CAPÍTULO 1-5
MEDICIÓN DE LA DISIMILITUD

1-5.1 MULTIDIMENSIONALIDAD, DISIMILITUD Y CONCENTRACIÓN

1-5.1.1 Problemática de la medición de la disimilitud

Vimos que una medición asociada a un concepto establece una correspondencia entre objetos y números. Lo que permite comparar objetos y determinar el valor de verdad de una o varias de las relaciones $=, \neq, <$ o $>$. Si un concepto contiene varias dimensiones, lo que es frecuente y queremos sin embargo tratarlo como un todo, vimos que al construir un índice se resuelve este problema.

No obstante, suele suceder que nos topemos con un concepto que no admite que se le asocie otra medición que no sea categórica; en estas condiciones, la construcción de un índice se vuelve imposible. Consideremos, por ejemplo, el concepto de estructura económica de una ciudad o de una región. Aunque definamos la estructura económica como la repartición del empleo entre las ramas de actividad, nunca podremos asociar a este concepto otra medición que no sea una clasificación (variable categórica): ciudad monoindustrial, ciudad de servicio, etc. Pero ¿cómo llegamos a construir

una clasificación que permita captar bien la realidad? Una manera de proceder consiste en comparar los objetos (en este caso, las estructuras económicas observadas) para constituir grupos de objetos lo suficientemente parecidos entre ellos, y claramente diferente de los objetos de los demás grupos. Una clasificación de este tipo puede, luego, servir de base para la elaboración de una tipología y para la definición de una variable categórica asociada al concepto.

Aprovechamos para mencionar que, aunque la construcción de un índice sea posible en principio, el proceso que acabamos de evocar puede ser de gran utilidad en caso de que no se pueda construir un índice que llene todas las expectativas del plan teórico. ¿Por ejemplo, al constituir una tipología de los países, podremos estudiar el desarrollo humano? Este tipo de tipología permitiría definir un índice apropiado para cada tipo de país con el objetivo de comparar únicamente países comparables con mediciones adaptadas a las características de estos países (es lo que ya efectúa el PNUD en relación con la medición de la pobreza: calcula dos "índices de la pobreza humana", uno para los países en vía de desarrollo y otro para los países desarrollados).

Formalizar el concepto de similitud y asociarle una medición no puede más que facilitar el proceso de clasificar objetos por tipos. De por sí existen procedimientos de clasificación automática basados en las mediciones de similitud[61]. Además, desearíamos, a veces, sólo tomar en cuenta un proceso heurístico con menos formalidades y examinar el grado de similitud entre los objetos sin tener que construir una tipología. Nuevamente, en este caso, la medición de la similitud puede convertirse en una herramienta de gran provecho. Vamos a hablar, por consiguiente, de la medición de similitud en esta parte.

[61] Dendrogramas, algoritmos de partición automáticos, etc. Vea Legendre y Legendre (1984 y 1998).

Para empezar, observemos que el concepto de similitud se aplica a un par de objetos. La similitud no es, por lo tanto, una propiedad de alguno de los dos objetos: es una propiedad del par.[62] Luego, el concepto de similitud es un concepto general que encierra en sí, una mirada de conceptos específicos; de hecho, examinar la similitud entre dos objetos es siempre *en relación con* un atributo en particular. Al momento de querer medir la similitud entre dos objetos, se define un concepto de similitud específico por el atributo que se selecciona para comparar estos objetos. En el caso de ciudades, por ejemplo, podemos considerar la similitud en relación con la estructura demográfica, con la tasa de criminalidad, con la calidad de vida, etcétera.

De entrada estamos de acuerdo para decir que la medición de similitud con relación a un atributo unidimensional es algo trivial puesto que no representa un gran problema para medir como, por ejemplo, la similitud entre dos países en cuanto al número de habitantes, a la tasa de criminalidad o al valor del IDH del PNUD.[63] Por el contrario, en caso de querer medir la similitud con relación a una propiedad multidimensional que no se resumió en un índice con anterioridad,[64] nos enfrentamos al mismo problema de construir un número índice. Por ejemplo:

> ¿Con relación a su estructura económica, cuál es el grado de similitud entre Quebec y Ontario?

[62] Se podría decir que el objeto al cual se aplica la similitud es un par de objetos.

[63] Este ejemplo es deliberadamente paradójico: si bien, el IDH es un indicador que permite medir una realidad multidimensional, la comparación de dos países en relación con el valor de este índice es, por su lado, unidimensional.

[64] O bien, y es exactamente lo mismo, medir al mismo tiempo la similitud bajo condiciones diferentes o, dicho de otra manera, medir la similitud entre dos objetos multidimensionales.

¿Con relación a su repartición en el territorio, cuál es el grado de similitud entre el cultivo de plátanos y la ganadería en Costa Rica?

Para medir la similitud en estos dos ejemplos, tenemos que tomar en cuenta más de una dimensión puesto que examinamos la similitud con relación a un concepto que contiene más de una dimensión.

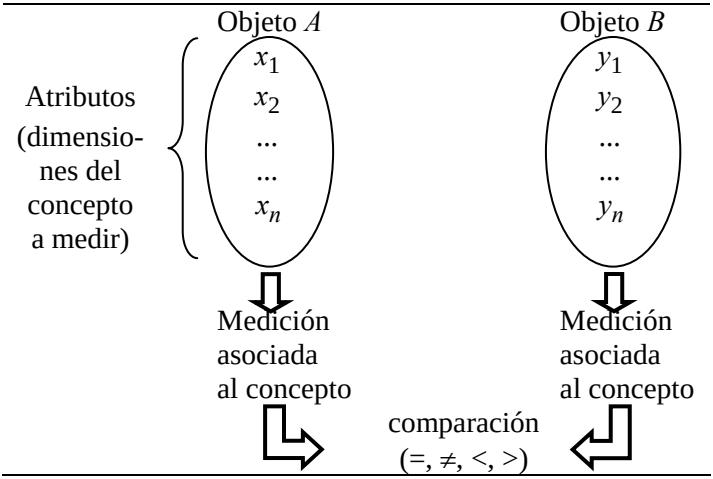
- En el caso de la similitud entre países en cuanto a su estructura económica, tenemos que considerar las diferentes ramas de producción.

En el caso de la similitud entre actividades en cuanto a la repartición espacial, tenemos que considerar las diferentes partes del territorio (zonas, distritos, provincias u otros en función del tipo de recorte geográfico que se usa).

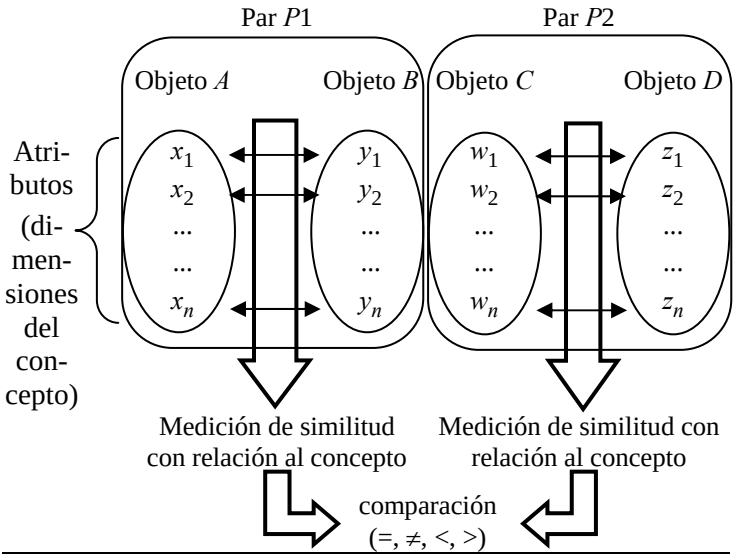
Así, las mediciones de la similitud entre objetos multidimensionales se parecen mucho a índices. A continuación, con el fin de resaltar las diferencias, vamos a centrar nuestro estudio en destacar las partes específicas de la medición de la similitud.

Como ya sabemos, un índice resume en una sola cifra los valores de los indicadores asociados a las múltiples dimensiones de un concepto. El índice es una medición porque permite comparar dos objetos en cuanto al grado que poseen de la propiedad definida por el concepto. Esto se resume con el diagrama siguiente.

Construcción de un número índice



En cambio, para medir la similitud, empezamos por comparar dos objetos detalle por detalle hasta obtener cuantas mediciones parciales de similitud como haya dimensiones que comparar. Es necesario, después, agregar todas estas mediciones parciales en una sola, lo que da por resultado una medición de la similitud. Esta medición permite comparar dos pares de objetos en cuanto a su similitud con relación a un atributo multidimensonal dado. Se resume esto en el diagrama que sigue, lo que al compararlo con el otro diagrama, permite entender las diferencias que existen entre la medición de la similitud entre objetos multidimensionales y la construcción de un índice.



De todas maneras, se trata efectivamente de una medición tal y como se definió en el capítulo 1-1. Recordemos que una medición asociada a un concepto establece una correspondencia entre los objetos y los números lo que permite comparar los objetos y determinar el valor de verdad de una o varias relaciones =, ≠, <, >. Por lo tanto, una medición de la similitud es una correspondencia que permite comparar dos pares de objetos cualesquiera en cuanto a su similitud con relación a un atributo dado. De manera formal, si convenimos que $f(A,B)$ es la medición de la similitud entre los objetos del par $[A,B]$ y que $f(C,D)$ es la medición de la similitud entre los objetos del par $[C,D]$, entonces, una medición de la similitud permite decidir el valor de verdad de una o varias de las relaciones siguientes:

$$f(A,B) = f(C,D)$$
$$f(A,B) \neq f(C,D)$$
$$f(A,B) < f(C,D)$$
$$f(A,B) > f(C,D)$$

Por ejemplo, si *A* es Nicaragua, *B* es Costa Rica, *C* es Costa Rica y *D* es Canadá,[65] una medición de la similitud permite contestar a la pregunta: "¿Con relación a la composición de su producción, Costa Rica se parece más a Nicaragua o a Canadá?" De igual manera, si *A* es el cultivo de plátano, *B* es la ganadería, *C* es el cultivo de plátano y *D* es el cultivo de cítricos, una medición de la similitud permite contestar a la pregunta: "¿Con relación a su repartición geográfica en Costa Rica, el cultivo de plátanos se parece más a la ganadería o al cultivo de cítricos?".

Notemos que nada de lo anterior mencionado implica que siempre tengamos que medir basándonos en una escala racional aunque las variables que se usan como medición de la similitud o de la disimilitud sean, a menudo, variables racionales. No obstante, el problema de la multidimensionalidad provoca que, por lo común, haya varias mediciones posibles sin que ninguna pueda, a priori, calificarse como la mejor. Es la razón por la cual, excepto en casos particulares, que es preferible interpretar las mediciones de similitud como mediciones ordinales y evitar interpretarlas de manera abusiva como mediciones de intervalo o racionales.

Además, y como lo veremos, las mediciones de similitud son, a menudo, mediciones inversas, es decir que son más bien mediciones de disimilitud. Debemos estar muy atentos a este aspecto que es causa de mucha confusión.

[65] Tal como nos lo muestra este ejemplo, puede suceder que $B = C$ (o $B = D$, o $A = C$, o $A = D$), sin embargo no sucede necesariamente.

1-5.1.2 La medición de la similitud entre distribuciones

Una distribución o una repartición es una propiedad (multidimensonal) de una población (en la aceptación general de una colección de personas u objetos) cuando se clasifica esta población en categorías: es el número de individuos o la fracción de la población que se encuentra en cada una de las categorías. En los ejemplos ya mencionados:

> Las personas empleadas en una economía constituyen una "población" que se puede clasificar en "categorías" como las ramas de actividad. Se puede describir la estructura económica de un país con una distribución que representa el número de personas empleadas por rama de actividad.
>
> Las hectáreas de un terreno que se dedican a una actividad dada (el cultivo de plátanos, por ejemplo) constituyen una "población" que se puede clasificar en las "categorías" como las subdivisiones (provincias u otras) de un territorio. Se pude describir la repartición espacial de las actividades con una distribución que representa el número de hectáreas que se dedican a la actividad en cada subdivisión del territorio.

Por consiguiente una distribución es un objeto multidimensional. No obstante, la comparación entre dos distribuciones no es tan complicada gracias a la existencia de una "regla de normalización" natural que marca lo siguiente: la medición asociada a cada una de las dimensiones de la distribución es simplemente la fracción de la población que pertenece a la categoría correspondiente. Ahora bien, en una distribución, la suma de las partes es necesariamente igual a 1. De entrada, esto elimina parte del problema de la multidimensionalidad que era, como lo vimos, el problema de los números índice y del peso que se le atribuye a cada una de las dimensiones.

Por el contrario, cuando intentamos comparar dos objetos que no sean distribuciones, la selección de la unidad de medición de cada dimensión de la comparación determina de manera implícita cuál será su peso en la medición de la disimilitud. Es entonces, en estas condiciones, que se intensifica el problema de la multidimensionalidad ya mencionado con relación a los números índice.

1-5.1.3 Disimilitud y desigualdad-concentración: ¿cuál es la diferencia?

En los ejemplos mencionados hasta el momento, sólo buscábamos examinar el grado de asociación entre dos fenómenos o a la inversa, el grado de segregación entre ellos. Sin embargo, existe otro uso de las mediciones de disimilitud entre dos distribuciones; este otro uso es la medición de la concentración o de la dispersión. Una medición de disimilitud se convierte en una medición de concentración cuando se compara la distribución estudiada con una distribución de referencia o teórica. Esta distribución teórica, que sirve de referencia, representa una concentración nula y sirve, de alguna manera, de patrón de medición (exhibiremos un ejemplo de esto más abajo).

Esto es del todo coherente con lo estudiado en el capítulo 1-4 puesto que, en general, una medición de la desigualdad compara la distribución observada con una distribución de referencia, la cual representa la igualdad perfecta, Por lo tanto, una medición de la desigualdad es una medición de disimilitud entre una distribución observada y una distribución de referencia.

De ahí que el índice de Gini resulte ser de igual manera adecuado como medición de disimilitud que como medición de desigualdad. De por sí, se mencionó con anterioridad que el índice de Gini, entre otras propiedades, era simétrico, es decir que los papeles de la distribución observada y de la dis-

tribución de referencia son intercambiables; en otras palabras, al intercambiar los papeles, el valor del coeficiente de Gini no cambia.

1-5.2 EL ÍNDICE DE DISIMILITUD

1-5.2.1 Un ejemplo numérico

Ahora, nos interesamos en una medición de disimilitud ampliamente usada, la cual se aplica a las distribuciones como, por ejemplo, la repartición geográfica del empleo. Mostramos un ejemplo numérico ficticio:

Empleo por zona y por rama

Ramo	*B1*	*B2*	*B3*	Total
Zona				
Z1	48	325	287	660
Z2	27	185	148	360
Z3	45	90	45	180
Total	120	600	480	1200

Se trata de una medición de similitud entre las ramas de actividades en cuanto a su repartición geográfica. Por lo tanto, nos interesa conocer la fracción del empleo de cada rama en cada zona:

Distribución del empleo entre zonas

Ramo	*B1*	*B2*	*B3*	Total
Zona				
Z1	0.400	0.542	0.598	0.550
Z2	0.225	0.308	0.308	0.300
Z3	0.375	0.150	0.094	0.150
Total	1.000	1.000	1.000	1.000

La solución más simple que podemos imaginar para examinar la similitud entre dos distribuciones consiste en obser-

var las diferencias entre estas fracciones zona por zona. Hagamos esta comparación entre las ramas *B1* y *B2*:

Comparación de la repartición geográfica de las ramas B1 y B2

Rama	*B1*	*B2*	Diferencia
Zona			
Z1	0.400	0.542	0.142
Z2	0.225	0.308	0.083
Z3	0.375	0.150	–0.225
Total	1.000	1.000	0.000

Cada una de las diferencias calculadas constituye una de las dimensiones de la disimilitud entre las dos reparticiones geográficas. Para medir la disimilitud, es necesario combinar las diferencias en una cifra única. Una simple suma será siempre igual a 0 por razones obvias.[66] Es el motivo por el cual haremos una suma de los valores absolutos:

$$|0.142| + |0.083| + |-0.225| = 0.142 + 0.083 + 0.225$$

(y no – 0.225)

Por razones que más adelante nos parecerán evidentes, dividimos el resultado entre dos y obtenemos así:

$$\frac{|0.142| + |0.083| + |-0.225|}{2} = 0.225$$

1-5.2.2 Definición del índice de disimilitud

La medición de la disimilitud y una tabla de contingencia: recuerdo de la simbología

Con el fin de formalizar la presentación, usaremos nueva-

[66] Puesto que $\sum_i v_i = \sum_i w_i = 1$, entonces $\sum_i (v_i - w_i) = \sum_i v_i - \sum_i w_i = 0$.

mente y esta vez generalizándola, la simbología que se manejó en el apartado 1-2.1.[67] Analizamos una tabla de contingencia de dos dimensiones. Convenimos que la líneas corresponden a *n* grupos diferentes mientras que las columnas corresponden a *m* categorías diferentes (en nuestro ejemplo como en el ejemplo del apartado 1-2.1, los *n* grupos son las tres ramas de actividad mientras que las *m* "categorías" son las tres zonas).

x_{ij}	Número de empleos de la rama *j* en la zona *i*
$x_{\bullet j} = \sum_i x_{ij}$	Número total de empleos de la rama *j*
$x_{i\bullet} = \sum_j x_{ij}$	Número total de empleos en la zona *i*
$x_{\bullet\bullet} = \sum_i \sum_i x_{ij}$	Número total de empleos de todas las ramas en todas las zonas
$p_{ij} = x_{ij} / x_{\bullet\bullet}$	Fracción del empleo total global que pertenece a la rama *j* y situado en la zona *i*
$p_{\bullet j} = \sum_i p_{ij}$	Fracción del empleo total global que pertenece a la rama *j*
$p_{i\bullet} = \sum_j p_{ij}$	Fracción del empleo total global situado en la zona *i*
$p_{j/i\bullet} = p_{ij} / p_{i\bullet}$	Fracción del empleo total de la zona *i* que pertenece a la rama *j*
$p_{i/\bullet j} = p_{ij} / p_{\bullet j}$	Fracción del empleo total de la rama *j* situado en la zona *i*

[67] Se invita al lector a referirse al apartado 1-2.1 para tener presente el enunciado de la identidades que se verifican en una tabla de contingencia.

En el ejemplo numérico arriba mencionado, aplicamos una medición de disimilitud entre dos divisiones geográficas, la primera de la rama B1 y la segunda de la B2. Según la simbología comúnmente usada, esto vuelve a aplicar una medición de disimilitud a las distribuciones que se representan por los vectores

$$Q_1 = \begin{bmatrix} p_{1/\bullet 1} \\ p_{2/\bullet 1} \\ \vdots \\ p_{m/\bullet 1} \end{bmatrix} \text{ y } Q_2 = \begin{bmatrix} p_{1/\bullet 2} \\ p_{2/\bullet 2} \\ \vdots \\ p_{m/\bullet 2} \end{bmatrix}$$

En un aspecto más general, comparamos las distribuciones

$$Q_h = \begin{bmatrix} p_{1/\bullet h} \\ p_{2/\bullet h} \\ \vdots \\ p_{m/\bullet h} \end{bmatrix} \text{ y } Q_k = \begin{bmatrix} p_{1/\bullet k} \\ p_{2/\bullet k} \\ \vdots \\ p_{m/\bullet k} \end{bmatrix}$$

o bien, las distribuciones

$$R_g = \begin{bmatrix} p_{1/g\bullet} & p_{2/g\bullet} & \cdots & p_{n/g\bullet} \end{bmatrix} \text{ y } R_i = \begin{bmatrix} p_{1/i\bullet} & p_{2/i\bullet} & \cdots & p_{n/i\bullet} \end{bmatrix}$$

Nota: Podemos trabajar tanto con fracciones como en la simbología arriba empleada como con porcentajes que obtenemos al multiplicar las fracciones por 100. En este momento, convenimos trabajar con fracciones para simplificar la escritura de las fórmulas. Sin embargo, en la práctica, con tal de simplificar las tablas al eliminar el punto de las decimales, se acostumbra presentar porcentajes.

Definición

A continuación, aplicamos el índice de disimilitud a una comparación de las distribuciones Q_h y Q_k. Se puede trasponer fácilmente todos los cálculos a una comparación entre las distribuciones R_g y R_i o, de hecho, a cualquier par de distribuciones que se pueda comparar de manera formal (es decir que tengan el mismo número de posibilidades).

Se define el índice de disimilitud como:

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

En el ejemplo numérico mencionado arriba,

$$Q_1 = \begin{bmatrix} 0.400 \\ 0.225 \\ 0.375 \end{bmatrix} \text{ y } Q_2 = \begin{bmatrix} 0.542 \\ 0.308 \\ 0.150 \end{bmatrix}$$

y

$$D = \frac{\left|0.400 - 0.542\right| + \left|0.225 - 0.308\right| + \left|0.375 - 0.150\right|}{2}$$

$$D = 0.225$$

Este índice de disimilitud y sus variantes cercanas se conocen con nombres diferentes según la disciplina que se trate. Por ejemplo, se conoce también el índice de disimilitud con las expresiones “índice de diferenciación” e “índice de disociación”.

Cuando una de las distribuciones es la repartición espacial de una actividad económica y la otra es el grupo de actividades, esta medición corresponde a lo que se conoce en ciencias regionales como el coeficiente de localización. No obstante, debemos hacer mención de que el coeficiente de localización no es exactamente un índice de disimilitud aunque se calculó con la misma fórmula; examinaremos el porqué más adelante.

En ciencias regionales, se usa también el coeficiente de especialización que compara la estructura económica de una zona (repartición del empleo entre las ramas de actividad) con la estructura del territorio completo que se estudia. Tampoco este índice es exactamente un índice de disimilitud.

En geografía, Taylor (1997, p. 180) nombra de varias maneras el índice de disimilitud; entre otros nombres, el de "coeficiente de asociación geográfica" nos parece particularmente inadecuado puesto que D es una medición de disimilitud o de disociación. Es hasta posible encontrar el término "índice de Gini" para designar el índice de disimilitud.

Los demógrafos y los sociólogos usan este mismo índice con el nombre "coeficiente de segregación residencial" o "índice de discriminación" para comparar las distribuciones espaciales residenciales de diferentes grupos étnicos o raciales (Mills y Hamilton, 1989, pp. 233-239; Waldorf, 1993).

¿Qué debemos entender de esta confusión de términos? Lo siguiente: al momento de enterarse de resultados de investigación que requieren el uso de índices de este tipo, verifique muy bien la fórmula matemática empleada.

Más allá de estas particularidades propias de cada disciplina, examinemos este índice de disimilitud como una medición de disimilitud de dos distribuciones.

1-5.2.3 El índice de disimilitud como medición de concentración o desigualdad

Hasta el momento hemos estudiado los diferentes usos del índice de disimilitud para medir la disimilitud entre dos distribuciones observadas. Pero podemos también usar el índice de disimilitud para medir la desigualdad o la concentración. De por sí y es importante recordarlo, las mediciones de desigualdad o de concentración son, en general, mediciones de disimilitud entre una distribución observada y una distribu-

ción de referencia. Para medir la desigualdad o la concentración, es necesario, por lo tanto, comparar una distribución observada con una distribución de referencia que representa la igualdad perfecta o una concentración nula (está claro que, en este caso, la tabla de datos no es una tabla de contingencia).

Ejemplo

Supongamos que queremos medir el grado de concentración geográfica de la población en un territorio dado que hubiéramos, con anterioridad, dividido en zonas (estados, provincias, distritos...). Una concentración nula corresponde a una situación donde la densidad de la población (habitantes / km^2) es, en todas partes, igual. De modo que podemos decir que la concentración es nula si la fracción de la población en cada zona es igual a la fracción del territorio contenida en esta zona.

Teniendo V, la distribución de la superficie del territorio y W, la distribución de la población,

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \text{ y } W = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}$$

v_i es la fracción de la superficie total que contiene la zona i y w_i es la fracción de la población que se encuentra en la zona i.

La concentración es nula si $w_i = v_i$ para todo i.

En este caso, la distribución observada del territorio sirve de distribución de referencia para la población: es la distribución teórica o hipotética de una población con una concentración nula.[68] Podemos, entonces, usar el índice de disimilitud entre la distribución del territorio y la distribución de la po-

[68] En otras palabras, la distribución V es una distribución observada cuando se trata del territorio pero se convierte en una distribución hipotética cuando la aplicamos a la población.

blación como medición de la concentración geográfica de la población. Tenemos:

$$D = \frac{1}{2}\sum_i |w_i - v_i|$$

La tabla que presentamos a continuación ilustra esta situación del índice de disimilitud. En ella se mide el grado de concentración de la población de la ciudad de Montreal. Se extrajeron los datos de población del censo de 1991. El territorio es dividido según los 54 barrios de planificación de la ciudad ordenados de manera decreciente de densidad. Se obtiene D = 0.2361, es decir que, para obtener una densidad uniforme, se tendría que desplazar 23.61% de la población de un barrio a otro.

Medición de la concentración de la población
por medio del índice de disimilitud
Ciudad de Montreal (54 barros de planificación),
población del censo de 1991

	Datos			Reparticiones		
Barrio	Pob. 1991	Superf. (km^2)	Densidad (hab/km^2)	Pob.	Superf.	Diferencia absoluta
11	29469	1.65	17860	2.90%	0.88%	0.0201
8	10604	0.72	14728	1.04%	0.38%	0.0066
18	27022	2.03	13311	2.66%	1.08%	0.0157
34	24258	1.85	13112	2.38%	0.99%	0.0140
13	30314	2.39	12684	2.98%	1.28%	0.0170
35	14187	1.24	11441	1.39%	0.66%	0.0073
31	19652	1.73	11360	1.93%	0.92%	0.0101
33	15752	1.40	11251	1.55%	0.75%	0.0080
42	25495	2.32	10989	2.51%	1.24%	0.0127
15	19126	1.75	10929	1.88%	0.93%	0.0095
16	15030	1.38	10891	1.48%	0.74%	0.0074
29	15606	1.46	10689	1.53%	0.78%	0.0075
9	21348	2.02	10568	2.10%	1.08%	0.0102
32	14737	1.48	9957	1.45%	0.79%	0.0066
40	20350	2.15	9465	2.00%	1.15%	0.0085
14	15973	1.80	8874	1.57%	0.96%	0.0061
10	14165	1.65	8585	1.39%	0.88%	0.0051
27	11592	1.41	8221	1.14%	0.75%	0.0039
17	16167	2.00	8084	1.59%	1.07%	0.0052
30	29664	3.69	8039	2.91%	1.97%	0.0095

Medición de la concentración de la población
por medio del índice de disimilitud (continuación)

Barrio	Datos			Reparticiones		
	Pob. 1991	Superf. (km^2)	Densidad (hab/km^2)	Pob.	Superf.	Diferencia absoluta
45	24738	3.23	7659	2.43%	1.72%	0.0071
46	19880	2.60	7646	1.95%	1.39%	0.0057
39	34906	4.85	7197	3.43%	2.59%	0.0084
51	8452	1.20	7043	0.83%	0.64%	0.0019
23	18672	2.67	6993	1.83%	1.43%	0.0041
12	14980	2.21	6778	1.47%	1.18%	0.0029
6	16785	2.48	6768	1.65%	1.32%	0.0033
19	11499	1.75	6571	1.13%	0.93%	0.0020
4	23636	3.70	6388	2.32%	1.98%	0.0035
44	18699	2.96	6317	1.84%	1.58%	0.0026
24	13665	2.22	6155	1.34%	1.19%	0.0016
21	20564	3.62	5681	2.02%	1.93%	0.0009
48	17038	3.02	5642	1.67%	1.61%	0.0006
41	20092	3.59	5597	1.97%	1.92%	0.0006
5	18478	3.36	5499	1.82%	1.79%	0.0002
49	14687	2.73	5380	1.44%	1.46%	0.0001
20	27819	5.22	5329	2.73%	2.79%	0.0005
43	24957	4.84	5156	2.45%	2.58%	0.0013
3	18052	3.56	5071	1.77%	1.90%	0.0013
28	17764	3.56	4990	1.75%	1.90%	0.0015
2	25181	5.25	4796	2.47%	2.80%	0.0033
26	19073	4.01	4756	1.87%	2.14%	0.0027
22	9651	2.18	4427	0.95%	1.16%	0.0022
38	12512	3.16	3959	1.23%	1.69%	0.0046
7	22660	5.84	3880	2.23%	3.12%	0.0089
1	22613	5.85	3865	2.22%	3.12%	0.0090
52	35098	9.50	3695	3.45%	5.07%	0.0162
50	14403	4.07	3539	1.42%	2.17%	0.0076
47	13111	4.45	2946	1.29%	2.38%	0.0109
54	47534	19.04	2497	4.67%	10.16%	0.0549
37	3546	2.06	1721	0.35%	1.10%	0.0075
25	4009	4.28	937	0.39%	2.28%	0.0189
53	11970	13.92	860	1.18%	7.43%	0.0625
36	431	4.24	102	0.04%	2.26%	0.0222
Total	*1017666*	*187.34*	*5432*	*100.00%*	*100.00%*	*0.4720*

Índice de disimilitud: 0.2361

1-5.2.4 Propiedades del índice de disimilitud

El índice de disimilitud y las propiedades de una medición de desigualdad

Puesto que una medición de desigualdad es una medición de disimilitud entre una distribución observada y una distribución de referencia, las propiedades deseables para una medición de desigualdad son de igual manera deseables para una medición de disimilitud. ¿En qué resulta esto con el índice de disimilitud *D*?

Recordemos las propiedades deseables de una medición de desigualdad según Valeyre (1993):

1. Una medición de desigualdad no puede tomar valores negativos ya que es una medición del alejamiento de la distribución observada en relación con la distribución de referencia.
2. Una medición de desigualdad debe tomar el valor 0 si y solamente si la distribución observada es idéntica a la distribución de referencia.
3. Se deben tratar a todas las observaciones de la misma manera.
4. Una medición de desigualdad debe ser independiente del valor promedio de la variable examinada; una medición de concentración debe ser independiente del tamaño de la población cuya distribución se estudia.
5. La agregación de observaciones que tienen el mismo grado de especificidad, no debe cambiar el valor de la medicion.[69]
6. Principio de transferencia de Pigou-Dalton: una medición de desigualdad debe disminuir si se modifica la

[69] Se puede demostrar fácilmente esta característica usando la interpretación geométrica del índice de disimilitud como la distancia vertical máxima entre la curva de Lorenz y la diagonal. Vea más abajo.

distribución de tal manera que reduce sin duda alguna la desigualdad.

El índice de disimilitud posee las cinco primeras propiedades pero no posee la sexta: su valor no cambia después de una transferencia entre dos categorías cuyas especificidades son ambas superiores o ambas inferiores a 1.

Campo de variación

Si alguien le anunciase que obtuvo una calificación de 18 en un examen, ¿estaría usted contento? ¿Es esta calificación una buena o una mala calificación? Para saberlo es necesario, ante todo, conocer la calificación máxima.[70] Si el examen se califica sobre 20, un 18 es, de seguro, una buena calificación; por el contrario, si el examen se califica sobre 100, es muy probable que usted no esté del todo contento.

Ésta es la razón por la cual el campo de variación nos interesa. El campo de variación de una medición es el conjunto de los valores que esta medición puede tomar. En el caso de una medición continua, se define el campo de variación por el valor mínimo y el valor máximo de la medición. Para saber si una calificación dada es "alta" o no, es necesario, por lo menos, conocer su campo de variación y verificar si este valor se acerca más el máximo o el mínimo.

En el caso del índice de disimilitud, su valor mínimo es 0; este índice toma el valor 0 cuando $p_{i/\bullet h} = p_{i/\bullet k}$ para todo i, o sea cuando las distribuciones son idénticas.

¿Cuál es su valor máximo?

[70] No es la única consideración. La interpretación de la calificación depende también de la calificación que los demás obtuvieron y de los criterios que se usan para su interpretación (como la calificación de aprobación).

Cuando comparamos las distribuciones de dos poblaciones perfectamente distintas,[71] el valor máximo que puede tomar el índice es 1: esto se produce cuando $p_{i/\bullet h} = 0$ y $p_{i/\bullet k} > 0$ y viceversa, es decir cuando la separación entre las dos poblaciones es completa, lo que significa que nunca aparecen juntas en la misma categoría. En efecto, en esta situación, para la categoría *i*, tenemos:

Teniendo $p_{i/\bullet h} = 0$ entonces:

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \left| 0 - p_{i/\bullet k} \right| = p_{i/\bullet k}$$

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| 0 + p_{i/\bullet k} = p_{i/\bullet h} + p_{i/\bullet k}$$

Teniendo $p_{i/\bullet k} = 0$, entonces:

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \left| p_{i/\bullet h} - 0 \right| = p_{i/\bullet h}$$

$$\left| p_{i/\bullet h} - p_{i/\bullet k} \right| = p_{i/\bullet h} + 0 = p_{i/\bullet h} + p_{i/\bullet k}$$

Por lo tanto tenemos:

$$D^{\max} = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2} \sum_i \left(p_{i/\bullet h} + p_{i/\bullet k} \right)$$

$$D^{\max} = \frac{1}{2} \left(\sum_i p_{i/\bullet h} + \sum_i p_{i/\bullet k} \right) = \frac{1+1}{2} = 1$$

La división entre 2 en la fórmula del cálculo del índice de disimilitud permite, por consiguiente, normalizar su campo de variación en el intervalo [0, 1].

¿Es posible que el índice de disimilitud tome un valor superior a 1? No. Para convencerse, basta preguntarse, refiriéndose a la situación de separación completa que se describió anteriormente, cuál sería la consecuencia de desplazar un individuo de una categoría a otra (el efecto es nulo si el indivi-

[71] Esto significa que ningún individuo puede pertenecer a dos poblaciones al mismo tiempo.

duo se queda con los mismos de su especie y si no, el valor del indicador disminuye)

Índice de disimilitud: ejemplo de segregación total

Etnia	Números			Reparticiones			
	Marcianos	Terrícolas	Total	Martianos	Terrícolas	Total	Diferencia
	x_{i1}	x_{i2}	$x_{i1}+x_{i2}$	$p_{i/\bullet 1}$	$p_{i/\bullet 2}$	$p_{i\bullet}$	$\lvert p_{i/\bullet h} - p_{i/\bullet k} \rvert$
Planeta							
Tierra	0	6	6	0.00	0.75	0.40	0.75
Luna	0	2	2	0.00	0.25	0.13	0.25
Marte	3	0	3	0.43	0.00	0.20	0.43
Júpiter	4	0	4	0.57	0.00	0.27	0.57
Total	7	8	15	1.00	1.00	1.00	

Índice de disimilitud:

$$\frac{0.75 + 0.25 + 0.43 + 0.57}{2} = 1.00$$

Interpretación metafórica

Aunque conozcamos perfectamente el campo de variación de una medición, es a veces difícil intuir con claridad lo que es un valor "alto", motivo por el cual puede ser útil una interpretación metafórica. Como su nombre lo indica, una interpretación metafórica se basa en una comparación, una metáfora del tipo "es como si", Cuidado en no tomar esta metáfora al pie de la letra.

En el caso del índice de disimilitud éste compara la distribución de dos grupos perfectamente distintos,[72] digamos *h* y *k*. Se puede interpretar el índice como una fracción del grupo

[72] Esto significa que ningún individuo puede pertenecer a dos poblaciones al mismo tiempo.

h que tendríamos que desplazar de una categoría a otra para que su distribución quedara idéntica a la distribución del grupo *k*.

En el ejemplo numérico que mencionamos al principio de este apartado, el índice de disimilitud entre la repartición espacial de los empleos de la rama *B1* y los empleos de la rama *B2* es de 0.225. Esto significa que, con tal de que las reparticiones espaciales *B1* y *B2* sean idénticas, se tendría que desplazar 22.5% de los empleos de *B1*.

Se puede demostrar fácilmente este resultado. Para empezar, determinamos cuál es la fracción del grupo *h* que tendríamos que desplazar para pasar de la distribución representada por los $p_{i/\bullet h}$ a la distribución representada por los $p_{i/\bullet k}$. Para lograr esto, sólo basta sumar las fracciones de población que debe quitarse de las categorías (zonas, regiones,...) "excedentes" para redistribuirlas en las categorías "deficitarias". Designemos por A, el conjunto de las categorías "excedentes", o sea cuando $p_{i/\bullet h} > p_{i/\bullet k}$. Para cada una de las categorías que pertenece al conjunto *A*, la fracción "excedente" de la población *H* es igual a $p_{i/\bullet h} - p_{i/\bullet k}$. La suma total de la fracción de la población *h* que debe quitarse de las categorías "excedentes" es, por lo tanto:

$$\sum_{i \in A} \left(p_{i/\bullet h} - p_{i/\bullet k} \right)$$

Es equivalente querer sumar las fracciones de población que debe añadirse en la categorías "deficitarias", es decir:

$$\sum_{i \notin A} \left(p_{i/\bullet k} - p_{i/\bullet h} \right)$$

Está claro que

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)=\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

Puesto que

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)-\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

$$=\sum_{i\in A}p_{i/\bullet h}-\sum_{i\in A}p_{i/\bullet k}-\sum_{i\notin A}p_{i/\bullet k}+\sum_{i\notin A}p_{i/\bullet h}$$

$$=\sum_{i}p_{i/\bullet h}-\sum_{i}p_{i/\bullet k}=0$$

¿Cuál es la relación con el índice de disimilitud? Bueno, si sumamos las dos sumatorias del miembro a la izquierda de la ecuación anterior (en lugar de restar la segunda de la primera), obtenemos

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)+\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)=\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|$$

Además, puesto que los dos términos de miembro de derecha son iguales, tenemos:

$$\sum_{i\in A}\left(p_{i/\bullet h}-p_{i/\bullet k}\right)=\sum_{i\notin A}\left(p_{i/\bullet k}-p_{i/\bullet h}\right)$$

$$=\frac{1}{2}\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|=D$$

Ahí está otra buena razón por dividir la suma entre 2.

Simetría

Es importante notar que el índice de disimilitud es simétrico con relación a los grupos *h* y *k*:

$$D=\frac{1}{2}\sum_{i}\left|p_{i/\bullet h}-p_{i/\bullet k}\right|=\frac{1}{2}\sum_{i}\left|p_{i/\bullet k}-p_{i/\bullet h}\right|$$

Por consiguiente, se puede, de igual manera, interpretar el indicador como la fracción del grupo *k* que debería desplazarse para que su distribución fuese idéntica a la distribución del grupo *h*. Importando el grupo que querramos desplazar para que su distribución sea idéntica a la distribución de otro grupo, la fracción que debe desplazarse es la misma (de esta manera, podríamos decir que se necesita desplazar 22.5% del empleo del ramo B2 para que su distribución sea idéntica a la distribución de la rama B1). No obstante, en cuanto al número de individuos que se necesita desplazar, éste es obviamente igual a esta fracción multiplicada por el número de la población. En caso de que dos grupos tengan tamaños diferentes, el número de individuos que se deba desplazar (de manera hipotética) difiere dependiendo de la fracción de uno u otro grupo que se quiere desplazar.

De nueva cuenta, no debemos olvidar el aspecto metafórico de esta interpretación. Para empezar, la similaridad de las distribuciones no es siempre buena (recordemos la controversia que suscitó el *busing* que se realizó en Estados Unidos para lograr la integración escolar de los blancos y los negros). Luego, el desplazamiento (a fuerza) de poblaciones no resulta ser una acción aceptable cuando se trata de poblaciones humanas.

Otras propiedades

El índice de disimilitud, tal como todos los índices, tiene sus límites. Además de no respectar el principio de Pigou-Dalton, mencionamos:

- Cuando los datos son agrupados, el índice de disimilitud, tal como el de Gini, es sensible a la definición y al número de las categorías utilizadas (clases, zonas). Esa debilidad no es tan grave si se escoge una clasificación bastante fina –o sea si comprende un gran número de

categorías– pero las comparaciones entre clasificaciones diversas no tienen ninguna significación.[73]

- Cuando se utiliza como medición de concentración espacial, el índice de disimilitud, tal como el de Gini, no toma en cuenta la contigüidad o la proximidad de las unidades espaciales.
- El índice de disimilitud no admite datos negativos. Por ejemplo, no se puede utilizar el índice de disimilitud para medir la similitud entre dos ramas de actividad respecto a las *variaciones* del número de empleos por zona, porque esas variaciones pueden ser negativas.

El índice de disimilitud y la curva de Lorenz

Acabamos de ver que, como el índice de Gini, el índice de disimilitud puede servir para medir la concentración aunque no posea todas las propiedades deseables del índice de Gini (le falta el principio de transferencia de Pigou-Dalton). Vimos, también, que se puede calcular el índice de Gini de manera geométrica basándonos en la curva de Lorenz. ¿Existe, por lo tanto, una relación entre el índice de disimilitud y la curva de Lorenz? ¡Pues sí!

En efecto, el índice de disimilitud es justamente igual a la diferencia vertical entre la curva de Lorenz y la diagonal

$$D = \underset{k}{\text{MAX}}\left[Cv_k - Cw_k\right]$$

Demostración:

Puesto que

$$\sum_i \left(v_i - w_i\right) = \sum_i v_i - \sum_i w_i = 1 - 1 = 0\,,$$

[73] Eso tema es conocido en geografía como "MAUP", es decir "Modifiable Areal Unit Problem".

esta suma contiene términos positivos y términos negativos (al menos que todos los términos sean igual a 0). Sin embargo, al ordenar las observaciones en un orden creciente de las razones w_i / v_i, los términos $(v_i - w_i)$ que son positivos preceden los términos negativos. En estas condiciones, está claro que la diferencia vertical

$$Cv_k - Cw_k = \sum_{i-1}^{k} v_i - \sum_{i-1}^{k} w_i = \sum_{i-1}^{k} (v_i - w_i)$$

alcanza su valor máximo cuando se escoge k de tal manera que, únicamente los términos positivos se incluyen en la sumatoria, excluyendo así todos los términos negativos. Por lo tanto

$$\underset{k}{\text{MAX}}\left[Cv_k - Cw_k\right] = \sum_{\substack{i \text{ tal que} \\ v_i > w_i}} (v_i - w_i)$$

y, puesto que $\sum_i (v_i - w_i) = 0$

$$\sum_{\substack{i \text{ tal que} \\ v_i > w_i}} (v_i - w_i) = \sum_{\substack{i \text{ tal que} \\ v_i < w_i}} |v_i - w_i| = \frac{1}{2} \sum_i |v_i - w_i| = D$$

Se tiene, así, una interpretación geométrica para el índice de disimilitud D: es la distancia máxima entre la diagonal y la curva de Lorenz asociada a la distribución V (vea el ejemplo numérico extraído de Taylor, 1997, y analizado en 1-4.3).

Con la ayuda de esta interpretación, es fácil constatar que el índice de disimilitud es insensible a toda distribución que no reduce la diferencia vertical máxima pero que, sin embargo, acerca la curva de Lorenz a la diagonal. Es justamente esta insensibilidad la que viola el principio de transferencia de Pigou-Dalton.

Sumario de las propiedades del índice de disimilitud

1. Posee las 5 primeras propiedades deseables de una medición de desigualdad pero le falta la última (el principio de transferencia de Pigou-Dalton; Valeyre, 1993)
2. Campo de variación (valores máximo y mínimo)
   – $D = 0$ cuando $p_{i/\bullet h} = p_{i/\bullet k}$ para todo i (las dos distribuciones son idénticas)
   – $D = 1$ cuando hay segregación completa:
     teniendo $p_{i/\bullet k} > 0$, entonces $p_{i/\bullet h} = 0$
     teniendo $p_{i/\bullet h} > 0$, entonces $p_{i/\bullet k} = 0$
3. D es simétrico con relación a los grupos h y k:
   $$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet k} - p_{i/\bullet h} \right|$$
4. Interpretación metafórica (grupos perfectamente distintos)
   D = fracción del grupo h que convendría desplazar para que su distribución fuera idéntica a la distribución del grupo k o viceversa.
5. Cuando los datos son agrupados, tanto D como G son sensibles a la definición y al número de categorías (clases, zonas). En particular, esto implica que la agregación de una o varias categorías significa una disminución del valor del índice de disimilitud.
6. Como medición de concentración espacial y al igual que el índice de Gini, el índice de disimilitud no toma en cuenta la proximidad en el espacio de diferentes zonas con fuerte densidad.
7. No se aplica con datos negativos (ejemplo: comparación de variación del empleo).
8. D es igual a la máxima diferencia vertical entre la curva de Lorenz y la diagonal.

1-5.2.5 Aplicación de índice de disimilitud a una dicotomía

Equivalencia de la fórmula de Duncan-Duncan (1995)

Cuando distinguimos solamente dos grupos, decimos que tratamos con una dicotomía; en estas condiciones, se compara un grupo *h* con el resto de la población (que toma el papel del grupo *k*). Para el grupo *k* tenemos entonces:

$$p_{i/\bullet k} = \frac{x_{i\bullet} - x_{ih}}{x_{\bullet\bullet} - x_{\bullet h}} = \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}}$$

Así que se puede escribir el índice de disimilitud con la fórmula

$$D = \frac{\sum_i p_{i\bullet} \left| p_{h/i\bullet} - p_{\bullet h} \right|}{2\, p_{\bullet h} \left(1 - p_{\bullet h}\right)}$$

Esta segunda definición que aparece en el artículo clásico de Duncan-Duncan (1995) es equivalente a la definición que dimos anteriormente cuando se aplica a una dicotomía.

Se demuestra esta equivalencia entre dos definiciones en el caso de una dicotomía como sigue:

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right|$$

$$D = \frac{1}{2} \sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2} \sum_i \left| \frac{p_{ih}}{p_{\bullet h}} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{ih}/p_{i\bullet}}{p_{\bullet h}} - \frac{1-\left(p_{ih}/p_{i\bullet}\right)}{1-p_{\bullet h}}\right|$$

$$D = \frac{1}{2}\sum_i p_{i\bullet}\left|\frac{p_{h/i\bullet}}{p_{\bullet h}} - \frac{1-p_{h/i\bullet}}{1-p_{\bullet h}}\right|$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet}(1-p_{\bullet h}) - (1-p_{h/i\bullet})p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{h/i\bullet}\,p_{\bullet h} - p_{\bullet h} + p_{h/i\bullet}\,p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

$$D = \frac{\sum_i p_{i\bullet}\left|p_{h/i\bullet} - p_{\bullet h}\right|}{2\,p_{\bullet h}(1-p_{\bullet h})}$$

Esta fórmula se presta a una interpretación interesante. En el numerador, aparece un promedio ponderado de las diferencias absolutas $\left|p_{h/i\bullet} - p_{\bullet h}\right|$ entre, por una parte, la fracción $p_{h/i\bullet}$ del grupo *h* en cada categoría *i* y, por otra parte, la fracción $p_{\bullet h}$ del grupo *h* en el total de la población, de modo que el peso $p_{i\bullet}$ de cada categoría es proporcional a su población, esto para cualquier grupo.

En cuanto a la expresión del denominador, ésta es igual a la diferencia absoluta promedio entre los individuos (y no entre las categorías) de la variable dicotómica de pertenencia al grupo *h*. Esta diferencia absoluta promedio es igual a dos veces la varianza de la misma variable.

En efecto, tenemos la variable dicotómica de pertenencia g_t

$$g_t \begin{cases} = 1 \text{ si el individuo } t \text{ pertenece al grupo } h \\ = 0 \text{ en otros casos} \end{cases}$$

donde el índice t se refiere a los individuos de los dos grupos: t varía entre 1 y $x_{\bullet\bullet}$.

La variable g_t tiene una distribución binómica cuyo promedio es

$$\mu_g = \frac{\sum_t g_t}{x_{\bullet\bullet}} = \frac{\sum_i x_{ih}}{x_{\bullet\bullet}} = \frac{x_{\bullet h}}{x_{\bullet\bullet}} = p_{\bullet h}$$

La diferencia absoluta promedio (desviación media) es entonces

$$d_g = \frac{\sum_t \left| g_t - \mu_g \right|}{x_{\bullet\bullet}} = \frac{\sum_t \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = \frac{\sum_{t \text{ tal que } g_t = 1} \left| g_t - p_{\bullet h} \right| + \sum_{t \text{ tal que } g_t = 0} \left| g_t - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = \frac{p_{\bullet h} x_{\bullet\bullet} \left| 1 - p_{\bullet h} \right| + (1 - p_{\bullet h}) x_{\bullet\bullet} \left| 0 - p_{\bullet h} \right|}{x_{\bullet\bullet}}$$

$$d_g = p_{\bullet h} \left| 1 - p_{\bullet h} \right| + (1 - p_{\bullet h}) \left| 0 - p_{\bullet h} \right|$$

$$d_g = 2 p_{\bullet h} (1 - p_{\bullet h})$$

La varianza, por su lado, se escribe con la fórmula

$$\sigma_g^2 = \frac{\sum_t (g_t - p_{\bullet h})^2}{x_{\bullet\bullet}} = \frac{\sum_t \left(g_t^2 - 2 s_t p_{\bullet h} + p_{\bullet h}^2 \right)}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{\sum_t g_t^2 - 2p_{\bullet h}\sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{\sum_t g_t - 2p_{\bullet h}\sum_t g_t + \sum_t p_{\bullet h}^2}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = \frac{p_{\bullet h}x_{\bullet\bullet} - 2p_{\bullet h}\left(p_{\bullet h}x_{\bullet\bullet}\right) + p_{\bullet h}^2 x_{\bullet\bullet}}{x_{\bullet\bullet}}$$

$$\sigma_g^2 = p_{\bullet h} - p_{\bullet h}^2 = p_{\bullet h}\left(1 - p_{\bullet h}\right)$$

El coeficiente de localización y el índice de disimilitud: ¡cuidado, que no son lo mismo!

En ciencia regional es común el uso del coeficiente de localización[74] para medir el grado de especificidad de la repartición espacial de una actividad económica con relación a la economía total.

En una tabla de contingencia del empleo por zona y por rama, $p_{i/\bullet h}$ designa la fracción del empleo total de la rama *h* que se sitúa en la zona *i*; y $p_{i\bullet}$ designa la fracción del empleo total del total de las ramas que se sitúa en la zona *i*. Se define el coeficiente de localización como

$$CL = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right|$$

[74] Según Isard (1960, p. 251) fue P. Sargant Florence quien introdujo el coeficiente de localización como nueva herramienta de la ciencia regional; Duncan y Duncan (1995) citan a P. Sargant Florence, W.G. Fritz y R.C. Gilles, "measure of industrial distribution", cap. 5 en *National Resources Planning Board Industrial Location and National Resources*, Washington, Government Prinding Office, 1943.

A primera vista, es un índice de disimilitud. No obstante, no lo es. Más bien, la relación entre el coeficiente de localización CL y el índice de disimilitud D es la siguiente:

$$CL = (1 - p_{\bullet h})D$$

Demostración:
Sabiendo que *D* se aplica a una dicotomía, tenemos:

$$D = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet k} \right| = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - \frac{p_{i\bullet} - p_{ih}}{1 - p_{\bullet h}} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h}(1 - p_{\bullet h}) - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{i/\bullet h}\, p_{\bullet h} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{ih} - p_{i\bullet} + p_{ih} \right|$$

$$D = \frac{1}{2(1 - p_{\bullet h})}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right| = \frac{CL}{(1 - p_{\bullet h})}$$

Esta diferencia se debe a que el coeficiente de localización compara la distribución de un grupo (una rama de actividad) con la distribución del grupo completo al cual pertenece, cuando el índice de disimilitud compara la distribución de un grupo con la distribución del resto de la población (las demás actividades). Esto implica que no podemos dar al coeficiente de localización la misma interpretación metafórica que al índice de disimilitud, a saber la fracción del grupo a desplazar para obtener distribuciones idénticas. Además, el campo de variación de *CL*, de 0 a $(1 - p_{\bullet h})$, es más estrecho para las ramas más importantes lo que dificulta la comparación entre coeficientes de ramas de tamaños diferentes. Por lo contrario, si queremos medir cómo la reparti-

ción espacial de cada actividad económica depende de esta misma actividad, entonces el índice de disimilitud tiene el inconveniente de usar una distribución de referencia diferente para cada rama. Esta distribución de referencia corresponde a la distribución del conjunto de las demás ramas y, por consiguiente, es diferente para cada rama.

Se pueden ilustrar estas diferencias con el ejemplo utilizado al inicio del capítulo.

Empleo por zona y por rama y distribución del empleo entre las zonas

	Empleo					Distribución entre las zonas				
Rama	*B*1	*B*2	*B*3	*B*1+2	Total	*B*1	*B*2	*B*3	*B*1+2	Total
Zona										
Z1	48	325	287	373	660	0.400	0.542	0.598	0.518	0.550
Z2	27	185	148	212	360	0.225	0.308	0.308	0.294	0.300
Z3	45	90	45	135	180	0.375	0.150	0.094	0.188	0.150
Total	120	600	480	720	1200	1.000	1.000	1.000	1.000	1.000

Comparación de la distribución geográfica de la rama *B*3 con la del conjunto de las tres ramas, pues con la suma de *B*1 y *B*2

Rama	*B*3	Total	Dif.absol.	*B*1+2	Dif.absol.
Zona					
Z1	0.598	0.550	0.048	0.518	0.080
Z2	0.308	0.300	0.008	0.294	0.014
Z3	0.094	0.150	0.056	0.188	0.094
Total	1.000	1.000	0.113	1.000	0.188

Apliquemos la fórmula de cálculo del índice de disimilitud a cada una de ambas comparaciones. En el primer caso (*B*3 y total), se obtiene el coeficiente de localización:

$$CL = \frac{|0.048| + |0.008| + |-0.056|}{2} = 0.056$$

En el segundo caso ($B3$ y $B1+2$), se obtiene el índice de disimilitud:

$$D = \frac{|0.080| + |0.014| + |-0.094|}{2} = 0.094$$

Tal como se esperaba, los resultados son realmente diferentes. Sin embargo, son vinculados por la relación

$$CL = \left(1 - \frac{480}{1200}\right) D = 0.6 \times 0.094 = 0.056$$

donde el factor 0.6 es igual a la parte del empleo de las ramas *demás de B*3.

Cuando el grupo de la población que se considera no es más que una pequeña fracción de la población aparente, $p_{\bullet h}$ es pequeño y el valor del coeficiente de localización es cercano al valor del índice de disimilitud.

En el caso particular donde hay segregación total, el índice de disimilitud D es igual a 1 y el coeficiente de localización es igual a la parte del empleo de las ramas *de más de B*3. Se pueden ilustrar estas diferencias con el ejemplo de segregación total ya mencionado.

Coeficiente de localización: ejemplo de segregación total

Etnia	Números		Reparticiones		
	Marcianos	Total	Marcianos	Total	Diferencia
	x_i	y_i	v_i	w_i	$\lvert v_i - w_i \rvert$
Planeta					
Tierra	0	6	0.00	0.40	0.40
Luna	0	2	0.00	0.13	0.13
Marte	3	3	0.43	0.20	0.23
Júpiter	4	4	0.57	0.27	0.30
Total	7	15	1.00	1.00	

Coeficiente de localización:

$$\frac{0.40 + 0.13 + 0.23 + 0.30}{2} = 0.53 = 1 - \frac{7}{15}$$

= fracción de no-marcianos en la población = fracción de terrícolas.

De la misma manera, se calcula un coeficiente de localización para los terrícolas

$$0.47 = 1 - \frac{8}{15}$$

Post scriptum: el coeficiente de localización y los cocientes de localización

El parecido entre los dos nombres puede ser factor para confundir el coeficiente de localización y el cociente de localización. Sin embargo, cuando el coeficiente de localización compara dos distribuciones, el cociente de localización compara dos partes (vea más arriba), es decir dos puntos que corresponden en dos distribuciones. No obstante, existe una relación entre ambos que se puede ver al desarrollar la definición del coeficiente de localización:

$$CL = \frac{1}{2}\sum_i \left| p_{i/\bullet h} - p_{i\bullet} \right|$$

$$= \frac{1}{2}\sum_i p_{i\bullet} \left| \left(\frac{p_{i/\bullet h}}{p_{i\bullet}} \right) - 1 \right| = \frac{1}{2}\sum_i p_{i\bullet} \left| QL_{ih} - 1 \right|$$

El coeficiente de localización es un promedio ponderado de las diferencias absolutas entre los cocientes de localización y el valor 1 de referencia.

1-5.2.6 Un último vistazo crítico

Como cualquier otro índice, el índice de disimilitud tiene límites. Además de no respetar el principio de transferencia de Pigou-Dalton, debemos mencionar:

El índice de disimilitud no admite datos negativos. Por ejemplo, sería imposible usar el índice de disimilitud para medir la similitud entre dos ramas de actividad en cuanto a la variación del número de empleos por zonas porque estas variaciones pueden ser negativas.

Al igual que el índice de Gini, el índice de disimilitud es sensible a la definición y al número de las categorías (clases, zonas). Este defecto no es tan grave siempre y cuando la división que se seleccione sea lo suficiente fina, o sea que considere un buen número de categorías, pero también que las comparaciones entre divisiones no tengan ningún peso significativo.[75]

Cuando se aplica a datos espaciales, el índice de disimilitud, tanto como el índice de Gini, no toma en cuenta la contigüidad o la proximidad de las unidades espaciales.

Como en el caso de los índices de Laspeyres y de Paasche, cuando vimos que se podía construir índices de precios con fundamentos teóricos más satisfactorios pero, en cambio, con más alto grado de complejidad, se puede definir indicadores de disimilitud más refinados, de los cuales Waldorf (1993) nos da un ejemplo. Sin embargo, hemos de interrogarnos si, según el contexto, tales refinamientos son del todo pertinentes y concretos. Además, la presentación de Waldorf no deja por completo de caer en la trampa que consiste en pasar de la metáfora a la interpretación literal; de hecho, en el contexto de un estudio de la segregación racial en los Estados

[75] Se examina este problema en todos los escritos de geografía en la rúbrica MAUP, que significa “Modifiable Areal Unit Problem”.

Unidos, menciona el "esfuerzo requerido" para un desplazamiento de la población.

1-5.3 DISTANCIA Y DISIMILITUD

La medición de distancias geográficas tiene una gran importancia en los estudios urbanos y regionales. Se puede considerar la medición de la distancia como un caso particular de la medición de la disimilitud: la distancia es una medición de la disimilitud entre dos objetos con relación a su situación en el espacio, o sea entre dos lugares en el espacio.

Una superficie (como la superficie a condición de ignorar el relieve) es un espacio de dos dimensiones. La especificación de una situación en el espacio tiene, por lo tanto, dos dimensiones: longitud y latitud o coordenadas cartesianas (x,y). En consecuencia, la medición de la distancia geográfica tiene también dos dimensiones. De hecho, aunque en la vida cotidiana acostumbramos usar la distancia euclidiana de manera automática, existen otras maneras de medir la distancia.[76]

Una medición de distancia debe satisfacer algunas condiciones: la función $d(a,b)$ es una función de distancia si y solamente si, para todo conjunto de lugares a, b y c, esta función satisface las cuatro condiciones siguientes:

(c1) no negativo
$d(a,b) \geq 0$

(c2) identidad
$d(a,b) = 0$ si y solamente si, $a = b$

(c3) simetría
$d(a,b) = d(b,a)$

(c4) desigualdad triangular
$d(a,c) \leq d(a,b) + d(b,c)$

[76] Vea Huriot y Perreur (1990 y 1994).

La medición de distancia que más usamos es la distancia euclidiana. La distancia euclidiana entre el punto a, de coordenadas (x_a,y_a) y el punto b de coordenadas (x_b,y_b) es:

$$d_e(a,b) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

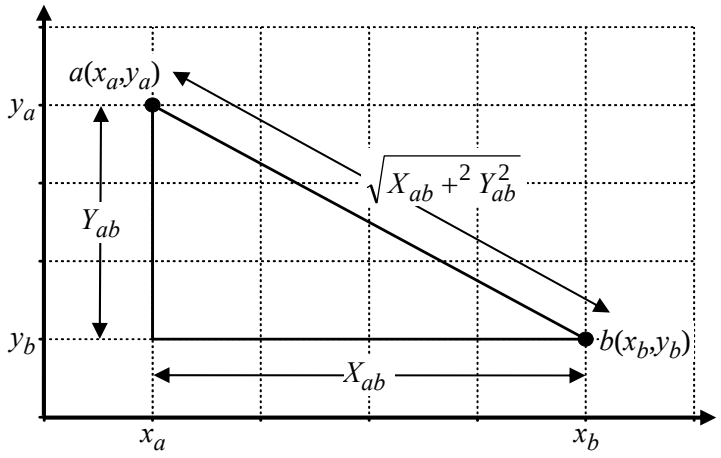
Entre las otras mediciones de distancia posibles, mencionemos la distancia rectilínea, conocida también como distancia según la métrica de Manhattan (vea Huriot y Perreur, 1994, p. 44):

$$d_r(a,b) = \left| x_a - x_b \right| + \left| y_a - y_b \right|$$

La distancia según la métrica de Manhattan es la distancia que se debe recorrer para ir del punto *a* al punto *b* siguiendo el trazo de las calles cuando estas forman una cuadra como en Manhattan.

Las dos métricas se enseñan en la figura que sigue teniendo

$X_{ab} = |x_a - x_b|$ et $Y_{ab} = |y_a - y_b|$



Puesto que se puede interpretar la distancia geográfica como una disimilitud, la recíproca es también verdadera: se

puede usar las mediciones de la distancia para medir las disimilitudes que no son distancias geográficas.

De esta manera, consideremos dos objetos que describimos con n variables que miden, cada una, una característica (dimensión) pertinente:

$x_{11}, x_{12}, ..., x_{1n}$ para el primer objeto y

$x_{21}, x_{22}, ..., x_{2n}$ para el segundo.

Ejemplo:

> Si los dos objetos fueran dos barrios de una ciudad, las características pertinentes podrían ser la densidad de la población, la proporción de la población menor de quince años, la proporción de la población que haya completado sus estudios de primaria, el ingreso promedio de los hogares, etcétera.

Para medir la disimilitud entre dos objetos multidimensionales, se usa a menudo la distancia euclidiana generalizada que tiene por fórmula

$$\sqrt{\sum_i (x_{1i} - x_{2i})^2}$$

Se usa también la distancia lineal generalizada o la distancia generalizada según la métrica de Manhattan, la cual se define con

$$\sum_i \left| x_{1i} - x_{2i} \right|$$

El lector perspicaz habrá observado la relación que existe entre el índice de disimilitud D y la distancia generalizada según la métrica de Manhattan. No obstante, en el presente contexto, los dos objetos que se comparan no son necesariamente distribuciones. Es de notar, en consecuencia, que no hay valor máximo inherente a la distancia recta generalizada (ni, tampoco, a la distancia euclidiana generalizada).

En general, el valor de una medición de distancia depende de las unidades de medición de las variables subyacentes. Por

este motivo, al momento de comparar dos objetos multidimensionales por medio de una distancia generalizada, no queda más que enfrentarnos a un problema parecido al problema encontrado cuando la construcción de un número índice. En efecto, la selección de la unidad de medición de cada variable determina de manera implícita cuál será el peso de la distancia-disimilitud en la medición. Sólo cuando los objetos que se comparan son distribuciones, el problema no se presenta.

1-5.4 La medición de la similitud en estadística

El problema de la medición de la similitud surge con frecuencia en estadística. Consideremos, por ejemplo, dos series de observaciones de dos variables:

$x_1, x_2, ..., x_n$ y $y_1, y_2, ..., y_n$

El coeficiente de correlación simple[77] es una medición de similitud entre dos series de datos.

Asimismo, para evaluar la exactitud de un modelo con relación a los datos que permitieron estimar sus parámetros, se mide la similitud entre los valores observados y los valores que la teoría predice. Una de las mediciones que más se emplea para este fin es el coeficiente de determinación múltiple R^2 (del cual hablaremos en la tercera parte de esta obra).

Finalmente, el Ji-cuadrado de Pearson[78] es una medición de la disimilitud entre los números observados y les números "teóricos" predichos por una hipótesis.

Todas estas mediciones pertenecen a la gran familia de las mediciones de similitud y disimilitud.

---

[77] Vea el anexo 2-A, "Recuerdo de algunas fórmulas usuales en estadísticas".

[78] Vea 4-1. Pero el Ji-cuadrado no es una medida simétrica : su valor varia si se intercambian lo papeles de los valores observados y teóricos.

En Webber (1984, pp. 41-45) se lleva a cabo una interesante discusión sobre el grado de pertinencia de diferentes mediciones de ajuste (en el contexto de la evaluación de la exactitud del modelo de repartición espacial de Lowry).

1-5.5 Otras mediciones de similitud y de disimilitud

Las mediciones de similitud y disimilitud existen en abundancia. Legendre y Legendre (1984, tomo 2, cap. 6, 1998) presentan y discuten numerosas mediciones que se usan en ecología numérica, las cuales se podrían emplear para el análisis espacial en ciencias sociales.