ESTIMADORES ROBUSTOS DE TENDENCIA CENTRAL

Lic. Valdez Blanco, Dindo

✉ dindovaldez@hotmail.com

RESUMEN

El presente artículo tiene por objeto, dar a conocer algunos estimadores robustos de tendencia central, y mostrar su aplicabilidad. En vista que el estimador de tendencia central más utilizado es la media muestral $\bar{x}$, es necesario considerar la presencia de datos atípicos en la muestra, ya que estos pueden distorsionar la estimación que se realiza con la media aritmética

PALABRAS CLAVE

Estimación robusta, mediana de Hodges-Lehmann, media de Takashi, trimedia de Tukey, Media de Huber.

1. INTRODUCCIÓN

Los estimadores robustos denominados también estimadores no paramétricos tienen la ventaja de disminuir la influencia de los valores extremos o de alguna manera ponderar los datos de tal forma que el estimador de posición central sea lo más representativo posible y el margen de error se minimice.

Los Estimadores Robustos, son estimadores libres de la suposición de la forma de distribución de la población de la cual se extrae la muestra. Al contrario de los estimadores clásicos que tienen asociada un tipo de distribución de la población. Así, por ejemplo, a la media aritmética se le asocia la distribución normal o mesocurtica.

Es más, a los Estimadores Clásicos se les asocia un criterio de óptimo prefijado, expresado por medio de las llamadas Normas Mínimas, basado en la distancia de las observaciones respecto el estimador. Las principales Normas Mínimas son las siguientes:

Norma L_1 : "la suma absoluta de los residuales es mínima."

$$Minimo\ L_1 = \sum_{i=1}^{n} |x_i - M|$$

La Norma L_1 está asociada a la Mediana y es conocida como la Norma de Laplace.

Norma L_2: " La suma de los cuadrados de los residuales es mínima."

$$Mínimo\ L_2 = \sum_{i=1}^{n} (x_i - M)$$

La norma L_2 está asociada a la media aritmética y es conocida como el principio de mínimos cuadrados

En contraposición los estimadores robustos no tiene asociados ninguna distribución y ninguna norma óptima. Los principales objetivos de usar los estimadores Robustos se pueden resumir en los siguientes puntos:

a. Construir una estimación segura ante una cantidad apreciable de datos atípicos.

b. Poner un límite a la influencia del sesgo escondido debido a la presencia de datos atípicos (los que se salen de una tolerancia).

c. Aislar de manera clara los datos atípicos para un tratamiento por separado.

d. Seguir cercanamente el sentido estricto del modelo Paramétrico.

2. ESTIMACION NO PARAMÉTRICA O ROBUSTA DE TENDENCIA CENTRAL

Los estimadores no paramétricos de tendencia central son los llamados estimadores de orden o estadísticos de orden, puesto que las observaciones o valores de la variable aleatoria X deben ser ordenados: $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$. Tal que debe cumplirse que: $x_{(1)} < x_{(2)} <... < x_{(n)}$.

De los diversos estimadores no paramétricos o robustos existentes, sólo se indicarán algunos de ellos.

2.1 LA MEDIANA DE HODGES – LEHMANN

Este estimador fue desarrollado por Joseph L. Hodges y Erich L. Lehmann en 1960, el mismo se basa en un algoritmo muy sencillo, es la mediana de los promedios de todos los pares sucesivos de observaciones de una muestra de n observaciones ordenadas.

Sea la serie de datos ordenados: $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$. En base a estos se definen los promedios sucesivos denominados: Y_1 , Y_2,..., Y_{n-1} tal que:

$$Y_1 = \frac{x_{(1)} + x_{(2)}}{2} \; ; \; Y_2 = \frac{x_{(2)} + x_{(3)}}{2} \; ; \; \dots \; ;$$

$$Y_{n-1} = \frac{x_{(n-1)} + x_{(n)}}{2}$$

De tal forma que se obtiene una nueva serie ordenada: $Y_1 < Y_2 <... < Y_{n-1}$. Siendo la mediana de Hodges – Lehmann la mediana de esta nueva serie.

Ejemplo: Suponer que 5 turistas se registran en un hotel, sus edades son: 18, 17, 18, 19 y 60 años. Para calcular el estimador de Hodges – Lehmann, se tiene la muestra ordenada:

$$x_{(1)} = 17 \quad x_{(2)} = 18 \quad x_{(3)} = 18$$

$$x_{(4)} = 19 \quad x_{(5)} = 60$$

En base a los que se calcula la serie de promedios:

$$Y_1 = 17{,}5 \quad Y_2 = 18$$

$$Y_3 = 18{,}5 \quad Y_4 = 39{,}5$$

Hallamos la mediana de los cuatro promedios y resulta:

$$M_{H-L} = \frac{18 + 18{,}5}{2} = 18{,}25$$

2.2 LA MEDIA DE TAKASHI

El Estimador de Takashi, presentado en 1969 por Takashi Yamagawa[1] toma la mediana sucesiva de las observaciones o mediciones y luego a esa nueva serie originada le aplica la media aritmética.

Sea la serie de datos ordenados: $x_{(1)}$, $x_{(2)}$,..., $x_{(n)}$. En base a estos se definen las medianas sucesivas denominadas: Y_1 , Y_2,..., Y_{n-1} tal que:

$$Y_1 = \frac{x_{(1)} + x_{(2)}}{2} \; ;$$

$$Y_2 = \frac{x_{(2)} + x_{(3)}}{2} \; ; \; \dots ;$$

$$Y_{n-1} = \frac{x_{(n-1)} + x_{(n)}}{2}$$

De tal forma que se obtiene una nueva serie ordenada: $Y_1 < Y_2 <... < Y_{n-1}$. Siendo la media de Takashi la media aritmética de esta nueva serie.

Ejemplo: El estimador de Takashi para los

1 Huber (1964) "Robust estimation of a location parameter"

datos de las edades de los turistas del ejemplo anterior se basa en la muestra ordenada de las edades:

$$x_{(1)} = 17 \quad x_{(2)} = 18 \quad x_{(3)} = 18$$

$$x_{(4)} = 19 \quad x_{(5)} = 60$$

En base a los que se calcula la serie de medianas:

$$Y_1 = 17{,}5 \quad Y_2 = 18$$

$$Y_3 = 18{,}5 \quad Y_4 = 39{,}5$$

Finalmente se calcula la media aritmética de las medianas:

$$M_T = \frac{17{,}5 + 18 + 18{,}5 + 39{,}5}{4} = 23{,}37$$

2.3 LA TRIMEDIA DE TUKEY

Este estimador fue desarrollado por John Tukey en 1960[2] y es un promedio ponderado del primer, segundo y tercer cuartil. Sean los cuartiles de una muesra aleatoria de X, entonces el estimador de Tukey se define como:

$$T = \frac{1}{4}Q_1 + \frac{1}{2}Q_2 + \frac{1}{4}Q_3$$

Ejemplo: Los cuartiles de los datos: 18, 17, 18, 19 y 60 son:

$$Q_1 = 17.5 \quad Q_2 = 18 \quad Q_3 = 39{,}5$$

Por tanto la trimedia de Tukey es:

$$T = \frac{1}{4}17{,}5 + \frac{1}{2}18 + \frac{1}{4}39{,}5 = 23{,}25$$

2.4 LA MEDIA ITERATIVA DE HUBER

Este estimador fue desarrollado por Peter J. Huber[3] en el año 1964. El estimador Huber se desarrolla en base a las funciones:

$$Min \sum_{i=1}^{n} (x_i - M)^2$$

Si se cumple:

$$|x_i - M| \leq K\sigma \;\; ; \;\; i = 1,2,\dots,n$$

$$Min \sum_{i=1}^{n} K\sigma(2|x_i - M| - K\sigma)$$

Si cumple $|x_i - M| \geq K\sigma \;\; ; \;\; i = 1,2,\dots,n$

Generalmente *K* adopta valores de 2 ó 3. Pues $K\sigma$ representa la tolerancia de la medición. El estimador de Huber utiliza una función de peso *P* de la siguiente forma:

$$P_i = 1 \;\; si \;\; |x_i - M| \leq K\sigma$$

$$P_i = \frac{K\sigma}{|x_i - M|} \;\; si \;\; |x_i - M| \geq K\sigma$$

De esta manera se otorga una ponderación más baja a las observaciones que se encuentran con mayor desviación de la deseada, lo que influirá directamente en la estimación final.

Ejemplo: Consideremos las observaciones x_i: 17, 18, 18, 19, 60. A simple vista la observación 60 aparenta ser un dato atípico. Tomando como desviación máxima σ = 10 y usando K = 2 se tiene que Kσ = 20 se procede a utilizar el proceso iterativo para la estimación.

Recordemos que la media aritmética ponderada está dada por:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i}$$

Utilizaremos la media aritmética de los datos como M=26,4 para la asignación de las ponderaciones, los residuales de las muestras son:

2 Huber (1964) Robust estimation of a location parameter.

3 Huber (1981) Robust Statistic.

$$|x_1 - M| = 6{,}44$$

$$|x_2 - M| = 5{,}44$$

$$|x_3 - M| = 5{,}44$$

$$|x_4 - M| = 4{,}44$$

$$|x_5 - M| = 36{,}56$$

De tal forma las ponderaciones de las observaciones resultan ser:

$$P_1 = 1 \quad P_2 = 1 \quad P_3 = 1$$

$$P_4 = 1 \quad P_5 = \frac{20}{33.6} = 0{,}5952$$

Así la media aritmética ponderada resulta:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i} = 23{,}44$$

Con esta media ponderada calculamos los nuevos residuales:

$$|x_1 - M| = 6{,}44$$

$$|x_2 - M| = 5{,}44$$

$$|x_3 - M| = 5{,}44$$

$$|x_4 - M| = 4{,}44$$

Y se calculan los nuevos pesos tomando en cuenta que si $|x_i\text{-}M| \leq 20$ tendrá un peso igual a la unidad y si es ≥ 20 se le calcula el peso tal como se indicó antes:

$$P_1 = 1 \quad P_2 = 1 \quad P_3 = 1$$

$$P_4 = 1 \quad P_5 = \frac{20}{36{,}56} = 0{,}547$$

La nueva media ponderada resulta:

$$\bar{X}_P = \frac{\sum_{i=1}^{n} P_i x_i}{\sum_{i=1}^{n} P_i} = 23{,}05$$

Este proceso iterativo se repite hasta que la media ponderada converja a un número el cuál será la media estimada de Huber.

Índice

3. CONCLUSIONES

Los estimadores robustos de tendencia central ofrecen la ventaja de que evitan el "uso y abuso" que se ha hecho de la media aritmética, puesto que "a todo" le aplicamos la estimación de la media aritmética. Por otra parte, dichos estimadores eliminan el uso arbitrario del rechazo de observaciones atípicas, en vista que minimizan su efecto en el cálculo del estimador de posición. Por último, asumir una distribución normal para la población de donde se extrae la muestra (lo cual implica usar la media aritmética como estimador) no es conveniente cuando el número de observaciones es muy pequeño.

BIBLIOGRAFÍA

Huber, P.J. 1964: *Robust estimation of a location parameter*. Annals of Mathematical Statistics. Vol.35; pág. 73-101.

Huber, P.J. 1981: *Robust Statistics*. New York. Wiley & Sons.