Standardizing data to analyse income inequality

Anthony Shorrocks, UNU-WIDER

United Nations, New York

6 May 2019

“Standardizing data” means …

- producing new data values and/or
- estimating new concepts

With a view towards providing **non-original** data that are:

- more reliable
- more useful
- more accessible to researchers

WIID attributes

Advantages:

- Coverage countries/timespan/sources
- Gini and **Lorenz** data

Disadvantages

- Variable quality/reliability
- Difficult to authenticate older records
- Too much extraneous detail: resource, pop coverage etc
- Missing item values: Gini (some) Lorenz (many)
- Countries/years missing observations
- Top shares underestimated (survey data)

WIID extensions

Improve quality and add new features

Overall strategic principles:

- Do no evil: original reliable observations kept or given high weight
- Transparency: each step documented and justified
- Reversibility: users can substitute alternative methods at each stage
- Gini values derived from Lorenz curve
- Gini and Lorenz values are consistent
- Observations not all equal: LIS preferred source
- Recognize not exact science: generated numbers are best estimate

WIID-EXTRA steps I

(1) Reorganize WIID into Lorenz format: share bottom 10%-80%, top 10%, 5%, 1%. Replace missing Lorenz values via “ungrouping” algorithm yielding synthetic income sample (any size) matching observed Lorenz pattern.

(2) Use synthetic sample to estimate Gini values. Useful to check:

- reliability of reported Gini values − some bad news
- efficacy of ungrouping algorithm − broadly encouraging

(3) Generate three **standardized** time series for each country/year reflecting:

- (1) net income (2) gross income (3) consumption
- distribution across persons of household resource per capita, based on LIS benchmark, full population and full geographical coverage.

WIID-EXTRA steps II

(4) Provide series on population and level of net income, gross income, and consumption (USD and PPP USD) to match the distribution series in (3).

(5) Construct regional and global income Lorenz curves and Gini values.

(6) Adjust top tail to match best estimates from other sources; recompute Lorenz values and Gini.

PLUS (?)

(7) Provide a globally representative synthetic income sample. Useful for:

- alternative inequality indices
- counterfactual experiments
- calculate income poverty indices under alternative scenarios