Estimation of Inequality Indices of the Cumulative Distribution Function

Ramses H. Abul Naga [*] Christopher Stapenhurst [†]

November 5, 2014

Abstract

Inequality indices for self-assessed health and life satisfaction are typically constructed as functions of the cumulative distribution function. We present a unified methodology for the estimation of the resulting inequality indices. We also obtain explicit standard error formulas in the context of two popular families of inequality indices that have emerged from this literature.

Keywords: Ordered response data, self-assessed health, multinomial sampling, large sample distributions, standard errors.

JEL codes: $D63, C43$, $I1$

[*] Business School and Health Economics Research Unit, University of Aberdeen, Aberdeen AB24 3QY, United Kingdom.

[†] Office of National Statistics, NewPort, United Kingdom.

A literature on the measurement of inequality in relation to ordered response data has emerged in the last ten years following the work of Allison and Foster (2004). A large body of theoretical literature has ensued, using the cumulative distribution function as the main argument of the underlying ethical index.

Some authors (e.g. Apouey 2007, Cowell and Flachaire 2012) have derived standard errors for the inequality indices they have introduced. The present work complements these papers in that it presents a unified methodology for the estimation of inequality indices of the cumulative distribution function.

1. Framework

Consider data on k ordered states of well-being (for example self-reported health status or more generally life satisfaction). We gather the responses $(n_1, ..., n_k)$ of n individuals from an underlying population $p = (p_1, ..., p_k)$ in a frequency distribution $\hat{p} = (\hat{p}_1, ..., \hat{p}_k)$ where $\hat{p}_i = n_i/n$ is the proportion of individuals who are in class i, and such that $\sum_{i=1}^{k} n_i = n$. We denote $\hat{P} = (\hat{P}_1, ..., \hat{P}_k)$ the resulting cumulative distribution, where $\hat{P}_j = \sum_{i=1}^{j} \hat{p}_i$, and we let $\mathbb{D}$ denote the set of cumulative distribution functions. An inequality index for ordered response data is then some function $F : \mathbb{D} \to \mathbb{R}_+$ with parameters reflecting some appropriately defined inequality aversion axiom and other ethical properties. Two difficulties arise in developing inference for ethical indices of the cumulative distribution. Firstly, the data analyst is confronted with functions of counts or frequencies rather than the usual moments of a continuous variable that are common in the income inequality literature (Cowell, 1999), and secondly the ethical index will rarely present itself in the form of a linear function of the cumulative distribution (though see below).

Let $m \in \{1, ..., k\}$ denote the median response state in some given distribution $P \in \mathbb{D}$. First, to give an example of an inequality index that is linear in the cumulative distribution function, consider the family of sub-group decomposable indices of Kobus and Miłoś (2012):

$$
\begin{aligned}
\Lambda_{a,b}(P) &= \frac{a\sum_{i=1}^{m-1} P_i - b\sum_{i=m}^{k} P_i + c_1\left(k,m,a,b\right)}{c_2\left(k,m,a,b\right)} \qquad a,b \geq 0 && (1.1)\\
c_1(k,m,a,b) &= b(k+1-m) && (1.2)\\
c_2(k,m,a,b) &= (m-1)\frac{a}{2} - (k+2)\frac{b}{2} + c_1 && (1.3)
\end{aligned}
$$

Here a and b are parameter values chosen by the data analyst in order to reflect different social value judgements regarding inequality below, and above, the median health state m, and c_1 and c_2 are normalization constants. Next consider the *alphabeta* family of inequality indices (Abul Naga and Yalcin, 2008):

$$
\begin{aligned}
\Delta_{\alpha,\beta}(P) &= \frac{\sum_{i=1}^{m-1} P_i^{\alpha} - \sum_{i=m}^{k} P_i^{\beta} - c_3(k,m,\alpha,\beta)}{c_4(k,m,\alpha,\beta)} \qquad \alpha,\beta \geq 1 && (1.4)\\
c_3(k,m,\alpha,\beta) &= k+1-m && (1.5)\\
c_4(k,m,\alpha,\beta) &= (m-1)\left(\frac{1}{2}\right)^{\alpha} - (k-m)\left(\frac{1}{2}\right)^{\beta} - 1 + c_3 && (1.6)
\end{aligned}
$$

Likewise, α and β are parameter values chosen to reflect social aversion to inequality below and above the median, and c_3 and c_4 are constants. Note that the index $\Delta_{\alpha,\beta}(P)$ is only linear in P in the specific case where $\alpha = \beta = 1$, and furthermore that $\Delta_{1,1}(P) = \Lambda_{1,1}(P)$ for any distribution P. The above indices feature in studies aimed at quantifying health inequality in multiple country contexts (e.g. Jones et al. 2011) and also in simulating the envisaged effect of policy interventions on health inequality in the context of specific pathologies (e.g. Arrighi et al. 2015).

2. Large sample distribution

The estimation of inequality indices of the type considered in this paper can be treated in a unified framework as an estimation of some non-linear function $F(.)$ of an unknown cumulative distribution function P_0, with associated probability distribution p_0. The Analogy Principle then guarantees that under appropriate assumptions $F(\hat{P})$ will result in a consistent estimator of $F(P_0)$.

Let cov(y) denote the covariance matrix of some random vector y and let Ω_0 :=cov$(n^{\frac{1}{2}}\hat{p})$. Since individuals select one and only one of k possible responses, the covariance matrix Ω_0 is said to arise from a context of *multinomial sampling.* That is, writing $p_0 = (p_1, \ldots, p_k)$, we have that Ω_0 is the following function of the vector p_0 :

$$\Omega_0 = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & \cdots & -p_1p_k \\ -p_2p_1 & p_2(1-p_2) & -p_2p_3 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ -p_kp_1 & -p_kp_2 & \cdots & p_k(1-p_k) \end{bmatrix}. \tag{2.1}$$

Observe then that, as a resulting of multinomial sampling, the covariance matrix Ω_0 will always be finite and positive semi-definite [1].

We next define the $k-$dimensional Jacobian vector of the transformation F as follows:

$$J = \begin{pmatrix} \partial F/\partial P_1 & \cdots & \partial F/\partial P_k \end{pmatrix}, \tag{2.2}$$

and throughout the paper we maintain the following assumptions:

$[A1]$ There is a finite number k of ordered states.

$[A2]$ The n independent responses $(n_1, ..., n_k)$ defining the vector of frequencies $\hat{p} = (\hat{p}_1, ..., \hat{p}_k)$ are jointly distributed from a multinomial distribution with parameters n and p_0, and such that cov$(n^{\frac{1}{2}}\hat{p}) = \Omega_0$, where Ω_0 is a $k \times k$ positive semi-definite matrix.

$[A3]$ The function $F : \mathbb{D} \to \mathbb{R}_+$ does not involve n and is continuously differentiable at the population distribution P_0.

Our purpose here is to obtain the large sample distribution of the sample estimator $F(\hat{P})$ as a function of $F(P_0)$. The following result (see for instance Anderson, 1996) will prove useful:

Lemma 1 *Under* $[A1]$ *and* $[A2]$ *the vector of frequencies* $\hat{p}$ *converges to a* $k-$*variate normal distribution such that:*

$$n^{1/2}(\hat{p} - p_0) \longrightarrow \mathcal{N}(0; \Omega_0). \tag{2.3}$$

[1] Specifically, because $p_1 + \cdots + p_k = 1$, the matrix Ω_0 will have a rank equal to $k - 1$.

Because by construction $\sum_{i=1}^{k} \hat{p}_i = 1$, the resulting large sample distribution of $\hat{p}$ in Lemma 1 above is degenerate. Nonetheless, the large sample distribution of the inequality index $F(\hat{P})$ is non-degenerate:

Proposition 2 *Under $[A1 - A3]$ the sample inequality index $F(\hat{P})$ converges to a univariate normal distribution such that:*

$$n^{1/2}\left(F(\hat{P}) - F(P_0)\right) \longrightarrow \mathcal{N}\left(0;\ J_0 L\Omega L' J_0'\right) \tag{2.4}$$

where $J_0 := J(P_0)$ is the Jacobian vector (2.2) *evaluated at P_0.*

Proof Define the $k \times k$ lower-triangular matrix $L = \{l_{st}\}$ such that $l_{st} = 0$ if $s < t$ and $l_{st} = 1$ if $s \geq t$. Then L is a summation matrix such that $\hat{P} = L\widehat{p}$ and it follows straightforwardly from Lemma 1 that $n^{1/2}\left(\hat{P} - P_0\right)$ converges to a $k-$variate normal distribution $\mathcal{N}\left(0;\ L\Omega_0 L'\right)$. Also, from $[A1 - A2]$, the Law of Large Numbers entails that $\hat{P}$ converges in probability to P_0, whilst $[A3]$ entails that $J(\hat{P})$ converges in probability to $J(P_0)$. It then follows from the *delta method* that $n^{1/2}\left(F(\hat{P}) - F(P_0)\right)$ converges to a normal distribution $\mathcal{N}\left(0;\ J_0 L\Omega L' J_0'\right)$. □

3. Jacobian vectors and standard errors

In the light of (2.4) in Proposition 2, the asymptotic distribution of $F(\hat{P})$ involves a quadratic form in the Jacobian vector $J(.)$, evaluated at P_0.

To clarify this point, define the matrix $V_0 = L\Omega_0 L'$. Then the asymptotic variance of $F(\hat{P})$ in Proposition 2 takes the form $J(P_0)V_0 J'(P_0)$, where V_0 is a positive semi-definite matrix. Consider the estimator $\hat{\Omega} = \{\hat{\omega}_{il}\}$ with generic element

$$\hat{\omega}_{il} = \begin{cases} \dfrac{n_i}{n}\left(1 - \dfrac{n_i}{n}\right) & i = l \\ -\dfrac{n_i n_l}{n^2} & i \neq l \end{cases} \tag{3.1}$$

and furthermore, define the matrix

$$\hat{V} = L\hat{\Omega}L'. \tag{3.2}$$

Then $\hat{\Omega}$ is a consistent estimator of Ω_0 and likewise $\hat{V}$ is a consistent estimator of V_0. In the light of Proposition 2 we can then write the asymptotic standard error of $F(\hat{P})$ as the following expression [2]:

$$\mathrm{se}[F(\hat{P})] = \left(\frac{1}{n}J(\hat{P})\hat{V}J'(\hat{P})\right)^{1/2}. \tag{3.3}$$

The estimator $\hat{V}$ is generally applicable in the context of inequality indices of the cumulative distribution function. However, the standard error formula (3.3) clearly requires differentiability of the function F so as to enable the researcher to evaluate the Jacobian vector $J(.)$. The Jacobian vector, however, will vary depending on the structure of the inequality index.

3.1. Indices that are decomposable by sub-groups

Let $m \in \{1, ..., k\}$ denote the median response state in some given distribution $P = (P_1, ..., P_n) \in \mathbb{D}$ and return to the class (1.1) of sub-group decomposable inequality indices of the cumulative distribution, introduced by Kobus and Miłoś (2012). As a corollary to Proposition 2, we derive the form of the Jacobian vector in relation to (1.1):

Corollary 3 *For the class of inequality indices of the cumulative distribution function* (1.1) *that are decomposable by population sub-groups, the Jacobian vector J evaluated at some distribution $P \in \mathbb{D}$ takes the form*

$$J(P) = \frac{1}{c_2(k, m, a, b)}\left(\ a\iota_{m-1}, \quad -b\iota_{k+1-m}\ \right) \tag{3.4}$$

where ι_q is a $q-$dimensional row vector of ones, and where c_2 is the constant defined under (1.3).

[2] Alternatively, we may write the standard error formula in expanded form. Let $J(\hat{P}) = (\hat{J}_1, ..., \hat{J}_k)$ and observe that the element $\hat{v}_{st}$ of the matrix $\hat{V}$ is of the form

$$\hat{v}_{st} = \sum_{i=1}^{s}\sum_{l=1}^{t}\hat{\omega}_{ij}$$

Then, the standard error (3.3) of the inequality index may be evaluated as follows

$$\mathrm{se}[F(\hat{P})] = \left(\frac{1}{n}\sum_{i=1}^{k-1}\sum_{l=1}^{k-1}\hat{J}_i\hat{J}_l\hat{v}_{il}\right)^{1/2}$$

Note that the double sum is evaluated from 1 to $k-1$ as the last row and column of $\hat{V}$ are both null vectors.

3.2. The alphabeta family of inequality indices

Unlike the class of decomposable inequality indices (1.1), the function $F(.)$ underlying the *alphabeta* family $\Delta_{\alpha,\beta}$ of (1.4) is a non-linear function of P when the parameters α and β are strictly greater than 1. There, the Jacobian vector will involve the distribution P :

Corollary 4 *For the alphabeta family of inequality indices of the cumulative distribution function* (1.4)*, the Jacobian vector J evaluated at some distribution $P \in \mathbb{D}$ takes the form*

$$J(P) = \frac{1}{c_4(k,m,a,b)} \left(\alpha P_1^{\alpha-1}, \ \cdots, \ \alpha P_{m-1}^{\alpha-1}, \ -\beta P_m^{\beta-1}, \ \cdots, \ -\beta P_k^{\beta-1} \right), \tag{3.5}$$

where c_4 is the constant defined under (1.6)*.*

4. An illustrative example

Consider data on five ordered nutritional health states from the Egyptian Integrated Household Survey of 1997-1999 [3]. The data refer to two statistical areas of Northern Egypt (also known as *Lower* Egypt), namely Metropolitan Lower Egypt (MLE) and Non-Metropolitan Lower Egypt (NMLE). The resulting cumulative distributions are respectively $\hat{P} = (0.075, 0.187, 0.430, 0.812\ 1.00)$ for the MLE data ($n = 107$) and $\hat{Q} = (0.040, 0.144, 0.363, 0.667\ 1.00)$ for the NMLE data ($n = 452$). Note also that the median response state is $m = 4$ in both distributions.

Table 1 reports inequality estimates and standard errors, calculated using the *alphabeta* family for various inequality aversion parameter combinations (α, β). Note that the first set of calculations pertaining to the pair $(\alpha, \beta) = (1, 1)$ are also estimates of inequality in the family of sub-group decomposable inequality indices $\Lambda_{a,b}(P)$ of (1.1), where $a = b = 1$.

Observe that all inequality estimates reported in the table are statistically significant. Furthermore, for all five (α, β) pairs, the inequality estimate is somewhat smaller in Metropolitan Lower Egypt. The inferential framework we have developed in this paper is well adapted to test for equality of the estimates pertaining to a given row of the table [4]. The test statistic of the last column of Table 1 does

[3] The health states in ascending order (from state 1 to state 5) are the following: type-III obese, type-II obese, type-I obese, overweight and not overweight.

[4] Consider two sample cumulative distributions $\hat{P}, \hat{Q} \in \mathbb{D}$, and some inequality index $F(.)$.

not reject the null hypothesis of equality of health dispersion levels in the two regions, in relation to the first three sets of estimates. However, it is only in the context of parameter values $(\alpha, \beta) = (2, 2)$ and $(\alpha, \beta) = (2, 4)$ that the difference in the inequality estimates is statistically different from zero.

5. References

Abul Naga R. and T. Yalcin (2008): ”Inequality Measurement for Ordered Response Health Data” *Journal of Health Economics* 27, 1614-1625.

Allison R. A. and Foster J. (2004): ”Measuring Health Inequalities Using Qualitative Data”, *Journal of Health Economics* 23, 505-524.

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

Apouey B. (2007): ”Measuring Health Polarization with Self-Assessed Health Data”, *Health Economics* 16, 875-894.

Arrighi Y., M. Abu-Zaineh and B. Ventelou (2015): “To Count or not to Count Deaths: Reranking Effects in Health Distribution Evaluation”, *Health Economics*, forthcoming.

Cowell, F. (1999) ”Estimation of inequality indices” in Silber, J. (ed.) *Handbook on Income Inequality Measurement,* Kluwer, Dewenter.

Cowell F. and E. Flachaire (2012): ”Inequality with Ordinal Data”, Manuscript.

Jones A., N. Rice, S. Robone and P. Rosa Dias (2011): ” Inequality and Polarization in Health Systems Responsiveness: a Cross-Country Analysis”, *Journal of Health Economics* 30, 616-625.

Kobus M. and P. Miłoś (2012): "Inequality Decomposition by Population Subgroups for Ordinal Data", *Journal of Health Economics* 31, 15-21.

Under the null hypothesis that health inequality levels are identical in P and Q ($H_0 : F(P) = F(Q)$), the test statistic z defined as follows:

$$z := \frac{F(\hat{P}) - F(\hat{Q})}{\left[\left(\mathrm{se}[F(\hat{P})]\right)^2 + \left(\mathrm{se}[F(\hat{Q})]\right)^2\right]^{1/2}} \tag{4.1}$$

is distributed in large samples as a $\mathcal{N}(0, 1)$ variate. The test thus rejects H_0 at the 5% level when $|z| > 1.96$.

Table 1: *Inequality in Nutritional Health in Lower Egypt*

(α, β)	MLE	$NMLE$	$Test$
$(1, 1)$	0.439 (0.042)	0.440 (0.018)	−0.022
$(1, 2)$	0.458 (0.040)	0.490 (0.016)	−0.746
$(2, 1)$	0.330 (0.041)	0.390 (0.017)	−1.387
$(2, 2)$	0.376 (0.042)	0.474 (0.017)	−2.165
$(2, 4)$	0.467 (0.046)	0.567 (0.014)	−2.108

1) The inequality estimates pertain to the inequality measure (1.4) with parameters (α, β). Standard errors are reported inside parentheses.

2) There are $n = 107$ observations pertaining the MLE sample (Metropolitan Lower Egypt) and $n = 452$ in the NMLE sample (non-Metropolitan Lower Egypt).

3) The test statistic of equality of estimates is as defined in (4.1) of footnote 4. The critical (absolute) value of the test at the 5% level is equal to 1.96.