OPHI

OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE

www.ophi.org.uk

UNIVERSITY OF OXFORD

Summer School on Multidimensional Poverty Analysis

11–23 August 2014

Oxford Department of International Development Queen Elizabeth House, University of Oxford

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

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

UNIVERSITY OF OXFORD

Associations across Deprivations

Sabina Alkire & Ana Vaz

11 – 23 August 2014
Oxford University, UK

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

Where we are…

We have defined:

- Purpose
- Unit of Analysis
- Dimensions

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

- Indicators
- Deprivation cutoffs

Today we will

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

- Reconsidering our selection of indicators
- Defining the categorization of indicators into Dimensions
- Defining 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
- Adjust your categorization of indicators into dimensions.

Multidimensionality & Association

View 1: High association favoured

- **Traditional composite marginal** measures
- 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, Giuo *et al.*)

Multidimensionality & Association

View 2: 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

Our view: not one or the other

- Value judgements are a fundamental element
- **If indicators are highly associated**, both may be retained for normative/policy reasons, or because their reduction over time differs
- **If indicators have a low association**, both may be retained if each is independently important

Sources of information

To study the "association"/similarity across deprivation indicators we will:

- Focus on dichotomised deprivation scores, 0 or 1.
- Use two different sources of information:
  - Uncensored deprivation scores
  - Censored deprivation scores

This class will:

- Explain limitations of correlation analysis
- Introduce measure of redundancy: measure of overlap

Definitions

Association for dichotomous variables - strength & direction

Similarity for dichotomous variables – strength

Similarity Coefficients in the Literature

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)

Describing Associations

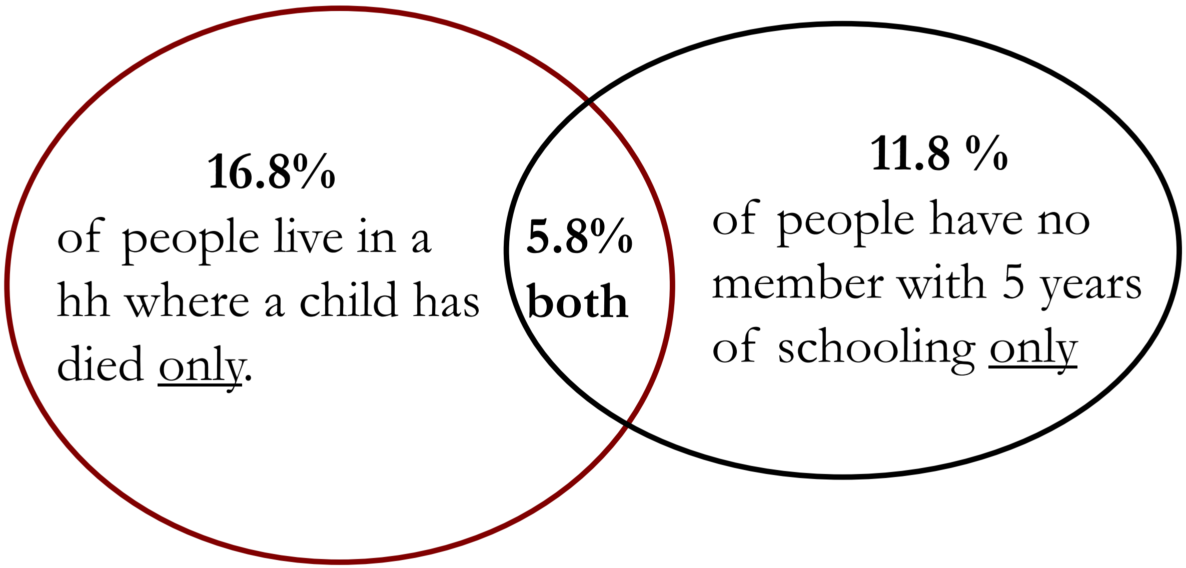
India NFHS data 2005-6 (sub-sample)

Raw headcount of child mortality

22.6%

Raw headcount of schooling

17.6%



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

The Contingency Table (Cross-tab)

When we are analysing two dichotomous variables…

	Child mortality		
Safe water	Non deprived = 0	Deprived = 1	Total
Non Deprived =0	4	2	6
Deprived = 1	1	3	4
Total	5	5	10

Headcount ratios: Safe water=50%, Child mortality= 40%

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

The Contingency Table

Formally:

Safe water (I)	Child mortality (J) Non deprived = 0	Deprived = 1	Total
Non deprived =0	n_{00}	n_{01}	n_{0+}
Deprived = 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

The Contingency Table

The contingency table gives information :

A) Joint distribution

Matches – two types

n_{00} number (percentage) of people who are not deprived

n_{11} number (percentage) of people who are deprived in both indicators

Mismatches – two types

n_{01} , n_{10} number (percentage) of people who are not deprived in one indicator but deprived in the other

B) Marginal distributions: headcount ratios n_{1+} , n_{+1}

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.

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

Correlation

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

The correlation is based on <u>all</u> of the elements of the cross-tab.

the raw headcount of each variable

the 'match' between deprivations

the 'match' between non-deprivations

the mismatches

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".

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

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.

Cramer's *V*

V uses "**entire** cross-tab"

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

Association is affected by:

- Extent to which deprivations between variables match (key)
- Values of the headcount ratios and their difference

Dilutes insights for redundancy.

Measure of Redundancy R^0

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:

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

Sources of information used by R^0:

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

n_{1+} , n_{+1} headcount ratios → Marginals

Redundancy: reflects the strength of the matches, but not the direction

Measure of Redundancy R^0

Meaning

Counts the number of observations which have the same status (both deprived/both poor) in both variables, adjusted by the "level" of deprivation (poverty for censored headcount)

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

This measure is sensitive to some distributional changes.

Interpreting R^0

If $R^0 = 90\%$, it shows that 90% of the people who are deprived in the indicator with the lowest headcount are also deprived in the other indicator.

This is a high association!

- That is not bad or good on its own – we need to think…
- Do we need both indicators or is one redundant?
- How do we justify keeping the two?
  - E.g. are they of independent value normatively or for monitoring purposes?

Example - Mozambique DHS

Case I

Years school. (I)	**School attendance (J)** Non deprived= 0	Deprived= 1	Total
Non deprived=0	47.15%	14.53%	61.68%
Deprived= 1	22.05%	16.27%	38.32%
Total	69.20%	30.80%	100%

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

Example – Mozambique & Bangladesh

Panel I: Mozambique		Attendance: Non deprived= 0	Attendance: Deprived=1	Total
Schooling	Non deprived=0	47.15%	14.52%	61.68%
	Deprived= 1	22.05%	**16.27%**	38.32%
	Total	69.20%	30.80%	100.00%

Panel II: Bangladesh		Attendance: Non deprived= 0	Attendance: Deprived=1	Total
Schooling	Non deprived=0	71.07%	9.43%	80.49%
	Deprived= 1	13.76%	**5.75%**	19.51%
	Total	84.82%	15.18%	100.00%

Example - Bangladesh DHS

Case I

Years school. (I)	**School attendance (J)** Non deprived= 0	Deprived= 1	Total
Non deprived=0	71.06%	9.43%	80.49%
Deprived= 1	13.76%	5.75%	19.51%
Total	84.82%	15.18%	100%

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

Two different countries with **completely different** patterns of deprivation show the **same association** coefficient **V**, but **different** measures of redundancy $\mathbf{R^0}$

Mozambique: Cramer's V vs. R^0

Correlation Matrix

	Schooling	Attendance	Safe water
Attendance	0.199	1.000	
Safe water	0.330	0.188	1.000
Cooking fuel	0.139	0.111	0.201

Overlap/Redundancy Measure

	Schooling	Attendance	Safe water
Attendance	0.529		
Safe water	0.776	0.708	
Cooking fuel	0.999	0.997	0.999

Mozambique: Cramer's V vs. R^0

Correlation Matrix

	Schooling	Attendance	Safe water
Attendance	0.199	1.000	
Safe water	0.330	0.188	1.000
Cooking fuel	0.139	0.111	0.201

Highest redundancy.
May suggest that cooking fuel is redundant, unless it is retained for other normative reasons.

Overlap/Redundancy

	Sc		
Attendance	[illegible]		
Safe water	0.776	0.708	
Cooking fuel	0.999	0.997	0.999

Association and Redundancy

Divergence reflects the different components of the cross-tab that they draw upon.

Measure of redundancy or overlap provides clear and precise information that should be considered when evaluating indicator redundancy

Multivariate Statistical Methods

- Multivariate techniques:
  - Principal component analysis (PCA),
  - Multiple correspondence analysis (MCA), and
  - Factor analysis (FA).

All three methods share a **common** view. This is to study the **association** (categorical variables) or **correlation** (cardinal variables) through a **multivariate input data matrix**, but they differ on the procedure use for that purpose

Input data matrices

Descriptive methods:

PCA: based on correlation or covariance matrix (cardinal)

MCA: based Burt or indicator tabulation (categorical)

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

Input matrix: 'correlation matrix' with:

pearson correlations for pairs of cardinal variables,
tetrachoric correlations for pais of binary variables,
biserial correlations for pairs of cardinal and binary variables

Consider assumptions re: shape of distribution

PCA

Is a **statistical** technique whose **primary aim** is to **reduce** the dimensionality of a data set. 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:

$$y_1 = a_{11}x_1 + a_{21}x_2 + \ldots + a_{d1}x_d$$
$$y_2 = a_{12}x_1 + a_{22}x_2 + \ldots + a_{d2}x_d$$
$$\vdots$$
$$y_d = a_{1d}x_1 + a_{2d}x_2 + \ldots + a_{dd}x_d$$

Reminder:

- 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)

Exercise

Analyse the relationships between your indicators:

a) Compute the cross-tabs
b) Compute Cramer's V
c) Compute the Measure of Redundancy R^0
d) Compare the measures (V-R^0) and interpret your results