Inferencia Estadística
Estimación Puntual

Gabriel Martos Venturini
gmartos@utdt.edu

UTDT

UNIVERSIDAD TORCUATO DI TELLA

- Un *estimador* $\widehat{\theta}_n(\underline{X})$ es una función de $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \theta)$.
  - El estimador (estadístico) $\widehat{\theta}_n$ es una variable aleatoria con la que pretendemos hacer inferencias sobre el parámetro desconocido θ.
- *Estimador vs Estimación*:
  - $\overline{X}_n$ es un estimador de $E(X) = \mu$ y $\overline{x}_n$ una estimación puntual de μ.
- Hoja de ruta:
  - Métodos generales para construir estimadores.
    - ★ Estimadores de Momentos.
    - ★ Estimadores Máximo Verosímiles (EMV) y principios de inferencia.
    - ★ Aspectos numéricos en torno a los EMV.
  - Cuantificación del riesgo de un estimador.
    - ★ Estimadores Insesgados y de Varianza Mínima
  - Propiedades en muestras finitas y asintóticas de los EMV.

Agenda

1. Métodos para construir estimadores
   - Métodos de momentos
   - Estimadores de máxima verosimilitud

Agenda

1. Métodos para construir estimadores
   - Métodos de momentos
   - Estimadores de máxima verosimilitud

El método de momentos (Pearson–1900)

- Sea $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x;\boldsymbol{\theta})$, donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta \subseteq \mathbb{R}^k$, los k primeros momentos muestrales y poblacionales se definen como:

$$
\begin{aligned}
M_1(\underline{X}) &= \frac{1}{n}\sum_{i=1}^{n} X_i, &\quad \text{y} \quad \mu_1(\boldsymbol{\theta}) &= \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x;\boldsymbol{\theta})\,\mathrm{d}x;\\
M_2(\underline{X}) &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, &\quad \text{y} \quad \mu_2(\boldsymbol{\theta}) &= \mathbb{E}(X^2) = \int_{-\infty}^{\infty} x^2 f(x;\boldsymbol{\theta})\,\mathrm{d}x;\\
&\vdots & &\vdots\\
M_k(\underline{X}) &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, &\quad \text{y} \quad \mu_k(\boldsymbol{\theta}) &= \mathbb{E}(X^k) = \int_{-\infty}^{\infty} x^k f(x;\boldsymbol{\theta})\,\mathrm{d}x.
\end{aligned}
$$

- M_i es una v.a., mientras que μ_i es una función de $\boldsymbol{\theta}$ (desconocido).

- Si la muestra es iid[1] $M_i \to_P \mu_i$ para $i = 1, \ldots, k$ (LGN).

[1] Asumiendo que los momentos poblacionales están bien definidos.

- Cuando $n \gg 0$ luego "$M_i \approx \mu_i$", entonces el estimador de momentos $\widetilde{\boldsymbol{\theta}}_n = (\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k)$ se obtiene resolviendo el sistema de ecuaciones:

$$\begin{array}{rcl} M_1(\underline{X}) & = & \mu_1(\boldsymbol{\theta}); \\ M_2(\underline{X}) & = & \mu_2(\boldsymbol{\theta}); \\ \vdots & \vdots & \vdots \\ M_k(\underline{X}) & = & \mu_k(\boldsymbol{\theta}). \end{array}$$

- Nota: Cuando la muestra se realiza $(M_1 = m_1, \ldots, M_k = m_k)$, tendremos **estimaciones de momentos** (solución del sistema).

- Ejemplos: Modelos Bernoulli y Normal.

- Inconvenientes:
  - Momentos poblacionales no dependen de θ o no están definidos.
  - No unicidad de $\widetilde{\theta}_n$ y en algunos casos $\widetilde{\theta}_n \notin \Theta$.
  - No garantizan que se cumplan los principios de inferencia.

Agenda

1. Métodos para construir estimadores
   - Métodos de momentos
   - Estimadores de máxima verosimilitud
     - Definición y algunos ejemplos
     - Principios de inferencia y EMV
     - Métodos numéricos y estimadores MV

Agenda

Refresh

- Sean $\underline{X} = \underline{x}$ los datos (realización de $\underline{X} \equiv \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \boldsymbol{\theta})$, con $\boldsymbol{\theta} \in \Theta$) definimos la función de verosimilitud como:

$$L(\boldsymbol{\theta}) \equiv L_n(\boldsymbol{\theta}|\underline{x}) = \underbrace{\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})}_{f(\underline{x};\theta)}.$$

- $L(\boldsymbol{\theta})$ debe entenderse como una función de $\boldsymbol{\theta}$.

- Podemos interpretar $L(\boldsymbol{\theta}) = P_\theta(\underline{X} = \underline{x}) = P_\theta(\text{Datos} \,|\, \text{Modelo})$.

- Ejemplo: $X \sim \text{Bern}(\theta)$ y $\underline{x} = \{x_1 = 1, x_2 = 1, x_3 = 0\}$:

$$L(\theta) = \theta^2 - \theta^3.$$

Si $L(\theta_1)/L(\theta_2) > 1$ entonces θ_1 es más factible/verosímil que θ_2 en relación a la evidencia empírica $\underline{x}$ (y el modelo probabilístico).

Estimación máximo verosímil

- Consideremos $\underline{X} = \{X_1, \ldots, X_n\} \overset{iid}{\sim} f(x; \boldsymbol{\theta})$, donde $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k) \in \Theta$.
- Dada $\underline{X} = \underline{x}$ definimos la estimación máximo verosímil:

$$\boxed{\widehat{\boldsymbol{\theta}}_n(\underline{x}) := \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\underline{X} = \underline{x}).}$$

  - $\widehat{\theta}_n$ es el valor de θ que maximiza $P_\theta(\text{Datos}\,|\,\text{Modelo})$.
- Por consiguiente, el estimador máximo verosímil (EMV):

$$\widehat{\boldsymbol{\theta}}_n(\underline{X}) = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}|\underline{X}).$$

  - Notar que $L(\boldsymbol{\theta}|\underline{X})$ es una función aleatoria de θ.
  - Por lo tanto $L(\widehat{\boldsymbol{\theta}}_n|\underline{X}) \geq L(T_n|\underline{X})$ para cualquier otro estadístico T_n.
- Veamos una ilustración de estos conceptos en la próxima diapositiva.

$X \sim N(\mu = 2, \sigma_0^2 = 1)$

- $\ell(\mu) \equiv \ln L(\mu, \sigma_0^2 = 1 | \underline{x})$.

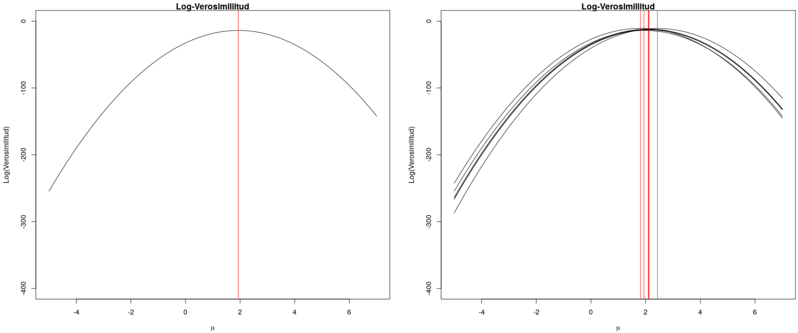


Figure: Izquieda: Estimación MV para una muestra concreta. Derecha: Diferentes realizaciones de $\ell(\mu)$ cuando muestreamos de $N(\mu = 2, \sigma_0^2 = 1)$.

- Notar: El argumento del máximo en ℓ (la estimación MV del parámetro μ) dependerá de la realización particular de la muestra.

(back–up código en R)

```
### Creo la función de Verosimilitud (sigma = 1)
l = function(mu,muestra){
  return( (-n/2)*log(2*pi) - sum(muestra^2)/2 - n*mu^2/2 +
    sum(muestra)*mu  )}

mu = 2; sigma = 1; n = 10
muestra = rnorm(n, mu, sigma)
plot(seq(-5,7,by=0.1), l(seq(-5,7,by=0.1),muestra))

for(i in 1:5){
  muestra = rnorm(n, mu, sigma)
  points(seq(-5,7,by=0.1), l(seq(-5,7,by=0.1),muestra),
  type = 'l')
  abline(v = mean(muestra), col = 'red')
}
```

- Si $\ell(\boldsymbol{\theta}) \equiv \log L(\boldsymbol{\theta})$, notar que:

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}),$$

  por ser el logaritmo una función monótona creciente.

- Obtenemos $\widehat{\boldsymbol{\theta}}_n$ resolviendo el sistema $S(\boldsymbol{\theta}) = \underbrace{\frac{\partial}{\partial\boldsymbol{\theta}}}_{\text{Score}}\ell(\boldsymbol{\theta}) = \mathbf{0}$.

- Ejemplo I: Modelos Binomial y Normal (varianza conocida).
  - Notar que ambos estimadores son funciones de estadísticos suficientes.

- Ejemplo II: El modelo de regresión lineal.

- Estimación máximo verosímil restringida:

$$\widehat{\boldsymbol{\theta}}_n^{(S)} = \arg\max_{\boldsymbol{\theta}} L(\boldsymbol{\theta}),\ \text{sujeto a: } \boldsymbol{\theta} \in S \subset \Theta$$

  - Modelos de regresión en alta dimensión (Ridge y Lasso).

Condiciones de segundo orden

- Encontrar valores de $\boldsymbol{\theta}$ para los cuales se cumple que $S(\boldsymbol{\theta}) = \mathbf{0}$ no garantizan necesariamente que se trate de un máximo de $L(\boldsymbol{\theta})$.

- Llamemos $H(\boldsymbol{\theta})$ al Hessiano asociado a $\ell(\boldsymbol{\theta})$, es decir que:

$$[H(\boldsymbol{\theta})]_{ij} = \frac{\partial^2}{\partial \theta_i \, \partial \theta_j} \ell(\boldsymbol{\theta}) \text{ para } i, j = 1, \ldots, k.$$

- CSO: $\boldsymbol{\theta}$ es un máximo global si $H(\boldsymbol{\theta})$ es una matriz definida negativa.
  - Modelos de 1 parámetro: $\ell''(\theta)\big|_{\theta=\widehat{\theta}_n} < 0$.

- Para los modelos de la familia exponencial, en general, los estimadores máximo verosímiles existen y son únicos.
  - En otras palabras, $L(\boldsymbol{\theta})$ es una función estrictamente cóncava si el modelo estadístico para los datos pertenece a la familia exponencial.
  - Discusión formal en VP §3.21 (pp 75).

Otros ejemplos de máxima verosimilitud

Example (Uniforme)

Sea $\{x_1, \ldots, x_n\} \overset{iid}{\sim}$ Unif$(0, \theta]$, la verosimilitud se define como:

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{\{(0,\theta]\}}(x_i),$$

que no es diferenciable respecto de θ (entonces?).

Otros ejemplos de máxima verosimilitud

Example (Uniforme)

Sea $\{x_1, \ldots, x_n\} \overset{iid}{\sim}$ Unif$(0, \theta]$, la verosimilitud se define como:

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{\{(0,\theta]\}}(x_i),$$

que no es diferenciable respecto de θ (entonces?).

- Notar que $L(\theta) \geq 0$ con $L(\theta) > 0 \iff \theta \geq \max(x_1, \ldots, x_n)$.
- Además, si $\theta \geq \max(x_1, \ldots, x_n)$, entonces $L(\theta)$ es decreciente en θ.
- La verosimilitud se maximiza en $\max(x_1, \ldots, x_n) = x_{(n)}$ y por lo tanto el estimador de máxima verosimilitud es

$$\widehat{\theta}_n = \max(X_1, \ldots, X_n) = X_{(n)}$$

Modelo de Laplace

Si $X \sim \text{Laplace}(\mu, b)$, entonces la densidad de la v.a. X se escribe como:

$$f(x; \mu, b) = \frac{1}{2b} \exp(-\frac{|x-\mu|}{b}).$$

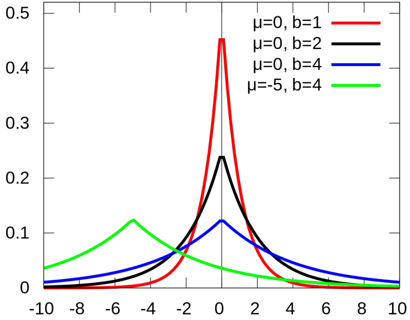


Figure: Densidad del modelo Laplace (el parámetro $b > 0$).

Otros ejemplos de máxima verosimilitud

Example (Laplace)

Sea $\{x_1, \ldots, x_n\} \overset{iid}{\sim}$ Laplace$(\theta, 1)$, obtener el estimador de máxima verosimilitud de θ. En primer término construimos la verosimilitud:

$$\begin{aligned} L(\theta) &= (1/2)\exp(-|X_1 - \theta|)\ldots(1/2)\exp(-|X_n - \theta|) \\ &= (1/2)^n \exp\left(-\sum_{i=1}^{n} |X_i - \theta|\right). \end{aligned}$$

Luego

$$\ell(\theta) = n\log(1/2) - \sum_{i=1}^{n} |X_i - \theta|,$$

que no es diferenciable, ya que el valor absoluto no es derivable en $\theta = 0$.

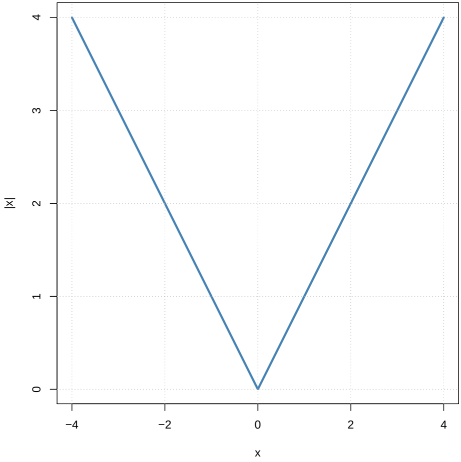


Figure: Gráfico de la función: $g(x) = |x|$.

Example (Laplace)

Definamos $\text{sign}(x)$ a la función que vale 1 si $x > 0$ y -1 si $x < 0$. La derivada de $g(x) = |x|$ es $g'(x) = \text{sign}(x)$ si $x \neq 0$. Si 'derivamos', informalmente, $\ell(\theta)$ respecto de θ e igualamos a cero obtenemos

$$\sum_{i=1}^{n} \text{sign}\left(x_i - \theta\right) = 0.$$

El valor de θ tiene que ser tal que 'la mitad' de las x_i tienen que ser menores que θ y 'la otra mitad' mayores que θ. Luego se tiene que:

$$\widehat{\theta}_n = \text{mediana}(X_1, \ldots, X_n).$$

Agenda

Suficiencia del EMV en familias exponenciales

- Asumiendo que el modelo pertenece a la familia exponencial:

$$
\begin{aligned}
L(\theta) &\overset{iid}{=} \Pi_{i=1}^{n} f(x_i|\theta) = \Pi_{i=1}^{n} h(x_i) c(\theta) \exp\Big(w(\theta) t(x_i)\Big) \\
&= \left[\Pi_{i=1}^{n} h(x_i)\right] c^n(\theta) \exp\left(w(\theta) \sum_{i=1}^{n} t(x_i)\right) \\
\ell(\theta) &= \sum_{i=1}^{n} \log(h(x_i)) + n \log(c(\theta)) + w(\theta) T(\underline{x}).
\end{aligned}
$$

- T es un estadístico (minimal) <u>suficiente y completo</u> para θ.

$$
S(\theta) = \frac{nc'(\theta)}{c(\theta)} + w'(\theta) T(\underline{x}) = 0.
$$

- El EMV es suficiente para θ ya que será una función de T.

Principios de Verosimilitud e Invarianza

- Dadas dos muestras $\underline{x}_1$ y $\underline{x}_2$ tales que $L(\theta|\underline{x}_1) \propto L(\theta|\underline{x}_2)$, luego:

$$\arg\max_{\theta\in\Theta} L(\theta|\underline{x}_1) = \arg\max_{\theta\in\Theta} L(\theta|\underline{x}_2).$$

  Por lo tanto los EMV cumplen el principio de verosimilitud.

- Si ψ es una función biyectiva (uno–a–uno) y $\widehat{\theta}_n$ es el estimador MV de θ, entonces $\psi(\widehat{\theta}_n)$ es el estimador máximo verosímil de $\psi(\theta)$.
  - $\{X_1, \ldots, X_n\} \overset{iid}{\sim} N(\mu, \sigma_0^2)$, luego $\widehat{\mu}_n = \overline{X}_n$.
  - Si nos interesa $\psi(\mu) = e^{\mu}$, luego $\widehat{\psi}_n = e^{\overline{X}_n}$.

- El principio de invarianza se cumple en contextos aún más generales y también es válido para el caso multiparámetro (CB §7.2.4).

Agenda

1. Métodos para construir estimadores
   - Métodos de momentos
   - Estimadores de máxima verosimilitud
     - Definición y algunos ejemplos
     - Principios de inferencia y EMV
     - Métodos numéricos y estimadores MV

- En general no existen soluciones analíticas para el EMV.
  - Parámetro de localización en un modelo Cauchy (ejercicio G2).
- Si podes computar derivadas respecto de $\ell(\boldsymbol{\theta})$ (o de $L(\boldsymbol{\theta})$), vas a poder implementar métodos numéricos para **aproximar** el valor de θ para el que se maximiza $\ell(\boldsymbol{\theta})$ (y por tanto $L(\boldsymbol{\theta})$).
- Discutimos un método numérico clásico de estimación.

Newton–Raphson

- Encontrar el cero de una función diferenciable $f(x)$.
- Hacemos expansión de Taylor en torno a $f(x^*) = 0$:

$$\text{Elijo } x_0 : \quad 0 = f(x^*) \approx f(x_0) + f'(x_0)(x^* - x_0),$$

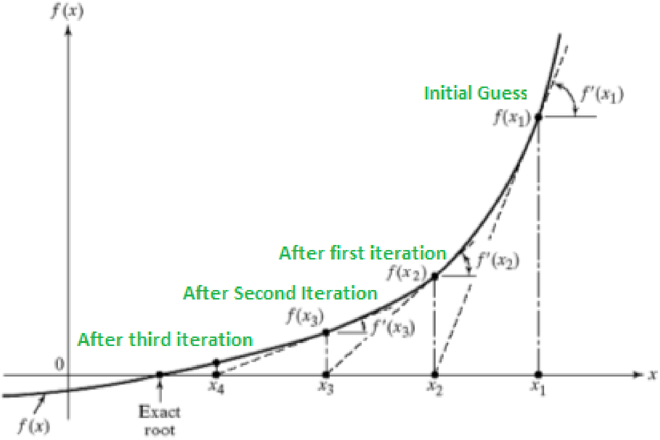
luego resolviendo para x^* obtenemos que:

$$x^* \approx x_0 - \frac{f(x_0)}{f'(x_0)}.$$

- En la práctica procedemos eligiendo x_0 e iteramos:

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})}, \text{ para } k = 1, 2, \ldots$$

  - hasta que $|x_k - x_{k-1}|$ sea pequeño y/o $f(x_k) \approx 0$.



Iteramos hasta que:

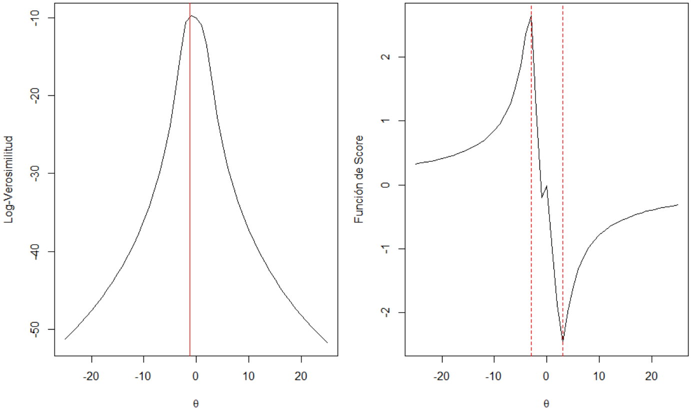
- $|x_k - x_{k-1}| \approx 0$ y/o
- $|f(x_k)| \approx 0$.

Newton–Raphson y estimación máximo verosímil

- $S(\theta) = \frac{\partial}{\partial\theta}\ell(\theta)$ (score) y $H(\theta) = \frac{\partial^2}{\partial\theta^2}\ell(\theta)$ (hessiano).
- Dado un candidato inicial para el EMV de θ al que llamamos $\widehat{\theta}_n^{(0)}$, en el paso k-ésimo, la aproximación numérica de la estimación MV será:

$$\widehat{\theta}_n^{(k)} = \widehat{\theta}_n^{(k-1)} - \frac{S(\widehat{\theta}_n^{(k-1)})}{H(\widehat{\theta}_n^{(k-1)})}, \text{ para } k = 1, 2, \ldots$$

- El procedimiento continua hasta que se verifica convergencia:
  - $|\widehat{\theta}_n^{(k)} - \widehat{\theta}_n^{(k-1)}| \leq \varepsilon$, y/o $|S(\widehat{\theta}_n^{(k)})| \leq \varepsilon$ (ε tan pequeño como quieras).
- Si la función de verosimilitud no es estrictamente cóncava pueden existir múltiples máximos / mínimos / puntos de ensilladura.
- Ejemplo (en R): $X \sim f(x;\theta) = \frac{1}{\sqrt{\pi}\Gamma(1/2)}(1 + (x-\theta)^2)^{-1}$.
  - Muestra: $X_1 = -1.5, X_2 = 0.5, X_3 = 2, X_4 = -2.5$.



- Investiga que ocurre cuando $X_4 = -20.5$.

Caso multiparámetro

- En este contexto S es un gradiente con k componentes:

$$S(\boldsymbol{\theta}) = \Big(\frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta})\Big)^T$$

- y H una matriz Hessiana de $k \times k$ con componentes

$$[H(\boldsymbol{\theta})]_{i,j} = \frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\ell(\boldsymbol{\theta}) \text{ para } i,j = 1\ldots,k.$$

- Dado un candidato inicial para el EMV de $\boldsymbol{\theta}$ al que llamamos $\widehat{\boldsymbol{\theta}}_n^{(0)}$, en el paso k-ésimo, la aproximación numérica del EMV será:

$$\widehat{\boldsymbol{\theta}}_n^{(k)} = \widehat{\boldsymbol{\theta}}_n^{(k-1)} - \big[H(\widehat{\boldsymbol{\theta}}_n^{(k-1)})\big]^{-1} S(\widehat{\boldsymbol{\theta}}_n^{(k-1)}), \text{ para } k = 1, 2, \ldots$$

- En contextos de muchos parámetros se suele optimizar de a una coordenada a la vez (*Coordinate Descent*, evitamos computar H).

Sobre la convergencia de NR

- El método de NR produce una secuencia que converge a la raíz $S(\widehat{\theta}_n) = 0$ a velocidad cuadrática si se cumplen las condiciones:
  1. θ_0 está suficientemente cerca de $\widehat{\theta}_n$.
  2. $H(\theta) \neq 0$ para todo θ en un entorno de $\widehat{\theta}_n$.
  3. $H'(\theta)$ es continua como función de θ en un entorno de $\widehat{\theta}_n$.
- Además, si la función de verosimilitud es estrictamente cóncava, entonces la solución numérica del método de NR se corresponde con el único máximo global (en otro caso, el método puede arribar a raíces que no se corresponden con dicho máximo global).

Recapitulación

- Discutimos dos métodos generales para construir estimadores.
- Los EMV son uno de los más utilizados en la práctica porque no solo cumplen los 3 principios de inferencia, sino también porque tienen interesantes propiedades en muestras grandes.
  - Asintóticamente, los EMV son parecidos a los estimadores insesgados de mínima varianza (menor error cuadrático medio).
- Newton–Raphson: Discutimos métodos numéricos para *aproximar* el valor de la *estimación* máximo verosímil.
- Siguiente: Medir el riesgo de un estimador para eventualmente poder comparar entre estimadores; y a partir de allí definir nociones de optimalidad. Luego discutiremos propiedades asintóticas de los EMV.