UNA MEDIDA DE DESIGUALDAD PARA DISTRIBUCIONES MULTIVARIANTES

M. Luisa Moltó Carbonell
Santiago Murgui Izquierdo
Ezequiel Uriel Jimenez
Facultad de Económicas
Universitat de Valencia

RESUMEN

Las medidas de desigualdad sobre distribuciones univariantes pueden ser planteadas desde distintas perspectivas. Una de ellas consiste en identificar la desigualdad como una distancia entre distribuciones. Profundizando en este planteamiento, es posible expandir los resultados a distribuciones multivariantes. Las medidas así construidas están expuestas a diversas críticas, pero en contrapartida gozan de las propiedades tradicionalmente exigidas a este tipo de indicadores.

Palabras clave: Medidas de desigualdad; distancia; multivariante

1.-INTRODUCCIÓN

Considérese un colectivo integrado por N unidades con pesos respectivos $p_1, p_2, \dots p_N$ que verifican la relación $\sum_{i=1}^{N} p_i = 1$. Sean $x_1, x_2, \dots x_N$ los valores que adopta una variable X sobre cada una de las unidades del colectivo. La construcción de índices para medir la desigualdad asociada con la distribución de X tiene largos antecedentes en la literatura estadística. Recientemente Goerlich(1998) ha efectuado una amplia recopilación de las aportaciones más interesantes sobre el tema.

La calidad de un índice de desigualdad debe medirse a través de un análisis de sus propiedades y de su sensibilidad frente a posibles variaciones planteadas sobre la distribución para la que se construyen. Las propiedades básicas habitualmente exigidas para un índice de desigualdad son:

- Independencia del tamaño absoluto de las unidades que integran el colectivo.

- Independencia de la escala o unidad de medida de la variable observada.

- Compatibilidad con el principio de Pigou-Dalton que asegura la minoración del valor de la desigualdad cuando disminuye la desviación entre los valores observados de X sobre dos unidades en las que no llega a invertirse el orden de prelación.

Al margen de las propiedades enunciadas, deben contemplarse otras referentes a la capacidad de descomposición de los índices. En ocasiones, el colectivo de referencia puede organizarse en grupos de unidades y resulta útil conocer la desagregación correspondiente de la desigualdad global. En otras ocasiones, la variable de interés tiene una definición compleja en la que participan un cierto número de variables simples. La posibilidad de descomponer el índice de desigualdad en factores aditivos, permite en estos casos analizar la aportación de las distintas componentes a la desigualdad global de la distribución.

Un primer conjunto de índices de desigualdad utilizados en la literatura es el que definen las medidas de dispersión estadística. Las más conocidas son:

- Rango relativo: $\frac{[max(x_i) - min(x_i)]}{m_x}$
- Ratio entre valores extremos: $1 - \frac{min(x_i)}{max(x_i)}$
- Coeficiente de variación: $C_v^2 = \frac{V[X]}{m_x}$ donde $V[X] = \sum_{i=1}^{N} (X_i - m_x)^2 p_i$
- Varianza logarítmica: $V \log = \sum_{i=1}^{N} (\log x_i - \log m_x)^2 p_i$

Algunas referencias acerca de las propiedades de estos índices pueden verse en Sen (1973) y Cowell(1995).

El segundo conjunto de índices de desigualdad al que haremos referencia está formado por las medidas construidas a partir de la curva de Lorenz. Los más utilizados son:

- Indice de Gini: cociente entre dos superficies delimitadas por la curva, con una expresión algebraica igual a:

$$I_g = \frac{\sum_i \sum_j |x_i - x_j| p_i p_j}{2m_x}$$

- Distancia máxima entre la curva y la distribución correspondiente a una situación con desigualdad mínima, con expresión algebraica igual a:

$$I_m = \frac{\sum_i |x_i - m_x| p_i}{2m_x}$$

Referencias clásicas a estos índices pueden encontrarse en Gini(1912) y Schultz(1951).

Finalmente, sin ánimo de ser exhaustivos ni de relativizar la trascendencia de otras aportaciones, como la que se deriva de la función de bienestar social, haremos referencia a la familia de indicadores de Theil generalizados. Esta familia se define sobre un parámetro b a través de la expresión:

$$T_b(X) = \frac{1}{b(b-1)} \sum_{i=1}^{N} \left[\left(\frac{x_i}{m_x} \right)^b - 1 \right] p_i$$

Como casos particulares, para $b = 0$ y $b = 1$ tomando límites se obtienen los índices más difundidos en las aplicaciones:

$$T_0(X) = \sum_{i=1}^{N} \log\left(\frac{m_x}{x_i} \right) p_i \qquad \text{y} \qquad T_1(X) = \sum_{i=1}^{N} \frac{x_i}{m_x} \log\left(\frac{x_i}{m_x} \right)$$

La primera referencia a estos índices es debida al propio Theil(1967). Posteriormente, otros autores como Shorrocks(1980) y Love y Wolfson(1976) han analizado la influencia del parámetro b en la sensibilidad de los índices.

En el ámbito económico es muy frecuente la utilización de las medidas de desigualdad sobre variables tales como la renta o el gasto. En estas circunstancias, las unidades del colectivo suelen ser familias cuyos pesos p_i están determinados por el cociente $\frac{M_i}{M}$ entre el número de miembros de la familia y el número total de individuos en el colectivo. Asimismo, la variable X suele hacer referencia a la renta o gasto per cápita de cada familia. De manera alternativa, es frecuente reescribir los indicadores de desigualdad haciendo referencia a los valores y_i que en cada unidad expresan la proporción de la renta familiar sobre la global o el gasto acumulado por todos sus miembros frente al total del colectivo.

La propia definición de los elementos anteriores garantiza las siguientes equivalencias entre los valores asociados a las distintas unidades del colectivo:

$$y_i = \frac{x_i M_i}{\sum_{i=1}^{N} x_i M_i} \qquad m_x = \frac{1}{M} \sum_{i=1}^{N} x_i M_i$$

$$y_i = \frac{x_i p_i}{m_x} \qquad \frac{x_i}{m_x} = \frac{y_i}{p_i}$$

Con referencia a las proporciones acumuladas y_i, el índice de Theil generalizado adopta la siguiente expresión

$$T_b(X) = \frac{1}{b(b-1)} \sum_{i=1}^{N} \left[\left(\frac{y_i}{p_i} \right)^b - 1 \right] p_i$$

y los índices particulares para $b = 0$ y $b = 1$ son:

$$T_0(X) = \sum_{i=1}^{N} \log\left(\frac{p_i}{y_i} \right) p_i = -\sum_{i=1}^{N} \log\left(\frac{y_i}{p_i} \right) p_i$$

$$T_1(X) = \sum_{i=1}^{N} \log\left(\frac{y_i}{p_i} \right) y_i$$

2. DISTANCIA ENTRE DISTRIBUCIONES

El concepto de distancia entre dos distribuciones ha sido ampliamente tratado en la literatura estadística en distintos contextos. No es el objetivo de este epígrafe efectuar una revisión crítica de las diferentes distancias habitualmente utilizadas. Por ahora, únicamente nos referiremos a una familia particular de ellas sin entrar en evaluaciones comparativas frente a otras posibles alternativas.

Considérese que sobre el colectivo propuesto anteriormente, además de la variable X se dispone de otra variable Z a cuyos valores observados los denotamos por $z_1, z_2, \; z_N$. Una familia de distancias entre las distribuciones asociadas a las dos variables es la que determina la siguiente expresión definida sobre un parámetro a :

$$d_a(X,Z) = \frac{1}{a(a+1)} \sum_{i=1}^{N} \frac{x_i}{m_x} \left[\left(\frac{x_i m_z}{m_x z_i} \right)^a - 1 \right] p_i$$

Dos casos particulares especialmente relevantes son los correspondientes a $a = -1$ y $a = 0$. Tomando límites para el primero se obtiene la distancia

$$d_{-1}(X,Z) = \sum_{i=1}^{N} \frac{z_i}{m_z} \log\left(\frac{z_i m_x}{m_z x_i} \right) p_i$$

y para el segundo

$$d_0(X,Z) = \sum_{i=1}^{N} \frac{x_i}{m_x} \log\left(\frac{x_i m_z}{m_x z_i} \right) p_i$$

Si en lugar de medir la distancia de la distribución de X a una única distribución, se pretendiera establecer una distancia con respecto a un conjunto de distribuciones asociadas con R variables (Z_1,Z_2, Z_R), sería necesario expandir la expresión anterior incorporando algunos elementos de la distribución conjunta del vector $\vec{Z}$ que tales variables definen.

Un procedimiento simple para construir una distancia global conjunta entre las distribuciones asociadas con la variable X y el vector $\vec{Z}$, consiste en agregar las distancias separadas entre la distribución de X y cada una de las componentes Z_r. Una primera propuesta analíticamente operativa y que permite resolver algunos problemas de optimización es la formulada por Burbea y Rao(1982). Como aspecto negativo, tiene el inconveniente de ignorar aspectos esenciales de la distribución conjunta del vector $\vec{Z}$, por lo que en trabajos posteriores, está previsto analizar otras propuestas más complejas. Por el momento limitaremos el planteamiento a una familia de distancias definida por la siguiente expresión:

$$d_a\left(X,\vec{Z}\right)=\sum_{r=1}^{R} a_r d_a\left(X,Z_r\right)$$

donde a_r expresa el peso asignado a la componente Z_r del vector, verificándose la relación $\sum_{r=1}^{R} a_r = 1$.

A los efectos prácticos, es posible determinar el valor de a_r a través del cociente $\frac{m_r}{\sum_{r=1}^{R} m_r}$, en el que m_r expresa el valor medio adoptado por la variable Z_r en el colectivo de interés.

Si para cada variable Z_r denotamos por w_{ri} a las proporciones acumuladas observadas sobre las distintas unidades del colectivo, es posible referir las distancias anteriores a los correspondientes valores acumulados de X y Z_r. En este caso, la distancia a una única componente del vector adopta la expresión:

$$d_a\left(X,Z_r\right)=\frac{1}{a\left(a+1\right)}\sum_{i=1}^{N} y_i\left[\left(\frac{y_i}{w_{ri}}\right)^{a-1}\right]$$

y la distancia global conjunta con respecto al vector $\vec{Z}$ puede escribirse como:

$$d_a\left(X,\vec{Z}\right)=\frac{1}{a\left(a+1\right)}\sum_{r=1}^{R}\sum_{i=1}^{N} a_r y_i\left[\left(\frac{y_i}{w_{ri}}\right)^{a}-1\right]$$

verificándose las relaciones $\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} w_{ri} = 1$ para r=1,2, ,R.

En particular, para $a = 0$ y $a = 1$ se obtienen las siguientes distancias:

$$d_{-1}(X, Z_r) = \sum_{i=1}^{N} w_{ri} \log\left(\frac{w_{ri}}{y_i}\right) \qquad d_0(X, Z_r) = \sum_{i=1}^{N} y_i \log\left(\frac{y_i}{w_{ri}}\right)$$

$$d_{-1}(X, \vec{Z}) = \sum_{r=1}^{R}\sum_{i=1}^{N} a_r w_{ri} \log\left(\frac{w_{ri}}{y_i}\right) \qquad d_0(X, \vec{Z}) = \sum_{r=1}^{R}\sum_{i=1}^{N} a_r y_i \log\left(\frac{y_i}{w_{ri}}\right)$$

Una vez definida una distancia global entre la distribución asociada con X y la conjunta del vector $\vec{Z}$, es posible plantearse la determinación de una distribución teórica y específica de X para la cual se minimice la citada distancia. Operando sobre la expresión de la distancia definida para proporciones acumuladas, minimizando $d_a(X, \vec{Z})$ con respecto a los valores de y_i se obtiene que el valor óptimo se consigue para:

$$Y_i = k\left(\sum a_r w_{ri}^{-a}\right)^{-\frac{1}{a}}$$

En particular, para las distancias asociadas con los valores paramétricos $a = -1$ y $a = 0$, el valor mínimo se consigue respectivamente cuando

$$y_i = \sum_{r=1}^{R} a_r w_{ri} \qquad \text{y} \qquad y_i = \prod_{r=1}^{R} w_{ri}^{a_r}$$

Observar que la solución óptima correspondiente al caso $a = 0$ está determinada por las medias geométricas de las distintas componentes del vector y en consecuencia, no esta normalizada. En la práctica, a los efectos de calcular distancias a dicha distribución sería necesario corregir tales valores medios utilizando los cocientes $\frac{y_i}{\sum_{i=1}^{N} y_i}$.

En el caso $a = -1$, las proporciones acumuladas obtenidas en el proceso de minimización si que están normalizadas y se comprueba que la solución coincidiría con la que se hubiese obtenido en el supuesto de haber adoptado una distancia cuadrática ordinaria, definida mediante la expresión $d(X, Z) = \sum_{i=1}^{N} (x_i - z_i)^2 p_i$.

En general, para un valor paramétrico a, si referimos las expresiones que minimizan la distancia a los valores directamente observados de las variables X y Z_r, se deduce que la distribución más próxima a la asociada con el vector $\vec{Z}$ es la que verifica la siguiente relación:

$$\frac{x_i\, p_i}{m_x} = k\left[\ \sum a_r \left(\frac{z_{ri}\, p_i}{m_z}\right)^{-a}\ \right]^{-\frac{1}{a}}$$

3.DISTANCIA Y DESIGUALDAD

En el epígrafe 1 se ha presentado la familia de los indicadores de Theil como un conjunto de alternativas para medir la desigualdad de una distribución. Analizando comparativamente las expresiones que determinan dichos índices y la distancia entre dos distribuciones definida en el epígrafe 2, se comprueba que el índice de desigualdad asociado con la distribución de una variable coincide con la distancia existente entre dicha distribución y la que correspondería a una variable que tomara un valor constante sobre todas las unidades del colectivo.

Si H es una variable que toma un valor constante sobre todas las unidades, la distancia entre la distribución asociada con X y la de H adopta la expresión:

$$d_a(X,H) = \frac{1}{a(a+1)} \sum_{i=1}^{N} \frac{x_i}{m_x} \left[\left(\frac{x_i}{m_x} \right)^a - 1 \right] p_i$$

Es fácil comprobar que esta expresión coincide con el índice de desigualdad $T_b(X)$ cuando $b = a + 1$.

De forma equivalente, reescribiendo el índice de desigualdad con referencia a las proporciones acumuladas puede comprobarse que:

$$T_b(X) = d_{b-1}(X,H) = \frac{1}{(b-1)b} \sum_{i=1}^{N} y_i \left[\left(\frac{y_i}{p_i} \right)^{b-1} - 1 \right]$$

siendo por lo tanto esta última expresión equivalente a la enunciada anteriormente.

Observar que una variable con valor constante sobre todas las unidades del colectivo, ofrecerá unas proporciones acumuladas idénticas a los pesos asignados a cada unidad y en consecuencia, representará una situación con desigualdad nula o de reparto uniforme. En particular, los índices de Theil correspondientes a los valores paramétricos $b = 0$ y $b = 1$ están determinados, respectivamente, por $d_{-1}(X,H)$ y $d_0(X,H)$.

Para construir un indicador de la desigualdad global asociada con la distribución conjunta del vector $\vec{Z}$, Maasoumi(1986) propone representar la citada distribución por una única distribución univariante convenientemente elegida, siendo la medida de desigualdad de está última la que se asociaría a todo el vector.

Como tal distribución se propone elegir la más próxima al vector en el sentido de que minimice la distancia $d_a\left(X,\bar{Z}\right)$. Consecuentemente, la desigualdad global asociada con el vector $\bar{Z}$ adoptará la expresión $T_b\left(\vec{Z}\right)=T_b\left(X\right))=d_{b-1}\left(X,H\right)$, siendo la distribución de X la que minimiza la distancia $d_a\left(X,\vec{Z}\right)$ y cuyas proporciones acumuladas están determinadas por $y_i \ = k\left(\sum_{r=1}^{R} a_r w_{ri}^{-a}\right)^{-\frac{1}{a}}$.

Es importante señalar que la propuesta formulada para medir la desigualdad global de un vector admite una doble flexibilización. Por una parte, el indicador está definido sobre un parámetro al que pueden asignarse distintos valores. Por otra, hay que destacar la amplitud de posibilidades que ofrece el proceso de elección de la distribución que representará al vector. Dicha distribución se determina minimizando la distancia global, pero esta distancia también se ha definido sobre un parámetro que puede adoptar una gran variedad de valores. Dependiendo del valor paramétrico asignado a la distancia, se obtendrá una solución distinta para la distribución que represente al vector y sobre ella procederá adoptar la medida de desigualdad que se juzgue más adecuada.

Puesto que la desigualdad global conjunta de una distribución multivariante se reduce a la desigualdad de una distribución univariante convenientemente elegida, las propiedades que posee el indicador así construido son las mismas que se atribuyen a los indicadores tradicionales de Theil definidos para distribuciones con una sola dimensión. En particular, es independiente de los tamaños absolutos de las unidades del colectivo y de la escala de medida de las variables. Igualmente verifica el principio de transferencia de Pigou-Dalton y admite una descomposición en factores cuando se efectúa un agrupamiento de las unidades.

Una atención especial merece el análisis de la aportación de cada una de las componentes Z_r a la desigualdad conjunta global.

En general, no siempre es posible obtener descomposiciones de $T_b\left(\vec{Z}\right)$ que faciliten el cálculo de tales aportaciones para cualquier valor paramétrico b y cualquiera que sea la distribución elegida para representar a la conjunta. A continuación se ofrecen dos resultados particulares sobre los índices de Theil más conocidos que ilustran estos comentarios.

Supóngase que la distribución elegida para representar a la asociada con el vector $\vec{Z}$ es la que minimiza la distancia $d_0\left(X,\vec{Z}\right)$, lo que conduce a tomar las medias geométricas de las distintas componentes. Supóngase además que se decide medir la desigualdad global a través de $T_0\left(\vec{Z}\right)=d_{-1}\left(X,H\right)$. En estas condiciones se verifica $T_o\left(\vec{Z}\right)=\sum_{r=1}^{R} a_r T_0\left(Z_r\right)$.

Como consecuencia de esta identidad se concluye que la desigualdad global asociada con todo el vector se descompone en la suma ponderada de las desigualdades correspondientes a cada una de las componentes, siendo su peso en el conjunto agregado el que determina su grado de aportación a la desigualdad global.

Como alternativa al caso anterior, si para representar a la distribución conjunta del vector $\vec{Z}$ se elige la univariante más próxima a la conjunta minimizando la distancia $d_{-1}\left(X,\vec{Z}\right)$ y la desigualdad global se define a través del indicador $T_1(X)=d_0(X,H)$, se verifica la relación $T_1\left(\vec{Z}\right)=\sum_{r=1}^{R}a_r T_1(Z_r)-d_0\left(X,\vec{Z}\right)$.

En este segundo caso, a la suma de las aportaciones de cada una de las componentes debe sustraerse la distancia entre la distribución elegida para representar al vector y este mismo, aunque con la particularidad de que el valor paramétrico correspondiente a dicha distancia no coincide con el utilizado en la determinación de la distribución más próxima al vector.

4. CONCLUSIONES

A lo largo del trabajo se ha conseguido elaborar un indicador de la desigualdad global asociada con una distribución multivariante que goza de las propiedades tradicionalmente exigidas a los índices de desigualdad definidos sobre distribuciones univariantes.

La equivalencia entre una medida de desigualdad y una distancia convenientemente elegida no es en absoluto ajena a la lógica de los argumentos habitualmente esgrimidos sobre estos tópicos. En un futuro inmediato, este planteamiento va a exigir una revisión crítica de las distancias a considerar. En la medida en que se utilicen métricas con mayor capacidad para captar la estructura de una distribución multivariante, será más fiel la representación univariante que se genere y en consecuencia más adecuada la medida de desigualdad construida.

Por el momento en esta primera aproximación, el interés por asegurar una solución al problema de optimización planteado, compatible con los indicadores de desigualdad de Theil, ha exigido recurrir a una familia de distancias particular que presenta notables limitaciones. Entre estas últimas resulta especialmente llamativa la ausencia de referencias a las posibles relaciones existentes entre las componentes del vector.

La viabilidad práctica de las expresiones que aquí se proponen ha sido puesta de manifiesto en una aplicación realizada al gasto familiar sobre datos de la Encuesta de Presupuestos Familiares de 91 realizada por Moltó, Murgui y Uriel(1998).

REFERENCIAS

Burbea j. and .R. Rao (1982) “Entropy differential metric, distance and divergence measures in probability spaces: a unified approach”. *Journal of Multivariate Analysis*, 12. 575-596

Cowell, F. (1995) *Measuring inequality*. 2^{nd} Edition, LSE Hanbooks in Economics, Prentice Hall. London, London (1^{st}. Edition 1977, Philip Alan Publiserhs Limited, London)

Goerlich, F. (1998) “Desigualdad, diversidad y convergencia: Algunos instrumentos de medida”. Instituto Valenciano de Investigaciones Económicas.

Gini, C. (1912) “Variabilità e mutabilità, contributo allo studio delle distribuzioni e relazioni statistiche”. *Studi Economico-Giuridici dell'Universiti de Cagliari*,3,part 2,1-158

Love, R. and Wolfson, M. C (1976) *Income Inequality: Statistical Methodology and Canadian Illustrations*. Otawa, Satatistics Canada.

Maasoumi, E. (1986) “The measurement and descomposition of multidimensional inequality”. *Econometrica*, 54, 4.991-997.

Moltó, M.L., Murgui, S. y Uriel, E. (1998) “Determinación de una tipología de hogares en el marco de una matriz de contabilidad social”. Instituto Valenciano de Investigaciones Económicas.

Schutz, R.R. (1951) "On the measurement of income inequality", *American Economic Review*, 41, (March), 107-122.

Sen, A. (1973) *On economic inequality*. Oxford University Press. Oxford.

Theil, H. (1967) *Economics and information theory*, Amsterdam, Noth-Holland.