econometrics

Article

Asymptotic Versus Bootstrap Inference for Inequality Indices of the Cumulative Distribution Function

Ramses Abul Naga [1,*], Christopher Stapenhurst [2] and Gaston Yalonetzky [3]

[1] Business School, University of Aberdeen, Old Aberdeen AB24 3UE, UK
[2] Department of Economics, University of Edinburgh, Edinburgh EH8 9AB, UK; c.stapenhurst@ed.ac.uk
[3] Business School, University of Leeds, Leeds LS2 9JT, UK; g.yalonetzky@leeds.ac.uk
* Correspondence: r.abulnaga@abdn.ac.uk.

Received: 12 December 2017; Accepted: 18 February 2020; Published: 26 February 2020

Abstract: We examine the performance of asymptotic inference as well as bootstrap tests for the Alphabeta and Kobus–Miłoś family of inequality indices for ordered response data. We use Monte Carlo experiments to compare the empirical size and statistical power of asymptotic inference and the Studentized bootstrap test. In a broad variety of settings, both tests are found to have similar rejection probabilities of true null hypotheses, and similar power. Nonetheless, the asymptotic test remains correctly sized in the presence of certain types of severe class imbalances exhibiting very low or very high levels of inequality, whereas the bootstrap test becomes somewhat oversized in these extreme settings.

Keywords: measurement of inequality; ordered response data; multinomial sampling; large sample distributions; Studentized bootstrap tests; monte carlo experiments

1. Introduction

Inequality indices expressed as functions of the cumulative distribution function (CDF) are routinely used in studies that quantify inequality in self-assessed health, happiness, and other life satisfaction variables that are collected in the form of ordered response data. The large sample distribution of inequality indices of the CDF has been obtained by Abul Naga and Stapenhurst (2015). It is, however, an open question as to how reliable the adoption of the large sample distribution in testing hypotheses in applied work involving finite samples sizes is. Such an investigation (the first of this kind according to our best knowledge) is the main purpose of this paper. The focus of our investigation will be the Alphabeta family of inequality indices (Abul Naga and Yalcin 2008) further extended by Kobus and Milos (2012). This family of indices has been associated with an important empirical literature (Arrighi et al. 2015; Dutta and Foster 2013; Jones et al. 2011; Madden 2010). This research is motivated by an important body of work, reviewed in Cowell and Flachaire (2015); Davidson and Duclos (2013), that documents poor finite sample performance of the *t*-statistic of income inequality measures.

The inequality indices we investigate in this paper are smooth statistics of multinomial distributions. It is well known that for smooth statistics of multinomial distributions, the normal approximation provided by the central limit theorem is generally very accurate as sample size increases with the number of probability categories fixed, and as long as the underlying probability distribution lies in the interior of its parameter space[1].

[1] Note that here the parameter space is a space of probability distributions.

Nonetheless, there are several reasons a priori why we may want to consider undertaking bootstrap type inference for inequality indices defined on ordered response data. Firstly, we note that such indices are typically non-linear functions. The analytical calculation of their standard errors involves linearization using the *delta* method. Davison and Hinkley (1997, p. 16) argue that simulation methods in practice can provide more accurate estimates of the distribution of test statistics than analytical methods that rely on the delta method. Also, under certain regularity assumptions, statistical theory will always recommend bootstrap inference over asymptotic tests in the context of *asymptotically pivotal* test statistics and in presence of finite samples (Horowitz 2001). Furthermore, in the income inequality literature, bootstrap methods often prove useful in cases where the exact distribution of the test is very complicated to obtain analytically (e.g., Barret et al. 2014).

Bootstrap methods, however, need not always be recommended over asymptotic inference in all sampling contexts. For instance, Athreya (1987) among others shows that bootstrap methods may fail when resampling occurs from long-tailed distributions. Likewise, in the context of income data, Russell and Flachaire (2007) show that the presence of outliers severely disrupts the performance of bootstrap inference in relation to income inequality measures.

A priori therefore, it is not clear which of bootstrap methods and asymptotic inference is to be preferred in economic investigations of inequality in relation to ordered response data. We therefore use Monte Carlo experiments to compare the empirical size and statistical power of asymptotic inference and the Studentized bootstrap test, in the context of a generic hypothesis test that a population has a given level of inequality t_0. We find that, in a broad variety of settings, both tests have similar rejection probabilities of true null hypotheses, and similar power. Nonetheless, the asymptotic test remains correctly sized in presence of certain types of severe class imbalances[2] exhibiting very low or very high levels of inequality, whereas the bootstrap test becomes somewhat oversized in these extreme settings. Given that the simulation results suggest that in practice the two tests perform very similarly, we are inclined to recommend the use of asymptotic inference in applied research involving the use of inequality indices of the cumulative distribution function.

The structure of the paper is as follows. Section 2 presents the two families of inequality indices for ordered response data that will be the subject of our investigation. Section 3 discusses the implementation of the Studentized bootstrap test in relation to the two families of inequality indices. In the discussion, we place particular emphasis on ensuring that sampling follows the two golden rules set out by Davidson (2007). Section 4 discusses the methods of investigation used in exploring the comparative properties of the two tests. Section 5 presents the results of the Monte Carlo simulations. Section 6 presents a brief application, while Section 7 concludes with a discussion of the limitations of the paper and directions for further research.

2. The Family of Inequality Indices

Let $\mathcal{S}$ denote a sample of observations on k ordered categories of well-being (for example life satisfaction, self-reported health status or obesity status). Assume n_1 individuals are reported to be in category 1, n_2 individuals are reported to be in category 2, etc., and define $n = \sum_{i=1}^{k} n_i$. The resulting sample then follows a multinomial distribution. If $\mathcal{S}$ is drawn from an underlying population probability mass function (PMF) $f = (f_1, ..., f_k)$, the likelihood of $\mathcal{S}$ takes the form

$$\mathbb{P}(\mathcal{S} \mid f) := \frac{n!}{n_1! \cdot \dots \cdot n_k!} f_1^{n_1} \cdot \dots \cdot f_k^{n_k}. \tag{1}$$

[2] Severe class imbalances occur when the share of probability mass allocated to the various probability states is unevenly distributed, and far from the uniform case presenting *perfectly balanced classes.*

We note that, in the context of multinomial sampling, the data counts $(n_1, ..., n_k)$ are jointly sufficient statistics for the sample $\mathcal{S}$ (see Yalonetzky 2013, for further detail).

Inequality indices for ordered response data are often expressed in terms of the cumulative distribution associated with the sample $\mathcal{S}$. We denote the sample's frequency distribution $x = (x_1, \cdots, x_k)$ where $x_i = n_i/n$ is the proportion of individuals who are in category i. We define $X = (X_1, ..., X_k)$ as the resulting empirical cumulative distribution, where $X_j := \sum_{i=1}^{j} n_i/n$, and we call X the empirical distribution function (EDF).

Let $\mathbb{D}$ denote the set of cumulative distributions defined over k ordered categories of well-being. An inequality index for ordered response data is then some function defined on $\mathbb{D}$, with parameters reflecting some appropriately defined inequality aversion axiom and other ethical properties. 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):

$$\Xi(X; m, a_1, a_2) := \frac{a_1 \sum_{i=1}^{m-1} X_i - a_2 \sum_{i=m}^{k} X_i + c_1(k, m, a_1, a_2)}{c_2(k, m, a_1, a_2)} \qquad a_1, a_2 \geq 0 \tag{2}$$

Here a_1 and a_2 are parameter values chosen by the data analyst in order to reflect different social value judgements regarding inequality below, and above, the median category m [3] and c_1 and c_2 are normalization constants that insure that the index takes values in the unit interval $[0, 1]$.

Next consider the *Alphabeta* family of inequality indices (Abul Naga and Yalcin, 2008):

$$\Delta(X; m, \omega_1, \omega_2) := \frac{\sum_{i=1}^{m-1} X_i^{\omega_1} - \sum_{i=m}^{k} X_i^{\omega_2} + c_3(k, m, \omega_1, \omega_2)}{c_4(k, m, \omega_1, \omega_2)} \qquad \omega_1, \omega_2 \geq 1 \tag{3}$$

Likewise ω_1 and ω_2 are parameter values chosen to reflect social aversion to inequality below and above the median, and $c_3 := k + 1 - m$ and $c_4 := (m-1)\left(\frac{1}{2}\right)^{\omega_1} - (k-m)\left(\frac{1}{2}\right)^{\omega_2} + (k-m)$ are normalization constants. Note that the index $\Delta(X; m, \omega_1, \omega_2)$ is only linear in X in the specific case where $\omega_1 = \omega_2 = 1$, and furthermore that $\Delta(X; m, 1, 1) = \Xi(X; m, 1, 1)$ for any distribution X.

The key property associated with the two families of inequality indices presented above is that they are increasing in median preserving spreads (Allison and Foster, 2004)[4]. The above indices feature in studies aimed at quantifying health inequality in multiple country contexts (e.g., Jones et al. 2011; Madden 2010) 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). They are also used in the quantification of happiness inequality in the United States (Dutta and Foster 2013).

3. The Bootstrap and Asymptotic Test

The purpose of this section is to detail the procedures used to implement the asymptotic and bootstrap tests. We can think of a generic statistical test as a function $\tau(H_o, \mathcal{S}) \mapsto [0, 1]$, where H_o is a hypothesis being investigated and $\mathcal{S}$ is a sample. The test function returns a p-value giving the lowest bound on the type-1 error rate for which we can reject the hypothesis of concern.

Let Λ_j denote the subset of $\mathbb{D}$ containing distributions with a median status j. Then $\Lambda_j := \{G \in \mathbb{D} : med(G) = j\}$ and it is clear that $\mathbb{D} = \Lambda_1 \cup \cdots \cup \Lambda_k$. Accordingly, for some sample $\mathcal{S}$ of n observations

[3] Set $X_0 = 0$, $X_k = 1$, and $X = (X_1, \cdots, X_k)$. The median socio-economic status category associated with X, $\text{med}(X) = m \in \{1, ..., k\}$, is an index such that $X_{m-1} \leq 0.5$ and $X_m \geq 0.5$. We note that the median status is uniquely defined in samples where n is an odd number, and accordingly we work with odd sample sizes throughout our simulation exercises.

[4] Consider two empirical distributions $X, Y \in \mathbb{D}$. The median preserving spreads partial ordering $(\mathbb{D}, \preceq_{AF})$ ranks X more egalitarian than Y, written $X \preceq_{AF} Y$, if three conditions are satisfied. These are (1) $med(X) = med(Y) = m$, (2) $X_i \leq Y_i$ for all $i < m$ and (3) $X_i \geq Y_i$ for all $i \geq m$.

with EDF X, drawn from a cumulative distribution function $F \in \Lambda_m$, and some inequality index $\Delta(.; m, \omega_1, \omega_2)$, consider testing the null hypothesis

$$H_o : \Delta(F; m, \omega_1, \omega_2) = t_o. \tag{4}$$

We define the null space associated with this null hypothesis as the set of cumulative distribution functions $G \in \Lambda_m$ that exhibit the level of inequality t_o, and whose median equals the median m of the cumulative distribution F :

$$E(t_o, \Lambda_m; \omega_1, \omega_2) := \{G \in \Lambda_m : \Delta(G; m, \omega_1, \omega_2) = t_o\} \tag{5}$$

Let $V[\Delta(X; m, \omega_1, \omega_2)]$ denote any consistent estimator of the asymptotic variance of the inequality index. Conditional on the median category m, an asymptotic test of the null hypothesis that $\Delta(F; m, \omega_1, \omega_2) = t_o$ involves computing the test statistic

$$z := \frac{\Delta(X; m, \omega_1, \omega_2) - t_o}{(V[\Delta(X; m, \omega_1, \omega_2)])^{1/2}}, \tag{6}$$

and approximating the exact distribution of z under the null hypothesis by a standard normal distribution[5]. Equivalently, for the asymptotic test, the critical values are obtained from the quantiles of the standard normal distribution.

As an alternative to asymptotic inference, bootstrap procedures simulate the distribution of the test via calculation of the test statistic in a large number of samples drawn from a distribution in the null space. Typically the null space (5) contains many distributions, yet one way of ensuring a successful implementation of the bootstrap is to insure that sampling follows the related two golden rules (Davidson 2007). The first of these rules requires that the data generating process underlying the bootstrap samples must belong to the model underlying the null hypothesis H_o. The second golden rule requires that the data generating process underlying the bootstrap samples be obtained under the null hypothesis from an efficient estimation procedure.

Because the Maximum Likelihood estimator is generally efficient, it is of particular relevance, if possible, to use an ML procedure to select the underlying model of the null hypothesis, and to generate the bootstrap samples from this data generating process. This method of investigation, known as the *parametric bootstrap*, ensures that both golden rules of bootstrap inference are satisfied (Davidson 2007).

Associate with the estimation sample $\mathcal{S}$ the vector of responses $(n_1, ..., n_k)$. The ML estimator of the data generating process underlying H_o is obtained by maximizing the sample likelihood (1) in the null space $E(t_o, \Lambda_m; \omega_1, \omega_2)$. Associate a probability mass function f with the cumulative distribution F. The ML estimator $\tilde{f}$ of the data generating process underlying the null hypothesis of interest is then chosen as the maximizer of $\mathbb{P}(\mathcal{S} \mid f)$ in the null space $E(t_o, \Lambda_m; \omega_1, \omega_2)$; that is,

From here on, the ML estimator of the null hypothesis is used to generate $b = 1, ..., B$ bootstrap samples $\widetilde{\mathcal{S}}_b$ resulting in empirical distributions $\widetilde{X}_1, ..., \widetilde{X}_B$ and a sequence of test statistics

$$\tilde{z}_b := \frac{\Delta(\widetilde{X}_b; m, \omega_1, \omega_2) - \Delta(X; m, \omega_1, \omega_2)}{(V^{BOOT})^{1/2}} \tag{7}$$

Here, V^{BOOT} is the empirical variance of the B values $\Delta(\widetilde{X}_b; j, \omega_1, \omega_2)$ of the level of inequality. The sample quantiles of the B bootstrap statistics $\tilde{z}_b$ are used instead of the quantiles of the standard

[5] See Abul Naga and Stapenhurst (2015) for a derivation of a consistent estimator of the asymptotic variance of the inequality index.

normal distribution in order to provide critical values for hypotheses tests related to the level of inequality in the underlying population.

4. Methods

Consider testing a null hypothesis of the generic form (4) using each of the asymptotic and bootstrap test discussed in the previous section. Call these two testing procedures τ^A and τ^B respectively. In this section we discuss how to use Monte Carlo simulation methods to evaluate the size and power properties of the asymptotic and bootstrap tests.

In order to obtain a good understanding of the comparative size and power properties of the two statistical tests, our interest in the Monte Carlo experiments will be to investigate for each of the two procedures, the effect of varying the parameters F, ω_1, ω_2 and t_o of the generic test (4), in relation to different sample sizes. Our interest in examining the effect of sample size is to explore which of the two estimation procedures is to be recommended in applied work, involving small samples.

We explore changing the data generating process F by varying its parameters, namely the number of response categories k, the median response category m[6], and the underlying level of inequality t. In varying these parameters of the DGP, we are particularly interested in investigating the effect of *severe class imbalances*. The interest underlying DGPs exhibiting severe class imbalances, is that they allow the researcher to explore the effect of sampling near the boundary of the parameter space of the underlying multinomial population, where we expect the normal approximation of the distribution of the asymptotic test to be less accurate.

Varying t_o in the generic test (4) allows us (i) to investigate test size (by setting $t_o = t$, where t is the level of inequality associated with the DGP F), and (ii) to investigate statistical power (by setting t_o to be different from t). We also explore variations in the null hypothesis (4) by varying the inequality aversion parameters ω_1 and ω_2 of the underlying inequality index.

To provide a unified view of the scope of the envisaged simulation exercises, it is useful to consider a general Monte Carlo procedure as a method for studying the size and power properties a given test τ (in this paper we consider $\tau = \tau^A, \tau^B$) given a prescribed null hypothesis H_o, a data generating process F, and a sample size n. A Monte-Carlo function $\mathcal{M}(\tau, H_o, F, n)$ is then an algorithmic procedure used to estimate, via simulation, the distribution of the resulting test statistic in a variety of contexts, such as those discussed above.

The strategy we pursue in the investigations is to define a baseline specification, and to explore variations from this benchmark case. The baseline case arises in relation to a sample size n of 499 observations, a DGP associated with $k = 5$ socioeconomic classes, a median socioeconomic status $m = 3$ and a uniform probability mass function $f^{base} = (1\,1\,1\,1\,1)/5$ (see Table 1).

Table 1. Baseline specification.

Parameter	Notation	Baseline Value
Sample size	n	499
Number of response categories	k	5
Probability mass function	$f = (f_1, ..., f_k)$	$f^{base}(1/5, ..., 1/5)$
Median response category	m	3
Inequality aversion parameters	(ω_1, ω_2)	$(2, 2)$
Level of inequality (size of test)	t_o	0.60
Level of inequality (power of test)		$t_o \pm 0.02$

[6] Observe from (2) and (3) that the functional form of the inequality index changes with the median socio-economic status m. In this sense, one could equally interpret the purpose of introducing variations in the parameter m as an exercise of exploring changes in the functional form of the inequality index in the null hypothesis being investigated.

The inequality aversion parameters are set at the value $(\omega_1, \omega_2) = (2, 2)$ and the resulting level of inequality t_o is equal to 0.60. The number B of bootstrap samples is set throughout equal to 999, while the number of Monte Carlo samples is set to the value 5000. The appendix provides further detail about the DGPs used in the Monte Carlo investigations.

Our chosen method of summarizing the simulation results will be to rely on graphical methods developed by Davidson and MacKinnon (1998). Firstly, we shall report p-*value curves* for both tests. The advantage of these graphical devices lies in allowing the researcher to investigate globally the size of tests, not just at key nominal values (such as 1% or 5%). We can furthermore identify a correctly sized test when its p-value curve lies below the 45 degree line.To investigate power, we shall report *size–power curves*. The advantage of the size–power curve is to quantify statistical power at a correct (i.e., consistently estimate) size, rather than nominal size. A test procedure τ^i is more powerful than τ^j in the context of a particular hypothesis of interest H_o, when the related size–power curve of τ^i lies above the size–power curve pertaining to τ^j.

5. Simulation Results

The focus in this section is on exploring the size and power properties of the asymptotic and bootstrap tests. With the exception of extreme cases of very low and very high inequality, we find that both tests are correctly sized in our investigations, and moreover, both tests have similar power even in the presence of small samples.

5.1. Sample Size

In the baseline specification of interest, we set the sample size at $n = 499$, and the other parameters of the DGP are chosen as indicated in Table 1 and the Appendix A of the paper. In the top panel of Figure 1 we plot p-value curves pertaining to both the asymptotic and bootstrap test, while the bottom panel plots size–power curves. The simulations pertain to sample sizes $n = 49$, $n = 99$, $n = 499$ (the baseline) and $n = 999$. We report on the horizontal axis of the p-value curves nominal sizes from 0% to 25%. For the size–power curves, we report power over the same range of *actual* sizes. For sample sizes of 49, 99 and 499 observations the p-value curves lie below diagonal (45 degree) line, indicating that the tests are correctly sized. At nominal sizes exceeding 15%, we note nonetheless that the p-value plots cross the diagonal axis in the context larger samples ($n = 999$). Our central concern being the comparison of the relative performance of the two tests, we note that they perform equally well in terms of their p-value curves.

Turning to the bottom panel where we report size–power curves, we observe that both testing methods also perform broadly similarly in terms of power, and the overall pattern is that power rises with sample size, as is of course expected to be the case.

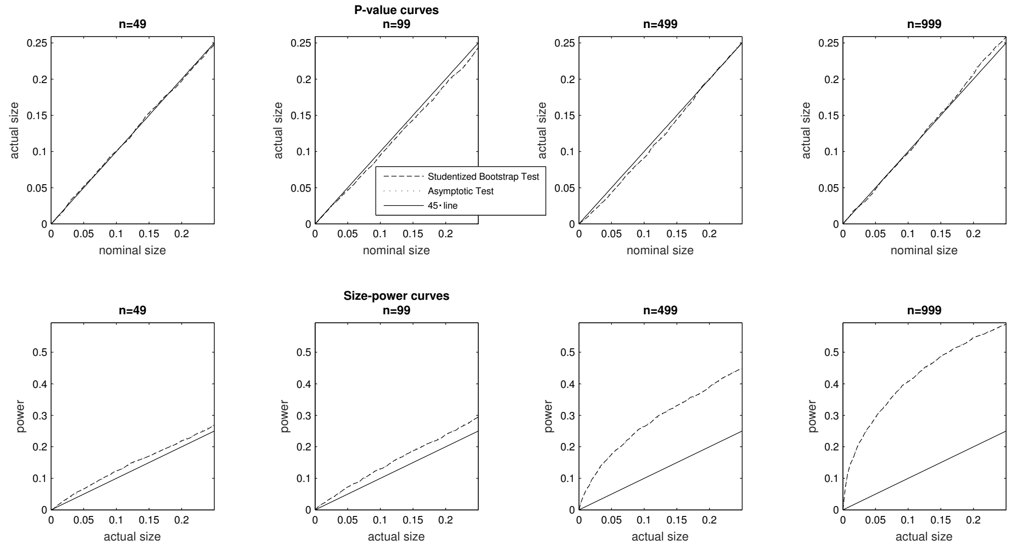

Figure 1. The effect of sample size.

5.2. The Number of Response Categories

We are interested in exploring how the choice of DGP influences the relative performance of the asymptotic and Studentized bootstrap test. Our first investigation in this respect is to explore the effect of changing the number of response categories k. The baseline specification chooses a uniform PMF defined on $k = 5$ socioeconomic categories, and we also examine DGPs pertaining to $k = 3$, and $k = 9$. For all three values of k, the evidence plotted in the top panel of Figure 2 supports the conclusion that both tests are correctly sized.

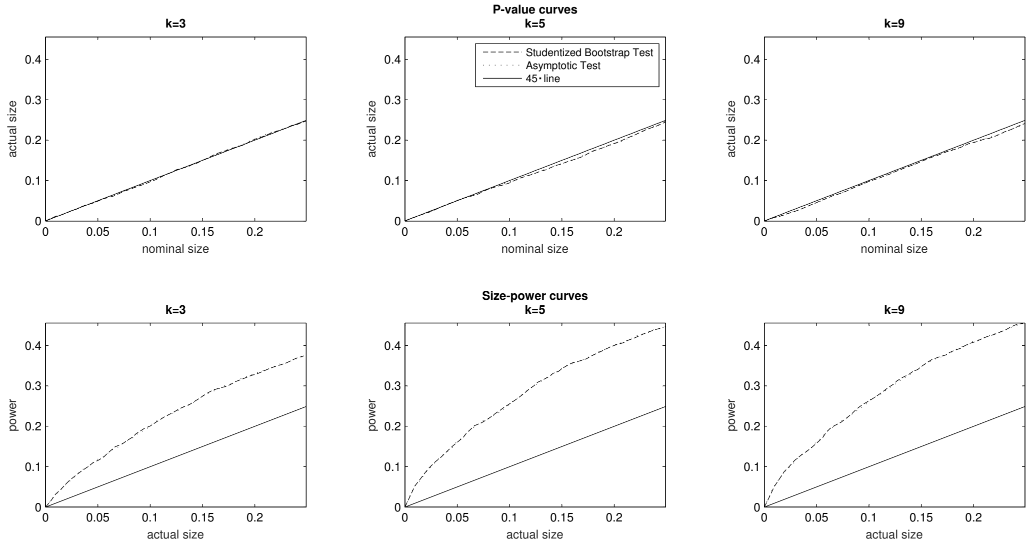

Figure 2. The number of response categories.

For a given sample size, it is not immediate to infer the effect of changing the number of response categories on the power of the tests. In the context of both tests, the findings of the bottom panel of Figure 2 suggest, however, that other things equal, increasing k from $k = 3$ to the baseline value $k = 5$ results in an increase of statistical power. The gain in power from increasing the number of response categories further to $k = 9$ is, however, less apparent in the context of both tests. In further investigations not reported in the paper, we explored the effect of further increasing k (for a range of values from $k = 2$ to $k = 49$), for different sample sizes. We did not find however any general monotonic relation emerging between the number of response categories and statistical power in either of the asymptotic or Studentized bootstrap test.

5.3. The Median Response Category

By varying the median response category m of the distribution we can explore some parametric changes in the shape of the underlying DGP. We consider first a PMF $f = (5710\ 1073\ 1073\ 1073\ 1073)/10,000$ with an associated median state $m = 1$. We then somewhat reduce the mass at the bottom tail of the distribution (i.e., at the bottom socioeconomic status), and consider a PMF $f = (2879\ 2879\ 1414\ 1414\ 1414)/10,000$ with median state $m = 2$. Finally, we use the baseline DGP to explore the effect of setting the median state at the value $m = 3$. Note also that the three PMFs exhibit a level of inequality equal to 0.60 (i.e., the baseline level), despite the different shapes of these probability mass functions and their varying median states.

For all three values of the median state m, the evidence plotted in the top panel of Figure 3 supports the conclusion that both tests are correctly sized. It is also the case that their p-value curves broadly overlap. Turning to the size–power curves, the findings of the bottom panel of Figure 3 suggest that upon varying the median response category, both tests perform equally well in terms of statistical power.

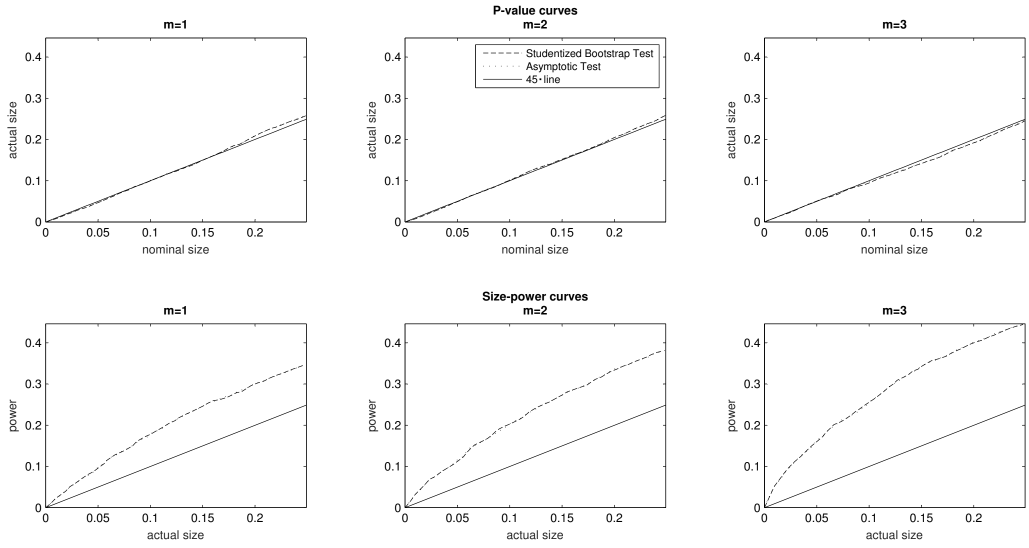

Figure 3. The median response category.

5.4. Inequality Aversion Parameters

We expect the asymptotic test to perform well in the context of a linear inequality index (i.e., when both inequality aversion parameters are set equal to one). When either inequality parameters are greater than unity, simulation methods in practice can provide more accurate estimates of the distribution of test statistics than analytical methods that rely on the *delta* method (Davison and Hinkley 1997, p. 16). For this reason, our interest here will be to investigate the practical advantage of adopting the Studentized bootstrap test, when the inequality aversion parameters are allowed to vary.

Specifically, in Figure 4 we begin by setting the inequality aversion parameters to $(\omega_1, \omega_2) = (1, 1)$. The resulting inequality index being linear, is a member of the Kobus–Miłoś family (2) of subgroup decomposable indices. We then examine the case pertaining to the baseline parameter values $(\omega_1, \omega_2) = (2, 2)$, followed by the case $(\omega_1, \omega_2) = (4, 4)$. For all three investigations, the evidence plotted in the top panel of Figure 4 supports the conclusion that both tests are correctly sized, and the p-value curves specific to each investigation broadly overlap. Turning to the size–power curves, the findings of the bottom panel of Figure 4 suggest that the two tests exhibit broadly similar power for each pair of parameter values, and furthermore that the power of each given test is sensitive to the specification of the inequality index.

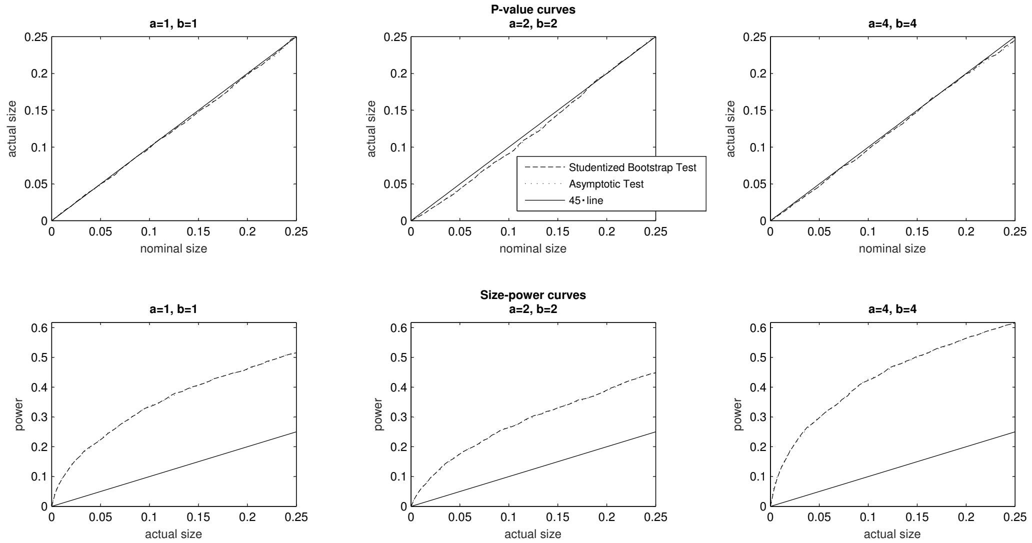

Figure 4. The effect of inequality aversion parameters.

In the investigations of Figure 5, we explore the effect of attaching different weight to the bottom and top tails of the distribution via setting ω_1 and ω_2 at different parameter values: a larger value of ω_1 renders the inequality index more sensitive to the bottom tail of the distribution, while increasing ω_2 makes the index more sensitive to the top tail of the distribution. We depict the p-value and size–power curves of the two tests in relation to the following sequence of inequality aversion parameters: $(\omega_1, \omega_2) = (1, 4)$, $(\omega_1, \omega_2) = (4, 1)$ and $(\omega_1, \omega_2) = (8, 1)$. For all three investigations, the evidence in the top panel of Figure 5 supports the conclusion that both tests are moderately over-sized, while the p-value curves of the two tests in each specific investigation broadly overlap. The bottom panel reveals that the power of the tests is sensitive to the specification of the inequality index. There is little to differentiate the asymptotic and Studentized bootstrap test in term of size–power curves, and, as in the findings of Figure 4, we are led to conclude that the power of each test is sensitive to the choice of inequality aversion parameters.

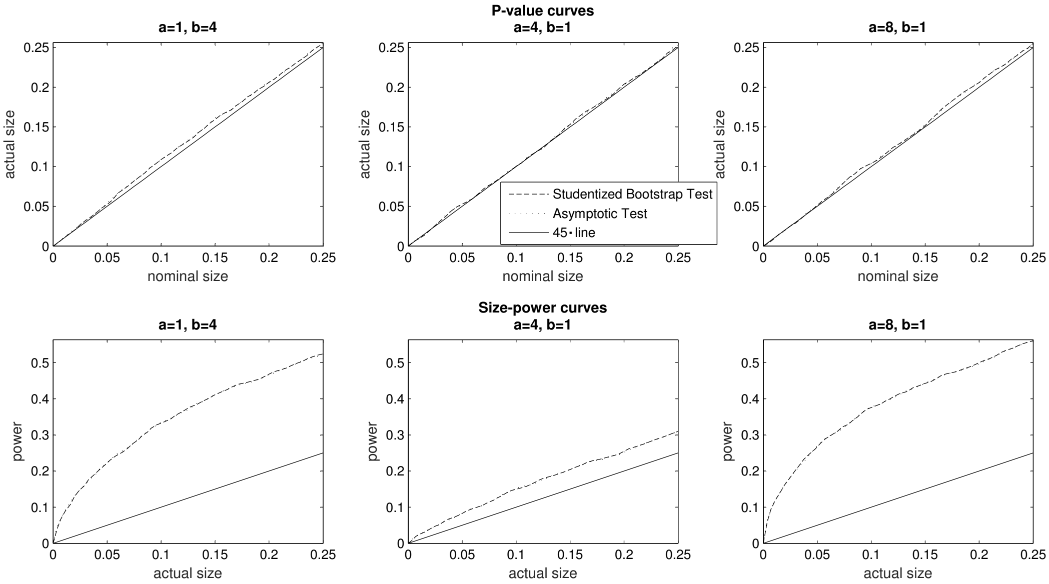

Figure 5. The effect of inequality aversion parameters (further results).

5.5. Severe Class Imbalances

To investigate the effect of severe class imbalances, our strategy here is to generate a sequence of distributions that are ordered by the $\preceq_{AF}$ relation. That is, we consider a sequence of PMFs f^1, f^2, f^3, f^4, f^5; where for $i = 1, 2, 3, 4$, f^{i+1} is obtained from f^i using a number of median preserving spreads. In increasing order, the level of inequality t associated with each of the distributions in the sequence equals 0.002, 0.3827, 0.444, 0.60, and 0.999. We note here however that associated with the two polar cases of extreme equality and inequality are two PMFs that exhibit *severe class imbalances*: $f^1 := (1\ 1\ 1000\ 1\ 1)/1004$, and $f^5 = (1000\ 1\ 1\ 1\ 1000)/2003$. On the other hand, f^4 coincides with the PMF of the baseline model, $f^4 := (1\ 1\ 1\ 1\ 1)/5$, which exhibits a property of *extremely well balanced classes. There are two other PMFs in the sequence,* $f^2 = (0\ 0\ 6\ 2\ 1)/9$ and $f^3 = (1\ 2\ 3\ 2\ 1)/9$. Both of f^2 and f^3 share a common median ($m = 3$), f^2 however exhibits a less balanced class structure than f^3, and furthermore lies at the boundary of the parameter space defining multinomial distributions.

We also consider a second sequence of PMFs ordered by the median preserving spreads relation, given by f^1, f^2, g^3, f^4, f^5, where g^3 replaces f^3, and such that the level of inequality at g^3 equals 0.5278. This new PMF, $g^3 = (0\ 0\ 3\ 2\ 1)/6$ is chosen to lie at the boundary between the subsets Λ_3 and Λ_4 pertaining to distributions with median states equal to 3 and 4 respectively[7].

The top panel of Figure 6 provides p-value plots for the tests associated with the (from left to right) the DGPs of f^1, f^2 and f^3. The PMF f^1 chosen to exhibit severe class imbalances is of particular interest, as it generates marked differences in the p-value curves associated with the asymptotic test and the Studentized bootstrap test. At low significance levels 0% to 5%, the asymptotic test is moderately oversized. We may however note that at higher significance levels (6% to 25%) the asymptotic test is correctly sized, rather than being oversized. The bootstrap test is oversized at all significance levels, and its size substantially exceeds that of the asymptotic test above the 4% significance level. It is equally important to observe that any differences between the p-value curves of the two tests become

[7] We note however that the new PMF g^3 is not comparable with f^3 in terms of median preserving spreads. The central feature defining the two sequences is therefore that within a sequence inequality rises according to any inequality index that is increasing in median preserving spreads.

hardly visible in relation to the DGPs associated with the PMFs f^2 and f^3. The bottom panel of Figure 6 reports the size–power curves of the two tests in relation to the DGPs associated with f^1, f^2 and f^3. In the case of severe class imbalances (the DGP associated with f^1) the two tests exhibit very similar power, and the conclusion is similar when inequality rises in the case of f^2 and f^3.

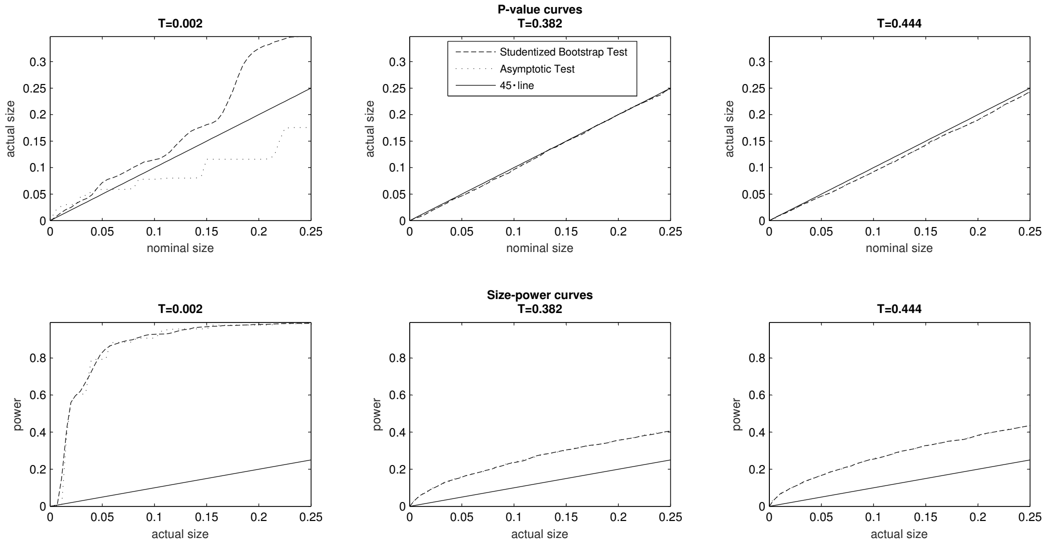

Figure 6. The effect of severe class imbalances.

Figure 7 similarly plots p-value and size–power curves for the DGPs associated with the remaining PMFs g^3, f^4 and f^5. In terms of test size and power, the plots of the top and bottom panel reveal that both the asymptotic and bootstrap test perform very similarly in relation to g^3 and f^4. Recalling that f^4 is the uniform PMF of the baseline model, we do not find in this investigation that severe class balances lead to oversized tests, comparatively low power, or overall differences in the p-value or size–power curves of the two tests.

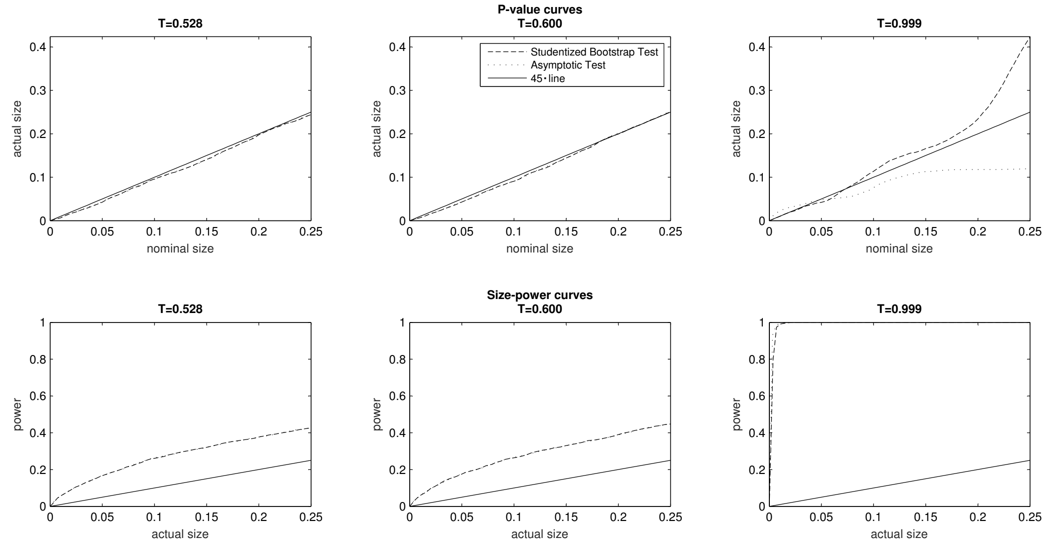

Figure 7. The effect of severe class imbalances (further results).

As discussed above, the probability mass function f^5 exhibits the polar case of high inequality ($t_0 = 0.999$), and is associated with severe class imbalances. The findings in relation to the DGP associated with f^5 are qualitatively similar to those of the DGP associated with f^1 which exhibits the opposite polar case of low inequality. Here we find that at low significance levels (0% to 4%), the asymptotic test is moderately oversized. We may however note that at higher significance levels (5% to 25%) the asymptotic test is correctly sized. The bootstrap test remains correctly sized at significance levels 0% to 6% but then becomes (increasingly) oversized at all levels above the 7% significance level. Nonetheless, we do not find any substantial differences in the size–power curves of the two tests in the bottom panel of Figure 7. What we can document, however, is that statistical power is considerably higher in the extreme cases of severe class imbalances than in the intermediate cases of the two sequences of PMFs, using either of the two tests.

6. An Illustrative Example

As an illustrative example, we return to the data application in Abul Naga and Stapenhurst (2015) pertaining to asymptotic inference in relation to the Alphabeta family of inequality indices (3). These are data on five ordered nutritional health states from the Egyptian Integrated Household Survey of 1997–1999.[8] The data refer to two statistical areas of Northern Egypt (also known as *Lower* Egypt), namely Metropolitan Lower Egypt (METLO) and Non-Metropolitan Lower Egypt (NMETLO). The resulting cumulative distributions are respectively $X_1 = (0.075, 0.187, 0.430, 0.812, 1.00)$ for the METLO data $(n_1 = 107)$ and $X_2 = (0.040, 0.144, 0.363, 0.667, 1.00)$ for the NMETLO data $(n_2 = 452)$. Note also that the median response category is $m = 4$ in both distributions. The data from the larger sample exhibit somewhat more class imbalances than the first sample, though neither sample exhibits the levels of class imbalances found to be problematic in the simulations of Section 5.

We begin by carrying out the same exercise as in Section 5, for the empirical distribution functions of the data. Using the baseline values $(\omega_1, \omega_2) = (2, 2)$ of the inequality aversion parameters, inequality is calculated at 0.376 and 0.474 in respectively the Metropolitan (METLO) and Non-Metropolitan (NMETLO) samples. We next investigate a hypothesis that inequality in a given sample—computed at parameter values $(\omega_1, \omega_2) = (2, 2)$—is equal to 0.456 (the level of inequality of the pooled sample, computed at the same parameter values.) On the basis of the simulations reported in the results section, we expect the asymptotic and bootstrap tests to exhibit very similar p-value curves as well as size–power curves. Inspecting the related curves in Figure 8, we find that this pattern indeed does emerge. It is worth noting however that the power of the test is considerably higher in the smaller METLO sample. This pattern is to be expected, as the hypothesized value of $T = 0.456$ is considerably closer to the level of inequality in the larger Non-Metropolitan sample $(T_2 = 0.474)$ than it is in the Metropolitan Lower Egypt $(T_1 = 0.376)$.

We report these inequality computations pertaining to parameter values $(\omega_1, \omega_2) = (2, 2)$ in Table 2, together with other computations of inequality in the two samples for various pairs of parameter values. Rows 4 and 5 of the Table furthermore report p-values arising from the bootstrap test of the hypothesis that each of the respective samples has the same level of inequality T of the combined sample. In the context of the computations pertaining to parameter values $(\omega_1, \omega_2) = (1, 1)$, both samples exhibit identical levels of inequality, a 0.440 figure, and hence the p-values of both tests are equal to 1. We observe that it is also the case that for the other pairs of parameter values, the bootstrap test fails to reject at the 5% level the hypothesis that either of the two populations has a level of inequality equal to that of the combined sample. Nonetheless, we note that in the context of the computations pertaining to parameter values $(\omega_1, \omega_2) = (2, 2)$, the test does reject at the 10% the hypothesis that METLO has the same level of inequality $T = 0.456$ as the combined sample (last column of the table).

[8] 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.

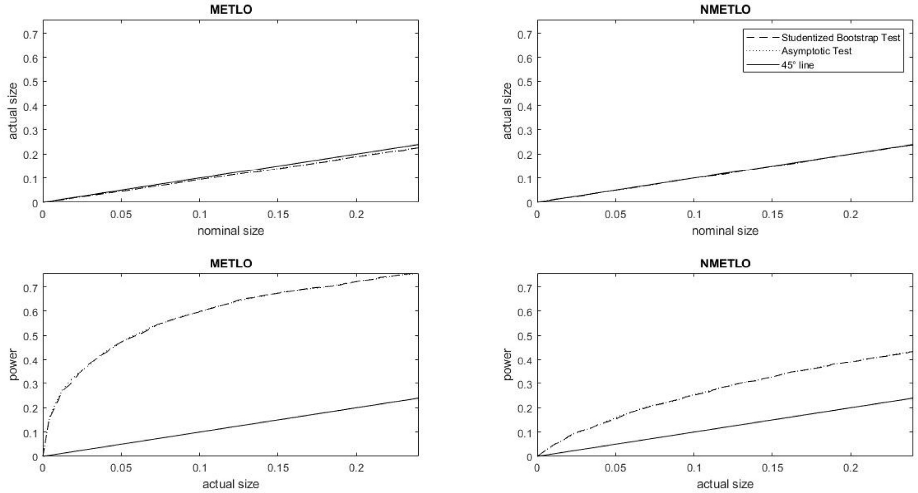

Figure 8. Inequality in nutritional health in Lower Egypt.

Table 2. Bootstrap inference for Inequality in Nutritional Health in Lower Egypt.

(ω_1, ω_2)	$(1,1)$	$(1,2)$	$(2,1)$	$(2,2)$
T_1	0.440	0.458	0.330	0.376
T_2	0.440	0.490	0.390	0.474
T	0.440	0.485	0.378	0.456
$p(T_1 = T)$	1	0.52	0.28	0.06
$p(T_2 = T)$	1	0.78	0.46	0.34
$T_1 BCI$	(0.353;0.545)	(0.373;0.553)	(0.243;0.437)	(0.287;0.470)
$T_1 ACI$	(0.357;0.521)	(0.380;0.536)	(0.250;0.410)	(0.294;0.458)
$T_2 BCI$	(0.402;0.481)	(0.453;0.527)	(0.352;0.426)	(0.432;0.507)
$T_2 ACI$	(0.405;0.475)	(0.459;0.521)	(0.357;0.423)	(0.440;0.507)

Notes: (1) T_1 denotes inequality in Metropolitan Lower Egypt ($n_1 = 107$);T_2 denotes inequality in Non-Metropolitan Lower Egypt ($n_2 = 452$); and T denotes inequality in the pooled sample. (2) $p(T_i = T)$ denotes the p-value associated with the null hypothesis $H_o : T_i = T$. (3) ACI and BCI are respectively asymptotic and bootstrap 95% confidence intervals.

In order to highlight one practical difference between the two tests, we calculate in rows 6 to 9 of Table 2 the 95% confidence intervals for inequality in the two samples, using both bootstrap inference (the rows starting with T_i BCI) and asymptotic inference (the rows starting with the cell T_i ACI)[9]. These intervals consist of all the levels of inequality that we would fail to reject at the 5% level. While the asymptotic inference confidence intervals are by construction symmetric about the sample value T_i of inequality, this symmetry need not arise in the context of the bootstrap confidence intervals. We find here that the bootstrap confidence intervals are generally larger than the asymptotic confidence intervals. These findings, obtained in the context of sample sizes of 100 to 500 observations,

[9] In the context of the bootstrap test, the confidence intervals are obtained by finding hypothesized values t_o which produce p-values equal to 0.025 and 0.975.

would suggest that the exact distribution of the alphabeta inequality index may exhibit thicker tails than those of the standard normal distribution.

7. Conclusions

We have used Monte Carlo experiments to compare the empirical size and statistical power of asymptotic inference and the Studentized bootstrap test. We have found that in all our investigations the asymptotic and bootstrap test exhibit very similar size–power curves. With the exception of extreme cases of very low and very high inequality, we also have found that both tests are correctly sized in our investigations. The experiments pertaining to extremely low and high levels of inequality (respective values of 0.002 and 0.999) present cases of several class imbalances in the underlying data generating process. In these two investigations, the asymptotic test remains correctly sized at all test sizes ranging between 0% and 25%, while the bootstrap test becomes increasingly oversized at all levels starting from the 6% value. Nonetheless, both tests remain correctly sized under other cases of severe class imbalances or other cases where the DGP lies at the boundary of the parameter space pertaining to multinomial distributions. Awaiting further investigations of this nature, given the numerical cost associated with implementing the bootstrap test, our broad recommendation to applied researchers would be to adopt asymptotic inference in the context of inequality indices defined on ordered response data. That is, the context of ordered response data would appear to be of a separate nature from income data, where asymptotic inference has been documented to often produce incorrectly sized tests.

These conclusions have been reached in the context of sampling where the population median, used in order to determine the functional form of the inequality index, was assumed to be known. It is therefore important in further investigations to develop a framework for exploring the performance of the two tests in sampling contexts where the median of the distribution is treated as a random quantity[10].

Author Contributions: Conceptualization, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; methodology, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; software, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; formal analysis, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; writing–original draft preparation, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; writing–review and editing,Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky; visualization, Ramses Abul Naga, Christopher Stapenhurst and Gaston Yalonetzky. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We are grateful to three anonymous referees, Karim Abadir, Martin Biewen and Emmanuel Flachaire for helpful comments and suggestions. The paper was written in part while Abul Naga was a visitor at Aix-Marseille University. The author wishes to thank the Iméra Institute of Advanced Studies and the Aix-Marseille School of Economics for their hospitality.

Conflicts of Interest: ‘The authors declare no conflict of interest.

Appendix A

In several Monte Carlo investigations, the baseline DGP is used, with an associated PMF $f^{base} = (1\ 1\ 1\ 1\ 1)/5$, (with a median response category $m = 3$). In other investigations, the adoption of alternative DGPs is required, for which we provide here further information.

(i) In the context of investigations pertaining to the effect of changes in *sample size*, the baseline DGP is used throughout the investigations. The sample sizes examined considered are $n = 49$, $n = 99$, $n = 499$, and finally $n = 999$.

(ii) In the context of investigations of changes in the *number of response categories*, the DGP in the context of $k = 3$ has a PMF given by $f = (3\ 4\ 3)/10$ (with a median response category $m = 2$). For the

[10] See Stapenhurst (2017) for further discussion.

case $k = 5$, the PMF of the baseline DGP is adopted. For the final case we examine, $k = 9$ and the PMF given is by $f = (1601\ 1103\ 811\ 665\ 1640\ 665\ 811\ 1103\ 1601)/10,000$ (with a median response category $m = 5$).

(iii) In the context of investigations of changes in the *median category*, the DGP in the context of $m = 1$ has a PMF given by $f = (5710\ 1073\ 1073\ 1073\ 1073)/10,000$. For the case $m = 2$, the PMF given is by $f = (2879\ 2879\ 1414\ 1414\ 1414)/10,000$. For the final case we examine, $m = 3$ the baseline DGP is chosen). We recall that all three DGPs share an identical level of inequality of 0.60.

(iv) In the context of investigations of changes in the *inequality aversion parameters*, the baseline DGP is used throughout the investigations. The inequality aversion parameters examined are given by the following sequence of ordered pairs $(\omega_1, \omega_2) : (1, 1), (2, 2), (4, 4), (1, 4), (4, 1), (8, 1)$.

(v) In the context of investigations of changes in the *level of inequality*, the DGP in the context of the investigations pertaining to the level of inequality $t_0 = 0.002$ has a PMF given by $f^1 = (1\ 1\ 1000\ 1\ 1)/1004$. The DGP in the context of $t_0 = 0.3827$ has a PMF given by $f^2 = (0\ 0\ 6\ 2\ 1)/9$. The DGP associated with the level of inequality $t_0 = 0.4444$ has a PMF given by $f^3 = (1\ 2\ 3\ 2\ 1)/9$. The DGP associated with the level of inequality $t_0 = 0.5278$ has a PMF $g^3 = (0\ 0\ 3\ 2\ 1)/6$. For the investigations pertaining to the level of inequality $t_0 = 0.60$, we adopt the baseline DGP; that is we set $f^4 := f^{base}$. The DGP associated with the level of inequality $t_0 = 0.999$ has a PMF given by $f^5 = (1000\ 1\ 1\ 1\ 1000)/2003$.

References

Abul Naga, Ramses, and Christopher Stapenhurst. 2015. Estimation of inequality indices of the cumulative distribution function. *Economic Letters* 130: 109–12.

Abul Naga, Ramses, and Tarik Yalcin. 2008. Inequality measurement for ordered response health data. *Journal of Health Economics* 27: 1614–25.

Arrighi, Y., M. Abu-Zaineh, and B. Ventelou. 2015. To count or not to count deaths: Reranking effects in health distribution evaluation. *Health Economics* 24: 193–205.

Athreya, K. 1987. Bootstrap of the mean in the infinite variance case. *The Annals of Statistics* 15: 724–31.

Barret, G., S. Donald, and D. Bhattacharya. 2014. Consistent nonparametrictests for lorenz dominance. *Journal of Business and Economics Statistics* 32: 1–13

Cowell, Frank, and Emmanuel Flachaire. 2015. Statistical methods for distributional analysis. *Handbook of Income Distribution*. Edited by A. Atkinson, and F. Bourguignon. Amsterdam: Elsevier, pp. 359–465. .

Davidson, Russell. 2007. Bootstrapping econometric models. *Quantile* 3: 13–36.

Davidson, Russell, and Jean-Yves Duclos. 2013. Testing for restricted stochastic dominance. *Econometric Reviews* 1: 84–125.

Davidson, R., and J. MacKinnon. 1998. Graphical methods for investigating the size and power of hypothesis tests. *Manchester School* 66: 1–26.

Davison, A.C., and D. Hinkley. 1997. *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Dutta, Indranil, and James Foster. 2013. Inequality of happiness in the united states 1972–2010. *Review of Income and Wealth* 59: 393–415.

Horowitz, J. 2001. The bootstrap. *Handbook of Econometrics*. Edited by J. Heckman, and E. Leamer. Amsterdam: Elsevier.

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: 616–25.

Kobus, Martyna, and Piotr Milos. 2012. Inequality decomposition by population subgroups for ordinal data. *Journal of Health Economics* 31: 15–21.

Madden, D. 2010. Ordinal and cardinal measures of health inequality: An empirical comparison. *Health Economics* 19: 243–250.

Russell, Davidson, and Emmanuel Flachaire. 2007. Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics* 141: 141–66.

Stapenhurst, Christopher. 2017. Testing for median preserving spreads of the multinomial distribution. Master's Thesis, Edinburgh: University of Edinburgh.

Yalonetzky, Gaston. 2013. Stochastic dominance with ordinal variables: Conditions and a test. *Econometric Reviews* 32: 126–63.