OPHI

OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE

www.ophi.org.uk

UNIVERSITY OF OXFORD

Asociación y Similitud

Sabina Alkire, Paola Ballón, Ana Vaz

Managua, 4 de Septiembre de 2013

El camino recorrido…

Comenzamos definiendo:

- El propósito
- la unidad de Análisis
- las Dimensiones

Luego nos detuvimos para analizar los datos, y posteriormente definimos y seleccionamos

- los indicadores, y
- los umbrales de privación

La clase de hoy

Hoy, nos detenemos nuevamente para analizar y comprender la **asociación** entre pares de indicadores de privación.

Esto nos permitirá posteriormente:

- Mejorar la definición de la clasificación de los indicadores en dimensiones

- Tener mayor solidez al momento de establecer ponderaciones o pesos tentativos para la construcción de medidas posibles.

Porque es necesario analizar la asociación ?

Para concluir sobre una posible 'redundancia'

Para identificar:

cuales indicadores están altamente asociados, y

cuales indicadores tienen asociación baja.

Para que sirve el análisis de asociación?

El análisis de asociación nos ofrece las siguientes opciones en el proceso de medición

- Modificar la ponderación de los indicadores altamente asociados
- Combinar algunos indicadores (sub-índice)
- Revisar la justificación de la selección de indicadores
- Adaptar la categorización de indicadores en dimensiones

Multidimensionalidad y Asociación: Una literatura en rápida expansión

La literatura del análisis de la asociación entre indicadores de privación múltiples es compleja e incluye perspectivas diversas

Perspectiva 1: Favorece una asociación baja

- **Alta correlación indica redundancia**

  el o los indicadores redundantes no deben ser incluidos
- **Baja redundancia** – justifica la construcción de una medida multidimensional
- Ranis, Samman, y Stewart, 2006; McGillivray y White, 1993.

Multidimensionalidad y Asociación

Perspectiva 2: Favorece la alta asociación

- Alta asociación favorece la construcción de medidas robustas
  Ejemplo: Índices compuestos tradicionales
  Estos se concentran en una medición marginal de pobreza e ignoran la distribución conjunta

- Por lo tanto indicadores con asociación baja no deben incluirse en la medida.

- Saisana, M., A. Saltelli, and S. Tarantola 2005, Foster, McGillivray, and Seth, 2012; *Handbook of Composite Indicators*; OECD, 2008;

Multidimensionalidad y Asociación

Nuestra perspectiva (tentativa): ni una ni la otra

Si dos indicadores tienen una asociacion alta y:
a) si existe una normativa o política pública que requiera incluir *ambos* indicadores, y
b) si esta es factible;

entonces se deben incluir los dos incadores pero con ponderaciones bajas.

En la ausencia de tal normativa factible, uno de los dos indicadores debe suprimirse de la medición.

Multidimensionalidad y Asociación

Nuestra perspectiva (tentativa): ni una ni la otra

Si dos indicadores tienen una asociación baja y si cada indicador es importante por separado entonces *ambos* deben incluirse en la medición.

Nota: En este caso suponemos que cada indicador contribuye directamente a la medición de pobreza o bienestar.

Caso contrario, debemos considerar usar ambos indicadores y combinarlos en un sub-indice.

Definiciones

Dos conceptos clave en el análisis de indicadores múltiples de privación son la asociación y la similitud.

Ambos conceptos se emplean cuando el análisis involucra variables dicotómicas o categóricas.

Asociación – es un indicador de la fuerza y la dirección de la relación existente en un par de indicadores dicotómicos, mientras que la

Similitud – es un indicador de la fuerza de dicha relación únicamente.

Fuentes de información

Para el análisis de la "asociación"/similitud entre indicadores de privación

a) Nos centraremos en los scores de privación dicotómicos, que toman dos valores, 0 o 1.

b) Emplearemos dos fuentes de información distintas:

Indicadores de privación brutos → tasas de privación brutas

Indicadores de privación censurados→ tasas de privación censuradas

c) Y usaremos una tabla de contingencia o **cross-tab.**

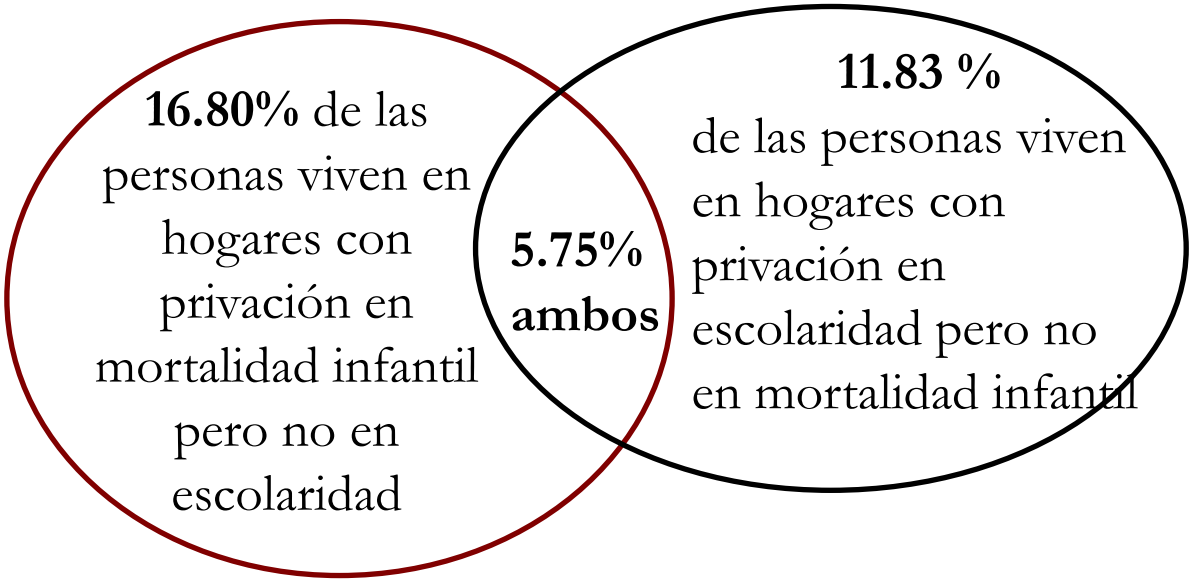
Esta constituye el instrumento principal para la representación de las relaciones entre indicadores dicotómicos, en nuestro caso de privación

Ejemplo

India NFHS submuestra (2005-6)

Tasa de privación bruta en mortalidad infantil **22.55%**

17.58% Tasa de privación bruta en escolaridad



Es que aquellos que sufren privación en un indicador son los mismos que sufren privación en el otro? Como podemos ver esto?
La tabla de contingencia nos permitirá estudiar la distribución conjunta

Tabla de Contingencia (cross-tab)
Tasa de Incidencia Bruta

		Mortalidad infantil (J)	
Agua potable (I)	No privado = 0	Privado = 1	Total
No privado =0	4 (67%, 80%)	2 (33%, 40%)	6
Privado = 1	1 (25%, 20%)	3 (75%, 60%)	4
Total	5	5	10

Tasas de incidencia bruta: Agua potable=40%, Mortalidad infantil= 50%

Pregunta: Que información de la tabla podemos utilizar para medir la asociación?

Tabla de contingencia (cross-tab)
Tasas de Incidencia Brutas
"P" = 75%

	Mortalidad infantil (J)		
Agua potable (I)	No Privado = 0	Privado = 1	Total
No Privado =0	4 (67%, 80%)	2 (33%, 40%)	6
Privado = 1	1 (25%, 20%)	3 (**75%,** 60%)	4
Total	5	5	10

Tasas de incidencia bruta: Agua potable=40%, Mortalidad infantil= 50%

Pregunta: Que información de la tabla podemos utilizar para medir la asociación?

Una medida de similitud *: "P"

Si dos indicadores de privación/pobreza no son independientes, y por lo menos una de las distribuciones marginales n_{1+} , n_{+1} es diferente de cero, P se define como:

$$P = \frac{n_{11}}{\min\left[n_{1+}, n_{+1}\right]} \in \left[0,1\right]$$

Fuentes de información utilizadas para calcular P:

n_{11} numero de personas que están privadas en los dos indicadores (concordancias) → Distribución conjunta

n_{1+} , n_{+1} tasa de privación (censurada o no) → Dist. marginales

*** Similitud** refleja la fuerza de las "concordancias";

Interpretando "P"

Un valor de P = 90% indica que 90% de las personas que están privadas en el indicador con la tasa de incidencia (bruta) más baja, también están privadas en el otro indicador.

Que podemos concluir sobre esta elevada similitud?

Un elevado valor de P no es bueno o malo por si mismo.

La conclusión sobre la inclusión/exclusión de los indicadores requiere reflexión. Esto conlleva analizar:

Redundancia

Justificación para su exclusión/inclusion:

Los indicadores tienen una justificación normativa o de monitoreo para ser incluidos de manera individual

Y en cuanto los indicadores de nivele de vida?

Analizando el combustible para cocinar:

		Fuel		
		Average P (%)	Number of Countries	Coefficient Variation of P
Indicator with the lowest Censored Headcount	Schooling	97	15	0.05
	Enrolment	94	15	0.12
	Ch.Mortality	94	15	0.10
	Nutrition	93	15	0.12
	Elect.	98	15	0.03
	Sanit	99	12	0.01
	Water	98	15	0.03
	Floor	99	15	0.02
	Assets	98	15	0.04

Niveles de P muy altos, pero coeficiente de correlación bajo

Redundancia?

Interpretando "P"

Un valor de P =10% indica que 10% de las personas que están privadas en el indicador con la tasa de incidencia bruta mas baja, están también privadas en el otro indicador

Que podemos concluir sobre esta baja similitud?

_ Un valor de P pequeño tampoco es bueno o malo por si mismo.

– Tenemos que reflexionar …

- esta es una relación esperada o no? cual es la intuición?
- medidas con unión van a ser mas altas que aquellas con un valor de *k* censurado
- medidas utilizando intersección van a ser menores que 10%
- cual es el lo valor de P con otros indicadores? (error de medición?)

Correlación o similitud?

Que sucedería si calculamos las correlaciones entre indicadores dicotómicos 0, 1 y obtenemos un patrón diferente de correlación comparado al obtenido con la medida "P".

Cómo podríamos explicarlo?

La correlación está basada en todas las entradas de la tabla de contingencia:

la tasa de incidencia bruta de cada indicador

la privaciones concordantes

las privaciones discordantes

Pero es correcto calcular una correlación con este tipo de variables?

La Tabla de Contingencias

Formalmente:

	Mortalidad Infantil		
Agua potable	No privado = 0	Privado = 1	Total
No privado =0	n_{00}	n_{01}	n_{0+}
Privado = 1	n_{10}	n_{11}	n_{1+}
Total	n_{+0}	n_{+1}	n

$$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$$

n_{ij} Denotan las frecuencias por celda

n_{i+}, n_{+j} Denotan las distribuciones marginales por fila y columna

La correlación

Cramer V es la medida más popular de asociación entre dos variables nominales, esto debido a su rango de variación.

$$V = \frac{\overbrace{n_{00}n_{11}}^{concordancias} - \overbrace{n_{01}n_{10}}^{discordancias}}{\underbrace{\left(n_{0+}n_{1+}n_{+0}n_{+1}\right)}_{distribuciones\,marginales}^{1/2}}, \in [-1,1]$$

En el caso 2 x 2, V varia entre 0 y ±1. Toma los valores extremos cuando las variables son (estadísticamente) independientes (0) o "completamente asociadas o disociadas" (±1).

Sin embargo, cuando las variables son **dicotómicas** (0-1), el coeficiente de **correlación** de Pearson es **igual** a la medida de asociación Cramer ***V***.

Testando la Independencia: $\chi 2$

Independencia está basada en las leyes de las probabilidades: dos variables son independientes si su distribución conjunta es igual al producto de sus distribuciones marginales.

Para concluir sobre la independencia usamos la estadística $\chi 2$.

Muchos de los coeficientes de asociación para variables nominales (Phi, Contingencia, Cramer's *V, Tschuprovw's T,* Lambda, y Incertidumbre) son función de la estadística $\chi 2$.

Fuentes de información utilizadas por Correlaciones 2x2/Cramer V

La fuerza de la relación está definida como el producto de las concordancias menos el producto de las discordancias, dividido por el producto de las distribuciones marginales de los indicadores

Concordancias **Discordancias**

$$V = \frac{\overbrace{n_{00}n_{11}} - \overbrace{n_{01}n_{10}}}{\underbrace{\left(n_{0+}n_{1+}n_{+0}n_{+1}\right)^{1/2}}_{\text{Distribuciones marginales}}}, \in [-1,1]$$

La correlación entre un **par** de **indicadores dicotómicos** utiliza **toda** la información de la tabla de contingencia

Ejemplo - Bangladesh DHS

Caso I

Años de escolaridad (I)	**Asistencia a la escuela (J)**		
	No privado= 0	Privado= 1	Total
No privado=0	55,049 71%	**7,301** 9%	62,351 80%
Privado= 1	10,657 14%	**4,455** 6%	15,112 20%
Total	65.706 85%	11,756 15%	77,463

$$P = \frac{n_{11}}{min[n_{1+}, n_{+1}]} = 0.379 \qquad V = \frac{n_{00}n_{11} - n_{01}n_{10}}{[n_{0+}n_{1+}n_{+0}n_{+1}]^{1/2}} = 0.196$$

Ejemplo - Mozambique DHS

Caso II

	Asistencia a la escuela (J)		
Años de escolaridad (I)	No Privado= 0	Privado= 1	Total
No Privado=0	28,722 47%	**8,845** 15%	37,567 62%
Privado= 1	**13,431** 22%	**9,913** 16%	23,344 38%
Total	42,153 69%	18,758 31%	60,911

$$P = \frac{n_{11}}{min[n_{1+}, n_{+1}]} = 0.528 \qquad V = \frac{n_{00}n_{11} - n_{01}n_{10}}{[n_{0+}n_{1+}n_{+0}n_{+1}]^{1/2}} = 0.199$$

Dos países con patrones de privación **muy distintos** tienen el mismo coeficiente de asociación V, pero medidas de similitud "P" diferentes.

Correlación vs. Similitud - "P"

Matriz de Correlaciones

	Escolaridad	Matricula	Agua	Combustible
Escolaridad	1.000	0.199	0.330	0.139
Matricula		1.000	0.188	0.111
Agua			1.000	0.201
Combustible				1.000

Medida "P"

Indicador con la tasa de privación más baja →

	Escolaridad	Matricula	Agua	Combustible
Escolaridad			0.776	0.999
Matricula	0.529		0.708	0.997
Agua				0.999
Combustible				

Corre "P"

La privación entre escolaridad y agua potable es más ALTA que la correlación entre escolaridad y combustible para cocinar.

Matriz de

		Agua	Combustible
Escolaridad		0.330	0.139
Matricula	1.000	0.188	0.111
Agua		1.000	0.201
Combustible			1.000

Medida "P"

	Escolaridad	Matricula	Agua	Combustible
Escolaridad			0.776	0.999
Matricula			0.708	0.997
Agua				0.999
Combustible				

Indicador con la tasa de privación más baja

La similitud entre escolaridad y agua potable es más BAJA que la similitud entre escolaridad y combustible para cocinar.

3. PCA, MCA y FA: Métodos Estadísticos en presencia de variables Múltiples

Estos tres métodos estudian la **asociación** (variables categóricas) o **correlación** (variables cardinales) a través de una **matriz de "información" multivariada** .

Los tres métodos utilizan **todos** los elementos de la tabla de contingencia.

Sin embargo emplean distintas matrices de información (matriz –*insumo*) y distintos procedimientos estadísticos y matemáticos.

Matrices “insumo” - información

PCA y MCA son técnicas **descriptivas**.

Las matrices insumo:

PCA: matriz de correlaciones

MCA: tabla de contingencia (todas entradas)

FA es un método basado en un **modelo**.

Matriz insumo: matriz de correlación ajustada por tipo de correlación

Pearson para pares de variables cardinales,
Tetrachorica/ polychorica para pares de variables binarias/categoricas
Biserial para pares de variables cardinales y binarias

PCA: Análisis de componentes principales

Es una técnica **estadística** utilizada para reducir el número de dimensiones de una base de datos. Esta técnica también es utilizada para analizar la estructura latente de los datos.

PCA **reemplaza** un grupo de variables correlacionadas (x) con un numero, más reducido, de 'nuevas' variables no correlacionadas, llamadas componentes (y), de manera que los componentes conserven la mayor parte de la información contenida en los datos.

Así:

$$y_1 = a_{11}x_1 + a_{21}x_2 + \ldots + a_{d1}x_d$$

$$y_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{d2}x_d$$

$$\vdots$$

$$y_d = a_{1d}x_1 + a_{2d}x_2 + \ldots + a_{dd}x_d$$

Como funciona?

- PCA incluye 3 pasos:

a) Cálculo de los componentes principales

   Esto implica hallar los coeficientes 'a'. Para ello se emplea la descomposición espectral de la matriz de correlaciones (valores y vectores propios)

b) Extracción o selección del número de componentes

c) Rotación de los componentes para facilitar la interpretación (algunas veces)

Análisis de Componentes Principales- Ejemplo

Filmer y Pritchett (1999, 2001) popularizaron el enfoque del índice de activos (asset index approach), que hace una aproximación (proxies) al estatus de bienestar de una población.

Desarrollaron su índice en el contexto de analizar las asociaciones entre el estatus económico de los hogares y los resultados de escolaridad cuando la información disponible no incluía información acerca del gasto de los hogares (encuestas DHS).

Análisis de Componentes Principales- Ejemplo

Este enfoque utiliza el análisis de componentes principales para calcular este índice de activos.

Desde entonces, el enfoque del índice de activos ha sido utilizado para una diversidad de propósitos, incluyendo en análisis de desigualdad, cambios en la pobreza (Sahn y Stifel 2000, Stifel y Christiaensen 2007, Mckenzie 2005).

Algunas aplicaciones del índice de Filmer y Pritchett : Sahn and Stifel 2000, Stifel and Christiaensen 2007, Mckenzie 2005

Un ejemplo

El índice de activos propuesto por Filmer y Pritchet (2000) es entonces:

$$A_i = a_1 x_{1i} + a_2 x_{2i} + \ldots + a_k x_{ki}$$

donde A_i es el índice de activos de un hogar i,

x_s son indicadores o variables de posesión de activos y de calidad de la vivienda.

a_s son las ponderaciones, obtenidas del primer Compomente Principal, utilizado para agregar los indicadores en un índice.

Nota: Filmer y Pritchet aplicaron el ACP a datos binarios. Una técnica más conveniente sería utilizar MCA. Aunque existe una equivalencia entre ACP y MCA, pero con valores cardinales diferentes.

Recapitulemos

En esta clase hemos revisado el debate existente sobre la asociación y correlación.

Vimos como la medida "P" puede ser utilizada para identificar similitudes entre los indicadores.

Vimos que las correlaciones y las técnicas PCA/FA/MCA utilizan **todas** las entradas de la tabla de contingencia. Esto puede levar a conclusiones diferentes, derivadas de relaciones ajenas a la similitud.

Para que nos sirve el análisis de correlaciones/asociaciones?

- Abandonar o modificar las ponderaciones en indicadores muy asociados
- Combinar algunos indicadores en un sub-índice
- Revisar la 'justificación' de los indicadores
- Ajustar la categorización de los indicadores en las dimensiones.

Ejercicio

Para cada par de sus indicadores

a) Obtenga las tablas de contingencia
b) Calcule Cramer V
c) Calcule la medida de similitud P
d) Compare ambas medidas y concluya

Muchas gracias ☺