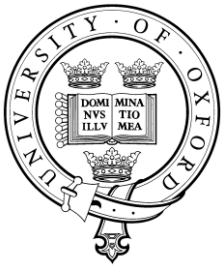
Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford



OPHI WORKING PAPER NO. 89

Multidimensional Poverty Measurement and Analysis: Chapter 8 – Robustness Analysis and Statistical Inference

Sabina Alkire*, James E. Foster**, Suman Seth***, Maria Emma Santos***, Jose M. Roche**** and Paola Ballon*****

February 2015

Abstract
The design of a poverty measure involves the selection of a set of parameters and poverty figures. In most cases the measures are estimated from sample surveys. This raises the question of how conclusive particular poverty comparisons are subject to both the set of selected parameters (or variations within a plausible range) and the sample datasets. This

* Sabina Alkire: Oxford Poverty & Human Development Initiative, Oxford Department of International Development, University of Oxford, 3 Mansfield Road, Oxford OX1 3TB, UK, +44-1865-271915, sabina.alkire@qeh.ox.ac.uk
** James E. Foster: Professor of Economics and International Affairs, Elliott School of International Affairs, 1957 E Street, NW, fosterje@gwu.edu.
*** Suman Seth: Oxford Poverty & Human Development Initiative (OPHI), Queen Elizabeth House (QEH), Department of International Development, University of Oxford, UK, +44 1865 618643, suman.seth@qeh.ox.ac.uk.
**** Maria Emma Santos: Instituto de Investigaciones Económicas y Sociales del Sur (IIES,), Departamento de Economía, Universidad Nacional del Sur (UNS) - Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), 12 de Octubre 1198, 7 Piso, 8000 Bahía Blanca, Argentina. Oxford Poverty and Human Development Initiative, University of Oxford. msantos@uns.edu.ar; maria.santos@qeh.ox.ac.uk.
***** Jose Manuel Roche: Save the Children UK, 1 St John's Lane, London EC1M 4AR, j.roche@savethechildren.org.uk.
****** Paola Ballon: Assistant Professor of Econometrics, Department of Economics, Universidad del Pacifico, Lima, Peru; Research Associate, OPHI, Department of International Development, Oxford University, Oxford, U.K, pm.ballonf@up.edu.pe.

This study has been prepared within the OPHI theme on multidimensional measurement.

OPHI gratefully acknowledges support from the German Federal Ministry for Economic Cooperation and Development (BMZ), Praus, national offices of the United Nations Development Programme (UNDP), national governments, the International Food Policy Research Institute (IFPRI), and private benefactors. For their past support OPHI acknowledges the UK Economic and Social Research Council (ESRC)/(DFID) Joint Scheme, the Robertson Foundation, the John Fell Oxford University Press (OUP) Research Fund, the Human Development Report Office (HDRO/UNDP), the International Development Research Council (IDRC) of Canada, the Canadian International Development Agency (CIDA), the UK Department of International Development (DFID), and AusAID.

ISSN 2040-8188 ISBN 978-19-0719-476-4

chapter shows how to apply dominance and rank robustness tests to assess comparisons as poverty cutoffs and other parameters changes. It presents ingredients of statistical inference, including standard errors, confidence intervals, and hypothesis tests. And it discusses how robustness and statistical inference tools can be used together to assert concrete policy conclusions. An appendix presents methods for computing standard errors, including the bootstrapped standard errors.

Keywords: robustness analysis, statistical inference, dominance analysis, rank robustness, standard errors, bootstrap

JEL classification: C10, C12, I32

Acknowledgements
We received very helpful comments, corrections, improvements, and suggestions from many across the years. We are also grateful for direct comments on this working paper from Tony Atkinson and Gaston Yalonetzky.

Citation: Alkire, S., Foster, J. E., Seth, S., Santos, M. E., Roche, J. M., and Ballon, P. (2015). *Multidimensional Poverty Measurement and Analysis*, Oxford: Oxford University Press, ch. 8.

The Oxford Poverty and Human Development Initiative (OPHI) is a research centre within the Oxford Department of International Development, Queen Elizabeth House, at the University of Oxford. Led by Sabina Alkire, OPHI aspires to build and advance a more systematic methodological and economic framework for reducing multidimensional poverty, grounded in people's experiences and values.

Oxford Poverty & Human Development Initiative (OPHI)
Oxford Department of International Development
Queen Elizabeth House (QEH), University of Oxford
3 Mansfield Road, Oxford OX1 3TB, UK
Tel. +44 (0)1865 271915 Fax +44 (0)1865 281801
ophi@qeh.ox.ac.uk http://www.ophi.org.uk

8 Robustness Analysis and Statistical Inference

Chapter 5 presented the methodology for the Adjusted Headcount Ratio poverty index M_0 and its different partial indices; Chapter 6 discussed how to design multidimensional poverty measures using this methodology in order to advance poverty reduction; and Chapter 7 explained novel empirical techniques required during implementation. Throughout, we have discussed how the index and its partial indices may be used for policy analysis and decision-making. For example, a central government may want to allocate resources to reduce poverty across its subnational regions or may want to claim credit for strong improvement in the situation of poor people using an implementation of the Adjusted Headcount Ratio. One is, however, entitled to question how conclusive any particular poverty comparisons are for two different reasons.

One reason is that the design of a poverty measure involves the selection of a set of parameters, and one may ask how sensitive policy prescriptions are to these parameter choices. Any comparison or ranking based on a particular poverty measure may alter when a different set of parameters, such as the poverty cutoff, deprivation cutoffs or weights, is used. We define an ordering as robust with respect to a particular parameter when the order is maintained despite a change in that parameter.[1] The ordering can refer to the poverty ordering of two aggregate entities, say two countries or other geographical entities, which is a *pairwise* comparison, but it can also refer to the order of more than two entities, what we refer to as a ranking. Clearly, the robustness of a ranking (of several entities) depends on the robustness of all possible pairwise comparisons. Thus, the robustness of poverty comparisons should be assessed for different, but reasonable, specifications of parameters. In many circumstances, the policy-relevant comparisons should be robust to a range of plausible parameter specifications. This process is referred as **robustness analysis**. There are different ways in which the robustness of an ordering can be assessed. This chapter presents the most widely implemented analyses; new procedures and tests may be developed in the near future.

The second reason for questioning claimed poverty comparisons is that poverty figures in most cases are estimated from sample surveys for drawing inferences about a

[1] This chapter is confined to assessing the robustness of rank ordering across groups. Naturally it is essential also to assess the sensitivity of key values (such as the values of M_0 and dimensional contributions) to parameter changes, in situations in which policies use these cardinal values.

population. Thus, it is crucial that inferential errors are also estimated and reported. This process of drawing conclusions about the population from the data that are subject to random variation is referred as **statistical inference**. Inferential errors affect the degree of certainty with which two and more entities may be compared in terms of poverty for a particular set of parameters' values. Essentially, the difference in poverty levels between two entities – states for example – may or may not be statistically significant. Statistical inference affects not only the poverty comparisons for a particular set of parameter values but also the *robustness* of such comparisons for a *range* of parameters' values.

In general, assessments of robustness should cohere with a measure's policy use. If the policy depends on levels of M_0, then the robustness of the respective levels (or ranks) of poverty should be the subject of robustness tests presented here. If the policy uses information on the dimensional composition of poverty, robustness tests should assess these—which lie beyond the scope of this chapter, but see Ura et al. (2012). Recall also from Chapter 6 people's values may generate plausible ranges of parameters. Robustness tests clarify the extent to which the same policies would be supported across that relevant range of parameters. In this way, robustness tests can be used for building consensus or for clarifying which points of dissensus have important policy implications.

This chapter is divided into two sections. Section 8.1 presents a number of useful tools for conducting different types of robustness analysis; section 8.2 presents various techniques for drawing statistical inferences and section 8.3 presents some ways in which the two types of techniques can be brought together.

8.1 Robustness Analysis

In monetary poverty measures, the parameters include (a) the set of indicators (components of income or consumption); (b) the price vectors used to construct the aggregate as well as any adjustments such as for inflation or urban/rural price differentials; (c) the poverty line; and (d) equivalence scales (if applied). The parameters that influence the multidimensional poverty estimates and poverty comparisons based on the Adjusted Headcount Ratio are (i) the set of indicators (denoted by subscript $j = 1, \ldots, d$); (ii) the set of deprivation cutoffs (denoted by vector z); (iii) the set of weights or deprivation values (denoted by vector w); and (iv) the poverty cutoff (denoted by k). A change in these parameters may affect the overall poverty estimate or comparisons across regions or countries.

This section introduces tools that can be used to test the robustness of pairwise comparisons as well as the robustness of overall rankings with respect to the initial choice of the parameters. We first introduce a tool to test the robustness of pairwise comparisons with respect to the choice of the poverty cutoff. This tool tests an extreme form of robustness, borrowing from the concept of stochastic dominance in the single-dimensional context (section 3.3.1).[2] When dominance conditions are satisfied, the strongest possible results are obtained. However, as dominance conditions are highly stringent and dominance tests may not hold for a large number of the pairwise comparisons, we present additional tools for assessing the robustness of country rankings using the correlation between different rankings. This second set of tools can be used with changes in any of the other parameters too, namely, weights, indicators and deprivation cutoffs.

8.1.1 Dominance Analysis for Changes in the Poverty Cutoff

Although measurement design begins with the selection of indicators, weights, and deprivation cutoffs, we begin our robustness analysis by assessing dominance with respect to changes in the poverty cutoff, which is applied to the weighted deprivation scores constructed using other parameters. We do this because as in the unidimensional context, it is the poverty cutoff that finally identifies who is poor, thereby defining the 'headcount ratio' and effectively setting the level of poverty. It is arguably most visibly debated.[3] We have introduced the concept of stochastic dominance in the uni- and multidimensional context in section 3.3.1. This part of the chapter builds on that concept and technique, focusing primarily on the **first-order stochastic dominance (FSD)** and showing how it can be applied to identify any unambiguous comparisons with respect to the poverty cutoff for our two most widely used poverty measures—Adjusted Headcount Ratio (M_0) and Multidimensional Headcount Ratio (H). Recall from section 3.3.1 the notation of two univariate distributions of achievements x and y with **cumulative distribution functions (CDF)** F_x and F_y, where $F_x(b)$ and $F_y(b)$ are the

[2] There is a well-developed literature on robustness and sensitivity analyses for composite indices rankings with respect to relative weights, normalization methods, aggregation methods, and measurement errors. See Nardo et al. (2005), Saisana et al. (2005), Cherchye et al. (2007), Cherchye et al. (2008), Foster, McGillivray and Seth (2009, 2013), Permanyer (2011, 2012), Wolff et al. (2011), and Høyland et al. (2012). These techniques may require adaptation to apply to normative, counting-based measures using ordinal data.

[3] More elaborative dominance analysis can be conducted with respect to the deprivation cutoffs and weights. For multivariate stochastic dominance analysis using ordinal variables, see Yalonetzky (2014).

shares of population in distributions x and y with achievement level less than $b \in \mathbb{R}_+$. Distribution x first-order stochastically dominates distribution y (or x FSD y) if and only if $F_x(b) \leq F_y(b)$ for all b and $F_x(b) < F_y(b)$ for some b. Strict FSD requires that $F_x(b) < F_y(b)$ for all b.[4] Interestingly, if distribution x FSD y, then y has no lower headcount ratio than x for all poverty lines.

Let us now explain how we can apply this concept for unanimous pairwise comparisons using M_0 and H between any two distributions of deprivation scores across the population. For a given deprivation cutoff vector z and a given weighting vector w, the FSD tool can be used to evaluate the sensitivity of any pairwise comparison to varying poverty cutoff k. Following the notation introduced in Chapter 2, we denote the (uncensored) deprivation score vector by c. Note that an element of c denotes the deprivation score and a larger deprivation score implies a lower level of well-being.

The FSD tool can be applied in two different ways: one is to convert deprivations into attainments by transforming the deprivation score vector c into an attainment score vector $1 - c$, and the other option is to use the tool directly on the deprivation score vector c. The first approach has been pursued in Alkire and Foster (2011a) and Lasso de la Vega (2010). In this section, because it is more direct, we present the results using the deprivation score vector and thus avoid any transformation. A person is identified as poor if the deprivation score is larger than or equal to the poverty cutoff k, unlike in the attainment space where a person is identified as poor if the person's attainment falls below a certain poverty cutoff. To do that, however, we need to introduce the **complementary cumulative distribution function (CCDF)**—the complement of a CDF.[5] For any distribution y with CDF F_y, the CCDF of the distribution is $\bar{F}_y = 1 - F_y$, which means that for any value b, the CCDF $\bar{F}_y(b)$ is the proportion of the population that has values larger than or equal to b. Naturally, CCDFs are downward sloping. The first-order stochastic dominance condition in terms of the CCDFs can be stated as follows. Any distribution y first order stochastically dominates distribution y' if and only

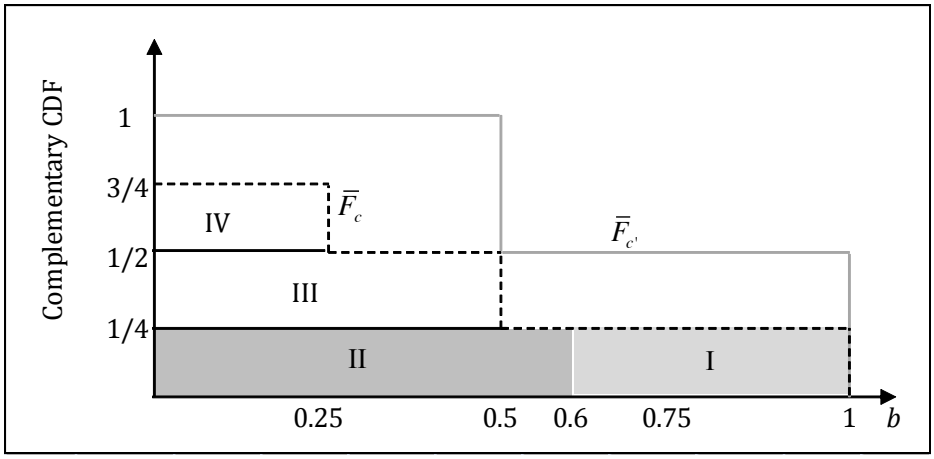
[4] In empirical applications, some statistical tests cannot discern between weak and strong dominance and thus assume x *first order stochastically dominates* distribution y, if $F_x(b) < F_y(b)$ for all b. See, for example, Davidson and Duclos (2012: 88–89).

[5] This is also variously known as survival function or reliability function in other branches of studies.

if $\bar{F}_y(b) \geq \bar{F}_{y'}(b)$ for all b and $\bar{F}_y(b) > \bar{F}_{y'}(b)$ for some b. For strict FSD, the strict inequality must hold for all b.

Now, suppose there are two distributions of deprivation scores, c and c', with CCDFs $\bar{F}_c$ and $\bar{F}_{c'}$. For poverty cutoff k, if $\bar{F}_c(k) \geq \bar{F}_{c'}(k)$, then distribution c has no lower multidimensional headcount ratio H than distribution c' at k. When is it possible to say that distribution c has no lower H than distribution c' for all poverty cutoffs? The answer is when distribution c first order stochastically dominates distribution c'. Let us provide an example in terms of two four-person vectors of deprivation scores: $c = (0, 0.25, 0.5, 1)$ and $c' = (0.5, 0.5, 1, 1)$. The corresponding CCDFs $\bar{F}_c$ and $\bar{F}_{c'}$ are denoted by a black dotted line and a solid grey line, respectively, in Figure 8.1. No part of $\bar{F}_c$ lies above that of $\bar{F}_{c'}$ and so $\bar{F}_{c'}$ first-order stochastically dominates $\bar{F}_c$ and we can conclude that c has unambiguously lower poverty than c', in terms of the multidimensional headcount ratio.

Figure 8.1 Complementary CDFs and Poverty Dominance

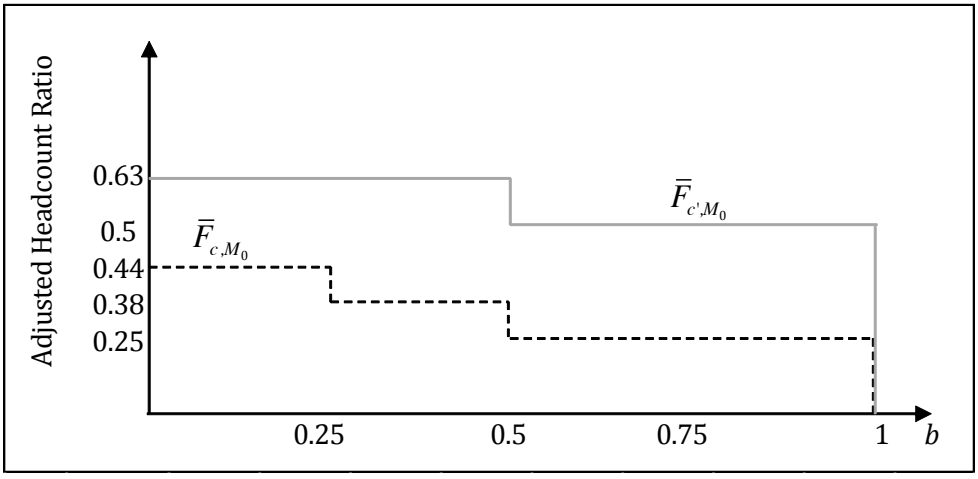


Let us now try to understand dominance in terms of M_0. In order to do so, first note that the area underneath a CCDF of a deprivation score vector is the average of its deprivation scores. Consider distribution c with CCDF $\bar{F}_c$ as in Figure 8.1. The area underneath $\bar{F}_c$ is the sum of areas I, II, III, and IV. Area IV is equal to $0.25 \times 1/4$, Area III is $0.5 \times 1/4$, and Areas I+II is $1 \times 1/4$, so essentially each area is a score times its frequency in the population. The sum of the four areas $(0.25 + 0.5 + 1)/4 = \sum_{i=1}^{4} c_i /4$, is simply the average of all elements in c and it coincides with the M_0 measure for a **union** approach. When an **intermediate** or **intersection** approach to identification is used, then the M_0 is the average of the censored deprivation score vector

$c(k)$. In other words, the deprivation scores of those who are not identified as poor are set to 0. For example, for a poverty cutoff $k = 0.6$, the censored deprivation score vector corresponding to c is $c(k) = (0, 0, 0, 1)$. Obtaining the average of censored deprivation scores is equivalent to ignoring areas III and IV in Figure 8.1. The M_0 of c for $k = 0.6$ is the sum of the remaining area I, II which is $1 \times 1/4 = 0.25$.[6]

We now compute the area underneath the censored CCDF for every $k \in (0,1]$ and plot the area on the vertical axis for each k on the horizontal axis and refer to it as an M_0 curve, depicted in Figure 8.2. We denote the M_0 curves of distributions c and c' by $\bar{F}_{c,M_0}$ and $\bar{F}_{c',M_0}$, respectively. Given that the M_0 curves are obtained by computing the areas underneath the CCDFs, the dominance of M_0 curves is referred as second-order stochastic dominance. Given that first-order stochastic dominance implies second-order dominance, if first-order dominance holds between two distributions, then M_0 dominance will also hold between them. However, the converse is not necessarily true, that is, even when there is M_0 dominance there may not be H dominance. Therefore, when the CCDFs of two distributions cross—i.e. there is not first-order (H) dominance—it is worth testing M_0 dominance between pairs of distributions, to which we refer as pairwise comparisons from now on, using the M_0 curves. Batana (2013) has used the M_0 curves for the purpose of robustness analysis while comparing multidimensional poverty among women in fourteen African countries.

Figure 8.2 The Adjusted Headcount Ratio Dominance Curves



[6] Technically, M_0 for poverty cutoff k can be expressed as $M_0 = \int_k^1 \bar{F}_c(x)dx + k\bar{F}_c(k)$. In our example, Area I is computed as $\int_k^1 \bar{F}_c(x)dx$ and Area II as $k\bar{F}_c(k)$.

The dominance requirement for all possible poverty cutoffs may be an excessively stringent requirement. Practically, one may seek to verify the unambiguity of comparison with respect to a limited variation the poverty cutoff, which can be referred to as **restricted dominance analysis**. For example, when making international comparisons in terms of the MPI, Alkire and Santos (2010, 2014) tested the robustness of pairwise comparisons for all poverty cutoffs $k \in [0.2, 0.4]$, in addition to the poverty cutoff of $k = 1/3$. In this case, if the *restricted* FSD holds between any two distributions, then dominance holds for the relevant particular range of poverty cutoffs for both H and M_0.

8.1.2 Rank Robustness Analysis

In situations in which dominance tests are too stringent, we may explore a milder form of robustness, which assesses the extent to which a ranking, that is, an ordering of more than two entities, obtained under a specific set of parameters' values, is preserved when the value of some parameter is modified. How should we assess the robustness of a ranking? One first intuitive measure is to compute the percentage of pairwise comparisons that are robust to changes in parameters – that is the proportion of pairwise comparisons that have the *same* ordering. As we shall see in section 8.3, whenever poverty computations are performed using a survey, the statistical inference tools need to be incorporated into the robustness analysis.

Another useful way to assess the robustness of a ranking is by computing a rank correlation coefficient between the original ranking of entities and the alternative rankings (i.e. those obtained with alternative parameters' values). There are various choices for a rank correlation coefficient. The two most commonly used rank correlation coefficients are the Spearman rank correlation coefficient (R^{ρ}) and the Kendall rank correlation coefficient (R^{τ}).[7]

Suppose, for a particular parametric specification, the set of ranks across m population subgroups is denoted by $r = (r_1, r_2, \ldots, r_m)$, where r_ℓ is the rank attributed to subgroup ℓ. The subgroups may be ranked by their level of multidimensional headcount ratio, the Adjusted Headcount Ratio, or any other partial indices. We present the rank correlation

[7] In this book, we only focus on bivariate rank correlation coefficients, but there are various methods to measure multivariate rank concordance that we do not cover. For such examples, see Boland and Proschan (1988), Joe (1990), and Kendall and Gibbons (1990). For an application of some of the multivariate concordance methods to examine multivariate concordance of MPI rankings, see Alkire et al. (2010).

measures using population subgroups, but they apply to ranking across countries as well. We denote the set of ranks for an alternative specification of parameters by r', where r'_ℓ is the rank attributed to subgroup ℓ. The alternative specification may be a different poverty cutoff, a different set of deprivation cutoffs, a different set of weights, or a combination of all three. If the initial and the alternative specification yield exactly the same set of rankings across subgroups, then $r_\ell = r'_\ell$ for all $\ell = 1, \ldots, m$. In this case, we state that the two sets of rankings are **perfectly positively** associated and the association is highest across the two specifications. In terms of the previous approach, 100% of the pairwise comparisons are robust to changes in one or more parameters' values. On the other hand, if the two specifications yield completely opposite sets of rankings, then $r_\ell = r'_{m-\ell}$ for all $\ell = 1, \ldots, m$. In this case, we state that the two sets of rankings are **perfectly negatively** associated and the association is lowest across the two specifications. In terms of the previous approach, 0% of the pairwise comparisons are robust to changes in one or more parameters' values.

The Spearman rank correlation coefficient can be expressed as

$$R^\rho = 1 - \frac{6\sum_{\ell=1}^{m}(r_\ell - r'_\ell)^2}{m(m^2-1)}. \qquad \textbf{(8.1)}$$

Intuitively, for the Spearman rank correlation coefficient, the square of the difference in the two ranks for each subgroup is computed and an average is taken across all subgroups. The R^ρ is bounded between -1 and $+1$. The lowest value of -1 is obtained when two rankings are perfectly negatively associated with each other whereas the largest value of $+1$ is obtained when two rankings are perfectly positively associated with each other.

The Kendall rank correlation coefficient is based on the number of concordant pairs and discordant pairs. A pair (ℓ, ℓ') is concordant if the comparisons between two objects are the same in both the initial and alternative specification, i.e. $r_\ell > r_{\ell'}$ and $r'_\ell > r'_{\ell'}$. In terms of the previously used terms, a concordant pair is equivalent to a robust pairwise comparison. A pair, on the other hand, is discordant if the comparisons between two objects are altered between the initial and the alternative specification such that $r_\ell > r_{\ell'}$ but $r'_\ell < r'_{\ell'}$. In terms of the previously used terms, a discordant pair is equivalent to a non-robust pairwise comparison. The R^τ is the difference in the number of concordant

and discordant pairs divided by the total number of pairwise comparisons. The Kendall rank correlation coefficient can be expressed as

$$R^{\tau} = \frac{\#\,\text{Concordant Pairs} - \#\,\text{Discordant Pairs}}{m(m-1)/2}. \qquad \textbf{(8.2)}$$

Like R^{ρ}, R^{τ} also lies between -1 and $+1$. The lowest value of -1 is obtained when two rankings are perfectly negatively associated with each other whereas the largest value of $+1$ is obtained when two rankings are perfectly positively associated with each other. Although both R^{ρ} and R^{τ} are used to assess rank robustness the Kendall rank correlation coefficient has an intuitive interpretation. Suppose the Kendall Tau correlation coefficient is 0.90, from equation (8.2), it can be deduced that this means that 95% of the pairwise comparisons are concordant (i.e. robust) and only 5% are discordant. Equations (8.1) and (8.2) are based on the assumption that there are no ties in the rankings. In other words, both expressions are applicable when no two entities have equal values. When there are ties, Kendall (1970) offers two adjustments in the denominator of both rank correlation coefficients (R^{ρ} and R^{τ}) to correct for tied ranks; these adjusted Kendall coefficients are commonly known as tau-b and tau-c.

Table 8.1 Correlation among Country Ranks for Different Weights[8]

		Equal Weights
Alternative Weights 1	Spearman	0.979
	Kendall	0.893
Alternative Weights 2	Spearman	0.987
	Kendall	0.918
Alternative Weights 3	Spearman	0.985
	Kendall	0.904

Let us present one empirical illustration showing how rank robustness tools may be used in practice. The first illustration presents the correlation between 2011 MPI rankings across 109 countries and the rankings for three alternative weighting vectors (Alkire et al. 2011). The MPI attaches equal weights across three dimensions: health, education, and standard of living. However, it is hard to argue with perfect confidence that the initial weight is the correct choice. Therefore, three alternative weighting schemes were considered. The first alternative assigns a 50% weight to the health dimension and then a 25% weight to each of the other two dimensions. Similarly, the second alternative assigns a 50% weight to the education dimension and then distributes the rest of the weight

[8] The computations of the Spearman and Kendall coefficients in the table have been adjusted for ties. For the exact formulation of tie-adjusted coefficients, see Kendall and Gibbons (1990).

equally across the other two dimensions. The third alternative specification attaches a 50% weight to the standard of living dimension and then 25% weights to each of the other two dimensions. Thus, we now have four different rankings of 109 countries, each involving 5,356 pairwise comparisons. Table 8.1 presents the rank correlation coefficient R^{ρ} and R^{τ} between the initial ranking and the ranking for each alternative specification. It can be seen that the Spearman coefficient is around 0.98 for all three alternatives. The Kendall coefficient is around 0.9 for each of the three cases, implying that around 80% of the comparisons are concordant in each case.

The same type of analysis has been done to changes in other parameters' values, such as the indicators used and deprivation cutoffs (Alkire and Santos 2014).

8.2 Statistical Inference

The last section showed how the robustness of claims made using the Adjusted Headcount Ratio and its partial indices may be assessed. Such assessments apply to changes in a country's performance over time, comparisons between different countries, and comparisons of different population subgroups within a country. Most frequently, the indices are estimated from sample surveys with the objective of estimating the unknown population parameters as accurately as possible. A sample survey, unlike a census that covers the entire population, consists of a representative fraction of the population.[9] Different sample surveys, even when conducted at the same time and despite having the same design, would most likely provide a different set of estimates for the same population parameters. Thus, it is crucial to compute a measure of confidence or reliability for each estimate from a sample survey. This is done by computing the standard deviation of an estimate. The standard deviation of an estimate is referred to as its **standard error**. The lower the magnitude of a standard error, the larger the reliability of the corresponding estimate. Standard errors are key for hypothesis testing and for the construction of confidence intervals, both of which are very helpful for robustness analysis and more generally for drawing policy conclusions. In what follows we briefly explain each of these statistical terms.

[9] Various sampling methods, such as simple random sampling, systematic sampling, stratified sampling, and proportional sampling, are used to conduct a sampling survey.

8.2.1 Standard Errors

There are different approaches to estimating standard errors. Two approaches are commonly followed:

- **Analytical Approach**: Formulas that provide either the exact or the asymptotic approximation of the standard error and thus confidence intervals[10]
- **Resampling Approach**: Standard errors and the confidence intervals may be computed through the bootstrap or similar techniques (as performed for the global MPI in Alkire and Santos 2014).

The Appendix to this chapter presents the formulas for computing standard errors with the analytical approach depending on the survey design.

The analytical approach is based on two assumptions. Such assumptions are based on the premise that the sample surveys used for estimating the population parameters are significantly smaller in size compared to the population size under consideration.[11] For example, the sample size of the Demographic and Health Survey of India in 2006 was only 0.04% of the Indian population. The first assumption is that the samples are drawn from a population that is infinitely large, so that *even* the finite population under study *is a sample* of an infinitely large superpopulation. This philosophical assumption is based on the **superpopulation approach**, which is different from the **finite population approach** (for further discussion see Deaton 1997). A finite population approach requires that a finite population correction factor should be used to deflate the standard error if the sample size is large relative to the population. However, if the sample size is significantly smaller than the finite population size, the finite population correction factor is approximately equal to one. In this case, the standard errors based on both approaches are almost the same.

The second assumption is that we treat each sample as drawn from the population *with replacement*. The practical motivation behind the assumption is the size of the sample survey compared to the population. The sample surveys are commonly conducted without replacement because, once a household is visited and interviewed, the same household is not visited again on purpose. When samples are drawn with replacement, the observations are independent of each other. However, if the samples are drawn

[10] Yalonetzky (2010).
[11] If the particular condition under study does not justify the assumptions made here, then these assumptions need to be relaxed and the standard error formulations are adjusted accordingly.

without replacement, then the samples are not independent of each other. It can be shown that in the absence of multistage sampling, a sampling without replacements needs a Finite Population Correction (FPC) factor for computing the sampling variance. The FPC factor is of the order $1 - \mathbb{m}/n$, where $\mathbb{m}$ is the sample size and n is the size of the population. The use of an FPC factor allows us to get a better estimate of the true population variance. However, when the sample size is small with respect to the population, i.e. $\mathbb{m}/n \to 0$, the use of an FPC factor will not make much difference to the estimation of the sampling variance as the FPC factor is closer to one (Duclos and Araar 2006: 276). These assumptions would be required in order to justify our assumption that each sample is independently and identically distributed.

We now illustrate relevant methods using the Adjusted Headcount Ratio (M_0) denoting its sample estimate by $\widehat{M}_0$ and standard error of the estimate by se_{M_0}. However, the methods are equally applicable to inferences for the multidimensional headcount ratio, the intensity, and the censored headcount ratios as long the standard errors are appropriately computed, as outlined in the Appendix of this chapter.

8.2.2 Confidence Intervals

A confidence interval of a point estimate is an interval that contains the true population parameter with some probability that is known as its confidence level. A significance level that is used is the complement of the confidence level. Let us denote the significance level[12] by ω, which by definition ranges between 0 and 100%. The level of confidence is $(1 - \omega)$ percent. Thus, for a given estimate, if one wants to be 95% confident about the range within which the true population parameter lies, then the significance is 5%. Similarly, if one wants to be 99% confident, then the significance level is 1%.

By the central limit theorem, we can say that the difference between the population parameter and the corresponding sample average divided by the standard error approximates the standard normal distribution (i.e. the normal distribution with a mean of 0 and a standard deviation of 1). Using the standard normal distribution one can

[12] The significance level is also referred to as the Type I error, which is the probability of rejecting the null hypothesis when it is true. See section 8.2.3 for the notion of null hypothesis. By statistical convention, the significance level is denoted with α. However, to avoid confusion with the use of this symbol for other purposes in this book, we denote it ω.

determine the *critical* value associated with that significance level, which is given by the **inverse of the standard normal distribution** at $\omega/2$. In other words, the critical value is the value at which the probability that the statistic is higher than that is precisely $\omega/2$.[13] The critical values to be used when one is interested in computing a 95% confidence interval are: $|\mathfrak{z}_{\omega/2}|=1.96$. If instead one is interested in computing a 99% or a 90% confidence interval, the corresponding critical values are $|\mathfrak{z}_{\omega/2}|=2.58$ and $|\mathfrak{z}_{\omega/2}|=1.645$, respectively.

For example, Table 8.2 presents the sample estimate of the Adjusted Headcount Ratio ($\widehat{M}_0$), the multidimensional headcount ratio ($\widehat{H}$), and the average deprivation share among the poor ($\hat{A}$) from the Demographic and Health Survey of 2005–2006. India's sample estimate of the population-Adjusted Headcount Ratio is $\widehat{M}_0 = 0.251$, with a standard error $se_{M_0} = 0.0026$. The 95% confidence interval is then $(\widehat{M}_0 - \mathfrak{z}_{\omega/2} \times se_{M_0}, \widehat{M}_0 + \mathfrak{z}_{\omega/2} \times se_{M_0}) = (0.245, 0.256)$. This means that with 95% confidence, the true population M_0 lies between 0.245 and 0.256. Similarly, the 99% confidence interval of India's $\widehat{M}_0$ is $(0.244, 0.257)$. The more one wants to be confident about the range within which the true population parameter lies, the larger the confidence interval will be.

Table 8.2 Confidence Intervals for $\widehat{M}_0$, $\widehat{H}$, and $\widehat{A}$

India 2005/6				
Estimate	**Value**	**Standard Error**	**Confidence Interval (95%)**	**Confidence Interval (99%)**
$\widehat{M}_0$	0.251	0.0026	(0.245, 0.256)	(0.244, 0.258)
$\widehat{H}$	48.5%	0.41%	(47.7%, 49.3%)	(47.4%, 49.6%)
$\hat{A}$	51.7%	0.20%	(51.3%, 52.1%)	(51.2%, 52.2%)

Source: Alkire and Seth (2013b)

Similar to $\widehat{M}_0$, the confidence interval for $\hat{A}$ is $(\hat{A} - \mathfrak{z}_{\omega/2} \times se_A, \hat{A} + \mathfrak{z}_{\omega/2} \times se_A)$, for $\widehat{H}$ is $(\widehat{H} - \mathfrak{z}_{\omega/2} \times se_H, \widehat{H} + \mathfrak{z}_{\omega/2} \times se_H)$, and for $\widehat{H}_j$ is $(\widehat{H}_j - \mathfrak{z}_{\omega/2} \times se_{H_j}, \widehat{H}_j + \mathfrak{z}_{\omega/2} \times se_{H_j})$ for all $j = 1, \ldots, d$. It can be seen from the table that the standard error of $\widehat{H}$ is 0.41%, whereas that of $\hat{A}$ is 0.20%.

8.2.3 Hypothesis Tests

Confidence intervals are useful for judging the statistical reliability of a point estimate when the population parameter is unknown. However, suppose that, somehow, we have

[13] Whenever the population standard deviation is unknown or when the sample size is small, one needs to use the Student-t distribution to compute the critical values rather than the standard normal distribution.

a hypothesis about what the population parameter is. For example, suppose the government hypothesizes that the Adjusted Headcount Ratio in India is 0.26. Thus, the null hypothesis is $\mathcal{H}_0: M_0 = 0.26$. This has to be tested against any of the three alternatives $\mathcal{H}_1: M_0 \neq 0.26$ or $\mathcal{H}_1: M_0 > 0.26$ or $\mathcal{H}_1: M_0 < 0.26$.[14] This is a **one-sample test**. Note that the first alternative requires a so-called two-tailed test, and each of the other two alternatives requires a so-called one-tailed test. Now, suppose a sample (either simple random or multistage stratified) $\hat{X}$ of size $\mathbb{m}$ is collected. We denote the estimated Adjusted Headcount Ratio by $\widehat{M}_0$. By the law of large numbers and by the central limit theorem, as $\mathbb{m} \to \infty$, $\left(\widehat{M}_0 - M_0\right) \xrightarrow{d} \text{Normal}(0, \sigma_0^2/\mathbb{m})$, where $\sigma_0^2 = E[\hat{c}_i(k) - M_0]^2$ is the population variance of M_0. The standard error se_{M_0} of $\widehat{M}_0$ can be estimated using either equation (8.11) or (8.30) in the Appendix, whichever is applicable.

In a two-tail test, the null hypothesis can be rejected against the alternative $\mathcal{H}_1: M_0 \neq 0.26$ with $(1-\omega)$ percent confidence if $|(\widehat{M}_0 - 0.26)/se_{M_0}| > |\mathfrak{z}_{\omega/2}|$.; in words, if the absolute value of the statistic is greater than the absolute value of the critical value. An equivalent procedure to reject or not the null hypothesis entails, rather than comparing the test statistic against the critical value, comparing the significance level against the so-called p-value. The p-value is defined as the actual probability that the test statistic assumes a value greater than the value observed, i.e. it is the probability of rejecting the null hypothesis when it is true.

Let us consider the example of India's Adjusted Headcount Ratio, reported in Table 8.2, where $\widehat{M}_0 = 0.251$ and $se_{M_0} = 0.0026$. Now, $|(\widehat{M}_0 - 0.26)/se_{M_0}| = 3.46 > 2.58 = \mathfrak{z}_{0.5}$. Thus, with 99% confidence, the null hypothesis can be rejected with respect to the alternative $M_0 \neq 0.26$ and the corresponding p-value is $2(1 - \Phi[(\widehat{M}_0 - 0.26)/se_{M_0}])$, where Φ stands for the cumulative standard normal distribution. Similarly, in a one-tail test to the right, the null hypothesis can be rejected against the alternative $\mathcal{H}_1: M_0 > 0.26$ with $(1-\omega)$ percent confidence if $(\widehat{M}_0 - 0.26)/se_{M_0} > \mathfrak{z}_{1-\omega}$. The corresponding p-value is $[1 - \Phi((\widehat{M}_0 - 0.26)/se_{M_0})]$. Finally, in a one tail test to the left, the null hypothesis can be rejected against the alternative $\mathcal{H}_1: M_0 < 0.26$ with $(1-\omega)$ percent

[14] We present the tests for country-level estimates but they are equally applicable to other population subgroups. Also, we only present the tests in terms of the M_0 measure, but again they are also applicable to A, H, and H_j for all j. and so we have chosen not to repeat the results.

confidence, if $(\widehat{M}_0 - 0.26)/se_{M_0} < \mathfrak{z}_\omega$, where the relevant p-value is $\Phi((\widehat{M}_0 - 0.26)/se_{M_0})$.[15]

Note that the conclusions based on the confidence intervals and the one-sample tests are identical. If the value at the null hypothesis lies outside of the confidence interval, then the test will also show that null hypothesis is rejected. On the other hand, if the value at the null hypothesis lies inside the confidence interval, then the test cannot reject the null hypothesis.

Formal tests are also required in order to understand whether a change in the estimate over time—or a difference between the estimates of two countries—has been statistically significant. The difference is that this is a **two-sample test**. We assume that the two estimates whose difference is of interest are estimated from two independent samples.[16] For example, when we are interested in testing the difference in M_0 across two countries, across rural and urban areas, or across population subgroups, it is safe to assume that the samples are drawn independently. A somewhat different situation may arise with a change over time. It is possible that the samples are drawn independently of each other or that the samples are drawn from the same population in order to track changes over time, as, for example, in panel datasets. This section restricts its attention to assessments in which we can assume independent samples.

Suppose there are two countries, Country I and Country II. The population achievement matrices are denoted by X^I and X^{II}, respectively, and the population Adjusted Headcount Ratios are denoted by $M_{0,I}$ and $M_{0,II}$, respectively. We seek to test the null hypothesis $\mathcal{H}_0: M_{0,I} - M_{0,II} = 0$, which implies that poverty in country I is not significantly different from poverty in country II in any of the three alternatives: $\mathcal{H}_1: M_{0,I} - M_{0,II} \neq 0$ which means that one of the two countries is significantly poorer than the other; or $\mathcal{H}_1: M_{0,I} - M_{0,II} > 0$, which means that country I is significantly poorer than country II; or $\mathcal{H}_1: M_{0,I} - M_{0,II} < 0$, which means the opposite. For the first alternative, we need to conduct a two-tailed test, and for the other two alternatives, we need to conduct a one-tailed test.

---

[15] See Bennett and Mitra (2013) for an exposition of hypothesis testing of M_0 and other AF partial sub-indices using a minimum p-value approach.

[16] See chapters 14 and 16 of Duclos and Araar (2006) for further discussion on non-independent samples for panel data analysis.

Now, suppose a sample (either simple random or multistage stratified) $\hat{X}^I$ of size $\mathbb{m}^I$ is collected from X^I and a sample $\hat{X}^{II}$ of size $\mathbb{m}^{II}$ is collected from X^{II}, where samples in $\hat{X}^I$ and $\hat{X}^{II}$ are assumed to have been drawn independently of each other. We denote the estimated Adjusted Headcount Ratios from the samples by $\widehat{M}_{0,I}$ and $\widehat{M}_{0,II}$, respectively. By the law of large numbers and the central limit theorem, $\left(\widehat{M}_{0,I} - M_{0,I}\right) \xrightarrow{d} \text{Normal}\left(0, \sigma^2_{0,I}/\hat{n}^I\right)$ and $\left(\widehat{M}_{0,II} - M_{0,II}\right) \xrightarrow{d} \text{Normal}\left(0, \sigma^2_{0,II}/\hat{n}^{II}\right)$. The difference of two normal distributions is a normal distribution as well. Thus,

$$\left(\left(\widehat{M}_{0,I} - \widehat{M}_{0,II}\right) - \left(M_{0,I} - M_{0,II}\right)\right) \xrightarrow{d} \text{Normal}\left(0, \sigma^2_{0,I-II}\right), \tag{8.3}$$

where $\sigma^2_{0,I-II} = \frac{\sigma^2_{0,I}}{\mathbb{m}^I} + \frac{\sigma^2_{0,II}}{\mathbb{m}^{II}}$. Note that, as we have assumed independent samples, the covariance between the two Adjusted Headcount Ratios is zero. Hence, the standard error of $\widehat{M}_{0,I} - \widehat{M}_{0,II}$, denoted by $se_{M_0,I-II}$, may be estimated using equations (8.11) or (8.30) in the Appendix, whichever is applicable, as:

$$se_{M_0,I-II} = \sqrt{se^2_{M_0,I} + se^2_{M_0,II}}, \tag{8.4}$$

where $se^2_{M_0,I}$ is the standard error of $\widehat{M}_{0,I}$ and $se^2_{M_0,II}$ is the standard error of $\widehat{M}_{0,II}$. Like the one-sample test discussed above, in the two-tail test, the null hypothesis can be rejected against the alternative $\mathcal{H}_1: M_{0,I} - M_{0,II} \neq 0$ with $(1-\omega)$ percent confidence, if $|[(\widehat{M}_{0,I} - \widehat{M}_{0,II}) - (M_{0,I} - M_{0,II})]/se_{M_0,I-II}| > |\mathfrak{z}_{\omega/2}|$. Given that at the null hypothesis $M_{0,I} - M_{0,II} = 0$, this implies requiring $|(\widehat{M}_{0,I} - \widehat{M}_{0,II})/se_{M_0,I-II}| > |\mathfrak{z}_{\omega/2}|$. Similarly, in order to reject the null hypothesis against $\mathcal{H}_1: M_{0,I} - M_{0,II} > 0$, we require $(\widehat{M}_{0,I} - \widehat{M}_{0,II})/se_{M_0,I-II} > \mathfrak{z}_{1-\omega}$ and against $\mathcal{H}_1: M_{0,I} - M_{0,II} < 0$, we require $(\widehat{M}_{0,I} - \widehat{M}_{0,II})/se_{M_0,I-II} < \mathfrak{z}_{\omega}$. The corresponding p-values can be computed as discussed in the one-sample test.

Table 8.3 presents an example of an estimation of MPI (an adaptation of M_0) in four Indian states: Goa, Punjab, Andhra Pradesh, and Tripura, with their corresponding standard errors, confidence intervals and hypothesis tests.[17] These results are computed from the Demographic and Health Survey of India for the years 2005–2006. In the table we can see that the MPI point estimate for Goa is 0.057, and with 95% confidence, we

[17] Alkire and Seth (2013b) use an MPI harmonized for strict comparability of indicator definitions across time.

can say that the MPI estimate of Goa lies somewhere between 0.045 and 0.069. Similarly, we can say with 95% confidence that Punjab's MPI is not larger than 0.103 and no less than 0.073, although the point estimate of MPI is 0.088. We can also state, after doing the corresponding hypothesis test, that Punjab is significantly poorer than Goa. However, we cannot draw the same kind of conclusion for the comparison between Andhra Pradesh and Tripura, although the difference between the MPI estimates of these two states (0.032) is similar to the difference between Goa and Punjab.

Table 8.3 Comparison of Indian States Using Standard Errors

States	MPI	Standard Error	95% Confidence Interval		Difference	
			Lower Bound	Upper Bound	MPI	Statistically Significant
Goa	0.057	0.0062	0.045	0.069	0.031	Yes
Punjab	0.088	0.0078	0.073	0.103		
Andhra Pradesh	0.194	0.0093	0.176	0.212	0.032	No
Tripura	0.226	0.0162	0.195	0.258		

Source: Alkire and Seth (2013b)

It is vital to understand that in two sample tests, conclusions about the statistical significance obtained with confidence intervals do not necessarily coincide with conclusions obtained using hypothesis testing. Let us formally examine the situation. Suppose, $\widehat{M}_{0,I} > \widehat{M}_{0,II}$. If the confidence intervals do not overlap, then the lower bound of $\widehat{M}_{0,I}$ is larger than the upper bound of $\widehat{M}_{0,II}$, i.e. $\widehat{M}_{0,I} - ȝ_{\omega/2} \times se_{M_{0,I}} > \widehat{M}_{0,II} + ȝ_{\omega/2} \times se_{M_{0,II}}$ or $[\widehat{M}_{0,I} - \widehat{M}_{0,II}]/[se_{M_{0,I}} + se_{M_{0,II}} > ȝ_{\omega/2}$. Given that for two independent samples, $se_{M_{0,I}} + se_{M_{0,II}} > se_{M_{0,I-II}}$, if the confidence intervals do not cross, a statistically significant comparison can be made. However, if the confidence intervals overlap, it does not necessarily mean that the comparison is *not* statistically significant at the same level of significance. It is thus essential to conduct statistical tests on *differences* when the confidence intervals overlap.

8.3 Robustness Analysis with Statistical Inference

In practice, the robustness analyses discussed in section 8.1 are typically performed with estimates from sample surveys. In at least two cases, it is necessary to *combine* the robustness analyses with the statistical inference tools just described. This section describes how to do so in practice.

The dominance analysis presented in section 8.1.1 assesses dominance between two CCDFs or two M_0 curves in order to conclude whether a pairwise ordering is robust to

the choice of all poverty cutoffs. But it is also crucial to examine if the pairwise dominance of the CCDFs or M_0 curves are statistically significant. For two entities in a pairwise ordering, one should perform one-tailed hypothesis tests of the difference in the two M_0 estimates for *each* possible k value, as described in section 8.2.3. This will determine whether the two countries' poverty estimates are not significantly different or whether one is significantly poorer than the other *regardless* of the poverty cutoff.[18] One may also construct confidence interval curves around each CCDF curve (or M_0 curve) and examine whether two corresponding confidence interval curves overlap or not, in order to conclude dominance. More specifically, if the lower confidence interval curve of a unit does not overlap with the upper confidence interval curve of another unit, then one may conclude that statistically significant dominance holds between two entities. However, as explained at the end of section 8.2.3, no conclusion on statistical significance can be made when the confidence intervals overlap. Thus a hypothesis test for dominance should be preferred.[19]

This need to *combine* methods also applies to the other type of robustness analysis presented in section 8.1.2, in the sense that one can implement this analysis to a ranking of entities and report the proportion of robust pairwise comparisons across the different k values. Moreover, the analysis described in section 8.2.3 (hypothesis testing or comparison of confidence intervals by pairs of entities) can be implemented not only with respect to the poverty cutoff but also with respect to changes in the other parameters, such as weights, deprivation cutoffs or alternative indicators.

As Alkire and Santos observe (2014: 260), the number of robust pairwise comparisons may be expressed in two ways. One may report the *proportion of the total* possible pairwise comparisons that are robust. A somewhat more precise option is to express it as a *proportion of the number of significant pair-wise comparisons in the baseline measure*, because a pairwise comparison that was not significant in the baseline M_0 cannot, by definition, be a *robust* pairwise comparison.

To interpret results meaningfully, it can be helpful to observe that the proportion of robust pairwise comparisons of alternative M_0 specifications is influenced by: the number of possible pairwise comparisons, the number of significant pairwise

[18] For formal tests on stochastic dominance in unidimensional poverty and welfare analysis, see Anderson (1996), Davidson and Duclos (2000), and Barrett and Donald (2003).
[19] Other new ways of testing robustness may be developed in the near future.

comparisons in the baseline distribution, and the number of alternative parameter specifications. Interpretation of the percentage of robust pairwise comparisons in light of these three factors illuminates the degree to which the poverty estimates and the policy recommendations they generate are valid across alternative plausible design specifications.

Alkire and Santos (2014) perform both types of robustness analysis with the global MPI (2010 estimates) for every possible pair of countries with respect to: (a) a restricted range of k values, namely, 20% to 40%; (b) four alternative sets of plausible weights; and (c) to subgroup-level MPI values.[20] The country rankings seem highly robust to alternative parameters' specifications.[21]

Chapter 9 further develops the techniques of multidimensional poverty measurement and analysis. Specifically, we present techniques for analysing poverty over time (with and without panel data) and for exploring distributional issues such as inequality among the poor.

Appendix: Methods for Computing Standard Errors

This appendix provides a technical outline of how standard errors may be computed. We first present the analytical approach and then the bootstrap method using the notation in Method I presented in Box 5.7. For the multidimensional and censored headcount ratios, we use the notation in Box 5.4. The M_0 and its partial indices are written as

$$M_0(X; z, w, k) = \frac{\sum_{i=1}^{n} c_i(k)}{n}. \quad \textbf{(8.5)}$$

$$A(X; z, w, k) = \frac{\sum_{i=1}^{q} c_i(k)}{q}. \quad \textbf{(8.6)}$$

$$H(X; z, w, k) = \frac{\sum_{i=1}^{n} \mathbb{I}[c_i \geq k]}{n}. \quad \textbf{(8.7)}$$

$$h_j(X; z, w, k) = \frac{\sum_{i=1}^{n} \mathbb{I}\left[(c_i \geq k) \wedge \left(g_{ij}^0 = 1\right)\right]}{n}. \quad \textbf{(8.8)}$$

[20] They compute the MPI for four population subgroups: children 0–14 years of age, women 15–49 years of age, women aged 50 years and older, and men 15 years and older, and test the rankings of subgroup MPIs across countries.

[21] Further methodological work is needed to propose overall robustness standards for measures that will be used for policy.

Note that $\bigwedge$ is the logical 'and' operator. The standard errors of the subgroups' M_0s and partial indices may be computed in the same way and so we only outline the standard errors of equations (8.5)–(8.8).

Simple Random Sampling with Analytical Approach

Suppose $\mathbb{m}$ samples have been collected through simple random sampling from the population. We denote the dataset by $\hat{X}$ and its ij^{th} element by $\hat{x}_{ij}$ for all $i = 1, \dots, \mathbb{m}$ and $j = 1, \dots, d$. We denote the deprivation status score for $\hat{x}_{ij}$ by $\hat{g}_{ij}^0$. For statistical inferences, our analysis focuses on the censored deprivation scores. The score, defined at the population level, becomes a random variable while performing statistical inference. We assume that a random sample (of size $\mathbb{m}$) of censored deprivation scores $\{c_1(k), \dots, c_{\mathbb{m}}(k)\}$ is a sequence of independently and identically distributed random variables with an expected value $E\big(c_i(k)\big) = M_0$ and $\text{Var}\,(c_i(k)) = \sigma_0^2$. Then as $\mathbb{m}$ approaches infinity, the random variable $\sqrt{\mathbb{m}}\big(\widehat{M}_0 - M_0\big)$ converges in distribution to $\text{Normal}(0, \sigma_0^2)$, where $\widehat{M}_0 = (\sum_{i=1}^{\mathbb{m}} c_i(k))/\mathbb{m}$. That is

$$\sqrt{\mathbb{m}}\big(\widehat{M}_0 - M_0\big) \xrightarrow{d} \text{Normal}(0, \sigma_0^2). \tag{8.9}$$

The unbiased sample estimate of σ_0^2 is

$$\hat{\sigma}_0^2 = \frac{1}{\mathbb{m} - 1}\sum\nolimits_{i=1}^{\mathbb{m}} \big[c_i(k) - \widehat{M}_0\big]^2, \tag{8.10}$$

and the standard error of the Adjusted Headcount Ratio is

$$se_{\widehat{M}_0} = \frac{\hat{\sigma}_0}{\sqrt{\mathbb{m} - 1}} = \frac{1}{\mathbb{m} - 1}\sqrt{\sum\nolimits_{i=1}^{\mathbb{m}} \big[c_i(k) - \widehat{M}_0\big]^2}. \tag{8.11}$$

The analytical approach based on the central limit theorem (CLT) also applies to the calculation of the standard errors of H, which leads to

$$\sqrt{\mathbb{m}}\big(\widehat{H} - H\big) \xrightarrow{d} \text{Normal}(0, \sigma_H^2), \tag{8.12}$$

where $\widehat{H} = [\sum_{i=1}^{\mathbb{m}} \mathbb{I}[c_i \geq k])/\mathbb{m}$ and $\sigma_H^2 = E[\mathbb{I}[\hat{c}_i \geq k] - H]^2$. Note that unlike M_0, H is an average across 0s and 1s, i.e. the mean is a proportion and σ_H^2 is estimated as

$$\hat{\sigma}_H^2 = \widehat{H}\big(1 - \widehat{H}\big), \tag{8.13}$$

and so the unbiased standard error is

$$se_{\hat{H}} = \frac{\hat{\sigma}_H}{\sqrt{\mathbb{m}-1}} = \sqrt{\frac{\hat{H}(1-\hat{H})}{\mathbb{m}-1}}. \tag{8.14}$$

With the same logic, the standard error for h_j, can be estimated as

$$se_{\hat{h}_j} = \frac{\hat{\sigma}_{h_j}}{\sqrt{\mathbb{m}-1}} = \sqrt{\frac{\hat{h}_j(1-\hat{h}_j)}{\mathbb{m}-1}}, \tag{8.15}$$

where, $\hat{h}_j = [\sum_{i=1}^{\mathbb{m}} \mathbb{I}[c_i \geq k \wedge (\hat{g}_{ij}^0 = 1)])/\mathbb{m}$.

The formulation of A is analogous to the formulation of M_0, and so the standard error of A) is computed as

$$se_{\hat{A}} = \frac{\hat{\sigma}_A}{\sqrt{\mathbb{q}-1}} = \frac{1}{\mathbb{q}-1}\sqrt{\sum_{i=1}^{\mathbb{q}}\left[c_i(k)-\hat{A}\right]^2}, \tag{8.16}$$

where $\hat{A} = (\sum_{i=1}^{\mathbb{q}} c_i(k))/\mathbb{q}$ and $\mathbb{q}$ is the number multidimensionally poor in the sample.

Note that if the number of multidimensionally poor is extremely low, the sample size for estimating $se_{\hat{A}}$ may not be large enough. This may affect the precision of $se_{\hat{A}}$ using (8.16). It may then be accurate to treat A as a ratio of M_0 and H for computing $se_{\hat{A}}$. By the Taylor series expansion (see the discussion in Casella and Berger 1990: 240–245), $\hat{A}$ can be approximated as $\hat{A} \approx \hat{M}_0/\hat{H}$ and $\sigma_{\hat{A}}^2$ can be estimated as

$$\hat{\sigma}_{\hat{A}}^2 \approx \left(\frac{\hat{M}_0}{\hat{H}}\right)^2\left[\frac{\hat{\sigma}_H^2}{\hat{H}^2} + \frac{\hat{\sigma}_0^2}{\hat{M}_0^2} - \frac{2\hat{\sigma}_{0,H}^2}{\hat{H}\hat{M}_0}\right]. \tag{8.17}$$

where $\hat{\sigma}_0^2$ and $\hat{\sigma}_H^2$ are based on (8.10) and (8.13), respectively, and $\hat{\sigma}_{0,H}^2$ can be estimated as

$$\hat{\sigma}_{0,H}^2 = \frac{1}{\mathbb{m}-1}\sum_{i=1}^{\mathbb{m}}\left[\mathbb{I}[c_i \geq k] - \hat{H}\right][c_i(k) - \hat{M}_0] = \hat{M}_0(1-\hat{H}). \tag{8.18}$$

By combining (8.17) and (8.18), the alternative formulation becomes

$$se_{\hat{A}} = \frac{\hat{\sigma}_A}{\sqrt{\mathbb{m}-1}} \approx \sqrt{\frac{1}{\mathbb{m}-1}\left(\frac{\hat{M}_0}{\hat{H}}\right)^2\left[\frac{\hat{\sigma}_H^2}{\hat{H}^2} + \frac{\hat{\sigma}_0^2}{\hat{M}_0^2} - \frac{2(1-\hat{H})}{\hat{H}}\right]}. \tag{8.19}$$

Stratified Sampling with an Analytical Approach

We next discuss the estimation of standard errors when samples are collected through two-stage stratification.[22] Using information on the population characteristics, the population is partitioned into several strata. The first stage, from each stratum, draws a sample of PSUs with or without replacement. The second stage draws samples either with or without replacement, from each PSU.

We suppose that the population is partitioned into $\mathcal{S} > 1$ strata and there are $\mathcal{P}_s$ PSUs in the s^{th} strata for all $s = 1, \dots, \mathcal{S}$. The population size of the j^{th} PSU in the s^{th} stratum is n_{js} so that $n = \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{P}_s} n_{js}$. We denote the total number of poor by q and the number of poor in the j^{th} PSU in the s^{th} strata by q_{js}. The population M_0 measure and its partial indices are presented in (8.20)–(8.23) with the same notation for the identity function as in (8.5)–(8.8).

$$M_0(X; z, w, k) = \frac{1}{n} \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{P}_s} \sum_{i=1}^{n_{js}} c_{ijs}(k) \qquad \textbf{(8.20)}$$

$$A(X; z, w, k) = \frac{1}{q} \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{P}_s} \sum_{i=1}^{q_{js}} c_{ijs}(k) \qquad \textbf{(8.21)}$$

$$H(X; z, w, k) = \frac{1}{n} \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{P}_s} \sum_{i=1}^{n_{js}} \mathbb{I}\left[c_{ijs} \geq k\right] \qquad \textbf{(8.22)}$$

$$h_j(X; z, w, k) = \frac{1}{n} \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{\mathcal{P}_s} \sum_{i=1}^{n_{js}} \mathbb{I}\left[\left(c_{ijs} \geq k\right) \wedge \left(g^0_{ijs,j} = 1\right)\right] \qquad \textbf{(8.23)}$$

Note that $g^0_{ijs,j} = 1$ if the i^{th} person from the j^{th} PSU in the s^{th} stratum is deprived in the j^{th} dimension and $g^0_{ijs,j} = 0$ otherwise; and c_{ijs} and $c_{ijs}(k)$ are the deprivation score and the censored deprivation score of the i^{th} person from the j^{th} PSU in the s^{th} stratum, respectively. Thus, $c_{ijs} = \sum_{j=1}^{d} w_j \, g^0_{ijs,j}$; and $c_{ijs}(k) = c_{ijs}$ if $c_{ijs} \geq k$ and $c_{ijs}(k) = 0$ otherwise.

Now, suppose a sample of size $\mathbb{m}$ is collected through a two-stage stratified sampling. The first stage selects p_s PSUs from the s^{th} stratum for all s. The second stage selects $\mathbb{m}_{js}$ samples from the j^{th} PSU in $^{\text{th}}$stratum s. So, $\mathbb{m} = \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \mathbb{m}_{sj}$. Each sample i in the j^{th} PSU in the s^{th} stratum is assigned a sampling weight W_{ijs}, which are summarized

[22] Appendix D of Seth (2013) gives an example of standard error estimation for one-stage sample stratification in the multidimensional welfare framework; for consumption/expenditure see Deaton (1997).

by an $\mathbb{m}$-dimensional vector W. The achievements are summarized by matrix $\hat{X}$, which is a typical sample dataset.

In order to estimate the measure from the sample, first, the total population and the total number of poor should be estimated from the sample. We denote the estimates of the population n by $\mathcal{N}$ and the estimate of the poor population q by $\mathcal{Q}$. Then,

$$\mathcal{N} = \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} \tag{8.24}$$

$$\mathcal{Q} = \sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{q}_{js}} W_{ijs} \tag{8.25}$$

The sample estimates of the population averages in (8.20)–(8.23) are presented in (8.26)–(8.29).

$$\widehat{M}_0 = \frac{1}{\mathcal{N}} \left[\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} c_{ijs}(k) \right] \tag{8.26}$$

$$\hat{A} = \frac{1}{\mathcal{Q}} \left[\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{q}_{js}} W_{ijs} c_{ijs}(k) \right] \tag{8.27}$$

$$\widehat{H} = \frac{1}{\mathcal{N}} \left[\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} \mathbb{I}\left[c_{ijs} \geq k\right] \right] \tag{8.28}$$

$$\hat{h}_j = \frac{1}{\mathcal{N}} \left[\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} \mathbb{I}\left[\left(c_{ijs} \geq k\right) \wedge \left(g^0_{ijs,j} = 1\right)\right] \right] \tag{8.29}$$

As each sample estimate is a ratio of two estimators, their standard errors are approximated using (8.17) and using equations (1.31) and (1.63) in Deaton (1997). The standard error for $\widehat{M}_0$ in (8.26) is

$$se_{\widehat{M}_0} = \frac{1}{\mathcal{N}} \sqrt{\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \left[\left(\sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} c_{ijs}(k) - \widehat{M}_0^s \right) - (W^{js} - \bar{W}^s)\widehat{M}_0 \right]^2}, \tag{8.30}$$

where $\widehat{M}_0^s = \left[\sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} c_{ijs}(k)\right] / \left[\sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs}\right]$, , $\bar{W}^s = \left[\sum_{j=1}^{p_s} \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs}\right] / \left[\sum_{j=1}^{p_s} \mathbb{m}_{js}\right]$ and $W^{js} = \sum_{i=1}^{\mathbb{m}_{js}} W_{ijs}$.

The standard errors of $\widehat{H}$ and $\hat{h}_j$ are

$$se_{\widehat{H}} = \frac{1}{\mathcal{N}} \sqrt{\sum_{s=1}^{\mathcal{S}} \sum_{j=1}^{p_s} \left[\left(\sum_{i=1}^{\mathbb{m}_{js}} W_{ijs} \mathbb{I}\left[c_{ijs} \geq k\right] - \widehat{H}^s \right) - (W^{js} - \bar{W}^s)\widehat{H} \right]^2} \tag{8.31}$$

$$se_{\hat{h}_j} = \frac{1}{\mathcal{N}}\sqrt{\sum_{s=1}^{S}\sum_{j=1}^{p_s}\left[\left(\sum_{i=1}^{m_{js}} W_{ijs}\mathbb{I}\left[\left(c_{ijs} \geq k\right) \wedge \left(g^0_{ijs,j} = 1\right)\right] - \hat{h}_j^s\right) - (W^{js} - \bar{W}^s)\hat{h}_j\right]^2}$$

(8.32)

where $\hat{H}^s = \left[\sum_{j=1}^{p_s}\sum_{i=1}^{m_{js}} W_{ijs}\mathbb{I}\left[c_{ijs} \geq k\right]\right] / \left[\sum_{j=1}^{p_s}\sum_{i=1}^{m_{js}} W_{ijs}\right]$ and $\hat{h}_j^s = \left[\sum_{j=1}^{p_s}\sum_{i=1}^{m_{js}} W_{ijs}\mathbb{I}\left[\left(c_{ijs} \geq k\right) \wedge \left(g^0_{ijs,j} = 1\right)\right]\right] / \left[\sum_{j=1}^{p_s}\sum_{i=1}^{m_{js}} W_{ijs}\right]$. Terms $\bar{W}^s$ and W^{js} are the same as in (8.30).

Finally, we present the standard error for $\hat{A}$ in (8.27), where the denominator is Q instead of $\mathcal{N}$ as

$$se_{\hat{A}} = \frac{1}{Q}\sqrt{\sum_{s=1}^{S}\sum_{j=1}^{p_s}\left[\left(\sum_{i=1}^{q_{js}} W_{ijs}c_i(k) - \hat{A}^s\right) - (\mathcal{W}^{js} - \bar{\mathcal{W}}^s)\hat{A}\right]^2}, \quad \textbf{(8.33)}$$

where $\hat{A}^s = \left[\sum_{j=1}^{p_s}\sum_{i=1}^{q_{js}} W_{ijs}c_{ijs}(k)\right] / \left[\sum_{j=1}^{p_s}\sum_{i=1}^{q_{js}} W_{ijs}\right]$, $\bar{\mathcal{W}}^s = \left[\sum_{j=1}^{p_s}\sum_{i=1}^{q_{js}} W_{ijs}\right] / \left[\sum_{j=1}^{p_s} q_{js}\right]$ and $\mathcal{W}^{js} = \sum_{i=1}^{\hat{q}_{js}} W_{ijs}$. Intuitively, $\hat{A}^s$ is the estimated average intensity for stratum s, $\bar{\mathcal{W}}^s$ is the average of sampling weights in stratum s across the poor, and $\mathcal{W}^{js}$ is the sum of all sampling weights in PSU j of stratum s also across the poor.

As a reasonably smaller sample size may affect the precision of the standard error of A the variance var_A can be approximated as in (8.17), but using (8.30) and (8.31) as

$$\widehat{var}_A \approx \left(\frac{\hat{M}_0}{\hat{H}}\right)^2\left[\frac{se_H^2}{\hat{H}^2} + \frac{se_{M_0}^2}{\hat{M}_0^2} - \frac{2\hat{\sigma}_{0,H}^2}{\hat{H}\hat{M}_0}\right], \quad \textbf{(8.34)}$$

where

$$\hat{\sigma}_{0,H}^2 = \frac{1}{\mathcal{N}^2}\sum_{s=1}^{S}\sum_{j=1}^{p_s}\left[\left(\sum_{i=1}^{m_{js}} W_{ijs}c_{ijs}(k) - \hat{M}_0^s\right) - (W^{js} - \bar{W}^s)\hat{M}_0\right]\left[\left(\sum_{i=1}^{m_{js}} W_{ijs}\mathbb{I}\left[c_{ijs} \geq k\right] - \hat{H}^s\right) - (W^{js} - \bar{W}^s)\hat{H}\right].$$

(8.35)

Hence, combining (8.34) and (8.35), we have

$$se_{\hat{A}} = \sqrt{\widehat{var}_A}. \quad \textbf{(8.36)}$$

Note that the analytical standard errors and confidence intervals may not serve too well when the sample sizes are small or when the estimates are too close to the natural upper

or lower bounds.[23] In these cases, resampling non-parametric methods, such as bootstrap, may be more suitable for computing standard errors and confidence intervals.

The Bootstrap Method

An alternative approach for statistical inference is the 'bootstrap', which is a data-based-simulation method for assessing statistical accuracy. Introduced in 1979, it provides an estimate of the sampling distribution of a given statistic θ, such as the standard error, by resampling from the original sample (cf. Efron 1979; Efron and Tibshirani 1993). It has certain advantages over the analytical approach. First, the inference on summary statistics does not rely on CLT as the analytical approach. In particular, for reasonably small sample size, standard errors/confidence intervals computed through the CLT-based asymptotic approximation may be inaccurate. Second, the bootstrap can automatically take into account the natural bounds of the measure. Confidence intervals using the analytical approach can lie outside natural bounds, which can be prevented when the bootstrap re-sampling distribution of the statistic is directly used.

Third, the computation of standard errors may become complex when the estimator and its standard error have a complicated form or have a no-closed expression. These types of complexities are common both in the context of statistical inference of inequality or poverty measurement and in tests where comparisons of group inequality or poverty (across gender or region) are of particular interest (Biewen 2002). Although, the delta-method can handle these analytical standard errors from stochastic dependencies, but when the number of time periods or groups increases, computing the standard errors analytically can easily become cumbersome (cf. Cowell 1989, Nygard and Sandström 1989). In practice, Monte Carlo evidence suggests that bootstrap methods are preferred for these analyses and shows that the simplest bootstrap procedure achieves the same accuracy as the delta-method (Biewen 2002; Davidson and Flachaire 2007). In developing economics, bootstrap has been used to draw statistical inferences for poverty and inequality measurement (Mills and Zandvakili 1997; Biewen 2002).

Here we briefly illustrate the use of the bootstrap for computing standard errors. Readers interested in using the bootstrap for confidence interval estimation and hypothesis testing can refer to Efron and Tibshirani (1993), chapters 12 and 16, respectively.

---

[23] When the estimate is too close to the natural upper and lower bounds (0 and 1), the confidence intervals using analytical standard error may fall outside these bounds. Different methods for adjustments are available. For a discussion of such methods, see Newcombe (1998).

The bootstrap algorithm can be described as a resampling technique, which is conducted $\mathbb{B}$ number of times by generating a random artificial sample each time, with replacement from the original sample, which is our dataset $\hat{X}$. The $\mathbb{b}^{th}$ resample produces an estimate $\hat{\theta}^{*\mathbb{b}}$ for all $= 1, \dots, \mathbb{b}, \dots, \mathbb{B}$. Thus, we have a set of $\mathbb{B}$ resample estimates of $\hat{\theta}$: $\{\hat{\theta}^{*1}, \dots, \hat{\theta}^{*\mathbb{b}}, \dots, \hat{\theta}^{*\mathbb{B}}\}$. If the artificial samples are independent and identically distributed (*i.i.d.)*, the bootstrap standard error estimator of $\hat{\theta}$, denoted $se_{\mathbb{b},\hat{\theta}}$, is defined as

$$se_{\mathbb{b},\hat{\theta}} = \left[\sum_{\mathbb{b}=1}^{\mathbb{B}} \frac{\left[\hat{\theta}^{*\mathbb{b}} - \hat{\theta}^{*}\right]^2}{\mathbb{B}-1}\right]^{1/2} \quad \textbf{(8.37)}$$

where $\hat{\theta}^{*}$ stands for the arithmetic mean over the artificial samples. Even if the artificial sample is drawn from a more complex but known sampling framework, the bootstrap standard error can be easily estimated from standard formulas (cf. Efron 1979; Efron and Tibshirani 1993). If the resampling is conducted on an empirical distribution of a given dataset $\hat{X}$, then it is referred to as a non-parametric bootstrap. In this case, each observation is sampled (with replacement) from the empirical distribution, with probability inversely proportional to the original sample size. However, the resampling can also be selected from a known distribution chosen on an empirical or theoretical basis. In this case, it is referred to as a parametric bootstrap.

Box 8.1 Bootstrap Standard Errors of Adjusted Headcount Ratio and Partial Indices

	Step 1: Bootstrap Samples	**Step 2: Bootstrap Replications of Estimates**
Empirical Distribution (Original Sample)	Resample 1 ⟶	$\left[\hat{M}_0^{*1}, \hat{H}^{*1}, \hat{A}^{*1}, \hat{h}_j^{*1}\right]$
	Resample 2 ⟶	$\left[\hat{M}_0^{*2}, \hat{H}^{*2}, \hat{A}^{*2}, \hat{h}_j^{*2}\right]$
	⋮	⋮
	Resample $\mathbb{b}$ ⟶	$\left[\hat{M}_0^{*\mathbb{b}}, \hat{H}^{*\mathbb{b}}, \hat{A}^{*\mathbb{b}}, \hat{h}_j^{*\mathbb{b}}\right]$
	⋮	⋮
	Resample $\mathbb{B}$ ⟶	$\left[\hat{M}_0^{*\mathbb{B}}, \hat{H}^{*\mathbb{B}}, \hat{A}^{*\mathbb{B}}, \hat{h}_j^{*\mathbb{B}}\right]$

Step 3: Standard Errors	$se_{\mathbb{b},\hat{M}_0} = \left[\frac{1}{\mathbb{B}-1}\sum_{\mathbb{b}=1}^{\mathbb{B}}\left[\hat{M}_0^{*\mathbb{b}} - \hat{M}_0^{*}\right]^2\right]^{1/2}, \hat{M}_0^{*} = \frac{1}{\mathbb{B}}\sum_{\mathbb{b}=1}^{\mathbb{B}} \hat{M}_0^{*\mathbb{b}}$
	$se_{\mathbb{b},\hat{H}} = \left[\frac{1}{\mathbb{B}-1}\sum_{\mathbb{b}=1}^{\mathbb{B}}\left[\hat{H}^{*\mathbb{b}} - \hat{H}^{*}\right]^2\right]^{1/2}, \hat{H}^{*} = \frac{1}{\mathbb{B}}\sum_{\mathbb{b}=1}^{\mathbb{B}} \hat{H}^{*\mathbb{b}}$
	$se_{\mathbb{b},\hat{A}} = \left[\frac{1}{\mathbb{B}-1}\sum_{\mathbb{b}=1}^{\mathbb{B}}\left[\hat{A}^{*\mathbb{b}} - \hat{A}^{*}\right]^2\right]^{1/2}, \hat{A}^{*} = \frac{1}{\mathbb{B}}\sum_{\mathbb{b}=1}^{\mathbb{B}} \hat{A}^{*\mathbb{b}}$
	$se_{\mathbb{b},\hat{h}_j} = \left[\frac{1}{\mathbb{B}-1}\sum_{\mathbb{b}=1}^{\mathbb{B}}\left[\hat{h}_j^{*\mathbb{b}} - \hat{h}_j^{*}\right]^2\right]^{1/2}, \hat{h}_j^{*} = \frac{1}{\mathbb{B}}\sum_{\mathbb{b}=1}^{\mathbb{B}} \hat{h}_j^{*\mathbb{b}}$

Source: Efron and Tibshirani (1993: 47).

Box 8.1 illustrates the use of the bootstrap for computing standard errors of the M_0 and its partial indices. In this case, the statistic θ comprises M_0, H, A, and h_j. Thus, the estimate $\hat{\theta}$ includes $\widehat{M}_0$, $\widehat{H}$, $\hat{A}$, or $\hat{h}_j$. To obtain the bootstrap standard errors, we need to pursue the following steps.

1. Draw $\mathbb{B}$ bootstrap resamples from the empirical distribution function.
2. Compute the set of $\mathbb{B}$ relevant bootstrap estimates of $\widehat{M}_0^{*\mathbb{b}}$, $\widehat{H}^{*\mathbb{b}}$, $\hat{A}^{*\mathbb{b}}$, or $\hat{h}_j^{*\mathbb{b}}$ from each bootstrap sample.
3. Estimate the standard errors by the sampling standard deviation of the $\mathbb{B}$ replications: $se_{\mathbb{b},\widehat{M}_0}$, $se_{\mathbb{b},\widehat{H}}$, $se_{\mathbb{b},\hat{A}}$, or $se_{\mathbb{b},\hat{h}_j}$ (cf. Efron and Tibshirani 1993: 47).

We have already discussed that the bootstrap approach has certain advantages—especially that it does not rely on the central limit theorem. Although the non-parametric bootstrap approach does not depend on any parametric assumptions, it does involve certain choices. The first is the number of replications. Indeed a larger number of replications increas the precision of the estimates, but is costly in terms of time. There are different approaches for selecting the appropriate number of replications (see Poi 2004). The second involves the choice of the bootstrap sample size being selected from the original sample. The third involves the choice of the resampling method. The bootstrap sample size in Efron's traditional bootstrap is equal to the number of observations in the original sample, but the use of smaller sample sizes has also been studied (for further theoretical discussions; see Swanepoel (1986) and Chung and Lee (2001).

Bibliography

Alkire et al. (2010): Alkire, S., Santos, M. E., Seth, S., and Yalonetzky, G. (2010). 'Is the Multidimensional Poverty Index Robust to Different Weights?' *OPHI Research in Progress* 22a, Oxford University.

Alkire, S. and Foster, J. (2011a). 'Counting and Multidimensional Poverty Measurement'. *Journal of Public Economics*, 95(7–8): 476–487.

Alkire, S. and Santos, M. E. (2010). 'Acute Multidimensional Poverty: A New Index for Developing Countries'. *OPHI Working Paper* 38, Oxford University; also published as *Human Development Research Paper* 2010/11.

Alkire, S. and Santos, M. E. (2014). 'Measuring Acute Poverty in the Developing World: Robustness and Scope of the Multidimensional Poverty Index'. *World Development,* 59: 251–274.

Alkire, S. and Seth, S. (2013b). 'Multidimensional Poverty Reduction in India between 1999 and 2006: Where and How?' *OPHI Working Paper* 60, Oxford University.

Anderson, G. (1996). 'Nonparametric Tests of Stochastic Dominance in Income Distributions'. *Econometrica*, 64(5): 1183–1193.

Barrett, G. and Donald, S. G. (2003). 'Consistent Tests for Stochastic Dominance'. *Econometrica*, 71(1): 71–103.

Batana, Y. M. (2013). 'Multidimensional Measurement of Poverty Among Women in Sub-Saharan Africa'. *Social Indicators Research*, 112(2): 337–362.

Biewen, M. (2002). 'Bootstrap Inference for Inequality, Mobility and Poverty Measurement'. *Journal of Econometrics*, 108(2): 317–342.

Boland, P. J. and Proschan, F. (1988). 'Multivariate Arrangement Increasing Functions with Applications in Probability and Statistics'. *Journal of Multivariate Analysis*, 25(2): 286–298.

Cherchye et al. (2007): Cherchye L., Moesen, W., Rogge, N., Puyenbroeck, T.V., Saisana, M., Saltelli, A., Liska, R., and Tarantola, S. (2007). 'Creating Composite Indicators with DEA and Robustness Analysis: The Case of the Technology Achievement Index'. *Journal of the Operational Research Society,* 59(2): 239–251.

Cherchye et al. (2008): Cherchye L., Ooghe, E., and Puyenbroeck, T. V. (2008). 'Robust Human Development Rankings'. *Journal of Economic Inequality,* 6(4): 287–321.

Chung, K. H. and Lee, S. (2001). 'Optimal Bootstrap Sample Size in Construction of Percentile Confidence Bounds'. *Scandinavian Journal of Statistics*, 28(1): 225–239.

Cowell, F. (1989). 'Sampling Variance and Decomposable Inequality Measures'. *Journal of Econometrics* 42(1): 27–41.

Davidson, R. and Duclos, J.-Y. (2000). 'Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality'. *Econometrica*, 68: 1435–1464.

Davidson, R. and Duclos, J.-Y. (2012). 'Testing for Restricted Stochastic Dominance'. *Econometric Reviews*, 32(1): 84–125.

Davidson, R. and Flachaire, E. (2007). 'Asymptotic and Bootstrap Inference for Inequality and Poverty Measures'. *Journal of Econometrics*, 141(1): 141–166.

Deaton, A. (1997). *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*. John Hopkins University Press.

Duclos, J.-Y. and Araar, A. (2006). *Poverty and Equity: Measurement, Policy and Estimation with DAD*. Springer.

Efron, B. (1979). 'Bootstrap Methods: Another Look at the Jackknife'. *The Annals of Statistics*, 7(1): 1–26.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Foster J., McGillivray, M., and Seth, S. (2009). 'Rank Robustness of Composite Indices'. *OPHI Working Paper* 26, Oxford University.

Foster, J. E., McGillivray, M., and Seth, S. (2013). 'Composite Indices: Rank Robustness, Statistical Association and Redundancy'. *Econometric Reviews,* 32(1): 35–56.

Høyland et al. (2012): Høyland, B., Moene, K., and Willumsen, F. (2012). 'The Tyranny of International Index Rankings'. *Journal of Development Economics,* 97(1): 1–14.

Joe, H. (1990). 'Multivariate Concordance'. *Journal of Multivariate Analysis*, 35(1): 12–30.

Kendall, M. G. (1970), *Rank Correlation Methods*, London: Griffin.

Kendall, M. G. and Gibbons, J. D. (1990). *Rank Correlation Method.* E. Arnold.

Lasso de la Vega, M. C. (2010). 'Counting Poverty Orderings and Deprivation Curves', in J. A. Bishop (ed.), *Studies in Applied Welfare Analysis: Papers from the Third ECINEQ Meeting. Research on Economic Inequality* 18, ch. 7.

Mills, A.M. and Zandvakili, S. (1997). 'Statistical Inference via Bootstrapping for Measures of Inequality'. *Journal of Applied Econometrics,* 12(2): 133–150.

Nardo et al. (2005): M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., and Giovannini, E. (2005). 'Handbook on Constructing Composite Indicators: Methodology and User's Guide'. *OECD Statistics Working Papers* 2005/3. OECD Publishing.

Newcombe, R. G. (1998). 'Two-sided Confidence Intervals for the Single Proportion: Comparison of Seven Methods'. *Statistics in Medicine*, 17(8): 857–872.

Nygård, F. and Sandström, A. (1989). 'Income Inequality Measures Based on Sample Surveys'. *Journal of Econometrics*, 42(1): 81–95.

Permanyer I. (2011). 'Assessing the Robustness of Composite Indices Rankings'. *Review of Income and Wealth* 57(2): 306–326.

Permanyer, I. (2012). 'Uncertainty and Robustness in Composite Indices Rankings'. *Oxford Economic Papers* 64(1): 57–79.

Poi, B. P. (2004). 'From the Help Desk: Some Bootstrapping Techniques'. *Stata Journal*, 4(3): 312–328.

Saisana et al. (2005): Saisana, M., Saltelli, A., and Tarantola, S. (2005). 'Uncertainty and Sensitivity Analysis as Tools for the Quality Assessment of Composite Indicators'. *Journal of the Royal Statistical Society: Ser. A* (*Statistics in Society*), 168(2): 307–323.

Seth, S. (2013). 'A Class of Distribution and Association Sensitive Multidimensional Welfare Indices'. *Journal of Economic Inequality*, 11(2): 133–162.

Swanepoel, J. W H. (1986). 'A Note on Proving that the (Modified) Bootstrap Works'. *Communications in Statistics* (*Theory and Methods*) 15(11): 3193–3203.

Ura *et al.* (2012): Ura, K., Alkire, S., Zangmo, T, and Wangdi, K. (2012). *An Extensive Analysis of The Gross National Happiness Index*. Centre of Bhutan Studies.

Wolff et al. (2011): Wolff, H., Chong, H., and Auffhammer, M. (2011). 'Classification, Detection and Consequences of Data Error: Evidence from the Human Development Index'. *The Economic Journal,* 121(553): 843–870.

Yalonetzky, G. (2014). 'Conditions for the Most Robust Multidimensional Poverty Comparisons Using Counting Measures and Ordinal Variables'. *Social Choice and Welfare*. Published online February.