Statistical Inference:
Poverty Indices and Poverty Decompositions

Roy van der Weide

DECP - Poverty and Inequality Research Group
The World Bank

Washington DC, March 1, 2012
References


Outline

- What we try to estimate: Poverty indices
- Estimates based on survey data
  - Where does the randomness come from? Two approaches
    - A simple random sample
    - Survey weights
    - Stratification and clustering
    - Using imputed data with an application to Morocco
- Decompositions
  - Poverty profiles and additivity
  - Growth versus redistribution
  - The sectoral decomposition of a change in poverty
What we try to estimate: Poverty indices

Let $y_i$ denote household's $i$ real consumption or income, and let $z$ denote the poverty line.

- Poverty head-count:

$$P_0 = \frac{1}{N} \sum_{i=1}^{i=N} 1(y_i \leq z) = \frac{q}{N}, \quad (1)$$

where $N$ denotes the total number of households in the population, and $q$ the number of households below the poverty line.

- Poverty gap:

$$P_1 = \frac{1}{N} \sum_{i=1}^{i=N} \left(1 - \frac{y_i}{z}\right) 1(y_i \leq z) = \bar{g}(y, z) P_0 \quad (2)$$
What we try to estimate: Poverty indices

General Foster-Greer-Thorbecke (FGT) poverty measure:

\[
P_{\alpha} = \frac{1}{N} \sum_{i=1}^{i=N} p(y_i, z),
\]

where the household poverty measure \( p(y_i, z) \) takes the form:

\[
p(y_i, z) = \left(1 - \frac{y_i}{z}\right)^{\alpha} 1(y_i \leq z)
\]

- Head-count index for \( \alpha = 0 \)
- Poverty gap for \( \alpha = 1 \)
- Poverty severity for \( \alpha = 2 \)
Survey estimates: A simple random sample

• Let $x_i$ denote the variable of interest, and suppose we wish to estimate the population average:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{i=N} x_i$$

(5)

• Suppose we have a simple random sample of $n$ observations from the total population of size $N$

• The sample mean $\bar{x}$ is the obvious estimator of $\bar{X}$:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{i=n} x_i$$

(6)

• Note that the sample mean $\bar{x}$ is a random variable
Where does the randomness come from?

- “To make inference using survey data we need a framework for thinking about how the data were generated, which means thinking thinking about the population from which the data came and about how data collection induces randomness into our sample” (Deaton, 2000)

- Finite populations approach (citations from Deaton 2000):
  - “The quantity of interest ... is a fixed number that could be measured with perfect accuracy from a census”
  - “No assumptions are made about the distribution of income”

- Superpopulations approach (citations from Deaton 2000):
  - Here, “we are less interested in the actual population ... regarding it as only one of many possible populations that might have existed”
  - The quantity of interest here is “the statistical law or economic process that generated income in the superpopulation”
Finite populations approach: A simple random sample

- Let $a_i$ be a random variable that indicates whether $i$ is in the sample, taking the value of 1 if so, and 0 otherwise.

- The sample mean $\bar{x}$ may then be rewritten as:

\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{i=N} a_i x_i \] (7)

- For a simple random sample (without replacement), we have:

\[ a_i = \begin{cases} 
1 & w.p. \frac{n}{N} \\
0 & w.p. \ 1 - \frac{n}{N} 
\end{cases} \] (8)

- We have $E[a_i] = \frac{n}{N}$, $var[a_i] = \frac{n}{N} \left( 1 - \frac{n}{N} \right)$, and $cov[a_i, a_j] = -\frac{n}{N(N-1)} \left( \frac{N-n}{N} \right)$.
Finite populations approach: A simple random sample

• The sample mean $\bar{x}$ is unbiased:

$$E[\bar{x}] = \frac{1}{n} \sum_{i=1}^{N} E[a_i] x_i$$

$$= \frac{1}{n} \sum_{i=1}^{N} \left( \frac{n}{N} \right) x_i$$

$$= \frac{1}{N} \sum_{i=1}^{N} x_i$$

• The variance solves:

$$var[\bar{x}] = \frac{1 - f}{n} S^2,$$  \hspace{1cm} (12)

where $1 - f = \frac{N-n}{N}$ is the finite population correction, and where

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{X})^2$$

\hspace{1cm} (13)
Superpopulations approach: A simple random sample

- Let $x_i$ be an independently and identically distributed random variable with mean $\mu$ and variance $\sigma^2$

- The sample mean is an unbiased estimator of $\mu$:

$$E[\bar{x}] = \frac{1}{n} \sum_{i=1}^{N} E[x_i]$$

$$= \frac{1}{n} \sum_{i=1}^{N} \mu = \mu \quad (14)$$

$$= \frac{1}{n} \sum_{i=1}^{N} \mu = \mu$$

$$= \frac{1}{n} \sum_{i=1}^{N} \mu = \mu \quad (15)$$

- The variance solves:

$$var[\bar{x}] = \frac{1}{n^2} \sum_{i=1}^{N} var[x_i]$$

$$= \left( \frac{1}{n^2} \right) n\sigma^2 = \frac{\sigma^2}{n} \quad (17)$$
Survey weights

• In most surveys different households have different probabilities of being selected into the sample

• “Depending on the purpose of the survey, some type of households are overrepresented relative to others, either deliberately as part of the design, or accidentally, for example because of differential response” (Deaton, 2000)

• Whether deliberate or accidental, “if the different types of households are different, [unadjusted] sample means will be biased estimators of population means” (Deaton, 2000)

• “To undo this bias, the sample data are “reweighted” to make them representative of the population” (Deaton, 2000)
Survey weights

• Let $\pi_i$ denote the probability household $i$ is selected at each draw (with replacement), and let $n$ be the total number of draws.

• The probability that $i$ is included in the sample is approx. $n\pi_i$.

• Define for each household the survey weight: $w_i = \frac{1}{n\pi_i}$.

• $w_i$ may be thought of as the number of population individuals represented by $i$.

• The sum of the survey weights may therefore be considered an estimate of the population size $N$:

$$\hat{N} = \sum_{i=1}^{i=n} w_i$$

(18)

• Let us assume that $N$ is known so that weights can be set to satisfy:

$$N = \sum_{i=1}^{i=n} w_i$$
Survey weights

• The probability-weighted sample mean $\bar{x}_w$ is given by:

$$\bar{x}_w = \frac{1}{N} \sum_{i=1}^{n} w_i x_i$$  \hspace{1cm} (19)

• Let $t_i$ be the number of times $i$ shows up in the sample

• The weighted sample mean may be rewritten as:

$$\bar{x}_w = \frac{1}{N} \sum_{i=1}^{N} t_i w_i x_i$$  \hspace{1cm} (20)

• $t_i$ is drawn from a multi-nominal distribution with: $E[t_i] = n\pi_i = \frac{1}{w_i}$, $var[t_i] = n\pi_i(1 - \pi_i)$, and $cov[t_i, t_j] = -n\pi_i\pi_j$
Survey weights

• The sample mean $\bar{x}_w$ is a consistent estimator for $\bar{X}$:

$$E[\bar{x}_w] = \frac{1}{N} \sum_{i=1}^{N} E[t_i] w_i x_i = \frac{1}{N} \sum_{i=1}^{N} x_i$$

(21)

• An estimate for the variance of $\bar{x}_w$ is (see Deaton, 2000):

$$var[\bar{x}_w] \approx \frac{n}{n-1} \sum_{i=1}^{n} \left( \frac{w_i}{N} \right)^2 (x_i - \bar{x}_w)^2$$

(22)

• Note that $\hat{N}$ is a consistent estimator for $N$:

$$E[\hat{N}] = \sum_{i=1}^{N} E[t_i] w_i = \sum_{i=1}^{N} \left( \frac{1}{w_i} \right) w_i = N$$

(23)
Stratification

- Here we break up a single survey into multiple independent surveys, one for each stratum.
- Let $N_s$ and $X_s$ denote the population size and mean for strata $s$.
- The population mean then solves:

$$\bar{X} = \sum_{s=1}^{S} \left( \frac{N_s}{N} \right) X_s,$$

which can be estimated by:

$$\bar{x} = \sum_{s=1}^{S} \left( \frac{N_s}{N} \right) \bar{x}_s,$$

- Without stratification, “the variability of the estimate will not only have a component from the variability of the stratum means ... but also a component from the variability of the fractions in each stratum” (Deaton, 2000).
Stratification

- In a simple random sample, the sample mean $\tilde{x}$ is given by:

$$\tilde{x} = \sum_{s=1}^{S} \left( \frac{n_s}{n} \right) \bar{x}_s$$

(26)

- Define $\alpha_s = N_s / N$, and $\hat{\alpha}_s = n_s / n$

- Substituting $\bar{x}_s = \bar{x} + (\bar{x}_s - \bar{x})$ and $\hat{\alpha}_s = \alpha_s + (\hat{\alpha}_s - \alpha)$ yields:

$$\tilde{x} = \bar{x} + \sum_{s=1}^{S} (\hat{\alpha}_s - \alpha)(\bar{x}_s - \bar{x})$$

(27)

- For the variance of $\tilde{x}$ this implies:

$$\text{var}[\tilde{x}] = \text{var}[\bar{x}] + \text{var}\left[ \sum_s (\hat{\alpha}_s - \alpha)(\bar{x}_s - \bar{x}) \right]$$

(28)

$$\approx \text{var}[\bar{x}] + \frac{1}{n} \sum_{s=1}^{S} \left( \frac{N_s}{N} \right) (\bar{x}_s - \bar{x})^2$$

(29)
Two-stage sampling: Clusters

- Survey data is often collected in two stages, first sampling clusters, and then sampling households from each cluster.

- Important to not treat a two-stage sample as a simple random sample: “the use of standard formulas can seriously overstate the precision of the estimates” (Deaton, 2000).

- Trade-off between statistical precision and costs of data collection.

- Assume there are an infinite number of households (living in an infinite number of clusters).

- Suppose that $x_i$ can be described by:

$$x_{ci} = \mu + \eta_c + \varepsilon_{ci},$$

(30)

where $ci$ denotes household $i$ living in cluster $c$. 
Two-stage sampling: Clusters

• Assume $\eta_c$ and $\varepsilon_{ci}$ are independent random variables with $E[\eta_c] = E[\varepsilon_{ci}] = 0$, $\text{var}[\eta_c] = \sigma^2_\eta$, and $\text{var}[\varepsilon_{ci}] = \sigma^2_\varepsilon$

• Define $\rho = \frac{\sigma^2_\eta}{\sigma^2_\eta + \sigma^2_\varepsilon}$

• Let $n$ denote the number of sampled clusters, and $m$ the number of sampled households per cluster

• It then follows that $E[\bar{x}] = \mu$, and:

$$\text{var}[\bar{x}] = \frac{\sigma^2_\eta}{n} + \frac{\sigma^2_\varepsilon}{nm} = \frac{\sigma^2_\eta + \sigma^2_\varepsilon}{nm} (1 + (m - 1)\rho) \quad (31)$$
Some practical issues I

- Is the master-sample complete?
  - In Vietnam for example, it is believed that migrants are not adequately covered by the master sample
  - Collectives of migrant workers sharing accommodation do not always make up officially registered households
  - As a result, the survey may be under-representing people employed in labour intensive industries

- Is the master-sample up-to-date?

- Is non-response an issue?
  - See e.g. Korinek, Mistiaen and Ravallion (2005)
Survey estimates using imputed data: Poverty

• Let the probability that \( i \) is poor be denoted by:

\[
\pi_i = 1(\dot z_i^T \theta + \varepsilon_i \leq 0)
\]

• Let \( F \) denote the distribution function of \( \varepsilon_i \)

• Now consider \( \hat \pi_i = F(-\dot z_i^T \hat \theta) \) as the imputed value for \( \pi_i \)
  – \( \pi \) will be imputed in **sample 1** which is of size \( n_1 \)
  – The model for \( \pi \) is estimated using **sample 2** which is of size \( n_2 \)
  – Both samples contain the independent variables \( z \)

• Consider \( \hat \mu \) as an estimator for head-count poverty \( \mu = E[\pi_i] \):

\[
\hat \mu = \frac{1}{n_1} \sum_i F(-\dot z_i^T \hat \theta)
\]
Survey estimates using imputed data: Poverty

• Note that $\hat{\mu}$ is subject to both sampling error and model error

• If $\hat{\theta}$ is a consistent estimator of $\theta$, then $\hat{\mu}$ is a consistent estimator of $\mu$

• Let $\Omega_\theta$ denote the asymptotic variance of $\hat{\theta}$

• Under standard assumptions, $\hat{V}_\mu$ provides a consistent estimate of the variance of $\hat{\mu}$:

$$\hat{V}_\mu = \frac{1}{n_1} svar[\hat{\pi}_i] + \frac{1}{n_2} q^T \hat{\Omega}_\theta q,$$

where $q = \frac{1}{n_1} \sum_i f(-z_i^T \hat{\theta}) z_i$ (where $f$ denotes the derivative of $F$)
Empirical example from Morocco: Background

- In the face of both negative and positive shocks, has poverty in Morocco increased or declined since 2007?

- On the one hand, recent years have seen a global crisis unfold and large swings in international food prices

- On the other hand, recently Morocco has been fortunate with advantageous rainfall which has boosted agricultural production
Empirical example from Morocco: Background

MASI Stock Market Index

0 5000 10000 15000


\( t \)
Empirical example from Morocco: Background
Empirical example from Morocco: Background

Morocco Head-Count Poverty (2001 & 2007)
Empirical example from Morocco: Background

• Unfortunately, poverty rates cannot be estimated directly from household consumption survey data since the last survey is from 2007.

• To fill this gap we propose to estimate poverty by imputing household consumption data into the Labor Force Survey (LFS).

• This approach offers quarterly poverty estimates that may be disaggregated by urban/rural.
Empirical example from Morocco: Modeling details

- **Key assumption for this application:** The model is not subject to change over time

- Exercise implicitly also an evaluation of comparability of consumption measures, comparability of LSMS vs LFS, and of temporal price adjustments

- Model urban and rural (and years) separately
  - Include domain/region fixed effects
  - 2001 and 2007 models are (slightly) different
  - Employment in finance matters in urban, employment in agriculture in rural Morocco

- Imputed consumption will be in constant prices (either 2001 or 2007 prices), i.e. the poverty line too will be constant over time
Empirical example from Morocco: Modeling details

- Consumption model descriptives

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>adj-$R^2$</td>
<td>0.59</td>
<td>0.43</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>number of vars</td>
<td>52</td>
<td>45</td>
<td>58</td>
<td>51</td>
</tr>
<tr>
<td>observations</td>
<td>7888</td>
<td>6355</td>
<td>4266</td>
<td>2796</td>
</tr>
</tbody>
</table>
Empirical example from Morocco: Results

- Imputing into the LSMS (a first test)

<table>
<thead>
<tr>
<th>Year</th>
<th>“true”</th>
<th>2001-M</th>
<th>2007-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>15.3</td>
<td>16.0</td>
<td>16.0</td>
</tr>
<tr>
<td>2007</td>
<td>8.9</td>
<td>9.6</td>
<td>9.9</td>
</tr>
</tbody>
</table>

- Imputing into the LFS (a second test)

<table>
<thead>
<tr>
<th>Year</th>
<th>“true”</th>
<th>2001-M</th>
<th>2007-M</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>15.3</td>
<td>15.2</td>
<td>15.1</td>
</tr>
<tr>
<td>2007</td>
<td>8.9</td>
<td>8.4</td>
<td>8.6</td>
</tr>
</tbody>
</table>
Empirical example from Morocco: Results

Quarterly Poverty Estimates

- 2007 model
- 2001 model

The chart shows the quarterly poverty estimates from 2000 to 2010, with two modeled lines representing different years.
Empirical example from Morocco: Urban-Rural

[Graph showing data on urban and rural poverty in Morocco, with trends from 2001 to 2009.]

mu_U, CI_U, CI_U, mu_R, CI_R, CI_R
Empirical example from Morocco: Region I

![Graph showing time series data for Region I in Morocco]
Empirical example from Morocco: Region II
Empirical example from Morocco: Region III
Decompositions: Poverty profiles and additivity

- Poverty profiles are essentially a cross-tabulation of poverty with selected regions, sectors, household characteristics etc.

- Helps to identify lagging regions/sectors; A tool for targeting

- ”Poverty profiles can be extremely useful in assessing how the sectoral or regional pattern of economic change is likely to affect aggregate poverty” (Ravallion, 1994)

- A poverty measure satisfies ”additivity” if the following holds:

\[ P = \sum_j \left( \frac{n_j}{n} \right) P_j \]  \hspace{1cm} (32)

- ”Additivity guarantees ”sub-group consistency” in that when poverty increases (decreases) in any sub-group of the population, aggregate poverty will also increase (decrease)” (Ravallion, 1994)
Decompositions: Poverty profiles and additivity

• “One possible objection of additivity is that it attaches no weight to one aspect of a poverty profile: the inequality between sub-groups in the extent of poverty” (Ravallion, 1994)

• An example with two equally sized sectors: urban and rural
  – Initial urban and rural poverty is 20% and 70%, respectively, with aggregate poverty equal to 45%
  – There is a choice between two policies: X and Y
  – Under X, urban poverty declines to 10%, rural poverty remains 70%
  – Under Y, urban poverty stays at 20%, while rural poverty is reduced to 60%
  – Both policies yield an aggregate poverty of 40%
  – Policy Y, however, also reduces urban-rural inequality
Some practical issues II

• When comparing poverty estimates across regions and sectors, take standard errors into consideration.

• Selected sub-groups may be “statistically invisible” when using survey data alone, think of ethnic minorities, disabled etc.

• An accurate account of the geographic variation in prices is particularly important when comparing poverty across regions.
Decompositions: Growth vs redistribution

- Let average real consumption or income be denoted by $E[y] = \mu$, and let $L$ denote the Lorenz curve.

- Let us define aggregate poverty as a function of $\mu$, $L$, and the poverty line $z$: $P(z, \mu, L)$.

- The change in poverty can then be decomposed as follows:

$$ P_{t+1} - P_t = G_{t,t+1}(r) + D_{t,t+1}(r) + R_{t,t+1}(r), \quad (33) $$

where:

$$ G_{t,t+1}(r) = P(z, \mu_{t+1}, L_r) - P(z, \mu_t, L_r) \quad (34) $$

$$ D_{t,t+1}(r) = P(z, \mu_r, L_{t+1}) - P(z, \mu_r, L_t), \quad (35) $$

and where $R_{t,t+1}(r)$ denotes a residual term.
The sectoral decomposition of a change in poverty

- Let $P_{it}$ denote an additive poverty measure (such as the FGT measure) for sector $i$ with population share $n_{it}$ at time $t$

- The change in poverty may then be decomposed as follows:

$$P_{t+1} - P_t = \sum_i (P_{i,t+1} - P_{it})n_{it}$$  \hspace{1cm} (36)

$$+ \sum_i (n_{i,t+1} - n_{it})P_{it}$$  \hspace{1cm} (37)

$$+ \sum_i (P_{i,t+1} - P_{it})(n_{i,t+1} - n_{it})$$  \hspace{1cm} (38)

- The last term arises from possible correlation between sectoral gains and population shifts; "the sign of the interaction effect tells us whether people tended to switch to the sectors where poverty was falling or not" (Ravallion, 1994)