Modelos Estadísticos para la Regresión y la Clasificación

Clase 3: Estimadores - Estimador MLE - Descenso por gradiente

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

18 de agosto de 2024

Plan

Glosario

- *Población*: conjunto de toda la información correspondiente a un valor de interés. La identificamos con una variable aleatoria X y distribución $F_X(x) = \mathbb{P}(X \leq x)$
- *Muestra*: un subconjunto de tamaño n de la población observada X. En general consideramos una familia de n variables aleatorias $X_1, \ldots, X_n$ independientes e identicamente distribuidas (iid) con la distribución de X.
- *Individuo*: un elemento de la población estudiada.
- *Parámetro*: valor de interés característico de X o de F_X (la media, la varianza,...)
- *Variables*: propiedad común a los individuos de una poblaci'ón. Diferenciamos:
  - *variables cualitativas*: *nominales*, por ejemplo el color del pétalo, el sexo, el color de los ojos, o *ordinales*, por ejemplo pequeño, mediano, grande. Se pueden codificar.
  - *variables cuantitativas* (numéricas): por ejemplo el largo del pétalo, el peso, el volumen, etc.
    - *continuas*: $\in \mathbb{R}$. Por ejemplo: temperatura, diamétro de un tronco de árbol, etc.
    - *discretas* o *categóricas*: por ejemplo la cantidad de adeptos al fútbol en un grupo de estudiantes.

  los valores observados para las variables son los *datos*.
- *Inferencia Estadística*: proceso mediante al que se llega a conclusiones sobre una población a partir de las observaciones realizadas sobre una muestra de individuos.
- *Estimador*: es una variable aleatoria, función de la muestra aleatoria, $T = T(X_1, X_2, \ldots, X_n)$ que se usa para aproximar el valor teórico del parámetro de interés.
- *Estimación*: la evaluación del estimador en una muestra observada $x_1, \ldots, x_n$ concreta y que proporciona un valor aproximado del valor del parámetro de interés.

Principio clave: se supone que la muestra es elegida al azar entre todos los individuos posibles para elegir.

Muestra, Población, estimador, estimación

- El **muestreo aleatorio** de tamaño n de una población dada por una variable aleatoria X, viene definido por un conjunto de n variables $X_1, \ldots, X_n$ independientes e idénticamente distribuidas con la misma distribución que X.
  Esto se corresponde con la idea más intuitiva de extraer una muestra al azar y con reposición de la población.
- La muestra observada corresponde a una realización concreta del muestreo. Para diferenciarla del muestreo aleatorio denotamos los valores observados por minúsculas:

$$(X_1)_{\text{obs}} = x_1, \ldots, (X_n)_{\text{obs}} = x_n$$

- Un **estimador** de un parámetro θ es una variable aleatoria $\hat{\theta}_n$ es una **función** del muestreo aleatorio: $\hat{\theta}_n = T(X_1, \ldots, X_n)$.
- Una **estimación** de θ es la evaluación de $\hat{\theta}_n$ en la muestra observada, es decir $(\hat{\theta}_n)_{\text{obs}} = T(x_1, \ldots, x_n)$.

Vamos a querer encontrar un buen estimador $T()$, de manera que $\hat{t}heta_n$ sea un valor cercano al verdadero valor θ.

Muestra, Población, estimador, estimación

Población	Estimador (aleatorio): $\hat{\theta}_n$	Estimación (observado): $(\hat{\theta}_n)_{\text{obs}}$
$F(x) = P(X \leq x)$	FD Empírica: $F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i, +\infty)}(x)$	$(F_n)_{\text{obs}}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[x_i, +\infty)}(x)$
$E(X) = \mu$	Media: $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$	$(\bar{X}_n)_{\text{obs}} = \bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i$
$Q(p) = F^{-1}(p)$	Cuantiles: $Q_n(p) = F_n^{-1}(p)$	$(Q_n(p))_{\text{obs}} = q_n(p) = (F_n)_{\text{obs}}^{-1}(p)$
$M = F^{-1}(1/2)$	Mediana: $M_n = F_n^{-1}(1/2)$	$(M)_{\text{obs}} = m_n = (F_n)_{\text{obs}}^{-1}(1/2)$
$var(X) = \sigma^2$	Varianza: $\sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$	$(\sigma_n^2)_{\text{obs}} = \sigma_n^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$

Estimación plug-in

Recordar que si X es una variable aleatoria entonces

$$\mathbb{E}(X) = \int_0^{+\infty} (1 - F(x))\, dx - \int_{-\infty}^{0} F(x)\, dx.$$

Observar que $\mathbb{E}(X)$ es una función de la función de distribución F_X de X por lo que podemos notar $\mathbb{E}(X) = \mathbb{E}(F_X)$. Si cambiamos la función de distribución F_X por su estimador, la distribución empírica F_n, entonces:

$$\begin{aligned}
\mathbb{E}(F_n) &= \int_0^{\infty} (1 - F_n(x))\, dx - \int_{-\infty}^{0} F_n(x)\, dx \\
&= \int_0^{\infty} \left(1 - \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,+\infty)}(x)\right) dx - \int_{-\infty}^{0} \frac{1}{n}\sum_{i=1}^{n} 1_{[X_i,+\infty)}(x)\, dx \\
&= \frac{1}{n}\sum_{i=1}^{n} \int_0^{\infty} \left(1 - 1_{[X_i,+\infty)}(x)\right) dx - \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{0} 1_{[X_i,+\infty)}(x)\, dx \\
&= \frac{1}{n}\sum_{X_i \geq 0} \int_0^{X_i} dx - \frac{1}{n}\sum_{X_i < 0} \int_{X_i}^{0} dx = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n
\end{aligned}$$

Estimación plug-in

Estimador plug-in

Sea $\mathcal{F}$ un conjunto de distribuciones y $T: \mathcal{F} \to \mathbb{R}, F \mapsto T(F)$ una función.
Si $F \in \mathcal{F}$ y $X_1, \ldots, X_n$ una muestra aleatoria de F, y denotemos por F_n la función de distribución empírica asociada. Entonces $T(F_n)$ se llama un estimador plug-in de $T(F)$.

Ejemplos anteriores

Todos los ejemplos anteriores son estimadores plug-in.

Más sobre la función de distribución empírica

Dvoretzky-Kiefer-Wolfowith (1956-1990-2021)

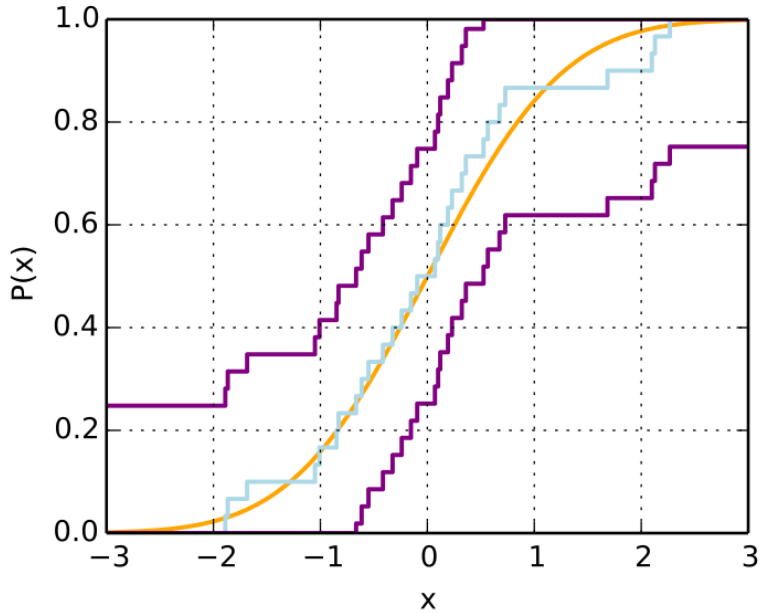
Para todo $\varepsilon > 0$ y todo n

$$P\left(\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \geq \varepsilon\right) \leq 2e^{-2n\varepsilon^2}$$

Bandas de confianza

La verdadera F se encuentra con probabilidad $1-\alpha$ en la siguiente banda

$$F_n(x) - \varepsilon < F(x) < F_n(x) + \varepsilon, \quad \varepsilon = \sqrt{\frac{\ln \frac{2}{\alpha}}{2n}}$$



Fuente: Wikipedia

Plan

Sesgo y varianza de un estimador

- el **sesgo** de un estimador $\hat{\theta}_n$ de θ es la diferencia entre el valor esperado de $\hat{\theta}_n$ y el valor del parámetro considerado:
$$\text{Sesgo}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$$
- un estimador es **insesgado** si $\text{Sesgo}(\hat{\theta}_n) = 0$
- Si $\boldsymbol{\theta} \in \mathbb{R}^p$ es un vector de parámetros y $\hat{\theta}_i$ es un estimador de θ_i para cada $i = 1, \ldots, p$ entonces $\boldsymbol{\theta}$ es **insesgado** si $\mathbb{E}(\hat{\boldsymbol{\theta}}_n) = \mathbb{E}(\boldsymbol{\theta})$.
- un estimador es **asíntoticamente insesgado** si $\text{Sesgo}(\hat{\theta}_n) \to 0$ ($\lim_n \mathbb{E}(\hat{\theta}_n) = \theta$).

Ejemplo: Si $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ independientes, entonces $\overline{X_n}$ es un estimador de μ:

1. $\overline{X}_n$ es un estimador insesgado de μ ya que $\mathbb{E}(\overline{X}_n) = \mu$
2. $\overline{X}_n$ tiene distribución normal ya que la suma de variables aleatorias normales e independientes es también normal.
3. $\text{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$
4. $\overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Estas propiedades van a ser las que querremos tener de los "buenos" estimadores.

Obs: σ_n^2 no es un un estimador insesgado de σ^2, pero si lo es $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ (Ejercicio)

Sesgo y varianza de un estimador

- la **varianza** de un estimador $\hat{\theta}_n$ de θ se define como $\text{Var}(\hat{\theta}_n) = \mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))^2\right]$
- un estimador **eficiente** de θ es un estimador insesgado con la menor varianza. Diremos que cumple con la propiedad de varianza mínima.
- Si $\boldsymbol{\theta} \in \mathbb{R}^p$ es un vector de parámetros y $\hat{\theta}_i$ es un estimador de θ_i para cada $i = 1, \ldots, p$ entonces $\hat{\boldsymbol{\theta}}$ cumple con la propiedad de varianza mínima si $\hat{\theta}_i$ cumple con la propiedad de varianza mínima para todo $i = 1, \ldots, p$.
- un estimador de θ es **asintóticamente eficiente** si en el límite alcanza la menor varianza.

¿Qué es la propiedad de varianza mínima? Veremos que para todo estimador insesgado $\hat{\theta}$ de θ se cumple que

$$\text{Var}(\hat{\theta}) \geq CR(\theta)$$

y cuando la varianza alcanza esta cota inferior, llamada **cota de Cramer Rao**, el estimador se dice entonces eficiente.

Descomposición sesgo varianza de un estimador

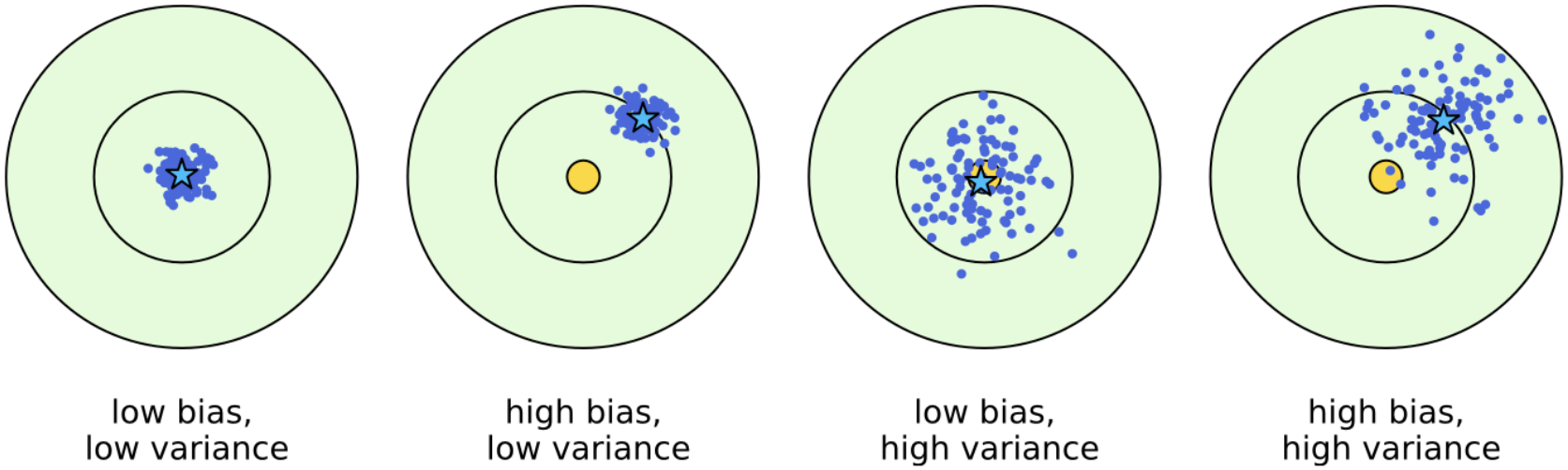


Figure 3: The classic dartboard analogy for explaining bias and variance.

Figura: A Unified Theory of Diversity in Ensemble Learning, Wood et al, 2023

Descomposición sesgo varianza de un estimador

Definición

El error cuadrático medio de un estimador $\hat{\theta}_n$ de θ es

$$ECM(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n - \theta)^2$$

Propiedad

$$ECM(\hat{\theta}_n) = \text{Sesgo}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$$

$$\text{ECM}(\hat{\theta}_n) = \mathbb{E}((\hat{\theta}_n - \theta)^2) = \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n) + \mathbb{E}(\hat{\theta}_n) - \theta\right)^2\right] \tag{1}$$

$$= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2 + 2\left((\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))(\mathbb{E}(\hat{\theta}_n) - \theta)\right) + \left(\mathbb{E}(\hat{\theta}_n) - \theta\right)^2\right] \tag{2}$$

$$= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right] + 2\mathbb{E}\left[(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))(\mathbb{E}(\hat{\theta}_n) - \theta)\right] + \mathbb{E}\left[\left(\mathbb{E}(\hat{\theta}_n) - \theta\right)^2\right] \tag{3}$$

$$= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right] + 2(\mathbb{E}(\hat{\theta}_n) - \theta)\ \overbrace{\mathbb{E}(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n))}^{=\mathbb{E}(\hat{\theta}_n)-\mathbb{E}(\hat{\theta}_n)=0}\ + \mathbb{E}\left[\left(\mathbb{E}(\hat{\theta}_n) - \theta\right)^2\right] \tag{4}$$

$$= \mathbb{E}\left[\left(\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}(\hat{\theta}_n) - \theta\right)^2\right] \tag{5}$$

$$= \text{Var}(\hat{\theta}_n) + \text{Sesgo}(\hat{\theta}_n)^2 \tag{6}$$

Observación: El error cuadrático medio de un estimador insesgado es igual a su varianza, por lo que minimizar el ECM (que puede ser díficil) equivale a minimizar varianza.

Plan

Estimador de mínima varianza

Vamos a querer buscar un estimador insesgado con varianza minima, es decir un $\hat{\theta}$ tal que

- $\mathbb{E}(\hat{\theta}) = \theta$
- $\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\hat{\theta}})$ para todo $\hat{\hat{\theta}}$ estimador insesgado de θ.

Este tipo de estimador se llama estimador insesgado de mínima varianza (MVU: Minimum Variance Unbiased estimator).

Observación: Este estimador puede llegar a no existir. Por ejemplo, si θ es el parámetro a estimar:

$$X \sim \mathcal{N}(\theta, 1) \quad Y \sim \begin{cases} \mathcal{N}(\theta, 1) & \text{si } \theta \geq 0 \\ \mathcal{N}(\theta, 2) & \text{si } \theta < 0 \end{cases}$$

Consideremos como estimadores $\hat{\theta}_1 = \frac{x+y}{2}$ y $\hat{\theta}_2 = \frac{2}{3}x + \frac{1}{3}y$ con x e y realizaciones de X e Y resp.

1. $\hat{\theta}_1$ y $\hat{\theta}_2$ son insesgados.
2. $\text{Var}(\hat{\theta}_1) = \frac{1}{4}\text{Var}(x) + \frac{1}{4}\text{Var}(y) = \begin{cases} 1/2 & \text{si } \theta \geq 0 \\ 3/4 & \text{si } \theta < 0 \end{cases}$

   $\text{Var}(\hat{\theta}_2) = \frac{4}{9}\text{Var}(x) + \frac{1}{9}\text{Var}(y) = \begin{cases} 5/9 & \text{si } \theta \geq 0 \\ 2/3 & \text{si } \theta < 0 \end{cases}$

Si $\theta \geq 0$ entonces $\hat{\theta}_1$ tiene menor varianza y si $\theta < 0$, $\hat{\theta}_2$ tiene menor varianza. No existe el MVU de θ.

Estimador de mínima varianza

Otro ejemplo:
Si $X \sim U[0, 1/\theta]$ siendo $\theta > 0$ y consideramos $\hat{\theta} = T(x)$ un estimador de θ. Supongamos que $\hat{\theta}$ es insesgado, es decir

$$\theta = \mathbb{E}(\hat{\theta}) = \int_0^{1/\theta} T(x) \frac{1}{1/\theta}\, dx = \theta \int_0^{1/\theta} T(x)\, dx$$

por lo que $\int_0^{1/\theta} T(x), dx = 1$ para cualquier valor de θ lo cual no pasa.

- no existe estimador insesgado para θ
- no tiene sentido hablar de MVU

Condición de regularidad

Recordamos la regla de Leibniz:

$$\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} f_\theta(x)\,dx = \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f_\theta(x)\,dx + f_\theta(b(\theta))b'(\theta) - f_\theta(a(\theta))a'(\theta)$$

Esto es porque:

$$\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} f_\theta(x)\,dx = \frac{\partial}{\partial\theta}\left[F_\theta(x)|_{a(\theta)}^{b(\theta)}\right] = \frac{\partial}{\partial\theta}\left(F_\theta(b(\theta)) - F_\theta(a(\theta))\right)$$

$$= \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial\theta} f_\theta(x)\,dx + f(b(\theta))b'(\theta) - f(a(\theta))a'(\theta)$$

Por lo tanto los ordenes de integración y de derivación se pueden intercambiar si los límites de integración no dependen de θ:

Condición de Regularidad

$$\frac{\partial}{\partial\theta}\int_{a}^{b} f_\theta(x)\,dx = \int_{a}^{b} \frac{\partial}{\partial\theta} f_\theta(x)\,dx$$

Cota de Cramer-Rao

Si $X \sim f_\theta$, x una realización de X y supongamos que $T(x)$ es un estimador insesgado de θ, entonces derivando respecto de θ la expresión:

$$\theta = \mathbb{E}(\hat{\theta}) = \int T(x) f_\theta(x)\, dx$$

y aplicando la condición de regularidad $\frac{\partial}{\partial \theta} \int T(x) f_\theta(x)\, dx = \int T(x) \frac{\partial}{\partial \theta} f_\theta(x)\, dx$ obtenemos

$$1 = \int T(x) f_\theta(x) \frac{\partial}{\partial \theta} \log f_\theta(x)\, dx$$

pues $\frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{\frac{\partial}{\partial \theta} f_\theta(x)}{f_\theta(x)}$.
Por otro lado $\theta \int f_\theta(x) \frac{\partial}{\partial \theta} \log f_\theta(x)\, dx = 0$ pues $\int f_\theta(x) \frac{\partial}{\partial \theta} \log f_\theta(x)\, dx = 0$ por la condición de regularidad, por lo que tenemos que

$$1 = \int f_\theta(x) \frac{\partial}{\partial \theta} \log f_\theta(x)(T(x) - \theta)\, dx$$

Elevando al cuadrado y usando la desigualdad de Cauchy-Schwarz [1], tenemos

$$1 = \mathbb{E}\left(\frac{\partial}{\partial \theta} \log f_\theta(x)(T(x) - \theta)\right)^2 \leq \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f_\theta(x)\right)^2\right] \mathbb{E}\left[((T(x) - \theta))^2\right]$$

[1] $\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$ y vale la igualdad sii $X = \alpha Y$ con α constante (ver Apéndice)

Cota de Cramer-Rao

La desigualdad anterior se convierte en

$$\mathbb{E}\left[((T(x)-\theta))^2\right] \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2\right]}$$

es decir

$$\text{Var}(T(x)) \geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_\theta(x)\right)^2\right]} = \frac{1}{-\mathbb{E}\left(\frac{\partial^2}{\partial\theta^2}\log f_\theta(x)\right)}$$

Recordar que vale la igualdad si y solo si $\frac{\partial}{\partial\theta}\log f_\theta(x) = I(\theta)(T(x)-\theta)$

Un estimador insesgado que alcanza la cota inferior de la desigualdad de Cramer-Rao se dice que es **eficiente** u **óptimo**.

Conclusión:

Si podemos factorizar $\frac{\partial}{\partial\theta}\log f_\theta(x)$ como $I(\theta)(T(x)-\theta)$ entonces $g(x)$ es este estimador eficiente.

Veremos más adelante que $I(\theta)$ es la información de Fischer.

Ejemplo

Si $X \sim \mathcal{N}(\mu, \sigma^2)$ con σ^2 conocido, un estimador insesgado de μ es cualquier realización x de X y su varianza es σ^2. Recordar que $\frac{\partial}{\partial \mu} \log f_\mu(x) = \frac{1}{\sigma^2}(X - \mu)$ y $\frac{\partial^2}{\partial \mu^2} \log f_\mu(x) = -\frac{1}{\sigma^2}$ por lo que la desigualdad de Cramer Rao es

$$\text{Var}(x) \geq \frac{1}{-(-1/\sigma^2)} = \sigma^2$$

para cualquier estimador insesgado de μ, por lo cual X es el MVU.

Observar que $\frac{\partial}{\partial \mu} \log f_\mu(x) = \frac{1}{\sigma^2}(X - \mu)$ por lo que la factorización nos indica que X es el MVU.

Plan

Consistencia de un estimador

Recordar que un estimador $\hat{\theta}_n$ de θ se basa sobre una muestra de tamaño n y por lo tanto depende de n.

Decimos que un estimador $\hat{\theta}_n$ de θ es **consistente** si para todo $\epsilon > 0$, $\lim_{n \to +\infty} \mathbb{P}(|\hat{\theta}_n - \theta| < \epsilon) = 1$

- Recordemos la desigualdad de Markov: Si X es una variable aleatoria positiva entonces para todo $\epsilon > 0$ vale que $\mathbb{P}(X \geq \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon}$
- Recordemos la desigualdad de Chebyshev: Si $\hat{\theta}_n$ es un estimador de θ entonces para todo $\epsilon > 0$ vale que $\mathbb{P}(|\hat{\theta}_n - \theta| \geq \epsilon) \leq \frac{ECM(\hat{\theta}_n)}{\epsilon^2}$
- Por lo tanto un criterio para probar que un estimador es consistente es probar que su error cuadrático medio tiende a cero cuando n tiende a infinito.
- Un estimador con error cuadrático medio que tiende a cero es asintóticamente insesgado. Esto es por que:

$$\left|\mathbb{E}(\hat{\theta}_n) - \theta\right| \leq \mathbb{E}\left(\left|\hat{\theta}_n - \theta\right|\right) \leq \sqrt{\mathbb{E}((\hat{\theta}_n - \theta)^2)} \to 0, \text{ si } n \to +\infty$$

- Si $\hat{\theta}_n$ es consistente para θ y g es continua entonces $g(\hat{\theta}_n)$ es consistente para $g(\theta)$

Plan

Método de máxima verosimilitud

Consideremos $X_1, \ldots, X_n$ una muestra i.i.d con función de distribución F_X la cual depende de un parámetro θ.
La **función de verosimilitud** (likelihood en inglés) de θ es la probabilidad de que halla ocurrido la muestra.

$$L_n(\theta) = \mathbb{P}_\theta(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}_\theta(X_1 = x_1) = \prod_{i=1}^{n} p_\theta(x_1)$$

siendo p_θ la densidad de X.

¿Cuál es la idea de la verosimilitud?

1. Dado que observamos la variable aleatoria, nos preguntamos qué tan probables son los distintos valores de θ.
2. Esto lo cuantificamos mediante la función de verosimilitud: si $L_n(\theta_1) > L_n(\theta_2)$ entonces la generación de las observaciones es más probable si tomamos $\theta = \theta_1$ que si tomamos $\theta = \theta_2$
3. Por ejemplo si tenemos $X \sim \mathcal{N}(\theta, 1)$ y tengo tres densidades normales $\mathcal{N}(-3, 1), \mathcal{N}(-2, 1)$ y $\mathcal{N}(3, 1)$ y observo $X = 2$, si nos tenemos que decidir por alguna de las tres, lo más probable es que θ valga 3.
4. La función de verosimilitud no es la probabilidad de θ, es la probabilidad de θ dada la muestra (aleatoria).

Método de máxima verosimilitud

El **estimador de máxima verosimilitud (MLE)** de θ es el valor de θ que maximiza la función $L_n(\theta)$, es decir

$$\hat{\theta}_{MLE,n} = \underset{\theta}{\text{Argmax}}\, L_n(\theta)$$

Resulta a veces más sencillo para hacer las cuentas utilizar el logaritmo de la función de verosimilitud y considerar $\ell_n(\theta) = \sum_{i=1}^{n} \log p_\theta(x_i)$, pues al ser log creciente se tiene

$$\hat{\theta}_{MLE,n} = \underset{\theta}{\text{Argmax}}\, L_n(\theta) = \underset{\theta}{\text{Argmax}}\, \ell_n(\theta).$$

Método de máxima verosimilitud - Ejemplo 1 (Bernoulli)

Ejemplo 1:
Consideramos una muestra i.i.d $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$ y queremos estimar $\theta = p \in (0,1)$.

El logaritmo de la función de verosimilitud es

$$\ell_n(p) = \sum_{i=1}^{n} X_i \log(p) + (1 - X_i) \log(1-p),$$

Derivando se obtiene que:

$$\ell_n'(p) = n \left[\frac{\overline{X}_n}{p} + \frac{1 - \overline{X}_n}{1-p} \right]$$

y

$$\ell_n'(p) = 0 \Leftrightarrow \widehat{p}_n = \overline{X}_n$$

Como $\ell_n''(p) = -n \left[\frac{\overline{X}_n}{p^2} + \frac{1-\overline{X}_n}{(1-p)^2} \right] < 0$ para todo $p \in (0,1)$, este punto crítico es un máximo.

Entonces el estimador por el método de máxima verosimilitud de p es el promedio $\hat{\theta}_{MLE,n} = \overline{X}_n$.

Método de máxima verosimilitud - Ejemplo 2 (Normal)

Ejemplo 2: Consideramos una muestra i.i.d $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, con μ y σ^2 que queremos estimar por el método de máxima verosimilitud. El parámetro a considerar es el vector $\theta = (\mu, \sigma^2)$.

La función de verosimilitud es

$$L_n(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-(X_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(X_i-\mu)^2}$$

y por lo tanto la logverosimilitud:

$$\ell_n(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Derivando respecto de μ y de σ^2:

$$\frac{\partial \ell_n}{\partial \mu}(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu); \quad \frac{\partial \ell_n}{\partial \sigma^2}(\mu, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2$$

Método de máxima verosimilitud - Ejemplo 2 (Normal)

Se tiene entonces:

$$\begin{cases} \frac{\partial \ell_n}{\partial \mu}(\mu, \sigma^2) = 0 \\ \frac{\partial \ell_n}{\partial \sigma^2}(\mu, \sigma^2) = 0 \end{cases} \Leftrightarrow \begin{cases} \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu) = 0 \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n (X_i - \mu)^2 = 0 \end{cases} \Leftrightarrow \begin{cases} \widehat{\mu} = \overline{X}_n \\ \widehat{\sigma}^2 = S_n^2 \end{cases}$$

Para verificar que este punto crítico es un máximo miramos la Hessiana:

$$H(\ell_n) = \begin{pmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \sigma^2 \partial \mu} \\ \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \ell}{\partial (\sigma^2)^2} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^n (X_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^n (X_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (X_i - \mu)^2 \end{pmatrix}$$

y la evaluamos en $\overline{X}_n$ y S_n^2. Como $H(\ell_n) = \begin{pmatrix} -\frac{n}{S_n^2} & 0 \\ 0 & -\frac{n}{2S_n^4} \end{pmatrix}$ es definida negativa, podemos concluir que $\hat{\boldsymbol{\theta}}_{MLE,n} = (\hat{\mu}, \hat{\sigma}^2)$ es el estimador de máxima verosimilitud de θ.

Teorema 1

Sea θ^* el parámetro de la distribución de las observaciones $X = \{X_i\}_{i=1}^n$ iid. Entonces

$$\forall\, \theta \neq \theta^* : \quad \mathbb{P}\left(L(\theta^*) > L(\theta)\right) \overset{\text{prob}}{\to} 1$$

Teorema 2: consistencia del estimador MLE

$$\hat{\theta}_{MLE,n} \overset{\text{prob}}{\to} \theta^*$$

Teorema del principio de invarianza

Si $\hat{\theta}$ es el estimador de máxima verosimilitud de θ, entonces $g(\hat{\theta})$ es el estimador de máxima verosimilitud de $g(\theta)$.

Teorema

Sea θ^* el parámetro de la distribución de las observaciones $X = \{X_i\}_{i=1}^n$ iid. Entonces

$$\forall\, \theta \neq \theta^* : \quad \mathbb{P}\left(L(\theta^*) > L(\theta)\right) \overset{\text{prob}}{\to} 1$$

Demostración: Vamos a mirar el cociente

$$\frac{\ell(\theta)}{\ell(\theta^*)} = \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}$$

y evaluar la probabilidad para que $\frac{\ell\theta}{\ell\theta^*} < 0$, lo cual equivale a

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}$$

Definimos $Y_i = \log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}$, estas variables son independientes y por lo tanto por la ley de los grandes números el promedio converge a la esperanza $\mathbb{E}\left[\log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}\right]$.
Por la desigualdad de Jensen se tiene que

$$\mathbb{E}\left[\log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}\right] < \log\left(\mathbb{E}\left[\frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}\right]\right)$$

Teorema 1

Como

$$\log\left(\mathbb{E}\left[\frac{f_\theta(X_i)}{f_{\theta^*}(X_i)}\right]\right) = \log\left(\int f_{\theta^*}(x)\frac{f_\theta(x)}{f_{\theta^*}(x)}\,dx\right) = \log(1) = 0$$

se tiene que existe μ tal que

$$\overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{1}{n}\sum_{i=1}^{n} \log\frac{f_\theta(X_i)}{f_{\theta^*}(X_i)} \to \mu < 0$$

Si $\epsilon = \frac{|\mu|}{2}$ entonces $\overline{Y}_n$ es negativo ya que por la ley de los grandes números tenemos

$$\mathbb{P}(|\overline{Y}_n - \mu| \leq \epsilon) \overset{\text{prob}}{\to} 1$$

con lo que se puede garantizar que $\ell(\theta^*) > \ell(\theta)$ con probabilidad que tiende a 1 cuando n es grande.

Teorema 2: consistencia del MLE

Recordamos del teorema anterior que $\forall\, \theta \neq \theta^* : \quad \mathbb{P}\left(\ell(\theta^*) > \ell(\theta)\right) \overset{\text{prob}}{\to} 1$
Vamos a mostrar que hay una secuencia de $\hat{\theta}_{MLE,n}$ que converge hacia θ^*. Sea $a > 0$ y consideramos el conjunto de observaciones S_n definido por

$$S_n = \{x : \ell(\theta^*, x) > \text{máx}(\ell(\theta^* - a, x), \ell(\theta^* + a, x)$$

Sabemos que por el teorema anterior

$$\mathbb{P}(S_n) \overset{\text{prob}}{\to} 1$$

es decir que casi todas las observaciones van a pertenecer a S_n.
En el intervalo $[\theta^* - a, \theta^* + a]$, como $\ell(\theta)$ es derivable y por lo tanto continua, por el teorema de Rolle, existe un valor de θ que anula $\frac{\partial}{\partial\theta}\ell(\theta)$ y la notamos por $\hat{\theta}_n$. Definimos el conjunto

$$\tilde{S}_n = \{x : \exists \hat{\theta}_n \text{ tal que } \frac{\partial}{\partial\theta}\ell(\hat{\theta}_n, x) = 0 \text{ y } ||\hat{\theta}_n - \theta^*|| < a\}$$

Entonces $S_n \subset \tilde{S}_n$ y por lo tanto $\mathbb{P}(S_n) \leq \mathbb{P}(\tilde{S}_n)$ por lo que

$$\mathbb{P}(\tilde{S}_n) \overset{\text{prob}}{\to} 1$$

Entonces para todos $a > 0$ en probabilidad se encuentra un $\hat{\theta}_n$ cercano a θ^* y realizando ésto para cada n, se encuentra una sucesión que converge hacia θ^*.

MLE y divergencia de Kullback-Leibler

Consideramos acá el parámetro verdadero θ^* y el estimador de máxima verosimilitud $\hat{\theta}_{MLE,n}$.

Divergencia de Kullback-Leibler

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx$$

Si bien se verifica que $D_{KL}(p||q) \geq 0$ y que $D_{KL}(p||q) = 0 \Leftrightarrow p = q$, no es una distancia ya que $D_{KL}(p||q||) \neq D_{KL}(q||p)$

La búsqueda del MLE está asociada a la **divergencia de Kullback-Leibler** ya que si $f_{\theta*}$ es la verdadera densidad entonces

$$D_{KL}(f_{\theta*}||f_\theta) = \mathbb{E}_{f_{\theta*}}[\log f_{\theta*}] - \mathbb{E}_{f_{\theta*}}[\log f_\theta]$$

y por lo tanto maximizar la verosimilitud equivale a minimizar la divergencia de Kullback-Leibler

Información de Fisher

Vamos a notar por $\ell(\theta) = l_1(\theta) = \log f_\theta(x)$ (x es una observación aleatoria).

- A la función $s(\theta, x) = \frac{\partial}{\partial \theta} \log f_\theta(x) = \frac{\partial}{\partial \theta} \ell(\theta)$ se le suele llamar *función de score* de θ. Esta función se anula cuando $\theta = \hat{\theta}_{MLE}$ ($s(\hat{\theta}_{MLE}, x) = 0$) y $\ell(\hat{\theta}_{MLE})$ es el máximo.
  Con esta función se visualiza los cambios de pendiente de la función de logverosimilitud
- Si $\theta = \theta^*$ (f_{θ^*} es la verdadera densidad) entonces

$$\mathbb{E}_{\theta^*}\left[\frac{\partial}{\partial \theta}\ell(\theta)\Big|_{\theta=\theta^*}\right] = \mathbb{E}_{x\sim f_{\theta^*}}\left[s(\theta^*, x)\right] = 0$$

ya que:

$$\mathbb{E}_{x\sim f_{\theta^*}}\left[s(\theta^*, x)\right] = \int \frac{\frac{\partial f_\theta}{\partial \theta}(x)\Big|_{\theta=\theta^*}}{f_{\theta^*}(x)} f_{\theta^*}(x)\, dx = \frac{\partial}{\partial \theta} \underbrace{\int f_{\theta^*}(x)\, dx}_{1}\Bigg|_{\theta=\theta^*} = 0$$

Acá hicimos el supuesto que derivada e integral son intercambiables.

Información de Fisher

La información de Fisher es una forma de medir la cantidad de información que una variable aleatoria observable X contiene respecto a un parámetro desconocido θ de su función de densidad.

Definición

La información de Fisher es la varianza del score $s(\theta, x) = \frac{\partial}{\partial \theta} \log f_\theta(x)$ en $\theta = \theta^*$, es decir

$$I(\theta^*) = \mathbb{E}_{x \sim f_{\theta^*}} \left[\left(\frac{\partial \log f_\theta(x)}{\partial \theta} \Big|_{\theta=\theta^*} \right)^2 \right] = \text{Var}_{x \sim f_{\theta^*}} \left[s(\theta^*, x) \right] = \text{Var}(\ell'(\theta^*))$$

Una variable aleatoria con información de Fisher elevada implica una gran sensibilidad a las variaciones de la estimación por máxima verosimilitud cuando se muestrea.

Propiedad

La información de Fisher está relacionada con la curvatura de la logverosimilitud en θ^*

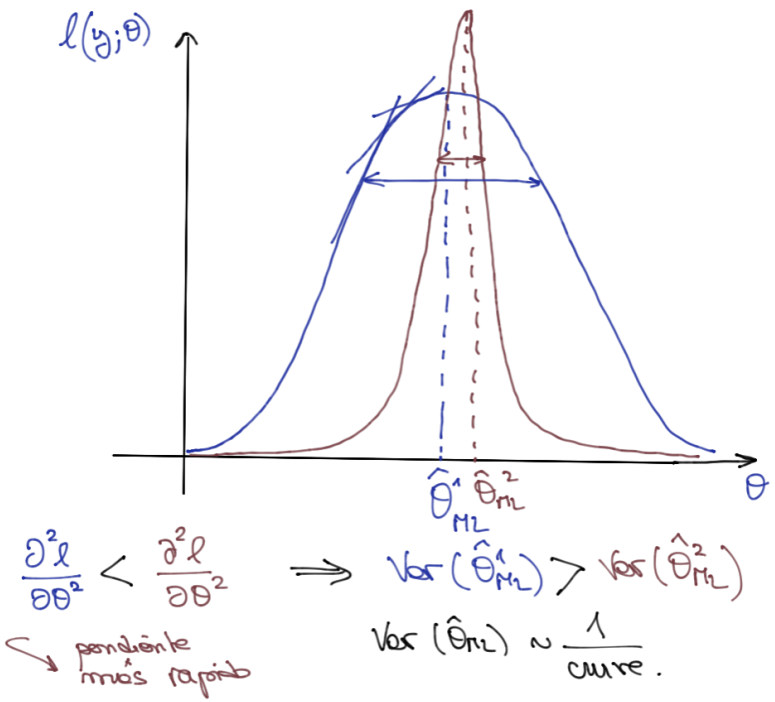
$$I(\theta^*) = -\mathbb{E}_{x\sim f_{\theta^*}}\left[\left.\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2}\right|_{\theta=\theta^*}\right] = -\mathbb{E}(\ell''(\theta^*))$$

$$\begin{aligned}
\mathbb{E}_{x\sim f_{\theta^*}}\left[\left.\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2}\right|_{\theta=\theta^*}\right] &= \int f_{\theta^*}(x) \times \left.\frac{\partial^2 \log f_\theta(x)}{\partial\theta^2}\right|_{\theta=\theta^*} dx \\
&= \int \left[-\frac{1}{f_{\theta^*}}(f'_\theta(\theta^*))^2 + f''_\theta(\theta^*)\right] dx \\
&= -\underbrace{\int f_{\theta^*}\left(\left.\frac{\partial \log f_\theta(x)}{\partial\theta}\right|_{\theta=\theta^*}\right)^2 dx}_{I(\theta^*)} + \underbrace{\int f''_\theta(\theta^*)dx}_{(\int f_\theta\, dx)''_{\theta=\theta^*}}
\end{aligned}$$

Acá hicimos el supuesto que podemos derivar dos veces debajo de la integral

Interpretación geométrica

- Observar que : $\ell''(\theta) = \frac{\partial}{\partial\theta}\left(\frac{\partial\ell}{\partial\theta}\right) = \frac{\partial}{\partial\theta}$ (gradiente)
  Intuitivamente: el gradiente a medida que θ crece cerca del máximo decrece y por lo tanto $\ell''(\theta) = \frac{\partial}{\partial\theta}$ (gradiente) < 0
  Valores negativos grandes de la derivada segunda de la función de verosimilitud indican que el estimador MLE está cerca del valor real.
- Podemos entonces mirar si este máximo es más o menos estrecho utilizando la curvatura de la función de verosimilitud la cual esta relacionada con la variabilidad de las pendientes cercanas (tasa de cambio de la derivada). Cuanto más grande es la curvatura, más cerca del valor real vamos a estar.



Aditividad de la información de Fisher

Si $x_1, \ldots, x_n$ son iid podemos escribir

$$\frac{\partial log f_\theta(x_1, \ldots, x_n)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial log f_\theta(x_i)}{\partial \theta}$$

y como $\frac{\partial log f_\theta(x_i)}{\partial \theta}$ son iid, la varianza se suma por lo que tenemos:

Aditividad información de Fisher

$$I(\theta^*, x_1, \ldots, x_n) = nI(\theta^*, x_i) = nI(\theta^*)$$

Desigualdad de Cramer Rao

Si $\hat{\theta}_n$ es un estimador insesgado de θ entonces

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}$$

Demostración: (es esencialmente la misma que vimos antes)

- Si $\hat{\theta}_n$ es un estimador insesgado de θ obtenido a partir de una muestra iid $X_1, \ldots, X_n$ con densidad f_θ, para cada $i = 1, \ldots, n$ definimos la variable aleatoria

$$Y_i = \frac{\partial \log f_\theta}{\partial \theta}(X_i) = \frac{1}{f_\theta(X_i)} \frac{\partial}{\partial \theta} f_\theta(X_i)$$

  Entonces $\mathbb{E}(Y_i) = \mathbb{E}(\ell'(\theta)) = 0$ para todo $i = 1, \ldots, n$

- Al ser $\hat{\theta}_n$ insesgado, se tiene que $\theta = \mathbb{E}(\hat{\theta}_n)$ y derivando respecto de θ se obtiene que:

$$1 = \int \hat{\theta}_n \sum_{i=1}^{n} \frac{\partial f_\theta}{\partial \theta}(x_i) \prod_{j \neq i} f_\theta(x_j) dx_1 \ldots dx_n$$

$$= \int \hat{\theta}_n \sum_{i=1}^{n} \frac{1}{f_\theta(x_i)} \frac{\partial}{\partial \theta} f_\theta(x_i) \prod_{j=1}^{n} f_\theta(x_j)\, dx_1 \ \ldots, dx_n = \mathbb{E}\left(\hat{\theta}_n \sum_{i=1}^{n} Y_i\right)$$

Desigualdad de Cramer Rao

- Por otro lado $\text{Cov}\left(\hat{\theta}_n, \sum_{i=1}^n Y_i\right) = \mathbb{E}\left(\hat{\theta}_n \sum_{i=1}^n Y_i\right) - \underbrace{\mathbb{E}\left(\hat{\theta}_n\right)\mathbb{E}\left(\sum_{i=1}^n Y_i\right)}_{=0} = 1$ y el cuadrado del coeficiente de correlación de Pearson es:

$$\rho^2 = \frac{\text{Cov}\left(\hat{\theta}_n, \sum_{i=1}^n Y_i\right)}{\text{Var}(\hat{\theta}_n)\text{Var}\left(\sum_{i=1}^n Y_i\right)} = \frac{1}{\text{Var}(\hat{\theta}_n)\text{Var}\left(\sum_{i=1}^n Y_i\right)} \leq 1$$

- Por lo tanto

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{\text{Var}\left(\sum_{i=1}^n Y_i\right)}$$

- Como $Y_1, \ldots, Y_n$ son independientes, se tiene que

$$\text{Var}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \text{Var}\left(Y_i\right) = \sum_{i=1}^n \mathbb{E}(Y_i^2) = n\mathbb{E}(\ell'(\theta)^2) = nI(\theta)$$

por lo que finalmente

$$\text{Var}(\hat{\theta}_n) \geq \frac{1}{nI(\theta)}$$

Desigualdad de Cramer Rao I

- La información de Fisher nos da una cota mínima sobre la capacidad de estimar el parámetro θ a partir de las observaciones.
- Si $\hat{\theta}_n$ es un estimador que cumple la igualdad $\text{Var}(\hat{\theta}_n) = \frac{1}{nI(\theta)}$, decimos que $\hat{\theta}_n$ es un estimador **eficiente** u **óptimo**: tiene la menor varianza posible entre todos los estimadores insesgados de θ.

Desigualdad de Cramer Rao II

Si $\hat{\theta}_n$ es un estimador θ con $\mathbb{E}_{x\sim f_\theta}(\hat{\theta}_n) = \tau(\theta)$ entonces

$$\text{Var}(\hat{\theta}_n) \geq \frac{|\tau'(\theta)|^2}{nI_1(\theta)}$$

Ejemplo

Ejemplo: Si $X_1, \ldots, X_n$ es una muestra iid de una variable aleatoria $X \sim Ber(p)$, vimos que el estimador MLE de p es $\overline{X}_n$.

1. Como $\mathbb{E}(\overline{X}_n) = E(X_1) = p$ se tiene que $\overline{X}_n$ es insesgado
2. Por la ley fuerte de los grandes números $\overline{X}_n$ converge en probabilidad a $\mathbb{E}(X) = p$ y por lo tanto es consistente.
3. Como la densidad se puede escribir como $f_p(x) = p^x(1-p)^{1-x}$ con $x \in \{0, 1\}$ entonces $\ell(p) = x \ln p + (1-x)\ln(1-p)$ y entonces

$$\ell'(p) = \frac{x}{p} - \frac{1-x}{1-p}; \quad \ell''(p) = -\frac{x}{p^2} + \frac{1-x}{(1-p)^2}$$

$$-\mathbb{E}(\ell''(p)) = -\left(-\frac{p}{p^2} + \frac{1-p}{(1-p)^2}\right) = \frac{1}{p(1-p)}$$

4. Por lo tanto

$$\text{Var}(\overline{X}_n) = \frac{1}{n}\text{Var}(X_1) = \frac{p(1-p)}{n} = \frac{1}{nI(p)}$$

y $\overline{X}_n$ es un estimador óptimo para p.

Otras propiedades del MLE

Vamos a verlas de manera empírica y a partir de una simulación. Consideramos la densidad siguiente:

$$p(x;\theta) = \frac{1}{2}(1+\theta x) \quad -1 \leq x \leq 1$$

con θ un parámetro desconocido que varía entre -1 y 1.

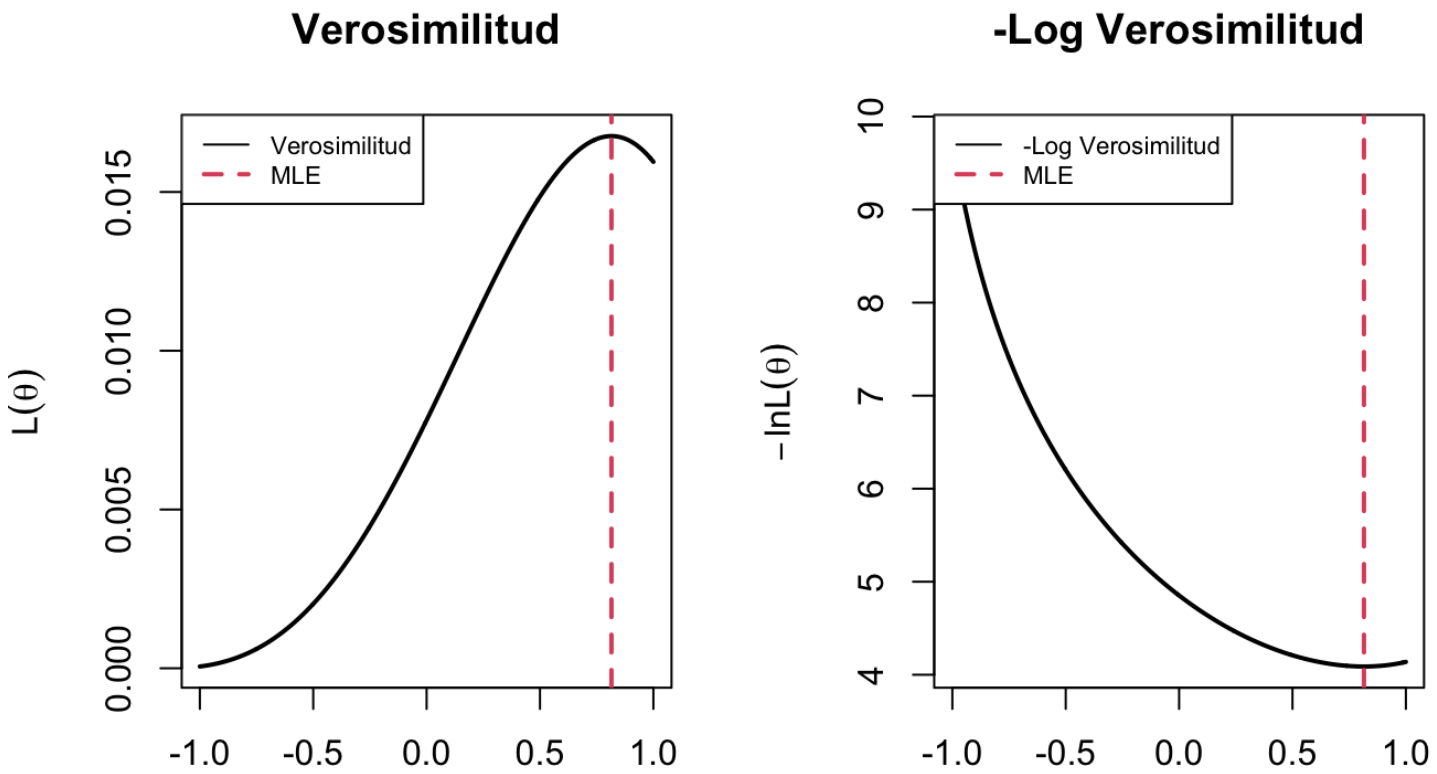
Supongamos que θ vale 0.5. Vamos a querer en primer lugar obtener datos a partir de esta distribución y obtener el MLE $\hat{\theta}_n$ de θ a partir de ellos dado que calcular a mano el estimador MLE no es sencillo.
Veamos a continuación un mecanismo para generar datos mediante el método de aceptación/rechazo.

Otras propiedades del MLE

Hemos obtenido 7 datos: $-0.47, 0.89, 0.26, -0.59, 0.37, 0.54$ y 0.87. La función de logverosimilitud es

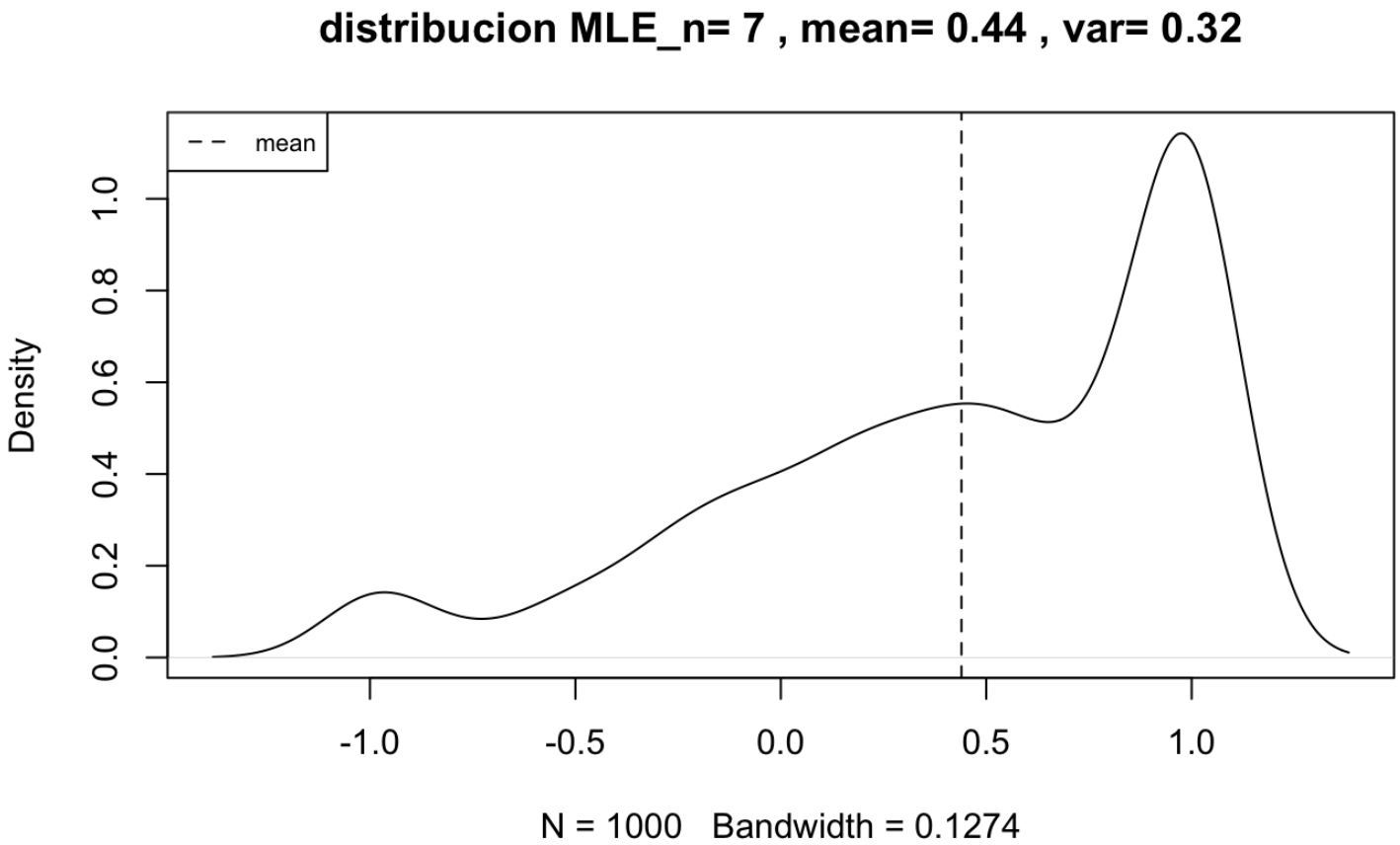
$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{2}(1+\theta x_i)\right) = n\log 2 + \sum_{i=1}^{n} \log\left((1+\theta x_i)\right)$$

Para la muestra dada, obtenemos que el valor que maximiza esta función es $\hat{\theta}_7 = 0.82$. En la figura siguientes, las gráficas:



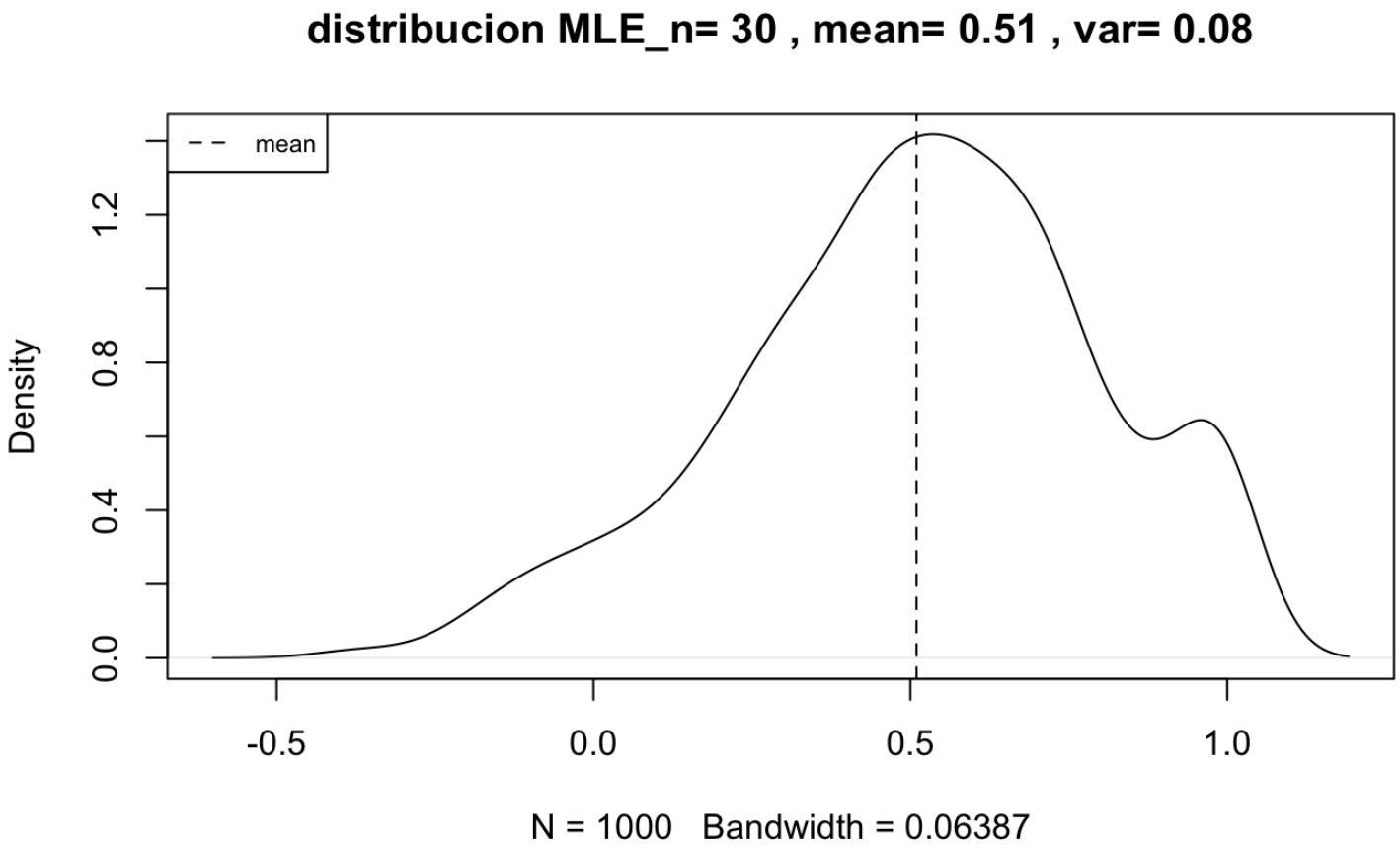
Otras propiedades del MLE

Ahora simulamos $N = 1000$ muestras de tamaño $n = 7$ y vemos la distribución del MLE El promedio es de 0.44 y la varianza 0.32



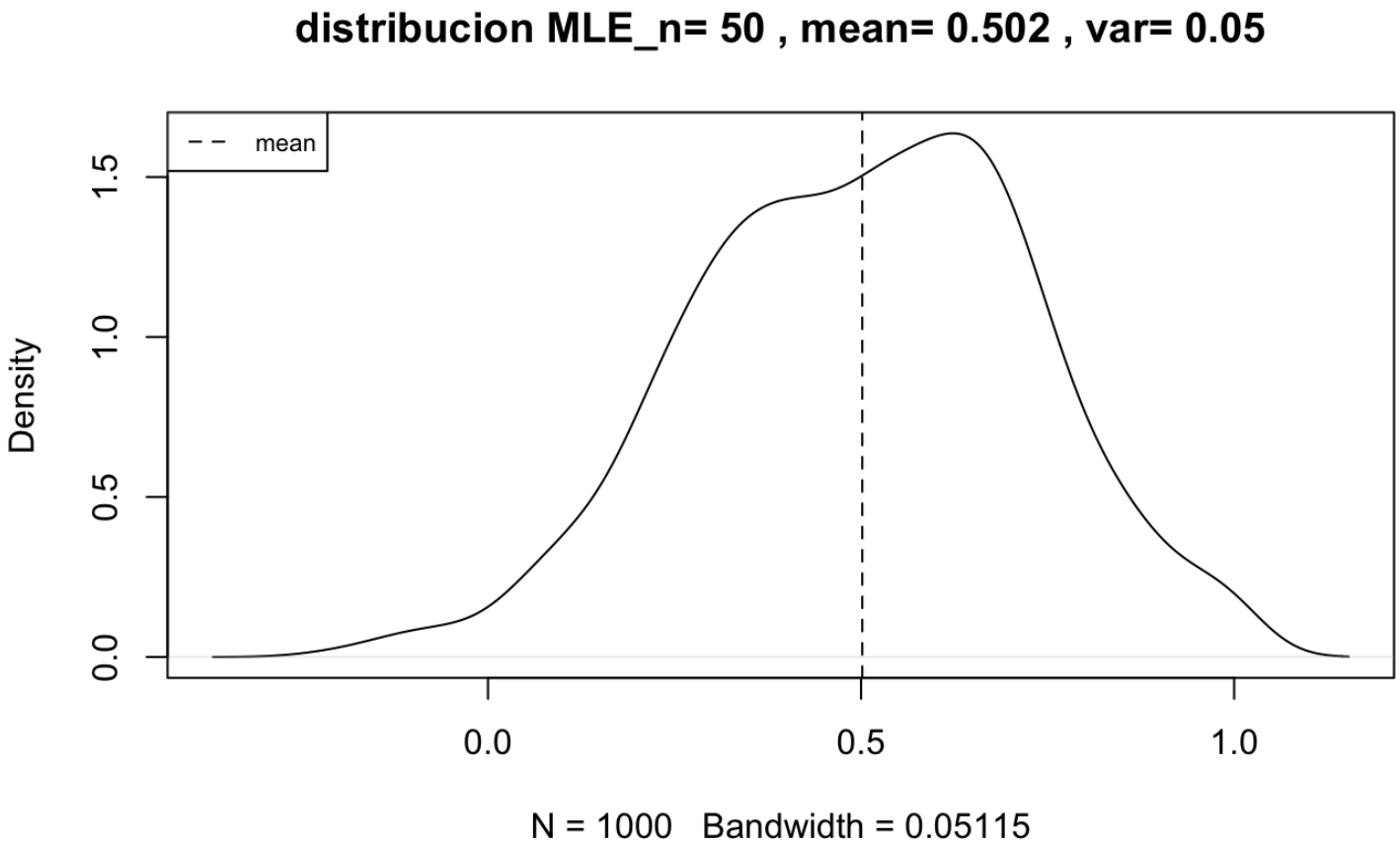
Otras propiedades del MLE

Cambiemos ahora el tamaño de la muestra a $n = 30$ y volvamos a simular $N = 1000$ muestras de tamaño $n = 30$ de manera independiente y vemos la distribución del MLE El promedio es de 0.51 y la varianza 0.05



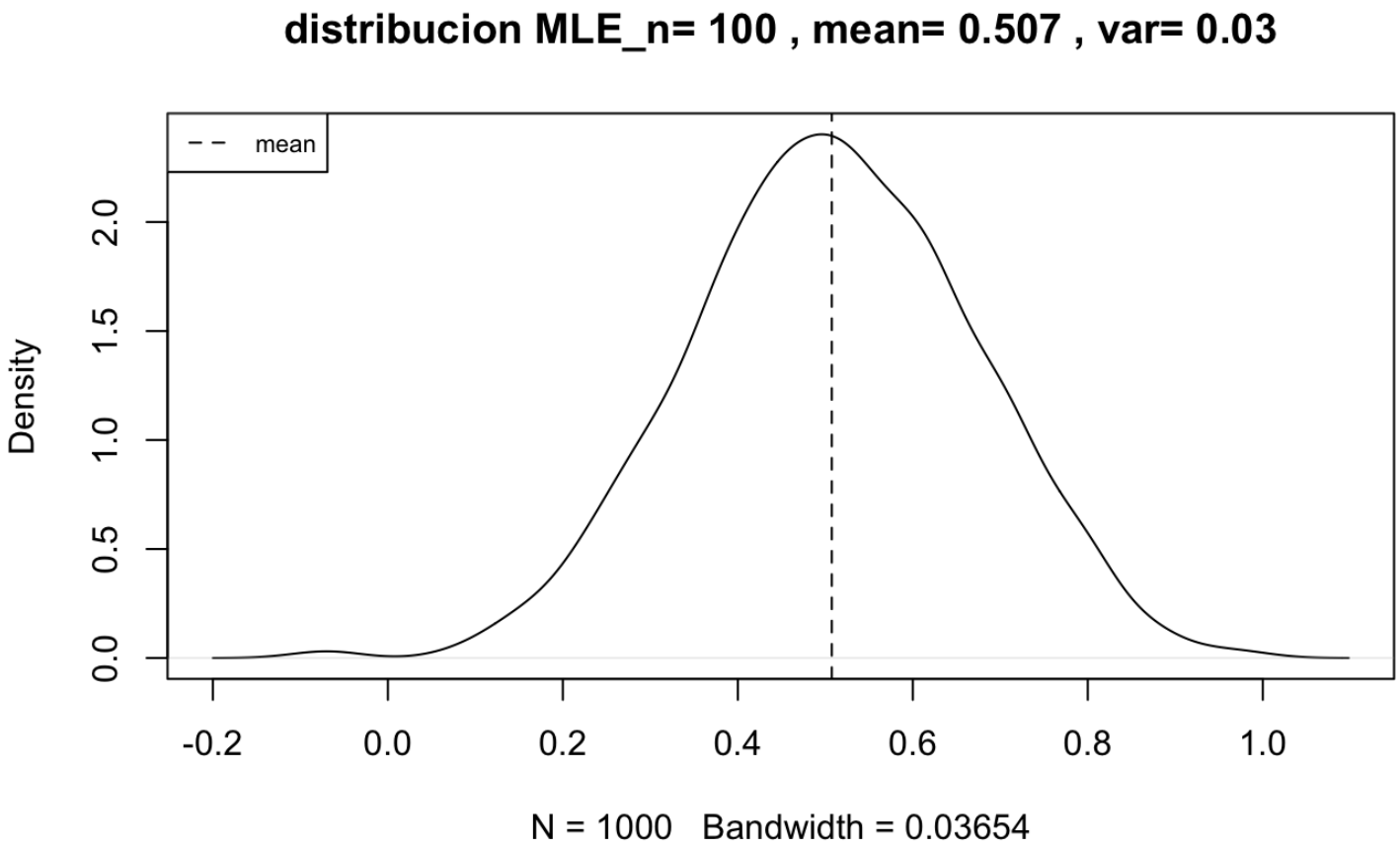
Otras propiedades del MLE

Cambiemos ahora el tamaño de la muestra a $n = 50$ y volvamos a simular $N = 1000$ muestras de tamaño $n = 50$ de manera independiente y vemos la distribución del MLE



Otras propiedades del MLE

Cambiemos ahora el tamaño de la muestra a $n = 100$ y volvamos a simular $N = 1000$ muestras de tamaño $n = 100$ de manera independiente y vemos la distribución del MLE El promedio es de 0.51 y la varianza 0.05



Otras propiedades del MLE

Observamos que a medida que n aumenta:

- Sesgo $(\hat{\theta}_{MLE,n}) \to 0$ (el MLE es asíntoticamente insesgado)
- Var $(\hat{\theta}_{MLE,n}) \to 0$

y por lo tanto $ECM(\hat{\theta}_{MLE,n}) \to 0$, es decir $\hat{\theta}_{MLE,n}$ es consistente.

Además la forma de la distribución de $\hat{\theta}_{MLE,n}$ se parece cada vez más a una normal. Se dice que el MLE es asíntoticamente normal.

Teorema normalidad asíntotica del MLE

Si la función de densidad $f_\theta(x)$ es derivable dos veces y cumple con la condición de regularidad $\mathbb{E}\left(\frac{\partial \log f_\theta(x)}{\partial \theta}\right) = 0, \forall \theta$ entonces la distribución de $\hat{\theta}_{MLE,n}$ tiende a:

$$\hat{\theta}_{MLE,n} \sim \mathcal{N}(\theta, I^{-1}(\theta))$$

siendo $I(\theta)$ la información de Fischer de θ.

El caso multivariado

Sea $\boldsymbol{\theta} \in \Theta$ un vector de parámetros. Su estimador de máxima verosimilitud es

$$\hat{\boldsymbol{\theta}}_{MLE,n} = \underset{\boldsymbol{\theta}\in\Theta}{\text{Argmax}} f(x, \boldsymbol{\theta})$$

La matriz de información de Fisher es $[\boldsymbol{I}(\boldsymbol{\theta})]_{ij} = -\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial^2 \log f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial\theta_i \partial\theta_j}\right]$

Teorema: normalidad asíntotica multivariada del MLE

Si la función de densidad $f_{\boldsymbol{\theta}}(x)$ es diferenciable dos veces y cumple con la condición de regularidad $\mathbb{E}\left(\frac{\partial \log f_{\boldsymbol{\theta}}(x)}{\partial \boldsymbol{\theta}}\right) = \mathbf{0}, \forall \boldsymbol{\theta}$ entonces la distribución de $\hat{\boldsymbol{\theta}}_{MLE,n}$ tiende a:

$$\hat{\boldsymbol{\theta}}_{MLE,n} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$$

siendo $\boldsymbol{I}(\boldsymbol{\theta})$ la matriz de Fisher de $\boldsymbol{\theta}$.

Existe un estimador que alcanza la cota para todo $\boldsymbol{\theta}$ si y solo si

$$\frac{\partial \log f_{\boldsymbol{\theta}}(\boldsymbol{x})}{\partial \boldsymbol{\theta}} = \boldsymbol{I}(\boldsymbol{\theta}))(\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{\theta})$$

para alguna función $T : \mathbb{R}^n \to \mathbb{R}^p$ e $I \in \mathcal{M}_{p\times p}$
Si se cumple la factorización entonces $\text{var}(\hat{\theta}_i) = [\boldsymbol{I}(\boldsymbol{\theta})]_{ii}$ y $\hat{\boldsymbol{\theta}} = \boldsymbol{T}(\boldsymbol{x})$ es el MVU.

Transformación de estimadores

Si $\hat{\theta}$ es un estimador eficiente de θ y f es una transformación afín del tipo $f(\theta) = a\theta + b$ entonces $f(\hat{\theta})$ es un estimador eficiente para $f(\theta)$.

- Si f no es afín, entonces $f(\hat{\theta})$ no tiene por qué ser eficiente ni tampoco insesgado.
- Pero $f(\hat{\theta}_n)$ es asintóticamente eficiente (y por lo tanto insesgado)

Plan

Ejemplo: descenso por gradiente

A continuación veamos el método de descenso por gradiente, muy utilizado en optimización. La idea es la siguiente: partiendo de un punto de una superficie, buscamos la pendiente más pronunciada calculando el gradiente y bajamos un pequeño escalón, recomenzamos desde este nuevo punto y así sucesivamente hasta llegar a un mínimo local.

Se deben elegir:

- un valor inicial θ_0,
- y un parámetro $\delta > 0$, llamado paso.

El método consiste en actualizar en cada iteración del algoritmo nuestra aproximación mediante la regla

$$\theta_{k+1} = \theta_k - \delta \cdot \ell'(\theta_k)$$

hasta llegar a una tolerancia deseada o a una situación de cierta estabilidad.

¿Por qué funciona?

$$f(\theta_{k+1}) = f(\theta_k - \delta f'(\theta_k)) \approx f(\theta_k) + f'(\theta_k)(-\delta f'(\theta_k)) = f(\theta_k) - \delta(f'(\theta_k))^2 < f(\theta_k)$$

Obtenemos entonces una sucesión tal que $f(\theta_n) < f(\theta_{n-1}) < \cdots < f(\theta_2) < f(\theta_1) < f(\theta_0)$

Ejemplo: descenso por gradiente

Con 7 datos: $-0.47, 0.89, 0.26, -0.59, 0.37, 0.54$ y 0.87 y la función de logverosimilitud es

$$\ell(\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{2}(1+\theta x_i)\right) = n\log 2 + \sum_{i=1}^{n} \log\left((1+\theta x_i)\right)$$

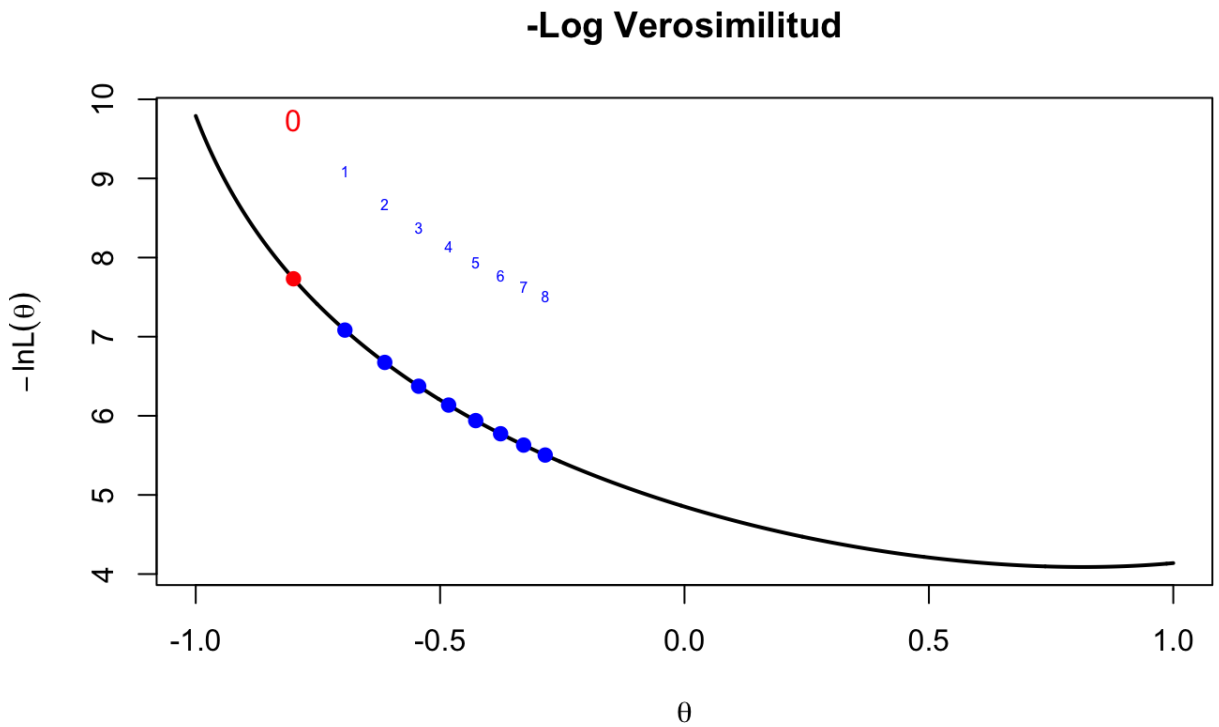


Figura: $\delta = 0{,}015$, iter=8, min=-0.2856

Ejemplo: descenso por gradiente

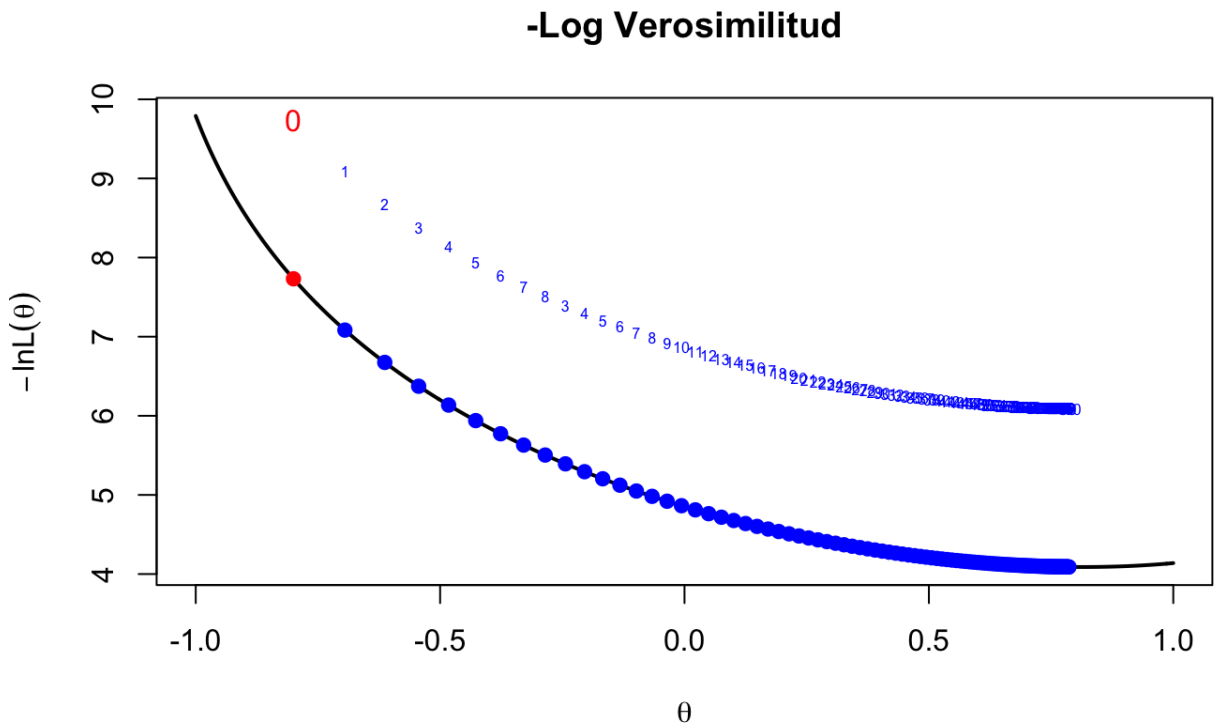


Figura: $\delta = 0{,}015$, iter=100, min=0.7865

Descenso por gradiente de una función de una variable

Otra justificación:

- Si f es creciente entonces $f'(\theta_k) > 0$ y $\theta_{k+1} = x_k - \delta f'(\theta_k) < x_k$ y como f es creciente entonces $f(\theta_{k+1}) < f(\theta_k)$
- Si f es decreciente entonces $f'(\theta_k) < 0$ y $x_{\theta+1} = x_k - \delta f'(\theta_k) > x_k$ y como f es decreciente entonces$f(\theta_{k+1}) < f(\theta_k)$

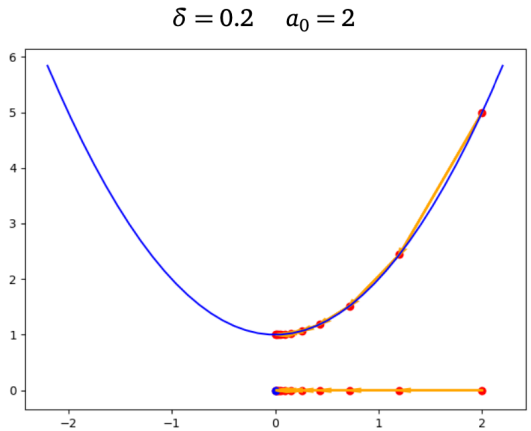
Consideramos la función

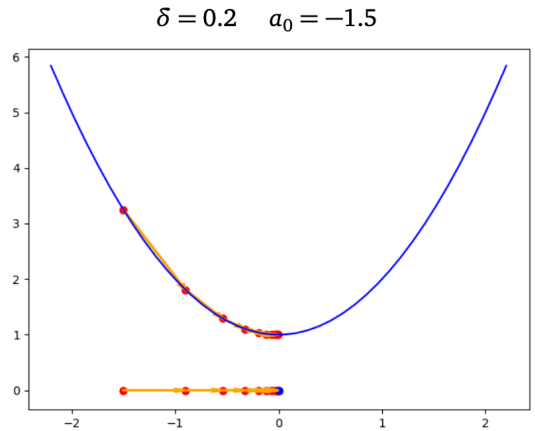
$$f(x) = x^2 + 1,$$

el descenso se hace por

$$a_{k+1} = a_k - \alpha f'(a_k)$$

k	a_k	$f'(a_k) = \text{grad} f(a_k)$	$f(a_k)$
0	2	4	5
1	1.2	2.4	2.44
2	0.72	1.44	1.5184
3	0.43	0.86	1.1866
4	0.25	0.5184	1.0671
5	0.15	0.31	1.0241
6	0.093	0.186	1.0087
7	0.055	0.111	1.0031
8	0.033	0.067	1.0011
9	0.020	0.040	1.0004
10	0.012	0.024	1.0001





Ejemplo: descenso por gradiente

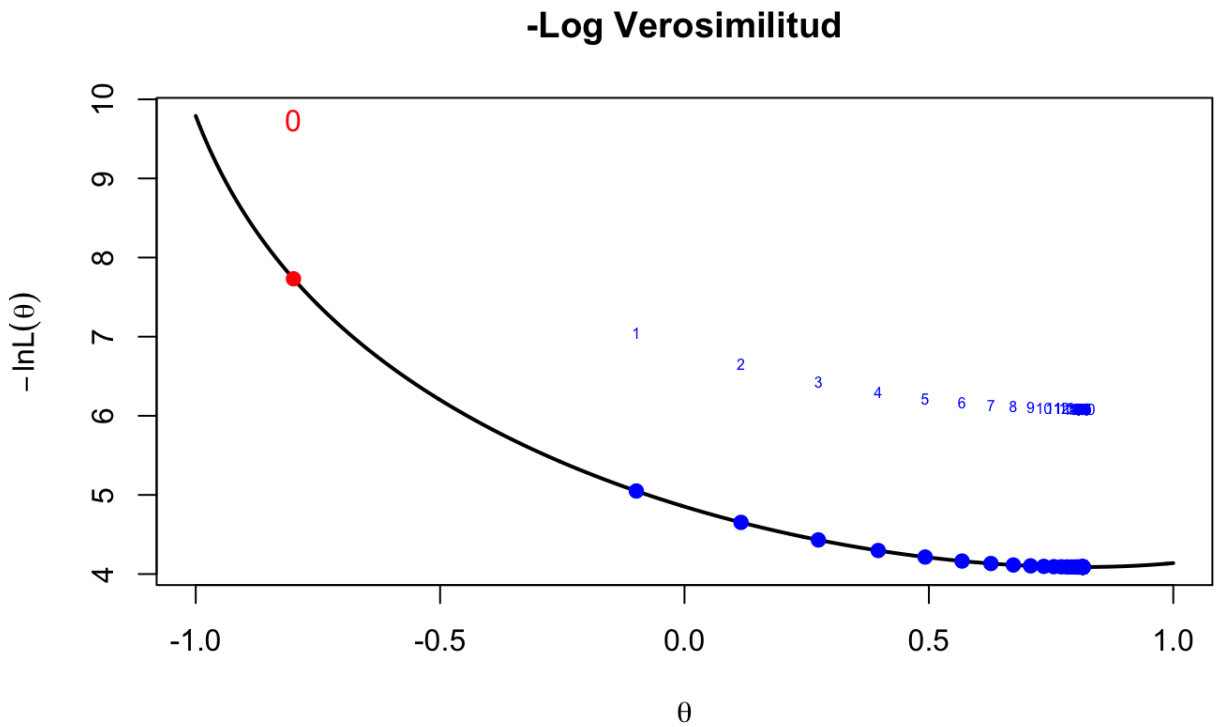


Figura: $\delta = 0{,}1$, iter=100, min=0.8151

Generalización del descenso por gradiente

El gradiente de una función $f : \mathbb{R}^n \to \mathbb{R}$ en $\mathbf{x_0}$ es el vector

$$\nabla f(\mathbf{x_0}) = \left(\frac{\partial f}{\partial x_1}(\mathbf{x_0}), \frac{\partial f}{\partial x_2}(\mathbf{x_0}), \ldots, \frac{\partial f}{\partial x_n}(\mathbf{x_0}) \right)$$

El gradiente es ortogonal a las curvas de nivel de f.

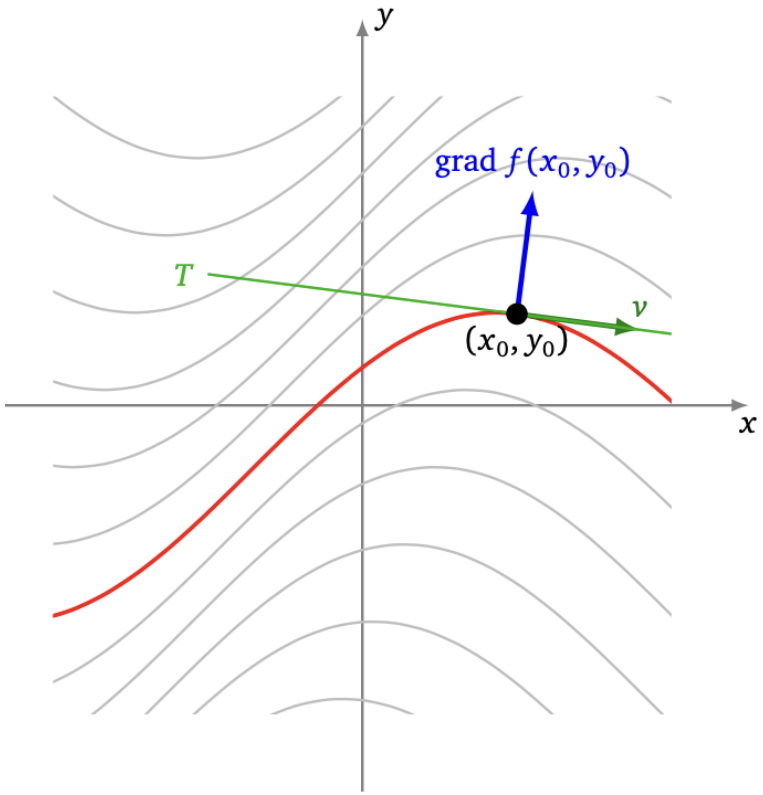


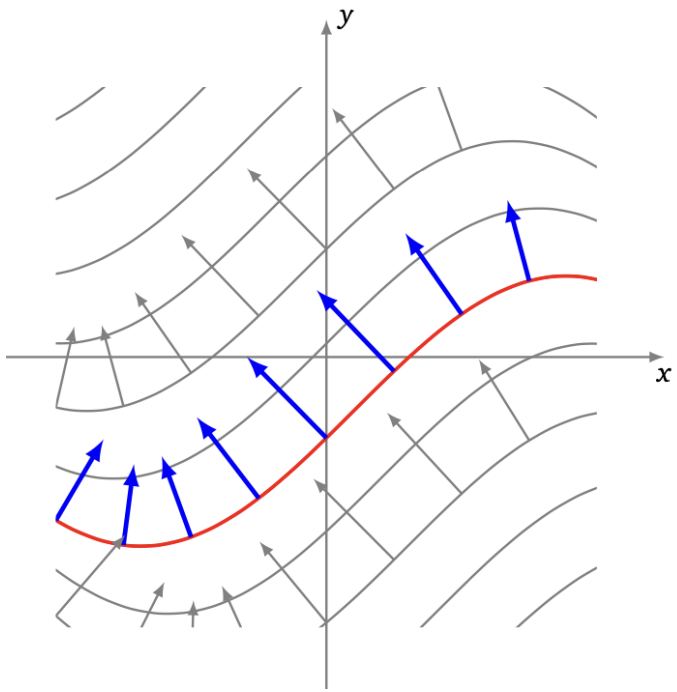
Figura: En rojo la curva de nivel que pasa por (x_0, y_0), en verde el vector tangente en este punto, en azul el vector gradiente.

Generalización del descenso por gradiente

El gradiente es ortogonal a las curvas de nivel de f

En efecto si $\mathcal{C}_k = \{\mathbf{x} : f(\mathbf{x}) = k\}$:

- se puede encontrar una parametrización $\gamma : [-1, 1] \to \mathbb{R} : \gamma(t) = (\gamma_1(t), \gamma_2(t))$ tal que $\gamma(0) = (x_0, y_0)$
- La tangente a la curva $\mathcal{C}$ en (x_0, y_0)es la recta que pasa por (x_0, y_0) y cuyo vector director es $\gamma'(0) = (\gamma_1'(0), \gamma_2'(0))$
- Un vector v es ortogonal a la curva $\mathcal{C}$ en (x_0, y_0) si es ortogonal a la tangente en este punto, es decir si $\langle v, \gamma'(0)\rangle = 0$.



Como para todo $t \in [-1, 1]$ se tiene que $f(\gamma(t)) = k$ se deriva esta expresión usando la regla de la cadena:

$$\mathbb{J}_f(\gamma(t))\mathbb{J}_\gamma(t) = 0$$

es decir

$$\begin{pmatrix} \frac{\partial f}{\partial x}(\gamma(t)) & \frac{\partial f}{\partial y}(\gamma(t)) \end{pmatrix} \begin{pmatrix} \gamma_1'(t) \\ \gamma_2'(t) \end{pmatrix} = 0$$

y evaluando en $t = 0$ encontramos que

$$\langle \nabla f(x_0, y_0), \gamma'(0) \rangle = 0$$

lo cual significa que el gradiente es ortogonal al vector tangente.

Generalización del descenso por gradiente

La dirección del gradiente $\nabla f(x_0, y_0)$ indica la dirección de mayor crecimiento de la función a partir del punto (x_0, y_0).

Esto quiere decir que si queremos pasar lo más rápidamente posible del nivel a a un nivel $b > a$, partiendo de (x_0, y_0) del nivel $f(x_0, y_0) = a$, entonces tenemos que empezar siguiendo la dirección del gradiente $\nabla f(x_0, y_0)$.

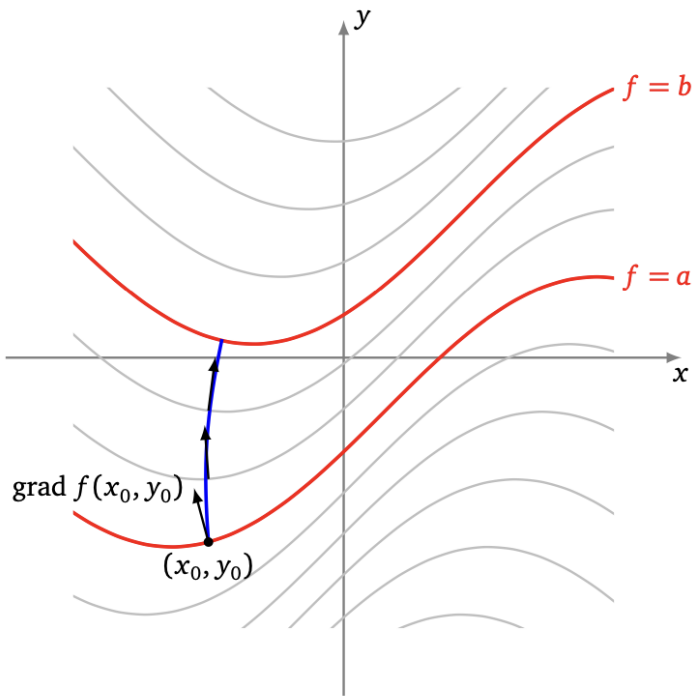


Figura: Si se quiere bajar lo más rapidamente de una montaña se elije la bajada más empinada en el punto y esta es la dirección opuesta al gradiente

Generalización del descenso por gradiente

La derivada direccional de una función diferenciable f según un vector v en (x_0, y_0) cuantifica cuanto varia la función cerca de este punto si uno se mueve en la dirección de v. La dirección en la que el crecimiento es más importante es la del gradiente de f, en efecto:

$$\frac{\partial f}{\partial v}(x_0, y_0) = \langle \nabla f(x_0, y_0), v \rangle = ||\nabla f(x_0, y_0)|| \times ||v|| \times \cos \phi$$

siendo ϕ el ángulo entre $\nabla f(x_0, y_0)$ y v. EL máximo se alcance cuando el ángulo es nulo, es decir que $\nabla f(x_0, y_0)$ y v apuntan en el mismo sentido.

Algoritmo del gradiente descendiente

Sea f una función de varias variables en la que podemos calcular el gradiente $\nabla f(\mathbf{x_0})$ en cualquier punto.

Parámetros de inicialización:

- un punto de partida $\mathbf{x_0}$ elegido al azar
- un paso δ (learning rate)
- un nivel de tolerancia ϵ o una cantidad de iteraciones K

Se calcula una sucesión $\mathbf{x_0}, \mathbf{x_1}, \ldots, \mathbf{x_K}$ de manera recurrente:

$$\mathbf{x_{k+1}} = \mathbf{x_k} - \delta \nabla f(\mathbf{x_k})$$

Se detiene el algoritmo cuando se llega a la cantidad K de iteraciones o cuando $||\nabla f(\mathbf{x_k})|| < \epsilon$

Observaciones:

- La elección de δ es crucial y se puede tomar distinta en cada paso.
- El criterio de parada garantiza que en $\mathbf{x_k}$ el gradiente es muy pequeño. No garantiza que este punto esté cerca de un mínimo local (y menos aún de un mínimo global).

Generalización del descenso por gradiente

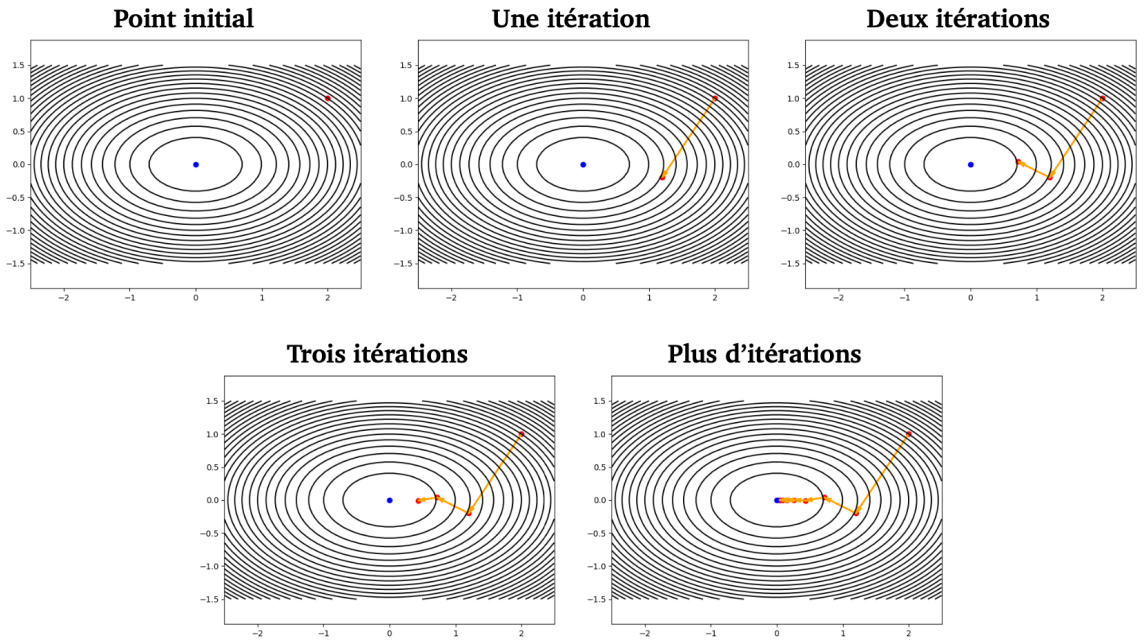
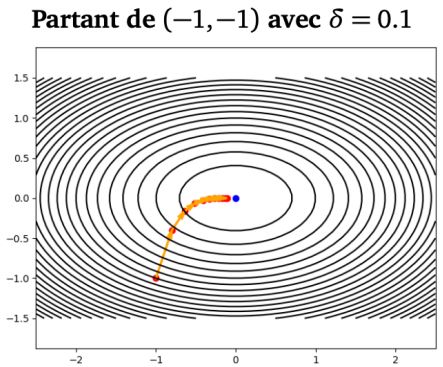


Figura: $\mathbf{x_0} = (2, 1)$ y paso $\delta = 0{,}2$



Generalización del descenso por gradiente

Existen otro métodos de optimización para funciones de varias variables

- El método de Newton Raphson que, en vez de utilizar la aproximación lineal del método por descenso por gradiente, utiliza la aproximación cuadrática y por lo tanto la matriz Hessiana de f. El método tiene como regla de actualización:

$$\mathbf{x_{k+1}} = \mathbf{x_k} - \mathbb{H}_f(\mathbf{x_k})^{-1}\nabla f(\mathbf{x_k})$$

Plan

Simulación de datos: método de inversión

El método de inversión Este método se basa en el siguiente teorema:

Teorema de inversión

Si X tiene distribución F_X entonces $U = F_X(X)$ tiene distribución $U[0,1]$.
Como consecuencia si $U \sim U[0,1]$ entonces $X = F_X^{-1}(U) \sim F_X$

Esto es muy facil de probar dado que si G es la distribución de U y $u \in (0,1)$ entonces

$$G(u) = \mathbb{P}(U \leq u) = \mathbb{P}(F_X(X) \leq u) \leq \mathbb{P}(X \leq F_X^{-1}(u)) = F_X(F_x^{-1}(u)) = u$$

El algoritmo del método de inversión para generar n datos proveniente de una VA X con distribución F_X consiste entonces en:

1. Generar n datos $u_1, \ldots, u_n$ proveniente de una distribución uniforme $U \sim U[0,1]$.
2. Definir $x_1 = F_X^{-1}(u_1), x_2 = F_X^{-1}(u_2), \ldots, x_n = F_X^{-1}(u_n)$

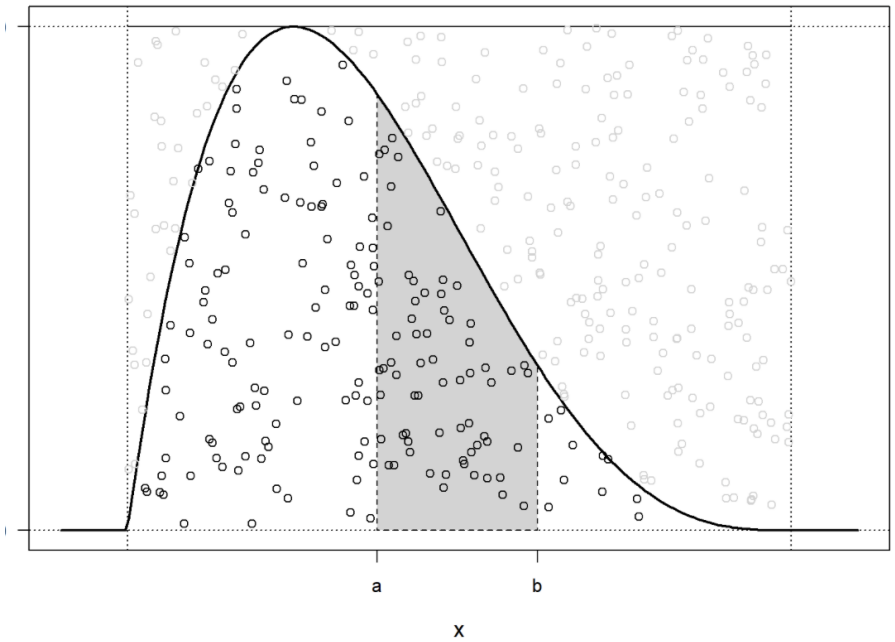
Sin embargo a veces no es sencillo ni hallar F_X (distribución normal), ni calcular la inversa de F_X.

Simulación de datos: método de Aceptación Rechazo

Este método se basa en el siguiente teorema:

Teorema de Aceptación/Rechazo

Sea X una variable aleatoria con función de densidad f y sea $U \sim U[0,1]$ otra variable aleatoria independiente de la anterior.
Entonces para cada $c > 0$ la variable aleatoria bidimensional $(X, cUf(X))$ tiene distribución uniforme en la región $A = \{(x,y) \in \mathbb{R}^2 : 0 \leq y \leq cf(x)\}$.
Reciprocamente si dada una función con densidad f, un vector aleatorio tiene distribución uniforme en el conjunto A, entonces su primera componente X es una variable aleatoria unidimensional con función de densidad f



Simulación de datos: método de Aceptación Rechazo

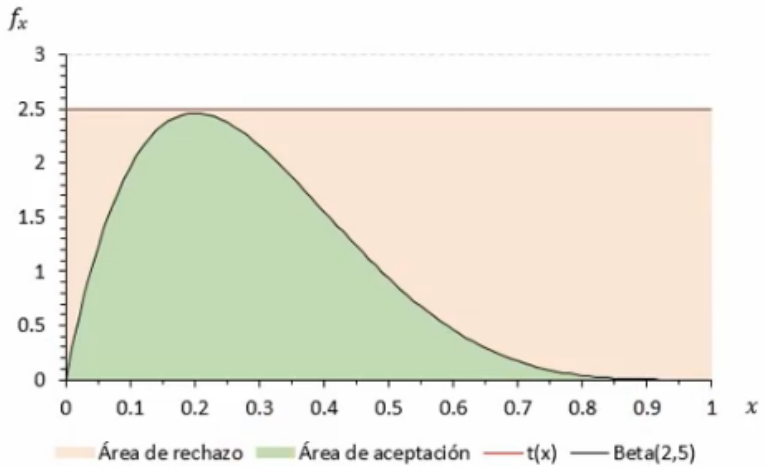
El teorema anterior establece la equivalencia entre simulacion de densidad unidimensionales y simulación de variables bidimensionales con distribución uniforme

$$A_{cf} = \{(x, y) : 0 \leq y \leq f(x)\}$$

.

Si f es una densidad con soporte en $[a, b]$ y M es el máximo de f en $[a, b]$, lo que hacemos es:

1. generar dos variables uniformes U y V con distribución $U[0, 1]$
2. Considerar la variable $T = a + (b - a)V$
3. Si $MU \leq f(T)$ devolver $X = T$ en caso contrario volver al paso 1.



Pasos:

1. Generar un x *al azar*
2. $P\ (\text{aceptar}\ x) = \dfrac{f_x(x)}{t(x)}$
3. Generar $u_x \sim U[0,1]$
4. Si $u_x \leq P\ (\text{aceptar}\ x)$ → devolver x
   Si $u_x > P\ (\text{aceptar}\ x)$ → ir a 1.

Desigualdad de Cauchy Schwarz

$$[\mathbb{E}(XY)]^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

Demostración:

- Supongamos que $\mathbb{E}(X^2) \neq 0$ y sea $\alpha = -\frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}$.

$$\begin{aligned} 0 \leq & \mathbb{E}\left[(\alpha X + Y)^2\right] = \mathbb{E}\left[\alpha^2 X^2 + 2\alpha XY + Y^2\right] = \alpha^2 \mathbb{E}(X^2) + 2\alpha \mathbb{E}(XY) + \mathbb{E}(Y^2) \\ & = \left(\frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}\right)^2 - 2\frac{\mathbb{E}(XY)}{\mathbb{E}(X^2)}\mathbb{E}(XY) + \mathbb{E}(Y^2) \\ & = -\frac{\mathbb{E}(XY)^2}{\mathbb{E}(X^2)} + \mathbb{E}(Y^2) \end{aligned}$$

  lo cual implica que

$$\mathbb{E}(XY)^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$$

- Si $\mathbb{E}(X^2) = 0$ entonces $\mathbb{P}(X^2 = 0) = 1$ y por lo tanto $\mathbb{P}(X = 0) = 1$ lo cual implica que $\mathbb{E}(XY) = 0$ y sigue valiendo la igualdad.