# Chapter 7

# Estimación Robusta

## 7.1 El problema de la robustez para el modelo de posición

Sea el modelo de posición y escala

$$x_i = \mu + \sigma u_i, 1 \leq i \leq n, \tag{7.1}$$

donde $\mu$ y $\sigma$ son parámetros de posición y escala respectivamente, $u_1, ..., u_n$ son variables i.i.d. con distribución $F$. En este caso, $x_1, ..., x_n$ resulta una muestra aleatoria de $F_{\mu\sigma}$, donde $F_{\mu\sigma}(x) = F((x-\mu)/\sigma)$ Por ejemplo las $x_i$ pueden ser distintas mediciones de una misma magnitud física $\mu$ medida con un error $\sigma u_i$ .

Si $F = \Phi$, la función de una distribución N(0,1), entonces las $x_i$ tienen distribución N$(\mu, \sigma^2)$. Por lo tanto, un estimador óptimo de $\mu$ es $\bar{x} = \sum_{i=1}^{n} x_i/n$. Efectivamente este estimador es IMVU y minimax. Es importante señalar que para que $\bar{x}$ tenga estas propiedades, la distribución de los $u_i$ debe ser exactamente N(0,1). Sin embargo, en la mayoría de las aplicaciones prácticas a lo sumo se puede asegurar los errores de medición tienen distribución *aproximadamente normal*. Por lo tanto, cabe preguntarse cual será el comportamiento de estimador $\bar{x}$ en este caso.

Una forma de determinar distribuciones aproximadamente normales es considerar entornos de contaminación de la función de distribución

$\Phi$ de la N(0,1). Un entorno de contaminación de tamaño $\epsilon$ de la distribución $\Phi$ se define por

$$\mathcal{V}_\epsilon = \{F : F = (1-\epsilon)\Phi + \epsilon H \text{ con } H \text{ arbitraria}\}. \tag{7.2}$$

La distribución $F = (1-\epsilon)\Phi + \epsilon H$ corresponde a que las observaciones con probabilidad $1-\epsilon$ provienen de la distribución $\Phi$ y con probabilidad $\epsilon$ de la distribución $H$.

En efecto supongamos que se tienen tres variables aleatoria independientes : $Z$ con distribución $\Phi$, $V$ con distribución $H$, y $W$ con distribución Bi$(1,\epsilon)$. Definamos entonces la variable aleatoria $U$ de la siguiente manera

$$U = \begin{cases} Z & \text{si} \quad W = 0 \\ V & \text{si} \quad W = 1 \end{cases}.$$

Luego

$$\begin{aligned} F_U(u) &= P(U \le u) = P(U \le u,\ W = 0) + P(U \le u,\ W = 1) \\ &= P(U \le u |\ W = 0)P(W = 0) + P(U \le u |\ W = 1)P(W = 1) \\ &= (1-\epsilon)\Phi(u) + \epsilon H(u). \end{aligned}$$

Con lo cual, si $\epsilon$ es pequeño (por ejemplo .05 o .10) esto significará que la gran mayoría de las observaciones se obtendrán a partir de la distribución $\Phi$, es decir serán normales. Por lo tanto, podemos afirmar que si $\epsilon$ es pequeño y $F \in \mathcal{V}_\epsilon$, entonces $F$ está cerca de $\Phi$. Supongamos que tenemos una muestra aleatoria $x_1, ..., x_n$ de $F \in V_\epsilon$ . Por lo tanto una proporción $(1-\epsilon)$ de las observaciones estarán dadas por (7.1) con $u_i$ proveniente de una distribución $\Phi$, y una proporción $\epsilon$ tendrán el correspondiente $u_i$ proveniente de la distribución $H$. Estas últimas observaciones serán denominadas puntos atípicos o *outliers,* y pueden ser debidas a realizaciones del experimento en circunstancias anormales u otros factores de error como, por ejemplo, una equivocación en la transcripción del dato.

Lo que vamos a mostrar a continuación es que aunque $\epsilon$ sea pequeño el comportamiento del estimador $\bar{x}$ puede ser muy ineficiente para distribuciones $F \in \mathcal{V}_\epsilon$.

Primero mostraremos que si

$$F = (1-\epsilon)\Phi + \epsilon H, \tag{7.3}$$

entonces

$$E_F(u) = (1-\epsilon)E_\Phi(u) + \epsilon E_H(u). \tag{7.4}$$

Además, si $E_H(u) = 0$, se tiene

$$\text{var}_F(u) = (1-\epsilon)\text{var}_\Phi(u) + \epsilon \text{var}_H(u). \tag{7.5}$$

Para mostrar (7.4) supongamos que la $H$ tiene densidad $h$ , y sea $\varphi$ la densidad correspondiente a $\Phi$. Luego la densidad de $F$ es

$$f = (1-\epsilon)\varphi + \epsilon h,$$

y luego

$$E_F(u) = \int_{-\infty}^{\infty} uf(u)du = (1-\epsilon)\int_{-\infty}^{\infty} u\varphi(u)du + \epsilon\int_{-\infty}^{\infty} uh(u)du = (1-\epsilon)E_\Phi(u) + \epsilon E_H(u).$$

Para mostrar (7.5), observemos que

$$\begin{aligned}
\text{var}_F(u) &= \int_{-\infty}^{\infty} u^2 f(u)du \\
&= (1-\epsilon)\int_{-\infty}^{\infty} u^2\varphi(u)du + \epsilon\int_{-\infty}^{\infty} u^2 h(u)du = \\
&= (1-\epsilon) + \epsilon\ \text{var}_H(u).
\end{aligned}$$

Consideremos ahora al estimador $\hat{\mu} = \bar{x}$, donde la muestra $x_1, ..., x_n$ son generadas por (7.1) donde las $u_i$ son independientes con distribución dada por (7.3) con $E_H(u) = 0$

Luego

$$\text{var}_F(\bar{x}) = \frac{\sigma^2\ \text{var}_F(u)}{n} = \frac{\sigma^2((1-\epsilon) + \epsilon\ \text{var}_H(u))}{n}.$$

Luego, si $\epsilon = 0$, entonces $\text{var}(\bar{x}) = \sigma^2/n$. En cambio una contaminación de tamaño $\epsilon$ puede producir un aumento de la varianza ilimitado, ya que $\text{var}_H(u)$ puede ser ilimitada, inclusive infinita.

Esta extrema sensibilidad de $\bar{x}$ a una contaminación con una proporción pequeña de outliers también puede verse de la siguiente forma. Supongamos que se tiene una muestra $x_1, ..., x_n$ y se agrega una observación $x_{n+1.}$ Si esta observación es un outlier, su influencia en $\bar{x}$ puede

ser ilimitada. En efecto sean $\bar{x}_n$ y $\bar{x}_{n+1}$ el promedio basado en $n$ y $n+1$ observaciones respectivamente. Luego se tiene

$$\bar{x}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}x_{n+1} = \bar{x}_n + \frac{1}{n+1}(x_{n+1} - \bar{x}_n),$$

y por lo tanto e $\bar{x}_{n+1}$ puede tomar valores tan altos ( o tan bajos) como se quiera con tal de tomar $x_{n+1}$ suficientemente lejos de $\bar{x}_n$.

Supongamos que tenemos el modelo de posición dado por (7.1) donde la distribución $F$ de los $u_i$ es simétrica respecto de 0. Como en este caso $\mu$ es también la mediana de las observaciones, un estimador alternativo será $\tilde{\mu}$ =mediana$(x_1, ..., x_n)$. Ordenemos los datos $x_1, ..., x_n$ de menor a mayor obteniendo los valores $x_{(1)} \leq ... \leq x_{(n)}$. Luego la mediana estará dada por

$$\tilde{\mu} = \begin{cases} x_{(m+1)} & \text{si} \quad n = 2m+1 \\ x_{(m)} + x_{(m+1)} & si \quad n = 2m \end{cases} .$$

Veamos que este estimador es mucho más resistente a outliers que la media. En efecto, para que la mediana tome un valor ilimitado no es suficiente agregar un outlier, sino se requiere por lo menos $n/2$ outliers.

Un estimador como la mediana que es poco sensible a outliers se denomina **robusto**

La distribución de $\tilde{\mu}$ para muestras finitas es muy complicada aún en el caso de muestras normales. Sin embargo, podremos derivar su distribución asintótica. Para ello necesitamos una version del Teorema Central del Límite para arreglos triangulares que enunciaremos sin demostración.

**Teorema Central del Límite**. Sean para cada $n$ natural, $v_{n1}, ...v_{nn}$, $v$ variables aleatoria independientes igualmente disribuidas. Supongamos que existan constantes $M > 0$ y $m > 0$, tales que $|v_{ni}| \leq M$ y $\lim_{n\to\infty} \text{var}(v_{ni}) \geq m$. Luego se tiene que

$$\frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{(v_{ni} - E(v_{ni}))}{\text{var}(v_{ni})^{1/2}} \xrightarrow{D} \text{N}(0, 1).$$

El siguiente Teorema establece la distribución asintótica de la mediana.

**Teorema 1**. Sea $x_1, ..., x_n$ una muestra aleatoria de una distribución $F$ con una única mediana $\mu$ y con una densidad $f$ tal que $f(\mu) > 0$. Entonces si $\tilde{\mu}_n$ es la mediana de la muestra, se tiene que

$$n^{1/2}(\tilde{\mu}_n - \mu) \xrightarrow{D} N\left(0, \frac{1}{4f^2(\mu)}\right).$$

Demostración: Para facilitar la demostración consideraremos solo el caso que $n = 2m + 1$. Tenemos que demostrar

$$\lim_{n\to\infty} P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = \Phi(2f(\mu)y), \tag{7.6}$$

donde $\Phi$ es la función de distribución correspondiente a N(0,1)

Es inmediato que

$$P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right). \tag{7.7}$$

Sea

$$v_{ni} = \begin{cases} 1 & \text{si} \quad x_i \leq \mu + \dfrac{y}{n^{1/2}} \\ 0 & \text{si} \quad x_i > \mu + \dfrac{y}{n^{1/2}} \end{cases}, \ 1 \leq i \leq n. \tag{7.8}$$

Como $v_{ni}$ tiene distribución Bi($F(\mu + yn^{-1/2}, 1)$ se tiene

$$E(v_{ni}) = \nu_n = F(\mu + \frac{y}{n^{1/2}}),$$

y

$$\text{var}(v_{ni}) = \nu_n(1 - \nu_n).$$

De acuerdo a la definición de mediana se tiene que

$$P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right) = P\left(\sum_{i=1}^{n} v_{ni} \geq \frac{n}{2}\right)$$

$$= P\left(\frac{1}{n^{1/2}} \sum_{i=1}^{n} \frac{(v_{ni} - \nu_n)}{(\nu_n(1 - \nu_n))^{1/2}} \geq \frac{(n/2 - n\nu_n)}{(n\nu_n(1 - \nu_n))^{1/2}}\right). \tag{7.9}$$

Como $|v_{ni}| \leq 1$, y $\lim_{n\to\infty}\text{var}(v_{ni}) = 1/4$. se cumplen las hipótesis del Teorema Central del Límite. Luego

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{(v_{ni}-\nu_n)}{(\nu_n(1-\nu_n))^{1/2}} \xrightarrow{D} N(0,1). \tag{7.10}$$

Usando el hecho de que $F(\mu) = 1/2$, y el Teorema del Valor Medio tenemos

$$\frac{(n/2-n\nu_n)}{n^{1/2}} = n^{1/2}\left(F(\mu) - F(\mu+\frac{y}{n^{1/2}})\right) = -n^{1/2}f(\mu_n^*)\frac{y}{n^{1/2}} = -yf(\mu_n^*),$$

donde $\mu_n^*$ es un punto intermedio entre $\mu$ y $\nu_n$. Luego usando el hecho que $\nu_n \to 1/2$ y $\mu_n^* \to \mu$, resulta

$$\frac{(n/2-n\nu_n)}{(n\nu_n(1-\nu_n))^{1/2}} \to -2yf(\mu). \tag{7.11}$$

Luego, usando (7.7), (7.9), (7.10) y (7.11) tenemos que

$$\begin{aligned}\lim_{n\to\infty} P(n^{1/2}(\tilde{\mu}_n-\mu)\leq y) &= P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right)\\ &= 1-\Phi(-2f(\mu)y) = \Phi(2f(\mu)y),\end{aligned}$$

y por lo tanto hemos probado (7.6). □

**Observación 1.** El Teorema 1 implica que $\tilde{\mu}_n \xrightarrow{p} \mu$. También puede probarse que $\tilde{\mu}_n \xrightarrow{a.s.} \mu$, pero no se dará la demostración.

Apliquemos ahora este resultado al modelo (7.1) y supongamos que la distribución $F$ de las $u_i$ sea simétrica respecto de 0 con densidad $f$. En este caso se tendrá que la mediana de la distribución $F_{\mu\sigma}$ es $\mu$ y

$$f_{\mu\sigma}(x) = \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right),$$

y por lo tanto,

$$f_{\mu\sigma}(\mu) = \frac{1}{\sigma}f(0).$$

Luego, de acuerdo al Teorema 1, se tendrá

$$n^{1/2}(\tilde{\mu}_n-\mu) \xrightarrow{D} N\left(0, \frac{\sigma^2}{4f^2(0)}\right).$$

Si $F = \Phi$, entonces $f(0) = 1/\sqrt{2\pi}$ y entonces

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N\left(0, \frac{\pi}{2}\sigma^2\right).$$

Por otro lado, $n^{1/2}(\bar{x}_n - \mu)$ tiene distribución N(0,$\sigma^2$). Por lo tanto la varianza asintótica de $\hat{\mu}_n$ es aproximadamente 57% más alta que la varianza de $\bar{x}_n$. Esto significa que la propiedad que tiene la mediana de ser poco sensible a observaciones atípicas tiene como contrapartida negativa ser 57% menos eficiente que $\bar{x}_n$ en el caso de errores normales. De todas maneras esto es menos grave que el comportamiento de $\bar{x}_n$ bajo una contaminación con outliers. En efecto, recordemos que en este caso una fracción de outliers tan pequeña como se quisiera podía provocar que la varianza se hiciese infinita.

Sin embargo, lo ideal sería tener un estimador robusto, es decir poco sensible a outliers y que simultáneamente fuera altamente eficiente cuando los datos son normales. En las secciones siguientes vamos a tratar entonces de encontrar estimadores con estas propiedades.

## 7.2 M-estimadores de posición

### 7.2.1 Definición de M-estimadores

Consideremos el modelo (7.1) y supongamos que conozcamos la distribución $F$ de las $u_i$. y el parámetro de escala $\sigma$. Estas hipótesis no son muy realistas y más adelante las eliminaremos. Sin embargo será conveniente suponerlas momentáneamente para simplificar el planteo del problema. Supongamos que $F$ tiene una densidad que llamaremos $f = F'$. Luego, la densidad de cada $x_i$ será

$$f_{\mu\sigma}(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right),$$

y luego la función de verosimilitud correspondiente a la muestra $x_1, ..., x_n$ será

$$L(\mu) = \frac{1}{\sigma^n} \prod_{i=1}^{n} f\left(\frac{x_i - \mu}{\sigma}\right).$$

Tomando logaritmos, como $\sigma$ se supone conocida, se tendrá que el estimador de máxima verosimilitud de $\mu$ que llamaremos $\hat{\mu}_f$( la $f$ como subscripto indica que corresponde a que las $u_i$ tienen densidad $f$) estará dado por el valor que maximiza

$$\sum_{i=1}^{n} \log f\left(\frac{x_i - \mu}{\sigma}\right).$$

Equivalentemente, podemos decir que $\hat{\mu}_f$ minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho_f\left(\frac{x_i - \mu}{\sigma}\right), \tag{7.12}$$

donde

$$\rho_f(u) = -\log f(u) + \log f(0).$$

Por ejemplo, si $f$ corresponde a la distribución N(0,1). Entonces $\rho_f(u) = u^2/2$, y entonces el estimador de máxima verosimilitud minimiza

$$S(\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2,$$

o equivalentemente, el que minimiza

$$S(\mu) = \sum_{i=1}^{n}(x_i - \mu)^2,$$

el cual es precisamente $\bar{x}_n$.

Si $f$ corresponde a la distribución doble exponencial, entonces

$$f(u) = \frac{1}{2}e^{-|u|},\ -\infty < u < \infty,$$

y por lo tanto $\rho_f(u) = |u|$. Entonces en este caso el estimador de máxima verosimilitud corresponde a minimizar

$$S(\mu) = \sum_{i=1}^{n} |x_i - \mu|, \tag{7.13}$$

y el valor que minimiza (7.13) es precisamente la mediana de la muestra.

En el párrafo anterior hemos visto los inconvenientes de media y la mediana muestral. Si conociéramos exactamente $f$, podríamos utilizar el estimador de máxima verosimilitud, del cual conocemos que tiene varianza asintótica mínima y que está dado por (7.12). Como en general se tiene sólo un conocimiento aproximado de $f$, por ejemplo que corresponde a una distribución de $\mathcal{V}_\epsilon$, Huber (1964) definió los M-estimadores para el modelo de posición como el valor $\hat{\mu}$ valor que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right), \tag{7.14}$$

donde la función $\rho$ es elegida independientemente de $f$ y de tal manera que tenga las propiedades deseadas:

1. El estimador es altamente eficiente cuando $f$ corresponde a la distribución N(0,1)

2. El estimador es poco sensible a contaminación por outliers, en particular es altamente eficiente para toda $f$ correspondiente a una distribución de $\mathcal{V}_\epsilon$.

A la función $\rho$ que define al M-estimador se le pedirá las siguientes propiedades

**A1** La función $\rho$ es derivable. Denominaremos $\psi = \rho'$.

**A2** La función $\rho$ es par.

**A3** La función $\rho(u)$ es monótona no decreciente en $|u|$.

**A4** Se cumple que $\rho(0) = 0$.

Huber (1964) propuso una familia de funciones $\rho$ intermedias entre las correspondientes a la distribución N(0,1) y a la doble exponencial. Esta funciones es cuadrática para valores de valor absoluto pequeños y lineal para valores absolutos grandes. Más precisamente, para cada $k \geq 0$ se define $\rho_k^H$ por

$$\rho_k^H(u) = \begin{cases} -ku - k^2/2 & \text{si} \quad u < -k \\ u^2/2 & \text{si} \quad |u| \leq k \\ ku - k^2/2 & \text{si} \quad u > k \end{cases} .$$

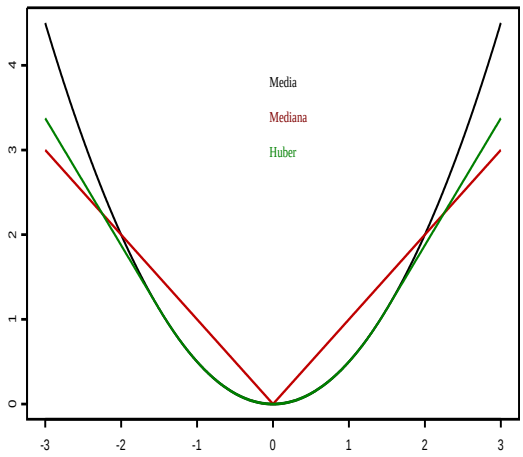


Figure 7.1: Funciones $\rho$ correspondientes a la Media (en negro), la mediana (en rojo) y el M-estimador con función de Huber (en verde)

En la Figura 7.1 se grafican las funciones $\rho$ correspondiente la media a la mediana y a la función de Huber. Obsérvese que las funciones $\rho_k^H$ resultan derivables en todos los puntos, incluidos los puntos de cambio $k$ y $-k$. Más adelante mostraremos que eligiendo $k$ convenientemente los M-estimadores basadas en estas funciones gozan de las propiedades 1 y 2 enunciadas en esta sección.

Para encontrar el valor mínimo de $S(\mu)$ en (7.14) que define el M-estimador podemos encontrar sus punto críticos derivando. De esta manera obtenemos la siguiente ecuación

$$A(\mu) = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0. \tag{7.15}$$

El siguiente Teorema muestra que bajo ciertas condiciones la ecuación 7.15 tiene solución y corresponde a un mínimo de $S(\mu)$.

**Teorema 2.** Supongamos que $\psi$ es continua impar, no decreciente y para algún $a$ se tiene $\psi(a) > 0$. Entonces

**(i)** La ecuación (7.15) tiene al menos una raíz.

**(ii)** Toda raíz de (7.15) corresponde a un mínimo de $S(\mu)$.

**(iii)** Las raíces de (7.15) forman un intervalo.

**(iv)** Si $\psi$ es estrictamente creciente hay una única raíz de (7.15).

Demostración. (i) Sea $M = \max_{1\leq i\leq n} x_i$ y $m = \min_{1\leq i\leq n} x_i$ . Sea $\mu_1 = m-\sigma a$ y $\mu_2 = M+\sigma a$. Luego $(x_i-\mu_1)/\sigma \geq a$ para todo $i$ y $(x_i-\mu_2)/\sigma \leq -a$ para todo $i$. Luego $\psi((x_i-\mu_1)/\sigma) \geq \psi(a) > 0$ para todo $i$ y $\psi((x_i-\mu_2)/\sigma) \leq \psi(-a) = -\psi(a) < 0$ para todo $i$. Luego $A(\mu_1) > 0$ y $A(\mu_2) < 0$. Como $A(\mu)$ es continua, existe un punto $\mu_0$ entre $\mu_2$ y $\mu_1$ tal que $A(\mu_0) = 0$.

(ii) Como $S'(\mu) = (-1/\sigma)A(\mu)$,es fácil ver que $S(\mu) - S(\mu_0) = (-1/\sigma)\int_{\mu_0}^{\mu} A(u)du$. Supongamos que $\mu_0$ es una raíz de $A(\mu)$. Supongamos que $\mu_0 > 0$. Habrá que mostrar que

$$S(\mu_0) \leq S(\mu), \forall \mu. \tag{7.16}$$

Vamos a mostrar (7.16) solamente para $\mu > \mu_0$. El caso $\mu < \mu_0$ se demostrará similarmente. Tomemos $\mu > \mu_0$, luego

$$S(\mu) = \frac{1}{\sigma}\int_{\mu_0}^{\mu} A(u)du.$$

Como $\psi$ es no decreciente resulta $A$ no creciente. Luego como $A(\mu_0) = 0$, resulta $A(\mu) \leq 0$ para $\mu > \mu_0$. Por lo tanto resulta $\int_{\mu_0}^{\mu} A(u)du \leq 0$, y por lo tanto

$$S(\mu) \geq S(\mu_0).$$

En el caso $\mu < \mu_0$ se demuestra similarmente que también vale (7.16).

(iii) Supongamos que $\mu_1 < \mu_2$ sean raíces de $A$, y sea un valor $\mu$ tal que $\mu_1 < \mu < \mu_2$. Tenemos que mostrar que también $A(\mu) = 0$. Como $A$ es no creciente se tendrá

$$0 = A(\mu_1) \geq A(\mu) \geq A(\mu_2) = 0.$$

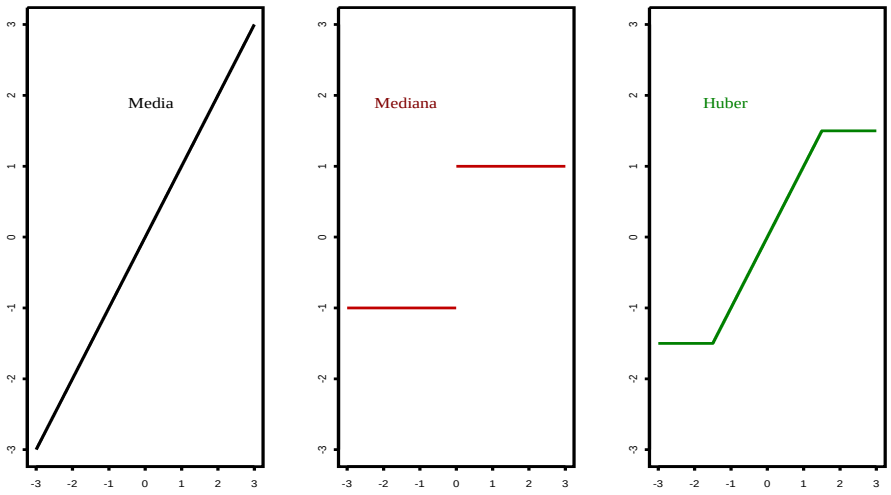


Figure 7.2: Funciones $\psi$ correspondientes a la Media (en negro), la mediana (en rojo) y el M-estimador con función de Huber (en verde)

y luego $A(\mu) = 0$.

(iv) Supongamos que $A(\mu) = 0$. Veremos que no puede haber otra raíz de $A$. Sea primero $\mu^* > \mu$, como en este caso $A$ es estrictamente decreciente se tendrá $A(\mu^*) < 0$. Similarmente se demuestra que si $\mu^* < \mu$, entonces $A(\mu^*) > 0$. □

Como vamos a ver más adelante la función $\psi$ cumple un papel muy importante en la teoría de M-estimadores. Para la función $\rho$ correspondiente a la media, resulta $\psi(u) = u$, para la función $\rho$ correspondiente mediana $\psi(u) = |u|$, y para la funciones $\rho_k^H$, las correspondientes derivadas $\psi_k^H$ están dadas por

$$\psi_k^H(u) = \begin{cases} -k & \text{si} \quad u < -k \\ u & \text{si} \quad |u| \leq k \\ k & \text{si} \quad u > k \end{cases} .$$

la cual corresponde a una identidad truncada. En Fig. 7.2 se grafican estas tres funciones $\psi$.

Como consecuencia de la propiedad A2, la función $\psi$ es impar .

Para que el M-estimador sea robusto como veremos más adelante se requerirá que la función $\psi$ sea acotada.

### 7.2.2 Propiedades asintóticas de M-estimadores

La condición de consistencia de Fisher, requerida para que el M-estimador converja a $\mu$ está dada por

$$E_{F_{\mu\sigma}}\left(\psi\left(\frac{x-\mu}{\sigma}\right)\right)=0,$$

y de acuerdo a (7.1), esto es equivalente a

$$E_F(\psi(u))=0. \tag{7.17}$$

Esta condición se cumple automaticamente si $F$ tiene una densidad simétrica respecto de 0 ya que en ese caso se tendrá

$$E_F(\psi(u))=\int_{\infty}^{\infty} uf(u)du=0,$$

ya que $uf(u)$ será una función impar.

Luego, se tendrá el siguiente Teorema que muestra la consistencia de los M-estimadores:

**Teorema 3.** Sean $x_1,...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.15), donde $\psi$ y $F$ satisfacen (7.17) . Luego $\hat{\mu}_n$ converge en casi todo punto a $\mu$ en cualquiera de los siguientes casos

1. La función $\psi$ es estrictamente creciente.
2. La función $\psi$ es no decreciente, $\psi(u)>\psi(0)$ y $F(u)>F(0)$ para todo $u>0$.

Demostración: Solamente mostraremos el Teorema para el caso 1. Consideremos $\epsilon>0$. Luego como $\psi$ es estrictamente creciente tenemos que $\psi(u-\epsilon)<\psi(u)$, y luego

$$E_F\psi(u-\epsilon)<E_F\psi(u)=0.$$

Por lo tanto

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu+\epsilon)}{\sigma}\right) = E_F\psi(u-\epsilon) < 0. \tag{7.18}$$

Similarmente se puede probar que

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu-\epsilon)}{\sigma}\right) = E_F\psi(u+\epsilon) > 0. \tag{7.19}$$

Sea ahora

$$G_n(\mu^*) = \frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{x_i-\mu^*}{\sigma}\right),$$

luego el M-estimador $\hat{\mu}_n$ satisface

$$G_n(\hat{\mu}_n) = 0. \tag{7.20}$$

Por otro lado usando la ley de los grandes números y (7.18) y (7.19) se tiene que con probabilidad 1 existe un $n_0$ tal que para todo $n > n_0$ se tiene que

$$G_n(\mu+\epsilon) < 0,\ G_n(\mu-\epsilon) > 0,$$

y por lo tanto como $G_n$ es monótona decreciente, se tiene que el valor $\hat{\mu}_n$ satisfaciendo (7.20) tendrá que satisfacer que

$$\mu - \epsilon < \hat{\mu}_n < \mu + \epsilon.$$

Esto prueba la consistencia de $\hat{\mu}_n$.

El siguiente teorema muestra la asintótica normalidad de los M-estimadores

**Teorema 4.** Sean $x_1, ...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.15), donde $\psi$ y $F$ satisfacen (7.17). Supongamos que $\hat{\mu}_n$ es consistente, y que además $\psi$ tiene dos derivadas continuas y $\psi''$ es acotada. Luego se tiene que

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N(0, \sigma^2 V(\psi, F)),$$

donde

$$V(\psi, F) = \frac{E_F\psi^2(u)}{(E_F\psi'(u))^2}. \tag{7.21}$$

Demostración. El M-estimador $\hat{\mu}_n$ satisface

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \hat{\mu}_n}{\sigma}\right) = 0,$$

y haciendo un desarrollo de Taylor en el punto $\mu$ se tiene

$$0 = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) - \sum_{i=1}^{n} \psi'\left(\frac{x_i - \mu}{\sigma}\right)\frac{\hat{\mu}_n - \mu}{\sigma} + \frac{1}{2}\sum_{i=1}^{n} \psi''\left(\frac{x_i - \mu_n^*}{\sigma}\right)\frac{(\hat{\mu}_n - \mu)^2}{\sigma^2},$$

donde $\mu_n^*$ es un punto intermedio entre $\hat{\mu}_n$ y $\mu$.

Luego, haciendo un despeje parcial de $(\hat{\mu}_n - \mu)$ se tiene

$$(\hat{\mu}_n - \mu) = \frac{\displaystyle\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right)}{\displaystyle\frac{1}{\sigma}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) - \frac{1}{2}\frac{(\hat{\mu}_n - \mu)}{\sigma^2}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right)},$$

y luego

$$n^{1/2}(\hat{\mu}_n - \mu) = \frac{\displaystyle\frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right)}{\displaystyle\frac{1}{n\sigma}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) - \frac{1}{2\sigma^2}(\hat{\mu}_n - \mu)\frac{1}{n}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right)}. \tag{7.22}$$

Sea

$$A_n = \frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left((x_i - \mu)/\sigma\right) = \frac{1}{n^{1/2}}\sum_{i=1}^{n} \psi\left(u_i\right),$$

$$B_n = \frac{1}{n}\sum_{i=1}^{n} \psi'\left((x_i - \mu)/\sigma\right) = \frac{1}{n}\sum_{i=1}^{n} \psi'\left(u_i\right),$$

y

$$C_n = \frac{1}{2}(\hat{\mu}_n - \mu)\frac{1}{n}\sum_{i=1}^{n} \psi''\left((x_i - \mu_n^*)/\sigma\right).$$

Luego

$$n^{1/2}(\hat{\mu}_n - \mu) = \frac{A_n}{\sigma^{-1}\ B_n + \sigma^{-2} C_n}. \tag{7.23}$$

Por el Teorema Central del Límite se tiene

$$A_n \xrightarrow{D} N(0, E_F(\psi^2(u))). \tag{7.24}$$

Por la Ley Fuerte de los Grandes Números se tiene

$$B_n \xrightarrow{p} E_F(\psi'(u)). \tag{7.25}$$

Finalmente, por hipótesis existe una constante $K$ tal que $|\psi''(u)| < K$. Luego $|C_n| < (K/2)(\hat{\mu}_n - \mu)$. Usando el hecho de que $\hat{\mu}_n \xrightarrow{p} \mu$, se tiene que

$$C_n \xrightarrow{p} 0. \tag{7.26}$$

Usando (7.23)-(7.26) se deduce el Teorema. □

### 7.2.3 M-estimador minimax para la varianza asintótica

El problema que vamos a desarrollar en esta sección es el de elegir la función $\rho$ o equivalentemente la función $\psi$ del M-estimador. En está sección vamos a utilizar como criterio minimizar la varianza asintótica del M-estimador dada en (7.21). Si conociéramos la distribución $F$ de las $u_i$, utilizaríamos el M-estimador que tiene como función $\psi$ la dada por

$$\psi(u) = \frac{d \log f(u)}{du},$$

es decir el estimador de máxima verosimilitud. Este estimador minimiza la varianza asintótica $V(\psi, F)$ dada en (7.21). Cuando existe la posibilidad de que hubieran outliers la distribución $F$ no es conocida exactamente y por lo tanto no podemos usar este estimador.

La solución que propuso Huber (1964) es la siguiente. supongamos que $F$ esté en el entorno de contaminación dado por (7.2), pero restringiendo $H$ a distribuciones simétricas respecto de 0. Para esto definimos un nuevo entorno de distribuciones de $\Phi$

$$\mathcal{V}_\epsilon^* = \{F : F = (1-\epsilon)\Phi + \epsilon H \text{ con } H \text{ simétrica}\}. \tag{7.27}$$

Luego, si usa el M-estimador basado en la función $\psi$. la mayor varianza posible en este entorno está dada por

$$V^*(\psi) = \sup_{F \in \mathcal{V}_\epsilon^*} V(\psi, F).$$

El criterio de Huber para elegir el M-estimador es utilizar la función $\psi^*$ que minimice $V^*(\psi)$. Estos estimadores se denominarán minimax (minimizan la máxima varianza asintótica en el entorno de contaminación $V_\varepsilon^*$. En Huber (1964) se muestra que $\psi^*$ está en la familia $\psi_k^H$, donde $k$ depende de la cantidad de contaminación $\epsilon$.

### 7.2.4 M-estimadores con escala desconocida

La definición de los M-estimadores dada en (7.14) supone que $\sigma$ es conocida. Sin embargo, en la práctica $\sigma$ es desconocida. En estos casos podemos reemplazar en esta ecuación $\sigma$ por un estimador $\hat{\sigma}$, y el M-estimador se definirá por el valor $\hat{\mu}$ que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}_n}\right). \tag{7.28}$$

Si queremos que el M-estimador resultante de $\mu$ sea robusto, será necesario que $\hat{\sigma}$ también lo sea. El estimador insesgado usual de $\sigma$ dado por

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}(x_i - \bar{x})^2$$

no es robusto. En efecto, es fácil ver que una observación lo pueda llevar fuera de todo límite. Un estimador robusto de $\sigma$ es el llamado MAD (median absolute deviation), que está definido por

$$\hat{\sigma}^2 = A \text{ mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \le i \le n\},$$

donde

$$\tilde{\mu}_n = \text{mediana}\{x_i : 1 \le i \le n\},$$

y donde $A$ es una constante que hace que el estimador sea consistente a $\sigma$ en el caso de que las observaciones sean una muestra aleatoria de una $\mathrm{N}(\mu, \sigma^2)$.

Vamos ahora a deducir cual debe ser el valor de $A$. Sean $x_1, ..., x_n$ una muestra de una distribución N($\mu$,$\sigma^2$). Entonces podemos escribir $x_i = \mu + \sigma u_i$, donde $u_1, ..., u_n$ es una muestra aleatoria de una distribución N(0,1). En este caso tenemos que

$$x_i - \tilde{\mu}_n = (\mu - \tilde{\mu}_n) + \sigma u_i$$

y

$$\text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \text{mediana}\{|(\mu - \tilde{\mu}_n) + \sigma u_i|,\ 1 \leq i \leq n\}.$$

Como de acuerdo a lo visto en Observación 1, $\lim(\mu - \tilde{\mu}_n) = 0$ casi seguramente, se tendrá que

$$\lim_{n\to\infty} \text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \lim_{n\to\infty} \text{mediana}\{|\sigma u_i|,\ 1 \leq i \leq n\}\ \}$$

$$= \sigma \lim_{n\to\infty} \text{mediana}\{|u_i|,\ 1 \leq i \leq n\},\ \text{c.s.}. \qquad (7.29)$$

Si $u$ es N(0,1), entonces $|u|$ tiene distribución $2\Phi - 1$. Sea entonces $B = \text{mediana}(2\Phi - 1)$, luego por lo visto en Observación 1 se tiene

$$\lim_{n\to\infty} \text{mediana}\{|u_i|,\ 1 \leq i \leq n\} = B,\ \text{c.s.}$$

y usando (7.29)

$$\lim_{n\to\infty} \text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\} = \sigma B \text{ c.s.}$$

Luego $A = 1/B$. La constante $B$ se calcula de la siguiente manera

$$2\Phi(B) - 1 = 0.5,$$

o sea

$$\Phi(B) = 0.75,\ B = \Phi^{-1}(0.75) = 0.675.$$

Luego se tendrá que el estimador MAD de $\sigma$ viene dado por

$$\hat{\sigma}^2 = \frac{1}{0.6745}\text{mediana}\{|x_i - \tilde{\mu}_n|,\ 1 \leq i \leq n\}.$$

Cuando el M-estimador se obtiene minimizando (7.28), la ecuación (7.15) se transforma en

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\hat{\sigma}}\right) = 0. \tag{7.30}$$

Las propiedades asintóticas del estimador $\hat{\mu}$ solución de (7.30) son similares a las del estimador correspondiente al caso de $\sigma$ conocida. El siguiente Teorema se dará sin demostración.

**Teorema 5.** Sean $x_1, ...x_n$ variables aleatorias independientes que satisfacen el modelo (7.1). Consideremos un estimador $\hat{\mu}_n$ solución de (7.30), donde $\psi$ es impar y $F$ es simétrica respecto de 0. Supongamos que $\hat{\mu}_n$ es consistente a $\mu$ y $\hat{\sigma}_n$ es consistente a $\sigma$, y que además $\psi$ tiene dos derivadas continuas y $\psi''$ es acotada. Luego se tiene que

$$n^{1/2}(\hat{\mu}_n - \mu) \xrightarrow{D} N(0, \sigma^2 V(\psi, F)),$$

donde $V$ está dada por (7.21)

### 7.2.5 Algoritmos para calcular M-estimadores

A continuación vamos a describir tres algoritmos para computar el M-estimador definido como la solución de (7.30).

#### Algoritmo basado en medias ponderadas iteradas (MPI)

Llamemos $w(u) = \psi(u)/u$. Luego la ecuación (7.30).se puede escribir como

$$\sum_{i=1}^{n} (x_i - \hat{\mu}) w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = 0,$$

o sea

$$\sum_{i=1}^{n} x_i w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right) = \hat{\mu} w\left(\frac{x_i - \hat{\mu}}{\hat{\sigma}}\right),$$

y haciendo un despeje "parcial"de $\hat{\mu}$ se tiene

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{7.31}$$

En realidad esta expresión no es un verdadero despeje, ya que el miembro derecho también aparece $\hat{\mu}$. Sin embargo esta fórmula nos va a sugerir un algoritmo iterativo para calcular $\hat{\mu}$.

En efecto, consideremos un estimador inicial $\hat{\mu}_0$ de $\mu$, como por ejemplo la mediana.Luego podemos definir

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu}_0)/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu}_0)/\hat{\sigma}\right)},$$

y en general si ya tememos definido $\hat{\mu}_h$, podemos definir $\hat{\mu}_{h+1}$ por

$$\hat{\mu}_{h+1} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \tag{7.32}$$

Se puede mostrar que este si $\psi$ es continua, entonces cuando este algoritmo iterativo converge, lo hace a una solución de (7.30). En efecto supongamos que $\lim_{h\to\infty} \hat{\mu}_h = \hat{\mu}$, luego tomando limite en ambos lados de (7.32), se tendrá

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{7.33}$$

Pero esta ecuación es precisamente (7.31) , que ya hemos visto es equivalente a (7.30).

La ecuación (7.33) muestra a $\hat{\mu}$ como promedio pesado de las $x_i$ y pesos proporcionales a $w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)$. Como en general $w(u)$ es una función par monótona no creciente en $|u|$, (7.33) se puede interpretar como que el M-estimador da a cada observación un peso que penaliza las observaciones para las cuales $|x_i - \hat{\mu}|/\hat{\sigma}$ es grande. Para la media se tiene $w(u) = 1$, y para el estimador basado en la función $\psi_k^H$, la correspondiente función de peso está dada por

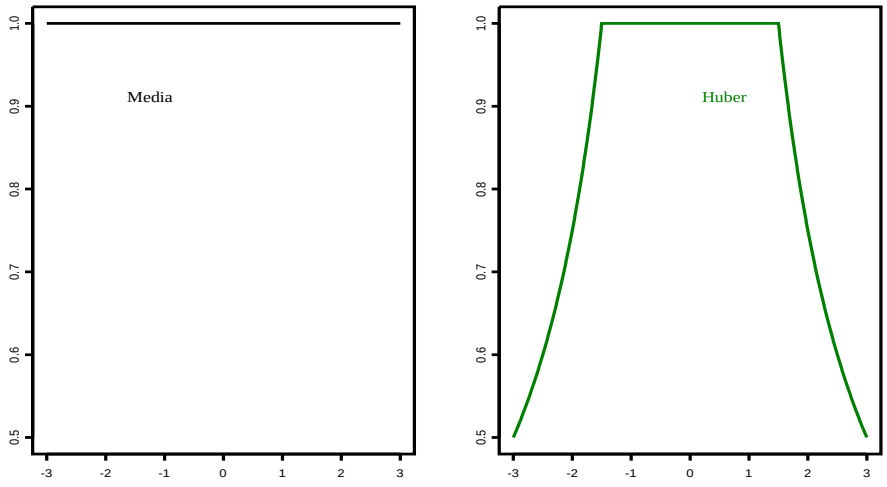


Figure 7.3: Funciones de peso $w$ correspondientes a la Media (en negro) y al M-estimador con función de Huber (en verde)

$$w_k^H(u) = \begin{cases} 1 & \text{si} \quad |u| \leq k \\ \frac{k}{|u|} & \text{si} \quad |u| > k \end{cases} .$$

El gráfico de esta función se encuentra en la Figura 7.3.

**Algoritmo basado en medias de pseudovalores iteradas (MPVI)**

Definamos el pseudovalor $x_i^*(\mu)$ por

$$x_i^*(\mu) = \mu + \hat{\sigma}\psi\left((x_i - \hat{\mu})/\hat{\sigma}\right).$$

Luego se tiene

$$\psi\left((x_i - \hat{\mu})/\hat{\sigma}\right) = (x_i^*(\mu) - \hat{\mu})/\hat{\sigma},$$

y reemplazando en (7.30) se tiene la ecuación para el M-estimador es

$$\sum_{i=1}^{n}(x_i^*(\hat{\mu}) - \hat{\mu})/\hat{\sigma} = 0.$$

Haciendo un despeje parcial de $\hat{\mu}$ se tiene

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}). \tag{7.34}$$

Es decir, se tiene expresado el M-estimador como promedio simple de los pseudo valores. Esta fórmula no permite calcular el M-estimador directamente, ya que el miembro derecho también depende de $\hat{\mu}$. Sin embargo, nos sugiere el siguiente algoritmo iterativo. Partiendo de un estimador inicial $\hat{\mu}_0$, consideramos la siguiente fórmula recursiva para $\hat{\mu}_h$

$$\hat{\mu}_{h+1} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}_h). \tag{7.35}$$

Es interesante calcular los pseudovalores correspondientes a $\psi_k^H$, los cuales están dados por

$$x_i^*(\mu) = \begin{cases} \mu - k\hat{\sigma} & \text{si} \quad x_i < \mu - k\hat{\sigma} \\ x_i & \text{si} \quad |x_i - \mu| \leq k\hat{\sigma} \\ \mu + k\hat{\sigma}\ V & \text{si} \quad x_i > \mu + k\hat{\sigma} \end{cases}.$$

Es decir, si $x_i$ pertenece al intervalo $[\mu - k\hat{\sigma},\ \mu + k\hat{\sigma}]$, el pseudovalor $x_i^*(\mu)$ es igual a la observación $x_i$. Si $x_i$ está fuera de este intervalo el psudovalor se define como el extremo del intervalo más cercano.

Vamos a ver ahora que si $\lim_{h\to\infty} \hat{\mu}_h = \hat{\mu}$ y $\psi$ es continua, entonces $\hat{\mu}$ es el M-estimador solución de (7.30). En efecto, tomando límite en ambos miembros de (7.35) se obtiene (7.34), que ya hemos visto es equivalente a (7.30).

### Algoritmo de Newton Raphson (NR)

De acuerdo a lo visto anteriormente, el algoritmo de Newton Raphson para calcular la raíz de (7.30) tiene la siguiente fórmula recursiva

$$\hat{\mu}_{h+1} = \hat{\mu}_h + \hat{\sigma}\frac{\sum_{i=1}^{n} \psi\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n} \psi'\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \tag{7.36}$$

Para el caso de que $\psi = \psi_k^H$, está formula toma una expresión particularmente interesante.

Para cada valor $\mu$ dividamos el conjunto de observaciones en tres conjuntos

$$\begin{aligned} D_- &= \{i : (x_i - \hat{\mu}_h)/\hat{\sigma} < -k\}, \\ D_0 &= \{i : |x_i - \hat{\mu}_h|/\hat{\sigma} \leq k\}, \\ D_+ &= \{i : (x_i - \hat{\mu}_h)/\hat{\sigma} > k\}. \end{aligned}$$

Es fácil ver que se tiene

$$\psi_k^H\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right) = \begin{cases} -k & \text{si} \quad i \in D_- \\ (x_i - \hat{\mu}_h)/\hat{\sigma} & \text{si} \quad i \in D_0 \\ k & \text{si} \quad i \in D_+ \end{cases},$$

y

$$\psi_k^{H\prime}\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right) = \begin{cases} 0 & \text{si} \quad i \in D_-(\hat{\mu}_h) \\ 1 & \text{si} \quad i \in D_0(\hat{\mu}_h) \\ 0 & \text{si} \quad i \in D_+(\hat{\mu}_h) \end{cases}.$$

Llamando $n_-$, $n_0$ y $n_-$, al número de elementos de $D_-$, $D_0$ y $D_+$ y reemplazando en (7.36), se tiene

$$\hat{\mu}_{h+1} = \hat{\mu}_h + \hat{\sigma}\frac{k(n_+ - n_-) + \sum\limits_{i \in D_0}(x_i - \hat{\mu}_h)/\hat{\sigma}}{n_0} = \frac{n_+ - n_-}{n_0}\hat{\sigma}k + \frac{1}{n_0}\sum_{i \in D_0} x_i.$$

Obsérvese que el miembro derecho de esta última fórmula solo depende de $D_-$, $D_0$ y $D_+$. Estos tres conjuntos forman una partición del conjunto $\{1, 2, ..., n\}$. Es claro que hay un número finito de estas particiones, y por lo tanto si $\hat{\mu}_h$ converge lo debe hacer en un número finito de pasos.

### Convergencia de los algoritmos iterativos

Se puede demostrar que los 3 algoritmos iterativos que hemos estudiado MPI, MPVI, y NR convergen a la raíz de (7.30) cuando $\psi$ es monótona no decreciente cuando ésta es única. Si (7.30) tiene más de una raíz, se puede demostrar que si $[\hat{\mu}_1,\ \hat{\mu}_2]$ es el intervalo de soluciones, entonces dado $\epsilon > 0$, existe $h_0$ tal que $\hat{\mu}_h \in [\hat{\mu}_1 - \epsilon,\ \hat{\mu}_2 + \epsilon]$ para todo $h > h_0$.