UNA MEDIDA DEL ACERCAMIENTO ENTRE DATOS DE DOS CARACTERÍSTICAS CON MISMO RANGO

Ernesto J. Veres Ferrer
Dpto. Economía Aplicada
Universidad de Valencia

RESUMEN

Se presenta una medida que expresa el grado de acercamiento o proximidad entre dos conjuntos de datos $\{X_i\}_{i=1}^{N}$ y $\{Y_i\}_{i=1}^{M}$, pertenecientes a sendas variables o atributos *X* e *Y* que tienen el mismo rango numérico de definición o mismo conjunto de modalidades posibles, y que puede entenderse en el sentido de hasta qué punto ambos conjuntos son concreciones de una misma característica. El conocido contraste no paramétrico de homogeneidad da respuesta a la pregunta planteada. No obstante, las exigencias de dichos contrastes a la hora de determinar un número mínimo de datos por categoría pueden invalidar, en algunos casos, su correcta aplicación. Por ello, planteamos en lo que sigue una alternativa en la que no es determinante contar con un número mínimo de datos por categoría. Se realizan dos aplicaciones de la medida propuesta sobre sendos conjuntos de datos: sobre las esperanzas de vida y sobre los PIB per cápita para un grupo de países.

1. Introducción

Sean $\{X_i\}_{i=1}^{N}$ y $\{Y_i\}_{i=1}^{M}$ dos conjuntos de datos que pertenecen a sendas variables o atributos *X* e *Y* que tienen el mismo rango de definición. El conocido contraste de la χ^2 de la bondad de ajuste podría ser la base para definir una medida del grado de acercamiento entre ambos conjuntos. Basta definir un conjunto de *K* categorías, determinar las frecuencias de los datos en cada una de ellas, dar a uno de los conjuntos la consideración de distribución teórica, y ajustar las frecuencias del otro conjunto, calculando el conocido estadístico

$$\chi^2_{K-1} = \sum_{i=1}^{K} \frac{\left(\frac{M_i}{M} \times N - N_i \right)^2}{N_i}$$

donde N_i y M_i es el número de datos de las variables X e Y , respectivamente, que pertenecen a la categoría *i-ésima*. Ese estadístico tiene asociado un *p-valor* que denotaremos por p_{xy}. Intercambiando el papel de las variables X e Y se obtendría el estadístico

$$\chi^2_{K-1} = \sum_{i=1}^{K} \frac{\left(\frac{N_i}{N} \times M - M_i \right)^2}{M_i}$$

que también tiene asociado otro *p-valor* que denotaremos por $p_{yx.}$ La media geométrica de ambos

$$p = \sqrt{p_{xy}\, p_{yx}}$$

define una medida, en términos de probabilidad, del posible acercamiento entre la estructura de los datos considerados. Debe tenerse en cuenta todas las limitaciones sobre la formación de las categorías (Cochran, 1952), cuyo número y extensión adolecen de cierta arbitrariedad, y la influencia de valores grandes de N, conocida como el efecto de la N inflada (Runyon & Haber, 1967).

Las exigencias de los contrastes de la χ^2 de la bondad de ajuste a la hora de determinar un número mínimo de datos por categoría pueden invalidar, en algunos casos, su correcta aplicación. Por ello, planteamos en lo que sigue una alternativa en la que no es determinante contar con un número mínimo de datos por categoría. Se basa en una medida que, en forma de índice, está definida en Beamonte *et al* (2004a) para valorar la calidad medioambiental del agua. Con el nombre de índice global de calidad, tiene la siguiente expresión, aplicada sobre un vector de componentes naturales *(a b c d)*:

$$I(a,b,c,d) = \frac{1}{6}(s_1^3 + 3s_1^2 + 2s_1) + \frac{1}{2}(s_2^2 + s_2) + v_1 + 1$$

con $s_1 = a + b + c$, $s_2 = a + b$ y $k = a + b + c + d$ constante.

Respeta la siguiente ordenación: siendo $(a_1\, b_1\, c_1\, d_1)$ y $(a_2\, b_2\, c_2\, d_2)$ dos vectores tales que $k = a_1 + b_1 + c_1 + d_1 = a_2 + b_2 + c_2 + d_2$, entonces:

1. $Si\ d_1 < d_2 \rightarrow I(a_1b_1c_1d_1) > I(a_2b_2c_2d_2)$
2. $Si\ d_1 = d_2\ y\ c_1 < c_2 \rightarrow I(a_1b_1c_1d_1) > I(a_2b_2c_2d_2)$
3. $Si\ d_1 = d_2\ y\ c_1 = c_2\ y\ b_1 < b_2 \rightarrow I(a_1b_1c_1d_1) > I(a_2b_2c_2d_2)$

Además, el índice anterior tiene como máximo valor $\frac{1}{6}(k+1)(k+2)(k+3)$, que corresponde al vector *(k, 0, 0, 0)* y como valor mínimo 1, que corresponde al vector *(0, 0, 0, k).*

En Beamonte *et al* (2004b) se aplica el índice anterior en una modelización bayesiana para el comportamiento de los datos, confirmando la extensión de la aplicabilidad del índice propuesto.

2. Definición de una medida de acercamiento entre estructuras de datos

La idea en la que se basa todo el desarrollo posterior es común con la de otros contrastes no paramétricos: *si las estructuras de datos de dos características que tienen el mismo rango son semejantes, cualquier categorización de los mismos proporcionará distribuciones de frecuencias absolutas parecidas*. Por ello, la comparación entre esas frecuencias resulta relevante a nuestros propósitos.

Sigamos considerando $\{X_i\}_{i=1}^{N}$ y $\{Y_i\}_{i=1}^{M}$ dos conjuntos de datos que pertenecen a sendas variables o atributos *X* e *Y* que tienen el mismo rango numérico de definición o el mismo conjunto de modalidades posibles. Sea el siguiente proceso:

a) *Categorización del rango de los datos*

En el caso de que *X* e *Y* sean variables, consideremos $a = \min_{i,j}(X_i, Y_j)$ y $b = \max_{i,j}(X_i, Y_j)$ con $1 \le i \le N$ y $1 \le j \le M$. Construyamos las siguientes *tres* categorías para la clasificación de los datos:

$$C_1 = \left[a, a + \frac{b-a}{3} \right[= \left[a, \frac{2a+b}{3} \right[$$

$$C_2 = \left[\frac{2a+b}{3}, \frac{2a+b}{3} + \frac{b-a}{3}\right[= \left[\frac{2a+b}{3}, \frac{a+2b}{3}\right[$$
$$C_3 = \left[\frac{a+2b}{3}, \frac{a+2b}{3} + \frac{b-a}{3}\right] = \left[\frac{a+2b}{3}, b\right]$$

Si *X* e *Y* son atributos ordenados, las categorías C_i se construyen agrupando en tres categorías un mismo número de modalidades adyacentes. Si se trata de atributos no ordenados, resultaría válida cualquier agrupación de modalidades que respetara cierta lógica. Por ejemplo, para el atributo *estado civil* podrían considerarse las categorías de *solteros*, *casados* y *resto*, como exponentes, respectivamente, de ausencia de vínculo matrimonial, existencia activa de dicho vínculo, y de ruptura por cualquier motivo del mismo.

b) Clasificación de los datos en las distintas categorías

La clasificación de los datos en esas tres categorías se recoge en la tabla siguiente:

	C_1	C_2	C_3
X	N_1	N_2	N_3
Y	M_1	M_2	M_3

donde

$$N_i = n^o\, datos\, de\, X \in C_i \quad y \quad \sum_{i=1}^{3} N_i = N$$
$$M_j = n^o\, datos\, de\, Y \in C_j \quad y \quad \sum_{j=1}^{3} M_j = M$$

c) Ajuste del número de datos M de Y a N al número de datos N de X

El papel de las características *X* e *Y* es intercambiable. Por ello, vamos a considerar a *X* como la variable referente, debiendo ajustar las frecuencias de *Y* para que su total coincida con *N*. Todo el proceso habrá que repetirlo posteriormente, intercambiando los papeles de *X* e *Y*. La tabla anterior se transforma en la siguiente:

	C_1	C_2	C_3
X	N_1	N_2	N_3
Y	$M_1^* = M_1 \times N / M$ M 1	$M_2^* = M_2 \times N / M$	$M_3^* = M_3 \times N / M$

donde

$$\sum_{j=1}^{3} M_j^* = N$$

Los valores M_i^* se redondean al entero más próximo.

d) Los errores y aciertos de clasificación

Sea $\varepsilon_i = error\ de\ clasificación\ en\ C_i = N_i - M_i^*$ con $\sum_{i=1}^{3} \varepsilon_i = 0$. Se cumple:

$$Si\ \exists i_0\ tal\ que\ \varepsilon_{i_0} \neq 0 \rightarrow \exists j_0 \neq i_0\ tal\ que\ \varepsilon_{j_0} \neq 0$$

esto es, si existen errores, al menos los hay en dos categorías de clasificación.

Sea $m_i = \min_i(N_i, M_i^*)\ i = 1,2,3$ los aciertos (coincidencias clasificación de *X* e *Y*) en C_i.

Se verifica:

$$A_x = total\ aciertos\ de\ X = \sum_{i=1}^{3} m_i = total\ aciertos\ de\ Y = A_y = A$$

Se cumple la siguiente *ecuación compensadora*:

$$A_x + A_y + \sum_{i=1}^{3} |\varepsilon_i| = 2N \qquad \text{ó} \qquad 2A + \sum_{i=1}^{3} |\varepsilon_i| = 2N$$

esto es, la suma de los aciertos o coincidencias de *X* e *Y* y de la suma de los valores absolutos de los errores es un número fijo igual al doble del número de datos.

e) Vector de discrepancias acumuladas

Sea el $V' = (v_1\ v_2\ v_3\ v_4)$ *el vector de discrepancias acumuladas* cuyas componentes son las siguientes:

$$v_1 = 6(A_x + A_y) = 12A$$
$$v_2 = 2|\varepsilon_1|$$
$$v_3 = 2|\varepsilon_1| + 3|\varepsilon_2|$$
$$v_4 = 2|\varepsilon_1| + 3|\varepsilon_2| + 6|\varepsilon_3|$$

La elección de las ponderaciones 2, 3 y 6 aplicadas sobre $|\varepsilon_1|$, $|\varepsilon_2|$ y $|\varepsilon_3|$ se justifica en el apartado siguiente.

f) Medida del acercamiento entre los datos de X a los de Y para las categorías de clasificación C_1, C_2 y C_3

Puede aplicarse al vector *V* el índice global de calidad definido en Beamonte *et al* (2004a), puesto que sus cuatro componentes son números naturales de suma constante, al ser:

$$\sum_{s=1}^{4} v_s = 12A + 6|\varepsilon_1| + 6|\varepsilon_2| + 6|\varepsilon_3| = 6\left(2A + \sum_{i=1}^{3}|\varepsilon_i|\right) = 12N$$

dado que las ponderaciones 2, 3 y 6 utilizadas aseguran ese resultado.

Por tanto, si en la expresión del índice global hacemos $a = v_1$, $b = v_2$, $c = v_3$ y $d = v_4$, podemos utilizar sus propiedades para definir una medida apropiada a nuestros propósitos. Sin embargo, previamente hay que determinar para nuestra situación el rango de posibles valores de la nueva medida. El valor máximo para $I(v_1\ v_2\ v_3\ v_4)$ sigue siendo $\frac{1}{6}(12N+1)(12N+2)(12N+3)$, que corresponde al posible vector de errores acumulados *(12N, 0, 0, 0)*, y que es precisamente el que expresa coincidencia total en la clasificación de los datos en las tres categorías, no existiendo por tanto discrepancias en la clasificación. Sin embargo, el mínimo se alcanzará cuando no haya ninguna coincidencia (todo sean errores), con al menos un error en la categoría C_1, $|\varepsilon_1| = 1$, $|\varepsilon_2| = N-1$ y $|\varepsilon_3| = N$, dado que debe haber datos, para una u otra variable, clasificados en las categorías extremas C_1 y C_3. De ahí que el vector de errores acumulados sea *(0, 2, 3N-1, 9N-1)*, sobre el que el índice toma su valor mínimo igual a

$$\frac{1}{6}(27N^3 + 54N^2 + 33N + 30)$$

De esta forma, el rango de posibles valores será la diferencia entre los dos anteriores e igual a

$$\frac{1}{6}(1701N^3 + 810N^2 + 99N - 24)$$

Definimos, pues, el acercamiento de la estructura de datos *Y* a la de *X*, una vez determinado el vector de errores acumulados $(v_1\ v_2\ v_3\ v_4)$, a partir de la siguiente transformación normalizadora del índice de Beamonte *et al* (2004a):

$$I_{y\to x}^{C_1C_2C_3} = \frac{I(v_1\, v_2 v_3\, v_4) - I(0,2,3N-1,9N-1)}{I(12N,0,0,0) - I(0,2,3N-1,9N-1)} \times 100$$

que resulta ser

$$I_{y\to x}^{C_1C_2C_3} = \frac{(s_1^3 + 3s_1^2 + 2s_1) + 3(s_2^2 + s_2) + 6v_1 + 6 - \left(27N^3 + 54N^2 + 33N + 30\right)}{1701N^3 + 810N^2 + 99N - 24} \times 100 \quad [1]$$

siendo $s_1 = v_1 + v_2 + v_3$ *y* $s_2 = v_1 + v_2$.

Por misma construcción, la medida anterior toma un valor comprendido entre 0 –que expresaría un máximo alejamiento entre las estructuras de datos- y 100 –que supondría la coincidencia de clasificación de las dos estructuras-.

g) Medida del acercamiento entre los datos de X a los de Y para las categorías de clasificación C_1, C_2 y C_3

El problema planteado presenta simetría, esto es, podemos preguntarnos también por el acercamiento de la estructura de datos *X* a la de *Y*. Repitiendo el proceso anterior, tomando como variable referente la *Y*, puede obtenerse de forma análoga la medida $I_{x\to y}^{C_1C_2C_3}$. En un primer momento, proponemos como medida del acercamiento entre ambas estructuras la media geométrica de las dos anteriores:

$$I_{x\Leftrightarrow y}^{C_1C_2C_3} = \sqrt{I_{x\to y}^{C_1C_2C_3} \times I_{y\to x}^{C_1C_2C_3}}$$

lo que implica la simetría de medición, siendo la identidad evidente por misma construcción.

h) Propiedades de la medida propuesta

El cambio de origen y escala utilizado en [1] no modifica la propiedad del índice definido en Beamonte *et al* (2004a) de respetar la ordenación entre vectores. Así pues, se respeta la ordenación siguiente: *dados dos vectores de discrepancias acumuladas* $V' = (v_1\, v_2\, v_3\, v_4)$ *y* $W' = (w_1\, w_2\, w_3\, w_4)$, *el índice expresará un mayor grado de acercamiento si, y sólo si:*

a) $v_4 < w_4$, *ó*

b) $v_4 = w_4$ *y* $v_3 < w_3$, *ó*

c) $v_4 = w_4$ *y* $v_3 = w_3$ *y* $v_2 < w_2$,

y sólo expresarán igual acercamiento si, y sólo si, $v_i = w_i \ \forall i$.

En definitiva, el primer criterio para decidir sobre el grado de acercamiento entre dos estructuras de datos es el número total de errores o discrepancias obtenidos en la clasificación de los mismos en las tres categorías consideradas. En segundo lugar, y a igualdad del número anterior, la decisión se basa en el número de discrepancias observadas en las dos primeras categorías. Para finalizar, a igualdad de los dos primeros criterios, se considera las diferencias de la primera categoría. Así pues, la medida propuesta no sólo tiene en cuenta el número total de errores o discrepancias, sino también cómo van distribuyéndose éstas de forma acumulada.

El número de categorías para clasificar los datos ha sido tres, dado que el índice propuesto por Beamonte *et al* (2004a) trabaja con vectores de cuatro componentes y una de ellas hay que dejarla como expresión de los aciertos. Pero también los autores citados advierten ya de la generalización del indicador a más de cuatro componentes, por lo que la medida aquí propuesta podría, a su vez, extenderse si fueran necesarias más categorías de clasificación para los datos. En cualquier caso, el número de categorías utilizado en la clasificación es arbitrario.

i) Medida del acercamiento entre los datos de X e Y

La definición final de la medida debe tener en cuenta la existencia de unas ponderaciones sobre los errores –2, 3 y 6- exigidas para su correcta construcción. Resulta evidente que la asignación de una ponderación, la 6 por ejemplo, sobre los errores existentes en la tercera categoría de clasificación no deja de ser espúrea. Por otra parte, en la acumulación de errores resulta irrelevante la consideración continua del rango empleado, pues la medida no hace referencia a ello. De ahí que la medida de acercamiento entre las estructuras de los datos de *X* y de *Y* sea, finalmente, la media geométrica de las medidas parciales $I_{x \Leftrightarrow y}^{C_1 C_2 C_3}$ definidas sobre las seis permutaciones posibles entre las tres categorías C_i:

$$I_{x \Leftrightarrow y} = \sqrt[3!]{\prod_{i, j \neq i, k \neq i, j} I_{x \Leftrightarrow y}^{C_i C_j C_k}}$$

que sigue respetando las propiedades de identidad y de simetría que verificaban sus factores.

De esta forma, dadas sendas estructuras de datos correspondientes a cuatro características *X*, *Y*, *Z* y *T* el acercamiento entre los datos de *X* e *Y* será mayor que el acercamiento de los datos de *Z* y *T* si, y sólo si, $I_{x \Leftrightarrow y} > I_{z \Leftrightarrow t}$.

3. Aplicación

Vamos a realizar dos aplicaciones de la medida propuesta: sobre un conjunto de esperanzas de vida y sobre otro conjunto de los PIB per cápita de un grupo de países.

a) Esperanzas de vida

La Tabla 1 del Anexo recoge, por países, sus esperanzas de vida al nacer correspondientes al periodo 1996-2000, así como su categoría de clasificación según el Indice de Desarrollo Humano publicado por la ONU. Los primeros descriptivos se recogen en el Cuadro 1:

Cuadro 1: Descriptivos para la Esperanza de vida al nacer

Desarrollo Humano	**Estadístico**	
Alto	Número de datos	46
	Media	76,048
	Intervalo confianza media al 95%: límite inferior	75,317
	Límite superior	76,779
	Mediana	76,750
	Desv. típ.	2,462
	Mínimo	68,7
	Máximo	80,0
	Amplitud intercuartil	3,675
	Asimetría	-,923
	Curtosis	,487
Medio	Número de datos	93
	Media	66,765
	Intervalo confianza media al 95%: límite inferior	65,353
	Límite superior	68,176
	Mediana	68,900
	Desv. típ.	6,852
	Mínimo	44,1
	Máximo	76,0
	Amplitud intercuartil	6,600
	Asimetría	-1,411

	Curtosis	1,546
Bajo	Número de datos	35
	Media	48,834
	Intervalo confianza media al 95%: límite inferior	46,760
	Límite superior	50,909
	Mediana	48,500
	Desv. típ.	6,040
	Mínimo	37,2
	Máximo	60,7
	Amplitud intercuartil	8,500
	Asimetría	,010
	Curtosis	-,721

La prueba no paramétrica de Kruskal-Wallis, con una significación asintótica de 0,000, confirma la desigualdad estadística de las esperanzas de vida al nacer de los países según su desarrollo humano.

En primer lugar comparamos las estructuras de las esperanzas de vida correspondientes a los países de alto desarrollo humano (IDH = 1, y 46 países) y desarrollo humano medio (IDH = 2, y 93 países). Para el conjunto de esos 139 países, la esperanza de vida al nacer máxima es de 80 años, y la mínima 44,1 años. Dividiendo el correspondiente rango entre 3, la clasificación de los países por categorías C_i se recoge en la tabla siguiente:

IDH	C_1 [44,1 , 56,07[	C_2 [56,07 , 68,03[	C_3 [68,03 , 80]
1	0	0	46
2	11	29	53

Transformando la tabla anterior para la coincidencia de los totales, tomando como referente el total de países con IDH = 1, se obtiene la tabla:

IDH	C_1 [44,1 , 56,07[	C_2 [56,07 , 68,03[	C_3 [68,03 , 80]
1	0	0	46
2	6	14	26

en la que se observan 26 coincidencias (todas en la categoría C_3), y unos errores por categoría de 6, 14 y 20, respectivamente. De ahí que el vector de errores acumulados sea *(312, 12, 54, 174)*, que tiene asociada una medida $I_{2\to1}^{C_1C_2C_3} = 31'09$.

Transformando la tabla de datos inicial tomando como referente el total de países con IDH = 2, se obtiene la información siguiente:

IDH	C_1 [44,1 , 56,07[	C_2 [56,07 , 68,03[	C_3 [68,03 , 80]
1	0	0	93
2	11	29	53

en la que se observan 53 coincidencias (todas en la categoría C_3), y unos errores por categoría de 11, 29 y 40, respectivamente. De ahí que el vector de errores acumulados sea *(636, 22, 109, 349)*, que tiene asociada una medida $I_{1\to 2}^{C_1C_2C_3} = 31'42$.

Finalmente, pues, la medida asociada a la permutación $C_1C_2C_3$ será $I_{1\Leftrightarrow 2}^{C_1C_2C_3} = \sqrt{31'09 \times 31'42} = 31'26$.

El proceso se repite para las 5 permutaciones restantes, obteniéndose $I_{1\Leftrightarrow 2}^{C_1C_3C_2} = 35'90$; $I_{1\Leftrightarrow 2}^{C_2C_1C_3} = 33'51$; $I_{1\Leftrightarrow 2}^{C_2C_3C_1} = 46'32$; $I_{1\Leftrightarrow 2}^{C_3C_1C_2} = 40'08$ y $I_{1\Leftrightarrow 2}^{C_3C_2C_1} = 48'26$. Finalmente, la medida de acercamiento/alejamiento entre las esperanzas de vida al nacer de los países de desarrollo humano alto y medio será la media geométrica de esas seis medidas parciales $I_{1\Leftrightarrow 2} = 38'72$, valor más cercano al mínimo 0 que al máximo 100.

Si repitiéramos todo el proceso anterior, pero esta vez comparando las esperanzas de vida al nacer de los países de desarrollo humano medio (IDH = 2) y bajo (IDH = 3) obtendríamos una medida de acercamiento/alejamiento de los datos igual a $I_{2\Leftrightarrow 3} = 11'60$, valor todavía más cercano a 0 y que confirma una mayor distancia entre las esperanzas de los países de bajo desarrollo con los de desarrollo medio, que entre la de éstos y los de alto desarrollo.

b) PIB per cápita

La Tabla 2 del Anexo recoge, por países y para tres zonas geográficas concretas de pertenencia, su PIB per cápita correspondiente (salvo excepciones) al año 2000. Las zonas geográficas consideradas -según la clasificación de zonas utilizada en los

Informes sobre Desarrollo Humano publicado por la ONU- son Asia y el Pacífico (que incluye Asia oriental, Asia sudoriental y el Pacífico, y Asia meridional), América Latina y el Caribe (incluido México), y Estados árabes. Los primeros descriptivos se recogen en el Cuadro 2:

Cuadro 2: Descriptivos para el PIB per cápita

Zona	**Estadístico**	
Asia y el Pacífico	Número de datos	22
	Media	3543,36
	Intervalo confianza media al 95%: límite inferior	2027,12
	Límite superior	5059,60
	Mediana	2505,00
	Desv. típ.	3419,77
	Mínimo	1157
	Máximo	16765
	Amplitud intercuartil	2469,25
	Asimetría	3,057
	Curtosis	10,984
América Latina y el Caribe	Número de datos	33
	Media	5881,55
	Intervalo confianza media al 95%: límite inferior	4773,21
	Límite superior	6989,88
	Mediana	5161,00
	Desv. típ.	3125,72
	Mínimo	1383
	Máximo	14614
	Amplitud intercuartil	3858,50
	Asimetría	1,091
	Curtosis	,874
Estados árabes	Número de datos	18
	Media	7646,06
	Intervalo confianza media al 95%: límite inferior	4054,12
	Límite superior	11238,00
	Mediana	4559,00
	Desv. típ.	7223,05
	Mínimo	719
	Máximo	25314
	Amplitud intercuartil	7892,50
	Asimetría	1,359
	Curtosis	,949

La prueba no paramétrica de Kruskal-Wallis, con una significación asintótica de 0,002, confirma la desigualdad estadística de los PIB per cápita según la zona geográfica de pertenencia.

En primer lugar comparamos las estructuras del PIB per cápita correspondientes a los países de Asia y el Pacífico (zona geográfica 4 y 23 países) y América Latina y el Caribe (incluyendo México, zona geográfica 5 y 33 países). Para el conjunto de esos 56 países, el PIB per cápita máximo es de 24210 $, y el mínimo 1157 $. Dividiendo el correspondiente rango entre 3, la clasificación de los países por categorías C_i se recoge en la tabla siguiente:

Zona	C_1 [1157 , 8841,3[	C_2 [8841,3 , 16525,7[	C_3 [16525,7 , 24210]
Asia y el Pacífico	21	0	2
América Latina y el Caribe	28	5	0

Transformando la tabla anterior para la coincidencia de los totales, tomando como referente el total de Asia y el Pacífico, se obtiene la tabla:

Zona	C_1 [1157 , 8841,3[	C_2 [8841,3 , 16525,7[	C_3 [16525,7 , 24210]
Asia y el Pacífico	21	0	2
América Latina y el Caribe	20	3	0

en la que se observan 20 coincidencias (todas en la categoría C_1), y unos errores por categoría de 1, 3 y 2, respectivamente. De ahí que el vector de errores acumulados sea *(240, 2, 11, 23)*, que tiene asociada una medida $I_{5\to4}^{C_1C_2C_3} = 76'71$

Transformando la tabla de datos inicial tomando como referente el total correspondiente a América Latina y el Caribe, se obtiene la información siguiente:

Zona	C_1 [1157 , 8841,3[	C_2 [8841,3 , 16525,7[	C_3 [16525,7 , 24210]
Asia y el Pacífico	30	0	3
América Latina y el Caribe	28	5	0

en la que se observan 28 coincidencias (todas en la categoría C_1), y unos errores por categoría de 2, 5 y 3, respectivamente. De ahí que el vector de errores acumulados sea *(336, 4, 19, 37)*, que tiene asociada una medida $I_{4\to5}^{C_1C_2C_3} = 74'14$.

Finalmente, pues, la medida asociada a la permutación $C_1C_2C_3$ será $I_{4\Leftrightarrow 5}^{C_1C_2C_3} = \sqrt{76'71 \times 74'14} = 75'41$.

El proceso se repite para las 5 permutaciones restantes, obteniéndose $I_{4\Leftrightarrow 5}^{C_1C_3C_2} = 72'19$; $I_{4\Leftrightarrow 5}^{C_2C_1C_3} = 77'31$; $I_{4\Leftrightarrow 5}^{C_2C_3C_1} = 79'68$; $I_{4\Leftrightarrow 5}^{C_3C_1C_2} = 72'96$ y $I_{4\Leftrightarrow 5}^{C_3C_2C_1} = 78'54$. Finalmente, la medida de acercamiento/alejamiento entre las esperanzas de vida al nacer de los países de desarrollo humano alto y medio será la media geométrica de esas seis medidas parciales $I_{4\Leftrightarrow 5} = 75'96$, valor más cercano al máximo 100 que al mínimo 0.

Si repitiéramos todo el proceso anterior, pero esta vez comparando los PIB per cápita de los países de América Latina y el Caribe y los Estados árabes (zonas geográficas 5 y 6) obtendríamos una medida de acercamiento/alejamiento de los datos igual a $I_{5\Leftrightarrow 6} = 70'87$, valor cercano al anterior, expresando pues estructuras de PIB per cápita semejantes al comparar los países de Asia y el Pacífico con América Latina y el Caribe, o los de esta última zona con los Estados árabes.

4. Bibliografía

Beamonte, E. *et al* (2004a): Un indicador global para la calidad del agua. Aplicación a las aguas superficiales de la Comunidad Valenciana. *Estadística Española, vol. 46, nº 156,* 357-384.

Beamonte, E. *et al* (2004b): A statistical study of the surface water quality in the surroundings of Valencia town (Spain). *Technical Reports del Departamento de Estadística e Investigación Operativa de la Facultad de Matemáticas de la Universidad de Valencia, TR08-2004* (http://matheron.uv.es/investigar/technicals.html).

Cochran, W.G. (1952): The χ^2 test of goodness of fit. *Ann. Math. Statist., 23*, 315-345.

Runyon, R. & Haber, A. (1967): *Fundamentals of Behavioral Statistics.* Addison-Wesley. Massachusetts.

ONU (2004): *Informe sobre el Desarrollo Humano 2002.*

5. Anexo

Tabla 1

País	*IDH*	e_0	*País*	*IDH*	e_0
Japón	1	80	Iran, Rep. Isl.	2	69,2
Canadá	1	79	El Salvador	2	69,1
Islandia	1	79	Turquía	2	69
Suiza	1	78,7	Argelia	2	68,9
Suecia	1	78,6	Cabo Verde	2	68,9
Hong Kong (China)	1	78,5	República Árabe	2	68,9
Australia	1	78,3	Tailandia	2	68,8
Italia	1	78,2	Ucrania	2	68,8
Francia	1	78,1	Estonia	1	68,7
Grecia	1	78,1	Letonia	2	68,4
Noruega	1	78,1	Filipinas	2	68,3
España	1	78	Perú	2	68,3
Países Bajos	1	77,9	Belarús	2	68
Chipre	1	77,8	Nicaragua	2	67,9
Israel	1	77,8	Kazajstán	2	67,6
Alemania	1	77,2	Kirguistán	2	67,6
Bélgica	1	77,2	Moldova, Rep. de	2	67,5
Malta	1	77,2	Uzbekistán	2	67,5
Reino Unido	1	77,2	Vanuatu	2	67,4
Singapur	1	77,1	Viet Nam	2	67,4
Austria	1	77	Tayikistán	2	67,2
Nueva Zelandia	1	76,9	Brasil	2	66,8
Finlandia	1	76,8	Federación de Ru	2	66,6
Estados Unidos	1	76,7	Marruecos	2	66,6
Luxemburgo	1	76,7	Egipto	2	66,3
Barbados	1	76,4	Mongolia	2	65,9
Irlanda	1	76,4	Turkmenistán	2	65,4
Antigua y Barbuda	1	76	Indonesia	2	65,1
Costa Rica	2	76	Maldivas	2	64,5
Dominica	2	76	Guyana	2	64,4
Kuwait	1	75,9	Guatemala	2	64
Dinamarca	1	75,7	Pakistán	2	64
Cuba	2	75,7	Santo Tomé y Príncipe	2	64
Brunei Darussala	1	75,5	India	2	62,6
Portugal	1	75,3	Iraq	2	62,4
Chile	1	74,9	Bolivia	2	61,4
Emiratos Árabes	1	74,9	Bhután	3	60,7
Jamaica	2	74,8	Swazilandia	2	60,2
Belice	2	74,7	Myanmar	2	60,1
Eslovenia	1	74,5	Ghana	2	60
República Checa	1	73,9	Comoras	2	58,8
Uruguay	1	73,9	Bangladesh	3	58,1
Bahamas	1	73,8	Yemen	3	58
Trinidad y Tobago	2	73,8	Papua Nueva Guinea	2	57,9
Panamá	2	73,6	Madagascar	3	57,5
Macedonia, ERY	2	73,1	Nepal	3	57,3
Sri Lanka	2	73,1	Lesotho	2	56
Eslovaquia	1	73	Sudán	3	55
San Vicente y las Granadina	2	73	Camerún	2	54,7

Argentina	1	72,9	Sudáfrica	2	54,7
Bahrein	1	72,9	Haití	3	53,8
Albania	2	72,8	Mauritania	3	53,5
Fiji	2	72,7	Camboya	2	53,4
Georgia	2	72,7	Benin	3	53,4
Croacia	2	72,6	Malí	3	53,3
Polonia	1	72,5	Lao, Rep. Dem. P	3	53,2
Corea, Rep. de	1	72,4	Gabón	2	52,4
Venezuela	2	72,4	Namibia	2	52,4
México	2	72,2	Senegal	3	52,3
Granada	2	72	Kenya	2	52
Malasia	2	72	Congo, Rep. Dem.	3	50,8
Qatar	1	71,7	Eritrea	3	50,8
Islas Salomón	2	71,7	Djibouti	3	50,4
Samoa (Occidental)	2	71,7	Nigeria	3	50,1
Arabia Saudita	2	71,4	Guinea Ecuatoria	2	50
Mauricio	2	71,4	Togo	3	48,8
Bulgaria	2	71,1	Congo	2	48,6
Seychelles	2	71	Níger	3	48,5
Hungría	1	70,9	Tanzanía, Rep. U	3	47,9
Omán	2	70,9	Botswana	2	47,4
República Dominicana	2	70,6	Chad	3	47,2
Armenia	2	70,5	Gambia	3	47
Colombia	2	70,4	Côte d'Ivoire	3	46,7
Jordania	2	70,2	Angola	3	46,5
Suriname	2	70,1	Guinea	3	46,5
Jamahiriya Árabe	2	70	Mozambique	3	45,2
Rumania	2	70	Guinea-Bissau	3	45
Saint Kitts y Ne	2	70	República Centro Africana	3	44,9
Santa Lucía	2	70	Burkina Faso	3	44,4
Azerbaiyán	2	69,9	Zimbabwe	2	44,1
Líbano	2	69,9	Etiopía	3	43,3
Lituania	2	69,9	Burundi	3	42,4
China	2	69,8	Rwanda	3	40,5
Paraguay	2	69,6	Zambia	3	40,1
Ecuador	2	69,5	Uganda	3	39,6
Túnez	2	69,5	Malawi	3	39,3
Honduras	2	69,4	Sierra Leona	3	37,2

Tabla 1: Indice de Desarrollo Humano (IDH) y esperanzas de vida, por países. Fuente: ONU, *Informe sobre el Desarrollo Humano 2002.*
IDH = 1 países de alto desarrollo humano; IDH = 2 países de desarrollo humano medio; IDH = 3 países de bajo desarrollo humano.

Tabla 2

País	*Zona*	*PIB*	*País*	*Zona*	*PIB*
Singapur	4	24210	Panamá	5	5249
Brunei Darussalam	4	16765	Venezuela	5	5808
Malasia	4	8137	Suriname	5	5161
Fiji	4	4231	Colombia	5	6006
Tailandia	4	5456	Brasil	5	6625
Filipinas	4	3555	San Vicente y las Granadi	5	4692
Sri Lanka	4	2979	Perú	5	4282
Maldivas	4	4083	Paraguay	5	4288
Samoa (Occidental)	4	3832	Jamaica	5	3389

Iran, Rep. Isl. de	4	5121	República Dominicana	5	4598
Viet Nam	4	1689	Santa Lucía	5	5183
Indonesia	4	2651	Ecuador	5	3003
Vanuatu	4	3120	Guyana	5	3403
Islas Salomón	4	1940	El Salvador	5	4036
Myanmar	4	1199	Honduras	5	2433
India	4	2077	Bolivia	5	2269
Papua Nueva Guinea	4	2359	Nicaragua	5	2142
Pakistán	4	1715	Guatemala	5	3505
Camboya	4	1257	Haití	5	1383
Lao, Rep. Dem. Pop.	4	1734	Kuwait	6	25314
Bhután	4	1536	Bahrein	6	13111
Nepal	4	1157	Qatar	6	20987
Bangladesh	4	1361	Emiratos Árabes Unidos	6	17719
Barbados	5	12001	Jamahiriya Árabe Libia	6	6697
Bahamas	5	14614	Arabia Saudita	6	10158
Argentina	5	12013	Líbano	6	4326
Antigua y Barbuda	5	9277	Omán	6	9960
Chile	5	8787	Jordania	6	3347
Uruguay	5	8623	Túnez	6	5404
Saint Kitts y Nevis	5	10672	Argelia	6	4792
Costa Rica	5	5987	República Árabe Siria	6	2892
Trinidad y Tobago	5	7485	Egipto	6	3041
Dominica	5	5102	Marruecos	6	3305
Granada	5	5838	Iraq	6	3197
México	5	7704	Sudán	6	1394
Cuba	5	3967	Yemen	6	719
Belice	5	4566	Djibouti	6	1266

Tabla 2: Zona geográfica y PIB per cápita, por países. Fuente: ONU, *Informe sobre el Desarrollo Humano 2002.*
Zona 4: Asia y el Pacífico; Zona 5: América Latina y el Caribe; Zona 6: Estados Arabes.