Economics Letters 130 (2015) 109–112

Contents lists available at ScienceDirect

Economics Letters

journal homepage: www.elsevier.com/locate/ecolet

Estimation of inequality indices of the cumulative distribution function

Ramses H. Abul Naga [a,*], Christopher Stapenhurst [b]

[a] Business School and Health Economics Research Unit, University of Aberdeen, Aberdeen AB24 3QY, United Kingdom
[b] Office of National Statistics, Cardiff Road, Newport NP10 8XG, United Kingdom

HIGHLIGHTS

- We study inequality indices of the cumulative distribution function.
- We derive their large sample distribution in a general framework.
- We provide a method for obtaining analytic standard errors.
- We give examples in the context of two families of inequality indices.
- We provide an application using Egyptian demographic and health data.

ARTICLE INFO

Article history:
Received 10 January 2015
Received in revised form
28 February 2015
Accepted 2 March 2015
Available online 8 March 2015

JEL classification:
D63
C43
I1

Keywords:
Ordered response data
Self-assessed health
Multinomial sampling
Large sample distributions
Standard errors

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.

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 $(n_1, \ldots, n_k)$ of n individuals from an underlying population $p = (p_1, \ldots, p_k)$ in a frequency distribution $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_k)$ 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 $\hat{P} = (\hat{P}_1, \ldots, \hat{P}_k)$ 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

* Corresponding author. Tel.: +44 1224 272709.
E-mail address: r.abulnaga@abdn.ac.uk (R.H. Abul Naga).

http://dx.doi.org/10.1016/j.econlet.2015.03.004
0165-1765/

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 $m \in \{1, \ldots, k\}$ 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):

$$\Lambda_{a,b}(P) = \frac{a \sum_{i=1}^{m-1} P_i - b \sum_{i=m}^{k} P_i + c_1(k, m, a, b)}{c_2(k, m, a, b)} \quad a, b \geq 0 \tag{1.1}$$

$$c_1(k, m, a, b) = b(k + 1 - m) \tag{1.2}$$

$$c_2(k, m, a, b) = (m - 1)\frac{a}{2} - (k + 2 - m)\frac{b}{2} + c_1. \tag{1.3}$$

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):

$$\Delta_{\alpha,\beta}(P) = \frac{\sum_{i=1}^{m-1} P_i^{\alpha} - \sum_{i=m}^{k} P_i^{\beta} + c_3(k, m, \alpha, \beta)}{c_4(k, m, \alpha, \beta)} \quad \alpha, \beta \geq 1 \tag{1.4}$$

$$c_3(k, m, \alpha, \beta) = k + 1 - m \tag{1.5}$$

$$c_4(k, m, \alpha, \beta) = (m - 1)\left(\frac{1}{2}\right)^{\alpha} - (k - m)\left(\frac{1}{2}\right)^{\beta} - 1 + c_3. \tag{1.6}$$

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 Continuous Mapping Theorem then guarantees that under appropriate assumptions $F(\hat{P})$ will result in a consistent estimator of $F(P_0)$.

Let $\text{cov}(y)$ denote the covariance matrix of some random vector y and let $\Omega_0 := \text{cov}(n^{\frac{1}{2}}\hat{p})$. 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:

$$J = \left(\partial F/\partial P_1 \quad \cdots \quad \partial F/\partial P_k\right), \tag{2.2}$$

and throughout the paper we maintain the following assumptions:

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

[A2] The n independent responses $(n_1, \ldots, n_k)$ defining the vector of frequencies $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_k)$ are jointly distributed from a multinomial distribution with parameters n and p_0, and such that $\text{cov}(n^{\frac{1}{2}}\hat{p}) = \Omega_0$, 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:*

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

Because by construction $\sum_{i=1}^{k} \hat{p}_i = 1$, the resulting large sample distribution of $\hat{p}$ in Lemma 1 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:*

$$n^{1/2}\left(F(\hat{P}) - F(P_0)\right) \longrightarrow \mathcal{N}\left(0;\ J_0 L \Omega_0 L' J_0'\right) \tag{2.4}$$

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 $\hat{P} = L\hat{p}$ 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 $\mathcal{N}\left(0;\ J_0 L \Omega_0 L' J_0'\right)$. □

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

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