejemplo (El juego de skat)**.** El skat es un juego de cartas alemán para tres jugadores, donde se utiliza una baraja de 32 cartas (donde hay 4 ases). Al inicio del juego, se reparten 10 cartas a cada jugador y 2 cartas se colocan boca abajo en el centro de la mesa (skat). Queremos modelar el reparto de cartas para conocer el número de ases que recibe cada jugador. Definimos primero los espacios muestrales

individuales $\Omega_i = \{0, 1, \ldots, 4\}$ para cada jugador $i = 1, 2, 3$, donde cada elemento representa el número de ases que recibe el jugador. Como hablamos de una baraja en la que distinguimos entre cartas normales (28) y ases (4), tenemos las siguientes distribuciones

$$\begin{aligned} P_1 &= \mathcal{H}_{10;4,28} && \text{distribución de j1} \\ P_{2|\omega_1} &= \mathcal{H}_{10;4-\omega_1,28-(10-\omega_1)} && \text{distribución de j2 dado que j1 tiene } \omega_1 \text{ ases} \\ P_{3|\omega_1,\omega_2} &= \mathcal{H}_{10;4-\omega_1-\omega_2,28-(20-\omega_1-\omega_2)} && \text{distribución de j3 dado que j1 tiene } \omega_1 \text{ ases y j2 tiene } \omega_2 \text{ ases} \end{aligned}$$

donde hemos usado la distribución hipergeométrica para modelar cada reparto de cartas.
Ahora, usando el teorema anterior, podemos construir el espacio de probabilidad completo $(\Omega, 2^\Omega, P)$ donde $\Omega = \Omega_1 \times \Omega_2 \times \Omega_3$, y por tanto para calcular la probabilidad de que cada jugador reciba un as, es decir, $P(\{(1,1,1)\})$, tenemos

$$P(\{(1,1,1)\}) = p_1(\{1\}) \cdot p_{2|1}(\{1\}) \cdot p_{3|1,1}(\{1\}) = \mathcal{H}_{10;4,28}(\{1\}) \cdot \mathcal{H}_{10;3,18}(\{1\}) \cdot \mathcal{H}_{10;2,8}(\{1\}) =$$
$$= \frac{\binom{4}{1}\binom{28}{9}}{\binom{32}{10}} \cdot \frac{\binom{3}{1}\binom{19}{9}}{\binom{22}{10}} \cdot \frac{\binom{2}{1}\binom{10}{9}}{\binom{12}{10}} \approx 0{,}0556$$

Definición 2.8.11 ($\sigma-$algebra prodotto)**.** Dati $(\Omega_i, \mathcal{F}_i)$ spazi misurabili per $i \in \mathbb{N}_+$, definiamo la $\sigma-$**algebra prodotto** in $\Omega = \prod_{i=1}^{n} \Omega_i$ come

$$\bigotimes_{i=1}^{\infty} \mathcal{F}_i := \sigma \left(\underbrace{\{A_1 \times \ldots \times A_k \times \Omega_{k+1} \times \ldots : k \geqslant 1, A_i \in \mathcal{F}_i \text{ para } i \leqslant k\}}_{\mathcal{G}} \right)$$

Observación. En el teorema anterior, hemos usado σ para construir la $\sigma-$álgebra más pequeña que contiene el conjunto definido. Otro detalle importante es que hemos definido una serie de eventeos $A_1, \ldots, A_k$ en los primeros k espacios muestrales, y luego hemos tomado los espacios muestrales siguientes, de manera que hemos indicado que solo importan los primeros k experimentos, y los demás no afectan al resultado.

Lema 2.8.11 (Lema di Dynkin)**.** *Sea una familia de conjuntos $\mathcal{G}$ tal que*

- $\Omega \in \mathcal{G}$
- $A \cap B \in \mathcal{G} \quad \forall A, B \in \mathcal{G}$

entonces, "en casi todos los casos" si una propiedad vale para los conjuntos en $\mathcal{G}$, entonces vale para todos los conjuntos en $\sigma(\mathcal{G})$.

Teorema 2.8.12 (Prodotto di infiniti numerabili Ω_i's)**.** *Sea $(\Omega_i)_{i\in\mathbb{N}}$ una colección de espacios muestrales discretos no vacíos, y dado*

1. $p_1 : \Omega_1 \to [0,1]$ *función de densidad (densità di probabilità), con* $\sum_{\omega_1 \in \Omega_1} p_1(\omega_1) = 1$.

2. Para cada $k \geqslant 2$, y para cada $\omega_i \in \Omega_i$ con $i < k$, tenemos

$$p_{k|\omega_1,\ldots,\omega_{k-1}} : \Omega_k \to [0,1] \quad \textit{función de densidad}$$

3. Definimos el espacio muestral completo como $\Omega = \prod_{i=1}^{\infty} \Omega_i$.

Entonces, $\exists!$ P medida de probabilidad en $(\Omega, \bigotimes_{i=1}^{\infty} \mathcal{F}_i)$ tal que

$$P(X_1 = \omega_1, \ldots, X_n = \omega_n) = p_1(\omega_1) \cdot p_{2|\omega_1}(\omega_2) \cdots p_{n|\omega_1,\ldots,\omega_{n-1}}(\omega_n) \quad \forall n \in \mathbb{N}, \forall \omega_i \in \Omega_i$$

donde $X_k : \Omega \to \Omega_k$ es la v.a. proyección en la k-ésima coordenada.

Ejemplo (Urna di Pólya)**.** Tenemos una urna de r bolas rojas y ω bolas blancas, donde en cada extracción de una bola, devolvemos la bola extraída y añadimos $c \geqslant 0$ bolas extras del mismo color. Es claro que el caso de $c = 0$ es el caso clásico de extracción con reemplazo, por lo que estamos interesados en estudiar el caso $c \in \mathbb{N}$. Dividiremos el estudio en varias fases, veámoslo.

Fase 1. Definimos el espacio muestral como $\Omega = \{0, 1\}$, donde 1 representa la extracción de una bola roja y 0 la extracción de una bola blanca, y definimos la siguiente distribución de probabilidad

$$p_1(1) = \frac{r}{r+\omega}, \quad p_1(0) = \frac{\omega}{r+\omega}$$

Fase $k \geqslant 2$. En esta fase, la distribución de probabilidad depende del resultado de las fases anteriores, por lo que tenemos

$$p_{k|\omega_1,\ldots,\omega_{k-1}} \text{ con } \omega_i \in \Omega_i$$

y sabiendo que $l := \sum_{i=1}^{k-1} \omega_i$ es el número de bolas rojas (recordemos que 1 es rojo y 0 es blanco) extraídas en las fases anteriores, definimos la distribución de probabilidad como

$$p_{k|\omega_1,\ldots,\omega_{k-1}}(\omega_k) = \begin{cases} \dfrac{\omega + (k-1-l)c}{r+\omega+(k-1)c} & \text{si } \omega_k = 0 \\ \dfrac{r+lc}{r+\omega+(k-1)c} & \text{si } \omega_k = 1 \end{cases}$$

Del Teorema 2.8.12, sabemos que existe una única medida de probabilidad P en $(\Omega, \bigotimes_{i=1}^{\infty} 2^{\Omega_i})$ con $\Omega = \{0,1\}^{\mathbb{N}}$ tal que

$$P(X_1 = \omega_1, \ldots, X_n = \omega_n) = \frac{\prod_{i=0}^{l-1}(r+ic) \cdot \prod_{j=0}^{n-l-1}(\omega+jc)}{\prod_{m=0}^{n-1}(r+\omega+mc)} \quad \forall n \in \mathbb{N}_+, \forall \omega_i \in \Omega_i$$

donde $l = \sum_{i=1}^{n} \omega_i$ es el número de bolas rojas extraídas en las primeras n fases. En este caso, el valor l se ha hecho sumando hasta n, pero esto se arregla haciendo el producto en el segundo productorio hasta $n - l - 1$, para evitar añadir a la urna el resultado obtenido en el caso n, ya que aqui acaba nuestro experimento.
Es importante destacar que en el numerador no se toma en cuenta el orden de extracción de las bolas, sino solo el número total de bolas rojas y blancas extraídas, puesto que gracias a la propiedad conmutativa de la multiplicación podemos reordenar los factores sin problema. Pensémoslo usando un simple ejemplo, suponiendo que primero sacamos una bola roja, luego una blanca y después dos rojas más, de modo que teniendo en cuenta el orden tendríamos

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) = \frac{r}{r+\omega} \cdot \frac{\omega}{r+\omega+c} \cdot \frac{r+c}{r+\omega+2c} \cdot \frac{r+2c}{r+\omega+3c}$$

pero si reordenamos los factores tenemos que

$$P(X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1) = \frac{r}{r+\omega} \cdot \frac{r+c}{r+\omega+c} \cdot \frac{r+2c}{r+\omega+2c} \cdot \frac{\omega}{r+\omega+3c}$$

y en ambos casos la probabilidad es la misma, pues al final el numerador es el mismo en ambos casos. Deducimos de este modo que la probabilidad de obtener la secuencia $(1, 0, 1, 1)$ es igual a la probabilidad de obtener cualquier otra secuencia con tres bolas rojas y una blanca, como por ejemplo $(1, 1, 0, 1)$ o $(0, 1, 1, 1)$.

Podemos calcular otro tipo de probabilidades, como por ejemplo la probabilidad de obtener l bolas rojas en n extracciones, sin importar el orden. Para ello, definimos la variable aleatoria

$$R_n = \sum_{i=1}^{n} X_i \quad \text{tal que} \quad R_n : \Omega \to \{0, \ldots, n\}$$

que cuenta el número de bolas rojas extraídas en las primeras n extracciones. Si nos damos cuenta ahora, que para todo $\omega \in \Omega$ con $R_n(\omega) = l$ la probabilidad es exactamente la misma (ya que como dijimos antes, no importa el orden de extracción, sino solo el número total de bolas rojas y blancas extraídas), podemos calcular la probabilidad de obtener l bolas rojas en n extracciones como

$$P(R_n = l) = \binom{n}{l} \cdot \frac{\prod_{i=0}^{l-1}(r+ic) \cdot \prod_{j=0}^{n-l-1}(\omega+jc)}{\prod_{m=0}^{n-1}(r+\omega+mc)} \quad \forall n \in \mathbb{N}_+, \forall l = 0, \ldots, n$$

donde para deducir el factor combinatorio, lo que hemos pensado es que enumerando las secuencias como $\{1, 2, \ldots, n\}$ queremos seleccionar un subconjunto de tamaño l sin que importe el orden en que eliges esos índices. Lo que realmente se ha hecho es contar las posibles formas de elegir las posiciones de las bolas rojas en la secuencia de extracciones, y luego multiplicar eso por la probabilidad de una de ellas (que tienen el mismo valor).

Veámos a continuación un resumen de los resultados para distintos valores de c, así como una expresión una alternativa abreviada para la probabilidad $P(R_n = l)$.

- **Para $c = 0$:** tenemos la distribución binomial con parámetros n y $p = \dfrac{r}{r+\omega}$, es decir, $\mathcal{B}_{n,\frac{r}{r+\omega}}$, pues se trata del caso de extracción con reemplazo.
- **Para $c > 0$:** tenemos el siguiente resultado

$$P(R_n = l) = \frac{\binom{-a}{l}\binom{-b}{n-l}}{\binom{-a-b}{n}} \quad \text{dove } a = \frac{r}{c},\ b = \frac{\omega}{c}$$

ya que teniendo presente que $\binom{-r}{k} = \frac{(-r)(-r-1)\ldots(-r-k+1)}{k!}$, desarrollamos de la siguiente manera

$$\prod_{i=0}^{l-1}(r+ic) = \prod_{i=0}^{l-1}(ac+ic) = c^l \prod_{i=0}^{l-1}(a+i) = c^l \cdot [a(a+1)\ldots(a+l-1)] =$$
$$= c^l \cdot (-1)^l \cdot [(-a)(-a-1)\ldots(-a-l+1)] = c^l \cdot (-1)^l \cdot l! \cdot \binom{-a}{l}$$

Si hacemos lo mismo para los otros dos productos, obtenemos

$$\prod_{j=0}^{n-l-1}(\omega+jc) = c^{n-l}\cdot(-1)^{n-l}\cdot(n-l)!\cdot\binom{-b}{n-l} \quad \text{y} \quad \prod_{m=0}^{n-1}(r+\omega+mc) = c^n\cdot(-1)^n\cdot n!\cdot\binom{-a-b}{n}$$

y sustituyendo en la expresión de $P(R_n = l)$, obtenemos el resultado deseado.

- **Para $c = -1$:** tenemos la distribución hipergeométrica $\mathcal{H}_{n;r,\omega}$, pues se trata del caso de extracción sin reemplazo.

2.9. Indipendenza

Definición 2.9.12 (Independenza di eventi)**.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, y sean $A, B \in \mathcal{F}$ dos eventos. Decimos que A y B son **independientes** si

$$P(A \cap B) = P(A) \cdot P(B)$$

Ejemplo (Urne con reinserimento e non)**.** Sea $A = \{\text{prima estrazione è rossa}\}$ y $B = \{\text{seconda estrazione è rossa}\}$, donde las extracciones se hacen de una urna con r bolas rojas y ω bolas blancas, de modo que el espacio muestral es $\Omega = \{1, \ldots, r, r+1, \ldots, r+\omega\}^2$. Veamos si los eventos A y B son independientes en los siguientes dos casos

- **Con reinserimento:** tenemos $P(A) = P(B) = \dfrac{r}{r+\omega}$, por haber reinserción. Veámos ahora $P(A \cap B)$

$$A\cap B = \{(i,j) : 1 \leqslant i,j \leqslant r\} \implies |A\cap B| = r^2 \implies P(A\cap B) = \frac{|A\cap B|}{|\Omega|} = \frac{r^2}{(r+\omega)^2} = P(A)\cdot P$$

- **Senza reinserimento:** tenemos $P(A) = \dfrac{r}{r+\omega}$ y el espacio muestral es

$$\Omega_{\neq} = \{(k,l) : 1 \leqslant k,l \leqslant r+\omega, k \neq l\} \subseteq \Omega$$

de modo que igual que antes tenemos $P(A) = P(B) = \dfrac{r}{r+\omega}$. Veámos ahora $P(A \cap B)$

$$P(A \cap B) = P(B \mid A) \cdot P(A) = \frac{r-1}{r+\omega-1} \cdot \frac{r}{r+\omega} < P(A) \cdot P(B)$$

por lo que en este caso los eventos A y B no son independientes.

En este ejemplo hemos visto que dos mismos eventos pueden ser dependientes respecto a una medida de probabilidad, y ser independientes respecto a otra medida de probabilidad distinta, es decir, la independencia de eventos en $\mathcal{F}$ depende de la medida de probabilidad

Definición 2.9.13 (Indipendenza della famiglia dagli eventi)**.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, y sea $(A_i)_{i \in I}$ una familia de eventos en $\mathcal{F}$, donde I es un conjunto. Decimos que la **familia** $(A_i)_{i \in I}$ **es independiente** si $\forall J \subseteq I$ finito se tiene que

$$P\left(\bigcap_{j \in J} A_j\right) = \prod_{j \in J} P(A_j)$$

En cuyo caso, tenemos que

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j) \quad \forall i \neq j$$

Ejemplo (Causalidad no implica dependencia)**.** Sea el experimento de lanzar dos dados uno detrás de otro, definimos el espacio muestral como $\Omega = \{1, \ldots, 6\}^2$ siguiendo la distribución uniforme $P = U_\Omega$, y las siguientes variables aleatorias

$$A = \{\text{suma de los valores es } 7\} = \{(k,l) \in \Omega : k + l = 7\}$$
$$B = \{\text{valor del primero dado es } 6\} = \{(k,l) \in \Omega : k = 6\}$$

entonces $|A| = |B| = 6$ y $|A \cap B| = 1$, por lo que

$$P(A \cap B) = \frac{1}{36} = P(A) \cdot P(B)$$

por tanto son independientes, aún sabiendo que la suma (evento B) depende del valor del primer dado (evento A), lo que viene a llamarse causalidad. Esto ocurre por haber elegido 7 como suma, ya que si hubiésemos elegido 12, tendríamos $P(B) = \dfrac{1}{36}$ y serían dependientes.

La clave en la independencia de eventos es la proporción de solapamiento de un evento sobre el otro, ya que si esta proporción (dependencia) es igual a la proporción respecto al espacio muestral total, entonces los eventos son independientes, ya que en este ejemplo lo que hemos visto es que la proporción de B sobre A es $1/6$ (de los 6 posibles valores del segundo dado necesitamos que salga el 1), mientras que la proporción de B respecto al espacio muestral total es también $1/6$ (de los 36 posibles valores necesitamos que salgan los 6 de ellos que suman 7). Por tanto, la proporción es la misma y los eventos son independientes.

Observación. La causalidad no implica dependencia, ya que un evento puede depender de otro, pero si la proporción de solapamiento es la misma que la proporción respecto al espacio muestral total (no consigo ayuda para predecir el evento B tras saber A), entonces los eventos son independientes.

Ejemplo (3 eventi non indip., ma a due a due independenti)**.**

Ejemplo (Libro 3.15)**.**

Definición 2.9.14 (Variabile aleatoria indipendenti)**.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad, y sean $(X_i)_{i\in I}$ una familia de variables aleatorias en Ω, donde I es un conjunto. Decimos que la **familia** $(X_i)_{i\in I}$ **es independiente** si $\forall J \subseteq I$ finito se tiene que

$$P\left(\bigcap_{j\in J}\{X_j \in A_j\}\right) = \prod_{j\in J} P(X_j \in A_j) \quad \forall A_j \in \mathcal{B}(\mathbb{R})$$

Es importante destacar que la independencia de variables aleatorias depende de la medida de probabilidad P, al igual que ocurría con la independencia de eventos. La notación $\{X_j \in A_j\}$ representa el siguiente conjunto de resultados del experimento $\{\omega \in \Omega : X_j(\omega) \in A_j\}$, es decir, el conjunto de resultados del experimento que hacen que la variable aleatoria X_j tome valores en el conjunto A_j.

2.10. Esperanza

Definición 2.10.15 (Valore atteso con v.a. discreta)**.** Sea X una variable aleatoria discreta, entonces tenemos

$$\mathbb{E}[|X|] = \sum_{z\in X(\Omega)} |z| \cdot P(X = z) \ \in]0, \infty[\text{ (caso discreto)}$$
$$\mathbb{E}[|X|] = \int_{\mathbb{R}} |z| \cdot f_X(z)\, dz \ \in]0, \infty[\text{ (caso continuo)}$$

donde si $\mathbb{E}[X] < \infty$, decimos que X es **integrable**, notado como $X \in \mathcal{L}^1 = \mathcal{L}^1(P)$ y definimos el **valor esperado** de X como

$$\mathbb{E}[X] = \sum_{z\in X(\Omega)} z \cdot P(X = z) \quad \text{(caso discreto)}$$
$$\mathbb{E}[X] = \int_{\mathbb{R}} z \cdot f_X(z)\, dz \quad \text{(caso continuo)}$$

Observación. Il valore atteso dipende solo della distribuzione di X.

Ejemplo (Función indicadora)**.** Para cada $A \in \mathcal{F}$, la función indicadora 1_A pertenece a $\mathcal{L}^1(P)$, y

$$\mathbb{E}[1_A] = 0 \cdot P(1_A = 0) + 1 \cdot P(1_A = 1) = P(A)$$

Esta relación conecta las nociones de esperanza y probabilidad.

Ejemplo.

Teorema 2.10.13 (Propiedades de la esperanza)**.** *Sean $X, Y, X_n, Y_n : \Omega \to \mathbb{R}$ variables aleatorias discretas en $\mathcal{L}^1(P)$. Entonces se cumplen las siguientes propiedades.*

(a) ***Monotonía.*** *Si $X \leqslant Y$, entonces $\mathbb{E}[X] \leqslant \mathbb{E}[Y]$.*

(b) ***Linealidad.*** *Para cada $c \in \mathbb{R}$, se tiene que $cX \in \mathcal{L}^1(P)$ y $\mathbb{E}[cX] = c\mathbb{E}[X]$. Además, $X + Y \in \mathcal{L}^1(P)$ y $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.*

(c) ***$\sigma-$aditividad y convergencia monótona.*** *Si cada $X_n \geqslant 0$ y $X = \sum_{n=1}^{\infty} X_n$, entonces*

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{E}[X_n]$$

Por otro lado, si $Y_n \nearrow Y$ para $n \to \infty$, se sigue que $\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[Y_n]$.

(d) ***Regla del producto en el caso de independencia.*** *Si X e Y son independientes, entonces $XY \in \mathcal{L}^1(P)$ y $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.*

Proposición 2.10.14 (Variables aleatorias con función de densidad)**.** *Sea X una variable aleatoria con valores en $\mathbb{R}^d$ y función de densidad f_X, es decir, suponemos que $P \circ X^{-1}$ tiene densidad de Lebesgue en $\mathbb{R}^d$. Para cualquier otra variable aleatoria $g : \mathbb{R}^d \to \mathbb{R}$, se sigue que $g \circ X \in \mathcal{L}^1$ si y solo si $\int_{\mathbb{R}^d} |g(x)| f_X(x)\, dx < \infty$, y en ese caso*

$$\mathbb{E}[g \circ X] = \int_{\mathbb{R}^d} g(x) f_X(x)\, dx$$

2.11. Varianza

Definición 2.11.16 (Varianza e deviazione standard)**.** Sea X una variable aleatoria tal que $\mathbb{E}[|X|^2] < \infty$, es decir, $X \in \mathcal{L}^2(P)$. Definimos la **varianza** de X como

$$\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

y la **desviación estándar** de X como $\sigma_X = \sqrt{\mathrm{Var}(X)}$.

Definición 2.11.17 (Covarianza)**.** Sea X, Y dos variables aleatorias tales que $\mathbb{E}[|X|^2], E[|Y|^2] < \infty$, entonces definimos la **covarianza** entre X e Y como

$$Cov(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

De hecho, si $Cov(X, Y) = 0$, decimos que X e Y son **no correladas**.

Teorema 2.11.15 (Proprietà di varianze e covarianze)**.** *Sean $X, Y, X_i \in \mathcal{L}^2$ y $a, b, c, d \in \mathbb{R}$, entonces se cumplen las siguientes propiedades*

(a) *$aX + b, cY + d \in \mathcal{L}^2$ y $Cov(aX + b, cY + d) = ac \cdot Cov(X, Y)$. En particular, $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$.*

(b) *$Cov(X, Y)^2 \leqslant \mathrm{Var}(X)\,\mathrm{Var}(Y)$ (desigualdad de Cauchy-Schwarz).*

(c) *Si X, Y son independientes, entonces X, Y son también no correladas.*

(d) Sea $X_1, \ldots, X_n \in \mathcal{L}^2$ y entonces

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i \neq j} Cov(X_i, X_j)$$

En particular, si $X_1, \ldots, X_n$ son no correladas, entonces $\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}(X_i)$.

Definición 2.11.18 (Mediana)**.** Sea $X : \Omega \to \mathbb{R}$ una variable aleatoria real, un numero $\mu \in \mathbb{R}$ se llama **mediana** de X si

$$P(X \leqslant \mu) \geqslant 1/2 \quad \text{y} \quad P(X \geqslant \mu) \geqslant 1/2$$

Una de las preguntas más habituales sobre la mediana es, ¿existe siempre una mediana?¿en ese caso, es única?¿que significa esta para el valor esperadode X? A continuación, respondemos a estas preguntas, usando como referencia la distribución normal representada en la Figura 2.1.

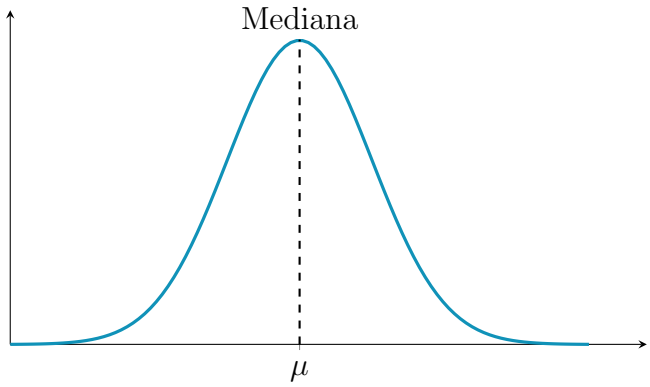

Figura 2.1: Distribución Normal con indicación de la mediana (μ).

¿Unicidad? Si consideramos un ejemplo donde tengamos $a < b$ con $a, b \in \mathbb{R}$ tales que $P(X \leqslant a) = P(X \geqslant b) = 1/2$ y $P(a < X < b) = 0$, entonces cualquier valor $\mu \in [a, b]$ es una mediana de X, por lo que la mediana no es única en este caso.

Para entender mejor esta demostración veremos un caso más concreto, suponiendo que $X = -\mathbb{1}_A + \mathbb{1}_{A^c}$ con $P(A) = P(A^c) = 1/2$. En este caso, tenemos que $P(X \leqslant -1) = P(X \geqslant 1) = 1/2$ y $P(-1 < X < 1) = 0$, por lo que cualquier valor $\mu \in [-1, 1]$ es una mediana de X, y por tanto la mediana no es única.

Podemos ver una representación gráfica de este caso en la Figura 2.2, donde la función de distribución acumulada F_X tiene una meseta en el intervalo $[-1, 1]$, y cualquier valor en ese intervalo cumple las condiciones para ser mediana (hemos incluido el 1, pues también cumple las condiciones para ser mediana).

¿Existencia? Si consideramos la función de distribución acumulada $F_X : \mathbb{R} \to [0, 1]$ definida como $F_X(x) = P(X \leqslant x)$, que es monótona creciente y continua por la derecha, entonces definimos el siguiente valor

$$\mu = \text{ínf}\{x \in \mathbb{R} : F_X(x) = P(X \leqslant x) \geqslant 1/2\}$$

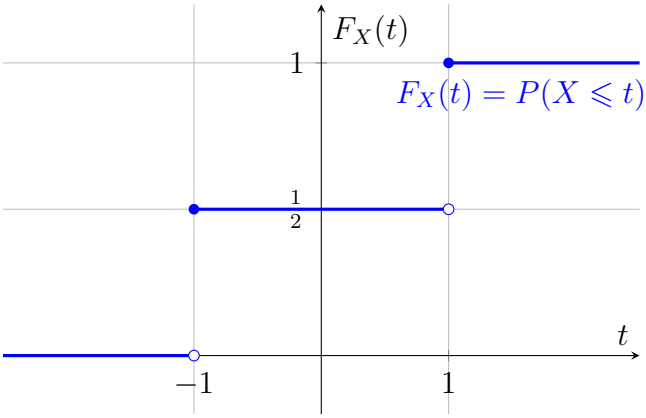

Figura 2.2: Representación de la función de distribución acumulativa escalonada.

que está bien definido, es decir, el conjunto $\{x : F_X(x) \geqslant 1/2\}$ no es vacío y está limitado inferiormente, pues F_X va desde 0 hasta 1. Veamos ahora que este valor μ es una mediana de X

1) Por definición de mínimo (en este caso ínfimo y mínimo son lo mismo, gracias a la continuidad por la derecha de F_X), tenemos que $P(X \leqslant \mu) = F_X(\mu) \geqslant 1/2$, por lo que se cumple la primera condición de la mediana
2) Por otro lado, falta demostrar que $P(X \geqslant \mu) \geqslant 1/2$, para ellos buscaremos demostrar que $P(X < \mu) \leqslant 1/2$, y por tanto tendríamos

$$P(X \geqslant \mu) = 1 - P(X < \mu) \geqslant 1 - \frac{1}{2} = \frac{1}{2}$$

Para demostrar lo querido, usando la definición de ínfimo, sabemos que no hay $v < \mu$ con $F_X(v) \geqslant 1/2$, por lo que

$$\forall v < \mu, \quad P(X \leqslant v) = F_X(v) < 1/2$$

entonces $\lim_{v\to\mu} F_X(v) \leqslant 1/2$, y por la continuidad por la derecha de F_X, tenemos que $\lim\limits_{v\to\mu} F_X(v) = P(X < \mu) \leqslant 1/2$, como queríamos demostrar.

2.12. Leggi dei grandi numeri

Definición 2.12.19 (Convergenza in probabilità)**.** Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Decimos que la sucesión **converge en probabilidad** a una variable aleatoria X, notado por $\{X_n\} \xrightarrow{P} X$, si:

$$\forall \varepsilon \in \mathbb{R}^+, \qquad \lim_{n\to\infty} P\left[|X_n - X| \geqslant \varepsilon\right] = 0$$

Definición 2.12.20 (Convergenza quasi certa)**.** Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, ,P)$. Decimos que la sucesión **converge casi seguramente** a una variable aleatoria X, que notaremos por $\{X_n\} \xrightarrow{c.s.} X$, si:

$$P\left[\lim_{n\to\infty} X_n = X\right] = 1$$

Análogamente, usando la probabilidad en $(\Omega, \mathcal{F}, P)$, se tiene que:

$$P\left[\omega \in \Omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)\right] = 1$$

Proposición 2.12.16. *Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$, entonces*

$$\{X_n\} \xrightarrow{c.s.} X \Longrightarrow \{X_n\} \xrightarrow{P} X$$

Demostración. □

Observación. La convergencia en probabilidad no implica la convergencia casi segura.

Proposición 2.12.17. *Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Si la sucesión converge en probabilidad, entonces*

$$\exists \text{ subsucesión } (n_j) \text{ tal que } \{X_{n_j}\} \xrightarrow{c.s.} X$$

Teorema 2.12.18 (Legge dei grandi numeri debole)**.** *Sea $(X_i)_{i\geqslant 1}$ una sucesión de variables aleatorias tales que*

- $\mathbb{E}[|X_i|^2] < \infty$ *para todo $i \geqslant 1$, es decir, $X_i \in \mathcal{L}^2(P).omega$*
- $Cov(X_i, X_j) = 0$ *para todo $i \neq j$, es decir, las variables aleatorias son no correladas dos a dos (por ejemplo, independientes).*
- $\sup_{i\geqslant 1} Var(X_i) < \infty$*, es decir, las varianzas están acotadas.*

Entonces

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \geqslant \varepsilon\right) = 0 \quad \textit{es decir} \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow{P} 0$$

De hecho, si $\mathbb{E}[X_i] = n$ para todo $i \geqslant 1$, entonces

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - n)\right| \geqslant \varepsilon\right) = 0 \quad \forall \varepsilon > 0$$

Demostración. □

Teorema 2.12.19 (Versione più generale)**.** *Sea $(X_i)_{i\geqslant 1}$ una sucesión de variables aleatorias independientes dos a dos e idénticamente distribuidas en $\mathcal{L}^1(P)$, entonces*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_1]$$

Ejemplo. Para el caso del lanzamiento de una moneda justa, tenemos la secuencia de variables aleatorias $(X_i)_{i\geqslant 1}$ donde la v.a. X_i representa el resultado del $i-$ésimo lanzamiento. Tenemos entonces

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = p \in]0, 1[\quad \text{casi seguro}$$

donde $\mathbb{E}[X_1] = p$.

Observación. La LGN débil es demasiado débil para poder capturar este tipo de convergencias.

Lema 2.12.20 (Lemma di Borel-Cantelli)**.** *Sea $(A_n)_{n\in\mathbb{N}}$ una sucesión de eventos en un espacio de probabilidad $(\Omega, \mathcal{F}, P)$, y sea*

$$A = \bigcap_{k\geqslant 1} \bigcup_{n=k}^{\infty} A_n = \{\omega \in \Omega : \omega \in A_k \text{ para } k \text{ valores } \infty \text{ de } k\}$$

entonces

$$si \sum_{n=1}^{\infty} P(A_n) < \infty \implies P(A) = 0$$

Demostración. □

Corolario 2.12.20.1. *Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $\{E_n\}_{n\in\mathbb{N}}$ una sucesión de eventos en $\mathcal{F}$. Si la suma de las probabilidades de estos eventos es finita*

$$\sum_{n=1}^{\infty} P(E_n) < \infty$$

Entonces, la probabilidad de que ocurran infinitos de estos eventos es cero

$$P(\limsup_{n\to\infty} E_n) = 0$$

Teorema 2.12.21 (Legge dei grandi numeri forte)**.** *Sea $(X_i)_{i\geqslant 1}$ una sucesión de variables aleatorias tales que*

- $\mathbb{E}[|X_i|^2] < \infty$ *para todo* $i \geqslant 1$*, es decir,* $X_i \in \mathcal{L}^2(P)$*.*
- $Cov(X_i, X_j) = 0$ *para todo* $i \neq j$*, es decir, las variables aleatorias son no correladas dos a dos (por ejemplo, independientes).*
- $\sup_{i\geqslant 1} Var(X_i) < \infty$*, es decir, las varianzas están acotadas.*

Entonces

$$P\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) = 0\right) = 1 \quad es\ decir \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \xrightarrow{c.s.} 0$$

Demostración. □

Observación. Es importante destacar que el conjunto

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) = 0$$

es realmente los $\omega \in \Omega$ tales que $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(X_i(\omega) - \mathbb{E}[X_i]) = 0$, y de este conjunto de ω's tenemos que la probabilidad es 1.

Corolario 2.12.21.1. *Sea $(X_i)_{i\geqslant 1}$ una sucesión de variables aleatorias idénticamente distribuidas y dos a dos no correladas en $\mathcal{L}^2(P)$ con $v := \sup_{i\geqslant 1} Var(X_i) < \infty$, entonces*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{c.s.} \mathbb{E}[X_1]$$

En el resultado anterior, la hipótesis de que las variables aleatorias son idénticamente distribuidas es lo que nos permite afirmar que $\mathbb{E}[X_i] = \mathbb{E}[X_1]$ para todo $i \geqslant 1$.

Proposición 2.12.22 (Composición de sucesiones convergentes c.s.)**.** *La composición de sucesiones convergentes casi seguramente también converge casi seguramente, es decir, si $\{X_n\} \xrightarrow{c.s.} X$ y $\{Y_n\} \xrightarrow{c.s.} Y$, entonces*

$$\{X_n + Y_n\} \xrightarrow{c.s.} X + Y \quad , \quad \{X_n Y_n\} \xrightarrow{c.s.} XY$$

2.13. Teorema Central del Límite

En la sección anterior hemos visto que

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \mathbb{E}[X_1] \quad \text{c.s.}$$

bajo las condiciones

- $(X_i)_{i\geqslant 1}$ son variables aleatorias no correladas dos a dos e idénticamente distribuidas, es decir, $P(X_i \in A)$ no depende de i.
- $\mathbb{E}[|X_1|^2] < \infty$, indicado solo para X_1 ya que todas las demás variables aleatorias tienen la misma distribución.

donde si además nombramos $m = \mathbb{E}[X_1]$, nos podemos preguntar cuál es el número $1-\alpha$ más grande tal que

$$\lim_{n\to\infty} n^{1-\alpha}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - m\right) = 0 \quad \text{c.s.}$$

Más adelante, veremos que el número más grande es $\alpha = 1/2$ gracias al Teorema Central del Límite, pero primero veremos unos enunciados que nos ayudarán a entender mejor este teorema.

Definición 2.13.21 (Convergenza in distribuzione)**.** Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$. Decimos que la sucesión **converge en distribución** a una variable aleatoria X, notado por $\{X_n\} \xrightarrow{d} X$, si:

$$\lim_{n\to\infty} F_{X_n}(t) = F_X(t) \quad \forall t \in C(F_X)$$

donde F_{X_n} y F_X son las funciones de distribución acumulada de X_n y X respectivamente, y $C(F_X)$ es el conjunto de puntos de continuidad de F_X.

Además, se dice que la sucesión **converge en distribución a una medida** μ de probabilidad en $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, notado por $\{X_n\} \xrightarrow{d} \mu$, si

$$\lim_{n\to\infty} F_{X_n}(t) = \mu(]-\infty, t]) \quad \forall t \in C(\mu)$$

donde $C(\mu)$ es el conjunto de puntos de continuidad de la función $\mu(]-\infty, t])$.

Notación. A veces en lugar de hablar de *convergencia en distribución* a una variable aleatoria X, se habla de convergencia en ley y se nota como $\{X_n\} \xrightarrow{L} X$.

Lema 2.13.23 (Caracterización de la convergencia en la distribución.)**.** *Sea $\{X_n\}_{n\in\mathbb{N}}$ una sucesión de variables aleatorias definidas sobre el mismo espacio de probabilidad $(\Omega, \mathcal{F}, P)$, las siguientes afirmaciones son equivalentes:*

1) $X_n \xrightarrow{d} X$

2) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \quad \forall f : \mathbb{R} \to \mathbb{R}$ *continua y acotada.*

Si F_X es continua en $\mathbb{R}$, entonces la siguiente afirmación también es equivalente

$$\lim_{n\to\infty} \sup_{c\in\mathbb{R}} |F_{X_n}(c) - F_X(c)| = 0$$

es decir, F_{X_n} converge uniformemente a F_X.

Demostración. Solo demostraremos la segunda parte del enunciado. □

Observación. Recordemos que $\mathbb{E}[f(X_n)] = \int_{\mathbb{R}} f(z) d\mu_n(z)$ donde $\mu_n(A) = P(X_n \in A)$.

Teorema 2.13.24 (Teorema Central del Límite)**.** *Sea $(X_i)_{i\geqslant 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas en $\mathcal{L}^2(P)$, es decir, tales que $\mathbb{E}[|X_i|^2] < \infty \ \forall i$. Si nombramos $m = \mathbb{E}[X_i]$ y $v = Var(X_i)$, entonces*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}} \xrightarrow{d} Z \sim N(0,1)$$

donde $N(0,1)$ es la distribución normal estándar con media 0 y varianza 1, es decir,

$$Z(]-\infty, t]) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{x^2}{2}}\, dx = \Phi(t) \quad \forall t \in \mathbb{R}$$

Además, como la función de distribución de la normal Φ es continua tenemos que

$$\lim_{n\to\infty} \sup_{c\in\mathbb{R}} |F_{S_n^*}(c) - \Phi(c)| = 0$$

dove $S_n^* = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} \frac{X_i - m}{\sqrt{v}}$

Demostración. □

Corolario 2.13.24.1. *Sean X, Y distribuciones gaussianas independientes con media 0 y varianza 1, entonces*

$$\mu_{\frac{X+Y}{\sqrt{2}}} \sim N(0,1)$$

De hecho, en general dadas X, Y variables aleatorias independientes con funciones de densidad f_X, f_Y, entonces la función de densidad de $X+Y$ es

$$f_{X+Y}(t) = \int_{\mathbb{R}} f_X(t-s) f_Y(s)\, ds$$

Dado el Teorema Central del Límite, vemos que este teorema parece ser una ley universal para todas las v.a. i.i.d. en $\mathcal{L}^2(P)$, por lo que es lógico preguntarse entonces que si de verdad funciona este teorema debería hacerlo también para v.a. que sigan una distribución normal de media 0 y varianza 1. A continuación, veremos un ejemplo que confirma esta hipótesis.

Ejemplo. Sea $(X_i)_{i\geqslant 1}$ una sucesión de v.a. independientes e idénticamente distribuidas con distribución $N(0,1)$, sabemos que su suma también sigue una distribución normal por su propiedad de reproductividad, es decir,

$$S_n = \sum_{i=1}^{n} X_i \sim N(0,n)$$

por lo que si normalizamos S_n como bien dice el Teorema Central del Límite, tenemos que

$$S_n^* = \frac{S_n - 0}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \sim N(0,1)$$

y por tanto S_n^* ya tiene la distribución límite que predice el Teorema Central del Límite. Hemos visto entonces, que como bien se preveía, este teorema funciona de forma general, incluso en el caso más obvio de v.a. con distribución normal.

3. Applicazioni

In questo capitolo vedremo alcune applicazioni pratiche dei concetti di statistica matematica studiati nei capitoli precedenti.

3.1. Regioni di confidenza

Imaginemos que tenemos una moneda trucada donde sale cara con probabilidad $p \in]0,1[$, ¿como puedo estimar p? Sabemos que

$$\lim_{n\to\infty} \frac{\#\text{caras}}{\#\text{lanzamientos}} = p \quad \text{c.s.}$$

y otra forma de hacerlo es lanzar la moneda n veces y contar el número de caras obtenidas, entonces una estimación natural de p es $p \approx \dfrac{\#\text{caras}}{\#\text{lanzamientos}}$.

Sin embargo, esta afirmación tiene dos problemas:

- no tiene en cuenta una posibilidad (improbable) que esté muy alejada de la media, como por ejemplo, que salgan todo caras.
- no cuantifica el error en función como de grande es n.

Definición 3.1.1 (Modello statistico)**.** Un **modelo estadístico** es una terna $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$, donde

- X es un conjunto.
- $\mathcal{F}$ es una σ-algebra sobre X.
- $(P_\theta : \theta \in \Theta)$ es una familia de medidas (al menos dos) de probabilidad sobre $(X, \mathcal{F})$ indexada por un conjunto de parámetros Θ.

Ejemplo. En el ejemplo de la moneda trucada con n lanzamientos tendríamos el modelo estadístico

$$(X, \mathcal{F}, P_p : p \in]0,1[)$$

donde

- $X = \{0, \ldots, n\}$ es el conjunto de posibles números de caras obtenidas en n lanzamientos.
- $\mathcal{F} = \mathcal{P}(X)$ es la σ-álgebra de partes de X.

- $P_\theta = B_{n,\theta}$, donde recordamos que

$$P(X = k) = B_{n,\theta}(\{k\}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

y tenemos $\Theta =]0, 1[$.

Definición 3.1.2 (Regione di confidenza)**.** Dado un modelo estadístico $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ y sea $\tau : \Theta \to \Sigma$ una característica del sistema que se quiere estimar, donde Σ es un conjunto arbitrario, entonces **una región de confianza para τ con nivel de error $\alpha \in]0, 1[$ (o nivel de confianza $1 - \alpha$)** es una aplicación

$$C : X \to \mathcal{P}(\Sigma) \quad \text{tal que} \quad \inf_{\theta\in\Theta} P_\theta(\{x \in X : \tau(\theta) \in C(x)\}) \geqslant 1 - \alpha$$

Si $\Sigma = \mathbb{R}$ y $C(x)$ es un intervalo, se dice que C es un **intervalo de confianza**.

Observación. Se puede usar indistintamente "'*nivel de error* α" o "*nivel de confianza* $1 - \alpha$".

Un caso trivial que podemos estudiar, es el caso en el que $C(x) = \Sigma$ para todo $x \in X$, ya que en este caso se cumple que

$$P_\theta(\{x \in X : \tau(\theta) \in C(x)\}) = P_\theta(\{x \in X : \tau(\theta) \in \Sigma\}) = P_\theta(\{x \in X\}) = P_\theta(X) = 1 \geqslant 1-\alpha \quad \forall\theta \in \Theta$$

por lo que en tal caso $\forall\alpha \in]0, 1[$, $C : X \to \{\Sigma\}$ es una región de confianza para τ con nivel de error α.

Observación. Si τ no se especifíca explícitamente, se asume que es la función identidad, es decir,

$$\tau : \Theta \to \Sigma = \Theta \ \text{ con } \ \tau(\theta) = \theta \quad \forall\theta \in \Theta$$

3.1.1. Metodo diretto per costruire regione di confidenza

Teorema 3.1.1 (Costruzione di intervalli di confidenza)**.** *Dado $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico y $\tau = id : \Theta \to \Theta$, la identidad, una característica del sistema que se quiere estimar. Entonces se siguen estos pasos para construir un intervalo de confianza para τ con nivel de error $\alpha \in]0, 1[$:*

- *Para cada $\theta \in \Theta$ encontrar el C_θ más pequeño tal que $P_\theta(C_\theta) \geqslant 1 - \alpha$ y $C_\theta \subseteq X$.*
- *$C : X \to \mathcal{P}(\Theta)$ está dada por $C(x) = \{\theta \in \Theta : x \in C_\theta\}$.*

Podemos ahora preguntarnos si realmente esta construcción de la aplicación C nos da un intervalo de confianza para τ con nivel de error α, lo que significa ver si cumple que

$$\inf_{\theta\in\Theta} P_\theta(\{x \in X : \tau(\theta) \in C(x)\}) \geqslant 1 - \alpha$$

Para ellos tomamos $\theta \in \Theta$ fijo, y tenemos que

$$y \in \{x \in X : \tau(\theta) = \theta \in C(x)\} \iff \theta \in C(y) = \{\theta \in \Theta \mid y \in C_\theta\} \iff y \in C_\theta$$

por lo que se cumple que $\{x \in X : \theta \in C(x)\} = C_\theta$, y por tanto para cada $\theta \in \Theta$ se cumple que

$$P_\theta(\{x \in X : \tau(\theta) \in C(x)\}) = P_\theta(C_\theta) \geqslant 1 - \alpha$$

es decir, se cumple $\inf_{\theta \in \Theta} P_\theta(\{x \in X : \tau(\theta) \in C(x)\}) \geqslant 1 - \alpha$, de manera que la construcción es totalmente correcta, pues nos da C que es un intervalo de confianza para τ con nivel de error α.

Intervalli di confidenza delle Binomiali

A continuación estudiaremos el caso particular de las distribuciones binomiales, y veremos cómo construir un intervalo de confianza para la característica del sistema $\tau = id : \Theta \to \Theta$, justificandolo mediante dos formas diferentes.
Sea $(X, \mathcal{F}, P_\mu : \theta \in \Theta)$ un modelo estadístico donde

- $X = \{0, \ldots, n\}$ cuenta las veces que salió cara en n lanzamientos de una moneda trucada.
- $P_\theta = B_{n,\theta}$ es la distribución binomial con parámetros n y θ.
- $\Theta =]0, 1[$.
- $\tau = id : \Theta \to \Theta$ es la identidad, una característica del sistema que se quiere estimar.

entonces un intervalo de confianza para τ con nivel de error $\alpha \in]0, 1[$ está dado por la aplicación $C : X \to \mathcal{P}(\Theta)$ definida como

$$C(x) = \left(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\right)$$

de manera que se intenta aproximar θ por la frecuencia relativa x/n, ya que sabemos que la mejor aproximación de θ es la frecuencia relativa de caras obtenidas en los lanzamientos, por lo que se busca un intervalo centrado en dicho punto. Buscaremos ahora como calcular el valor de $\varepsilon > 0$, mediante dos métodos.

Chebyshev. Sabiendo la restricción que debe cumplir P_θ, tenemos lo siguiente

$$P_\theta\left(\{x \in X : \theta \in C(x)\}\right) = P_\theta\left(\left\{x \in X : \theta \in \left(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\right)\right\}\right) = B_{n,\theta}\left(\left\{x \in X : \theta \in \left(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\right.\right.\right.$$

veámos a continuación cómo simplificar la expresión para θ

$$\theta \in \left(\frac{x}{n} - \varepsilon, \frac{x}{n} + \varepsilon\right) \iff \frac{x}{n} - \varepsilon < \theta < \frac{x}{n} + \varepsilon \iff -\varepsilon < \theta - \frac{x}{n} < \varepsilon \iff \left|\theta - \frac{x}{n}\right| < \varepsilon$$

tenemos por tanto

$$P_\theta\left(\{x \in X : \theta \in C(x)\}\right) = B_{n,\theta}\left(\left\{x \in X : \left|\theta - \frac{x}{n}\right| < \varepsilon\right\}\right) = 1 - B_{n,\theta}\left(\left\{x \in X : \left|\theta - \frac{x}{n}\right| \geqslant \varepsilon\right\}\right) \geqslant 1 -$$

$$\iff B_{n,\theta}\left(\left\{x \in X : \left|\theta - \frac{x}{n}\right| \geqslant \varepsilon\right\}\right) \leqslant \alpha$$

Definiremos ahora $A_1, \ldots, A_n$ eventos independientes con $P(A_i) = \theta$ de forma que la v.a. X viene dada como

$$X = \sum_{i=1}^{n} \mathbb{1}_{A_i} \sim B_{n,\theta}$$

entonces tenemos que

$$B_{n,\theta}\left(\left\{x \in X : \left|\theta - \frac{x}{n}\right| \geqslant \varepsilon\right\}\right) = P\left(\left|\frac{X}{n} - \theta\right| \geqslant \varepsilon\right) \overset{(*)}{\leqslant} \frac{Var\left[\frac{X}{n}\right]}{\varepsilon^2} = \frac{Var[X]}{n^2\varepsilon^2}$$

donde en $(*)$ hemos usado la desigualdad de Chebyshev, y sabemos que $\mathbb{E}[X/n] = \theta$. Desarrollando la varianza de X, obtenemos su valor simplificado

$$Var[X] = Var[\sum_{i=1}^{n} \mathbb{1}_{A_i}] \overset{(*)}{=} \sum_{i=1}^{n} Var[\mathbb{1}_{A_i}] = \sum_{i=1}^{n} \theta(1-\theta) = n\theta(1-\theta)$$

donde en $(*)$ hemos usado la independencia de los eventos A_i, y además es fácil ver que $Var[\mathbb{1}_{A_i}] = \theta(1-\theta)$. Finalmente, obtenemos la desigualdad esperada para ε

$$P_\theta\left(\{x \in X : \theta \in C(x)\}\right) \geqslant 1-\alpha \iff B_{n,\theta}\left(\left\{x \in X : |\theta - \frac{x}{n}| \geqslant \varepsilon\right\}\right) \leqslant \alpha \iff \frac{n\theta(1-\theta)}{n^2\varepsilon^2} = \frac{\theta(1-}{n\varepsilon^2}$$

Como no quiero que ε dependa de θ, usaré la cota superior de $\theta(1-\theta)$ en $]0,1[$, que es $1/4$, por lo que obtenemos la siguiente cota para ε

$$\frac{1}{4n\varepsilon^2} \leqslant \alpha \iff \varepsilon \geqslant \frac{1}{2\sqrt{n\alpha}}$$

Es importante destacar que el haber hecho esta cota extra y la anterior al tener en cuenta la varianza, hace que el intervalo de confianza obtenido sea más grande de lo necesario, pero a cambio no depende de θ. Notemos ahora lo siguiente sobre el valor de ε obtenido:

- Si $\alpha \ll 1$, entonces ε es grande, y el intervalo de confianza $C(x)$ es grande, lo cual es lógico ya que si queremos un nivel de error muy pequeño, el intervalo de confianza debe ser grande para asegurar que contiene el valor real de θ.
- Si n es grande, entonces ε es pequeño, y el intervalo de confianza $C(x)$ es pequeño como queremos, ya que si hacemos muchos lanzamientos, la frecuencia relativa se acerca mucho a θ, por lo que el intervalo de confianza debe ser pequeño.

Teorema del Límite Central. Recordemos que el TCL en el caso de la bernoulli nos dice que

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{1}_{A_i} - \theta}{\sqrt{\theta(1-\theta)}} \xrightarrow{d} N(0,1) \quad \text{cuando } n \to \infty$$

lo que implica que para $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ tenemos

$$F_{S_n^*} \to \Phi \quad \text{uniformemente en } \mathbb{R} \text{ cuando } n \to \infty \qquad \text{donde } S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\mathbb{1}_{A_i} - \theta}{\sqrt{\theta(1-\theta)}}$$

Antes de iniciar con los cálculos para obtener la aproximación de ε, veámos dos desarrollos importantes que usaremos después

- $S_n^* = \dfrac{1}{\sqrt{n}} \displaystyle\sum_{i=1}^{n} \frac{\mathbb{1}_{A_i} - \theta}{\sqrt{\theta(1-\theta)}} = \frac{1}{\sqrt{n\theta(1-\theta)}} \left(\sum_{i=1}^{n} \mathbb{1}_{A_i} - n\theta \right) = \frac{X - n\theta}{\sqrt{n\theta(1-\theta)}}$
- $\left| \dfrac{x}{n} - \theta \right| < \varepsilon \iff \sqrt{\dfrac{n}{\theta(1-\theta)}} \left| \dfrac{x - n\theta}{n} \right| < \varepsilon \sqrt{\dfrac{n}{\theta(1-\theta)}} \overset{(*)}{\iff} \dfrac{1}{\sqrt{n}} \left| \dfrac{x - n\theta}{\sqrt{\theta(1-\theta)}} \right| < \varepsilon \sqrt{\dfrac{n}{\theta(1-\theta)}}$

donde en $(*)$ hemos usado que $\dfrac{\sqrt{n}}{|n|} = \dfrac{1}{\sqrt{n}}$, ya que también sabemos que $n \in \mathbb{N}$. Ahora sí, podemos empezar a calcular la cota para ε

$$1 - \alpha \leqslant B_{n,\theta} \left(\left\{ x \in X : \left| \frac{x}{n} - \theta \right| < \varepsilon \right\} \right) = B_{n,\theta} \left(\left\{ x \in X : \frac{1}{\sqrt{n}} \left| \frac{x - n\theta}{\sqrt{\theta(1-\theta)}} \right| < \varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right\} \right) =$$

$$= P \left(\frac{1}{\sqrt{n}} \left| \frac{X - n\theta}{\sqrt{\theta(1-\theta)}} \right| < \varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) = P \left(|S_n^*| < \varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) \overset{TCL}{\approx}$$

$$\overset{TCL}{\approx} \Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) - \Phi \left(-\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) = 2\Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) - 1$$

Tenemos finalmente que

$$1 - \alpha \leqslant 2\Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) - 1 \implies \Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) \geqslant \frac{2-\alpha}{2} \implies$$

$$\implies \varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \geqslant \Phi^{-1} \left(\frac{2-\alpha}{2} \right) \implies \varepsilon \geqslant \Phi^{-1} \left(\frac{2-\alpha}{2} \right) \sqrt{\frac{\theta(1-\theta)}{n}}$$

Como bien sabemos si queremos un nivel de error α muy bajo, debemos por tanto tener un valor de ε grande, es decir, un intervalor C grande. Veámos que esto se cumple en la expresión obtenida, suponiendo que $\alpha \to 0$

$$2\Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) - 1 \approx 1 \iff \Phi \left(\varepsilon \sqrt{\frac{n}{\theta(1-\theta)}} \right) \approx 1 \iff \varepsilon \to +\infty$$

donde claramente para el último resultado hemos usado la definición de Φ, es decir, de función de distribución de la normal estándar.

Observación. Asumimos que C es una aplicación tal que $\{x \in X \mid \tau(\theta) \in C(x)\} \in \mathcal{F}$.

3.1.2. Metodo alternativo per costruire regione di confidenza

Fijado el modelo estadístico $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$, supongamos que existe una función T_θ tal que $P_\theta \circ T_\theta^{-1}$ es constante, ¿como podemos usar esta función para construir C la región de confianza para τ con nivel de error α?
La idea es poder definir $Q = P_\theta \circ T_\theta^{-1}$, de manera que dado $\alpha \in]0,1[$ y I_α tal que $Q(I_\alpha) \geqslant 1 - \alpha$, entonces $P_\theta(T_\theta^{-1}(I_\alpha)) = Q(I_\alpha) \geqslant 1 - \alpha$, por lo que demostrando $\{\theta \in \Theta : T_\theta(x) \in I_\alpha\} = T_\theta^{-1}(I_\alpha)$ ya habríamos terminado. Veámos primero varias definiciones y enunciados que nos llevarán a dicho resultado.

Definición 3.1.3 (Pivot per una caratteristica)**.** Dado un modelo estadístico $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ y sea $\tau : \Theta \to \Sigma$ una característica del sistema que se quiere estimar, entonces se dice que $(Q, T_\mu : \mu \in \Sigma)$ es un **pivote para** τ si

- $\forall \mu \in \Sigma$, $T_\mu : X \to \Sigma'$ es una función medible, donde $(\Sigma', \mathcal{F}')$ es un espacio medible.
- $Q = P_\theta \circ T^{-1}_{\tau(\theta)} \quad \forall \theta \in \Theta$.

Observación. Importante notar que Q no depende de $\theta \in \Theta$, es decir, para todo $\theta \in \Theta$ sabemos que Q es la misma, y además es una medida de probabilidad sobre $(\Sigma', \mathcal{F}')$.

El pivote es una herramienta que "estandariza" el problema para que siempre trabajes con la misma distribución Q, facilitando enormemente el cálculo del intervalo.

Notación. $T^{-1}_{\tau(\theta)}(A) = \{x \in X : T_{\tau(\theta)}(x) \in A\}$ para todo $A \in \mathcal{F}'$.

Lema 3.1.2. *Dado $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico. Sea $(Q, T_\mu : \mu \in \Sigma)$ un pivote para la característica $\tau : \Theta \to \Sigma$, $\alpha \in]0, 1[$ y $I \subseteq \Sigma'$ tal que $Q(I) \geqslant 1 - \alpha$, entonces la aplicación*

$$C : X \to \mathcal{P}(\Sigma), \quad C(x) = \{\mu \in \Sigma : x \in T^{-1}_\mu(I)\}$$

es una región de confianza para τ con nivel de error α.

Demostración. □

Ejemplo (Prodotto Gaussiano)**.** Consideremos el modelo estadístico dado por el producto Gaussiano $n-$veces

$$(X, \mathcal{F}, P_\mu : \mu \in \mathbb{R}) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), N^{\otimes n}_{m,v} : m \in \mathbb{R}, v > 0)$$

como viene dado dicho modelo queremos obtener una región de confianza $1 - \alpha$ para poder estimar la media y la varianza, por lo que tenemos que los parámetros son $\Theta = \mathbb{R} \times \mathbb{R}^+$, donde $\theta = (m, v)$ y la característica del sistema que queremos estimar es $\tau : \Theta \to \Sigma = \Theta$, con $\tau(\theta) = \theta$.

Buscamos por tanto, si existe, un pivote para τ, es decir, un mapeo T_θ tal que $N^{\otimes n}_{m,v} \circ T^{-1}_{m,v}$ sea constante para m, v. Una buena elección para $T_{m,v}$ es la siguiente

$$T_{m,v} : \mathbb{R}^n \to \mathbb{R}^n \quad \text{dada por} \quad T_{m,v}(x) = \left(\frac{x_i - m}{\sqrt{v}}\right)_{i=1,\dots,n} = \frac{x - m}{\sqrt{v}}$$

Veámos ahora que efectivamente $N^{\otimes n}_{m,v} \circ T^{-1}_{m,v} = N^{\otimes n}_{0,1}$ para todo $m \in \mathbb{R}$ y $v > 0$, y por tanto $Q = N^{\otimes n}_{0,1}$ no depende de θ. Basta demostrarlo para $A_1, \dots, A_n \in \mathcal{B}(\mathbb{R})$, ya que los rectángulos generan la σ-álgebra de Borel en $\mathbb{R}^n$, de modo que tenemos

$$\begin{aligned}
&N^{\otimes n}_{m,v} \circ T^{-1}_{m,v}(A_1 \times \ldots \times A_n) = N^{\otimes n}_{m,v}\left(\{x \in \mathbb{R}^n : T_{m,v}(x) \in A_1 \times \ldots \times A_n\}\right) = \\
&= N^{\otimes n}_{m,v}\left(\left\{x \in \mathbb{R}^n : \frac{x_i - m}{\sqrt{v}} \in A_i, i = 1, \dots, n\right\}\right) = N^{\otimes n}_{m,v}\left(\prod_{i=1}^n (\sqrt{v}A_i + m)\right) = \\
&= \prod_{i=1}^n N_{m,v}(\sqrt{v}A_i + m) = \prod_{i=1}^n N_{0,1}(A_i) = N^{\otimes n}_{0,1}(A_1 \times \ldots \times A_n)
\end{aligned}$$

NO ESTA TERMINADO

Observación. En la demostración se ha usado una idea muy importante sobre $\sigma-$álgebras de Borel en $\mathbb{R}$, que es que los intervalos abiertos generan la $\sigma-$álgebra de Borel en $\mathbb{R}$, de ahí que baste demostrar la igualdad en los rectángulos (los generadores) para que se cumpla en toda la $\sigma-$álgebra de Borel en $\mathbb{R}^n$.

4. Inferenza Statistica

A continuación veremos los elementos básicos para la inferencia estadística, es decir, cómo a partir de una muestra aleatoria podemos inferir propiedades de la población de la que procede dicha muestra. La inferencia estadística es la rama de la estadística que intenta entender un fenómeno gobernado por leyes probabilísticas a través de observaciones, es decir, de tomar pequeñas muestras de toda la población.

Normalmente los datos son:

- **Relaciones input-output** (*problema supervisado*): el objetivo es obtener una relación entre las variables de entrada y las variables de salida, es decir, se intenta estimar $y = \hat{f}(x)$ a partir de un conjunto de datos $D = \{(x^A, y^A)\}_{i=1}^M$, donde cada par $(x^A, y^A) \in X \times Y \subseteq \mathbb{R}^p \times \mathbb{R}^d$ es una observación del fenómeno que se quiere estudiar y M es el número de observaciones.
  Se trata de un **problema de regresión**.

- **Observación de una sola magnitud de interés** (*problema no supervisado*): el objetivo es definir la probabilidad de observar un dato x, o, mejor dicho, encontrar una función de probabilidad P que tiene en cuenta los datos observados para estimar la probabilidad de ver un dato x. Toma como base un conjunto de datos $D = \{x^A\}_{i=1}^M$, donde cada $x^A \in X \subseteq \mathbb{R}^p$ es una observación del fenómeno que se quiere estudiar y M es el número de observaciones.
  Se trata de un **problema de maximización de la entropía/likelihood**.

- **Series temporales**: el objetivo es predecir la evolución futura de una variable en función de su evolución pasada, es decir, obtener la distribución P de forma que se intenta calcular la probabilidad de darse un suceso x_t en el momento t en función de lo ya visto, lo que equivale a $P(x_t \mid x_1, \ldots, x_{t-1})$. En este caso, la información a usar se codifica como una secuencia $x^A = (x_1^A, x_2^A, \ldots, x_T^A)$, donde cada $x_t^A \in X \subseteq \mathbb{R}^p$ es la observación en el tiempo t del fenómeno que se quiere estudiar y T es el número de observaciones temporales.
  Se trata de un **problema autorregresivo**.

Hay varias maneras de justificar el por qué usamos la probabilidad para modelar los datos observados

- **Error en la medida**: los datos pueden contener errores de medida. Por ejemplo, en el caso del problema de regresión, los datos obtenidos tienen la siguiente forma
$$y^A = f(x^A) + \varepsilon$$

donde f es la función de verdad que se quiere estimar y ε es un término de error que se modela como una variable aleatoria. El objetivo es obtener $\hat{f} : X \to Y$ de modo que $y = \hat{f}$ aproxime f lo mejor posible.

NO está terminado.

Definición 1 (Statistica)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico, y $(\Sigma, \mathcal{F}')$ un espacio medible, llamamos **estadística** a una función $S : X \to \Sigma$ medible.

Definición 2 (Stimatore)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico, $(\Sigma, \mathcal{F}')$ un espacio medible y $\tau : \Theta \to \Sigma$ una función que asigna a cada $\theta \in \Theta$ una cierta característica $\tau(\theta) \in \Sigma$. Llamamos **estimador de** τ a una estadística $T_\tau : X \to \Sigma$ tal que T_τ aproxime a $\tau(\theta)$ para cada $\theta \in \Theta$.

Ejemplo. Supongamos que estamos estudiando la altura de los estudiantes. Asumimos que siguen una Normal $N(\mu, \sigma^2)$. Aquí, el parámetro es el vector $\theta = (\mu, \sigma^2)$.

- **Escenario A**: Queremos estimar la media. Definimos $\tau(\theta) = \mu$. (Esta es la çaracterística"que queremos). El estimador T_τ sería la media muestral: $\bar{X} = \frac{1}{n}\sum X_i$. ¿Qué estima T_τ? Estima el valor de μ.
- **Escenario B**: Queremos estimar la varianza. Definimos $\tau(\theta) = \sigma^2$. El estimador T_τ sería la varianza muestral: $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$. ¿Qué estima T_τ? Estima el valor de σ^2.

Definición 3 (Stimatore non distorto)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico y $(\Sigma, \mathcal{F}')$ un espacio medible. Si tenemos $\tau : \Theta \to \Sigma$ una función que asigna a cada $\theta \in \Theta$ una cierta característica $\tau(\theta) \in \Sigma$ y $T_\tau : X \to \Sigma$ un estimador de τ, decimos que T_τ es un **estimador no distorsionado** (o no sesgado) de τ si se cumple que

$$E_\theta[T_\tau] = \int_X T_\tau(x) \cdot f(x;\theta)\, dx = \tau(\theta) \quad \forall \theta \in \Theta$$

Definición 4 (Estimador asintóticamente no sesgado)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico y $(\Sigma, \mathcal{F}')$ un espacio medible. Si tenemos $\tau : \Theta \to \Sigma$ una función que asigna a cada $\theta \in \Theta$ una cierta característica $\tau(\theta) \in \Sigma$ y $(T_\tau^{(n)})_{n\in\mathbb{N}}$ una sucesión de estimadores de τ, decimos que la sucesión $(T_\tau^{(n)})_{n\in\mathbb{N}}$ es un **estimador asintóticamente no sesgado** de τ si se cumple que

$$\lim_{n\to\infty} E_\theta[T_\tau^{(n)}] = \tau(\theta) \quad \forall \theta \in \Theta$$

Definición 5 (Stimatore consistente)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico y $(\Sigma, \mathcal{F}')$ un espacio medible. Si tenemos $\tau : \Theta \to \Sigma$ una función que asigna a cada $\theta \in \Theta$ una cierta característica $\tau(\theta) \in \Sigma$ y $(T_\tau^{(n)})_{n\in\mathbb{N}}$ una sucesión de estimadores de τ, decimos que la sucesión $(T_\tau^{(n)})_{n\in\mathbb{N}}$ es un **estimador consistente** de τ si se cumple que

$$T_\tau^{(n)} \xrightarrow{n\to\infty} \tau(\theta) \quad \text{casi seguro} \quad \forall \theta \in \Theta$$

4.1. Valores muestrales

En esta sección trabajaremos con $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas (i.i.d.) cn $\mathbb{E}[X_i] = \mu$ y $\text{Var}(X_i) = \sigma^2 < +\infty$. Un estimador de la media μ es la **media muestral** (media empirica)

$$T_\mu = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

y como $\mathbb{E}[\bar{X}] = \frac{1}{n}\sum \mathbb{E}[X_i] = \mu$, tenemos que $\bar{X}$ es un **estimador no distorsionado** de μ. Además, por la ley fuerte de los grandes números, tenemos que $\bar{X} \xrightarrow{q.c.} \mu$ cuando $n \to \infty$, por lo que $\bar{X}$ es un **estimador consistente** de μ.

Observación. Es importante destacar que aunque no se haya indicado en el estimador $T_\mu = \bar{X}$, este depende del tamaño muestral n, por lo que en general escribimos $T_\mu^{(n)} = \bar{X}^{(n)}$.

Buscaremos ahora un estimador para la varianza σ^2. Suponiendo no conocido μ, un estimador natural es la **varianza muestral** (varianza campionaria)

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Calculemos ahora su esperanza, donde usaremos el hecho de que

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + (\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \implies \mathbb{E}[X^2] = Var(X) + (\mathbb{E}[X]$$

Veámos a continuación que S_n^2 es un **estimador distorsionado** de σ^2, calculando su esperanza

$$\begin{aligned}
\mathbb{E}[S_n^2] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \frac{2}{n}\bar{X}\sum_{i=1}^{n}X_i + \bar{X}^2\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - 2\bar{X}^2 + \bar{X}^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}X_i^2 - \bar{X}^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] - \mathbb{E}[\bar{X}^2] \\
&= \frac{1}{n}\sum_{i=1}^{n}(Var(X_i) + (\mathbb{E}[X_i])^2) - \mathbb{E}[\bar{X}^2] = \sigma^2 + \mu^2 - \mathbb{E}[\bar{X}^2] \overset{(*)}{=} \sigma^2 + \mu^2 - \frac{\sigma^2}{n} - \mu^2 = \frac{n-1}{n}\sigma^2
\end{aligned}$$

donde en $(*)$ hemos usado el siguiente desarrollo

$$\begin{aligned}
\mathbb{E}[\bar{X}^2] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2\right] = \mathbb{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n}X_iX_j\right] = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbb{E}[X_iX_j] \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \sum_{\substack{i,j=1 \\ i\neq j}}^{n}\underbrace{\mathbb{E}[X_i]\mathbb{E}[X_j]}_{\mu^2}\right) = \frac{1}{n^2}\sum_{i=1}^{n}(Var(X_i) + (\mathbb{E}[X_i])^2) + \frac{1}{n^2}n(n-1)\mu^2 \\
&= \frac{1}{n^2}n(\sigma^2 + \mu^2) + \frac{n(n-1)}{n^2}\mu^2 = \frac{\sigma^2}{n} + \mu^2
\end{aligned}$$

Como podemos ver la varianza muestral S_n^2 subestima el valor real de la varianza σ^2. Para corregir este sesgo definimos la **varianza muestral corregida o no distorsionada** (varianza campionaria corretta o non distorta)

$$S^2 = \frac{n}{n-1} S_n^2 \quad \text{y} \quad \mathbb{E}[S^2] = \sigma^2$$

Observación. Para los casos que se tiene como característica de interés la media o la varianza, es decir, $\tau_1(\mu, \sigma^2) = \mu, \tau_2(\mu, \sigma^2) = \sigma^2$, los estimadores no distorsionados y consistentes son respectivamente la media muestral $\bar{X}$ y la varianza muestral corregida S^2.

Definición 4.1.6. La **función error** $erf : \mathbb{R} \to \mathbb{R}$ se define como sigue

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$$

Proposición 4.1.1. *Dada la función de error erf, se cumplen las siguientes propiedades*

- $erf(-x) = -erf(x) \quad \forall x \in \mathbb{R}$ *(es una función impar)*
- $\lim_{x \to +\infty} erf(x) = 1$ *y* $\lim_{x \to -\infty} erf(x) = -1$
- *La derivada de la función de error es* $erf'(x) = \frac{2}{\sqrt{\pi}} e^{-x^2} \quad \forall x \in \mathbb{R}$

donde destacamos que en la segunda propiedad se ha hecho uso del hecho de la integral $\int_{-\infty}^{+\infty} e^{-a(x+b)^2} dt = \sqrt{\frac{\pi}{a}}$, *con* $a \in \mathbb{R}^+, b \in \mathbb{R}$.

Observación. Dada la función de error, podemos redefinir la **función acumulativa gaussiana** de una distribución normal estándar $\mathcal{N}(0,1)$ como sigue

$$\Phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x}{\sqrt{2}}\right)\right) \quad \forall x \in \mathbb{R}$$

Ejemplo. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias i.i.d. con distribución $\mathcal{N}(\mu, 1)$, queremos estudiar como estimar el intervalo de confianza para la media μ con un nivel de confianza $1 - \alpha$, sabiendo que el intervalo tiene la forma

$$C(X) = \left[\bar{X} - t, \bar{X} + t\right]$$

por lo que estamos interesados en estudiar $P(\mu \in C(X)) = P(|\bar{X} - \mu| \leqslant t)$. Usando las propiedades de la distribución normal, tenemos que

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{1}{n}\right) \Longrightarrow Z = \frac{\bar{X} - \mu}{1/\sqrt{n}} = \sqrt{n}(\bar{X} - \mu) \sim \mathcal{N}(0,1)$$

donde en el último paso usado la normalización de la variable aleatoria. Estudiemos por tanto ahora esta nueva v.a. usando la función acumulativa gaussiana, ya que ahora si podemos usarla al tratarse de una normal estándar

$$P(\sqrt{n}|\bar{X} - \mu| \leqslant \varepsilon) = P(|Z| \leqslant \varepsilon) = \Phi(\varepsilon) - \Phi(-\varepsilon) = \frac{1}{2}\left(1 + erf\left(\frac{\varepsilon}{\sqrt{2}}\right) - 1 - erf\left(-\frac{\varepsilon}{\sqrt{2}}\right)\right)$$
$$= \frac{1}{2}\left(2erf\left(\frac{\varepsilon}{\sqrt{2}}\right)\right) = erf\left(\frac{\varepsilon}{\sqrt{2}}\right)$$

Veámos ahora como podemos acotar superiormente $erf(x)$, ya que nos ayudará a definir nuestro intervalo de confianza, con confianza superior ($\geqslant$) a $1-\alpha$

$$1-erf(x) \overset{(*)}{=} \frac{2}{\sqrt{\pi}}\int_0^{+\infty} e^{-t^2}dt - \frac{2}{\sqrt{\pi}}\int_0^{x} e^{-t^2}dt = \frac{2}{\sqrt{\pi}}\int_x^{+\infty} e^{-t^2}dt = \frac{2}{\sqrt{\pi}}\int_x^{+\infty} \frac{t}{t}e^{-t^2}dt \leqslant \tag{4.1}$$

$$\overset{(**)}{\leqslant} \frac{2}{\sqrt{\pi}}\int_x^{+\infty} \frac{t}{x}e^{-t^2}dt = \frac{2}{2x\sqrt{\pi}}\int_x^{+\infty} 2te^{-t^2}dt = \frac{1}{x\sqrt{\pi}}\left[-e^{-t^2}\right]_x^{+\infty} = \frac{e^{-x^2}}{x\sqrt{\pi}}$$

donde en $(*)$ hemos usado el hecho de que $\int_0^{+\infty} e^{-t^2}dt = \dfrac{\sqrt{\pi}}{2}$ y en $(**)$ hemos usado que $t \geqslant x$, es decir, $\frac{1}{x} \geqslant \frac{1}{t}$ en el intervalo de integración. Usando esta cota superior, tenemos que

$$erf(x) \geqslant 1 - \frac{e^{-x^2}}{x\sqrt{\pi}} \quad \forall x>0 \Longrightarrow P(\sqrt{n}|\bar{X}-\mu| \leqslant \varepsilon) \geqslant 1 - \frac{\sqrt{2}e^{-\varepsilon^2/2}}{\varepsilon\sqrt{\pi}}$$

y tomando por tanto $\varepsilon = t\sqrt{n}$, podemos calcular como se quería el intervalo de confianza para la media μ con un nivel de confianza $1-\alpha$, lo que equivale a la siguiente probabilidad (recordemos que sabemos como era el intervalo de confianza)

$$P(|\bar{X}-\mu| \leqslant t) \geqslant 1 - \frac{\sqrt{2}e^{-n\frac{t^2}{2}}}{t\sqrt{n\pi}}$$

Dado ya esto, veámos dos ejemplos concretos

- Sea $n=100$ y queremos un nivel de confianza del 95 %, es decir, $\alpha = 0{,}05$. Busquemos el valor aproximado de t usando la cota superior obtenida

  $$0{,}05 = \alpha = \frac{\sqrt{2}e^{-100\frac{t^2}{2}}}{t\cdot 10\sqrt{\pi}} \Longrightarrow t \approx 0{,}18$$

  por lo que el intervalo de confianza de 95 % para $n=100$ es $C(X) = [\bar{X} - 0{,}18, \bar{X} + 0{,}18]$.

- Sea $n=1000$ y queremos un nivel de confianza del 95 %, es decir, $\alpha = 0{,}05$. Busquemos el valor aproximado de t usando la cota superior obtenida

  $$0{,}05 = \alpha = \frac{\sqrt{2}e^{-1000\frac{t^2}{2}}}{t\cdot \sqrt{1000\pi}} \Longrightarrow t \approx 0{,}06$$

  por lo que el intervalo de confianza de 95 % para $n=1000$ es $C(X) = [\bar{X} - 0{,}06, \bar{X} + 0{,}06]$.

Definición 4.1.7 (Error estimador)**.** Sea $(X, \mathcal{F}, P_\theta : \theta \in \Theta)$ un modelo estadístico y $(\Sigma, \mathcal{F}')$ un espacio medible. Si tenemos $\tau : \Theta \to \Sigma$ una función que asigna a cada $\theta \in \Theta$ una cierta característica $\tau(\theta) \in \Sigma$ y $T_\tau : X \to \Sigma$ un estimador de τ, definimos el **error cuadrático medio del estimador** como el valor esperado de la diferencia al cuadrado entre el estimador y el parámetro

$$MSE(T_\tau) = E_\theta\left[(T_\tau - \tau(\theta))^2\right] \quad \equiv \quad MSE(\hat{\theta}) = E_\theta\left[(\hat{\theta}-\theta)^2\right]$$

que desarrolado queda como

$$MSE(T_\tau) = Var_\theta(T_\tau) + (Bias_\theta(T_\tau))^2 \quad \equiv \quad MSE(\hat{\theta}) = Var(\hat{\theta}) + \Big[Bias(\hat{\theta})\Big]^2$$

donde $Bias_\theta(T_\tau) = E_\theta[T_\tau] - \tau(\theta)$ es el **sesgo del estimador**.

Definición 4.1.8 (Minimum variance unbiased)**.** Sea T_τ un estimador no distorsionado de $\tau(\theta)$, es decir, $Bias_\theta(T_\tau) = 0$. Decimos que T_τ es un **estimador MVUB (varianza mínima no sesgada)** si se cumple que

$$Var(T_\tau) \leqslant Var(S_\tau) \quad \forall S_\tau \text{ estimador no distorsionado de } \tau(\theta)$$

Ejemplo (Libro 7.3)**.** En un programa de televisión, el presentador muestra una máquina que genera números aleatorios en el intervalo $[0;\theta]$, si el presentador la ha ajustado al valor $\theta > 0$. A dos jugadores se les permite usar la máquina $n = 10$ veces y luego deben adivinar θ. Quien se acerque más gana.

Es claro que la v.a. viene dada como $X = [0,\theta]$ con distribución uniforme $U(0,\theta)$, por lo que el modelo estadístico es $(X, \mathcal{B}([0,\theta]), P_\theta : \theta > 0)$, donde P_θ tiene función de densidad

$$f(x;\theta) : \mathbb{R} \to [0,1], \quad f(x;\theta) = \begin{cases} \dfrac{1}{\theta} & x \in [0,\theta] \\ 0 & x \notin [0,\theta] \end{cases}$$

Ahora veremos dos posibles estimadores para θ:

- **Estimador 1**: el primer jugador para calcular su estimador usará la ley de los grandes números, por lo que primero calcularemos la esperanza de la distribución uniforme
$$\mathbb{E}[X] = \int_0^\theta x \cdot \frac{1}{\theta} dx = \frac{1}{\theta} \cdot \frac{\theta^2}{2} = \frac{\theta}{2}$$
de modo que gracias a la ley de los grandes números tenemos que
$$\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \xrightarrow{q.c.} \frac{\theta}{2}$$
entonces el primer jugador usará como estimador
$$T_n^{(A)} = \frac{2}{n}\sum_{i=1}^n X_i$$
Es importante destacar que $T_n^{(A)}$ es un estimador no distorsionado, veámoslo
$$\mathbb{E}[T_n^{(A)}] = \frac{2}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{2}{n} \cdot n \cdot \frac{\theta}{2} = \theta$$

- **Estimador 2**: el segundo jugador usará como estimador el máximo de la muestra, es decir,
$$T_n^{(B)} = \text{máx}\{X_1, X_2, \ldots, X_n\}$$

que claramente también converge a θ casi seguro cuando $n \to \infty$ (al igual que el otro estimador). Veámos ahora si es un estimador distorsionado o no distorsionado calculando su esperanza

$$\mathbb{E}[T_n^{(B)}] = \int_{-\infty}^{\infty} x \cdot f_{T_n^{(B)}}(x;\theta)dx$$

para ello necesitamos calcular la función de densidad de $T_n^{(B)}$. Primero calculamos su función acumulativa

$$F_{T_n^{(B)}}(x;\theta) = P(T_n^{(B)} \leqslant x) = P(\text{máx}\{X_1, X_2, \ldots, X_n\} \leqslant x) = P(X_1 \leqslant x, X_2 \leqslant x, \ldots, X_n \leqslant x) =$$
$$\overset{(*)}{=} \prod_{i=1}^{n} P(X_i \leqslant x) = (F_X(x;\theta))^n$$

donde en $(*)$ hemos usado la independencia de las v.a. Ahora, como la función de acumulación viene dada así

$$F_X(x;\theta) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{\theta} & x \in [0,\theta] \\ 1 & x > \theta \end{cases}$$

podemos calcular la función de densidad de $T_n^{(B)}$ derivando la función de acumulación (recordemos que aplicando el la continuidad local de la función de acumulación, la función de densidad es su derivada casi en todas partes)

$$f_{T_n^{(B)}}(x;\theta) = \frac{d}{dx}F_{T_n^{(B)}}(x;\theta) = \begin{cases} 0 & x < 0 \\ \dfrac{nx^{n-1}}{\theta^n} & x \in]0,\theta[\\ 0 & x > \theta \end{cases}$$

notar que en $x = 0$ y $x = \theta$ no está definida, pero da igual ya que son puntos de medida nula, por lo que podemos darle cualquier valor, por convenio se introducen en el intervalo $[0,\theta]$. Ahora sí podemos calcular la esperanza

$$\mathbb{E}[T_n^{(B)}] = \int_0^{\theta} x \cdot f_{T_n^{(B)}}(x;\theta)dx = \int_0^{\theta} x \cdot \frac{nx^{n-1}}{\theta^n}dx = \frac{n}{\theta^n}\int_0^{\theta} x^n dx =$$
$$= \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta$$

es decir, se trata de un estimador distorsionado, por lo que para corregir el sesgo definimos el siguiente estimador no distorsionado

$$\tilde{T}_n^{(B)} = \frac{n+1}{n}T_n^{(B)} = \frac{n+1}{n}\,\text{máx}\{X_1, X_2, \ldots, X_n\}$$

Dados los dos estimadores no distorsionados $T_n^{(A)}$ y $\tilde{T}_n^{(B)}$, calculemos ahora su error cuadrático medio para compararlos, y por tanto ver cual es mejor. Tengamos en

cuenta primero que como ambos son no distorsionados tenemos que su sesgo es nulo ($Bias_\theta(T_n^{(A)}) = Bias_\theta(\tilde{T}_n^{(B)}) = 0$),por lo que calculemos ambas varianzas

$$Var_\theta(T_n^{(A)}) = Var_\theta\left(\frac{2}{n}\sum_{i=1}^{n} X_i\right) = \frac{4}{n^2}\sum_{i=1}^{n} Var_\theta(X_i) = \frac{4}{n^2}\cdot n \cdot Var_\theta(X) = \frac{4}{n}\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3n}$$

$$Var_\theta(\tilde{T}_n^{(B)}) = \left(\frac{n+1}{n}\right)^2 Var_\theta(T_n^{(B)}) = \left(\frac{n+1}{n}\right)^2 \left(E_\theta[(T_n^{(B)})^2] - (E_\theta[T_n^{(B)}])^2\right) = \frac{\theta^2}{n(n+2)}$$

donde hemos usado que las v.a. son independientes e idénticamente distribuidas, y además

$$Var_\theta(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_0^\theta \left(x - \frac{\theta}{2}\right)^2 \cdot \frac{1}{\theta} dx = \frac{1}{\theta}\int_0^\theta \left(x^2 - \theta x + \frac{\theta^2}{4}\right) dx = \frac{\theta^2}{12}$$

$$E[T_n^{(B)}] = \int_0^\theta x \cdot \frac{n x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n}\int_0^\theta x^n dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1}\theta$$

$$E[(T_n^{(B)})^2] = \int_0^\theta x^2 \cdot \frac{n x^{n-1}}{\theta^n} dx = \frac{n}{\theta^n}\int_0^\theta x^{n+1} dx = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{n+2}\theta^2$$

Finalmente, calculamos los errores cuadráticos medios para el caso $n = 10$

$$MSE(T_n^{(A)}) = Var_\theta(T_n^{(A)}) = \frac{\theta^2}{3\cdot 10} = \frac{\theta^2}{30}$$
$$MSE(\tilde{T}_n^{(B)}) = Var_\theta(\tilde{T}_n^{(B)}) = \frac{\theta^2}{10\cdot 12} = \frac{\theta^2}{120}$$

por lo que el segundo estimador es mejor que el primero, ya que su error cuadrático medio es menor.

4.2. Regressione lineare semplice

Trabajaremos con un problema supervisado de regresión lineal simple, es decir, con datos de la forma $(x_i, y_i) \in \mathbb{R}^2$, donde x_i es la variable independiente (predictora) y y_i la variable dependiente (respuesta). Supondremos que existe una relación lineal entre ambas variables, es decir, que

$$y_i = \theta_2 + \theta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, n$$

donde los datos con los que trabajaremos serán $D = \{(x_i, y_i)\}_{i=1}^n$, y ε_i es un término de error que modelaremos como una variable aleatoria con distribución normal $\mathcal{N}(0, \sigma^2)$, es decir, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Nuestro objetivo es estimar los parámetros $\theta = (\theta_1, \theta_2)$ a partir de los datos observados D.

Para ello, como sabemos la distribución que sigue el error, tenemos que $\varepsilon_i = y_i - (\theta_2 + \theta_1 x_i) \sim \mathcal{N}(0, \sigma^2$ por lo que la probabilidad para cada observación i es

$$P_\theta(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\theta_2 + \theta_1 x_i))^2}{2\sigma^2}\right)$$

donde $\theta = (\theta_1, \theta_2)$. Suponiendo que las observaciones son independientes, la probabilidad conjunta de todas las observaciones es

$$P_\theta(ETICHETTE \mid DATI) = \prod_{i=1}^{n} P_\theta(y_i|x_i) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - (\theta_2 + \theta_1 x_i))^2\right)$$

donde se ha usado que $P(y_i, y_j \mid x_i, x_j) = P(y_i \mid x_i) \cdot P(y_j \mid x_j)$ con $i \neq j$ por la independencia.

Para obtener estos estimadores de θ_1 y θ_2, que son $T_{\tau_1} = \theta_1^*$ es **estimador de la pendiente** (stimatore di dependenza) y $T_{\tau_2} = \theta_2^*$ el **estimador de intersección** (stimatore di intercetta), usaremos el **método de los mínimos cuadrados**, que consiste en minimizar la suma de los errores al cuadrado, es decir, realizar este cálculo

$$T_{\tau_1}, T_{\tau_2} = \arg\min_{\theta_1,\theta_2} \frac{1}{2}\sum_{i=1}^{n}(y_i - (\theta_2 + \theta_1 x_i))^2$$

donde se ha usado que las características de interés son $\tau_1(\theta_1, \theta_2) = \theta_1$ y $\tau_2(\theta_1, \theta_2) = \theta_2$.

Definición 4.2.9 (Stimatore di pendenza e di intercetta)**.** En las condiciones del problema supervisado nombrado anteriormente, los estimadores T_{τ_1} y T_{τ_2} se denominan respectivamente **estimador de la pendiente** y **estimador de intersección**, y el probelema se resuelve mediante el **método de los mínimos cuadrados** como sigue

$$T_{\tau_1}, T_{\tau_2} = \arg\min_{\theta_1,\theta_2} \frac{1}{2}\sum_{i=1}^{n}(y_i - (\theta_2 + \theta_1 x_i))^2$$

Ejercicio 4.2.1. Demostrar en las condiciones anteriores que los estimadores de mínimos cuadrados vienen dados por las siguientes fórmulas

$$T_{\tau_1} = \theta_1^* = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{covarianza empírica}}{\text{varianza empírica}} \quad , \quad T_{\tau_2} = \theta_2^* = \bar{y} - \theta_1^*\bar{x}$$

donde $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ y $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ son las medias muestrales de las variables x e y. En este caso, T_{τ_1} es la media de la pendiente y T_{τ_2} es la desviación media de la línea con respecto a la inte

Teorema 4.2.2 (Teorema de Gauss-Markov)**.** *Sea $D = \{(x_i, y_i)\}_{i=1}^n$ un conjunto de datos generado según el modelo supervisado de regresión lineal simple, es decir, existe una relación lineal entre ambas variables dada por*

$$y_i = \theta_2 + \theta_1 x_i + \varepsilon_i \quad i = 1, 2, \ldots, n$$

donde los errores ε_i son variables aleatorias independientes con media cero y varianza σ^2. Entonces, bajo estas condiciones, el teorema de Gauss-Markov establece que

los estimadores de mínimos cuadrados T_{τ_1} y T_{τ_2} son los ***mejores estimadores lineales insesgados*** *(BLUE, best linear unbiased estimator) de los parámetros θ_1 y θ_2 respectivamente.*

Observación. Si los errores ε_i siguen una distribución normal, el teorema de Gauss-Markov sigue siendo válido, y además, se tiene que los estimadores de mínimos cuadrados son los mejores estimadores de varianza mínima insesgados (MVUB) de los parámetros θ_1 y θ_2.

Observación. Puede haber otros estimadores no distorsionados que tengan menor error cuadrático medio, pero el teorema de Gauss-Markov solo se refiere a los estimadores lineales no distorsionados.

A continuación demostraremos que los estimadores de mínimos cuadrados son *estimadores no distorsionados* de los parámetros θ_1 y θ_2, pero antes debemos tener en cuenta que $\mathbb{E} = \mathbb{E}_{y\sim P(y|x)}$ es la esperanza respecto a la distribución condicional $P(y \mid x)$, por lo que x se considera fijo en este caso, ya que lo que queremos es estimar es el valor de la variable dependiente y en función de la variable independiente x. Notado este detalle, calculamos la esperanza de ambos estimadores

$$E[T_{\tau_1}] = E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\mathbb{E}\left[(y_i-\bar{y})\right]}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\left(\mathbb{E}[y_i]-\mathbb{E}[\bar{y}]\right)}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$
$$\overset{(*)}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\left((\theta_2+\theta_1 x_i)-(\theta_2+\theta_1\bar{x})\right)}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\theta_1(x_i-\bar{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \theta_1$$
$$E[T_{\tau_2}] = E[\bar{y}-T_{\tau_1}\bar{x}] = E[\bar{y}] - E[T_{\tau_1}]\bar{x} = \theta_2+\theta_1\bar{x}-\theta_1\bar{x} = \theta_2$$

donde en $(*)$ hemos usado el siguiente desarrollo

$$\mathbb{E}_{y\sim P(y|x)}[y_i] = \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}[\theta_2+\theta_1 x_i+\varepsilon_i] = \theta_2+\theta_1 x_i+\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2)}[\varepsilon_i] = \theta_2+\theta_1 x_i \Longrightarrow$$
$$\Longrightarrow \mathbb{E}_{y\sim P(y|x)}[\bar{y}] = \mathbb{E}_{y\sim P(y|x)}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{y\sim P(y|x)}[y_i] = \frac{1}{n}\sum_{i=1}^{n}(\theta_2+\theta_1 x_i) = \theta_2+\theta_1\bar{x}$$

Por tanto, hemos demostrado que ambos estimadores son no distorsionados, es decir, $E[T_{\tau_1}] = \theta_1$ y $E[T_{\tau_2}] = \theta_2$.

Finalmente, vamos a demostrar que ambos estimadores son consistentes, es decir, que convergen casi seguro a los parámetros reales cuando el tamaño de la muestra

n tiende a infinito. Para ello, primero veamos este desarrollo

$$T_{\tau_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})y_i - \frac{\bar{y}}{n}\cancelto{0}{\sum_{i=1}^{n}(x_i-\bar{x})}}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(\theta_1 x_i+\theta_2+\varepsilon_i)}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} =$$

$$= \frac{\theta_1\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})x_i + \theta_2\frac{1}{n}\cancelto{0}{\sum_{i=1}^{n}(x_i-\bar{x})} + \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} \overset{(*)}{=} \frac{\theta_1\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} =$$

$$= \theta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \theta_1 + \frac{1}{S_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})\varepsilon_i$$

donde hemos usado en $(*)$ que

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})x_i - \bar{x}\cancelto{0}{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})x_i$$

Notar que $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ es la varianza muestral de la variable independiente x, y que $S_x^2 > 0$, pues si no fuera así, todos los x_i serían iguales y no tendría sentido hacer una regresión lineal (tendríamos una línea vertical).

Sabemos también que por la ley de los grandes números, $S_x^2 \xrightarrow{c.s.} Var(x) = \sigma_x^2$ cuando $n \to \infty$, donde $\sigma_x^2 > 0$ ya que la varianza de x es positiva (lo justificamos igual que antes). Para ver más claro esta convergencia casi segura dividimos la varianza muestral en dos términos y vemos que ambos convergen casi seguro

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2$$

donde por la ley de los grandes números tenemos que

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i \xrightarrow{c.s.} E[x]$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2 \xrightarrow{c.s.} E[x^2]$$

donde para la segunda expresión hemos usado que si X_i son v.a. i.i.d., entonces X_i^2 también son v.a. i.i.d., y por tanto podemos aplicar la ley de los grandes números a la muestra $\{X_i^2\}_{i=1}^n$. Tenemos finalmente que

$$S_x^2 \xrightarrow{c.s.} E[x^2] - (E[x])^2 = Var(x) = \sigma_x^2 \quad \text{cuando } n \to \infty$$

Por otro lado, definimos la siguiente v.a.

$$\tilde{\varepsilon}_i = (x_i - \bar{x})\varepsilon_i \sim \mathcal{N}(0, (x_i - \bar{x})^2\sigma^2) \quad i = 1, 2, \ldots, n$$

donde viendo que se cumplen las condiciones de la ley de los grandes números, es decir,

- $\mathbb{E}[\tilde{\varepsilon}_i^2] = (x_i - \bar{x})^2 E[\varepsilon_i^2] = (x_i - \bar{x})^2\sigma^2$, que es finito ya que los x_i son fijos, y donde hemos obtenido el segundo multiplicando así

$$E[\varepsilon_i^2] = Var(\varepsilon_i) + (E[\varepsilon_i])^2 = \sigma^2 + 0 = \sigma^2$$

- Las v.a. $\tilde{\varepsilon}_i$ son independientes, ya que las ε_i lo son y los x_i son fijos.
- Las varianzas de $\tilde{\varepsilon}_i$ son acotadas, ya que $(x_i - \bar{x})^2\sigma^2 \leqslant C$ para cierto $C > 0$, pues los x_i no crecen hasta infinito (supesto del problema).

entonces podemos aplicar dicha ley y tenemos que

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i \xrightarrow{c.s.} E[\tilde{\varepsilon}_i] = 0 \quad \text{cuando } n \to \infty$$

Por tanto, usando estos dos resultados en el desarrollo anterior, tenemos que

$$T_{\tau_1} = \theta_1 + \frac{1}{S_x^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i \xrightarrow{c.s.} \theta_1 \quad \text{cuando } n \to \infty$$

El desarrollo para T_{τ_2} es análogo.

4.3. Likelihood/Verosimiglianza

En este apartado introduciremos el concepto de *likelihood* (verosimilitud) y veremos cómo se usa para estimar parámetros en modelos estadísticos.

Definición 4.3.10. Dado un conjunto de datos observados $D = \{(x_i, y_i)\}_{i=1}^n$ y un modelo estadístico parametrizado por $\theta \in \Theta$, la **función de verosimilitud** (likelihood/Verosimiglianza) se define como la probabilidad conjunta de los datos observados dado el parámetro θ

$$L(\theta; D) = P_\theta(D) = P_\theta\left(y \mid x\right)$$

donde P_θ es la distribución de probabilidad del modelo estadístico para el parámetro θ. De hecho definimos también la **log-verosimilitud** como

$$\ell(\theta; D) = \frac{1}{n}\log L(\theta; D) = \frac{1}{n}\log P_\theta(D)$$

en otras ocasiones será más conveniente usar $\ell(\theta; D) = \log L(\theta; D)$, pero los parámetros estimados serán los mismos.

Desarrollemos un poco más la función de log-verosimilitud en el caso de que las observaciones sean independientes, ya que es el caso más común, y tomando el caso de regresión lineal simple anterior. Tenemos, por tanto, que $P(y \mid x) = \prod_{i=1}^{n} P_\theta(y_i \mid x_i)$, por lo que la función de log-verosimilitud queda como

$$\begin{aligned}\ell(\theta; D) = \frac{1}{n}\log L(\theta; D) = \frac{1}{n}\log\left(\prod_{i=1}^{n} P_\theta(y_i \mid x_i)\right) = \frac{1}{n}\sum_{i=1}^{n}\log P_\theta(y_i \mid x_i) = \\ = \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_i - \theta_1 x_i - \theta_2)}{2\sigma^2}\right)\right) = \\ = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{\sigma^2}\underbrace{\frac{1}{2n}\sum_{i=1}^{n}(y_i - \theta_1 x_i - \theta_2)^2}_{MSE}\end{aligned}$$

por lo que maximizar la función de log-verosimilitud es equivalente a minimizar la suma de los errores al cuadrado (MSE), puesto que la segunda parte de la expresión está restando y depende de los parámetros θ_1 y θ_2, mientras que la primera parte es constante.

Definición 4.3.11. Dado un conjunto de datos observados $D = \{(x_i, y_i)\}_{i=1}^n$ y un modelo estadístico parametrizado por $\theta \in \Theta$, el **estimador de máxima verosimilitud** (maximum likelihood estimator/Stimatore di massima verosimiglianza) para el parámetro θ se define como

$$T_\theta = \arg\max_{\theta\in\Theta} \ell(\theta; D) = \arg\max_{\theta\in\Theta} P_\theta(D)$$

donde $\ell(\theta; D)$ es la función de log-verosimilitud. Se suele notar como $T_\theta = \theta_{MLE}$.

Ejemplo. Supongamos que tenemos $X_1, X_2, \ldots, X_n \sim^{i.i.d.} \mathcal{N}(\mu, \sigma^2)$, es decir, una muestra aleatoria de tamaño n de una distribución normal con media μ y varianza σ^2. Nuestro objetivo es estimar los parámetros μ y σ^2 usando el método de máxima verosimilitud.

En primer lugar, trabajaremos con el vector de variables aleatorias

$$\mathbf{X} = (X_1, X_2, \ldots, X_n) \sim \mathcal{N}(\underline{\mu}, \sigma^2 I_n)$$

donde $\underline{\mu} = (\mu, \ldots, \mu)$ tal que $|\underline{\mu}| = n$ y I_n es la matriz identidad de tamaño n, de forma que la matriz de covarianzas Σ tiene diagonal con σ^2 y el resto cero (las v.a.s son independientes por lo que su covarianza es 0). La función de densidad conjunta de $\mathbf{X}$ es

$$\begin{aligned}P_{\mu,\sigma^2}(X) = f(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2}(\mathbf{X} - \underline{\mu})\Sigma^{-1}(\mathbf{X} - \underline{\mu})^T\right) = \\ = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right)\end{aligned}$$

donde hemos usado que $\Sigma^{-1} = \frac{1}{\sigma^2} I_n$. Por tanto, la función de log-verosimilitud es

$$\ell(\mu, \sigma^2; X) = \frac{1}{n} \log P_{\mu,\sigma^2}(X) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2n\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 \Longrightarrow$$
$$\Longrightarrow (T_\mu, T_{\sigma^2}) = \arg\max_{\mu,\sigma^2} \ell(\mu, \sigma^2; X)$$

Teniendo ya definido el problema de maximización, procedemos a calcular los estimadores. Primero calculamos el estimador de la media μ

$$\frac{\partial \ell(\mu, \sigma^2; X)}{\partial \mu} = -\frac{1}{2n\sigma^2} \cdot 2\sum_{i=1}^{n}(X_i - \mu)\cdot(-1) = \frac{1}{n\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0 \Longleftrightarrow$$
$$\Longleftrightarrow \sum_{i=1}^{n}(X_i - \mu) = 0 \Longleftrightarrow n\mu = \sum_{i=1}^{n} X_i \Longrightarrow \mu_{MLE} = T_\mu = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ (no distorsiona}$$

Ahora calculamos el estimador de la varianza σ^2

$$\frac{\partial \ell(\mu, \sigma^2; X)}{\partial \sigma^2} = -\frac{1}{2}\cdot\frac{1}{\sigma^2} - (-1)\frac{1}{2n(\sigma^2)^2}\sum_{i=1}^{n}(X_i - \mu)^2 = 0 \Longleftrightarrow$$
$$\Longleftrightarrow -n\sigma^2 + \sum_{i=1}^{n}(X_i - \mu)^2 = 0 \Longrightarrow \sigma^2_{MLE} = T_{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - T_\mu)^2 \text{ (distorsionado)}$$

Por tanto, los estimadores de máxima verosimilitud para la media y la varianza son

$$\mu_{MLE} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad \sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Ejemplo. Supongamos una v.a. X que sigue una distribución binomial $X \sim Bin(n, p)$, donde n es conocido y p es desconocido. Nuestro objetivo es estimar el parámetro p usando el método de máxima verosimilitud.

Sabemos que su función de masa de probabilidad es

$$P_{p,n}(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \ldots, n$$

Por tanto, la función de log-verosimilitud es

$$\ell(p; n, x) = \log P_{p,n}(x) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p) \Longrightarrow$$
$$\Longrightarrow T_p = \arg\max_{p} \ell(p; X)$$

Procedemos a calcular el estimador de máxima verosimilitud para p

$$\frac{d\ell(p; n, X)}{dp} = \frac{X}{p} - \frac{n-X}{1-p} = 0 \Longleftrightarrow X(1-p) - (n-X)p = 0 \Longleftrightarrow$$
$$\Longleftrightarrow X - Xp - np + Xp = 0 \Longleftrightarrow np = X \Longrightarrow T_p = \frac{X}{n} \text{ (no distorsionado)}$$

donde sabemos que es no distorsionado usando que $E[X] = np$.

Sin embargo, existe un estimador alternativo para p, para el cual el error cuadrático medio es menor que el del estimador de máxima verosimilitud, para ciertos valores de p. Este estimador es

$$\tilde{T_p} = \frac{X+1}{n+2} \text{ (distorsionado)}$$

Calculemos su error cuadrático medio para compararlo con el del estimador de máxima verosimilitud. Primero calculamos su sesgo

$$E[\tilde{T_p}] = E\left[\frac{X+1}{n+2}\right] = \frac{1}{n+2}E[X] + \frac{1}{n+2} = \frac{1}{n+2}(np+1) = \frac{np+1}{n+2}$$
$$Bias_p(\tilde{T_p}) = E[\tilde{T_p}] - p = \frac{np+1}{n+2} - p = \frac{np+1-p(n+2)}{n+2} = \frac{1-2p}{n+2}$$

Ahora calculamos su varianza

$$Var_p(\tilde{T_p}) = Var_p\left(\frac{X+1}{n+2}\right) = \frac{1}{(n+2)^2}Var_p(X) = \frac{1}{(n+2)^2}(np(1-p)) = \frac{np(1-p)}{(n+2)^2}$$

Finalmente, calculamos su error cuadrático medio

$$MSE_p(\tilde{T_p}) = Var_p(\tilde{T_p}) + (Bias_p(\tilde{T_p}))^2 = \frac{np(1-p) + (1-2p)^2}{(n+2)^2}$$

Por otro lado, el error cuadrático medio del estimador de máxima verosimilitud es

$$MSE_p(T_p) = Var_p(T_p) + (Bias_p(T_p))^2 = Var_p(T_p) = \frac{p(1-p)}{n}$$

donde hemos usado que $Var[X] = np(1-p)$ y que el sesgo es nulo.

Comparando ambos errores cuadráticos medios, graficado en la Figura 4.1 para $n = 10$, vemos que existe un rango de valores de p (cercas a $1/2$) para los cuales el estimador alternativo $\tilde{T_p}$ tiene un error cuadrático medio menor que el del estimador de máxima verosimilitud T_p, a pesar de ser distorsionado. Esto ilustra que en algunos casos, un estimador distorsionado puede ser preferible si su varianza es suficientemente baja.

Observación. Aqui podemos ver un claro ejemplo donde el estimador de máxima verosimilitud no es el mejor estimador en términos de error cuadrático medio, ya que existe otro estimador distorsionado con menor MSE para ciertos valores del parámetro a estimar. La clave del estimador de máxima verosimilitud es que a la larga ($n \to \infty$) convergerá al **valor real del parámetro**, pero para muestras pequeñas puede no ser el mejor en términos de MSE.

Ejercicio 4.3.1. Supongamos que observamos n eventos de decadencia de una esfera radiactiva, registrando los tiempos de cada desintegración $(t_1, t_2, \ldots, t_n)$. Sabiendo que la desintegración de las partículas sigue una ley física exponencial, encuentra el

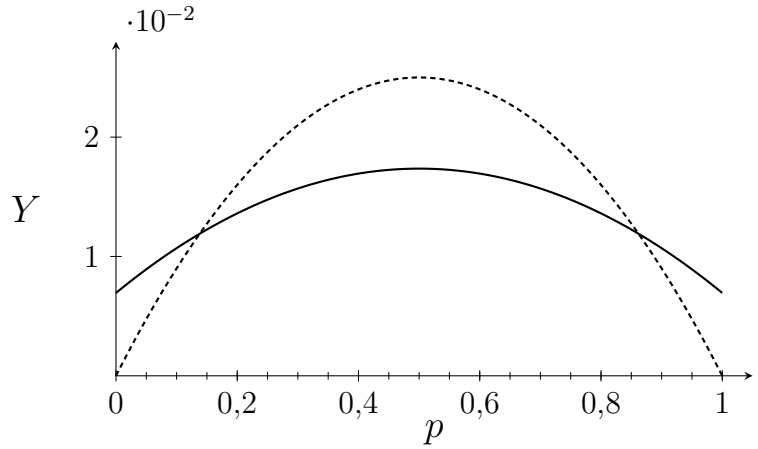

Figura 4.1: T_p (discontinua) vs. $\tilde{T}_p$ (continua) para $n = 10$

Estimador de Máxima Verosimilitud (MLE) para el parámetro τ (tiempo de decadencia o vida media).

Utilizando conceptos de física nuclear (que no serán relevantes) sabemos que las partículas radiactivas siguen una distribución exponencial, es decir, $t_i \sim exp\left(\frac{1}{\tau}\right)$, con función de densidad de probabilidad dada por

$$P(t_i)_\tau = \frac{1}{\tau} e^{-t_i/\tau} \quad t \geqslant 0$$

donde $\tau > 0$ es el tiempo de decadencia o vida media. Suponiendo observado el conjunto de datos $D = \{t_1, t_2, \ldots, t_n\}$, la función de log-verosimilitud es

$$\ell(\tau; D) = \frac{1}{n} \log L(\tau; D) = \frac{1}{n} \log \left(\prod_{i=1}^{n} P(t_i)_\tau \right) = \frac{1}{n} \log \left(\frac{1}{\tau^n} \exp \left(-\frac{1}{\tau} \sum_{i=1}^{n} t_i \right) \right) =$$

$$= -\log \tau - \frac{1}{n\tau} \sum_{i=1}^{n} t_i \Longrightarrow T_\tau = \tau_{MLE} = \arg \operatorname*{máx}_\tau \ell(\tau; D)$$

donde hemos usado que $L(\tau; D) = P_\tau(D) = \prod_{i=1}^{n} P(t_i)_\tau$, al ser las observaciones independientes. Procedemos a calcular el estimador de máxima verosimilitud para τ, derivando la función de log-verosimilitud e igualando a cero

$$\frac{d\ell(\tau; D)}{d\tau} = -\frac{1}{\tau} + \frac{1}{n\tau^2} \sum_{i=1}^{n} t_i = 0 \Longleftrightarrow -\tau + \frac{1}{n} \sum_{i=1}^{n} t_i = 0 \Longleftrightarrow \tau = \frac{1}{n} \sum_{i=1}^{n} t_i \Longrightarrow T_\tau = \bar{t}$$

Es importante notar que hemos hecho un abuso de notación al escribir t_i, tanto para la variable aleatoria como para el valor observado, pero es fácil de distinguir si sabemos lo que estamos haciendo en cada momento. Por ejemplo, en la función de densidad $P(t_i)_\tau$, t_i es la variable aleatoria, mientras que cuando se esta definiendo la función de log-verosimilitud

$$\ell(\tau; D) = \frac{1}{n} \log \left(\prod_{i=1}^{n} P(t_i)_\tau \right)$$

t_i es el valor observado.

Ejercicio 4.3.2. Sea X una v.a. discreta con función de masa de probabilidad

$$P(X=0) = P(X=1) = p, \quad P(X=2) = 1-2p, \quad p \in (0, 1/2)$$

estima el parámetro p a partir de una muestra aleatoria $x_1, x_2, \ldots, x_n$ usando el método de máxima verosimilitud.

RESUELTO EN LOS APUNTES.

Ejercicio 4.3.3. Sean $X_1, X_2, \ldots, X_n$ una muestra aleatoria de una v.a. con función de densidad de probabilidad

$$P_\theta(x) = (\theta+1)x^\theta \mathbb{1}_{[0,1]}, \quad x \in \mathbb{R}, \quad \theta \geqslant -1$$

estima el parámetro θ usando el método de máxima verosimilitud. Notar la importancia de tener $\theta \geqslant -1$ para que sea una función de densidad válida, pues en otro caso la función de densidad podría dar un valor negativo.

RESUELTO EN LOS APUNTES.

Ejercicio 4.3.4. Sean $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas (i.i.d.) que siguen una distribución uniforme en el intervalo $[-\theta, 2\theta]$, con $\theta > 0$, es decir,

$$X_1, \ldots, X_n \overset{iid}{\sim} U([-\theta, 2\theta])$$

Se nos pide que:

1) Estimar el parámetro θ mediante el método de máxima verosimilitud ($\hat{\theta}_{MLE}$).

2) Calcular la esperanza del estimador obtenido, $E[\hat{\theta}_{MLE}]$.

. RESUELTO EN LOS APUNTES.

4.4. Entropia

Definición 4.4.12 (Entropia di Shannon)**.** La **entropía de Shannon** de una variable aleatoria discreta X con función de masa de probabilidad $P(X = x_i) = p_i$ es una medida de la incertidumbre asociada a la variable aleatoria, y se define como

$$S(P) = -k \sum_i p_i \log p_i$$

donde $k > 0$ es una constante positiva.
De forma análoga, para una variable aleatoria continua X con función de densidad de probabilidad $p(x)$, la **entropía de Shannon** se define como

$$S(P) = -k \int p(x) \log p(x) dx$$

donde $k > 0$ es una constante positiva.

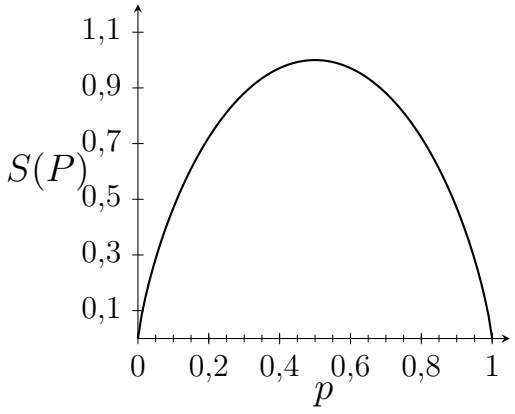

Figura 4.2: Entropía de Shannon para v.a. $X \sim Ber(p)$

Ejemplo. Para una variable aleatoria discreta $X = \{0, 1\} \sim Ber(p)$, la entropía de Shannon es

$$S(P) \propto p \log p + (1 - p) \log(1 - p)$$

donde se ha usado que $P(X = 1) = p$ y $P(X = 0) = 1 - p$. Notar que la entropía es máxima cuando $p = 1/2$, es decir, cuando la incertidumbre es máxima. Podemos ver una representación en la Figura 4.2.

Ahora vamos a ver como maximizar la entropía de Shannon bajo ciertas restricciones, usando el *método de los multiplicadores de Lagrange*, pero, ¿por qué se hace esto? Porque el **Principio de Máxima Entropía** establece que la mejor distribución que podemos elegir es aquella que maximiza la incertidumbre (Entropía de Shannon) lo que es equivalente a encontrar la distribución de probabilidad más ïmparcial.º üniforme"posible, sujeta a lo que ya sabemos (los vínculos o restricciones). Esto es útil en situaciones donde queremos evitar introducir sesgos innecesarios en nuestras inferencias estadísticas.

Dado por tanto la definición de entropía de Shannon para una v.a. continua, vamos a intentar maximizarla bajo las siguientes restricciones

- Para obtener una función de densidad de probabilidad válida, la integral de la función de densidad debe ser igual a 1, es decir,

$$\int p(x)dx = 1$$

a estas restricciones podemos añadir otras adicionales, que se conozcan del problema en cuestión, como por ejemplo

- La media de la distribución es igual a un valor conocido μ, es decir,

$$\int xp(x)dx = \mu$$

- La varianza de la distribución es igual a un valor conocido σ^2, es decir,

$$\int (x - \mu)^2 p(x)dx = \sigma^2$$

pero para el desarrollo que haremos a continuación, solo usaremos la primera restricción, que es la que conocemos en general siempre, y desarrollaremos lo que se llama **Optimización con restricciones** (extremizzazione vincolata) usando el método de los multiplicadores de Lagrange. Veámoslo

$$\widehat{S}_\lambda(P) = S + \lambda\left(\int p(x)dx - 1\right) = -k\int p(x)\log p(x)dx + \lambda\left(\int p(x)dx - 1\right)$$

COPIAR DESARROLLO Y EXPLICACIÓN.

Finalmente, el resultado de este desarrollo es que la distribución que maximiza la entropía de Shannon bajo la restricción de que la integral de la función de densidad es igual a 1, es la distribución uniforme en el intervalo considerado. Esto tiene sentido intuitivamente, ya que la distribución uniforme es la que tiene la mayor incertidumbre posible dentro de un intervalo dado, ya que todos los resultados son igualmente probables.
Tenemos por tanto que para un problema con soporte en el intervalo $[a, b]$ sin conocer nada más, la distribución que maximiza la entropía de Shannon es

$$p(x) = \frac{1}{b-a} \quad x \in [a, b]$$

es decir, la distribución uniforme en $[a, b]$, $X \sim \mathcal{U}(a, b)$.

Ahora vamos a ver otro ejemplo, en este caso añadiendo las restricciones de media y varianza conocidas, y obtener por tanto la distribución que maximiza la entropía de Shannon bajo estas restricciones.

COPIAR DESARROLLO Y EXPLICACIÓN.

Finalmente, el resultado de este desarrollo es que la distribución que maximiza la entropía de Shannon bajo las restricciones de media y varianza conocidas, es la distribución normal con dicha media y varianza. Esto tiene sentido intuitivamente, ya que la distribución normal es la que tiene la mayor incertidumbre posible dado un valor fijo de media y varianza.
Tenemos por tanto que para un problema con media μ y varianza σ^2 conocidas, la distribución que maximiza la entropía de Shannon es

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad x \in \mathbb{R}$$

es decir, la distribución normal, $X \sim \mathcal{N}(\mu, \sigma^2)$.

Definición 4.4.13 (Cross entropy)**.** Dadas dos distribuciones de probabilidad p y q sobre el mismo espacio de eventos, la **entropía cruzada** entre las dos distribuciones calcula el error que cometo al usar la distribución p para aproximar la distribución real q. Se define como

$$H(q, p) = -\sum_i q(x_i)\log p(x_i)$$

en el caso discreto, y como

$$H(q,p) = -\int_X q(x)\log p(x)dx$$

en el caso continuo.

En nuestro caso concreto, cuando trabajamos con un problema de datos supervisados, es decir, $D = \{(x_i, y_i)\}_{i=1}^n$, y tenemos una distribución de probabilidad P con la que queremos aproximar la distribución real de los datos Q, usando la función de log-verosimilitud tenemos que

$$\ell(\theta; D) = \frac{1}{n}\sum_i \log P_\theta(y_i \mid x_i) \approx_{n\to\infty} \int q(x,y)\log P_\theta(y \mid x)dxdy =$$

$$= \int q(x)\left(\underbrace{\int q(y \mid x)\log P_\theta(y \mid x)dy}_{CE[Q,P]}\right)dx \Longrightarrow \ell(\theta; D) \approx -\int q(x)CE[Q,P]dx$$

donde hemos usado que los datos (x_i, y_i) son i.i.d. según la distribución real Q, y por tanto, $q(x,y) = q(x)q(y \mid x)$. Tenemos entonces que maximizar la función de log-verosimilitud es equivalente a minimizar la entropía cruzada entre la distribución real Q y la distribución del modelo P_θ.

4.5. Distribución Gaussiana multivariante

Definición 4.5.14 (Difeomorfismo)**.** Sea $X \subseteq R^n$ un abierto. Una función $g : X \to Y \subseteq R^n$ es un **difeomorfismo** si es biyectiva, diferenciable y su inversa $g^{-1} : Y \to X$ también es diferenciable.

Proposición 4.5.3 (Transformazioni di densità)**.** *Sea $X \subseteq R^n$ un abierto y P una distribución de probabilidad en $(X, \mathcal{B}_X)$ con función de densidad $p(x)$. Sea $g : X \to Y \subseteq R^n$ un difeomorfismo de clase C^1, entonces la distribución de probabilidad $P \circ g^{-1}$ en Y tiene función de densidad*

$$q(y) = p(g^{-1}(y))\,|\det J_{g^{-1}}(y)| \quad \forall y \in Y$$

donde $J_{g^{-1}}(y)$ es la matriz jacobiana de la función g^{-1} evaluada en y.

Demostración. □

Teorema 4.5.4 (Vettore di gaussiane standard)**.** *Sean $X_1, X_2, \ldots, X_n$ v.a.s independientes e idénticamente distribuidas según una distribución normal estándar, $X_i \sim \mathcal{N}(0,1)$. Dado $B \in \mathbb{R}^{n\times n}$ una matriz invertible y $m \in \mathbb{R}^n$ un vector, entonces la v.a. $Y = BX + m$ tiene función de densidad*

$$p_Y(y) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det(C)|}}\exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right)$$

donde $C := BB^T$ y $X = (X_1, X_2, \ldots, X_n)^T$.

Demostración. □

Observación. En el teorema anterior, la matriz C es simétrica y definida positiva.

Definición 4.5.15 (Matriz definida positiva)**.** Una matriz simétrica $C \in \mathbb{R}^{n\times n}$ se dice **definida positiva** si $\forall v \in \mathbb{R}^n, v \neq 0$ se cumple que

$$v^T C v > 0$$

Proposición 4.5.5 (Caracterización de matrices definidas positivas)**.** *Una matriz es definida positiva si y solo si todos sus autovalores son estrictamente positivos ($\lambda_i > 0$ para todo i)*

Definición 4.5.16 (Gaussiana multivariata)**.** Dada una v.a. $Y = (Y_1, Y_2, \ldots, Y_n)^T$ se dice **variable aleatoria gaussiana multivariante** si $\exists m \in \mathbb{R}^n, \exists C \in \mathbb{R}^{n\times n}$ simétrica y definida positiva tal que distribución de Y tiene función de densidad

$$p_Y(y) = \frac{1}{(2\pi)^{n/2}\sqrt{|\det(C)|}} \exp\left(-\frac{1}{2}(x-m)^T C^{-1}(x-m)\right)$$

y se escribe $Y \sim \mathcal{N}(m, C)$.

Observación. Si $Y \sim \mathcal{N}_n(m, C)$, entonces

$$\mathbb{E}[Y_i] = m \quad \text{y} \quad Cov(Y_i, Y_j) = C_{ij} \quad i, j = 1, 2, \ldots, n$$

Proposición 4.5.6. *Sea $Y \sim \mathcal{N}_n(m, C)$ una v.a. gaussiana multivariante con C matriz simétrica y definida positiva, entonces $\exists X \sim \mathcal{N}_n(0, I_n)$ tal que*

$$Y = BX + m$$

donde $C = BB^T$, es decir, $B = \sqrt{C}$.

4.6. Distribuciones relacionadas con la normal

Distribución Gamma

Previamente al estudio de la distribución Gamma, vamos a estudiar la función Gamma, que es la función que da nombre a la distribución. Esta es:

$$\begin{array}{rccc} \Gamma: & \mathbb{R}^+ & \longrightarrow & \mathbb{R}^+ \\ & x & \longmapsto & \displaystyle\int_0^\infty t^{x-1}e^{-t}\,dt \end{array}$$

Algunas propiedades son:

1. $\Gamma(1) = 1$.
2. $\Gamma(x+1) = x\Gamma(x), \qquad \forall x \in \mathbb{R}^+$.
3. $\Gamma(n) = (n-1)!, \qquad \forall n \in \mathbb{N}$.

Definición 4.6.17. Dada X una variable aleatoria y $\alpha, r > 0$, se dice que X sigue una **distribución gamma** con parámetros α y r si $\Gamma_{\alpha,r}$ es una medidad de probabilidad en $(\mathbb{R}^+, \mathcal{B}(\mathbb{R}^+))$ con función de densidad

$$p_{\alpha,r}(x) = \frac{\alpha^r}{\Gamma(r)} x^{r-1} e^{-\alpha x} \quad x > 0$$

donde $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}dt$ es la función gamma, y se denota como $X \sim \Gamma(\alpha, r)$.

Distribución χ^2

Definición 4.6.18. Dada X una variable aleatoria, se dice que X sigue una **distribución chi-cuadrado** si X tiene distribución gamma con parámetros $\alpha = 1/2$ y $r = 1/2$, es decir, $\chi^2 = \Gamma_{\frac{1}{2},\frac{1}{2}}$. La función de densidad es

$$p_{1/2,1/2}(x) = \frac{1}{\sqrt{2}\Gamma(1/2)} x^{-1/2} e^{-x/2} \quad x > 0$$

y se denota como $X \sim \chi^2$.

Definición 4.6.19. Dada X una variable aleatoria, se dice que X sigue una **distribución chi-cuadrado con n grados de libertad** si X tiene distribución gamma con parámetros $\alpha = 1/2$ y $r = n/2$, es decir, $\chi^2_n = \Gamma_{1/2,n/2}$. La función de densidad es

$$p_{1/2,n/2}(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} \quad x > 0$$

y se denota como $X \sim \chi^2_n$.

Proposición 4.6.7. *Sea $X \sim \mathcal{N}(0,1)$ una variable aleatoria con distribución normal estándar, entonces la variable aleatoria $Y = X^2$ sigue una distribución chi-cuadrado, es decir, $Y \sim \chi^2$.*

Demostración. □

Teorema 4.6.8. *Sean $X_1, X_2, \ldots, X_n$ v.a.s independientes e idénticamente distribuidas según una distribución normal estándar, $X_i \sim \mathcal{N}(0,1)$. Entonces la variable aleatoria*

$$\sum_{i=1}^{n} X_i^2 \sim \Gamma\left(\frac{1}{2}, \frac{n}{2}\right) = \chi^2_n$$

Demostración. □

Distribución Beta

Definición 4.6.20 (Función Beta)**.** La **función Beta de Euler** es una función definida como:

$$\begin{array}{rccl} \beta: & \mathbb{R}^+ \times \mathbb{R}^+ & \longrightarrow & \mathbb{R}^+ \\ & (p,q) & \longmapsto & \displaystyle\int_0^1 t^{p-1}(1-t)^{q-1}\, dt \end{array}$$

Algunas propiedades son:

1. Es simétrica. Es decir, $\forall p, q \in \mathbb{R}^+$, se cumple que $\beta(p,q) = \beta(q,p)$.
2. $\beta(p,q) = \dfrac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}$.

Definición 4.6.21 (Distribución Beta)**.** Una variable aleatoria continua X sigue una **distribución Beta** con parámetros $p, q \in \mathbb{R}^+$ en $]0,1[$, si su función de densidad es:

$$f_X(x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1} \quad x \in [0,1]$$

Lo denotaremos como $X \sim \beta(p,q)$.

Proposición 4.6.9. *Dados $r, s > 0$ y $X \sim \Gamma(\alpha, r)$, $Y \sim \Gamma(\alpha, s)$ dos variables aleatorias independientes, entonces*

$$X + Y \sim \Gamma(\alpha, r+s) \quad y \quad \frac{X}{X+Y} \sim \beta(r,s)$$

Demostración. □

Distribución t-Student

Definición 4.6.22. Una variable aleatoria X sigue una **distribución t-Student con n grados de libertad** sobre $]0,+\infty[$ si su función de densidad es

$$\tau_n(x) = \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \left(\sqrt{n} B\left(\frac{1}{2}, \frac{n}{2}\right)\right)^{-1} \quad x \in \mathbb{R}^+$$

y se denota como $X \sim t_n$.

Corolario 4.6.9.1. *Sean las variables aleatorias $X, Y_1, \ldots, Y_n$ i.i.d. con distribución común $\mathcal{N}(0,1)$, entonces*

$$Z = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}} \sim t_n$$

Demostración. □

Teorema 4.6.10. *Sea el modelo estadístico de producto gaussiano n-ario*

$$(\mathbb{R}^n, B(\mathbb{R}^n), N_{m,v}^{\otimes n}; m \in \mathbb{R}, v > 0)$$

y sean las v.a.s. $(X_i)_{i=1}^n$ como las proyecciones, es decir, son i.i.d. y $X_i \sim \mathcal{N}(m,v)$. Definimos las siguientes variables aleatorias

$$M = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad V^* = \frac{1}{n}\sum_{i=1}^{n-1}(X_i - M)^2$$

las siguientes afirmaciones son ciertas

- *M y V^* son independientes.*
- *$M \sim N(m, \frac{v}{n})$ y $\frac{n-1}{v} V^* \sim \chi^2_{n-1}$.*
- *$\frac{\sqrt{n}(M-m)}{\sqrt{V^*}} \sim t_{n-1}$.*

4.7. Teorema de estimación de Cramér-Rao-Fisher

Definición 4.7.23 (Modello statistico regolare)**.** Sea un modelo estadístico $(X, \mathcal{B}_X, P_\theta; \theta \in \Theta)$ con función de densidad $p_\theta(x)$ y $\Theta \subseteq \mathbb{R}$, se dice que es un **modelo estadístico regular** si se cumplen las siguientes condiciones:

(1) Θ es un abierto en $\mathbb{R}$.

(2) $p_\theta(x)$ es estrictamente positiva y continuamente diferenciable respecto a θ. En particular, existe la función llamada **función de puntuación** (score function)

$$U_\theta(x) = \frac{d}{d\theta} \log p_\theta(x)$$

(3) p permite intercambiar la diferenciación respecto a θ y la integración sobre x, es decir,

$$\frac{d}{d\theta} \int_X p_\theta(x)dx = \int_X \frac{d}{d\theta} p_\theta(x)dx$$

(como siempre para X discreto la integral se sustituye por una suma).

(4) Para cada $\theta \in \Theta$, la varianza $I(\theta) := V_\theta(U_\theta)$ existe y no es nula.

La función $I : \theta \to I(\theta)$ se llama **información de Fisher**.

Proposición 4.7.11. *Sea un modelo estadístico regular $(X, \mathcal{F}, P_\theta; \theta \in \Theta)$, entonces se cumple que*

$$\mathbb{E}_\theta[U_\theta] = 0 \quad y \quad I(\theta) = -\mathbb{E}_\theta\left[\frac{d}{d\theta} U_\theta\right] = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2} \log p_\theta\right]$$

Demostración. Primero vamos a demostrar que $\mathbb{E}_\theta[U_\theta] = 0$. Por definición de esperanza tenemos que

$$\begin{aligned}\mathbb{E}_\theta[U_\theta] = \mathbb{E}\left[\frac{d}{d\theta} \log p_\theta(X)\right] = \int_X \frac{d}{d\theta} \log p_\theta(x) p_\theta(x)dx = \int_X \frac{p'_\theta(x)}{p_\theta(x)} p_\theta(x)dx = \\ = \int_X p'_\theta(x)dx \overset{(*)}{=} \frac{d}{d\theta} \int_X p_\theta(x)dx = \frac{d}{d\theta} 1 = 0\end{aligned}$$

donde en $(*)$ hemos usado la propiedad (3) de modelo estadístico regular y en la última igualdad hemos usado que la integral de una función de densidad es 1.
Veámos ahora la segunda igualdad

$$\begin{aligned}0 = \frac{d}{d\theta} \int_X p_\theta(x) \frac{d}{d\theta} \log p_\theta(x)dx \overset{(3)}{=} \int_X \frac{d}{d\theta}\left(p_\theta(x) \frac{d}{d\theta} \log p_\theta(x)\right) dx = \\ \overset{(*)}{=} \int_X \frac{d}{d\theta} p_\theta(x) \frac{d}{d\theta} \log p_\theta(x)dx + \int_X p_\theta(x) \frac{d^2}{d\theta^2} \log p_\theta(x)dx = \\ = \underbrace{\int_X U_\theta(x) p_\theta(x) U_\theta(x)dx}_{I(\theta)} + \int_X p_\theta(x) \frac{d^2}{d\theta^2} \log p_\theta(x)dx = 0 \Longrightarrow \\ \overset{(**)}{\Longrightarrow} I(\theta) = -\mathbb{E}_\theta\left[\frac{d^2}{d\theta^2} \log p_\theta\right] = -\mathbb{E}_\theta\left[\frac{d}{d\theta} U_\theta\right]\end{aligned}$$

donde en $(*)$ hemos usado la regla del producto para derivadas y en $(**)$ hemos usado que por el resultado anterior ahora sabemos que $I(\theta) = \mathbb{E}[U_\theta^2]$. □

Proposición 4.7.12 (Aditividad de la información de Fisher)**.** *Sea un modelo estadístico regular $(X, \mathcal{B}_X, P_\theta; \theta \in \Theta)$ con información de Fisher $I(\theta)$. Entonces el modelo estadístico de producto n-ario*

$$(X^n, \mathcal{F}^{\otimes n}, P_\theta^{\otimes n}; \theta \in \Theta)$$

es un modelo estadístico regular con información de Fisher

$$I_n(\theta) = nI(\theta)$$

donde además la función de puntuación del modelo estadístico de producto n-ario es

$$U_\theta^{(n)}(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} U_\theta(x_i)$$

Ejemplo. 1. Sea $X \sim \mathcal{N}(\mu, \sigma^2)$ una variable aleatoria, donde conocemos σ^2 y queremos estimar μ, es decir, tenemos el modelo estadístico

$$(\mathbb{R}, B(\mathbb{R}), N_{\mu,\sigma^2}; \mu \in \mathbb{R})$$

El modelo estadístico es regular, y para poder obtener la información de Fisher tenemos que calcular la función de puntuación

$$p_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \implies \log p_\mu(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} \implies$$
$$\implies U_\mu(x) = \frac{d}{d\mu}\log p_\mu(x) = \frac{x-\mu}{\sigma^2}$$

Calculada la función de puntuación, obtengamos la información de Fisher

$$I(\mu) = \mathbb{E}[U_\mu^2] = \mathbb{E}\left[\frac{(X-\mu)^2}{\sigma^4}\right] = \frac{1}{\sigma^4}\mathbb{E}[(X-\mu)^2] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$$

Para el caso del modelo estadístico de producto n-ario, tenemos que la información de Fisher es

$$I_n(\mu) = nI(\mu) = \frac{n}{\sigma^2} = \frac{1}{\sigma^2/n} = \frac{1}{Var(\bar{X})} = \frac{1}{\sigma_{\bar{X}}^2}$$

2. Sea $X \sim Ber(p)$ una variable aleatoria, donde queremos estimar p, es decir, tenemos el modelo estadístico

$$(\{0,1\}, \mathcal{P}(\{0,1\}), Ber(p); p \in (0,1))$$

El modelo estadístico es regular, y sabemos que $P(X=1) = p$ y $P(X=0) = 1-p$. Por tanto, para poder obtener la información de Fisher tenemos que calcular la función de puntuación

$$U_p(X=1) = \frac{d}{dp}\log P(X=1) = \frac{d}{dp}\log p = \frac{1}{p}$$
$$U_p(X=0) = \frac{d}{dp}\log P(X=0) = \frac{d}{dp}\log(1-p) = -\frac{1}{1-p}$$

entonces la información de Fisher es

$$I(p) = \mathbb{E}[U_p^2] = P(X=1)U_p(X=1)^2 + P(X=0)U_p(X=0)^2 = p\left(\frac{1}{p}\right)^2 + (1-p)\left(-\frac{1}{1-p}\right)^2$$
$$= \frac{1}{p} + \frac{1}{1-p} = \frac{1}{p(1-p)}$$

donde como al tratarse de una variable aleatoria discreta, la esperanza se calcula como suma ponderada de las probabilidades.
Si ahora calculamos la entropía de Shannon para esta distribución, tenemos que

$$S(P) = -\sum_{x\in\{0,1\}} P(X=x)\log P(X=x) = -p\log p - (1-p)\log(1-p)$$

podemos observar que la información de Fisher y la entropía de Shannon están relacionadas, ya que se comportan de forma contraria, como bien podemos ver en la Figura 4.3.
En este sentido, son inversas en su significado, ya que una mayor entropía implica una mayor incertidumbre en la distribución, lo que a su vez implica una menor información disponible para estimar el parámetro p.

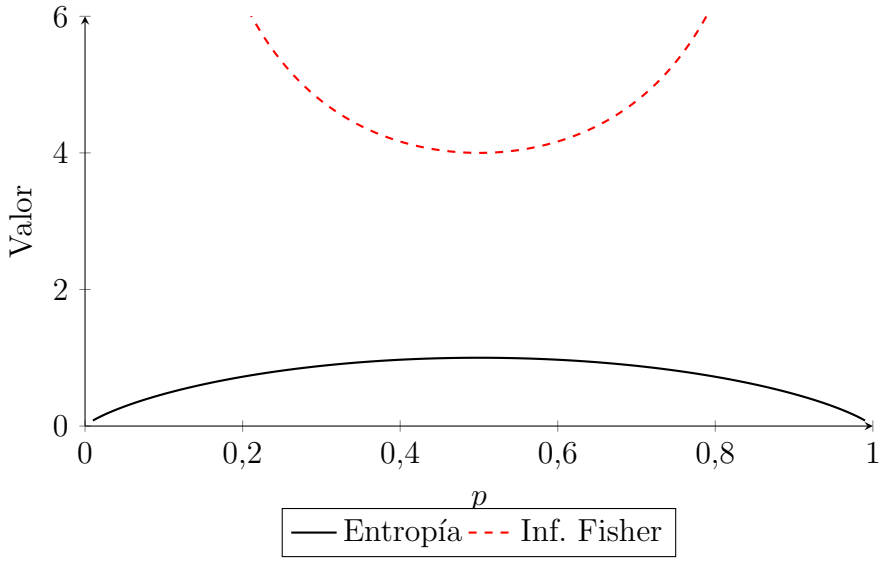

Figura 4.3: Entropía e Información de Fisher ($X \sim Ber(p)$).

3. Sea $X \sim P(\lambda)$ una variable aleatoria, donde queremos estimar λ, es decir, tenemos el modelo estadístico
NO ESTA TERMINADO.

Teorema 4.7.13 (Teorema Fisher/Cramer/Rao)**.** *Sea un modelo estadístico regular* $(X, \mathcal{B}_X, P_\theta; \theta \in \Theta)$ *y sea* T_τ *un estimador insesgado de una característica* $\tau : \Theta \to \mathbb{R}$, *es decir,* $\mathbb{E}_\theta[T(X)] = \tau(\theta)$ *para todo* $\theta \in \Theta$. *Entonces se tiene*

1) Se cumple el ***límite de Cramer-Rao***

$$Var(T_\tau) \geqslant \frac{(\tau'(\theta))^2}{I(\theta)} \quad \forall \theta \in \Theta$$

En particular, si $\tau(\theta) = \theta$*, entonces*

$$Var(T_\tau) \geqslant \frac{1}{I(\theta)} \quad \forall \theta \in \Theta$$

2) Si $Var(T_\tau) = \dfrac{(\tau'(\theta))^2}{I(\theta)}$ *para todo* $\theta \in \Theta$*, entonces*

$$T_\tau(x) - \tau(\theta) = \frac{\tau'(\theta)}{I(\theta)} U_\theta(x)$$

que equivale a

$$p_\theta(x) = h(x) \exp\left(a(\theta) T_\tau(x) - b(\theta)\right)$$

que se llama ***familia exponencial respecto al estimador*** T_τ*, donde tenemos*

$$a(\theta) = \int \frac{I(\theta)}{I'(\theta)} d\theta$$
$$b(\theta) = \log \int_X h(x) \exp\left(a(\theta) T_\tau(x)\right) dx$$
$$h(x) \text{ función medible oportuna}$$

Demostración. □

Dado este teorema, tenemos que si podemos representar la función de densidad del modelo estadístico en la forma de familia exponencial respecto al estimador T_τ, entonces el estimador es eficiente, es decir, alcanza el límite de Cramer-Rao, y tenemos por tanto (identificando a, b, h) lo siguiente

$$\tau(\theta) = \frac{b'(\theta)}{a'(\theta)} = \mathbb{E}[T_\tau]$$
$$I(\theta) = a'(\theta)\tau'(\theta)$$
$$Var(T_\tau) = \frac{(\tau'(\theta))^2}{I(\theta)}$$

y en general para n extracciones distintas tenemos

$$p_\theta^n(x) = \left(\prod_{i=1}^n h(x_i)\right) \exp\left(a(\theta) \sum_{i=1}^n T_\tau(x_i) - b(\theta)\right)$$

4.8. Inferenza Bayesiana

Antes de introducir el ejemplo que dará sentido a las siguientes definiciones, recordemos el teorema de Bayes en su forma más simple, es decir, el caso discreto.

Teorema 4.8.14 (Regla de Bayes)**.** *Sea* $(\Omega, \mathcal{A}, P)$ *un espacio de probabilidad y sea* $\{A_n\}_{n\in\mathbb{N}} \subset \mathcal{A}$ *una partición de* Ω *con* $P(A_n) > 0 \ \ \forall n \in \mathbb{N}$*. Entonces,* $\forall B \in \mathcal{A}$*:*

$$P(A_j \mid B) = \frac{P(B \mid A_j) P(A_j)}{\sum_{i=1}^{\infty} P(B \mid A_i) P(A_i)} \qquad j \in \mathbb{N}$$

Ejemplo (Seguro de coche)**.** Un cliente se acerca a una compañía de seguros para obtener una cita para su nuevo seguro de coche. Tiene un estilo de conducción particular que provoca, con probabilidad $\theta \in [0,1]$, al menos un reclamo por año. Obviamente, el agente de seguros no conoce θ, pero tiene estadísticas sobre las frecuencias de reclamaciones para la población total. Estos le proporcionan una evaluación de riesgo a priori, es decir, una medida de probabilidad π_0 en $\Theta = [0,1]$, llamada medida de probabilidad a priori. Supongamos por simplicidad que π es 'suave', en el sentido de que tiene una densidad de Lebesgue α, lo que significa que

$$\pi_0(\theta) = 1 \quad \forall \theta \in [0,1]$$

es decir, todas las frecuencias de reclamaciones son igualmente probables a priori. Podemos ver una representación de la función de densidad a priori en la Figura 4.4. Ahora, el agente de seguros presume (subjetivamente) que la probabilidad de que el cliente tenga k años de reclamaciones en n años está dada por

$$P_n(X = k) = \int_0^1 \pi_0(\theta) B_{n,\theta}(\{k\}) d\theta$$

donde $B_{n,\theta}(\{k\}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$ es la función de probabilidad de una distribución binomial con parámetros n y θ. Desarrollemos

$$\begin{aligned} P_n(X = k) &= \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k} d\theta = \binom{n}{k}\int_0^1 \theta^k(1-\theta)^{n-k} d\theta = \\ &= \binom{n}{k}\beta(k+1, n-k+1) = \binom{n}{k}\frac{\Gamma(k+1)\Gamma(n-k+1)}{\Gamma(n+2)} = \\ &= \frac{n!}{k!(n-k)!}\frac{k!(n-k)!}{(n+1)!} = \frac{1}{n+1} \end{aligned}$$

lo que tiene total sentido, pues el agente de seguros no tiene información sobre el cliente, por lo que todas las frecuencias de reclamaciones son igualmente probables, de modo que sabiendo que k puede tomar valores entre 0 y n, la probabilidad de cada uno es $\frac{1}{n+1}$. Podemos ver una representación gráfica de esta distribución en la Figura 4.5.

Ahora pasados estos n años, el agente de seguros observa que el cliente ha hecho $x \in \{0, \dots, n\}$ reclamaciones. ¿Cómo debe actualizar su evaluación de riesgo a priori π_0? Utilizando el teorema de Bayes, la respuesta es la siguiente: la nueva evaluación de riesgo, llamada evaluación de riesgo a posteriori, es la función de densidad π_X que reemplaza a π_0 y está dada por

$$P(\theta \mid X = x, n) = \pi_X(\theta) = \frac{\pi_0(\theta)B_{n,\theta}(\{x\})}{\int_0^1 \pi_0(p)B_{n,p}(\{x\})dp} = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{P_n(X=x)} = (n+1)\binom{n}{x}\theta^x(1-\theta)^{n-x}$$

Es decir, la evaluación de riesgo a posteriori es la distribución condicional de θ dado el evento 'el cliente ha hecho x reclamaciones en n años'. De este modo la nueva función de probabilidad es la siguiente

$$\begin{aligned} P(X = l) &= \int_0^1 \pi_X(\theta) B_{n,\theta}(\{l\}) d\theta = (n+1)\binom{n}{x}\binom{n}{l}\int_0^1 \theta^{x+l}(1-\theta)^{2n-(x+l)} d\theta = \\ &= (n+1)\binom{n}{x}\binom{n}{l}\beta(x+l+1, 2n-(x+l)+1) \end{aligned}$$

Veamos ahora la esperanza en ambos casos, es decir, usando ambas funciones de densidad π_0 y π_X

$$\begin{aligned}
\mathbb{E}_{\pi_0}[\theta] &= \int_0^1 \theta\pi_0(\theta)d\theta = \int_0^1 \theta d\theta = \frac{1}{2} \\
\mathbb{E}_{\pi_X}[\theta] &= \int_0^1 \theta\pi_X(\theta)d\theta = (n+1)\binom{n}{k}\int_0^1 \theta^{x+1}(1-\theta)^{n-x}d\theta = \\
&= (n+1)\binom{n}{k}\beta(x+2, n-x+1) = (n+1)\binom{n}{x}\frac{\Gamma(x+2)\Gamma(n-x+1)}{\Gamma(n+3)} = \\
&= (n+1)\binom{n}{x}\frac{(x+1)!(n-x)!}{(n+2)!} = \frac{x+1}{n+2}
\end{aligned}$$

Dadas estas dos situaciones (conociendo o no el número de reclamaciones hechas por el cliente), podemos observar que la esperanza de la probabilidad de que el cliente haga una reclamación en un año es $\frac{1}{2}$ si no conocemos el número de reclamaciones hechas, mientras que si conocemos que ha hecho x reclamaciones en n años, la esperanza se actualiza a $\frac{x+1}{n+2}$, que es una media ponderada entre la esperanza a priori y la frecuencia observada x/n. Esto tiene sentido, ya que si el cliente ha hecho muchas reclamaciones, la probabilidad de que haga una en el próximo año debería ser mayor que si no hubiera hecho ninguna. Por tanto, la inferencia bayesiana nos permite actualizar nuestras creencias sobre la probabilidad de que un evento ocurra en función de la evidencia observada.

De hecho, si definimos dos estimadores de la siguiente manera

$$\begin{aligned}
T_\theta(X) &= \frac{X}{n} \quad \text{(estimador no sesgado)} \\
S_\theta(X) &= \frac{X+1}{n+2} \quad \text{(estimador sesgado)}
\end{aligned}$$

viendo la gráfica del error cuadrático medio de ambos estimadores en la Figura 4.6, podemos observar que el estimador sesgado S_θ tiene un error cuadrático medio menor que el estimador no sesgado T_θ para ciertos valores de θ. Esto ilustra cómo la incorporación de información a priori puede mejorar la precisión de las estimaciones en ciertos contextos.

Definición 4.8.24 (Distribución a priori y a posteriori)**.** Sea un modelo estadístico paramétrico estándar $(X, \mathcal{F}, P_\theta; \theta \in \Theta)$ con función de probabilidad $p_\theta : X \to [0, \infty[$ para cada $\theta \in \Theta$. Entonces, para toda densidad de probabilidad π_0 en $(\Theta, \mathcal{F}')$, que se llama **densidad a priori**, y conjunto de datos observados $\{x_1, \ldots, x_n\} \subseteq X$, la **densidad a posteriori** π_X está dada por

$$\pi_X(\theta) = \frac{\pi_0(\theta)\prod_{i=1}^n p_\theta(x_i)}{\int_\Theta \pi_0(p)\prod_{i=1}^n p_p(x_i)dp} = \frac{\pi_0(\theta)L(\theta; D)}{\int_\Theta \pi_0(p)L(p; D)dp} = \frac{\pi_0(\theta)\exp\left(\frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i)\right)}{\int_\Theta \pi_0(p)\exp\left(\frac{1}{n}\sum_{i=1}^n \log p_p(x_i)\right)dp}$$

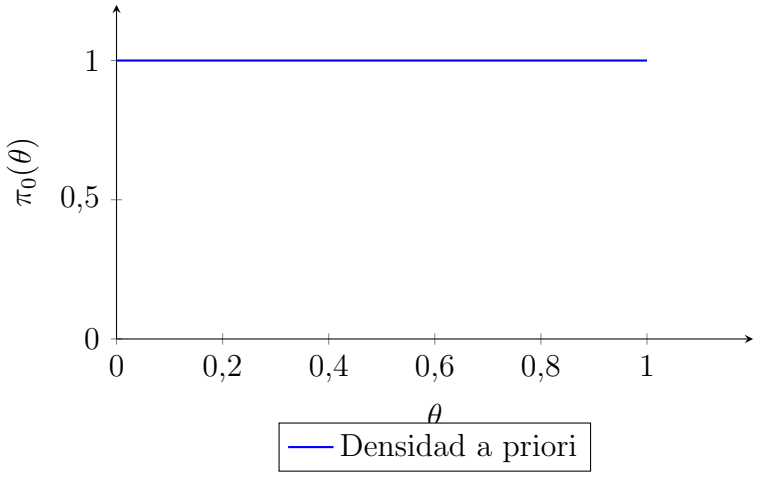

Figura 4.4: Función de densidad a priori $\pi_0(\theta)$ para el seguro de coche.

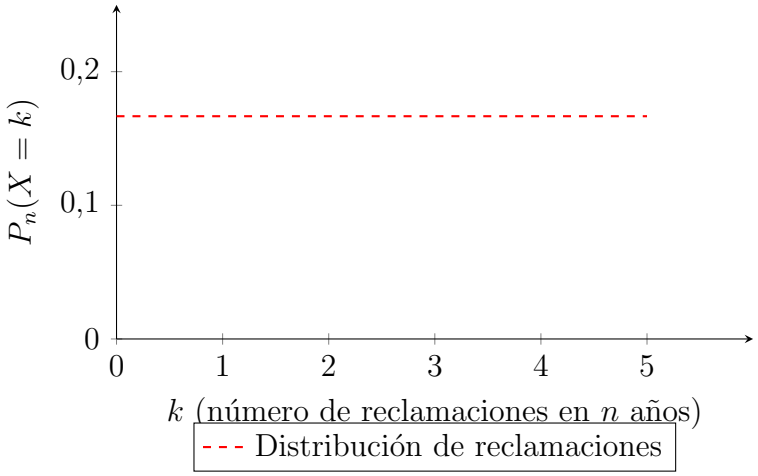

Figura 4.5: Distribución de reclamaciones $P_n(X = k)$ para el seguro de coche con $n = 5$.

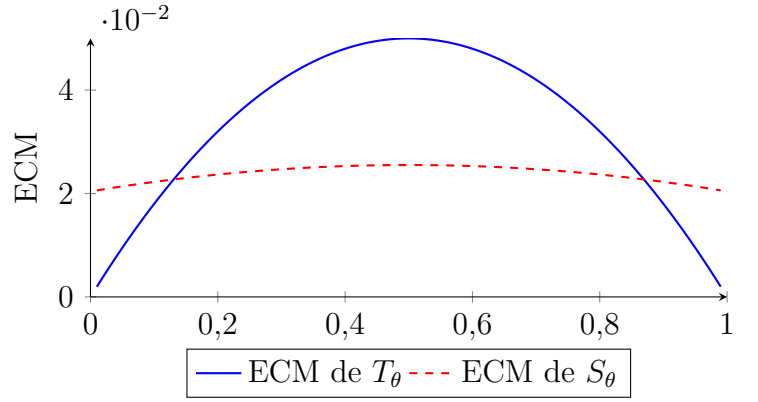

Figura 4.6: Error cuadrático medio (ECM) de los estimadores T_0 y T_1 para el seguro de coche con $n = 5$.

Observación. Es importante recalcar que $P_\theta(X_i) = p_\theta(x_i)$ es la función de probabilidad de la variable aleatoria X_i evaluada en el dato observado x_i.

Definición 4.8.25 (MSE Bayesiano)**.** Sea $(X; \mathcal{F}; P_\theta; \theta \in \Theta)$ un modelo estadístico paramétrico estándar con función de verosimilitud $p > 0$ que es medible conjuntamente en ambas variables, y sea π una distribución a priori en Θ con densidad π_0. Además, sea $\tau : X \times \Theta \to \mathbb{R}$ una característica real medible con $E_{\pi_0}(\tau^2) < \infty$. Dado un estimador $T : X \to \mathbb{R}$ de τ su **error cuadrático medio (MSE) bayesiano** cuando hay una observación $n = 1$ está dado por

$$MSE(T) = \int_\Theta \int_X (T(x) - \tau(\theta))^2 p_\theta(x) dx \pi_0(\theta) d\theta = \frac{1}{C} \int_\Theta \int_X (T(x) - \tau(\theta))^2 \pi_X(\theta) dx d\theta$$

Para el caso en el que haya $n > 1$ observaciones, el MSE bayesiano se define como

$$\begin{aligned} MSE(T(x_1, \ldots, x_n)) &= \int_\Theta \int_{X^n} \prod_{i=1}^{n} p_\theta(x_i)(T(x_1, \ldots, x_n) - \tau(\theta))^2 dx_1 \cdot \ldots \cdot dx_n \pi_0(\theta) d\theta = \\ &= \frac{1}{C} \int_\Theta \int_{X^n} (T(D) - \tau(\theta))^2 \pi_D(\theta) dx_1 \cdot \ldots \cdot dx_n d\theta \end{aligned}$$

donde en ambos casos C es una constante de normalización.

Definición 4.8.26 (Stimatore di Bayes Ottimale)**.** En las condiciones anteriores, un estimador $\hat{T} : X \to \mathbb{R}$ se llama **estimador de Bayes óptimo** (OBE) si minimiza el MSE bayesiano, es decir,

$$\hat{T} = \arg\min_T MSE_{\pi_0, p_\theta(x)}(T)$$

esto equivale a decir que

$$\hat{T}(x) = E_{\pi_X}[\tau(\theta)] = \int_\Theta \tau(\theta) \pi_X(\theta) d\theta$$

donde π_X es la distribución a posteriori dada la muestra observada $X = x$. En muchas ocasiones el estimador de Bayes óptimo se representará como θ_{BAYES}, donde θ es el parámetro que se está estimando.

Observación. La diferencia clave con el estimador de máxima verosimilitud (MLE) es que el estimador de Bayes óptimo incorpora información a priori sobre el parámetro a través de la distribución a priori π_0. Esto puede ser especialmente útil cuando se dispone de información previa o cuando los datos son limitados, ya que permite ajustar las estimaciones en función de dicha información adicional, mientras en el caso del estimador MLE solo se basa en los datos observados.

Ejemplo.

Definición 4.8.27 (Función de Dirac)**.** La **función de Dirac** $\delta_a : \mathbb{R} \to \mathbb{R}$ en el punto $a \in \mathbb{R}$ se define como

$$\delta(x - a) = \delta_a(x) = \begin{cases} +\infty & x = a \\ 0 & x \neq a \end{cases}$$

y cumple la propiedad de que

$$\int_{-\infty}^{\infty} f(x)\delta_a(x)dx = f(a)$$

para toda función continua $f : \mathbb{R} \to \mathbb{R}$.

Proposición 4.8.15. *Una propiedad importante de la función de Dirac es la siguiente:*

$$\delta(g(x)) = \sum_i \frac{\delta(x - x_i)}{|g'(x_i)|}$$

donde los x_i son las raíces de la función $g(x)$.

5. Ejercicios

5.1. Sección 1

5.2. Sección 2

Ejercicio 5.2.1 (Libro 2.10)**.**

5.3. Sección 3

Ejercicio 5.3.1 (Libro 3.6)**.** Sea $(\Omega, \mathcal{F}, P)$ un espacio de probabilidad y $A, B, C \in \mathcal{F}$. Demuestre directamente (sin usar el Corolario (3.20) ni el Teorema (3.24)):

(a) Si A y B son independientes, entonces también lo son A y B^c.

(b) Si A, B, C son independientes, entonces también lo son $A \cup B$ y C.

Haremos el ejercicio por partes, veámoslo.

(a) Sabemos que A y B son independientes, por lo que se cumple que $P(A \cap B) = P(A) \cdot P(B)$, veámos que tenemos

$$P(A \cap B^c) = P(A) - P(A \cap B) = P(A) - P(A) \cdot P(B) = P(A) \cdot (1 - P(B)) = P(A) \cdot P(B^c)$$

donde hemos usado la independencia de A y B.

(b) Sabemos que A, B, C son independientes, por lo que se cumple que $P(A \cap B \cap C) = P(A)P(B)P(C)$, veámos que tenemos

$$P((A \cup B) \cap C) = P((A \setminus B \cup B) \cap C) = P((A \setminus B \cap C) \cup (B \cap C)) = P(A \setminus B \cap C) + P(B \cap$$
$$= P(A \cap B^c \cap C) + P(B \cap C) = P(A)P(B^c)P(C) + P(B)P(C) =$$
$$= P(A)(1 - P(B))P(C) + P(B)P(C) = P(C)[P(A)P(B^c) + P(B)] \overset{(*)}{=} P(C)P(A \cup B)$$

donde en $(*)$ hemos usado lo siguiente

$$P(A)P(B^c) = P(A \cap B^c) = P(A \setminus B)$$
$$P(A)P(B^c) + P(B) = P(A \setminus B) + P(B) = P((A \setminus B) \cup B) = P(A \cup B)$$

Ejercicio 5.3.2 (Libro 3.10)**.**

Ejercicio 5.3.3 (Libro 3.17 Thinning of a Poisson distribution)**.** Supongamos que el número de huevos N que pone un insecto sigue una distribución de Poisson con parámetro $\lambda > 0$:

$$N \sim \text{Poisson}(\lambda)$$

De cada huevo nace una larva con probabilidad $p \in (0,1)$, de forma independiente para cada huevo. Encuentre la distribución del número total de larvas j.

Para la resolución definiremos la v.a. N que representa el número de huevos que pone el insecto, de manera que tenemos

$$N : \Omega \to \mathbb{N}_0, \quad N(w) = n \text{ si el insecto pone } n \text{ huevos en el experimento } w \in \Omega$$

es importante destacar que el espacio muestral Ω es el conjunto de todos los posibles experimentos que podemos realizar, y la variable aleatoria N asigna a cada experimento el número de huevos que pone el insecto en dicho experimento. Además, sabemos que N sigue una distribución de Poisson con parámetro $\lambda > 0$, es decir, tenemos que

$$P(N = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \in \mathbb{N}_0$$

Por otra parte, definimos los eventos A_i como el evento en el que nace una larva del huevo i-ésimo, de modo que tenemos la variable aleatoria S que representa el número de larvas que nacen en un experimento $w \in \Omega$, es decir,

$$S : \Omega \to \mathbb{N}_0, \quad S(w) = \sum_{i=1}^{N(w)} \mathbb{1}_{A_i}(w) \ \forall w \in \Omega$$

donde $\mathbb{1}_{A_i}$ es la función indicadora del evento A_i, es decir,

$$\mathbb{1}_{A_i}(w) = \begin{cases} 1, & \text{si nace una larva del huevo } i\text{-ésimo en el experimento } w \in \Omega \\ 0, & \text{si no nace una larva del huevo } i\text{-ésimo en el experimento } w \in \Omega \end{cases}$$

donde tenemos que $P(A_i) = p$ para todo $i \in \mathbb{N}$, y los eventos A_i son independientes entre sí.

Nuestro objetivo es encontrar la distribución de la variable aleatoria S, es decir, encontrar $P(S = j)$ para todo $j \in \mathbb{N}_0$. Para ello primero fijaremos $n \in \mathbb{N}_0$ como el número de huevos que pone el insecto en un experimento $w \in \Omega$, es decir, $N(w) = n$, y definiremos la siguiente medida de probabilidad para S_n, que representa el número de larvas que nacen cuando el insecto pone n huevos, que sigue claramente una distribución binomial con parámetros n y p

$$P(S_n = j) = \binom{n}{j} p^j (1-p)^{n-j}, \quad \forall j \leqslant n, \ j \in \mathbb{N}_0$$

Ahora sí, aplicando la fórmula de la probabilidad total, tenemos que

$$P(S = j) = \sum_{k=j}^{\infty} P(S = j | N = k) P(N = k) \overset{(*)}{=} \sum_{k=j}^{\infty} P(S_k = j) P(N = k) = \sum_{k=j}^{\infty} \binom{k}{j} p^j (1-p)^{k-j} e^{-\lambda} \frac{\lambda^k}{k!}$$

$$= e^{-\lambda}\frac{p^j}{j!} \sum_{k=j}^{\infty} \frac{(1-p)^{k-j}\lambda^k}{(k-j)!} = e^{-\lambda}\frac{p^j}{j!}\lambda^j \sum_{m=0}^{\infty} \frac{((1-p)\lambda)^m}{m!} \overset{(**)}{=} e^{-\lambda}\frac{(p\lambda)^j}{j!} e^{(1-p)\lambda} = e^{-p\lambda}\frac{(p\lambda)^j}{j!}$$

donde en $(**)$ hemos usado la serie de Taylor de la función exponencial, y en $(*)$ hemos de usado el siguiente desarrollo

$$P(S=j \mid N=k) = P(\sum_{i=1}^{k} \mathbb{1}_{A_i} = j \mid N=k) = \frac{P(\{\sum_{i=1}^{k} \mathbb{1}_{A_i} = j\} \cap \{N=k\})}{P(N=k)} \overset{(***)}{=} P(\sum_{i=1}^{k} \mathbb{1}_{A_i} = j) = P$$

donde se ha usado en $(***)$ que el número de huevos N es independiente del nacimiento de las larvas en cada huevo, es decir, $A_1, \ldots, A_k$. Finalmente se tiene que la distribución del número total de larvas j es una distribución de Poisson con parámetro $p\lambda$, es decir,

$$S \sim \text{Poisson}(p\lambda) \quad \text{donde } P(S=j) = e^{-p\lambda}\frac{(p\lambda)^j}{j!}, \quad j \in \mathbb{N}_0$$

Ejercicio 5.3.4 (Libro 3.20)**.**

5.4. Sección 4

Ejercicio 5.4.1 (Libro 4.15 b))**.**

5.5. Sección 5

Ejercicio 5.5.1 (Libro 5.4)**.** Sea $(X_i)_{i\geq 1}$ una sucesión v.a. de Bernoulli con $0 < p < 1$. Demuestra que para todo $p < a < 1$ se cumple que:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right) \leq e^{-nh(a;p)}, \tag{5.1}$$

donde

$$h(a;p) = a\log\frac{a}{p} + (1-a)\log\frac{1-a}{1-p}. \tag{5.2}$$

Pista: Demuestra primero que para todo $s \geq 0$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq a\right) \leq e^{-nas}\left(\mathbb{E}[e^{sX_1}]\right)^n. \tag{5.3}$$

Ejercicio 5.5.2 (Libro 5.10)**.**

5.6. Sección 8

Ejercicio 5.6.1 (Libro 8.1)**.**

Ejercicio 5.6.2 (Libro 8.6)**.**

5.7. Sección 9

Ejercicio 5.7.1 (Libro 9.2)**.**

Ejercicio 5.7.2 (Libro 9.17)**.**

5.8. PDF extra