OPHI

OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE

www.ophi.org.uk

UNIVERSITY OF OXFORD

Summer School on Multidimensional Poverty Analysis

1-13 August 2016

2016 Summer School, IPRCC, Beijing, China

Tabita, Kenya | Rabiya, India | Stéphanie, Madagascar | Agathe, 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 & Paola Ballon

Presented by Gisela Robles

OPHI

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

Associations across dimensions

To see which indicators have high / low associations
To identify ‘redundancy’

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
- Highly associated indicators 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

We will focus on:

- Dichotomised deprivation scores, 0 or 1.
- Joint distribution (cross tabulations)
- Could use two different sources of information:
  - **Uncensored deprivation scores**
  - Censored deprivation scores

This class will:

- Explain strengths and limitations of association analysis
- Introduce measure of redundancy: a measure of overlap/ similarity

Definitions

Association for dichotomous variables - strength & direction

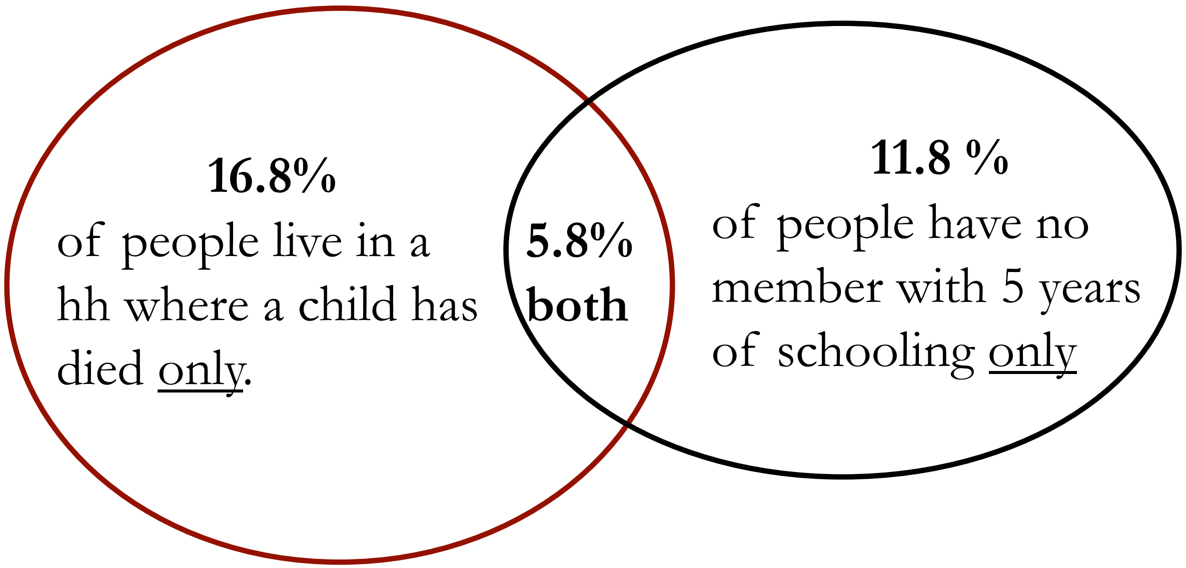
Similarity for dichotomous variables – strength

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…

Safe water	**Child mortality** 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=40%, Child mortality= 50%

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

The Contingency Table

Formally:

	Child mortality (J)		
Safe water (I)	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.

The Contingency Table

An independent distribution…

Safe water	Child mortality: Non deprived = 0	Deprived = 1	Total
Non Deprived =0	2.5	2.5	5
Deprived = 1	2.5	2.5	5
Total	5	5	10

Association

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".

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

Association

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".

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

For 0-1 variables, the Cramer's V measure is useful and similar to the correlation coefficient for continuous variables.

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.

Limitations of Association

Association of 0-1 deprivations, as 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

Cramer's V

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

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

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:

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

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

The Contingency Table

We divide the matches over the mínimum marginal headcount ratio:

	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

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%

$V = n_{00} n_{11} - n_{01} n_{10} / [n_{0+} n_{1+} n_{+0} n_{+1}]^{1/2} = 0.199$

$R_0 = n_{11} / \min[n_{1+}, n_{+1}] = 0.528$

Example – Mozambique & Bangladesh

Panel I: Mozambique		Attendance		
		Non deprived= 0	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	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%

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

$R^0 = \frac{n_{11}}{\min[n_{1+}, n_{+1}]} = 0.379$

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

Mozambique: Cramer's V vs. R^0

Association 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

Association 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 [illegible]

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

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

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

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

Preparation of data of the cross-table

```
foreach var1 in hh_d_school hh_d_nutri hh_d_water hh_d_assets {

    foreach var2 in hh_d_school hh_d_nutri hh_d_water hh_d_assets {


        /* Temporal variables for the matches and mismatches of two variables */
        generate temp_1 = (`var1'==0 & `var2'==0) if sample==1
        sum temp_1 [aw = weight]
        gen dd00 = r(mean)

        generate temp_2 = (`var1'==0 & `var2'==1) if sample==1
        sum temp_2 [aw = weight]
        gen dd01 = r(mean)

        generate temp_3 = (`var1'==1 & `var2'==0) if sample==1
        sum temp_3 [aw = weight]
        gen dd10 = r(mean)

        generate temp_4 = (`var1'==1 & `var2'==1) if sample==1
        sum temp_4 [aw = weight]
        gen dd11 = r(mean)


        /* Compute uncensored headcount ratios */
        sum `var1' [aw = weight] if sample==1
        gen h1_`var1' = r(mean)
        sum `var2' [aw = weight] if sample==1
        gen h2_`var2' = r(mean)

```

Cramer V and Redundancy

```
        /* Compute redundancy (coefficient P) */
        egen h_min_`var1'_`var2' = rowmin(h1_`var1' h2_`var2')
        gen P_`var1'_`var2' = dd11 / h_min_`var1'_`var2'


        /* Alternative way to compute Cramer Vs (with weights) from these temporal variables */
        gen CV_`var1'_`var2' = (dd00*dd11 - dd01 *dd10)/ sqrt(h1_`var1'*(1-h1_`var1')*h2_`var2'*(1-h2_`var2'))

        drop temp* dd* h1_* h2_* h_min*

        }


    /* Present results */
    sum P_* CV_*
    }
```