Inferencia Estadística

Apuntes

Andrés Herrera Poyatos
Universidad de Granada
andreshp9@gmail.com

Índice

1. Familias de distribuciones

1.1. Distribuciones discretas

1.2. Distribuciones continuas

1.2.1. Distribución uniforme

La distribución uniforme asigna una credibilidad uniforme a todos los puntos de un intervalo $[a, b]$. Esto es, su función de densidad viene dada por

$$f(x|a,b) = \begin{cases} \frac{1}{b-a} \text{ si } x \in [a,b], \\ 0 \text{ en otro caso.} \end{cases}$$

Claramente tenemos que $\int_{-\infty}^{\infty} f(x|a,b)dx = 1$. Además, podemos calcular fácilmente sus momentos como sigue (y, por tanto, también su varianza)

$$E[X^j] = \int_a^b \frac{x^j}{b-a}dx = \frac{b^{j+1} - a^{j+1}}{(b-a)(j+1)},$$

$$Var(X) = E[X^2] - E[X]^2 = \frac{a^2 + ab + b^2}{3} - \frac{(a+b^2)}{4} = \frac{(b-a)^2}{12}.$$

1.2.2. Distribución normal

La distribución normal, también llamada distribución gaussiana, es la distribución más importante de la estadística. Esto se debe a sus numerosas aplicaciones en análisis de poblaciones y al teorema central del límite.

Definición 1.1. Sean $\mu \in \mathbb{R}$ y $\sigma^2 > 0$. Definimos la distribución $N(x|\mu, \sigma^2)$ como la distribución que tiene función de densidad

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, x \in \mathbb{R}.$$

La distribución normal está bien definida como consecuencia del siguiente lema.

Lema 1.1. *Sean $\mu \in \mathbb{R}$ y $\sigma > 0$. Tenemos que $\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}dx = \sqrt{2\pi}\sigma$.*

Demostración. En primer lugar, vamos a calcular la integral para $\mu = 0$ y $\sigma = 1$. La demostración consiste en reducir el problema en calcular una integral en dos variables. Para ello, elevamos al cuadrado y obtenemos

$$\left(\int_{-\infty}^{\infty} e^{-x^2/2}dx\right)^2 = \left(\int_{-\infty}^{\infty} e^{-t^2/2}dt\right)\left(\int_{-\infty}^{\infty} e^{-s^2/2}ds\right) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(t^2+s^2)/2}dtds.$$

Resolvemos esta última integral mediante un cambio a polares

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(t^2+s^2)/2}dtds = \int_{-\pi}^{\pi}\left(\int_0^{\infty} \rho e^{-\rho^2/2}d\rho\right)d\theta = 2\pi\int_0^{\infty} \rho e^{-\rho^2/2}d\rho = 2\pi.$$

Por último, utilizamos el cambio de variable $y = (x-\mu)/\sigma$ para obtener

$$\int_{-\infty}^{\infty} e^{-(x-\mu)^2/(2\sigma^2)}dx = \int_{-\infty}^{\infty} \sigma e^{-y^2/2}dy = \sqrt{2\pi}\sigma. \qquad \square$$

Nótese que si $X \sim N(x|\mu,\sigma^2)$, entonces $Y = (X-\mu)/\sigma$ sigue una distribución $N(x|0,1)$.

Proposición 1.2. *La función característica de la distribución $N(x|\mu,\sigma^2)$ viene dada por $\varphi_X(t) = e^{it\mu - t^2\sigma^2/2}$.*

Demostración. En primer lugar, tenemos que

$$\varphi_X(t) = E[e^{itX}] = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{itx-(x-\mu)^2/(2\sigma^2)}dx = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-((x-\mu)^2-2itx\sigma^2)/(2\sigma^2)}dx.$$

Completamos cuadrados como sigue

$$(x-\mu)^2 - 2itx\sigma^2 = (x-(it\sigma^2+\mu))^2 + t^2\sigma^4 - 2it\sigma^2\mu.$$

Esto sugiere utilizar el cambio de variable $g(y) = y + it\sigma^2$. Obtenemos

$$\begin{aligned}\sqrt{2\pi}\sigma\varphi_X(t) = \int_{-\infty}^{\infty} e^{-((x-\mu)^2-2itx\sigma^2)/(2\sigma^2)} &= e^{it\mu-t^2\sigma^2/2}\int_{-\infty}^{\infty} e^{-((x-(it\sigma^2+\mu))^2)/(2\sigma^2)}dx \\ &= e^{it\mu-t^2\sigma^2/2}\int_{-\infty}^{\infty} e^{-(y-\mu)^2)/(2\sigma^2)}dy = \sqrt{2\pi}\sigma e^{it\mu-t^2\sigma^2/2},\end{aligned}$$

como se quería. Nótese que a pesar de ser una integral de contorno compleja el cambio de variable es válido. En efecto, el cambio de variable es afín y la función a integrar es entera. Por tanto, utilizando el camino cerrado $g([0,\infty]) + [\infty, 0]$ se puede probar que el cambio es válido. □

Análogamente se puede probar el siguiente resultado.

Proposición 1.3. *La función generatriz de momentos de la distribución $N(x|\mu,\sigma^2)$ viene dada por $\varphi_X(t) = e^{t\mu-t^2\sigma^2/2}$.*

Corolario 1.4. *Los momentos de la distribución $N(x|\mu,\sigma^2)$ verifican la ecuación recurrente*

$$E[X^k] = -(k-1)\sigma^2E[X^{k-2}] + (\mu - t\sigma^2)E[X^{k-1}], \quad k \geq 2.$$

Demostración. Sabemos que $E[X^k] = \varphi_X^{(k)}(t)$. Tenemos $\varphi_X^{(1)}(t) = (\mu - t\sigma^2)\varphi_X(t)$. Consecuentemente,

$$\varphi_X^{(2)}(t) = -\sigma^2\varphi_X(t) + (\mu - t\sigma^2)\varphi_X^{(1)}(t).$$

Por inducción se extiende el resultado fácilmente para $k \geq 2$. □

Corolario 1.5. *Si $X \sim N(x|\mu,\sigma^2)$, entonces $E[X] = \mu$ y $E[X^2] = \sigma^2 + \mu^2$. Consecuentemente, $Var(X) = \sigma^2$. Como consecuencia de este resultado al parámetro μ se le llama media y al parámetro σ^2 varianza.*

Podemos utilizar los dos corolarios anteriores para calcular los momentos de la distribución normal resolviendo una ecuación recurrente de segundo orden. Evidentemente, la fórmula obtenida será bastante larga. Sin embargo, esta ecuación se simplifica en el caso de los momentos centrados, como pone de manifiesto el siguiente resultado, que se puede demostrar fácilmente por inducción a partir del Corolario 1.4.

Corolario 1.6. *Si $X \sim N(x|0,\sigma^2)$, entonces*

$$E[X^k] = \begin{cases} 0 & \text{si } k \text{ es impar;} \\ (k-1)!!\sigma^k & \text{si } k \text{ es par;} \end{cases}$$

donde $n!!$ denota al doble factorial, definido como el producto de los números desde 1 hasta n con la misma paridad que n.

La Figura 1 muestra la función de densidad de una distribución normal. Podemos ver que la densidad se concentra en torno a la media. De hecho, $P(|X-\mu| \geq 2\sigma) \approx 0{,}046$. Es más, $P(|X-\mu| \geq 3\sigma) \approx 0{,}03$.

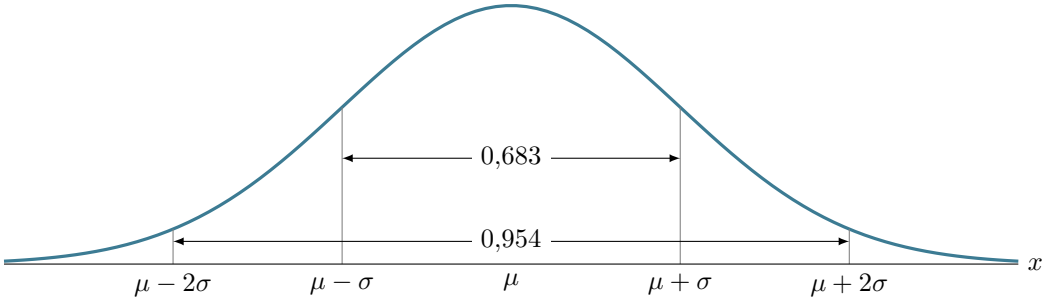

Figura 1: Función de densidad de una distribución normal.

Proposición 1.7. *Sean X_1 e Y_2 dos variables aleatorias independientes que siguen una distribución $N(x|\mu_1,\sigma_1^2)$ y $N(x|\mu_2,\sigma_2^2)$ respectivamente. Entonces $X+Y$ sigue una distribución $N(x|\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$.*
Demostración. Basta darse cuenta de que $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{t(\mu_1+\mu_2)-t^2(\sigma_1^2+\sigma_2^2)/2}$ es la función característica asociada a la distribución $N(x|\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$. Recordemos que la función característica determina de forma unívoca a la distribución. □

El recíproco del resultado anterior también es cierto.

Teorema 1.8 (Cramer)**.** *Sean X e Y dos variables aleatorias independientes. Si $X+Y$ es normal, entonces X e Y son normales.*

1.2.3. Distribución gamma

La famila de distribuciones gamma se encuentra definida sobre el intervalo $[0,\infty)$. En su definición entra en juego la famosa función gamma, de ahí su nombre.

Definición 1.2. Se define la función gamma como la aplicación $\Gamma : (0,\infty) \to (0,\infty)$ dada por

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Proposición 1.9. *La función gamma está bien definida.*
Demostración. Sea $\alpha > 0$. Tenemos que probar que $\int_0^\infty t^{\alpha-1}e^{-t}dt < \infty$. Tomando $b > 0$, escribimos

$$\int_0^\infty t^{\alpha-1}e^{-t}dt = \int_0^b t^{\alpha-1}e^{-t}dt + \int_b^\infty t^{\alpha-1}e^{-t}dt.$$

Sabemos que la función $t^{\alpha-1}$ tiene a t^α/α como primitiva y, por tanto, es integrable en $[0,b]$. Puesto que $t^{\alpha-1}e^{-t} \leq t^{\alpha-1}$, obtenemos que $t^{\alpha-1}e^{-t}$ es integrable en $[0,b]$. Por otro lado tenemos que

$$\lim_{t\to\infty} \frac{t^{\alpha-1}e^{-t}}{e^{-t/2}} = 0.$$

Consecuentemente, para cierto $b > 0$ se verifica $t^{\alpha-1}e^{-t} \leq e^{-t/2}$ para todo $t \geq b$. Puesto que $e^{-t/2}$ es integrable en $[b,\infty)$, deducimos que $t^{\alpha-1}e^{-t}$ también lo es, lo que termina la demostración. □

Proposición 1.10 (Propiedades de la función gamma)**.** *Sea $\alpha > 0$. Se verifica:*

a) $\Gamma(1) = 1$*;*

b) $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$*;*

c) $\Gamma(n + 1) = n!$ *para cualquier* $n \in \mathbb{N}$*;*

d) (Fórmula de reflexión de Euler) si $0 < \alpha < 1$*, entonces* $\Gamma(\alpha)\Gamma(1 - \alpha) = \frac{\pi}{\sin(\alpha\pi)}$*;*

e) $\Gamma(1/2) = \sqrt{\pi}$*;*

f) $\Gamma(\alpha) = \beta^\alpha \int_0^\infty t^{\alpha-1}e^{-\beta t}dt$ *para todo* $\beta > 0$*.*

Demostración.

a) Es fácil ver que $\int_0^\infty e^{-t}dt = 1$.

b) Integrando por partes obtemos

$$\Gamma(\alpha + 1) = \int_0^\infty t^\alpha e^{-t}dt = \Big[- e^{-t}t^\alpha \Big]_0^\infty + \int_0^\infty xt^{\alpha-1}e^{-t}dt = \alpha \int_0^\infty t^{\alpha-1}e^{-t}dt = \alpha\Gamma(\alpha).$$

c) Es consecuencia directa de los apartados a) y b).

d) Se obtiene utilizando definiciones alternativas de la función gamma tras extenderla a $\mathbb{C} \setminus \mathbb{Z}_0^-$. Para más información véase [1]. No desarrollamos esta demostración pues solo la necesitamos para el siguiente apartado.

e) Se obtiene al evaluar la fórmula de reflexión en $\alpha = 1/2$.

f) Se obtiene realizando el cambio de variable $t = \beta s$. □

Definición 1.3. Sean $\alpha, \beta > 0$. Definimos la distribución $Gamma(x|\alpha, \beta)$ como la distribución que tiene función de densidad

$$f(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1}e^{-\beta x}, x > 0.$$

El parámetro α se conoce como parámetro de forma ya que influencia la forma de la distribución, como muestra el siguiente resultado.

Proposición 1.11. *La función de densidad de la distribución* $Gamma(\alpha, \beta)$ *verifica las siguientes propiedades:*

- *Si* $0 < \alpha < 1$*, entonces* $f(x|\alpha, \beta)$ *es decreciente y* $f(x) \to \infty$ *para* $x \to 0$*.*
- *Si* $\alpha = 1$*, entonces* $f(x|\alpha, \beta)$ *es decreciente con* $f(0) = 1$*.*
- *Si* $\alpha > 1$*, entonces* $f(x|\alpha, \beta)$ *crece en* $[0, (\alpha - 1)/\beta]$ *y decrece en* $[(\alpha - 1)/\beta, \infty]$*.*
- *Si* $0 < \alpha \leq 1$*, entonces* $f(x|\alpha, \beta)$ *es convexa.*
- *Si* $1 < \alpha \leq 2$*, entonces* $f(x|\alpha, \beta)$ *es cóncava en* $[0, (\alpha - 1 + \sqrt{\alpha - 1})/\beta]$ *y convexa en* $[(\alpha - 1 + \sqrt{\alpha - 1})/\beta, \infty]$*.*
- *Si* $2 < \alpha$*, entonces* $f(x|\alpha, \beta)$ *es cóncava en* $[(\alpha - 1 - \sqrt{\alpha - 1})/\beta, (\alpha - 1 + \sqrt{\alpha - 1})/\beta]$ *y convexa en* $[0, (\alpha - 1 - \sqrt{\alpha - 1})/\beta]$ *y* $[(\alpha - 1 + \sqrt{\alpha - 1})/\beta, \infty]$*.*

Demostración. Los resultados se obtienen mediante las herramientas habituales del cálculo. Basta estudiar la derivada primera y la derivada segunda

$$f'(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-2} e^{-\beta x}[(\alpha-1) - \beta x];$$
$$f''(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-3} e^{-\beta x}\left[(\alpha-1)(\alpha-2) - 2\beta(\alpha-1)x + \beta^2 x^2\right]. \qquad \square$$

La Figura 2 muestra la función de densidad de la distribución gamma para distintos valores de α. El parámetro β se denomina parámetro de escala debido a su influencia en la escala de la función de densidad. La Figura 3 muestra la función de densidad de la distribución gamma para distintos valores de β.

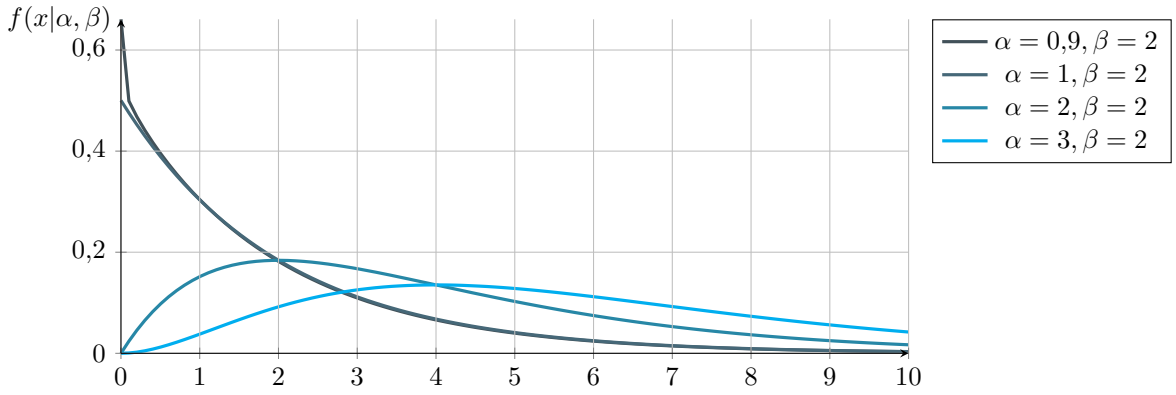

Figura 2: Densidad de la distribución gamma con distintos valores de α.

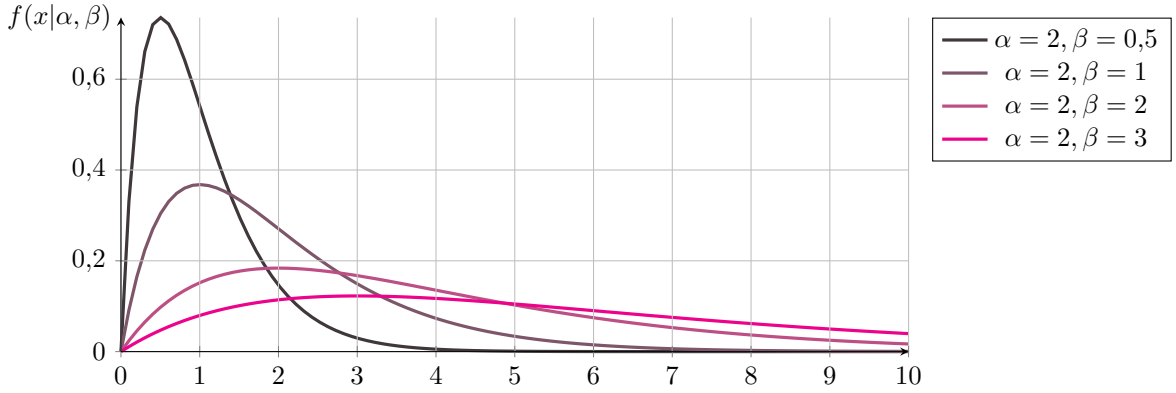

Figura 3: Densidad de la distribución gamma con distintos valores de β.

Proposición 1.12. *La función característica de la distribución Gamma$(x|\alpha,\beta)$ viene dada por* $\varphi_X(t) = \left(\frac{\beta}{\beta - it}\right)^\alpha$.

Demostración. Basta utilizar el cambio de variable $g(y) = y/(\beta - it)$ como sigue

$$E[e^{itX}] = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-(\beta-it)x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)(\beta-it)^\alpha} \int_0^\infty y^{\alpha-1} e^{-y} dy = \left(\frac{\beta}{\beta - it}\right)^\alpha.$$

Nótese que a pesar de ser una integral de contorno compleja el cambio de variable afín es válido como se comentó en la Proposición 1.2. □

Corolario 1.13. *El momento k-ésimo de la distribución $Gamma(x|\alpha,\beta)$ es $\alpha(\alpha+1)\dots(\alpha+k-1)/\beta^k$.*
Demostración. Tenemos que $i^k E[X^k] = \varphi_X^{(k)}(t) = i^k\alpha(\alpha+1)\dots(\alpha+k-1)/\beta^k$. □

Proposición 1.14. *La función generatriz de momentos de la distribución $Gamma(x|\alpha,\beta)$ viene dada por $\varphi_X(t) = \left(\frac{\beta}{\beta-t}\right)^\alpha$.*
Demostración. La demostración es análoga a la dada en la Proposición 1.12. □

Corolario 1.15. *La distribución $Gamma(x|\alpha,\beta)$ tiene media α/β y varianza α/β^2.*

Proposición 1.16. *Sea $n \geq 1$. Consideremos $X_1,\dots,X_n$ variables aleatorias independientes tales que X_j sigue una distribución $Gamma(x|\alpha_i,\beta)$. Entonces, $\sum_{i=1}^n X_j$ sigue una distribución $Gamma(x|\sum_{i=1}^n \alpha_i,\beta)$.*
Demostración. En primer lugar, calculamos la función característica de $\sum_{i=1}^n X_j$ como sigue

$$E[e^{i\sum X_j}] = E[\prod e^{iX_j}] = \prod E[e^{iX_j}] = \left(\frac{\beta}{\beta-it}\right)^{\sum \alpha_j},$$

donde se ha utilizado que la esperanza del producto de dos variables aleatorias independientes es el producto de las esperanzas. Por último, nótese que la función característica de la variable $\sum X_j$ es la función característica de $Gamma(x|\sum_{i=1}^n \alpha_i,\beta)$. El hecho de que la función característica de una distribución la determina de forma unívoca finaliza la prueba. □

Proposición 1.17. *Sea $X \sim N(x|0,\sigma^2)$. La variable aleatoria $Y = X^2$ sigue una distribución $Gamma(y,1/2,1/(2\sigma^2))$. En particular, para $\sigma = 1$, $Y = X^2$ sigue una distribución χ_1^2.*
Demostración. Sean F y G las funciones de distribución de las variables X e Y respectivamente. Tenemos que $G(y) = P(X^2 \leq y) = P(-\sqrt{y} \leq X \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y})$. Derivando, obtenemos

$$G'(y) = \frac{F'(\sqrt{y}) + F'(-\sqrt{y})}{2\sqrt{y}} = \frac{1}{\sqrt{y}}\frac{1}{\sqrt{2\pi}\sigma}e^{-y/(2\sigma^2)} = \frac{(1/(2\sigma^2))^{1/2}}{\Gamma(1/2)}y^{-1/2}e^{-y/(2\sigma^2)}.$$

Por último, basta darse cuenta de que $G'(y)$ es la función de densidad de $Gamma(y,1/2,1/(2\sigma^2))$. □

1.2.4. Distribución beta

La famila de distribuciones beta se encuentra definida sobre el intervalo $(0,1)$. En su definición entra en juego la denominada función beta, de ahí su nombre.

Definición 1.4. Se define la función beta como la aplicación $\beta : (0,\infty)\times(0,\infty) \to (0,\infty)$ dada por

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt.$$

Proposición 1.18. *Para cada $x,y > 0$ se tiene que $\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = \beta(x,y)$. Como consecuencia, la función beta está bien definida.*
Demostración. En primer lugar escribimos $\Gamma(x)\Gamma(y)$ como una integral doble

$$\Gamma(x)\Gamma(y) = \int_0^\infty e^{-u}u^{x-1}\,du\int_0^\infty e^{-v}v^{y-1}\,dv = \int_0^\infty\int_0^\infty e^{-u-v}u^{x-1}v^{y-1}\,du\,dv.$$

La expresión anterior nos sugiere utilizar el cambio de variable $(u, v) = J(t, s) = (st, (1-t)s)$. Nótese que $|J(t, s)| = s$. Aplicamos el cambio a continuación

$$\Gamma(x)\Gamma(y) = \int_0^\infty \left(\int_0^1 e^{-s}(st)^{x-1}(s(1-t))^{y-1}|J(t,s)|\, dt \right) ds$$
$$= \int_0^\infty e^{-s}s^{x+y-2}s \left(\int_0^1 t^{x-1}(1-t)^{y-1}\, dt \right) ds = \Gamma(x+y)\beta(x,y). \qquad \square$$

En la práctica siempre se utiliza la función gamma para evaluar la función beta. Ya podemos definir la distribución beta.

Definición 1.5. Sean $p, q > 0$. Definimos la distribución $beta(x|p, q)$ como la distribución que tiene función de densidad

$$f(x|p, q) = \frac{1}{\beta(p, q)} x^{p-1}(1-x)^{q-1}, 0 < x < 1.$$

Claramente, la función de densidad integra 1. Esta distribución asigna probabilidad 1 al intervalo $(0, 1)$. Por ello, es útil en modelos de proporciones. Las Figuras 4 y 5 muestran la distribución beta cambiando los valores p y q respectivamente. Podemos observar que las funciones de densidad de $beta(x|p, q)$ y $beta(x|q, p)$ son simétricas respecto del punto $1/2$. Esto se puede demostrar fácilmente a partir de la definición. La Figura 6 muestra la distribución beta con iguales valores de p y q. Vemos que las densidades son simétricas en el eje $x = 1/2$, hecho que también puede demostrarse fácilmente a partir de la definición.

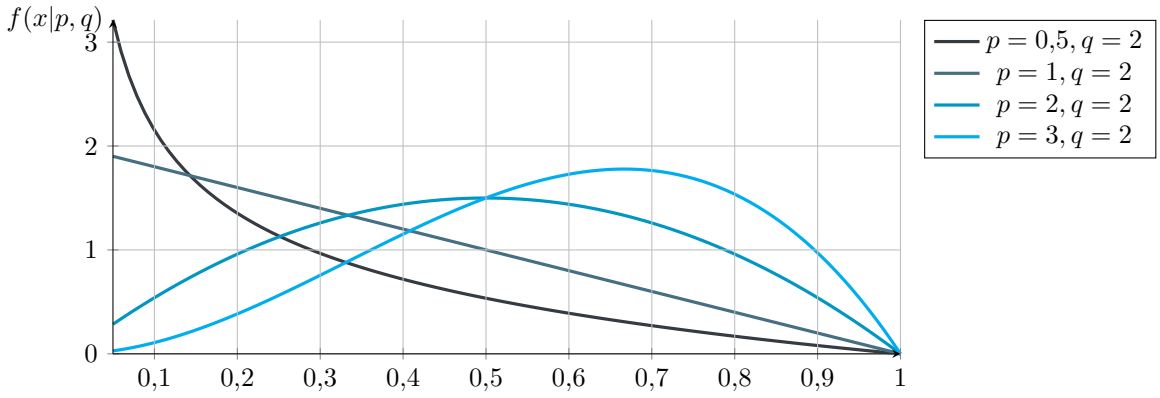

Figura 4: Densidad de la distribución beta con distintos valores de p.

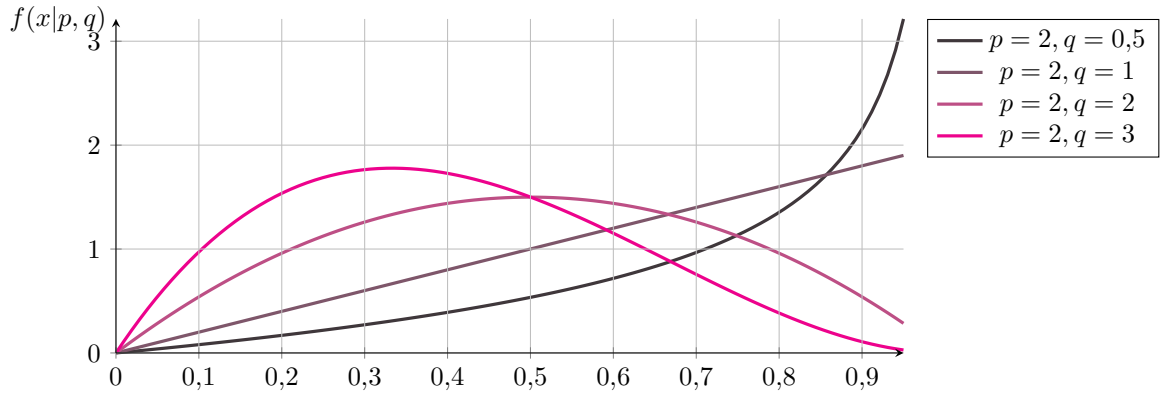

Figura 5: Densidad de la distribución beta con distintos valores de q.

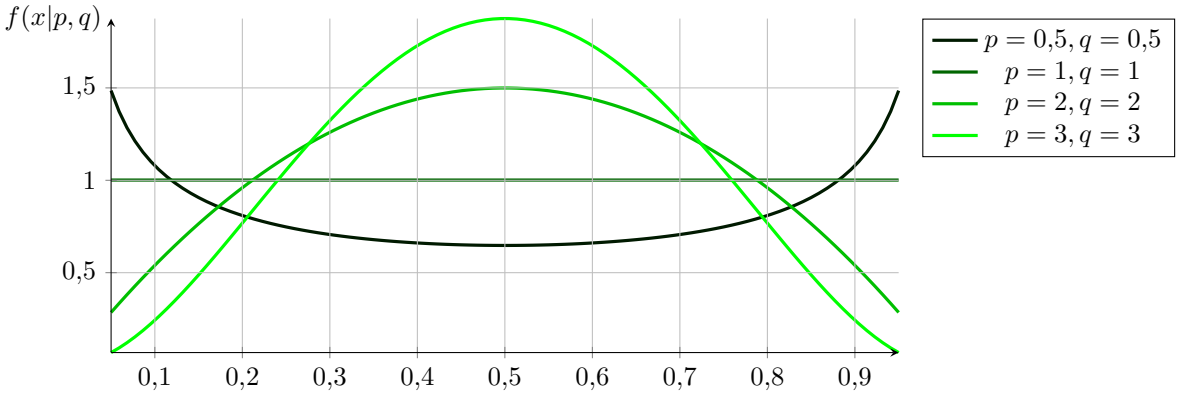

Figura 6: Densidad de la distribución beta con $p = q$.

Proposición 1.19. *La función característica de la distribución beta$(x|p,q)$ viene dada por*

$$\varphi_X(t) = 1 + \sum_{k=1}^{\infty} \frac{(it)^k}{k!} \frac{\beta(p+k,q)}{\beta(p,q)}.$$

Demostración. Desarrollamos la función característica utilizando el desarrollo de la exponencial

$$E[e^{itX}] = \frac{1}{\beta(p,q)} \int_0^1 e^{itx} x^{p-1}(1-x)^{q-1} dx = \frac{1}{\beta(p,q)} \int_0^1 \sum_{k=0}^{\infty} \frac{(itx)^k}{k!} x^{p-1}(1-x)^{q-1} dx$$

$$= \frac{1}{\beta(p,q)} \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \int_0^1 x^{p+k-1}(1-x)^{q-1} dx = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \frac{\beta(p+k,q)}{\beta(p,q)} = 1 + \sum_{k=1}^{\infty} \frac{(it)^k}{k!} \frac{\beta(p+k,q)}{\beta(p,q)}. \quad \square$$

Corolario 1.20. *Sea $X \sim$ beta$(x|p,q)$. Entonces, para cada $k \geq 1$ se tiene que*

$$E[X^k] = \frac{\beta(p+k,q)}{\beta(p,q)}.$$

Corolario 1.21. *Sea $X \sim$ beta$(x|p,q)$. Entonces, $E[X] = \frac{p}{p+q}$ y $Var(X) = \frac{pq}{(p+q)^2(p+q+1)}$.*

Demostración. Por el Corolario 1.20 y la Proposición 1.18 tenemos que

$$E[X] = \frac{\beta(p+1,q)}{\beta p,q} = \frac{\Gamma(p+1)\Gamma(q)}{\Gamma(p+q)}\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} = \frac{\Gamma(p+1)}{\Gamma(p)}\frac{\Gamma(p+q)}{\Gamma(p+q+1)} = \frac{p}{p+q}$$

y

$$E[X^2] = \frac{\beta(p+2,q)}{\beta p,q} = \frac{\Gamma(p+2)\Gamma(q)}{\Gamma(p+q+2)}\frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} = \frac{\Gamma(p+2)}{\Gamma(p)}\frac{\Gamma(p+q)}{\Gamma(p+q+2)} = \frac{p(p+1)}{(p+q)(p+q+1)}.$$

Por último, es directo calcular $Var(X) = E[X^2] - E[X]^2$. □

1.2.5. Distribución de Cauchy

Definición 1.6. Sea $\mu \in \mathbb{R}$ y $\sigma > 0$. Definimos la distribución $Cauchy(x|\mu,\sigma)$ como la distribución que tiene función de densidad

$$f(x|\mu,\sigma) = \frac{1}{\sigma\pi}\frac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{\sigma}{\pi(\sigma^2+(x-\mu)^2)},\ x \in \mathbb{R}.$$

La distribución está bien definida. En efecto, utilizando el cambio de variable $x = \sigma y + \mu$ obtenemos

$$\int_{-\infty}^{\infty} \frac{1}{1+\left(\frac{x-\mu}{\sigma}\right)^2}dx = \int_{-\infty}^{\infty}\frac{\sigma}{1+y^2}dy = \sigma\pi.$$

Proposición 1.22. *La función característica de la distribución $Cauchy(x|\mu,\sigma)$ viene dada por*

$$\varphi_X(t) = e^{i\mu t - \sigma|t|}.$$

Demostración. En primer lugar, demostramos el resultado para $Cauchy(x|0,1)$. Tenemos que

$$\varphi_X(t) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{e^{itz}}{1+z^2}dz = e^{-|t|},$$

donde la última igualdad se explica en [2]. Ahora, si $X \sim Cauchy(x|\mu,\sigma)$, entonces $Y = (X-\mu)/\sigma \sim Cauchy(x|0,1)$ y, por tanto, obtenmos

$$\varphi_X(t) = E[e^{itX}] = E[e^{it(\sigma Y+\mu)}] = e^{it\mu}\varphi_Y(\sigma t) = e^{i\mu t - \sigma|t|}.$$ □

Nótese que la función característica de la distribución de Cauchy no es diferenciable en 0. Consecuentemente, esta distribución no tiene momentos de orden mayor o igual que 1.

Proposición 1.23. *Sean X e Y dos variables aleatorias independientes con distribuciones $Cauchy(x|\mu_1,\sigma_1)$ y $Cauchy(x|\mu_2,\sigma_2)$ respectivamente. Entonces, $X+Y \sim Cauchy(x|\mu_1+\mu_2,\sigma_1+\sigma_2)$.*

Demostración. Nótese que $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{it(\mu_1+\mu_2)-|t|(\sigma_1+\sigma_2)}$. La prueba finaliza al darse cuenta de que ésta es la función característica de $Cauchy(x|\mu_1+\mu_2,\sigma_1+\sigma_2)$. □

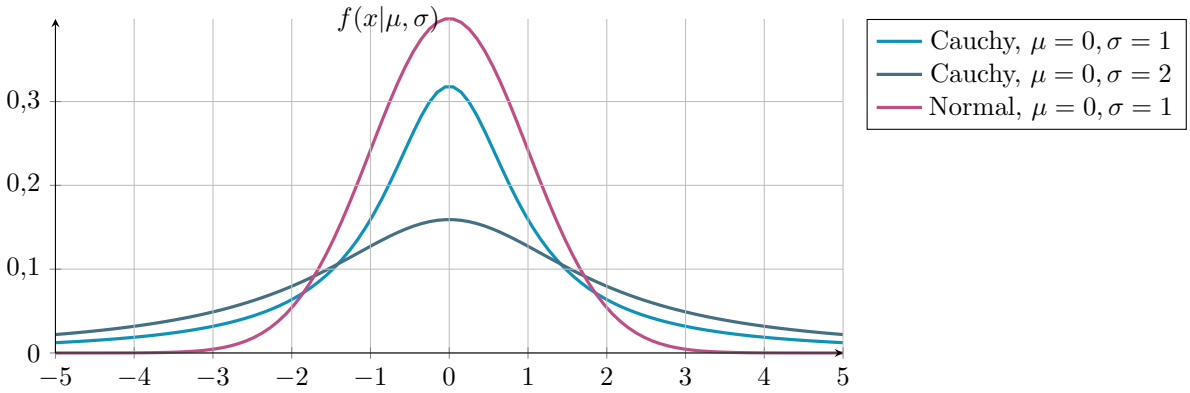

Figura 7: Densidad de la distribución de Cauchy comparada con la distribución normal.

1.2.6. Distribución de Laplace

Definición 1.7. Sea $\mu \in \mathbb{R}$ y $\sigma > 0$. Definimos la distribución de Laplace, y la denotamos $Laplace(x|\mu,\sigma)$ como la distribución que tiene función de densidad

$$f(x|\mu,\sigma) = 2\sigma e^{-|x-\mu|/\sigma},\ x \in \mathbb{R}.$$

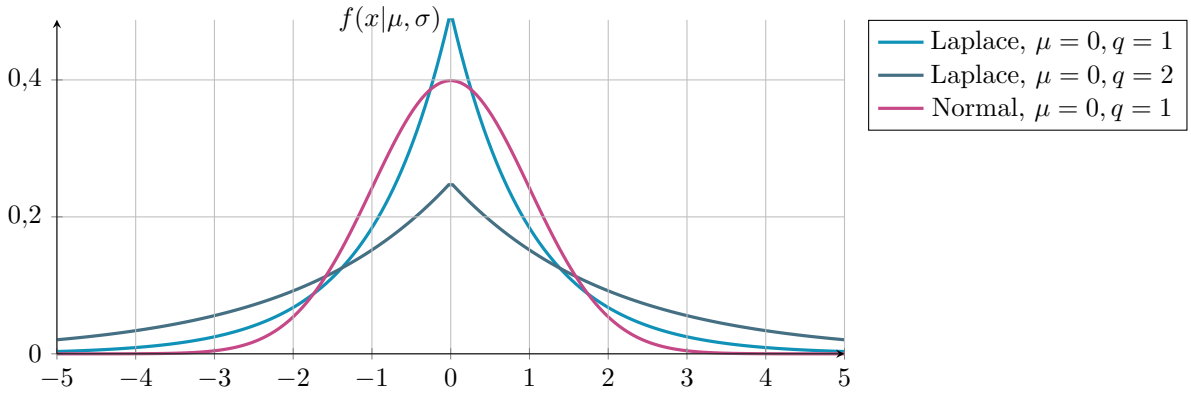

Figura 8: Densidad de la distribución de Laplace comparada con la densidad de la distribución normal.

1.2.7. Distribución T de Student

1.2.8. Distribución de Dirichlet

2. Estimación de parámetros

Supongamos que estamos estudiando un fenómeno aleatorio que sabemos que sigue una distribución $f(X|\theta_0)$, donde $\theta_0 \in \Omega$ es un parámetro que no es conocido. Nuestro objetivo es estimar el parámetro θ_0 a partir de una muestra $x_1, \ldots, x_n$. Para ello buscamos una función T_n de manera que podamos decir $\theta_0 \approx T_n(x_1, \ldots, x_n)$.

Definición 2.1. Un estimador puntual es una función medible $T_n(X_1, \dots, X_n)$ que toma valores en Ω, donde Ω es el dominio del parámetro a estimar. Una estimación es la evaluación obtenida por un estimador sobre una muestra $x_1, \dots, x_n$, esto es, $T_n(x_1, \dots, x_n)$.

Nótese que la nomenclatura es ambigua. Para nosotros una muestra es una secuencia finita de variables aleatorias independientes e idénticamente distribuidas. Sin embargo, a los valores $x_1, \dots, x_n$ obtenidos en la práctica también se le denomina muestra. Algunos autores evitan esta abigüedad denominando a $x_1, \dots, x_n$ realización de la muestra. Nosotros distinguiremos entre ambos casos mediante el uso de mayúsculas para denotar variables aleatorias y el uso de minúsculas para denotar valores concretos.

En múltiples situaciones encontramos estimadores de calidad de forma natural. Por ejemplo, imaginemos que el parámetro θ_0 se corresponde con la media de la distribución $f(X|\theta_0)$. En tal caso, parece claro que el mejor estimador para θ_0 será la media muestral $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$. Sin embargo, en general no sabemos qué estimador hay que utilizar. Buscamos técnicas que nos proporcionen estimadores que sean razonables. En ocasiones querremos estimar $g(\theta_0)$, donde g es determinada transformación de Ω en otro espacio más manejable.

2.1. Método de los momentos

El método de los momentos es, probablemente, el método más antiguo para estimar parámetros. Fue propuesto por Pearson al finales del siglo XIX. En muchos casos los resultados de este método son mejorables. Sin embargo, siempre es un último recurso en el caso de que no podamos aplicar otros métodos.

Sea $X_1, \dots, X_n$ una muestra de un fenómeno con función de distribución $f(X|\theta)$ con $\theta = (\theta_1, \dots, \theta_m) \in \Omega \subset \mathbb{R}^m$. Definimos los momentos de la muestra como $m_j = \frac{1}{n}\sum_{i=1}^n X_i^j$. En media se debería cumplir que $m_j = E_\theta X^j$ para todo j tal que $E_\theta X^j$ existe. Nótese que $E_\theta X^j = \mu_j(\theta_1, \dots, \theta_m)$ es una función que depende de $\theta_1, \theta_2, \dots, \theta_k$. El método de los momentos propone como estimador a una solución del sistema de ecuaciones

$$\begin{aligned} m_1 &= \mu_1(\theta_1, \dots, \theta_k), \\ m_2 &= \mu_2(\theta_1, \dots, \theta_k), \\ &\vdots \\ m_k &= \mu_k(\theta_1, \dots, \theta_k). \end{aligned} \tag{1}$$

Ejemplo 2.1: [Distribución normal] Supongamos que $X_1, \dots, X_n$ son muestras de una distribución normal $N(\theta, \sigma^2)$. En el contexto anterior, los parámetros a estimar son $\theta_1 = \theta, \theta_2 = \sigma^2$. En este caso el sistema (1) viene dado por las ecuaciones $\overline{X} = \theta$ y $m_2 = \theta^2 + \sigma^2$. La solución claramente es $\theta = \overline{X}$ y

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

En este caso, los estimadores obtenidos coinciden con nuestra intuición. Este método es más útil cuando no disponemos de un estimador intuitivo. △

2.2. Método de la máxima verosimilitud de Fisher

El método de la máxima verosimilitud es una de las técnicas más utilizada para obtener estimadores de calidad.

Definición 2.2. Sea $x_1, \dots, x_n$ una muestra de un fenómeno con función de distribución $f(X|\theta_0)$, donde $\theta_0 \in \Omega$. Se define la función de verosimilitud para cada $\theta \in \Omega$ como $L(\theta|x) = \prod_{i=1}^n f(x_i|\theta)$.

Para cada posible valor θ del parámetro a estimar, la verosimilitud proporciona la credibilidad que se le da a θ para los datos $x_1, \dots, x_n$. Buscamos una aproximación $\hat{\theta}$ de θ_0 en base a la muestra obtenida. Parece lógico que si asumimos que los datos son correctos, entonces una buena aproximación será aquella en la que los datos sean coherentes, esto es, la probabilidad de que se den datos similares a la muestra observada debe ser lo más alta posible.

Definición 2.3. Para cada elemento $x = (x_1, \dots, x_n)$ del espacio muestral, definimos $\hat{\theta}(x) \in \Omega$ como un máximo global de $L(\theta|x)$. El estimador máximo verosímil (EMV) de una muestra X se define como $\hat{\theta}(X)$.

El estimador máximo verosímil presenta principalmente dos problemas.

- Cálculo del estimador. Para calcular $\hat{\theta}(X)$ es necesario maximizar una función. Muchas veces esto es complejo incluso para funciones de densidad comunes. Es más, puede suceder que la verosimilitud presente múltiples máximos globales y, por tanto, el estimador máximo verosímil no está bien definido. Necesitaremos condiciones sobre la distribución que nos permitan asegurar la buena definición del estimador máximo verosímil.
- Sensibilidad numérica. El valor $\hat{\theta}(x)$ puede cambiar considerablemente para pequeñas variaciones de x. Nos preguntamos qué condiciones debe verificar la función de distribución para evitar este comportamiento.

Para adentrarnos en el estudio de estos problemas necesitaremos teoría general de estimadores. Antes de desarrollarla realizaremos varios ejemplos de cálculo de estimadores máximo verosímiles.

Comentario 2.2. Los máximos globales de la función $L(\theta|x)$ se corresponden con los máximos globales de la función $\log L(\theta|x) = \sum_{i=1}^n \log f(x_i|\theta)$. En múltiples ocasiones es más sencillo maximizar esta última expresión.

Ejemplo 2.3: [Distribución normal] Consideremos una muestra $X_1, \dots, X_n$ de un fenómeno con distribución $N(\theta, 1)$. En primer lugar, calculamos la función de verosimilitud

$$L(\theta|x) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\theta)^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{i=1}^n (x_i-\theta)^2/2}.$$

En virtud del Comentario 2.2 maximizamos la función $-\sum_{i=1}^n (x_i-\theta)^2/2 - n/2 \log(2\pi)$. Maximizar esta función equivale a minimizar $h(\theta) = \sum_{i=1}^n (x_i-\theta)^2$. Derivando, obtenemos que $h'(\theta) = 0$ si, y solo si, $\theta = \overline{x}$. Además, es rutinario comprobar que $\overline{x}$ es el mínimo absoluto de h. Por tanto, $\overline{x}$ es el máximo absoluto de $L(\theta|x)$. Tenemos pues $\hat{\theta}(x) = \overline{x}$. △

A continuación pretendemos extender el método de la máxima verosimilitud para estimar $g(\theta)$, donde $g : \Omega \to \Omega'$ sobreyectiva. Si la aplicación g fuese inyectiva, entonces podemos definir de norma natural la verosimilitud de $\eta \in \Omega'$ como $L^*(\eta|x) = L(g^{-1}(\eta)|x)$. Claramente, el valor que maximiza $L^*(\eta|x)$, que denotaremos $\hat{g}(x)$, es $g(\hat{\theta}(x))$. Sin embargo, los casos que presentan relevancia práctica son aquellos en los que g no es inyectiva ya que de esta forma conseguimos reducir la dimensionalidad del espacio de parámetros. Necesitamos extender la definición de verosimilitud para abordar esta problemática.

Definición 2.4. En el contexto anterior, definimos la verosimilitud inducida por g como

$$L^*(\eta|x) = \sup\{L(\theta|x) : \theta \in g^{-1}(\eta)\}.$$

El valor $\hat{g}(x)$ que maximiza $L^*(\eta|x)$ se denomina estimador maximo verosímil de $g(\theta)$.

La definición anterior es artificial en el sentido de que se realiza con el fin de poder mantener la propiedad de invarianza del estimador máximo verosímil, que se recoge en el siguiente teorema.

Teorema 2.4 (Invarianza de Zehna)**.** *Para cualquier aplicación sobreyectiva $g : \Omega \to \Omega'$ se tiene que $\hat{g}(X) = g(\hat{\theta}(X))$.*
Demostración. En primer lugar, la definición de la verosimilitud inducida proporciona

$$\sup_{\eta \in \Omega'} L^*(\eta|x) = \sup_{\eta \in \Omega'} \sup\{L(\theta|x) : \theta \in g^{-1}(\eta)\} = \sup_{\theta \in \Omega} L(\theta|x).$$

Por tanto, si la verosimilitud tiene un máximo global $\hat{\theta}(x)$, entonces lo tiene la verosimilitud inducida (el recíproco puede no ser cierto) y se alcanza en $g(\hat{\theta}(x))$. □

2.3. Teoría general de estimadores

En esta sección introduciremos conceptos y definiciones relacionados con estimadores arbitrarios. El objetivo de esta teoría es dotarnos de herramientas que nos permitan abordar el estudio práctico de estimadores concretos, como el estimador máximo verosímil.

2.3.1. Estadísticos suficientes

Fijemos $\{f(x|\theta) : \theta \in \Omega\}$ una familia de distribuciones. Sea $X = (X_1, \dots, X_n)$ una muestra de la distribución $f(x|\theta_0)$. Nuestro objetivo es inferir el parámetro θ_0 a partir de la muestra. El concepto de estadístico suficiente nos permitirá separar la información contenida en X en dos partes. Una parte contiene toda la información útil sobre θ_0 mientras que la otra parte no dependerá del parámetro θ_0. Consecuentemente, podemos ignorar esta última parte.

Intuitivamente, un estadístico T es suficiente para la familia de distribuciones considerada si $T(X)$ nos permite estimar θ_0 tan bien como lo permite toda la muestra X. Procedemos a dar la definición matemática.

Definición 2.5. Un estadístico $T(X_1, X_2, \dots, X_n)$ es suficiente si para cada $\theta \in \Omega$ y t la distribución condicional de $X_1, X_2, \dots, X_n$ respecto de θ y $T(X) = t$ no depende de θ.

El teorema de factorización de Neyman nos proporciona un criterio práctio para ver si un estadístico es suficiente.

Teorema 2.5. *Sea $\{f(x|\theta) : \theta \in \Omega\}$ una familia de distribuciones. Sea $X = (X_1, \dots, X_n)$ una muestra de la distribución $f(x|\theta)$. Sea $T(X_1, X_2, \dots, X_n)$ un estadístico. Entonces, T es suficiente si y solo si la función de verosimilitud puede factorizarse de la siguiente forma*

$$L(x_1, x_2, \dots, x_n) = h(t; \theta) g(x_1, x_2, \dots, x_n),$$

donde $t = T(x_1, \dots, x_n)$ y $g(x_1, x_2, \dots, x_n)$ no depende de θ.

Si encontramos un estadístico suficiente, entonces podemos inferir el parámetro θ_0 utilizando solamente la función $h(t; \theta)$. Interesa pues que el codominio del estadístico suficiente sea lo más simple posible. Los estadísticos suficientes son especialmente interesantes al aplicar el método de la máxima verosimilitud.

Ejemplo 2.6: [Distribución normal, media desconocida] Sea $X = (X_1, \dots, X_n)$ una muestra de $N(x|\mu, \sigma^2)$ donde solamente σ^2 es conocido ($\theta = \mu$). Es fácil verificar que

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2 = (n-1)S^2 + n(\overline{x} - \mu)^2.$$

A partir de la igualdad anterior obtenemos

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-((n-1)S^2 + n(\overline{x} - \mu)^2)/(2\sigma^2)).$$

Definimos

$$g(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(n-1)S^2/(2\sigma^2)) \text{ y } h(t|\mu) = \exp(n(t-\mu)^2)/(2\sigma^2)).$$

Tenemos que $f(x|\theta) = h(\overline{x}|\theta)g(x_1, x_2, \dots, x_n)$ y, por tanto, $T(x) = \overline{x}$ es suficiente. △

Ejemplo 2.7: [Distribución normal, ambos parámetros son desconocidos] Sea $X = (X_1, \dots, X_n)$ una muestra de $N(x|\mu, \sigma^2)$ donde μ y σ^2 son desconocidos ($\theta = (\mu, \sigma^2)$). Tenemos que

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2)/(2\sigma^2)).$$

Por tanto, el estadístico $T(X_1, \dots, X_n) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ es suficiente (tomamos $g(x_1, x_2, \dots, x_n) = 1$). También podemos desarrollar $f(x|\theta)$ como sigue

$$f(x|\theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp(-(\sum_{i=1}^{n}(x_i - \overline{x})^2 + n(\overline{x} - \mu)^2)/(2\sigma^2)).$$

Consecuentemente, el estadístico $T(X_1, \dots, X_n) = (\overline{x}, S^2)$ también es suficiente. Nótese que el estadístico $T(X_1, \dots, X_n) = (X_1, \dots, X_n)$ es trivialmente suficiente, pero no aporta ninguna información. △

2.3.2. Score, hipótesis de regularidad y función de información de Fisher

Definición 2.6. Sea $\Omega \subset \mathbb{R}^m$ un abierto y sea $\{f(X|\theta) : \theta \in \Omega\}$ una familia de funciones de densidad. Sea $X = (X_1, \dots, X_n)$ una muestra que sigue una distribución con función de densidad $f(X|\theta_0)$. Si la función de verosimilitud para los valores $x = (x_1, \dots, x_n)$ es diferenciable en $\theta \in \Omega$, entonces definimos el score de θ como el gradiente de la función $\log L(x; \theta)$ y lo denotamos $S(x; \theta)$.

Intuitivamente el score indica la sensibilidad de la verosimilitud en un punto. Nos centraremos en el estudio del score cuando Ω es un abierto de $\mathbb{R}$. En tal caso

$$S(x; \theta) = \frac{\partial}{\partial\theta} \log f(x|\theta) = \frac{\frac{\partial}{\partial\theta} f(x|\theta)}{f(x|\theta)}.$$

Supongamos que el score de θ existe para cualesquiera valores de la muestra $x_1, \dots, x_n$. En tal caso es natural considerar la función $E_{X|\theta}[S(X; \theta)]$, que depende solamente de θ. Si θ fuese el parámetro a estimar, entonces $E_{X|\theta}[S(X; \theta)]$ mide la sensibilidad media de la verosimilitud en θ.

Lema 2.8. *Sea $\Omega \subset \mathbb{R}$ un abierto y sea $\{f(X|\theta) : \theta \in \Omega\}$ una familia de funciones de densidad que verifican las siguientes condiciones de regularidad:*

a) Para cualesquier muestra $x = (x_1, \dots, x_n)$ la función $L(x;\theta)$ es diferenciable para todo $\theta \in \Omega$.

b) Se verifica

$$\frac{\partial}{\partial\theta}\int_X f(x|\theta)dx = \int_X \frac{\partial}{\partial\theta} f(x|\theta)dx.$$

Entonces, $E_{X|\theta}[S(X;\theta)] = 0$.
Demostración. Tenemos que

$$E_{X|\theta}[S(X;\theta)] = \int_X \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta)dx = \int_X \frac{\partial}{\partial\theta} f(x|\theta)dx = \frac{\partial}{\partial\theta}\int_X f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0. \qquad \square$$

En el resultado anterior aparecen por primera vez hipótesis de regularidad sobre las distribuciones a estudiar. Nótese que en la práctica normalmente vamos a trabajar con distribuciones que satisfagan estas hipótesis. La hipótesis b) se verifica si la derivada de la verosimilitud es continua [3]. Consecuentemente, todas las distribuciones continuas estudiadas, exceptuando la distribución de Laplace, cumplen estas hipótesis de regularidad (su función de densidad es de clase infinito con respecto a θ).

Definición 2.7. Sea $\Omega \subset \mathbb{R}$ un abierto y sea $\{f(X|\theta) : \theta \in \Omega\}$ una familia de funciones de densidad para la cual siempre existe el score. Dado $\theta \in \Omega$, definimos la función de información de Fisher en θ como el segundo momento de la variable aleatoria $S(X;\theta)$, donde $X = (X_1, \dots, X_n)$ es una muestra de la distribución con función de densidad $f(X|\theta)$. Se denota $\mathcal{I}(\theta) := E_{X|\theta}[S(X;\theta)^2] \geq 0$.

Si en determinado contexto no está clara la muestra X para la cual calculamos la información de Fisher, entonces la denotamos $\mathcal{I}_X$ o $\mathcal{I}^X$.

El siguiente resultado nos permite explicar por qué se define de esta forma la información de Fisher.

Corolario 2.9. *Bajo las hipótesis de regularidad del Lema 2.8, tenemos que $\mathcal{I}(\theta) = Var_{X|\theta}(S(X;\theta))$.*
Demostración. Nótese que $Var_{X|\theta}(S(X;\theta)) = \mathcal{I}(\theta) - E_{X|\theta}[S(X;\theta)]^2$. El Lema 2.8 nos indica que $E_{X|\theta}[S(X;\theta)] = 0$. $\square$

Como consecuencia, la información de Fisher nos informa de cómo varía la sensibilidad de la verosimilitud en θ. Si la información de Fisher es pequeña, entonces la sensibilidad de la verosimilitud en θ no depende prácticamente de la muestra utilizada y, por tanto, siempre será cercana a cero. Si por el contrario la información de Fisher es muy grande, entonces la sensibilidad de la verosimilitud en θ varía mucho en función de la muestra con la que se trabaje. Si utilizamos el estimador máximo verosímil, entonces estamos maximizando el logaritmo de la verosimilitud. Buscamos pues aquellos θ que sean extremos relativos de $\log L(x;\theta)$ y, por tanto, verifiquen $S(x;\theta) = 0$. Consecuentemente, nos interesa que $\mathcal{I}(\theta)$ sea grande para todo θ ya que de esta forma podremos discriminar aquellos θ que tengan score no nulo (no son extremos relativos de $\log L(x;\theta)$). Si en determinado θ la información de Fisher es muy pequeña, obtendremos que θ es un candidato a estimador máximo verosímil para casi cualquier muestra, incluso para muestras poco probables bajo ese parámetro, lo cual dificulta el correcto cómputo del estimador.

En lo que sigue habitualmente exigiremos unas hipótesis de regularidad más fuertes, denominadas hipótesis o condiciones de regularidad de Cramer-Rao. Estas hipótesis son las siguientes:

i) Ω es un abierto de $\mathbb{R}$.

ii) Para cualquier muestra $x = (x_1, \dots, x_n)$, la verosimilitud $L(x|\theta)$ es dos veces derivable en Ω.

iii) $\frac{\partial^i}{\partial\theta^i}\int_X f(x|\theta)dx = \int_X \frac{\partial^i}{\partial\theta^i} f(x|\theta)dx$ para $i = 1, 2$.

iv) Para cada $\theta \in \Omega$ se tiene $0 < \mathcal{I}(\theta) < +\infty$.

Todas las distribuciones continuas estudiadas, exceptuando la distribuciónd de Laplace, verifican estas hipótesis de regularidad.

El siguiente lema profundiza en nuestro entendimiento de la función de información de Fisher.

Lema 2.10. *Bajo hipótesis de regularidad de Cramer-Rao tenemos que*

$$\mathcal{I}(\theta) = E_{X|\theta}\left[-\frac{\partial^2}{\partial\theta^2}\log f(X;\theta)\right].$$

Demostración. En primer lugar, podemos escribir

$$\frac{\partial^2}{\partial\theta^2}\log f(X|\theta) = \frac{\frac{\partial^2}{\partial\theta^2}f(X|\theta)}{f(X|\theta)} - \left(\frac{\frac{\partial}{\partial\theta}f(X|\theta)}{f(X|\theta)}\right)^2 = \frac{\frac{\partial^2}{\partial\theta^2}f(X|\theta)}{f(X|\theta)} - \left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2.$$

La demostración finaliza al tomar esperanzas en la expresión anterior y darse cuenta de que

$$E_{X|\theta}\left[\frac{\frac{\partial^2}{\partial\theta^2}f(X|\theta)}{f(X|\theta)}\right] = \int_X \frac{\partial^2}{\partial\theta^2}f(x|\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\int_X f(x|\theta)\,dx = \frac{\partial^2}{\partial\theta^2}\,1 = 0. \qquad \square$$

Como consecuencia, la información de Fisher también indica cuál es la curvatura media de la función $\log L(x;\theta)$, que como vemos, en media es negativa ($\mathcal{I}(\theta) \geq 0$). Para calcular el estimador máximo verosímil intentamos maximizar $\log L(x;\theta)$. Si la función de información de Fisher es habitualmente grande, entonces en media tendremos máximos relativos muy claros.

El Lema 2.10 nos permite calcular la información de Fisher de forma más sencilla, como muestra el siguiente ejemplo.

Ejemplo 2.11: Calculamos la función de información de Fisher de $X \sim N(x|\mu,\sigma^2)$ para varias configuraciones de la distribución normal.

- El parámetro σ^2 es conocido. Tenemos que $\log f(X|\mu,\sigma^2) = -(X-\mu)^2/(2\sigma^2) - \log(\sqrt{2\pi}) - \log(\sigma^2)/2$. Consecuentemente, deducimos que

  $$\frac{\partial^2}{\partial\mu^2}f(X|\mu,\sigma^2) = \frac{-1}{\sigma^2}.$$

  Por tanto, $\mathcal{I}(\mu) = 1/\sigma^2$.

- El parámetro μ es conocido. Obtenemos que

  $$\frac{\partial^2}{\partial(\sigma^2)^2}f(X|\mu,\sigma^2) = -\frac{(X-\mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}.$$

  Por tanto, podemos calcular $\mathcal{I}(\sigma^2)$ utilizando que $E[(X-\mu)^2] = Var(X) = \sigma^2$. Obtenemos que

  $$\mathcal{I}(\sigma^2) = E_{X|\sigma^2}[\frac{(X-\mu)^2}{\sigma^6} - \frac{1}{2\sigma^4}] = \frac{1}{\sigma^6}Var((X-\mu)^2) - \frac{1}{2\sigma^4} = \frac{1}{2\sigma^4}. \qquad \triangle$$

Comentario 2.12. Bajo hipótesis de regularidad de Cramer-Rao, si $X = (X_1,\ldots,X_n)$ es una muestra de $f(X|\theta)$, entonces tenemos que

$$\frac{\partial^2}{\partial\theta^2}\log(f(X;\theta)) = \frac{\partial^2}{\partial\theta^2}\left(\sum_{i=1}^n \log(f(X_i;\theta))\right) = \sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log(f(X_i;\theta)).$$

Consecuentemente, $\mathcal{I}^{X_1,\ldots,X_n}(\theta) = \sum_{i=1}^n \mathcal{I}^{X_i}(\theta) = n\mathcal{I}^{X_i}(\theta)$.

Lema 2.13. *Bajo hipótesis de regularidad de Cramer-Rao, sea $T(X_1, \dots, X_n)$ un estadístico tal que su ditribución inducida también verifica las hipótesis de regularidad de Cramer-Rao. Entonces, para cualquier $\theta \in \Omega$ se tiene*

$$\mathcal{I}_{T(X)}(\theta) \leq \mathcal{I}_X(\theta).$$

Además, la igualdad se da para todo $\theta \in \Omega$ si, y solo si, T es suficiente.

En lo que sigue necesitaremos el siguiente lema.

Lema 2.14 (Desigualdad de Jenssen)**.** *Sean X una variable aleatoria cuya imagen está contenida en un intervalo I. Sea $g : I \to \mathbb{R}$ una función.*

a) Si g es convexa, entonces $E[g(X)] \geq g(E[X])$.

b) Si g es cóncava, entonces $E[g(X)] \leq g(E[X])$

Proposición 2.15. *Sea $X = (X_1, \dots, X_n)$ una muestra de $f(X; \theta_0)$. Entonces, para cada $\theta_1 \in \Omega$ se tiene*

$$E_{X|\theta_0} \log f(X|\theta_0) \geq E_{X|\theta_1} \log f(X|\theta_0).$$

Demostración. Por la desigualdad de Jenssen obtenemos

$$E_{X|\theta_0} \log \frac{f(X|\theta_1)}{f(X|\theta_0)} \leq \log \int_X \frac{f(X|\theta_1)}{f(X|\theta_0)} f(X|\theta_0)\, dx = \log \int_X f(X|\theta_1)\, dx = 0,$$

de donde se deduce el resultado. □

Cabe mencionar que la información de Fisher puede definirse cuando $\Omega \subset \mathbb{R}^m$. Incluimos la definición por complitud, aunque no entraremos en ella a fondo.

Definición 2.8. Sea $\Omega \subset \mathbb{R}^m$ un abierto y sea $\{f(X|\theta) : \theta \in \Omega\}$ una familia de funciones de densidad para la cual siempre existe el score. Dado $\theta \in \Omega$, definimos la función de información de Fisher en θ como

$$\left(\mathcal{I}(\theta)\right)_{i,j} = E_{X|\theta}\left[\left(\frac{\partial}{\partial \theta_i} \log f(X;\theta)\right)\left(\frac{\partial}{\partial \theta_j} \log f(X;\theta)\right)\right],\ 1 \leq i, j \leq m,$$

donde $X = (X_1, \dots, X_n)$ es una muestra de la distribución con función de densidad $f(X|\theta)$.

Bajo determinadas hipótesis de regularidad se puede probar que para cada $1 \leq i, j \leq n$ se verifica

$$\left(\mathcal{I}(\theta)\right)_{i,j} = -E_{X|\theta}\left[\frac{\partial^2}{\partial \theta_i\, \partial \theta_j} \log f(X;\theta)\right].$$

2.3.3. Estimadores insesgados

Para comprobar cómo de bueno es un estimador T podemos definir una función de pérdida $L(\theta, T(X))$ que indique la pérdida asociada a estimar un parámetro mediante T si su verdadero valor es θ. A partir de la función de pérdida definimos la función de riesgo, que asocia a cada posible valor del parámetro la pérdida media producida por el estimador. La función de riesgo viene dada por

$$R_T^L(\theta) = E_{X|\theta}[L(\theta, T(X))].$$

Un estimador T que "minimice uniformemente" la función de riesgo hará mejores estimaciones en media. Con minimizar uniformemente queremos decir que para cada estimador T' se tiene que

$$R_T^L(\theta) \leq R_{T'}^L(\theta)\ \forall\ \theta \in \Omega.$$

En esta sección introducimos un tipo particular de estimadores que minimizan determinada función de riesgo.

Definición 2.9. Se denomina sesgo de un estimador T de $g(\theta)$ a la diferencia entre la esperanza del estimador y el verdadero valor del parámetro a estimar. Diremos que un estimador es insesgado si para cualquier posible valor del parámetro a estimar su sesgo es nulo.

Nótese que el sesgo de un estimador es la función de riesgo asociada a la pérdida $L(\theta, T(X)) = g(\theta) - T(X)$. Un estimador insesgado verifica $0 = g(\theta) - E_{X|\theta}T(X)$ y, por tanto, minimiza uniformemente la función de riesgo. Aunque esta propiedad puede parecer a priori interesante, puede suceder que en la práctica el estimador insesgado no proporcione estimaciones de calidad si la varianza $Var_{X|\theta}(T(X))$ es muy alta.

Claramente, la media muestral es un estimador insesgado de la media de la distribución. El siguiente resultado nos muestra otro ejemplo de un estimador insesgado.

Proposición 2.16. *Sea $X_1, \dots, X_n$ una muestra de alguna población con función de densidad $f(X|\theta_0)$. Definimos la varianza muestral como*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Entonces, S^2 es un estimador insesgado de la varianza de la distribución.

Demostración. Nótese que $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2$. Consecuentemente tenemos

$$E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \sum_{i=1}^{n} E\left[X_i^2\right] - nE[\overline{X}^2] = n(E\left[X_i^2\right] - E[\overline{X}^2]).$$

Utilizando que $Var(\overline{X}) = Var(X_i)/n$ y $E[\overline{X}] = E[X_i]$ obtenemos

$$E[X_i^2] - E[\overline{X}^2] = Var(X_i) + E[X_i]^2 - Var(\overline{X}) - E[\overline{X}]^2 = \frac{n-1}{n}Var(X_i).$$

Por tanto, $E[S^2] = Var(X_i)$ como se quería. □

La función de información de Fisher juega un papel importante en el estudio de los estimadores insesgados como muestra el siguiente Teorema.

Teorema 2.17 (Cota de Cramer-Rao)**.** *Supongamos que se verifican las hipótesis de regularidad de Cramer-Rao. Sea $g : \Omega \to \mathbb{R}$ de clase 1. Sea $\hat{\theta}$ un estimador insesgado de $g(\theta)$ tal que*

$$\int_X \left|\hat{\theta}(x)\frac{\partial}{\partial\theta}f(x|\theta)\right| dx < \infty.$$

Entonces, para todo $\theta \in \Omega$

$$Var_{X|\theta}(\hat{\theta}) \geq \frac{g'(\theta)^2}{\mathcal{I}(\theta)}.$$

Demostración. Puesto que $\hat{\theta}$ es insesgado tenemos que

$$g(\theta) = \int_X \hat{\theta}(x)f(x|\theta)\,dx.$$

Podemos derivar respecto de θ la expresión anterior y utilizar que $\int_X \frac{\partial}{\partial\theta}f(x|\theta)\,dx = 0$, obteniendo

$$g'(\theta) = \int_X \hat{\theta}(x)\frac{\partial}{\partial\theta}f(x|\theta)\,dx = \int_X \left(\hat{\theta}(x) - g(\theta)\right)\frac{\partial}{\partial\theta}f(x|\theta)\,dx. \tag{2}$$

Aplicamos la desigualdad de Cauchy-Schwarz al miembro de la derecha de (2), obteniendo

$$g'(\theta)^2 \leq \left(\int \left(\hat{\theta}(x) - g(\theta)\right)^2 f(x|\theta)\, dx\right)\left(\int \left(\frac{\partial}{\partial\theta}(\log f(x|\theta))\right)^2 f(x|\theta)\, dx\right) = Var_{X|\theta}(\hat{\theta})\, \mathcal{I}(\theta). \quad \square$$

Corolario 2.18. *Supongamos que se verifican las hipótesis de regularidad de Cramer-Rao. Sea $\hat{\theta}$ un estimador insesgado de θ tal que*

$$\int_X \left|\hat{\theta}(x)\frac{\partial}{\partial\theta} f(x|\theta)\right| dx < \infty.$$

Entonces, para todo $\theta \in \Omega$

$$Var_{X|\theta}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_X(\theta)} = \frac{1}{n\mathcal{I}_{X_i}(\theta)}.$$

La cota de Cramer-Rao nos dice que si la información de Ficher en θ es pequeña, entonces cualquier estimador insesgado tendrá una gran varianza y, por tanto, será inestable ante pequeños cambios en la muestra.

Definición 2.10. Un estimador se dice eficiente si alcanza la cota de Cramer-Rao para todo $\theta \in \Omega$.

2.3.4. Consistencia de sucesiones de estimadores

Nos interesa que los estimadores tiendan al parámetro de la distribución cuando el tamaño de la muestra diverge. En tal caso, podemos mejorar el resultado del estimador recurriendo a una mayor muestra de la población.

Definición 2.11. Consideremos una familia de densidades $\{f(x|\theta) : \theta \in \Omega\}$. Una sucesión de estimadores $\hat{\theta}_n$ de $g(\theta)$ es consistente para $\theta_0 \in \Omega$ si toda sucesión X_n de variables aleatorias independientes e idénticamente distribuidas con función de distribución $f(x|\theta_0)$ la sucesión $\hat{\theta}_n(X_1, \ldots, X_n)$ converge en probabilidad (P_{θ_0}) a $g(\theta_0)$. Si $\hat{\theta}_n$ es consitente para todo $\theta_0 \in \Omega$, entonces decimos que es consistente.

Teorema 2.19. *Sea $\hat{\theta}_n$ una sucesión de estimadores de $g(\theta)$ verificando*

a) $\lim_{n\to\infty} E_\theta[\hat{\theta}_n] = g(\theta)$*;*

b) $\lim_{n\to\infty} Var_\theta(\hat{\theta}_n) = 0$.

Entonces, $\hat{\theta}_n$ es consistente para θ.

Demostración. Sea $\varepsilon > 0$. La desigualdad de Markov nos proporciona

$$P_\theta[|\hat{\theta}_n - g(\theta)| \geq \varepsilon] \leq \varepsilon^{-2} E[(\hat{\theta}_n - g(\theta))^2] = \varepsilon^{-2}\left(Var(\hat{\theta}_n) + (E\hat{\theta}_n - g(\theta))^2\right).$$

La prueba finaliza al recordar que el último término converge a 0. $\square$

Otra propiedad interesante de un estimador cuando la muestra tiene a infinito es la siguiente.

Definición 2.12. Consideremos una familia de densidades $\{f(x|\theta) : \theta \in \Omega\}$. Una sucesión de estimadores $\hat{\theta}_n$ de $g(\theta)$ es asintóticamente normal para $\theta_0 \in \Omega$ si toda sucesión X_n de variables aleatorias independientes e idénticamente distribuidas con función de distribución $f(x|\theta_0)$ la sucesión $\sqrt{n}(\hat{\theta}_n(X_1, \ldots, X_n) - g(\theta_0))$ converge en ley a una distribución $N(X|0, \sigma^2)$ para cierto $\sigma^2 > 0$.

Proposición 2.20. *Todo estimador asintóticamente normal es consistente.*

Demostración. Sea $\hat{\theta}_n$ un estimador asintóticamente normal de $g(\theta)$. Tenemos que $n(\hat{\theta}_n - g(\theta))^2$ converge en ley a $Gamma(X|1/2, 1/(2\sigma^2))$ por la Proposición 1.17. Sea F la función de distribución de $Gamma(X|1/2, 1/(2\sigma^2))$ y sea $\varepsilon > 0$. Vamos a probar que $P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \to 1$. En efecto, sea $1 > \alpha \geq 0$. Tomamos y tal que $F(y) = \alpha$. Tenemos que

$$P[n(\hat{\theta}_n - g(\theta))^2 < y] \to F(y) = \alpha.$$

Por tanto, para cada $\delta > 0$ existe n_0 tal que $\varepsilon > y/n$ y para cada $n \geq n_0$ tenemos

$$P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \geq P[n(\hat{\theta}_n - g(\theta))^2 < y] \geq \alpha - \delta.$$

Deducimos que $\liminf P[(\hat{\theta}_n - g(\theta))^2 < \varepsilon] \geq \alpha - \delta$. De la arbitrariedad de δ y α se deduce el resultado. □

Como es natural, el recíproco del anterior no es cierto.

2.4. Estudio teórico del estimador máximo verosímil

En este punto nos preguntamos cuándo está bien definido el estimador máximo verosímil. En tal caso nos interesa saber si el método de la máxima verosimilitud nos proporciona un estimador consistente. Para ello aplicamos los resultados teóricos vistos en la sección anterior.

Proposición 2.21. *Si existe un estadístico suficiente T par ala familia de distribuciones $\{f(X|\theta) : \theta \in \Omega\}$ y $\hat{\theta}$ es un estimador máximo verosímil, entonces $\hat{\theta}$ depende solamente de $T(X)$.*

Demostración. Por el teorema de factorización de estimadores suficientes podemos escribir $f(X|\theta) = h(X)g(T(X), \theta)$. Maximizar $L(x;\theta) = h(x)g(T(x);\theta)$ equivale a maximizar $g(T(x);\theta)$. Por tanto, $\hat{\theta}$ depende solamente de $T(x)$. □

Teorema 2.22. *Bajo las hipótesis de regularidad de Cramer-Rao se verifican las siguientes afirmaciones:*

a) Existe n_0 tal que para cada $n \geq n_0$ la ecuación en probabilidad $0 = \sum_{i=0}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$ tiene solución única. A esta solución se le llama $\hat{\theta}_n(X_1, \ldots, X_n)$. En dicho punto se maximiza la verosimilitud.

b) $\hat{\theta}(X_1, \ldots, X_n)$ es consistente. De hecho, se puede probar que la convergencia a θ_0 es casi segura.

Demostración. Para cualquier muestra X de $f(X|\theta_0)$ tenemos que

$$0 > -\mathcal{I}(\theta_0) = E_{X|\theta_0}\left[\frac{\partial^2}{\partial\theta^2} \log f(X;\theta_0)\right] = \frac{\partial}{\partial\theta} E_{X|\theta_0}\left[\frac{\partial}{\partial\theta} \log f(X;\theta_0)\right].$$

Consecuentemente, $E_{X|\theta_0}[S(X;\theta)]$ es decreciente en un entorno de θ_0. Recordemos que $E_{X|\theta_0}[S(X;\theta_0)] = 0$ por el Lema 2.8. Por tanto, existe $\varepsilon > 0$ tal que

- $E_{X|\theta_0}[S(X;\theta)] > 0$ para todo $\theta \in (\theta_0 - \varepsilon, \theta_0)$;
- $E_{X|\theta_0}[S(X;\theta)] < 0$ para todo $\theta \in (\theta_0, \theta_0 + \varepsilon)$.

Esto implica que θ_0 es un máximo relativo de $E_{X|\theta_0}[\log f(X|\theta)]$. INCOMPLETO. □

Teorema 2.23. *Bajo hipótesis de regularidad de Cramer-Rao, si $\hat{\theta}(X_1, \ldots, X_n)$ es un estimador máximo verosímil consistente, entonces es asintóticamente normal. Además, la varianza de la distribución normal asociada es $1/\mathcal{I}_{X_1}(\theta_0)$.*

Demostración. Escribimos $L_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$. Tenemos que $L_n'(\hat{\theta}_n) = 0$. El teorema del valor medio nos proporciona un θ_n entre θ_0 y $\hat{\theta}_n$ tal que $0 = L_n'(\hat{\theta}) = L_n'(\theta_0) + L_n''(\theta_n)(\hat{\theta}_n - \theta_0)$. Por tanto,

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) = \frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\theta_n)}. \tag{3}$$

Por el teorema central del límite tenemos que

$$\sqrt{n}L_n'(\theta_0) = \sqrt{n}(L_n'(\theta_0) - E_{\theta_0}[S(X_1;\theta_0)]) \to N(0, \mathcal{I}_{X_1}(\theta_0)),$$

donde hemos utilizado que $Var_{\theta_0}(S(X_1;\theta_0)) = \mathcal{I}_{X_1}(\theta_0)$. Estudiamos ahora el denominador de (3). Puesto que $\hat{\theta}_n$ converge en probabilidad a θ_0 y θ_n se encuentra entre ambos, tenemos que θ_n converge en probabilidad a θ_0. Además, la ley uniforme de los grandes números nos garantiza que

$$L_n''(\theta) \xrightarrow{P_{\theta_0}} E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} f(X_1|\theta)\right],$$

siendo la convergencia uniforme en espacios de parámetros compactos. Por tanto, tomando un compacto que contenga una cola de θ_n obtenemos la mencionada convergencia uniforme. De esta convergencia uniforme se desprende que

$$L_n''(\theta_n) \xrightarrow{P_{\theta_0}} E_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2} f(X_1|\theta_0)\right] = \mathcal{I}_{X_1}(\theta_0).$$

Hemos obtenido pues

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) = \frac{\sqrt{n}L_n'(\theta_0)}{L_n''(\theta_n)} \to N(0, \frac{1}{\mathcal{I}_{X_1}(\theta_0)}).$$ □

Referencias

[1] Proof Wiki, Euler's Reflection Formula, https://proofwiki.org/wiki/Euler%27s_Reflection_Formula.

[2] Wikipedia, Residue theorem, https://en.wikipedia.org/wiki/Residue_theorem#Example.

[3] Wikipedia, Leibniz integral rule, https://en.wikipedia.org/wiki/Leibniz_integral_rule.