Chapter 10

Estimación Robusta

10.1 El problema de la robustez para el modelo de posición

Sea el modelo de posición y escala

$$x_i = \mu + \sigma u_i, 1 \leq i \leq n, \tag{10.1}$$

donde μ y σ 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 μ medida con un error σu_i .

Si $F = \Phi$, la función de distribución N(0,1), entonces las x_i tienen distribución N(μ, σ^2). Por lo tanto, un estimador óptimo de μ 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 que los errores de medición tienen distribución *aproximadamente normal*. Por lo tanto cabe preguntarse cual será el comportamiento del estimador $\bar{x}$ en este caso.

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

Φ de la N(0,1). Un entorno de contaminación de tamaño ϵ de la distribución Φ se define por

$$\mathcal{V}_\epsilon = \{F : F = (1-\epsilon)\Phi + \epsilon^* F^* \text{ con } F^* \text{ arbitraria}\}. \tag{10.2}$$

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

En efecto supongamos que se tienen tres variables aleatoria independientes : V con distribución Φ, V^* con distribución F^*, y W con distribución $\mathrm{Bi}(\epsilon, 1)$. Definamos entonces la variable aleatoria U de la siguiente manera

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

Luego

$$F_U(u) = P(U \leq u) = P(U \leq u,\ W = 0) + P(U \leq u,\ W = 1)$$

$$= P(U \leq u|\ W = 0)P(W = 0) + P(U \leq u|\ W = 1)P(W = 1)$$

$$= (1-\epsilon)\Phi(u) + \epsilon F^*(u).$$

Por lo tanto si ϵ es pequeño (por ejemplo .05 o .10) esto significará que la gran mayoría de las observaciones se obtendrán a partir de la distribución Φ, es decir serán normales. Por lo tanto podemos afirmar que si ϵ es pequeño y $F \in \mathcal{V}_\epsilon$, entonces F está cerca de Φ. 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 (10.1)con u_i proveniente de una distribución Φ, y una proporción ϵ tendrán el correspondiente u_i proveniente de la distribución F^*. 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 ϵ sea pequeño el comportamiento del estimador $\bar{x}$ puede ser muy ineficiente para distribuciones $F \in \mathcal{V}_\epsilon$.

Primero mostraremos que si

$$F = (1-\epsilon)\Phi + \epsilon^* F^*, \tag{10.3}$$

entonces

$$E_F(u) = (1-\epsilon)E_\Phi(u) + \epsilon E_{F^*}(u). \tag{10.4}$$

Además si $E_{F^*}(u) = 0$, se tiene

$$\mathrm{var}_F(u) = (1-\epsilon)\mathrm{var}_\Phi(u) + \epsilon \mathrm{var}_{F^*}(u). \tag{10.5}$$

Para mostrar (10.4) supongamos que la F^* tiene densidad f^* , y sea φ la densidad correspondiente a Φ. Luego la densidad de F es

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

y luego

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

Para mostrar (10.5), observemos que

$$\mathrm{var}_F(u) = \textstyle\int_{-\infty}^{\infty} u^2 f(u)du = (1-\epsilon)\int_{-\infty}^{\infty} u^2\varphi(u)du + \epsilon\int_{-\infty}^{\infty} u^2 f^*(u)du =$$

$$(1-\epsilon) + \epsilon var_{F^*}(u).$$

Consideremos ahora al estimador $\hat{\mu} = \bar{x}$, donde la muestra $x_1, ..., x_n$ son generadas por (10.1) donde las u_i son independientes con distribución dada por (10.3) con $E_{F^*}(u) = 0$

Luego

$$\mathrm{var}_F(\bar{x}) = \frac{\sigma^2\ \mathrm{var}_F(u)}{n} = \frac{\sigma^2((1-\epsilon) + \epsilon\mathrm{var}_{F^*}(u))}{n}.$$

Luego si $\epsilon = 0$, entonces $\mathrm{var}(\bar{x}) = \sigma^2/n$. En cambio una contaminación de tamaño ϵ puede producir un aumento de la varianza ilimitado, ya que $\mathrm{var}_{F^*}(u)$ 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 $\bar{x}_{n+1}$ 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 (10.1) donde la distribución F de los u_i es simétrica respecto de 0. Como en este caso μ 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

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

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 que 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 aun 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 aleatorias independientes igualmente distribuidas. 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

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

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 μ y con una densidad f tal que es continua y positiva en μ Entonces si $\tilde{\mu}_n$ es la mediana de la muestra, se tiene que

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

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

$$\lim_{n\to\infty} P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = \Phi(2f(\mu)y), \tag{10.6}$$

donde Φ es la función de distribución correspondiente a N(0,1)

Es inmediato que

$$P(n^{1/2}(\tilde{\mu}_n - \mu) \leq y) = P\left(\tilde{\mu}_n \leq \mu + \frac{y}{n^{1/2}}\right). \tag{10.7}$$

Sea

$$v_{ni} = \begin{cases} 1 & \text{si} \quad x_i \leq \mu + \frac{y}{n^{1/2}} \\ 0 & \text{si} \quad x_i > \mu + \frac{y}{n^{1/2}} \end{cases}, \ 1 \leq i \leq n. \tag{10.8}$$

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

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

$$= P\left(\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{(v_{ni} - \nu_n)}{(\nu_n(1-\nu_n))^{1/2}} \geq \frac{(n/2 - n\nu_n)}{(n\nu_n(1-\nu_n))^{1/2}}\right). \tag{10.9}$$

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

$$\frac{1}{n^{1/2}}\sum_{i=1}^{n}\frac{(v_{ni}-\nu_n)}{(\nu_n(1-\nu_n))^{1/2}} \to_D N(0,1). \tag{10.10}$$

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 μ_n^* es un punto intermedio entre μ y $\mu+n^{-1/2}y$. Luego usando el hecho que $\nu_n \to 1/2$ y $\mu_n^* \to \mu$, resulta

$$\frac{(n/2-n\nu_n)}{(n\nu_n(1-\nu_n))^{1/2}} \to -2yf(\mu). \tag{10.11}$$

Luego, usando (10.7), (10.9), (10.10) y (10.11) tenemos que

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

y por lo tanto hemos probado (10.6)

Observación 1. El Teorema 1 implica que $\tilde{\mu}_n \to_p \mu$. También puede probarse que $\tilde{\mu}_n \to_{\text{c.s.}} \mu$, pero no se dará la demostración.

Apliquemos ahora este resultado al modelo (10.1) 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 μ y

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

y por lo tanto

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

Luego de acuerdo al Teorema 1 se tendrá

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

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

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

Por otro lado $n^{1/2}(\bar{x}_n - \mu)$ tiene distribución N(0,σ^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.

10.2 M-estimadores de posición

10.2.1 Definición de M-estimadores

Consideremos el modelo (10.1) y supongamos que conozcamos la distribución F de las u_i. y el parámetro de escala σ. 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 σ es supuesto conocida, se tendrá que el estimador de máxima verosimilitud de μ 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

$$S(\mu) = \sum_{i=1}^{n} \rho_f\left(\frac{x_i - \mu}{\sigma}\right), \tag{10.12}$$

donde

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

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

$$S(\mu) = \sum_{i=1}^{n} |x_i - \mu|, \tag{10.13}$$

y el valor que minimiza (10.13) es precisamente la mediana de la muestra.

En el párrafo anterior hemos visto los inconvenientes de la 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 10.12. Como en general se tiene solo 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 por un valor que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\sigma}\right), \tag{10.14}$$

donde la función ρ 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 ρ que define al M-estimador se le pedirá las siguientes propiedades

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

A2 La función ρ es par.

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

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

Huber propuso una familia de funciones ρ intermedias entre las correspondientes a la distribución N(0,1) y a la doble exponencial. Esta funciones son cuadráticas para valores de valor absoluto pequeños y líneas para valores absolutos grandes. Mas precisamente para cada $k \geq 0$ se define ρ_k^H por

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

En la Figura 1 se grafican las funciones ρ correspondientes a la media, a la mediana y a la función de Huber. Obsérvese que las funciones ρ_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 basados en estas funciones gozan de las propiedades 1 y 2 enunciadas en esta sección.

Para encontrar el valor mínimo de $S(\mu)$ en (10.14) que define el M-estimador podemos encontrar sus punto críticos derivando. De esta manera obtenemos la siguiente ecuación

$$A(\mu) = \sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\sigma}\right) = 0. \tag{10.15}$$

El siguiente Teorema muestra que bajo ciertas condiciones la ecuación 10.15 tiene solución y corresponde a un mínimo de $S(\mu)$.

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

(i) La ecuación (10.15) tiene al menos una raíz.

(ii) Toda raíz de (10.15) corresponde a un mínimo de $S(\mu)$.

(iii) Las raíces de (10.15) forman un intervalo.

(iv) Si ψ es estrictamente creciente hay una única raíz de 10.15.

Demostración. (i) Sea $M = \max_{1\le i\le n} x_i$ y $m = \min_{1\le i\le n} x_i$. Sea $\mu_1 = m - \sigma a$ y $\mu_2 = M + \sigma a$. Luego $(x_i - \mu_1)/\sigma \ge a$ para todo i y $(x_i - \mu_2)/\sigma \le -a$ para todo i. Luego $\psi((x_i - \mu_1)/\sigma) \ge \psi(a) > 0$ para todo i y $\psi((x_i - \mu_2)/\sigma) \le \psi(-a) = -\psi(a) < 0$ para todo i. Luego $A(\mu_1) > 0$ y $A(\mu_2) < 0$. Como $A(\mu)$ es continua, existe un punto μ_0 entre μ_2 y μ_1 tal que $A(\mu_0) = 0$.

(ii) Como $S'(\mu) = (-1/\sigma)A(\mu)$,es fácil ver que $S(\mu) = (-1/\sigma)\int_0^\mu A(u)du$. Supongamos que μ_0 es una raíz de $A(\mu)$. Supongamos que $\mu_0 > 0$. Habrá que mostrar que

$$S(\mu_0) \le S(\mu), \forall \mu. \tag{10.16}$$

Vamos a mostrar (10.16) solamente para $\mu > \mu_0$. El caso $\mu < \mu_0$ se demostrará similarmente. Tomemos $\mu > \mu_0$, luego

$$S(\mu) = -\frac{1}{\sigma}\int_0^\mu A(u)du = -\frac{1}{\sigma}\int_0^{\mu_0} A(u)du - \frac{1}{\sigma}\int_{\mu_0}^{\mu} A(u)du.$$

Como ψ 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

$$S(\mu) \geq -\frac{1}{\sigma}\int_0^{\mu_0} A(u)du = S(\mu_0).$$

En el caso $\mu < \mu_0$ se demuestra similarmente que también vale (10.16).

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

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

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 ψ cumple un papel muy importante en la teoría de M-estimadores. Para la función ρ correspondiente a la media , resulta $\psi(u) = u$, para la función ρ correspondiente a la mediana $\psi(u) = |u|$, y para la funciones ρ_k^H, las correspondientes derivadas ψ_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. 2 se grafican estas tres funciones ψ .

Como consecuencia de la propiedad A2, la función ψ es impar . Para que el M-estimador sea robusto como veremos más adelante se requerirá que la función ψ sea acotada.

10.2.2 Propiedades asintóticas de M-estimadores

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

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

y de acuerdo a (10.1), esto es equivalente a

$$E_F(\psi(u))=0. \tag{10.17}$$

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

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

ya que $\psi(u)f(u)$ 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 (10.1). Consideremos un estimador $\hat{\mu}_n$ solución de (10.15), donde ψ y F satisfacen (10.17) . Luego $\hat{\mu}_n$ converge en casi todo punto a μ en cualquiera de los siguientes casos

1. La función ψ es estrictamente creciente.

2. La función ψ 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 ψ 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

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu+\epsilon)}{\sigma}\right)=E_F\psi\left(u-\ \frac{\epsilon}{\sigma}\right)<0. \tag{10.18}$$

Similarmente se puede probar que

$$E_{F_{\mu\sigma}}\psi\left(\frac{x-(\mu-\epsilon)}{\sigma}\right)=E_F(\left(u+\frac{\epsilon}{\sigma}\right)>0. \tag{10.19}$$

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

$$G_n(\hat{\mu}_n) = 0. \tag{10.20}$$

Por otro lado usando la ley de los grandes números y (10.18) y (10.19) se tiene que con probabilidad 1 existe un n_0 tal que para todo $n > n_0$ se tiene que

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

y por lo tanto como G_n es monótona decreciente, se tiene que el valor $\hat{\mu}_n$ satisfaciendo (10.20) 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 (10.1). Consideremos un estimador $\hat{\mu}_n$ solución de (10.15), donde ψ y F satisfacen (10.17). Supongamos que $\hat{\mu}_n$ es consistente, y que además ψ tiene dos derivadas continuas y ψ'' es acotada. Luego se tiene que

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

donde

$$V(\psi, F) = \frac{E_F\psi^2(u)}{(E_F\psi'(u))^2}. \tag{10.21}$$

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 μ se tiene

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

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

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

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

y luego

$$n^{1/2}(\hat{\mu}_n-\mu)=\frac{\sigma n^{-1/2}\sum_{i=1}^{n}\psi\left((x_i-\mu)/\sigma\right)}{\frac{1}{n}\sum_{i=1}^{n}\psi'\left((x_i-\mu)/\sigma\right)-\ ((\hat{\mu}_n-\mu)/(2\sigma))\frac{1}{n}\sum_{i=1}^{n}\psi''\left((x_i-\mu_n^*)/\sigma\right)}. \tag{10.22}$$

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

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

Luego

$$n^{1/2}(\hat{\mu}_n-\mu)=\frac{\sigma A_n}{B_n+C_n}. \tag{10.23}$$

Por el Teorema Central del Límite se tiene

$$A_n\rightarrow_D N(0,E_F(\psi^2(u))). \tag{10.24}$$

Por la Ley Fuerte de los Grandes Números se tiene

$$B_n\rightarrow_{c.s.} E_F(\psi'(u)). \tag{10.25}$$

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

$$C_n \to_P 0. \tag{10.26}$$

Usando (10.23)-(10.26) se deduce el Teorema.

10.2.3 M-estimador minimax para la varianza asintótica

El problema que vamos a desarrollar en esta sección es el de elegir la función ρ o equivalentemente la función ψ del M-estimador. En esta sección vamos a utilizar como criterio minimizar la varianza asintótica del M-estimador dada en (10.21). Si conociéramos la distribución F de las u_i, utilizaríamos el M-estimador que tiene como función ψ 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 (10.21). 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 (10.2), pero restringiendo F^* a distribuciones simétricas respecto de 0. Para esto definimos un nuevo entorno de distribuciones de Φ

$$\mathcal{V}_\epsilon^* = \{F : F = (1-\epsilon)\Phi + \epsilon^* F^* \text{ con } F^* \text{ simétrica}\}. \tag{10.27}$$

Luego, si se usa el M-estimador basado en la función ψ. 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 ψ^* que minimice $V^*(\psi)$. Estos estimadores se denominarán minimax (minimizan la máxima varianza asintótica en el entorno de contaminación V_ε^*. En Huber (1964) se muestra que ψ^* está en la familia ψ_k^H, donde k depende de la cantidad de contaminación ϵ.

10.2.4 M-estimadores con escala desconocida

La definición de los M-estimadores dada en (10.14) supone que σ es conocida. Sin embargo en la práctica σ es desconocida. En estos casos podemos reemplazar en esta ecuación σ por un estimador $\hat{\sigma}$, y el M-estimador se definirá por el valor $\hat{\mu}$ que minimiza

$$S(\mu) = \sum_{i=1}^{n} \rho\left(\frac{x_i - \mu}{\hat{\sigma}_n}\right). \tag{10.28}$$

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

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

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

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

donde

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

y donde A es una constante que hace que el estimador sea consistente a σ 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 $\mathrm{N}(\mu,\sigma^2)$. 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 $\lim(\mu - \tilde{\mu}_n) = 0$ c.s., se tendrá que

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

$$= \sigma \lim_{n\to\infty}\ \text{mediana}\{|u_i|,\ 1 \leq i \leq n\},\ \text{c.s.}. \qquad (10.29)$$

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 (10.29)

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

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

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

o sea

$$\Phi(B) = 0.75,\ B = \Phi^{-1}(0.75) = 0.6745.$$

Luego se tendrá que el estimador MAD de σ viene dado por

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

Cuando el M-estimador se obtiene minimizando (10.28), la ecuación (10.15) se transforma en

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - \mu}{\hat{\sigma}}\right) = 0. \qquad (10.30)$$

Las propiedades asintóticas del estimador $\hat{\mu}$ solución de (10.30) son similares a las del estimador correspondiente al caso de σ conocida. El siguiente Teorema se dará sin demostración.

Teorema 5. Sean $x_1, ...x_n$ variables aleatorias independientes que satisfacen el modelo (10.1). Consideremos un estimador $\hat{\mu}_n$ solución de (10.15), donde ψ es impar y F es simétrica respecto de 0. Supongamos que $\hat{\mu}_n$ es consistente a μ y $\hat{\sigma}_n$ es consistente a σ, y que además ψ tiene dos derivadas continuas y ψ'' es acotada. Luego se tiene que

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

donde V está dada por (10.21)

10.2.5 Algoritmos para calcular M-estimadores

A continuación vamos a describir tres algoritmos para computar el M-estimador definido como la solución de (10.30).

Algoritmo basado en medias ponderadas iteradas (MPI)

Llamemos $w(u) = \psi(u)/u$. Luego la ecuación (10.30).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

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

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

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{10.31}$$

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 μ, 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 tenemos definido $\hat{\mu}_h$, podemos definir $\hat{\mu}_{h+1}$ por

$$\hat{\mu}_{h+1} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \tag{10.32}$$

Se puede mostrar que si ψ es continua, entonces cuando este algoritmo iterativo converge, lo hace a una solución de (10.30). En efecto supongamos que $\lim_{h\to\infty} \hat{\mu}_h = \hat{\mu}$, luego tomando limite en ambos lados de (10.32), se tendrá

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i w\left((x_i - \hat{\mu})/\hat{\sigma}\right)}{\sum_{i=1}^{n} w\left(\ (x_i - \hat{\mu})/\hat{\sigma}\right)}. \tag{10.33}$$

Pero esta ecuación es precisamente (10.31) , que ya hemos visto es equivalente a (10.30).

La ecuación (10.33) 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 $|u|$, (10.33) 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 ψ_k^H, la correspondiente función de peso está dada por

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

El gráfico de esta función se encuentra en la Figura 3.

Algoritmo basado en medias de pseudovalores iteradas (MPVI)

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

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

Luego se tiene

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

y reemplazando en (10.30) se tiene que 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

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}). \tag{10.34}$$

Es decir se tiene expresado el M-estimador como promedio simple de los pseudovalores. 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 $\hat{\mu}_h$

$$\hat{\mu}_{h+1} = \frac{1}{n}\sum_{i=1}^{n} x_i^*(\hat{\mu}_h). \tag{10.35}$$

Es interesante calcular los pseudovalores correspondientes a ψ_k^H, los cuales están dados por

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

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 pseudovalor 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 ψ es continua, entonces $\hat{\mu}$ es el M-estimador solución de (10.30). En efecto tomando límite en ambos miembros de (10.35) se obtiene (10.34), que ya hemos visto es equivalente a (10.30).

Algoritmo de Newton Raphson (NR)

De acuerdo a lo visto anteriormente, el algoritmo de Newton Raphson para calcular la raíz de (10.30) tiene la siguiente fórmula recursiva

$$\hat{\mu}_{h+1} = \hat{\mu}_h + \hat{\sigma}\frac{\sum_{i=1}^{n}\psi\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}{\sum_{i=1}^{n}\psi'\left((x_i - \hat{\mu}_h)/\hat{\sigma}\right)}. \quad (10.36)$$

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

Para cada valor μ dividamos el conjunto de observaciones en tres subconjuntos

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

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

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

Llamando n_-, n_0 y n_-, al número de elementos de D_-, D_0 y D_+ y reemplazando en (10.36), se tiene

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

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 (10.30) cuando ψ es monótona no decreciente cuando esta es única. Si (10.30) 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 $h > h_0$.

10.3 Medidas de robustez

10.3.1 Estimadores como funcionales

Dada una muestra $x_1, ..., x_n$, la distribución empírica se define de la siguiente manera

$$F_n(x) = \frac{\#\{i : x_i \leq x\}}{n},$$

es decir, como la proporción de elementos menores o iguales que x que se observa en la muestra.

Supongamos que la muestra contenga $m \leq n$ elementos distintos $y_1 \leq ... \leq y_m$, y sea n_i el número de observaciones x_j iguales a y_i. Luego es fácil ver que F_n es la distribución que asigna al valor y_i, $1 \leq i \leq m$, probabilidad $p_i = n_i/n$. En efecto, supongamos una variable Y que tiene esta distribución, luego se tiene

$$P(Y \leq y) = \sum_{y_i \leq y} \frac{n_i}{n} = \frac{1}{n} \sum_{y_i \leq y} n_i = \frac{\#\{i : x_i \leq y\}}{n} = F_n(y).$$

Vamos a ver cual es la esperanza de $g(X)$,donde X es una variable aleatoria que tiene función de distribución igual a F_n. Se tendrá

$$E_{F_n}(g(X)) = \sum_{i=1}^{m} g(y_i)\frac{n_i}{n} = \frac{1}{n}\sum_{i=1}^{m} n_i g(y_i) = \frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

En particular

$$E_{F_n}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

Es decir la esperanza de la distribución empírica coincide con la media muestral.

Si definimos mediana principal (medprin) de una distribución como el punto medio del intervalo de medianas también se tiene

$$\text{medprin}(F_n) = \text{mediana}(x_1, ..., x_n).$$

Esto resultado queda como ejercicio.

Supongamos ahora que $x_1, ..., x_n$ es una muestra aleatoria de una distribución F, luego es fácil ver que

$$\lim_{n\to\infty} F_n(x) = F(x) \text{ c.s..} \tag{10.37}$$

En efecto, es inmediato que

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} z_i,$$

donde

$$z_i = \begin{cases} 1 & \text{si} \quad x_i \le x \\ 0 & \text{si} \quad x_i > x \end{cases}.$$

Las variables z_i, $1 \le i \le n$ son i.i.d con $E(z_i) = P(x_i \le x) = F(x)$, luego (10.37) se obtiene de la Ley Fuerte de los Grandes Números.

Se puede probar un resultado aun más fuerte: la convergencia de F_n a F es uniforme en x. Este resultado queda establecido en el Teorema de Glivenko Cantelli que enunciaremos sin demostración

Teorema de Glivenko Cantelli. Sea $x_1, ..., x_n$ una muestra aleatoria de una distribución F, y sea F_n su distribución empírica. Luego se tiene

$$\sup_x |F_n(x) - F(x)| \to 0 \text{ c.s..}$$

Muchos estimadores pueden ser expresados como una funcional de la distribución empírica. Se dan a continuación algunos ejemplos. Por ejemplo se tiene $\bar{x}_n = T(F_n)$, donde$T(F) = E_F(x)$. También mediana$(x_1, ..., x_n) = T(F_n)$, donde $T(F)$ =medprin(F). Finalmente, los M-estimadores de posición se definen como

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

o equivalentemente

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

Luego $\hat{\mu}_n = T(F_n)$ donde $T(F)$ se define implicitamente por

$$E_F\left(\psi\left(\frac{x-T(F)}{\sigma}\right)\right)=0. \tag{10.38}$$

Supongamos que se tiene ahora una muestra aleatoria $x_1,...,x_n$ de una distribución F y supongamos que se tiene un estimador $\hat{\theta}_n=T(F_n)$, donde T está definido para un conjunto de distribuciones $\mathcal{F}$ que incluyen las empíricas y la propia F. Supongamos que el funcional T sea continuo, en el sentido de que si $F_n^*\to F$, entonces $T(F_n^*)\to T(F)$. Luego de acuerdo al Teorema de Glivenko Cantelli tendrá que $\hat{\theta}_n=T(F_n)\to T(F)$ c.s.. Luego $T(F)$ se puede interpretar como el valor límite al cual converge el estimador $\hat{\theta}_n$ cuando la muestra proviene de la distribución F.

Consideremos ahora la situación donde se tiene un modelo paramétrico dado por la familia de distribuciones F_θ, donde $\theta\in\Theta\subset R$. Consideremos una muestra aleatoria $x_1,...,x_n$ de F_θ y sea un estimador $\hat{\theta}_n=T(F_n)$, donde T es un funcional continuo. Luego $T(F_n)\to T(F_\theta)$ c.s., y la siguiente definición está motivada por el hecho de que interesa que $T(F_n)\to\theta$.

Definición Se dirá que un funcional T es *Fisher-consistente* para la familia de distribuciones F_θ si

$$T(F_\theta)=\theta.$$

Veamos como se traduce esta condición para el modelo de posición dado en (10.1). En este caso se deberá tener

$$E_{F_{\mu\sigma}}\left(\psi\left(\frac{x_i-\mu}{\sigma}\right)\right)=E_F\left(\psi\left(u\right)\right)=0.$$

que es la condición usada cuando se estudió consistencia

10.3.2 Función de influencia

Supongamos ahora que se tiene una familia de distribuciones F_θ, y consideremos un estimador dado por un funcional $T(F)$ que es Fisher-consistente. Dada una muestra aleatoria $x_1,...,x_n$ definimos $\hat{\theta}_n=T(F_n)$. Entonces si la distribución de la muestra es F_θ, se tendrá que $\hat{\theta}_n\to T(F_\theta)=\theta$. En cambio si la distribución de los elementos de la muestra es $F \neq F_\theta$, en general se tendrá que $\hat{\theta}_n \to T(F) \neq \theta$. Consideremos un entorno de contaminación de F_θ

$$\mathcal{V}_{\theta,\epsilon} = \{F : F = (1-\epsilon)F_\theta + \epsilon F^*,\ F^* \text{ arbitraria}\}.$$

Luego para el estimador basado en T sea robusto, este funcional debería ser cercano a θ para toda $F \in V_{\theta,\epsilon}$. El sesgo asintótico $S(T,\epsilon,\theta,F^*)$ del funcional T se define de la siguiente manera

$$S(T,\epsilon,\theta,F^*) = T((1-\epsilon)F_\theta + \epsilon F^*) - T(F_\theta) = T((1-\epsilon)F_\theta + \epsilon F^*) - \theta.$$

Como este sesgo puede ser complicado de calcular, vamos a utilizar una aproximación lineal usando el teorema del valor medio que vale para valores pequeños de ϵ. En efecto podemos escribir

$$S(T,\epsilon,\theta,F^*) \cong IC^*(T,\theta,F^*)\epsilon, \tag{10.39}$$

donde

$$IC^*(T,\theta,F^*) = \left.\frac{\partial T((1-\epsilon)F_\theta + \epsilon F^*)}{\partial \epsilon}\right|_{\epsilon=0}. \tag{10.40}$$

Definición. Sea δ_x la distribución que asigna probabilidad 1 al punto x. Luego la *curva de influencia* del funcional T se define por

$$IC_{T,\theta}(x) = IC^*(T,\theta,\delta_x). \tag{10.41}$$

El significado de la curva de influencia es el siguiente: Una proporción pequeña ϵ de outliers en el punto x produce un sesgo asintótico en el funcional T aproximadamente igual a $IC_{T,\theta}(x)\epsilon$.

Se puede mostrar que bajo condiciones muy generales se tiene el siguiente resultado. Para cualquier F^*

$$IC^*(T,\theta,F^*) = \int_{-\infty}^{\infty} IC_{T,\theta}(x) f_\theta(x)dx = E_\theta(IC_{T,\theta}(x)). \tag{10.42}$$

En el caso discreto la integral se reemplaza por la correspondiente sumatoria. En estas notas está fórmula será demostrada para M-estimadores del modelo de posición.

Supongamos que se tiene el siguiente modelo de posición

$$x = \mu + \sigma u, \tag{10.43}$$

donde u es una variable aleatoria con distribución F_0, y donde σ es conocida. Luego la distribución de x es $F_\mu = F_0((x-\mu)/\sigma)$, y el funcional T de un M-estimador se define por

$$E_F\left(\psi\left(\frac{x-T(F)}{\sigma}\right)\right) = 0, \qquad (10.44)$$

y como

$$E_{F_\mu}\left(\psi\left(\frac{x-\mu}{\sigma}\right)\right) = E_{F_0}\psi\left(u\right)$$

la condición de Fisher-Consistencia se puede escribir como

$$E_{F_0}\psi\left(u\right) = 0. \qquad (10.45)$$

El siguiente teorema nos permite calcular la curva de influencia. de un M-estimador para el modelo de posición

Teorema 6. Sea el M-estimador cuyo funcional $T(F)$ está dado por (10.44). Supongamos que ψ es impar, continua y estrictamente creciente. Luego se tiene que

(i) Dado $\delta > 0$, existe ϵ_0, tal que si $\epsilon \le \epsilon_0$, entonces para todo F^* se tiene

$$|T((1-\epsilon)F_\mu + \epsilon F^*) - \mu| \le \delta.$$

(ii)

$$IC^*(T,\mu,F^*) = \sigma\frac{E_{F^*}\left(\psi\left((x-\mu)/\sigma\right)\right)}{E_{F_0}\left(\psi'\left(u\right)\right)}. \qquad (10.46)$$

(iii)

$$IC_{T.\mu}(x) = \sigma\frac{\psi\left((x-\mu)/\sigma\right)}{E_{F_0}\left(\psi'\left(u\right)\right)}. \qquad (10.47)$$

y luego la formula (10.42) vale.

Demostración: (i) Pongamos $G_\epsilon = (1-\epsilon)F_\mu + \varepsilon F^*$. Luego por definición del M-estimador se tiene

$$E_{G_\epsilon}\left(\psi\left(\frac{x-T(G_\epsilon)}{\sigma}\right)\right) = (1-\epsilon)E_{F_\mu}\left(\psi\left(\frac{x-T(G_\epsilon)}{\sigma}\right)\right) + \epsilon E_{F^*}\left(\psi\left(\frac{x-T(G_\epsilon)}{\sigma}\right)\right) = 0$$

y usando (10.43), se tiene

$$(1-\epsilon)E_{F_0}\left(\psi\left(u+\frac{\mu-T(G_\epsilon)}{\sigma}\right)\right) + \epsilon E_{F^*}\left(\psi\left(\frac{x-T(G_\epsilon)}{\sigma}\right)\right) = 0. \qquad (10.48)$$

Luego, si llamamos

$$H(z) = (1-\epsilon)E_{F_0}\left(\psi\left(u + \frac{\mu - z}{\sigma}\right)\right) + \epsilon E_{F^*}\left(\psi\left(\frac{x-z}{\sigma}\right)\right),$$

$T(G_\epsilon)$ esta definido por

$$H(T(G_\epsilon)) = 0. \tag{10.49}$$

Se puede demostrar que H es continua y estrictamente decreciente

Por otro lado por la condición (10.45) se tiene que como ψ es estrictamente creciente

$$E_{F_0}(\psi(u - \frac{\delta}{\sigma}) < E_{F_0}(\psi(u\ ) = 0 < E_{F_0}(\psi(u + \frac{\delta}{\sigma}).$$

Sea $M = \max \psi$ y

$$\epsilon_0 = \frac{\min(-E_{F_0}(\psi(u-(\delta/\sigma))), E_{F_0}(\psi(u+(\delta/\sigma))))}{2M}.$$

Luego se tiene

$$H(\mu{+}\delta) = (1{-}\epsilon)E_{F_0}\left(\psi\left(u - \frac{\delta}{\sigma}\right)\right){+}\epsilon E_{F^*}\left(\psi\left(\frac{x-(\mu+\delta)}{\sigma}\right)\right) > -\frac{2\varepsilon_0}{M}{+}\frac{\varepsilon_0}{M} < 0.$$

Similarmente se demuestra que

$$H(\mu{-}\delta) = (1{-}\epsilon)E\left({}_{F_0}(\psi\left(u - \frac{\delta}{\sigma}\right)\right){+}\epsilon E_{F^*}\left(\psi\left(\frac{x-(\mu+\delta)}{\sigma}\right)\right) > \frac{2\varepsilon_0}{M}{-}\frac{\varepsilon_0}{M} > 0.$$

Luego como H es decreciente por (10.49), se obtiene que $\mu - \delta < T(G_\varepsilon) < \mu + \delta$, y la parte (i) del Teorema queda demostrada

Ahora demostraremos (ii). Derivando (10.48) con respecto a ϵ, se obtiene

$$-E_{F_0}\left((\psi\left(u + \tfrac{\mu - T(G_\epsilon)}{\sigma}\right)\right) - \tfrac{(1-\epsilon)}{\sigma}E_{F_0}\left(\psi'\left(u + \tfrac{\mu - T(G_\epsilon)}{\sigma}\right)\right)\tfrac{\partial T(G_\epsilon)}{\partial \epsilon} +$$

$$E_{F^*}\left(\psi\left(\tfrac{x - T(G_\epsilon)}{\sigma}\right)\right) - \tfrac{\varepsilon}{\sigma}E_{F^*}\left(\psi'\left(\tfrac{x - T(G_\epsilon)}{\sigma}\right)\right) = 0.$$

Tomando $\epsilon = 0$ y usando (10.45) y el hecho que $T(G_0) = \mu$, se obtiene

$$-\frac{1}{\sigma} E_{F_0}(\psi'(u) \left.\frac{\partial T(G_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} + E_{F^*}(\psi(\frac{x-\mu}{\sigma})) = 0,$$

y

$$IC^*(T, \mu, F^*) = \left.\frac{\partial T(G_\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} = \sigma \frac{E_{F^*}(\psi((x-\mu)/\sigma))}{E_{F_0}(\psi'(u))}.$$

Poniendo $F^* = \delta_x$, se obtiene (iii).

Obsérvese que de acuerdo con el Teorema 4 y Teorema 6(iii), para M-estimadores del modelo de posición se tiene

$$n^{1/2}(T(F_n) - \mu) \to_D N(0, E_{F_\mu}(IC^2_{T,\mu}(x)). \tag{10.50}$$

Está formula vale en general para un estimador basado en un funcional Fisher-consistente T del parámetro θ de una familia de distribuciones F_θ. En efecto, bajo condiciones bastante generales se tendrá

$$n^{1/2}(T(F_n) - \mu) \to_D N(0, E_\theta(IC^2_{T,\theta}(x)). \tag{10.51}$$

Vamos a definir ahora una medida de robustez de un estimador basado en un funcional T.

Definición. Se llama *sensibilidad a errores groseros* de un estimador basado en un funcional T para θ, cuando se observa una variable x con distribución en la familia F_θ a

$$\gamma_{T,\theta} = \sup_x |IC_{T,\theta}(x)|.$$

Obsérvese, que como consecuencia de (10.42) también se tendrá también

$$\gamma_{T,\theta} = \sup_{F^*} |IC^*(T, \theta, F^*)|. \tag{10.52}$$

De acuerdo al Teorema 6 (iii), para M estimadores del modelo de posición se tiene que

$$\gamma_{T,\mu} = \frac{\sigma \sup_x |\psi(x)|}{E_{F_0}(\psi'(u))}, \tag{10.53}$$

y por lo tanto, para que $\gamma_{T,\theta} < \infty$ es condición necesaria y suficiente que la función ψ sea acotada.

Hampel propuso un criterio para elegir un M-estimador que tuviera simultaneamente propiedades de robustez y eficiencia bajo el modelo normal. Para eso supongamos $F_0 = \Phi$. Luego para garantizar que el estimador tiene un grado de robustez adecuado, se le exige la restricción que

$$\gamma_{T,\mu} \leq c \tag{10.54}$$

donde c se elige de acuerdo al grado de robustez que se requiera. El estimador óptimo de acuerdo al criterio de Hampel sera aquel que minimice $V(\psi,\Phi)$ dada en el Teorema 4, entre todos los que satisfagan (10.54). Hampel demostró que los M- estimadores óptimos de acuerdo a este criterio son los que tienen función ψ en la familia de Huber ψ_k^H. El siguiente Teorema establece la optimalidad de estos estimadores

Teorema 7. Sea

$$c = \frac{\sigma k}{E_\Phi\left(\psi_k^{H\prime}(u)\right)}. \tag{10.55}$$

Luego si T_k^H es el funcional correspondiente al M-estimador basado en ψ_k^H, se tiene

(i)

$$\gamma_{T_k^H,\mu} = c.$$

(ii)Si T es otro funcional correspondiente a un M-estimador basado en una función ψ satisfaciendo (10.54), entonces

$$V(\psi,\Phi) \geq V(\psi_k^H,\Phi).$$

Demostración: (i) es inmediato a partir de (10.53), (10.55) y de la definición de ψ_k^H

Para demostrar (ii) de acuerdo a (10.53) y (10.21), tenemos que mostrar que si ψ es una función tal que

$$\frac{\sigma \sup \psi(x)}{E_\Phi\left(\psi'(u)\right)} \leq \frac{\sigma k}{E_\Phi\left(\psi_k^{H\prime}(u)\right)}, \tag{10.56}$$

entonces

$$\frac{\sigma^2 E_\Phi\left(\psi^2(u)\right)}{\left(E_\Phi\left(\psi'(u)\right)\right)^2} \geq \frac{\sigma^2 E_\Phi\left(\psi_k^{H2}(u)\right)}{\left(E_\Phi\left(\psi_k^{H\prime}(u)\right)\right)^2}. \tag{10.57}$$

Es fácil ver que si m es una constante, el M-estimador definido por $\psi^* = m\psi$ es el mismo que el definido por ψ. Por lo tanto sin pérdida de generalidad podemos suponer que la función ψ satisface

$$E_\Phi\left(\psi'\left(u\right)\right) = E_\Phi\left(\psi_k^{H\prime}\left(u\right)\right), \tag{10.58}$$

ya que si no la satisface, se la puede multiplicar por una constante adecuada, de tal manera que (10.58) se cumpla. Pero entonces de acuerdo a (10.56) y (10.57) tenemos que demostrar si ψ satisface (10.58) y

$$\sup|\psi(x)| \leq k, \tag{10.59}$$

entonces

$$E_\Phi\left(\psi^2\left(u\right)\right) \geq E_\Phi\left(\psi_k^{H2}\left(u\right)\right).$$

Se tiene que

$$E_\Phi(\psi^2(u)) = E_\Phi(((\psi(u)-u)+u)^2) =$$

$$E_\Phi((\psi(u)-u)^2) + E_\Phi(u^2) + 2E_\Phi((\psi(u)-u)u) =$$

$$E_\Phi((\psi(u)-u)^2) - E_\Phi(u^2) - 2E_\Phi(\psi(u)u). \tag{10.60}$$

Por otro lado usando el hecho de que para densidad N(0,1) φ se verifica que $\varphi'(u) = -u\varphi(u)$, obtenemos

$$E_\Phi(\psi(u)u) = \int_{-\infty}^{\infty} \psi(u)u\varphi(u)du = -\int_{-\infty}^{\infty} \psi(u)\varphi'(u)du.$$

Luego, integrando por partes resulta

$$E_\Phi(\psi(u)u) = -[\psi(u)\varphi(u)]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \psi'(u)\varphi(u)du.$$

Usando el hecho que ψ tiene que ser acotada y de que

$$\lim_{x\to\infty} \varphi(x) = \lim_{x\to-\infty} \varphi(x) = 0,$$

resulta$-[\psi(u)\varphi(u)]_{-\infty}^{\infty} = 0$. Luego usando (10.58) resulta

$$E_\Phi(\psi(u)u) = E_\Phi(\psi'(u)) = E_\Phi(\psi_k^{H\prime}(u)).$$

Reemplazando en (10.60) se obtiene

$$E_\Phi(\psi^2(u)) = E_\Phi((\psi(u) - u)^2) - E_\Phi(u^2) - 2E_\Phi(\psi_k^{H\prime}(u)) \; .$$

Como los dos últimos términos del segundo miembro esta igualdad no dependen de ψ, entonces (10.57) se transforma en

$$E_\Phi((\psi(u) - u)^2) \geq E_\Phi((\psi_k^H(u) - u)^2).$$

Para demostrar esta desigualdad será suficiente mostrar que para toda ψ satisfaciendo (10.59) se tiene

$$|\psi(u) - u| \geq |\psi_k^H(u) - u| \;\forall u. \tag{10.61}$$

Obsérvese que si ψ satisface (10.59) se debe tener necesariamente que

$$|\psi(u) - u| \geq \begin{cases} -u-k & \text{si} \quad u < -k \\ 0 & \text{si} \quad |u| \leq k \\ u-k & \text{si} \quad u > k \end{cases} ,$$

y como es inmediato que

$$|\psi_k^H(u) - u| = \begin{cases} -u-k & \text{si} \quad u < -k \\ 0 & \text{si} \quad |u| \leq k \\ u-k & \text{si} \quad u > k \end{cases} ,$$

entonces (10.61) se satisface. Esto prueba el Teorema.

También se podría hacer una formulación dual del problema. En vez de fijar una cota superior para $\gamma_{T,\mu}$, y minimizar la varianza asintótica $V(\psi, \Phi)$, se podría fijar una cota superior para la varianza asintótica y minimizar $\gamma_{T,\mu}$. Luego tendríamos el siguiente teorema, cuya demostración se deja como ejercicio

Teorema 8. Sea

$$v = \frac{E_\Phi\left(\psi_k^{H2}(u)\right)}{E_\Phi^2\left(\psi_k^{H\prime}(u)\right)} \tag{10.62}$$

Luego si T_k^H es el funcional correspondiente al M-estimador basado en ψ_k^H, se tiene

(i)

$$V(\psi_k^H, \Phi) = v$$

(ii)Si T es otro funcional correspondiente a un M-estimador basado en una función ψ, satisfaciendo

$$V(\psi, \Phi) \leq v$$

entonces

$$\gamma_{T,\mu} \geq \gamma_{T_k^H,\mu}$$

En la práctica se fija v,(generalmente 1.05) y se busca el M-estimador minimizando $V(\psi, \Phi)$. Este estimador corresponde a ψ_k^H con $k = 1.345$.

10.3.3 Punto de ruptura

La sensibilidad a errores groseros es una medida de la robustez de un estimador para una proporción de contaminación ϵ pequeña. En efecto de acuerdo a (10.39) y a (10.52) para valores pequeños de ϵ se tendrá

$$|S(T, \epsilon, \theta, F^*)| \cong |IC^*(T, \theta, F^*)|\epsilon \leq \gamma_{T,\theta}\epsilon.$$

Vamos a definir ahora una medida de la robustez del estimador frente para proporciones de contaminación ϵ grandes. Esta medida se denomina punto de ruptura. Hay dos versiones de esta medida, una asintótica, que mide la robustez de un estimador para muestras grandes, y otra para muestras finitas. Aquí daremos solamente el punto de ruptura para muestras finitas.

Supongamos que tenemos un estimador $\hat{\theta}_n = \delta(x_1, ..., x_n)$ definido para muestras de tamaño n. Tomemos una muestra fija $\mathbf{x} = (x_1, ..., x_n)$, sea $m < n$, y llamemos Z_m al conjunto de todas muestras $\mathbf{z} = (z_1, ..., z_n)$ tales que $\#\{i : x_i = z_i\} \leq m$. La interpretación de Z_m es como el conjunto de todas las muestras que se obtienen reemplazando a lo sumo m observaciones de $\mathbf{x}$ por outliers. Vamos ahora a definir el sesgo máximo causado por m outliers como

$$SM(\hat{\theta}_n, \mathbf{x}, m) = \sup_{\mathbf{z} \in Z_m} |\delta(\mathbf{z}) - \delta(\mathbf{x})|$$

Vamos a definir m^* como el mínimo número de outliers que puede provocar un sesgo infinito, más precisamente

$$m^* = \min\{m : SM(\hat{\theta}, \mathbf{x}, m) = \infty\}$$

Definición El *punto de ruptura finito* de $\hat{\theta}$ en la muestra $\mathbf{x}$ se define como

$$\epsilon^*(\hat{\theta}, \mathbf{x}) = \frac{m^*}{n}$$

Es decir es la mínima proporción de outlier que puede dar un sesgo ∞. Luego si la proporción de contaminaciónminación $\epsilon < \epsilon^*(\hat{\theta}, \mathbf{x})$, el estimador es informativo, dejando de serlo cuando $\epsilon > \epsilon^*(\hat{\theta}, \mathbf{x})$. En general, el máximo punto de ruptura es menor o igual que $1/2$. Cuando la muestra tiene mas de un 50 % de outliers, es imposible saber cuales son las observaciones normales y cuales los outliers

El siguiente Teorema muestra que los M-estimadores de posición basados en una ψ acotada tienen punto de ruptura $1/2$.

Teorema 9.Sea un M -estimador $\hat{\mu}_n = \delta_n(x_1, ..., x_n)$ basado en una función ψ definido por (10.15). Supongamos que ψ es impar, continua, monótona no decreciente y acotada. Luego dada cualquier muestra $\mathbf{x} = (x_1, ..., x_n)$ y $m < n/2$ se tiene que $SM(\hat{\mu}_n, \mathbf{x}, m) < \infty$. Luego $\epsilon^*(\hat{\theta}, \mathbf{x}) \geq \mathbf{0.5}$.

Demostración: Comenzaremos mostrando que existe M tal que para toda muestra $\mathbf{z} \in Z_m$ se tiene $|\delta_n(\mathbf{z})| < M$. Sea $k = \sup |\psi(u)|$, como $n - 2m > 0$, podemos elegir $\delta > 0$ tal que

$$\delta < \frac{k(n-2m)}{n-m}, \tag{10.63}$$

y elijamos $x_0 > 0$ tal que

$$\psi(x_0) = k - \delta. \tag{10.64}$$

Sea ahora

$$m = \max_{1 \leq i \leq n} |x_i|.$$

Veremos que podemos tomar $M = m + \sigma x_0$. En efecto tomemos $\mathbf{z} = (z_1, ..., z_n) \in \mathbf{Z}_m$, vamos a mostrar que no puede existir μ tal que $|\mu| > M$ y tal que

$$\sum_{i=1}^{n} \psi\left(\frac{z_i - \mu}{\sigma}\right) = 0. \tag{10.65}$$

Mostraremos primero que no es posible que $\mu > M$. En efecto supongamos que $\mu > M$ y que satisface (10.65) Luego como ψ es no decreciente se debería tener

$$0=\sum_{i=1}^{n}\psi\left(\frac{z_i-\mu}{\sigma}\right)\leq\sum_{i=1}^{n}\psi\left(\frac{z_i-M}{\sigma}\right)=\sum_{i=1}^{n}\psi\left(\frac{z_i-m-\sigma x_0}{\sigma}\right)=\sum_{i=1}^{n}\psi\left(\frac{z_i-m}{\sigma}-x_0\right) \tag{10.66}$$

Sea $D=\{i:z_i=x_i\}$. Luego $\#D=n-m$ y $\#D'=m$. Luego podemos escribir

$$\sum_{i\in D}\psi\left(\frac{x_i-m}{\sigma}-x_0\right)+\sum_{i\in D'}\psi\left(\frac{z_i-m}{\sigma}-x_0\right)\geq 0.$$

Como $(x_i-m)/\sigma\leq 0$ para todo $i\in D$, usando nuevamente el hecho de que ψ es no decreciente, tenemos

$$\sum_{i\in D}\psi\left(-x_0\right)+\sum_{i\in D'}\psi\left(\frac{z_i-m}{\sigma}-x_0\right)\geq 0. \tag{10.67}$$

Como por (10.64) $\psi\left(-x_0\right)=-\psi\left(x_0\right)=-(k-\delta)$, y $\psi(v)\leq k$ for all v, reemplazando en (10.67) y usando nuevamente que ψ es no decreciente, obtenemos

$$-(n-m)(k-\delta)+mk=-(n-m)k+\delta(n-m)+mk=-(n-2m)k+\delta(n-m)\geq 0.$$

Luego despejando se obtiene

$$\delta\geq\frac{k(n-2m)}{n-m},$$

contrariamente a lo supuesto en (10.63). Luego no puede ser $\mu>M$. En forma similar se prueba que no puede ser $\mu<-M$, y por lo tanto $|\mu|\leq M$.

Consideremos ahora $\mathbf{z}\in Z_n$

$$|\delta_n(\mathbf{z})-\delta_n(\mathbf{x})|\leq|\delta_n(\mathbf{z})|+|\delta_n(\mathbf{x})|\leq M+|\delta_n(\mathbf{x})|,$$

y luego

$$S(\hat{\mu},\mathbf{x},m)=\sup_{z\in Z_n}|\delta_n(\mathbf{z})-\delta_n(\mathbf{x})|\leq M+|\delta_n(\mathbf{x})|<\infty,$$

con lo que se prueba el Teorema.