DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico II

Probabilidad y Estadística
Primer Cuatrimestre de 2019

Integrante	LU	Correo electrónico
Facundo Linlaud	561/16	facundolinlaud@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

1. Estimaciones del parámetro b

1.1. Método de momentos

Planteemos el momento muestral de primer orden:

$$E(X) = \frac{\sum_{i=1}^{n} X_i}{n}$$

Como $X_1, \ldots, X_n$ es una muestra aleatoria de una distribución $U[0, b]$, podemos igualar el primer momento muestral a la esperanza de la distribución con parámetros 0 y b:

$$\begin{aligned} \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X} &= E(X) \\ &= \frac{b}{2} \end{aligned}$$

De esta manera, obtenemos la expresión del estimador del parámetro b:

$$\hat{b}_{mom} = 2\bar{X}$$

1.2. Estimador de Máxima Verosimilitud

Sabemos que la función de densidad de una distribución uniforme es:

$$f_X(x) = \frac{1}{b-a} I_{(a,b)}(x)$$

Sabiendo que $a = 0$ y que $a < b$ por propiedades de distribución uniforme, procederemos a maximizar $f_X(x)$ para estimar el valor de máxima verosimilitud de b:

$$\begin{aligned} L(b) &= \prod_{i=1}^{n} \frac{1}{b-a} I_{(a,b)}(x) \\ &= \prod_{i=1}^{n} \frac{1}{b-0} I_{(a,b)}(x) \\ &= \prod_{i=1}^{n} \frac{1}{b} I_{(x_i,+\infty)}(b) \end{aligned}$$

Por lo tanto:

$$L(b) = \begin{cases} \frac{1}{b^n} & \text{si } max\{x_1, \ldots, x_n\} < b \\ 0 & \text{sino} \end{cases}$$

Esto implica que $\frac{1}{b^n}$ es máximo cuando b^n es mínimo. Luego, el valor más chico posible de b es el valor más grande que haya tomado algún resultado en la muestra $X_1, \ldots, X_n$. Porque si b fuese menor a alguno de ellos, la probabilidad total sería nula (por contener un elemento fuera del rango de la distribución). Finalmente, tenemos un valor para nuestro estimador:

$$\hat{b}_{mv} = max\{x_1, \ldots, x_n\}$$

1.3. Estimador de la doble mediana

Dada una muestra $X_1, ..., X_n$, estimar el doble de su mediana implica tomar los únicos datos que tenemos, es decir la muestra, calcular su mediana muestral y multiplicarla por dos. Su implementación puede ser observada en el archivo **tp2.py** y el estimador queda definido de la siguiente manera:

$$\hat{b}_{med} = 2\tilde{X}$$

2. Valores y errores de los estimadores para muestra de tamaño 15

Con una muestra aleatoria de tamaño 15 de una distrubición uniforme con parámetros $[0, 2]$, calculamos los estimadores correspondientes y sus errores. Este último paso es realizado restando el estimador por el valor real del parámetro que el estimador intenta justamente aproximar. Los resultados son:

Estimador	Valor	Error
$\hat{b}_{mv}$	0.838	$0{,}161$
$\hat{b}_{mom}$	0.920	$0{,}079$
$\hat{b}_{med}$	1.135	$-0{,}135$

3. Simulación de mil muestras de tamaño 15

A continuación se exhiben los valores promedio de los estimadores con sus sesgos, sus varianzas y sus errores cuadráticos medios para una simulación de mil muestras de tamaño 15:

Estimador	**Promedio**	**Sesgo**	**Varianza**	**ECM**
$\hat{b}_{mv}$	$0{,}938$	$-0{,}064$	$0{,}003$	$0{,}007$
$\hat{b}_{mom}$	$0{,}998$	$-0{,}001$	$0{,}022$	$0{,}022$
$\hat{b}_{med}$	$0{,}997$	$-0{,}002$	$0{,}058$	$0{,}058$

Como cada estimador es una variable aleatoria, para facilitar la notación definiremos $X = \hat{b}$ para cualquier estimador de b. Luego, el promedio de una muestra de tamaño n de cualquier estimador de b es $\bar{X}_n$, que a su vez equivale a:

$$\bar{X}_n = \sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{n} \frac{\hat{b}_i}{n}$$

El sesgo del estimador fue calculado utilizando:

$$Sesgo(\bar{X}) = \bar{X} - 2 \text{ (siendo 2 el valor real del parámetro } b)$$

Mientras tanto, para calcular la variabilidad de cada estimador, se utilizó la expresión de la varianza muestral (que es insesgada) sobre las muestras de mil estimadores previamente calculados para cada método:

$$V(X_n) = \sum_{i=1}^{n} \frac{(x_i - \bar{X})^2}{n-1} \implies$$
$$V(X_{1000}) = \sum_{i=1}^{1000} \frac{(x_i - \bar{X})^2}{999}$$

4. Análisis de las propiedades de los estimadores

A continuación, procederemos a graficar los sesgos, varianzas y errores cuadráticos medios de los estimadores en función de distintos valores de **b** en el rango $(0, 2)$.

4.1. Sesgo

Analizaremos el sesgo de los estimadores utilizando la siguiente definición:

$$Sesgo(\hat{b}) = \hat{b} - b$$

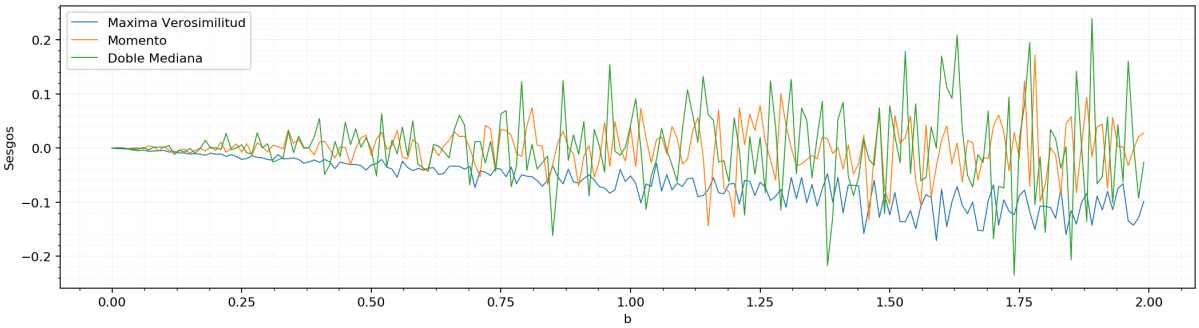

Figura 1: Sesgos de los estimadores en función de los valores de **b**. $a = 0, n = 15$

En la figura 1 se puede observar que, a medida que el valor de b aumenta, el valor absoluto del sesgo del estimador de b también incrementa. Esto tiene sentido porque, por ejemplo, un error del 5 % en un estimador de $b = 1$ sería $0,95$ o $1,05$, mientras que un error del 5 % en un estimador de $b = 100$ sería 95 o 105. Por eso, si bien la calidad del estimador no varía en función de b, sí lo hace la magnitud de sus sesgos. Grafiquemos la diferencia porcentual entre $\hat{b}$ y b en función de b:

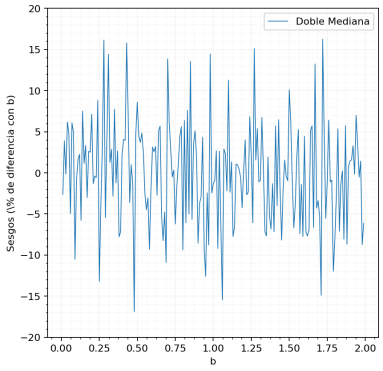

Figura 2: Dif. % entre b y $\hat{b}_{med}$

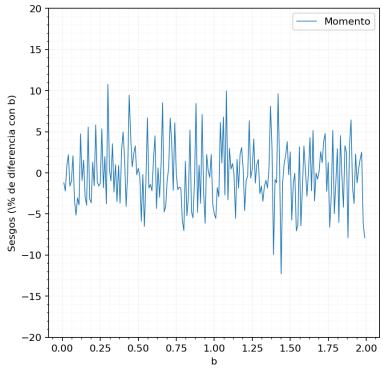

Figura 3: Dif. % entre b y $\hat{b}_{mom}$

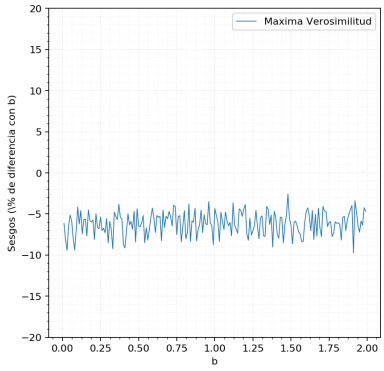

Figura 4: Dif. % entre b y $\hat{b}_{mv}$

Podemos observar ahora que el porcentaje de diferencia entre b y cada $\hat{b}$ tiene un comportamiento con cierta aleatoriedad pero siempre dentro de una cota, que tenderá a un módulo 0 si el estimador respectivo es asintóticamente insesgado. Definimos la diferencia porcentual así: $dif(\hat{b}) = Sesgo(\hat{b}) \cdot \frac{100}{b}$.

También podemos ver que para el estimador $\hat{b}_{med}$, su diferencia porcentual con b está entre el $-15\,\%$ y el $15\,\%$. La diferencia porcentual para el estimador $\hat{b}_{mom}$ está entre el $-10\,\%$ y el $10\,\%$ mientras que la diferencia porcentual para el estimador $\hat{b}_{mv}$ está entre el $-10\,\%$ y el $-5\,\%$. Esta cota está contenida en los números negativos debido a la naturaleza del estimador de máxima verosimilitud, que estima el valor de b como el elemento de valor máximo dentro de la muestra analizada. Como la muestra en cuestión tiene distribución uniforme de parámetros $[0, b]$, un experimento uniforme jamás producirá un resultado mayor a b, por lo tanto el valor máximo de $\hat{b}_{mv}$ es b. Luego, el $sesgo(\hat{b}_{mv}) \leq 0$.

4.1.1. Estimador de Máxima Verosimilitud

Veamos si el estimador $\hat{b}_{mv}$ es insesgado. Para que lo sea, debemos ver que $E(\hat{b}_{mv}) = b$.

$$E(\hat{b}_{mv}) = b \iff E(max\{x_1, \ldots, x_n\}) = b$$

Para esto, debemos encontrar la función de densidad de la variable aleatoria $max\{x_1, \ldots, x_n\}$:

$$Y = max\{x_1, \ldots, x_n\} \implies P(Y \leq y) = \Big(P(X \leq y)\Big)^n = F_Y(y)$$

Sabiendo que X $U[0, b]$:

$$F_Y(y) = \begin{cases} 0 & \text{si } ny \leq 0 \\ \left(\frac{y}{b}\right)^n & \text{si } 0 < y < b \\ 1 & \text{si } y \geq b \end{cases}$$

Luego, podemos derivar la función acumulada y obtener la función de densidad de $\hat{b}_{mv}$:

$$f_Y(y) = \frac{n}{b} \cdot \left(\frac{y}{b}\right)^{n-1} I_{(0,b)}(y)$$

Ahora podemos calcular la esperanza del estimador:

$$\begin{aligned} E(\hat{b}_{mv}) = E(Y) &= \int_0^b y \cdot \frac{n}{b} \cdot \left(\frac{y}{b}\right)^{n-1} \\ &= \left(\frac{n}{n+1}\right) \cdot b \neq b \end{aligned}$$

Por lo tanto, podemos concluir que el estimador de máxima verosimilitud **no es insesgado**. Sin embargo, el estimador **sí es asintóticamente insesgado** porque su límite tiende al parámetro b cuando $n \to \infty$.

4.1.2. Estimador de Momentos

Averiguemos si el estimador de momentos es insesgado o asintóticamente insesgado:

$$\begin{aligned} E(\hat{b}_{mom}) &= E(2\bar{X}) \\ &= 2 \cdot E\Big(\frac{\sum_{i=1}^n X_i}{n}\Big) \\ &= 2 \cdot \left(\frac{1}{n}\right) \cdot n \cdot E(x_i) \text{ (por indep.)} \\ &= 2 \cdot b \neq b \end{aligned}$$

Por lo tanto, el estimador de momentos no es insesgado. Tampoco es asintóticamente insesgado porque:

$$\lim_{n \to \infty} 2 \cdot b = 2 \cdot b \neq b$$

4.1.3. Estimador de la Doble Mediana

Averigüemos si el estimador de la doble mediana es insesgado o asintóticamente insesgado:

$$\begin{aligned} E(\hat{b}_{med}) &= E(2\tilde{X}) \\ &= 2 \cdot E(\tilde{X}) \\ &= 2 \cdot \left(\frac{b-a}{2}\right) \text{ (por indep. y L.G.N.)} \\ &= 2 \cdot \left(\frac{b}{2}\right) \\ &= b \end{aligned}$$

Por lo tanto, el estimador de la doble mediana es insesgado. Esto implica que también es asintóticamente insesgado.

4.2. Varianza

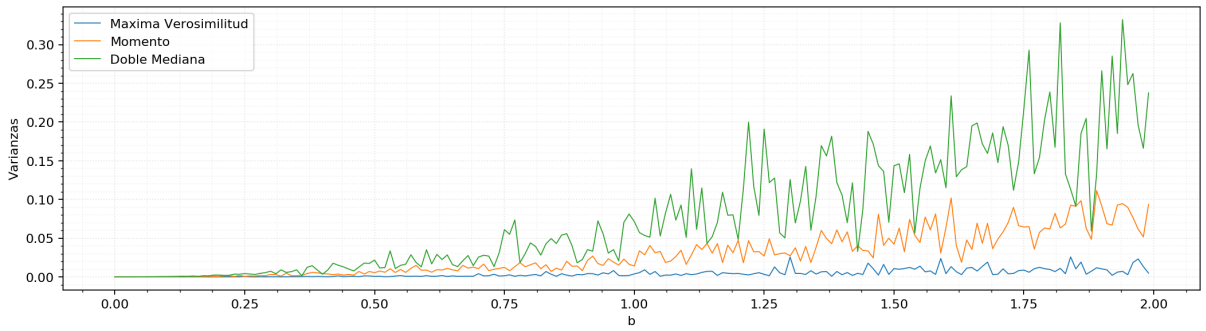

Figura 5: Varianzas de los estimadores en función de los valores de **b**. $a = 0, n = 15$

La figura 5 nos muestra el comportamiento de las varianzas. Podemos observar que la varianza del estimador de la doble mediana $\hat{b}_{med}$ tiene el mayor crecimiento al variar b, seguida por la varianza del estimador $\hat{b}_{mom}$ y en último lugar, la varianza del estimador $\hat{b}_{mv}$.

4.3. Error Cuadrático Medio

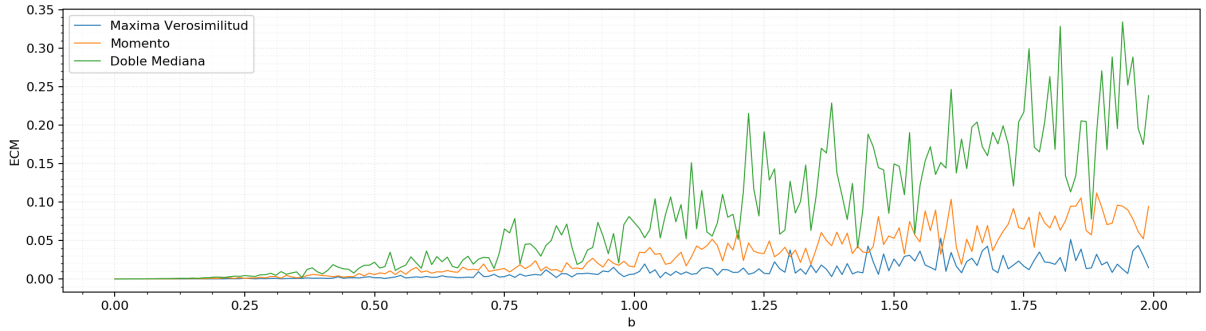

Figura 6: ECM de los estimadores en función de los valores de **b**. $a = 0, n = 15$

La figura 6 exhibe la tendencia del **Error Cuadrático Medio** a aumentar cuando el valor de b lo hace también. Esta figura nos muestra que el estimador $\hat{b}_{mv}$ provee la mejor calidad, dada una muestra de suficientes elementos, debido a su bajo error cuadrático medio y por ser asintóticamente insesgado. La definición del error cuadrático medio es:

$$ECM(\hat{\theta}) = V(\hat{\theta}) + Sesgo(\hat{\theta})^2$$

Sabremos de antemano que el **E.C.M.** del estimador de momentos no será cero porque sus sesgos no son nulos. Sin embargo, el **E.C.M.** de los estimadores de la doble mediana y de máxima verosimilitud pueden llegar a serlo si $\lim_{n\to\infty} V(\hat{b}) = 0$.

A continuación presentamos los errores cuadráticos medios para cada estimador en función de n, con n siendo la cantidad de muestras de tamaño 15:

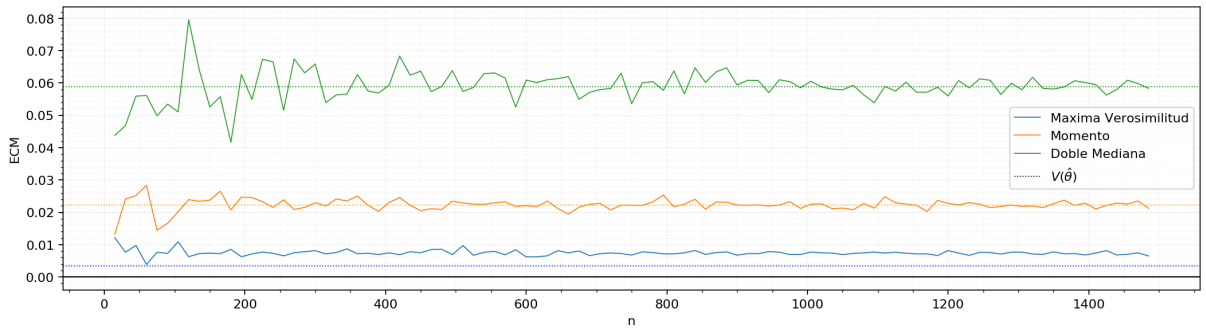

Figura 7: ECM de los estimadores en función de n. $a = 0, b = 1$

La figura 7 sugiere que el error cuadrático medio de los estimadores $\hat{b}_{mom}$ y $\hat{b}_{med}$ no tienden a cero. Sin embargo, el estimador de máxima verosimilitud muestra una significativa evidencia de tendencia al valor nulo. Por esta razón, tenemos razones para creer que su Error Cuadrático Medio tiende a cero a medida que se aumenta el valor de n. Veamos qué implicancias tiene:

$$\begin{aligned}\lim_{n\to\infty} ECM(\hat{b}_{mv}) &= \lim_{n\to\infty} V(\hat{b}_{mv}) + Sesgo(\hat{b}_{mv})^2 \\ &= \lim_{n\to\infty} V(\hat{b}_{mv}) \text{ (porque asint. insesgado)} \\ &= 0\end{aligned}$$

Por lo tanto, se tiene que:

$$\lim_{n\to\infty} V(\hat{b}_{mv}) = 0$$

Al tener varianza cero y ser asintóticamente insesgado, el estimador cumple todas las hipótesis de consistencia. Además, por presentar la menor varianza de los estimadores insesgados, el principio de estimación insesgada de mínima varianza nos asegura que $\hat{b}_{mv}$ es el estimador que provee la mejor calidad de los tres.

5. Análisis de una muestra específica

Se nos pide calcular los Estimadores de la siguiente muestra

$0,917, 0,247, 0,384, 0,530, 0,798, 0,912, 0,096, 0,684, 0,394, 20,1, 0,769, 0,137, 0,352, 0,332, 0,670$

Y estos son los estimadores de la muestra, asumiendo que proviene de una distribución uniforme de parámetros $a = 0$ y b:

Estimador	**Valor**
$\hat{b}_{mv}$	$20,1$
$\hat{b}_{mom}$	3,64293
$\hat{b}_{med}$	1,06

Podemos observar que en la muestra, todos los elementos están entre 0 y 1, con la excepción de uno que vale $20,1$. Debido a que $\hat{b}_{mv} = max\{x_1, \ldots, x_n\}$, es decir, el estimador de máxima verosimilitud toma la forma del elemento más grande en la muestra, este estimador brindará el valor $20,1$. Esto significa que este estimador es **muy susceptible** a valores atípicos en la muestra. Por otro lado, $\hat{b}_{mom}$ toma en cuenta todos los valores de la muestra para determinar su valor, brindando cierta resistencia ante un valor grande y aislado como $20,1$. Pero este estimador no es insesgado. Finalmente, nos queda el estimador de la doble mediana, que es insesgado y – como el estimador de momentos – ofrece resistencia frente a la aparición de valores atípicos, que de existir, estarán en los extremos del esquema de tallo y hoja y sabemos que la mediana define sus valores a partir de los elementos del medio.

El estimador más confiable para este caso es el de la **doble mediana**.

6. Muestra contaminada

Nos piden analizar aproximar sesgo, varianza y error cuadrático medio para los estimadores de una muestra aleatoria con $b = 1$ y $n = 15$ donde, de manera independiente, un elemento tiene probabilidad $0,005$ de ser *contaminado* (multiplicado por 100).

Esto nos dice que tenemos, para cada valor x_i, una variable aleatoria b_i de distribución Bernoulli con parámetro $p = 0,005$. Como tenemos n valores en nuestra muestra, tenemos n Bernoullis. Luego, definimos la cantidad de valores contaminados en X_n como $Y = \sum_{i=1}^{n} b_i \sim Bi(n, p = 0,005)$ porque es una sumatoria de variables aleatorias de Bernoulli, y sabemos su función de densidad:

$$p_Y(k) = \binom{n}{k} p^k (1-p)^{n-p}$$

De esta manera, obtenemos la probabilidad de una muestra $X_1, \ldots, X_{15}$ esté contaminada:

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y < 1) \\ &= 1 - P(Y = 0) \\ &= 0{,}07243103119 \end{aligned}$$

Para aproximar el sesgo, la varianza y el error cuadrático medio de cada estimador, procederemos (como en el ejercicio 4) a hacer 1000 repeticiones en las que calculamos los estimadores (momentos, doble mediana y máxima verosimilitud). A continuación mostramos los resultados obtenidos:

Estimador	**Promedio**	**Sesgo**	**Varianza**	**ECM**
$\hat{b}_{mv}$	4,309	3,309	206,119	217,069
$\hat{b}_{mom}$	1,462	0,462	3,938	4,152
$\hat{b}_{med}$	1,005	0,005	0,060	0,060

Como ya discutimos en el punto anterior, el estimador de momentos y el estimador de la doble mediana diluirán la gravedad ejercida por los elementos contaminados (que en probabilidad son pocos) con los valores no contaminados (que serán la gran mayoría). Como con $\hat{b}_{med}$, si hay un elemento contaminado en la muestra, será altamente probable que se produzca una estimación de b menor a alguno de los elementos contaminados y por ende resulte incompatible la estimación del parámetro con la muestra. Si esto no es un problema, y al contrario, se desea estimar b con la menor influencia posible por parte de valores atípicos, el estimador de la doble mediana es la mejor opción al ser insesgado.

Por otro lado, el estimador de máxima verosimilitud se adaptará a cualquier aparición de elementos contaminados. Es decir, el parámetro estimado será siempre el más grande en la muestra, por lo tanto cualquiera estimación de b será compatible con los elementos de la muestra. Es decir, el estimador de máxima verosimilitud es altamente sensible a la aparición de elementos atípicos. Si este tipo de comportamiento es el buscado, este estimador es el indicado.

6.1. Conclusiones

Para este caso, con qué estimador quedarse depende de varias cosas. ¿Una muestra contaminada sigue siendo importante o debería ser deshechada? Si la respuesta es que es importante: ¿Qué tanto queremos arriesgarnos a producir un estimador incompatible con la muestra? y ¿Cuánto nos preocupa que nuestra estimación sea consistente? En nuestra opinión, un estimador debe representar con precisión el parámetro de una distribución de donde se extrae una muestra, por lo tanto concluimos que el mejor estimador ante la posibilidad de valores atípicos es el **estimador de la doble mediana**.