OPHI

OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE
www.ophi.org.uk

UNIVERSITY OF OXFORD

Associations across Deprivations

Sabina Alkire, Paola Ballón, Jose Manuel Roche, Ana Vaz

Washington DC 13 July 2013

Tabita, Kenya | Rabiya, India | Stéphanie, Madagascar | Agathe, Madagascar | Dalma, Kenya | Ann-Sophia, Kenya | Valérie, Madagascar

Where we are…

You have defined:

- Purpose
- Unit of Analysis
- Dimensions

Then you took a pause and described the data, before defining

- Indicators
- Deprivation cutoffs

Now, we take another pause, to describe and understand the associations between deprivations, before defining:

- Reconsider your selection of indicators
- Categorization of indicators into Dimensions
- Tentative Weights for trial measures

Why this pause?

To identify 'redundancy'

To see which indicators are highly associated

which indicators have low associations

What might you do based on an analysis of associations?

- Drop or modify weights on highly associated indicators
- Combine some indicators into a sub-index
- Revise your 'justification' of indicators
- Adjust your categorization of indicators into dimensions.

Multidimensionality & Association: A rapidly-changing literature

The study of the association across multiple indicators of deprivation engages literatures with diverse views on association.

View 1: Low association favoured

- **High correlation signals redundancy**
- redundant indicator(s) could be dropped
- **Low redundancy** – justifies multidimensional measure
- Ranis, Samman, and Stewart, 2006; McGillivray and White, 1993.

Multidimensionality & Association

View 2: High association favoured

- **Traditional composite marginal** measures (not joint distribution)
- Aggregate indicators having high association
  - to generate a robust measure.
- Do not include indicators having low association
- Saisana, M., A. Saltelli, and S. Tarantola 2005, Foster, McGillivray, and Seth, 2012; *Handbook of Composite Indicators*; OECD, 2008;

Our view (tentative): <u>not one or the other</u>

If indicators are highly associated, *if* there is a normative/policy need to include *both* indicators it is possible, but their weights may be less. Otherwise one might be dropped.

If indicators have a low association, and if each is independently important, then *both* can be entered in the index.

Note: we presume that each indicator contributes directly to poverty or well-being; if *always* requires another (left shoe), consider a sub-index.

Sources of information

We focus on dichotomised deprivation scores, 0 or 1.

To study the "association"/similarity across deprivation indicators you might use two different sources of information:

Raw deprivation indices → headcounts (**you have these**)

Censored deprivation indices→censored head counts (**not yet**)

This class:

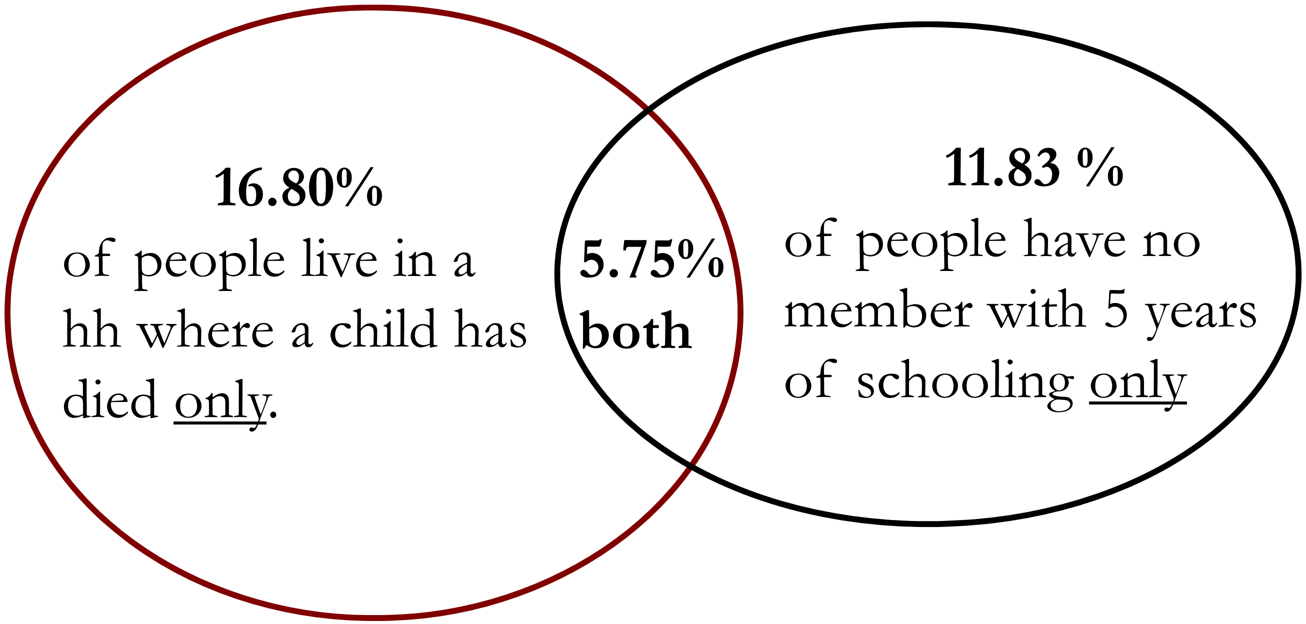
1. **Cross – tab** (the basic tool for displaying the relationships across indicators) – and a linked measure **'P'**
2. **Correlation** (Cramer's *V*)
3. **PCA/**FA/MCA

Describing Associations

Recall: India NFHS data 2005-6 (sub-sample)

Raw headcount of child mortality — **22.55%**

Raw headcount of schooling — **17.58%**



Are they mostly the same people? **Less than one-third of the time.**

Cross-tabs are a basic way to view a joint distribution

The Cross-tab (Contingency Table) – Raw Headcounts

	Child mortality (J)		
Safe water (I)	Non deprived = 0	Deprived = 1	Total
Non Deprived =0	4 (67%, 80%)	2 (33%, 40%)	6
Water Deprived = 1	1 (25%, 20%)	3 (75%, 60%)	4
Total	5	5	10

Raw headcount ratios: Safe water=40%, Child mortality= 50%

Question: What information of this cross tab do we use to assess association?

The Cross-tab (Contingency Table) – Raw Headcounts

"P" = 75%

	Child mortality (J)		
Safe water (I)	Non deprived = 0	Deprived = 1	Total
Non Deprived =0	4 (67%, 80%)	2 (33%, 40%)	6
Water Deprived = 1	1 (25%, 20%)	3 **(75%, 60%)**	4
Total	5	5	10

Raw headcount ratios: Safe water=40%, Child mortality= 50%

Question: What information of this cross tab do we use to think about association?

A measure of similarity*: "P"

If two deprivation/poverty indicators are not independent, and if at least one of the marginal distributions n_{1+} , n_{+1} is different from zero P is defined as:

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

Sources of information used by P:

n_{11} number of people who are MD poor in both indicators → Joint

n_{1+} , n_{+1} censored headcount ratios ("levels") → Marginals

* **Similarity** reflects strength of the matches;
Association reflects strength and direction

Interpreting 'P'

If P = 90%, it shows that 90% of the people who are deprived in the indicator with the lower raw headcount are also deprived in the other indicator (The match for the other indicator will *always* be lower, mechanically, because it has a higher raw headcount).

Observation: this is a high association!

- that is not 'bad' or 'good' – we need to think…
- do I need both of these indicators or is one redundant?
- how do I justify having both in? –

e.g. are they of independent value
normatively or for monitoring purposes?

Example: Let's say sanitation and cooking fuel have high associations. Why might you keep both indicators? Why might you drop one?

Illustration: "P" Coefficient
Average over 15 countries

		Sch.	Enrol.	Ch.Mort.	Nut.	Elect.	Sanit	Water	Floor	Fuel	Assets
Indicator with the lowest Censored Headcount	Schooling		35	31	28	89	93	61	81	97	80
	Enrolment	45		45	41	85	88	57	79	94	68
	Ch.Mortality	51	54		46	82	88	55	73	94	67
	Nutrition	39	37	53		82	87	54	68	93	64
	Elect.	39	37	53	0		95	0	93	98	92
	Sanit	0	0	0	0	95		0	96	99	94
	Water	60	48	48	48	92	93		89	98	81
	Floor	52	39	49	0	95	94	67		99	85
	Fuel	0	0	0	0	0	96	0	0		0
	Assets	0	0	49	39	93	94	69	89	98	

3. What about Living Standard Indicators?

Let's look at Fuel:

		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

Very high values of P across 15 countries, very small C.V

Redundancy?

Interpreting 'P'

If P = 10%, it shows that 10% of the people who are deprived in the indicator with the lower raw headcount are also deprived in the other indicator (The match for the other indicator will *always* be even lower, mechanically, because it has a higher raw headcount).

Observation: this is a low association

- that is not 'bad' or 'good' – we need to think…
- is this relationship expected or unexpected? Intuition?
- <u>union</u> will be higher than a censored *k* value
- measures using <u>intersection</u> will be lower than 10%
- what are P values with other indicators? (meas. error?)

Illustration: "P" Coefficient

		Sch.	Enrol.	Ch.Mort.	Nut.	Elect.	Sanit	Water	Floor	Fuel	Assets
Indicator with the lowest Censored Headcount	Schooling		35	31	28	89	93	61	81	97	80
	Enrolment	45		45	41	85	88	57	79	94	68
	Ch.Mortality	51	54		46	82	88	55	73	94	67
	Nutrition	39	37	53		82	87	54	68	93	64
	Elect.	39	37	53	0		95	0	93	98	92
	Sanit	0	0	0	0	95		0	96	99	94
	Water	60	48	48	48	92	93		89	98	81
	Floor	52	39	49	0	95	94	67		99	85
	Fuel	0	0	0	0	0	96	0	0		0
	Assets	0	0	49	39	93	94	69	89	98	

2. What about Correlation?

Now let's correlate the 0-1 deprivations. What happens?

Correlation coefficients may not have the same pattern as P. Why?

The correlation is based on all of the elements of the cross-tab.

- the raw headcount of each variable
- the 'match' between deprivations
- the 'match' between non-deprivations
- the mismatches

The Cross-tab or Contingency table

Formally:

	Child mortality		
Safe water	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	n_{00}	n_{01}	n_{0+}
MD Poor = 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} are the cell count frequencies

n_{i+}, n_{+j} are the row, and column marginal totals

Correlation

For 0-1 variables, the correlation coefficient is the same as the Cramer's *V* measure.

Cramer's *V* is the most popular measure of association between two nominal variables because of its norming range

In the 2x2 case, V ranges from 0 to ±1, and take the extreme values under (statistical) independence and "complete association".

$$V = \frac{n_{00}n_{11} - n_{01}n_{10}}{(n_{0+}n_{1+}n_{+0}n_{+1})^{1/2}} \quad , \in [-1,1]$$

Meaning and interpretability of Correlation Coefficients / *V*

V^2 is the mean square canonical correlation between two variables.
2x2 correlation coefficients/V could be viewed as the percentage of the maximum possible variation between two variables.

Testing for Independence: $\chi 2$

Independence is based on the laws of probability: i.e. two variables are independent if their joint distribution equals the product of marginals.

This is tested through the $\chi 2$ statistic.

Most coefficients of association for nominal variables like, Phi, Contingency, Cramer's *V* (2x2 correlation coefficients), *Tschuprovw's T*, Lambda, and Uncertainty rely on the $\chi 2$ statistic.

Sources of information used by 2x2 Correlations/Cramer's V

Strength of the relationship is defined as the product of matches minus product of mismatches adjusting for the marginal distribution of the variables.

$$V = \frac{\overbrace{n_{00}n_{11}}^{matches} - \overbrace{n_{01}n_{10}}^{mismatches}}{\underbrace{(n_{0+}n_{1+}n_{+0}n_{+1})}_{marginal\ distributions}{}^{1/2}}, \in [-1,1]$$

This is, correlations use the "**entire** cross-tab"

What are the implications for MD poverty analysis?

Example - Bangladesh DHS

Case I

Years school. (I)	**School attendance (J)** Non deprived= 0	Deprived= 1	Total
Non deprived=0	55,049 71%	**7,301** 9%	62,351 80%
Deprived= 1	10,657 14%	**4,455** 6%	15,112 20%
Total	65.706 85%	11,756 15%	77,463

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

Example - Mozambique DHS

Case I

Years school. (I)	School attendance (J): Non deprived= 0	Deprived= 1	Total
Non deprived=0	28,722 47%	**8,845** 15%	37,567 62%
Deprived= 1	**13,431** 22%	**9,913** 16%	23,344 38%
Total	42,153 69%	18,758 31%	60,911

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

Two different countries with **completely different** patterns of deprivation show the **same association** coefficient **V**, but **different** « P » measures

Correlation vs. "P" Measure

Correlation Matrix

	Schooling	Enrolment	Water	Cooking fuel
Schooling	1.000			
Enrolment	0.199	1.000		
Water	0.330	0.188	1.000	
Cooking fuel	0.139	0.111	0.201	1.000

"P" Measure

	Schooling	Enrolment	Water	Cooking fuel
Schooling				
Enrolment	0.529			
Water	0.776	0.708		
Cooking fuel	0.999	0.997	0.999	

Correlation vs. "P" Measure

Correlation Matrix

	Schooling	Enr	el
Schooling	1.000		
Enrolment	0.199	1	
Water	0.330		
Cooking fuel	0.139	0	

Water is more highly correlated with schooling deprivations than cooking fuel.

"P" Measure

	Schooling	Enr	l
Schooling			
Enrolment	0.529		
Water	0.776		
Cooking fuel	0.999	0	

Water is less highly correlated with schooling deprivations than cooking fuel.

Which is right?

Correlation vs. "P" Measure

Correlations Matrix

	Schooling	Enrolment	Water	Cooking fuel
Schooling	1.000			
Enrolment	0.199	1.000		
Water	0.330	0.188	1.000	
Cooking fuel	0.139	0.111	0.201	1.000

"P" Measure **Denominator (min value) is on the line**

	Schooling	Enrolment	Water	Cooking fuel
Schooling			0.776	0.999
Enrolment	0.529		0.708	0.997
Water				0.999
Cooking fuel				

3. PCA, MCA and FA: Multivariate Statistical methods

These three methods study the **association** (categorical variables) or **correlation** (cardinal variables) through a **multivariate input data matrix**.

All three methods use all elements of the cross-tab.

However the input data matrices and the way they are used, differ.

Input data matrices

PCA and MCA are **descriptive** techniques.

Input matrices:

PCA: correlation matrix | MCA: cross-tabs (all elements)

FA is a **model-based** method.

Input matrix: 'correlation matrix' with:

Pearson correlations for pairs of cardinal variables,
Tetrachoric correlations for pairs of binary variables,
Polychoric if not a dichotomous variable
Biserial correlations for pairs of cardinal and binary variables

PCA

Is a **statistical** technique whose **primary aim** is to **reduce** the dimensionality of a data set or. Another aim is to **interpret** the underlying structure of the data.

PCA **replaces** a set of correlated variables (*x*) by a much smaller number of uncorrelated 'new' variables, called components (*y*) , that **retain 'most'** of the information of the data set.

This is:

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

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

$$\vdots$$

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

How does it work?

- PCA includes 3 successive steps:

a) Computation of the principal components

   Find the 'a's through the *eigen* decomposition of the correlation matrix (spectral decomposition)

b) Extraction or selection of the number of components

c) Rotation of retained components to facilitate interpretation (sometimes)

What have we done?

Observed the debates presently active on associations

Looked at the use of "P" to identify similarity
to see which indicators are similar
which indicators are dissimilar

Pointed out that correlations and PCA/FA/MCA use all elements of the cross-tab. This can lead to diverse conclusions, which are influenced by relationships not related to similarity.

What might you do based on an analysis of associations?

- Drop or modify weights on highly associated indicators
- Combine some indicators into a sub-index
- Revise your 'justification' of indicators
- Adjust your categorization of indicators into dimensions.

Thank you