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

UNIVERSITY OF OXFORD

Understanding Associations Across Deprivation Indicators in MP

Research in-progress

Sabina Alkire & Paola Ballón
OPHI, University of Oxford

Oxford, November 23rd 2012

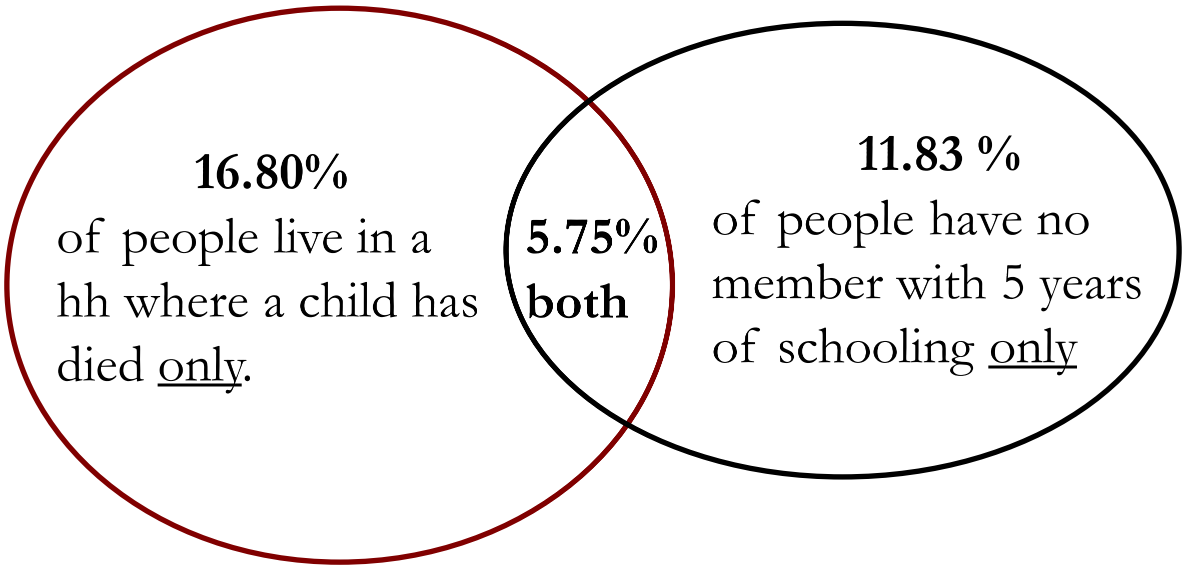
Tabita, Kenya · Rabiya, India · Stephanie, Madagascar · Agatha, Madagascar · Dalma, Kenya · Ann-Sophia, Kenya · Valérie, Madagascar

Why Joint Distribution Matters?

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

Raw headcount of mortality — 22.55%

Raw headcount of schooling — 17.58%



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

What implications does this have for a multidimensional measure?

Multidimensionality & Association

Debate:

Low association: to avoid redundancy

- HDI Debates

High association: to create stability

- Composite indicators
- Strong political message
- Techniques vary with data: PCA, MCA, FA, reliability, MD Scaling, Cluster, item response theory

Our practice to date

This Paper

The aim of this paper is to:

Consider, which techniques to use to assess similarity (strength) and association (strength and direction) of potential variables for inclusion in a multidimensional poverty index.

Clarify how to interpret them in the context of deprivation indicators (dichotomous variables) for a counting index.

Many techniques are surveyed and assessed which do not appear in this presentation.

1. Sources of information

Dichotomised deprivation scores, 0 or 1.

Raw headcounts → all deprivations

Censored headcounts → deprivations of the poor

The Contingency Table

Formally:

Years of Schooling	**Child mortality** 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

2. Traditional Measures of Association

Association (**affinity**) between two (or more) nominal (dichotomous) variables refers to a "coefficient" that measures the strength and direction(sign) of the relationship between the two variables.

Most coefficients of association define absence of association ("null" relationship) as independence.

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, Tschuprovw's T,* Lambda, and Uncertainty rely on the $\chi 2$ statistic..

2.A Cramer's *V* - Coefficient of Association

Cramer's *V* : popular because of its norming range for 0-1 variables

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

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

Meaning and interpretability of *V*

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

Reported in many tables in papers in this workshop

2.A Cramer's *V*

Sources of information used by 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, V uses "**entire** cross-tab"

What are the implications for MD poverty analysis?

Examples: Cramer V

Case I

Safe water (I)	**Child mortality (J)** Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**4** 40%	**2** 20%	6 60%
MD Poor = 1	**1** 10%	**3** 30%	4 40%
Total	5 50%	5 50%	10

$$V = \frac{n_{00}n_{11} - n_{01}n_{10}}{(n_{0+}n_{1+}n_{+0}n_{+1})^{1/2}} = \frac{4*3-1*2}{(5*6*5*4)^{1/2}} = + \ 0.41$$

Note the + value of V - both indicators move in the same direction

Ch Mort: 50%-50% (constant) ; Saf wat. 60% - 40% (decrease)

How sensitive V is to changes in the joint distribution?

Examples: Cramer V

Case II

Safe water (I)	Child mortality (J): Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	1 10%	3 30%	4 40%
MD Poor = 1	4 40%	2 **20%**	6 60%
Total	5 50%	5 50%	10

$$V = \frac{n_{00}n_{11} - n_{01}n_{10}}{(n_{0+}n_{1+}n_{+0}n_{+1})^{1/2}} = \frac{1*2-4*3}{(5*4*5*6)^{1/2}} = -0.41$$

Note the - value of V - both indicators move in opposite directions

Ch Mort: 50%-50% (**still constant**) ; Saf wat. 40% - 60% (**now increase**)

V does not reflect the change in 'poor-poor' cell

Examples: Cramer V

Case III: Absence of poverty (both indicators)

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**3** 30%	**3** 30%	6 **60**%
MD Poor = 1	**4** 40%	**0** **0%**	4 **40**%
Total	7 70%	3 30%	10

$$V = \frac{n_{00}n_{11-}n_{01}n_{10}}{(n_{0+}n_{1+}n_{+0}n_{+1})^{1/2}} = \frac{3*0-4*3}{(7*6*3*4)^{1/2}} = -0.53$$

Non-overlap leads to a CV= -0.53

Examples: Cramer V

Case IV: Absence of Non poverty (both indicators)

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	0 **0%**	**3** 30%	3 **30%**
MD Poor = 1	**4** 40%	3 **30%**	7 **70%**
Total	4 40%	6 60%	10

$$V = \frac{n_{00}n_{11} - n_{01}n_{10}}{(n_{0+}n_{1+}n_{+0}n_{+1})^{1/2}} = \frac{0 * 3 - 4 * 3}{(4 * 3 * 6 * 7)^{1/2}} = -0.53$$

Greater poor-poor leads to the <u>**same**</u> CV= -0.53

Conclusion: Insufficient for our purposes

2. Similarity Coefficients

There is an extensive list of binary similarity coefficients.

Hubalek (1982) surveys 43 similarity coefficients for binary/dichotomous data

Two simple and very intuitive ones are:

a) The Simple Matching Coefficient - *SM*
   Sokal & Sneath, (1963)

b) The Jaccard Coefficient – *J*
   Jaccard, (1901); Sneath, (1957)

2. Jaccard Similarity Coefficient

Meaning and interpretability

Counts the number of observations (households/individuals) which have the same status (only poor) in both variables

Strength of the relationship is defined as the proportion of "matches" in poverty **only**

Sources of information used by SM: Entire cross-tab

n_{00} number of people who are not MD poor

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

n joint distribution of matches and mismatches

$$J = \frac{n_{11}}{n - n_{oo}} \quad , \in [0,1]$$

What are the implications for MD poverty analysis?

Examples: J

Case I

Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	4 40%	2 20%	6 60%
MD Poor = 1	1 10%	3 30%	4 40%
Total	5 50%	5 50%	10

Child mortality (J) (column heading)

$$J = \frac{n_{11}}{n - n_{oo}} = \frac{3}{10 - 4} = 0.5$$

How sensitive these are to changes in the joint distribution?

Examples: J

Case III: Absence of poverty (both indicators)

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**3** 30%	**3** 30%	6 60%
MD Poor = 1	**4** 40%	**0** 0%	4 **40%**
Total	7 70%	3 **30%**	10

$$J = \frac{n_{11}}{n - n_{oo}} = \frac{0}{10-3} = 0$$

Note the levels of poverty: 30% in Ch. Mort; 40% in Safe water

Examples: J

Case IV: Absence of Non poverty (both indicators)

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**0** 0%	**3** 30%	3 30%
MD Poor = 1	**4** 40%	**3** 30%	7 **70**%
Total	4 40%	6 **60%**	10

$$J = \frac{n_{11}}{n - n_{oo}} = \frac{3}{10 - 0} = 0.3$$

Full non poverty leads to different J

What about the "levels"? These have increased, but J is not sensitive.

A: J = (2/(10-6))=50%
B: J = (1/(10-8))=50%

- Not sensitive to level;
- Not sensitive to overlap

A

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**6** 60%	**1** 10%	7 70%
MD Poor = 1	**1** 10%	**2** **20%**	3 30%
Total	7 70%	3 30%	10

B

	Child mortality (J)		
Safe water (I)	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**8** 0%	**0** 0%	8 80%
MD Poor = 1	**1** 10%	**1** **10%**	2 30%
Total	9 90%	**1** **10%**	10

An Alternative Measure "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:

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

Meaning and interpretability

Counts the number of observations (households/individuals) which have the same status (both poor or both deprived) in both variables, adjusted by the "level" of poverty

Strength of the relationship is defined as the proportion of "poverty matches" in the lowest level of poverty

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

Examples: P

Case I

Safe water (I)	**Child mortality (J)** Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	4 40%	**2** 20%	6 60%
MD Poor = 1	**1** 10%	**3** 30%	4 40%
Total	5 50%	5 50%	10

$$P = \frac{n_{11}}{\min[n_{1+}, n_{+1}]} = \frac{3}{\min[5,4]} = \frac{3}{4} = 0.75$$

50% of people are poor in Ch.Mort, 40% in safe water, 30% both

75% of poor people in Safe water are poor in both

How sensitive these are to changes in the joint distribution?

Examples: P

Case V

Safe water (I) \ **Child mortality (J)**	Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	4 40%	**3** 30%	7 70%
MD Poor = 1	**1** 10%	**2** 20%	**3** 30%
Total	5 50%	5 50%	10

$$P = \frac{n_{11}}{\min[n_{1+}, n_{+1}]} = \frac{2}{\min[5,3]} = \frac{2}{3} = 0.66$$

Decrease in the level of poverty

50% of people are poor in Ch.Mort, 30% in safe water, 20% both

66% of poor people in Safe water are poor in both

Examples: P

Case IV

Safe water (I)	Child mortality (J): Non MD poor = 0	MD poor = 1	Total
Non MD poor =0	**0** 0%	**3** 30%	3 30%
MD Poor = 1	**4** 40%	**3** 30%	7 **70**%
Total	4 40%	6 **60%**	10

$$P = \frac{n_{11}}{\min[n_{1+}, n_{+1}]} = \frac{3}{\min[6,7]} = \frac{3}{6} = 0.50$$

60% of people are poor in Ch.Mort, 70% in safe water, 30% both

50% of poor people in Ch.Mortality are poor in both

3. Illustration of "P" - Countries

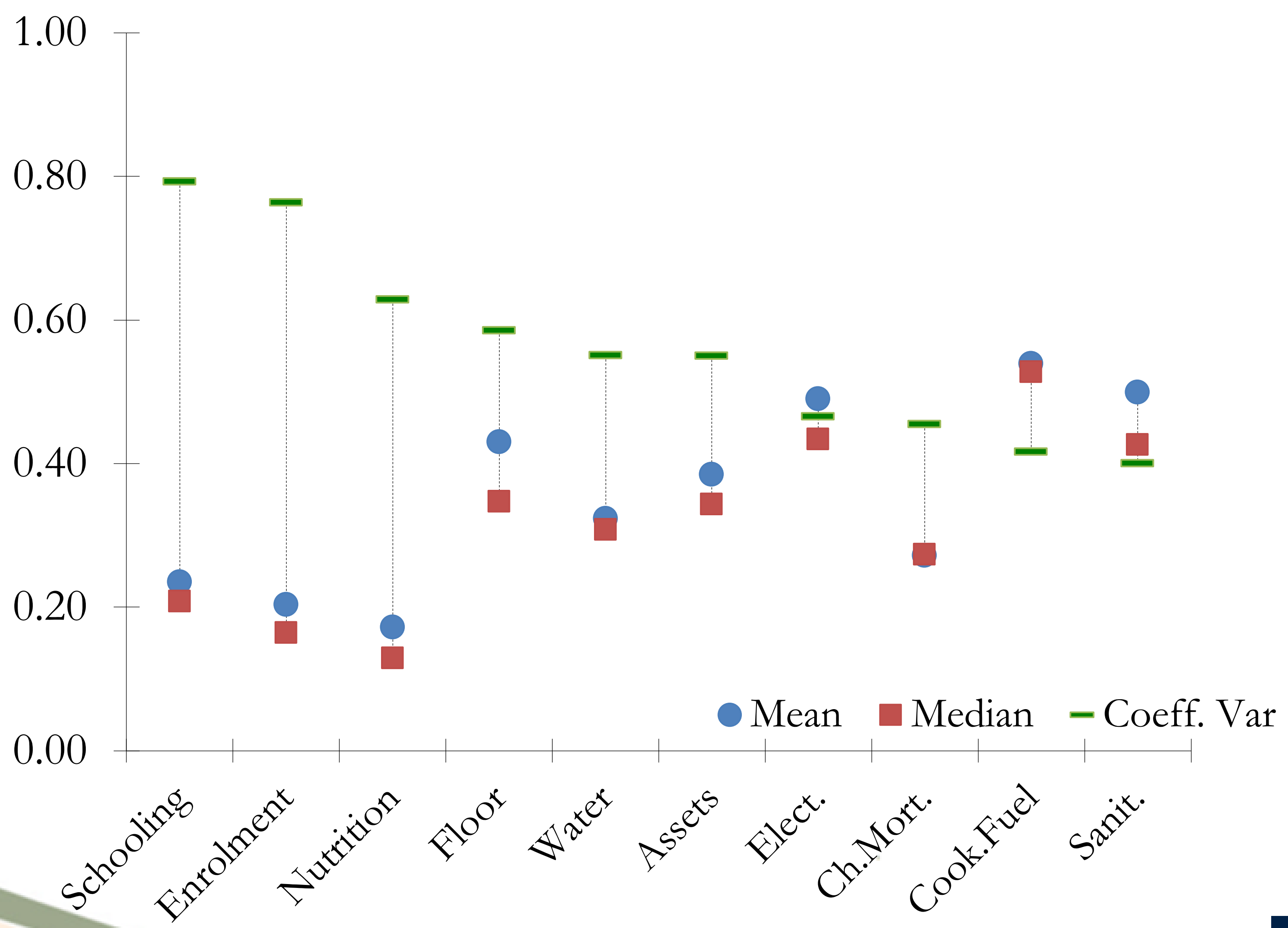
Country	DHS Year	Country	DHS Year
Bolivia	2008		
Ethiopia	2005	Namibia	2007
Gabon	2000	Nepal	2006
Ghana	2008	Nigeria	2008
Haiti	2006	Rwanda	2005
Kenya	2009	Swaziland	2007
Malawi	2004	Uganda	2006
Mali	2006	Zimbabwe	2006

Criteria of selection:

Information on all 10 censored headcount indicators

Variability across indicators

3. Censored Headcount Ratios



3. “P” Coefficient - Average over 15 countries

Indicator with the lowest Censored Headcount

"P" Coefficient (%)

	Sch.	Enrol.	Ch.Mort.	Nut.
Schooling		35	31	28
Enrolment	45		45	41
Ch.Mortality	51	54		46
Nutrition	39	37	53	

Coefficient of Variation of "P"

	Sch.	Enrol.	Ch.Mort.	Nut.
Schooling		0.49	0.38	0.61
Enrolment	0.43		0.28	0.44
Ch.Mortality	0.35	0.42		0.29
Nutrition	0.45	0.49	0.19	

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?

4. Concluding Remarks

Redundancy?

This **still** needs to be verified for a **larger** number of countries

This illustration considers countries with very similar profiles of deprivation/poverty

Our hypothesis:

If high values of P are found, we might need to:

Consider a restrained version of "acute poverty", and alternative weighs.

Thank you