Associations across Deprivations

Sabina Alkire, Paola Ballón, Jose Manuel Roche, Ana Vaz
Washington DC 13 July 2013
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.
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): not one or the other**
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)

<table>
<thead>
<tr>
<th>Raw headcount of child mortality</th>
<th>Raw headcount of schooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>22.55%</td>
<td>17.58%</td>
</tr>
</tbody>
</table>

16.80% of people live in a hh where a child has died only.

5.75% both

11.83% of people have no member with 5 years of schooling only

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

<table>
<thead>
<tr>
<th>Safe water (I)</th>
<th>Child mortality (J)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non deprived = 0</td>
<td>Deprived = 1</td>
</tr>
<tr>
<td>Non Deprived =0</td>
<td>4 (67%, 80%)</td>
<td>2 (33%, 40%)</td>
</tr>
<tr>
<td>Water Deprived = 1</td>
<td>1 (25%, 20%)</td>
<td>3 (75%, 60%)</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

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%

<table>
<thead>
<tr>
<th>Safe water (I)</th>
<th>Child mortality (J)</th>
<th>Non deprived = 0</th>
<th>Deprived = 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non Deprived  = 0</td>
<td>4 (67%, 80%)</td>
<td>2 (33%, 40%)</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Water Deprived = 1</td>
<td>1 (25%, 20%)</td>
<td>3 (75%, 60%)</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>5</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

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:

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

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

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>35</td>
<td>31</td>
<td>28</td>
<td>89</td>
<td>93</td>
<td>61</td>
<td>81</td>
<td>97</td>
<td>80</td>
</tr>
<tr>
<td>Enrolment</td>
<td>45</td>
<td>45</td>
<td>41</td>
<td>85</td>
<td>88</td>
<td>57</td>
<td>79</td>
<td>94</td>
<td>68</td>
</tr>
<tr>
<td>Ch. Mortality</td>
<td>51</td>
<td>54</td>
<td>46</td>
<td>82</td>
<td>88</td>
<td>55</td>
<td>73</td>
<td>94</td>
<td>67</td>
</tr>
<tr>
<td>Nutrition</td>
<td>39</td>
<td>37</td>
<td>53</td>
<td>82</td>
<td>87</td>
<td>54</td>
<td>68</td>
<td>93</td>
<td>64</td>
</tr>
<tr>
<td>Elect.</td>
<td>39</td>
<td>37</td>
<td>53</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>93</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Sanit</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>96</td>
<td>99</td>
<td>94</td>
</tr>
<tr>
<td>Water</td>
<td>60</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>92</td>
<td>93</td>
<td>89</td>
<td>98</td>
<td>81</td>
</tr>
<tr>
<td>Floor</td>
<td>52</td>
<td>39</td>
<td>49</td>
<td>0</td>
<td>95</td>
<td>94</td>
<td>67</td>
<td>99</td>
<td>85</td>
</tr>
<tr>
<td>Fuel</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Assets</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td>39</td>
<td>93</td>
<td>94</td>
<td>69</td>
<td>89</td>
<td>98</td>
</tr>
</tbody>
</table>
3. What about Living Standard Indicators?

Let’s look at Fuel:

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Average Number of Countries</th>
<th>Coefficient Variation of P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>97</td>
<td>0.05</td>
</tr>
<tr>
<td>Enrolment</td>
<td>94</td>
<td>0.12</td>
</tr>
<tr>
<td>Ch. Mortality</td>
<td>94</td>
<td>0.10</td>
</tr>
<tr>
<td>Nutrition</td>
<td>93</td>
<td>0.12</td>
</tr>
<tr>
<td>Elect.</td>
<td>98</td>
<td>0.03</td>
</tr>
<tr>
<td>Sanit</td>
<td>99</td>
<td>0.01</td>
</tr>
<tr>
<td>Water</td>
<td>98</td>
<td>0.03</td>
</tr>
<tr>
<td>Floor</td>
<td>99</td>
<td>0.02</td>
</tr>
<tr>
<td>Assets</td>
<td>98</td>
<td>0.04</td>
</tr>
</tbody>
</table>

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?
- union will be higher than a censored $k$ value
- measures using intersection will be lower than 10%
- what are P values with other indicators? (meas. error?)
### Illustration: “P” Coefficient

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>35</td>
<td>31</td>
<td>28</td>
<td>89</td>
<td>93</td>
<td>61</td>
<td>81</td>
<td>97</td>
<td>80</td>
</tr>
<tr>
<td>Enrolment</td>
<td>45</td>
<td>45</td>
<td>41</td>
<td>85</td>
<td>88</td>
<td>57</td>
<td>79</td>
<td>94</td>
<td>68</td>
</tr>
<tr>
<td>Ch.Mortality</td>
<td>51</td>
<td>54</td>
<td>46</td>
<td>82</td>
<td>88</td>
<td>55</td>
<td>73</td>
<td>94</td>
<td>67</td>
</tr>
<tr>
<td>Nutrition</td>
<td>39</td>
<td>37</td>
<td>53</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>93</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Elect.</td>
<td>39</td>
<td>37</td>
<td>53</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>93</td>
<td>98</td>
<td>92</td>
</tr>
<tr>
<td>Sanit</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>95</td>
<td>0</td>
<td>96</td>
<td>99</td>
<td>94</td>
</tr>
<tr>
<td>Water</td>
<td>60</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>92</td>
<td>93</td>
<td>89</td>
<td>98</td>
<td>81</td>
</tr>
<tr>
<td>Floor</td>
<td>52</td>
<td>39</td>
<td>49</td>
<td>0</td>
<td>95</td>
<td>94</td>
<td>67</td>
<td>99</td>
<td>85</td>
</tr>
<tr>
<td>Fuel</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>96</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Assets</td>
<td>0</td>
<td>0</td>
<td>49</td>
<td>39</td>
<td>93</td>
<td>94</td>
<td>69</td>
<td>89</td>
<td>98</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Safe water</th>
<th>Child mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Non MD poor = 0</td>
</tr>
<tr>
<td>Non MD poor = 0</td>
<td>$n_{00}$</td>
</tr>
<tr>
<td>MD Poor = 1</td>
<td>$n_{10}$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_{+0}$</td>
</tr>
</tbody>
</table>

$n_{ij}$ are the cell count frequencies

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

$n = \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}$
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.
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), Tschuprov’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:

\[ V = \frac{\text{matches} \cdot \text{mismatches}}{\left( n_{00}n_{11} - n_{01}n_{10} \right)^{1/2}} \]

\[ \in [\mathbf{0,1}] \]

This is, correlations use the “entire cross-tab”

What are the implications for MD poverty analysis?
## Case I

### School attendance (J)

<table>
<thead>
<tr>
<th>Years school. (I)</th>
<th>Non deprived= 0</th>
<th>Deprived= 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non deprived=0</td>
<td>55,049</td>
<td>7,301</td>
<td>62,351</td>
</tr>
<tr>
<td></td>
<td>71%</td>
<td>9%</td>
<td>80%</td>
</tr>
<tr>
<td>Deprived= 1</td>
<td>10,657</td>
<td>4,455</td>
<td>15,112</td>
</tr>
<tr>
<td></td>
<td>14%</td>
<td>6%</td>
<td>20%</td>
</tr>
<tr>
<td>Total</td>
<td>65,706</td>
<td>11,756</td>
<td>77,463</td>
</tr>
<tr>
<td></td>
<td>85%</td>
<td>15%</td>
<td></td>
</tr>
</tbody>
</table>

\[ P = \frac{n_{11}}{\min[n_{1+}, n_{+1}]} = 0.379 \]

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

**Case I**

<table>
<thead>
<tr>
<th>Years school. (I)</th>
<th>Non deprived= 0</th>
<th>Deprived= 1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non deprived=0</td>
<td>28,722</td>
<td>8,845</td>
<td>37,567</td>
</tr>
<tr>
<td></td>
<td>47%</td>
<td>15%</td>
<td>62%</td>
</tr>
<tr>
<td>Deprived=1</td>
<td>13,431</td>
<td>9,913</td>
<td>23,344</td>
</tr>
<tr>
<td></td>
<td>22%</td>
<td>16%</td>
<td>38%</td>
</tr>
<tr>
<td>Total</td>
<td>42,153</td>
<td>18,758</td>
<td>60,911</td>
</tr>
<tr>
<td></td>
<td>69%</td>
<td>31%</td>
<td></td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.199</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>0.330</td>
<td>0.188</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Cooking fuel</td>
<td>0.139</td>
<td>0.111</td>
<td>0.201</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### “P” Measure

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.529</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>0.776</td>
<td>0.708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cooking fuel</td>
<td>0.999</td>
<td>0.997</td>
<td>0.999</td>
<td></td>
</tr>
</tbody>
</table>
## Correlation vs. “P” Measure

### Correlation Matrix

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.199</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>0.330</td>
<td>0.188</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Cooking fuel</td>
<td>0.139</td>
<td>0.111</td>
<td>0.201</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

### “P” Measure

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.529</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>0.776</td>
<td>0.708</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Cooking fuel</td>
<td>0.999</td>
<td>0.997</td>
<td>0.999</td>
<td>1.000</td>
</tr>
</tbody>
</table>

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

Which is right?
## Correlation vs. “P” Measure

### Correlations Matrix

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.199</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water</td>
<td>0.330</td>
<td>0.188</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>Cooking fuel</td>
<td>0.139</td>
<td>0.111</td>
<td>0.201</td>
<td>1.000</td>
</tr>
</tbody>
</table>

### “P” Measure Denominator (min value) is on the line

<table>
<thead>
<tr>
<th></th>
<th>Schooling</th>
<th>Enrolment</th>
<th>Water</th>
<th>Cooking fuel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schooling</td>
<td></td>
<td>0.776</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>Enrolment</td>
<td>0.529</td>
<td></td>
<td>0.708</td>
<td>0.997</td>
</tr>
<tr>
<td>Water</td>
<td></td>
<td></td>
<td></td>
<td>0.999</td>
</tr>
<tr>
<td>Cooking fuel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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