THE
Teoría e Historia Económica
Working Paper Series

Inferring Inequality: Testing for Median-Preserving Spreads in Ordinal Data

Ramses H. Abul Naga, Christopher Stapenhurstz and Gaston Yalonetzky

WP 2021-01
November 2021

Departamento de Teoría e Historia Económica
Facultad de Ciencias Económicas y Empresariales
Universidad de Málaga
ISSN 1989-6908

Inferring Inequality: Testing for Median-Preserving Spreads in Ordinal Data*

Ramses Abul Naga†, Christopher Stapenhurst‡ and Gaston Yalonetzky§

October 29, 2021

Abstract

The median-preserving spread (MPS) ordering for ordinal variables (Allison and Foster, 2004) has become ubiquitous in the inequality literature. However, the literature lacks an explicit frequentist method for inferring whether an ordered multinomial distribution $\boldsymbol{G}$ is more unequal than $\boldsymbol{F}$ according to the MPS criterion. We devise formal statistical tests of the hypothesis that $\boldsymbol{G}$ is *not* an MPS of $\boldsymbol{F}$. Rejection of this hypothesis enables the conclusion that $\boldsymbol{G}$ is robustly more unequal than $\boldsymbol{F}$. Using Monte Carlo simulations and novel graphical techniques, we find that the choice between Z and Likelihood Ratio test statistics does not have a large impact on the properties of the tests, but that the method of inference does: bootstrap inference has generally better size and power properties than asymptotic inference. We illustrate the usefulness of our tests with three applications: (i) happiness inequality in the United States, (ii) self-assessed health in Europe and (iii) sanitation ladders in Pakistan.

Keywords: inequality measurement; hypothesis testing; median preserving spread; ordinal data

*Acknowledgments: We would like to thank seminar participants at Heriot-Watt and conference participants at the 9th ECINEQ Meeting for helpful comments and suggestions.

†University of Malaga, University of Aberdeen, Pan African Scientific Research Council.

‡University of Edinburgh.

§University of Leeds.

1 Introduction

Despite the prominence of ordinal data in the social sciences (e.g. subjective wellbeing, self-assessed health, dwelling conditions, etc.), inequality assessments of ordered multinomial distributions pose significant measurement challenges. Chiefly, the differences between the ordered categories are not commensurable (Stevens, 1946); therefore any numerical scales attributed to them will be arbitrary. Then, the robustness of these inequality comparisons to alternative scales is not guaranteed (Mendelson, 1987; Allison and Foster, 2004). In order to solve this fundamental problem, suitable measurement tools have been proposed (see Silber and Yalonetzky, 2021, for a recent review).

One such tool is a set of partial orderings known as the quantile-preserving spreads (Mendelson, 1987), of which the median-preserving spread is by far the most popular in the literature, especially the rendering by Allison and Foster (2004) and many key works thereafter. Median-preserving spreads rank ordered multinomial distributions sharing a common median category in terms of the size of their tails. Distributions with 'thicker' tails, i.e. with a higher proportion of the population further away from the common median, are deemed more unequal. In that sense, the median-preserving spread (henceforth MPS) partial ordering is an extension of the mean-preserving spread relation proposed by Rothschild and Stiglitz (1970) into the world of ordinal variables. Naturally, the orderings are partial because not all distributions with a common median have unambiguously thicker or thinner tails across the whole domain of categories.

Besides being intrinsically interesting, the MPS partial ordering provides a useful criterion to test the robustness of inequality comparisons with ordinal data to alternative choices of inequality indices characterised by an 'aversion to MPS'. Introduced by Apouey (2007) and Abul Naga and Yalcin (2008), this property requires any inequality index to rank distribution $\boldsymbol{g}$ as more unequal than $\boldsymbol{f}$ whenever the former is obtained from the latter through a (sequence of) MPS transferring probability mass toward the tails and away from the pre-

served median category. Kobus (2015) showed that all inequality indices averse to MPS transformations agree in ranking any pair of distributions belonging in the MPS partial ordering.

Ever since Allison and Foster (2004) popularised the notion of MPS, empirical illustrations have followed. For instance, based on data from the American National Health Interview Survey, Allison and Foster (2004) conclude that self-reported health in Texas is an MPS of Pennsylvania's, implying that health levels in Pennsylvania are less unequal than in Texas. Dutta and Foster (2013) find that happiness inequality in the US fell over the 1970s and 80s, but has increased again since the 90s. Madden (2010) and Balestra and Ruiz (2015) have performed similar assessments in the realms of self-assessed health, education levels and subjective wellbeing. Such conclusions are potentially of interest to both researchers and policymakers.

However, these empirical studies do not rule out the possibility that the observed orderings are a result of random sampling. Our contribution is to devise statistical tests that can allow us to conclude that the populations underlying the samples are also ordered by MPS.

Our first goal is to derive an appropriate null hypothesis. Davidson and Duclos (2013) derive tests for the partial ordering of first-order stochastic dominance (FOSD) involving cardinal variables, such as income or consumption expenditure. They choose to specify a null hypothesis that the pair of distributions is *not* ordered in a specific direction, because rejection of this hypothesis leads to the conclusion that the two population distributions *are* ordered, which is normally the outcome of most interest. For the same reason, in section 2 we posit the hypothesis that a given pair of distributions is *not ordered* by MPS. Thus a pair of distributions, $(\boldsymbol{F}, \boldsymbol{G})$, lies in our null space if it has least one of three properties: either $\boldsymbol{F}$ and $\boldsymbol{G}$ do not share the same median; or the potentially more equal distribution, $\boldsymbol{F}$, does not FOSD the potentially more unequal population, $\boldsymbol{G}$, below the median; or $\boldsymbol{G}$ does not FOSD $\boldsymbol{F}$ above the median. We develop a novel graphical depiction of the null and alternative spaces, which is useful for visually checking whether a pair of samples is ordered.

The FOSD conditions stem from the definition of the MPS partial ordering (Allison and Foster, 2004) and are analogous to those characterising the null space of Davidson and Duclos (2013). The median condition is new and gives our null and alternative spaces rather unique characteristics. Specifically, the boundary of the null space will be shown to be the union of two sets, that we shall refer to as the 'dominance' and 'median' boundaries. In turn this characteristic of the boundary, arising as the union of two sets, has further implications for how we compute the closest null distribution. This is important for constructing the Quasi Maximum Likelihood Estimator (QMLE) (section 3). Furthermore the derivation of the closest null distribution is used to investigate the power of the tests (section 4).

In section 3 we define a likelihood ratio (LR) statistic and a standardised (Z) statistic. We characterise their respective asymptotic distributions under the null, yielding two tests: the asymptotic LR test and the asymptotic Z test. We then construct two corresponding bootstrap tests, the bootstrap LR test and the bootstrap Z test, and state a result on their inference properties. Thus the paper provides a family of four tests for empirical applications.

In section 4 we use Monte Carlo simulations to compare the tests' size and power properties. Regarding size, our main finding is that our tests are mostly correctly sized, even when one or both samples comprise as few as 10 observations. We also find that the empirical size of bootstrap tests is often closer to the nominal size than asymptotic tests', and that the choice of tests statistic is usually inconsequential. Notable exceptions to these results arise when asymptotic inference is used in circumstances where the size of the sample drawn from the more equal distribution $\boldsymbol{F}$ is an order of magnitude smaller than that drawn from the more unequal distribution $\boldsymbol{G}$, and the distributions have a pair of similar cumulants (i.e. $F_i \approx G_i$ for some i). Regarding power, our main result is that the tests have broadly the same *actual* power (i.e. the rate of rejection of a false hypothesis relative to the rate of rejection of the closest true hypothesis is very similar across all tests), but the exceptions are slightly different than for size. Specifically, we find that bootstrap inference is more powerful against almost all alternatives when the size of the sample drawn from $\boldsymbol{G}$ is very

small (close to 10), but that asymptotic inference can be more powerful when the size of the sample drawn from $\boldsymbol{F}$ is very small and the median of $\boldsymbol{G}$ is close to the edge of the median category (i.e. $G_m \approx \frac{1}{2}$ where m is the median state of $\boldsymbol{G}$). All the tests have low power against certain sets of circumstances: firstly, when both true cumulative distributions take values close to 0.5 at the median category ($F_i \approx G_i \approx \frac{1}{2}$ for some i); and secondly when the size of the sample drawn from $\boldsymbol{F}$ is small in absolute terms. On the basis of this result, we conclude that asymptotic inference is generally adequate when sample sizes drawn from the two distributions are relatively well matched, but bootstrap inference should be considered when sample sizes differ by an order of magnitude or more.

The graphical tools deployed to illustrate the size and power properties of the tests are themselves novel, and, we hope, useful to researchers. A standard size curve plots the actual (empirical) rejection rate of a test against its nominal size for a given null distribution. However the intricacies of our null space are difficult to capture with a small selection of null distributions. Instead, we hold the nominal size constant (at levels 1, 5 and 10%) and plot the empirical size against a continuum of distributions in the boundary of the null space, thereby obtaining an upper bound on the actual size of the tests. Similarly, a standard size-power curve[1] plots the rejection rate for a distribution in the alternative space against the rejection for the closest corresponding distribution in the null space. Our power-locus curve plots the rejection rate for a continuum of distributions in the alternative space, which we call the 'interior locus'. We argue that the power properties against this set of alternatives is indicative of properties against all other distributions in the alternative space.

In section 5 we demonstrate the broad usefulness of our tests in three diverse areas of applications covering subjective wellbeing (happiness in the United States), health economics (self-assessed health in Europe) and development economics (sanitation ladders in Pakistan).

This paper can be deemed a sequel to Mendelson (1987); Allison and Foster (2004); and Kobus (2015) in so far as we devise formal tests for their proposed inequality relations. In

[1] We follow the approach of MacKinnon and Davidson (1996) and Davidson and MacKinnon (1998).

fact, all the methods discussed in this paper generalise to accommodate Mendelson (1987)'s more general family of quantile preserving spreads, of which MPS and FOSD are special cases. Yalonetzky (2013) devises a similar asymptotic test for first-order stochastic dominance with ordinal variables, which we extend to suit our purposes. As Davidson and Duclos (2013), we adopt a likelihood ratio statistic combined with bootstrap inference, but we additionally consider both a Z statistic and asymptotic inference. Our work is also related to Abul Naga and Stapenhurst (2015) and Abul Naga et al. (2020): while they perform inference on a random variable derived from a particular class of indices consistent with the MPS ordering, we perform inference on the binary outcome given by the partial ordering itself.

The rest of the paper proceeds as follows. Section 2 introduces the MPS partial ordering along with the required notation, motivates the null hypothesis of no ordering for a pair of distributions, and introduces a novel graphical representation of the parameter space. Section 3 develops our four proposed tests for the MPS partial ordering, combining two test statistics with two methods to compute p-values from the sampling distribution under the null. Section 4 studies and compares the size and power properties of the four tests, aided by novel size-power curves specifically tailored for our testing problem. Section 5 provides our empirical illustrations. Finally section 6 offers some concluding remarks.

2 Median Preserving Spreads and the Null Hypothesis of No Ordering

Following the required introduction of notation, this section defines Allison and Foster (2004)'s MPS partial ordering and describes our null and alternative hypotheses. Then we introduce a novel graphical technique for locating pairs of distributions relative to the null and alternative spaces.

2.1 Notation

Let $k \in \mathbb{N}$ denote the number of ordered categories and $[k] := \{1, \ldots, k\}$ denote the set of categories. We focus on a pair of samples $(\boldsymbol{x}, \boldsymbol{y})$ of respective sizes n_x and n_y. Each sample is a vector of frequencies which add up to the sample size, for example $\boldsymbol{x} = (x_1, \ldots, x_k) \in \mathbb{N}^k$ and $\sum_{i=1}^k x_i = n_x$. Since the states are ordered we can define the cumulants $\boldsymbol{X} = (\sum_{j=1}^1 x_j, \ldots, \sum_{j=1}^k x_j = n_x)$ of $\boldsymbol{x}$; the vector of cumulants $\boldsymbol{Y}$ is defined analogously for $\boldsymbol{y}$. We use $X_{[i]} := (X_1, \ldots, X_i)$ to denote the first i cumulants of $\boldsymbol{X}$. The sample space is then $\mathcal{X}(k, n_x, n_y) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{N}^k \times \mathbb{N}^k \mid X_k = n_x \text{ and } Y_k = n_y\}$. The combined sample is denoted by $\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{Y}$ with combined sample size $n_x + n_y$.

Our ultimate goal is to perform inference on the pair of distributions $(\boldsymbol{f}, \boldsymbol{g})$ underlying the pair $(\boldsymbol{x}, \boldsymbol{y})$. Specifically, $\boldsymbol{x} \sim \boldsymbol{f}$ and $\boldsymbol{y} \sim \boldsymbol{g}$ where f_i (respectively g_i) denotes the probability that any particular observation from population $\boldsymbol{f}$ (respectively $\boldsymbol{g}$) falls into category i. We denote their cumulative distribution functions (henceforth CDF) by $\boldsymbol{F}$ and $\boldsymbol{G}$ respectively. Let $\boldsymbol{L} = (L_1, \ldots, L_k) := \boldsymbol{W}/(n_x + n_y)$ be the sample-weighted average empirical distribution function (EDF).

The parameter space involving all possible pairs of ordered distributions with a given natural number of categories, $k > 1$, is defined by:

$$\Theta := \{(\boldsymbol{f}, \boldsymbol{g}) \in [0,1]^k \times [0,1]^k \mid F_k = G_k = 1\}.^2$$

A generic parameter vector is denoted by $\boldsymbol{\theta} = (\boldsymbol{f}, \boldsymbol{g}) \in \Theta$. Samples $\boldsymbol{x}$ and $\boldsymbol{y}$ are drawn from independent ordered multinomial distributions, so the likelihood of $(\boldsymbol{x}, \boldsymbol{y})$ given θ is:

$$\mathbb{P}_\theta[\boldsymbol{x}, \boldsymbol{y}] := \frac{n_x!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k f_i^{x_i} \frac{n_y!}{\prod_{i=1}^k y_i!} \prod_{i=1}^k g_i^{y_i}.$$

Also note that $(\frac{\boldsymbol{x}}{n_x}, \frac{\boldsymbol{y}}{n_y}) \in \Theta$.

[2] Even though the sample size (n_x, n_y) is normally considered a parameter of the multinomial distribution, we do not consider it as such because in our applications it is normally fixed (e.g. by survey design).

Finally, the following sample-size-weighted Euclidean metric for the distance between two distributions will prove useful in several instances below:

$$d(\boldsymbol{\theta}, \boldsymbol{\theta'}) = \sqrt{\frac{n_x}{n_x + n_y} \sum_{i=1}^{k} (f_i - f'_i)^2 + \frac{n_y}{n_x + n_y} \sum_{i=1}^{k} (g_i - g'_i)^2}.$$

2.2 Median Preserving Spreads

We define the median of $\boldsymbol{F}$ to be a category $m \in [k]$ such that $F_{m-1} < 0.5$ and $F_m \geq 0.5$. We assume a unique median category for the purposes of exposition, but all the results generalise to cases with multiple median categories.[3] Allison and Foster (2004) discuss the difficulties of defining suitable measures of dispersion for ordinal variables. They propose the partial ordering over the sample space $\mathcal{X}$ which ranks distributions according to their spread. Here we define the analogous partial ordering for a pair of distributions:

Definition 1 (Median Preserving Spread)**.** Let $(\boldsymbol{f}, \boldsymbol{g}) \in \Theta$. We say that $\boldsymbol{g}$ is a strict *median preserving spread* (MPS) of $\boldsymbol{f}$, or that $\boldsymbol{f}$ and $\boldsymbol{g}$ are ordered, and write $\boldsymbol{g} \succ \boldsymbol{f}$, if and only if there exists a category m such that all the following conditions hold:

[M1] $G_{m-1} < \frac{1}{2}$

[M2] $\frac{1}{2} < G_m$

[D1] $G_i > F_i$ for all $i \in [m-1]$

[D2] $G_i < F_i$ for all $i \in [k-1] \setminus [m-1]$.

We call $\boldsymbol{f}$ the *concentrated* distribution and $\boldsymbol{g}$ the *spread* distribution. If $\boldsymbol{g}$ is not an MPS of $\boldsymbol{f}$ then $\boldsymbol{f}$ and $\boldsymbol{g}$ are unordered, and $\boldsymbol{g} \not\succ \boldsymbol{f}$. If one or more of the inequalities holds with equality then we say that $\boldsymbol{g}$ is a *weak MPS* of $\boldsymbol{f}$ and write $\boldsymbol{g} \succeq \boldsymbol{f}$. A pair of samples $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}$ is ordered if and only if the distributions $\frac{\boldsymbol{x}}{n_x}$ and $\frac{\boldsymbol{y}}{n_y}$ are ordered.

[3] For the case of median-preserving spreads with multiple median categories see Kobus (2015).

2.3 The null hypothesis of no ordering

We propose tests of the null hypothesis that $\boldsymbol{g}$ is not a strict MPS of $\boldsymbol{f}$ because we are mainly interested in situations where $\boldsymbol{x}$ and $\boldsymbol{y}$ are ordered and want to confirm whether this is indicative of an ordering in the underlying populations. Following Davidson and Duclos (2013), if we reject the null hypothesis that the populations are *not* ordered, then we logically conclude that they *are* ordered.

The *null space* is the subset of all parameter values such that g is not a strict MPS of f:

$$\Theta_0 = \{(\boldsymbol{f}, \boldsymbol{g}) \in \Theta \mid \boldsymbol{g} \not\succ \boldsymbol{f}\}. \ ^{4}$$

Any distributional pair in the null space is denoted by $\boldsymbol{\theta_0} \in \Theta_0$. The alternative space is the complement of the null space, which is equivalent to the set of all ordered pairs:

$$\Theta_1 := \Theta_0^c = \{(\boldsymbol{f}, \boldsymbol{g}) \in \Theta \mid \boldsymbol{g} \succ \boldsymbol{f}\}$$

A generic element of Θ_1 is denoted by $\boldsymbol{\theta}_1$.

We can graphically depict a two dimensional projection of the null and alternative spaces relative to the whole parameter space. Specifically, a pair of distributions $(\boldsymbol{f}, \boldsymbol{g}) \in \Theta$ can be written as a set of k pairs of cumulants (F_i, G_i). Figure 1 plots the pairs of coordinates of distributions $\boldsymbol{F} = (3/24, 9/24, 17/24, 21/24, 1)$ and $\boldsymbol{G} = (7/24, 11/24, 15/24, 17/24, 1)$ in the unit square. Every set of coordinates must include the point $(1, 1)$ and the set of cumulant coordinates is necessarily non-decreasing as we move from left to right because both $\boldsymbol{F}$ and $\boldsymbol{G}$ must be individually non-decreasing. The median category of $\boldsymbol{F}$ (resp. $\boldsymbol{G}$) is given by the state corresponding to the first coordinate to the right of the vertical (resp. horizontal) line at $\frac{1}{2}$. We know the two distributions share the same median category ($m = 3$) because all the coordinates lie in the south-west and north-east quadrants. Any pair of distributions with a

[4]The set Θ_0 is rotationally symmetric, meaning that reversing the ordering of the categories does not alter the MPS partial ordering of the original distributions. Therefore, all the tests we propose are invariant to reverse ordering of the categories.

coordinate in either the north-west or south-east quadrants do not share the same median and therefore cannot be ordered. Similarly, we can see that $\boldsymbol{f}$ first-order dominates $\boldsymbol{g}$ below the median and $\boldsymbol{g}$ first-order dominates $\boldsymbol{f}$ at and above the median because all the coordinates lie in the interiors of the two triangles with vertices $(0,0),(\frac{1}{2},\frac{1}{2}),(0,\frac{1}{2})$ and $(1,1),(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2})$, labelled Θ_1 in figure 1. It follows from definition 1 that $\boldsymbol{g}$ is an MPS of $\boldsymbol{f}$. In general, $(\boldsymbol{f},\boldsymbol{g}) \in \Theta_1$ if and only if their coordinates are *all* contained in the triangles labelled Θ_1. Conversely $(\boldsymbol{f},\boldsymbol{g}) \in \Theta_0$ if and only if *at least one* of their coordinates is contained outside of these triangles, in either of the parallelograms with vertices $(0,\frac{1}{2}),(0,1),(1,1),(\frac{1}{2},\frac{1}{2})$ and $(0,0),(1,0),(1,\frac{1}{2}),(\frac{1}{2},\frac{1}{2})$, labelled Θ_0 in figure 1.

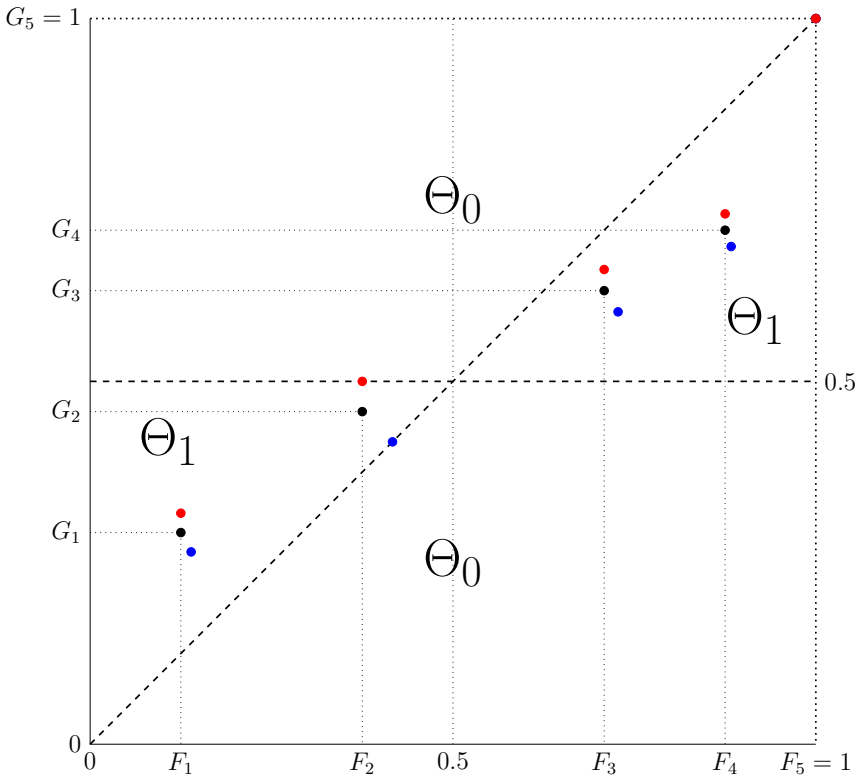

Figure 1: The parameter space, null space and alternative space projected onto the unit square. The red and blue dots illustrate, respectively, the closest distributions on the median and dominance boundaries to the distribution represented by black dots.

The boundary of the null space plays an important role in the proposed tests. Firstly,

the QMLE lies in the boundary of the null space. We will use the QMLE both to calculate the likelihood ratio statistic, and to draw bootstrap samples when carrying out bootstrap inference. Secondly, when we study the empirical size of the tests, we will choose data-generating processes in the boundary of the null space in order to provide an upper bound on the size of the tests of distributions in the interior of the null space. Thirdly, when we study the power of the tests, we will compare the rejection rates of distributions in the alternative space with rejection rates of the corresponding 'closest null' distribution, which lie in the boundary of Θ_0.

We characterise the boundary of the null space as the union of two sets:

Definition 2. The *median subset of the boundary* of Θ_0 (henceforth 'median boundary') is the set of all weakly ordered distributions for which at least one of the median constraints in definition 1 hold with equality:

$$\bar{M} = \{(\boldsymbol{F}, \boldsymbol{G}) \in (\Delta[k])^2 \mid \boldsymbol{F} \succeq \boldsymbol{G} \text{ and } G_i = \frac{1}{2} \text{ for some } i \in [k]\}.$$

The *dominance subset of the boundary* of Θ_0 (henceforth 'dominance boundary') is the set of all weakly ordered distributions for which at least one of the dominance constraints in definition 1 hold with equality:

$$\bar{D} = \{(\boldsymbol{F}, \boldsymbol{G}) \in (\Delta[k])^2 \mid \boldsymbol{F} \succeq \boldsymbol{G} \text{ and } F_i = G_i \text{ for some } i \in [k]\}.$$

The boundary can now be characterised:

Lemma 1. *The* boundary *of the null space is equal to the union of the median and dominance boundaries:*

$$\partial\Theta_0 = \bar{M} \cup \bar{D}.$$

Proof. See appendix A. □

A pair of distributions lies on the median boundary if and only if it has one or more coordinates lying on the horizontal dashed line intersecting the vertical axis at (0,0.5) in

figure 1, and all other coordinates lying in the interior of Θ_0 triangles. Similarly, a pair of distributions lies on the dominance boundary if and only if it has one or more coordinates lying on the 45° dashed line in figure 1, and all other coordinates lying in the interior of Θ_0 triangles. We illustrate examples of distributions on the median and dominance boundaries in figure 1.

3 Statistical Tests of No Ordering

A statistical test can be regarded as a function $P : \mathcal{X} \to [0, 1]$ returning a P value for every sample in the sample space $\mathcal{X}$. The P value describes the likelihood of observing a sample 'as extreme' as $(\boldsymbol{x}, \boldsymbol{y})$ when the null hypothesis is true, so a low P value can be taken as evidence that the null hypothesis is false. If the P-value is less than $\alpha \in (0, 1)$ then we 'reject the null hypothesis at the α% level.'

A test statistic is a function $\mathcal{S} : \mathcal{X} \to \mathbb{R}$ which formalises what it means for one sample to be 'as extreme' as another by associating each sample with a real number: sample $(\boldsymbol{x}', \boldsymbol{y}')$ is more extreme under the null than $(\boldsymbol{x}, \boldsymbol{y})$ if $\mathcal{S}(\boldsymbol{x}', \boldsymbol{y}') \geq \mathcal{S}(\boldsymbol{x}, \boldsymbol{y})$. Thus we consider four tests of the form $T(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{P}_{\theta_0}[\mathcal{S}(\boldsymbol{x}', \boldsymbol{y}') \geq \mathcal{S}(\boldsymbol{x}, \boldsymbol{y})]$. The remainder of this section discusses our choices of test statistic (LR or Z) and the method of inference (asymptotic or bootstrap).

3.1 Test statistics

The LR statistic The log likelihood ratio (LR) statistic is a natural choice due to its intuitive construction and well-known optimality in terms of uniform power (see section 4.2). The LR statistic of a sample $(\boldsymbol{x}, \boldsymbol{y})$ is the ratio of its unconstrained maximum likelihood function to its constrained counterpart, the QMLE.

Definition 3. The log likelihood ratio (LR) statistic of a sample $(\boldsymbol{x}, \boldsymbol{y})$ is given by:

$$\mathrm{LR}(\boldsymbol{x}, \boldsymbol{y}) := 2[\ln(\mathbb{P}_{\theta^*}[\boldsymbol{x}, \boldsymbol{y}]) - \ln(\mathbb{P}_{\tilde{\theta}}[\boldsymbol{x}, \boldsymbol{y}])] \tag{1}$$

where $\theta^* \in \arg\max_{\theta\in\Theta} \mathbb{P}_\theta[\boldsymbol{x},\boldsymbol{y}]$ is the maximum likelihood estimator (MLE), and $\tilde{\theta} \in \arg\max_{\theta\in\Theta_0} \mathbb{P}_\theta[\boldsymbol{x},\boldsymbol{y}]$ is the QMLE.

We necessarily have $\mathrm{LR}(\boldsymbol{x},\boldsymbol{y}) \geq 0$. Lemma 2 derives closed form expressions for the MLE and QMLE.

Lemma 2.

1. *The likelihood maximiser (unconstrained) of a sample* $(\boldsymbol{x},\boldsymbol{y})$ *is given by*

$$\boldsymbol{\theta}^* = (\boldsymbol{x}/n_x, \boldsymbol{y}/n_y).$$

2. *If* $\boldsymbol{x}$ *is not a strict MPS of* $\boldsymbol{y}$*, then the quasi maximum likelihood estimator (QMLE; constrained), is given by*

$$\tilde{\boldsymbol{\theta}} = (\boldsymbol{x}/n_x, \boldsymbol{y}/n_y)$$

and $LR(\boldsymbol{x},\boldsymbol{y}) = 0$.

3. *Otherwise, if* $\boldsymbol{x}$ *is a strict MPS of* $\boldsymbol{y}$*, then the QMLE is given by either one of the following* $k-1$ dominance-constrained *distributions* $\{\tilde{\boldsymbol{\theta}}^{D_j}\}_{j\in[k-1]} \in \bar{D}$ *defined by:*

$$\tilde{\theta}_i^{D_j} = (\tilde{f}_i^{D_j}, \tilde{g}_i^{D_j}) = \begin{cases} \left(\frac{x_i}{X_j}L_j, \frac{y_i}{Y_j}L_j\right) & \text{if } i \leq j \\ \left(\frac{x_i}{n_x - X_j}(1-L_j), \frac{y_i}{n_y - Y_j}(1-L_j)\right) & \text{otherwise;} \end{cases}$$

or else it is one of the following two median-constrained *distributions,* $\{\tilde{\boldsymbol{\theta}}^{M_j}\}_{j=m-1,m} \in \bar{M}$ *defined by:*

$$\tilde{\theta}_i^{M_j} = (\tilde{f}_i^{M_j}, \tilde{g}_i^{M_j}) = \begin{cases} \left(\frac{x_i}{n_x}, \frac{y_i}{2Y_j}\right) & \text{if } i \leq j \\ \left(\frac{x_i}{n_x}, \frac{y_i}{2(n_y - Y_j)}\right) & \text{otherwise.} \end{cases}$$

The likelihood ratio statistic is then given by

$$LR(\boldsymbol{x},\boldsymbol{y}) = 2\ln\{\frac{\mathbb{P}_{\theta^*}[\boldsymbol{x},\boldsymbol{y}]}{\max\{\mathbb{P}_{\tilde{\theta}^{D_1}}[\boldsymbol{x},\boldsymbol{y}], ..., \mathbb{P}_{\tilde{\theta}^{D_{k-1}}}[\boldsymbol{x},\boldsymbol{y}], \mathbb{P}_{\tilde{\theta}^{M_{m-1}}}[\boldsymbol{x},\boldsymbol{y}], \mathbb{P}_{\tilde{\theta}^{M_m}}[\boldsymbol{x},\boldsymbol{y}]\}}\}.$$

Proof. See appendix A □

In lemma 2, the multiple cases arise from a Kuhn-Tucker optimization problem where, depending on the regime of binding constraints, we obtain the various solutions above. A large likelihood ratio is evidence that the constraint is hard to satisfy therefore rendering the null hypothesis unlikely to be true.

The Z statistic Z statistics have been used in test of stochastic dominance for multivariate distributions of ordinal variables (e.g. Yalonetzky, 2013). Let $\sigma_i^L = \sqrt{L_i(1-L_i)\frac{n_x+n_y}{n_x n_y}}$ be the standard error of the pooled sample's cumulative frequency L_i, and $\sigma_i^Y = \sqrt{[(Y_i/n_y)(1-Y_i/n_y)]/n_y}$ be the standard error of the sample cumulative frequency Y_i. Then consider the Z statistic in definition 4:

Definition 4. The Z statistic for a multinomial sample $(\boldsymbol{x}, \boldsymbol{y})$ is given by:

$$Z(\boldsymbol{x}, \boldsymbol{y}) = \min\left\{Z_D^<, Z_D^\geq, Z_M\right\},$$

where $Z_D^< := \min\{(Y_i/n_y - X_i/n_x)/\sigma_i^L |\ i < m_y\}$, $Z_D^\geq := \min\{(X_i/n_x - Y_i/n_y)/\sigma_i^L |\ i \geq m_y\}$ and $Z_M := \min\{(0.5 - Y_{m_x-1}/n_y)/\sigma_{m_x-1}^Y, (Y_{m_x}/n_y - 0.5)/\sigma_{m_x}^Y\}$.

The term Z_M is positive if and only if $m_x = m_y$ (corresponding to conditions [M1] and [M2] in definition 1 for the population counterparts). Hence they are helpful to test the equality of the population medians, which is necessary (but insufficient) to establish an MPS ordering.

The term $Z_D^<$ is the minimum among all the standarised distances of sample cumulative frequencies **Y**-**X** below m_y; whereas $Z_D^\geq$ is the minimum among all the standarised distances of sample cumulative frequencies **X**-**Y** at and above m_y. Note the similarities with their (unstandarised) population counterparts in conditions [D1] and [D2], respectively. The three statistics are jointly positive, and hence Z is positive, if and only if the sample counterparts of conditions [D1], [D2], [M1] and [M2] hold together. That is, Z is positive *if and only if* $\mathbf{Y} \succ \mathbf{X}$.

3.2 Method of Inference

The ideal choice of null distribution (Lehmann and Romano, 2005) is that which maximises the probability of the upper contour set $\{\boldsymbol{x}', \boldsymbol{y}' \mid |\mathcal{S}(\boldsymbol{x}', \boldsymbol{y}')| \geq |\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})|\}$, namely $\boldsymbol{\theta_0} \in \arg\max_{\boldsymbol{\theta} \in \Theta_0} \mathbb{P}_\theta[|\mathcal{S}(\boldsymbol{x}', \boldsymbol{y}')| \geq |\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})|]$, because this choice ensures that the test always has the correct size (see section 4.1). To the best of our knowledge, there is no analytical expression for it in the context of tests involving our specific null space, and numerical solutions are computationally intensive. Instead, we follow the standard approach (e.g. Davidson and Duclos (2013)) of using the QMLE of the observed sample $\boldsymbol{\theta_0} = \tilde{\boldsymbol{\theta}}$, characterised in lemma 2.

We approximate the probability $\mathbb{P}_{\theta_0}[|\mathcal{S}(\boldsymbol{x}', \boldsymbol{y}')| \geq |\mathcal{S}(\boldsymbol{x}, \boldsymbol{y})|]$ by using either the asymptotic or the bootstrapped distribution of the test statistics. The following theorem will be important for the purpose of asymptotic inference.

Theorem 1 (Asymptotic distributions of test statistics under the null)**.** *Suppose the true distribution pair lies in the boundary, so that* $(\boldsymbol{x}, \boldsymbol{y}) \sim \boldsymbol{\theta_0} \in \partial\Theta_0$*, then:*

1. $LR(\boldsymbol{x}, \boldsymbol{y}) \xrightarrow{d} \chi^2(1)$*, and*
2. $Z(\boldsymbol{x}, \boldsymbol{y}) \xrightarrow{d} \mathcal{N}(0, 1)$.

Hence if $\boldsymbol{\theta_0}$ lies in the boundary of the null space, point 1 of the theorem states that the LR statistics converges to a chi-squared variable with one degree of freedom. The one degree of freedom in the chi-squared distribution stems from the difference between the dimensions of the constrained and unconstrained maximum likelihood solutions ((Mood et al., 1974, p. 440); see lemma 2 in appendix A). In the case where $\boldsymbol{\theta_0}$ lies in the interior of the null space, then the LR statistic will generally be lower than for distributions in the boundary, therefore the distribution of the statistic will be first-order stochastically dominated by the $\chi^2(1)$ distribution. We refer the reader to Davidson and Duclos (2013, p. 105). Likewise, the Z statistic is asymptotically standard normal when $\boldsymbol{\theta_0}$ lies in $\partial\Theta_0$, but otherwise is bounded by $\mathcal{N}(0, 1)$. We suggest in practice to approximate the P value of a sample $(\boldsymbol{x}, \boldsymbol{y})$

by $1-\zeta[\mathcal{S}(\boldsymbol{x},\boldsymbol{y})]$, where ζ denotes the CDF of $\mathcal{N}(0,1)$ and $\chi^2(1)$ distributions, respectively. Thus, as we document in our Monte Carlo investigations, the size of the test can be expected to be smaller than the associated nominal value.

Instead of calculating the test statistic of all the samples in the sample space, bootstrap tests approximate the distribution of the test statistic by its empirical distribution in a sample of B samples $\{(\boldsymbol{x}^i,\boldsymbol{y}^i)\}_{i\in[B]}$, each independently drawn from $(\boldsymbol{x},\boldsymbol{y})$. The P value of a sample $(\boldsymbol{x},\boldsymbol{y})$ is then approximated by $\#\{(\boldsymbol{x}^i,\boldsymbol{y}^i) \mid \mathcal{S}(\boldsymbol{x}^i,\boldsymbol{y}^i) \geq \mathcal{S}(\boldsymbol{x},\boldsymbol{y}), i\in[B]\}/B$.

Proposition 1 (Bootstrap p-values under the null; Davison and Hinkley (1997))**.** *Suppose the true distribution pair lies in the null space, so that* $(\boldsymbol{x},\boldsymbol{y}) \sim \boldsymbol{\theta_0} \in \Theta_0$ *as well as* $(\boldsymbol{x}^i,\boldsymbol{y}^i) \sim \boldsymbol{\theta_0} \in \Theta_0$ *for all* $i \in [B]$ *where* $B \in \mathbb{N}\setminus\{0\}$*, then the bootstrap p-values for a given statistic* $\mathcal{S}(\boldsymbol{x},\boldsymbol{y})$ *are:*

$$T_{B\mathcal{S}}(\boldsymbol{x},\boldsymbol{y}) = \#\{(\boldsymbol{x}^i,\boldsymbol{y}^i) \mid |\mathcal{S}(\boldsymbol{x}^i,\boldsymbol{y}^i)| \geq |\mathcal{S}(\boldsymbol{x},\boldsymbol{y})|, i\in[B]\}/B.$$

Combining the two test statistics with these two methods of approximation gives a family of four tests and respective P values:

1. Asymptotic Z test: $T_{AZ}(\boldsymbol{x},\boldsymbol{y}) = 1-\Phi(Z(\boldsymbol{x},\boldsymbol{y}))$.
2. Asymptotic LR test: $T_{A\mathrm{LR}}(\boldsymbol{x},\boldsymbol{y}) = 1-\chi^2(\mathrm{LR}(\boldsymbol{x},\boldsymbol{y});1)$.
3. Bootstrap Z test: $T_{BZ}(\boldsymbol{x},\boldsymbol{y}) = \#\{(\boldsymbol{x}^i,\boldsymbol{y}^i) \mid Z(\boldsymbol{x}^i,\boldsymbol{y}^i) \leq Z(\boldsymbol{x},\boldsymbol{y}), i\in[B]\}/B$.
4. Bootstrap LR test: $T_{B\mathrm{LR}}(\boldsymbol{x},\boldsymbol{y}) = \#\{(\boldsymbol{x}^i,\boldsymbol{y}^i) \mid LR(\boldsymbol{x}^i,\boldsymbol{y}^i) \geq LR(\boldsymbol{x},\boldsymbol{y}), i\in[B]\}/B$.

In the next section we use Monte Carlo simulations to investigate the size and power properties of these four tests.

4 Size and Power Properties

In this section we introduce novel graphical tools, namely the size-boundary curve for the study of test size, and power-locus curves for for the study of test power. We adopt the

standard practice of using Monte Carlo experiments to measure the empirical distribution of P values produced by the tests. Specifically, we draw $M = 100,000$ independent samples $(\boldsymbol{x}^i, \boldsymbol{y}^i)$ from sets of judiciously chosen *data generating processes* (DGPs) $\theta \in \Theta$, and calculate all the P-values, $\{T(\boldsymbol{x}^i, \boldsymbol{y}^i)\}_{i\in[M]}$ for each test T. The rejection rate of a nominal size α test at θ is then estimated by $\#\{\boldsymbol{x}^i, \boldsymbol{y}^i) \mid T(\boldsymbol{x}^i, \boldsymbol{y}^i) \leq \alpha\}/M$.

We focus on DGPs with just two categories, i.e. $k = 2$. This class of DGPs is easy to visualise because it is mathematically equivalent to the unit square, with f_1 on one axis and g_1 on the other. The median boundary is mathematically equivalent to the horizontal line intersecting the vertical axis at $(0, 0.5)$ and the dominance boundary is mathematically equivalent to the 45°line. Because the boundary is unidimensional it is easy to show how rejection rates vary along it. Similarly, we are able to identify a unidimensional 'interior locus' which allows us to illustrate how power varies against different DGPs in the alternative space. We argue in section 4.3 that tests' behaviour in the $k = 2$ case is indicative of behaviour in higher dimensions.

4.1 Size

If T satisfies the inequality $\mathbb{P}_{\boldsymbol{\theta}_0}[T(\boldsymbol{x}, \boldsymbol{y}) < \alpha] \leq \alpha$ for all null distributions $\theta \in \Theta_0$, then we say that the test is *correctly sized* at level α; otherwise it is *oversized*. Our tests will have higher rejection rates on the boundary than anywhere else in the null space, therefore it is sufficient to study their behaviour of the tests on the boundary alone. In order to build a comprehensive picture of behaviour in the boundary, we calculate the rejection rates of our tests along a grid of different DGPs in the boundary of the $k = 2$ null space, for a range of sample sizes. Figure 2 illustrates the DGP's used for the cases $n_x = n_y = 10, 100, 1000$.[5] Our interest in small sample sizes, and more specifically in small ratio of n_x to n_y and n_y to n_x is three fold: (a) to investigate the relative merits of the bootstrap versus asymptotic

[5]The precise choices of boundary DGPs for these and other sample sizes are listed in appendix B.

inference in relation to size and power of Z and LR tests; (b) to explore lower bounds on sample size in relation to the performance of the tests; and (c) to highlight the asymmetric role of sample sizes of the spread distribution (n_y) and the concentrated distribution (n_x) in the statistical performance of the four tests.

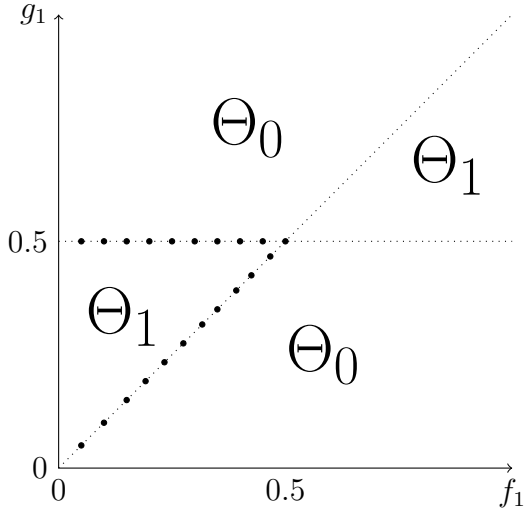

Figure 2: Boundary DGPs used to generate size curves for cases $k = 2$ and $n_x = n_y$.

Results Figure 3 shows the rejection rate of all four tests at the 5% nominal level, as a function of the first coordinate of the boundary DGPs. In each panel, the first half of the horizontal axis, from 0 to 0.5, corresponds to the median boundary (moving along the horizontal dotted line from coordinate (0,0.5) to (0.5,0.5) in figure 2) ; and the second half of the horizontal axis, from 0.5 to 1, corresponds to the dominance boundary (moving along the diagonal dotted line from coordinate (0.5,0.5) to the origin in figure 2). The intersection of the median and dominance boundaries, coincides with the point 0.5 (the kink of the dotted line in the middle of figure 2). From top-left downward and rightward, panels in row i show results for $n_x = 10^i$ while panels in column i show results for $n_y = 10^i$, where $i = 1, 2, 3$.

All the tests are correctly sized in most cases. Exceptions arise on the dominance boundary when n_x, the size of the sample drawn from the more concentrated distribution, is small

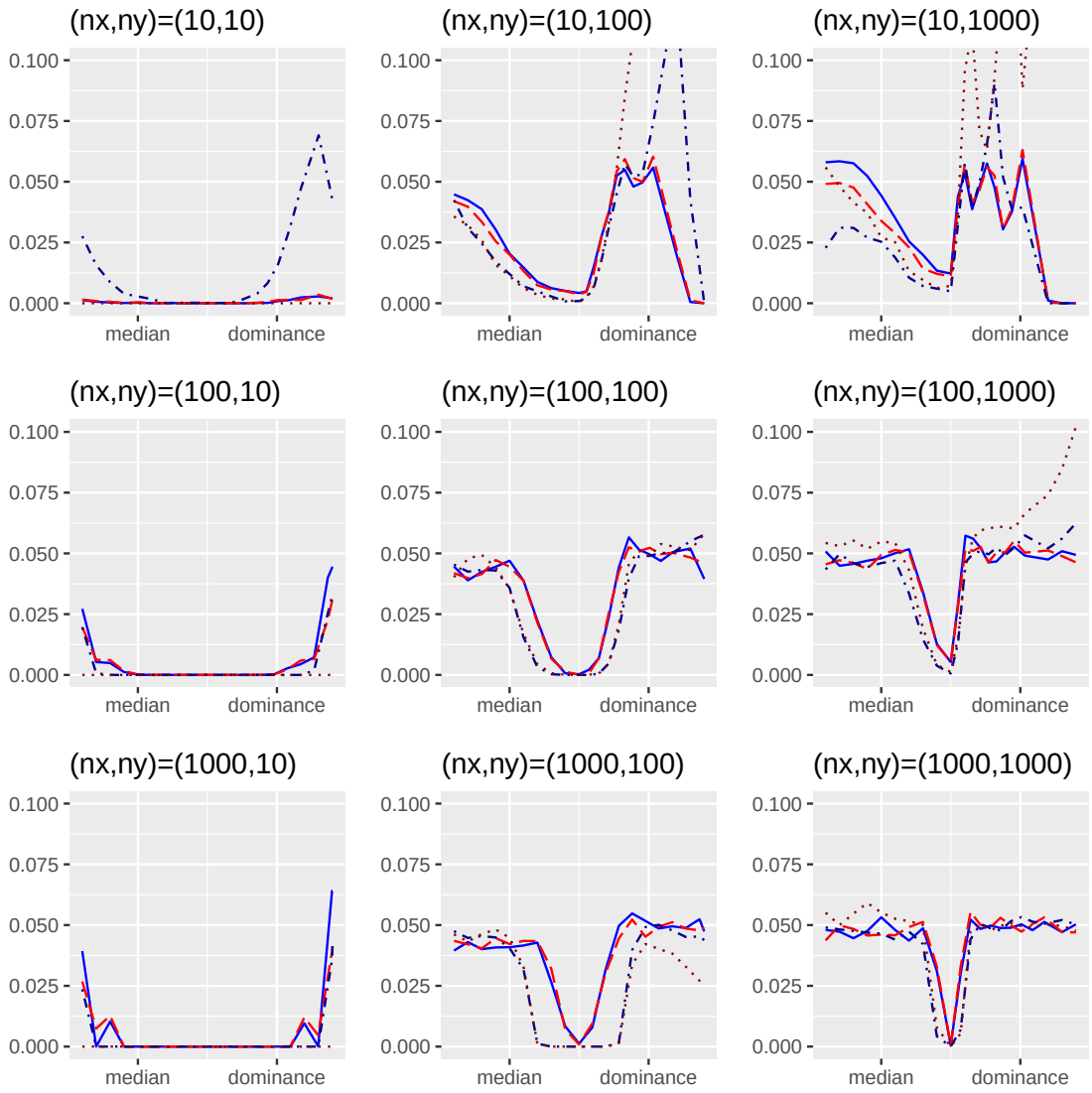

Figure 3: Size-boundary curves (for nominal 5% tests). Key: Solid/light blue — bootstrap LR; dashed/light red — bootstrap Z; dotdash/dark blue — asymptotic LR; dotted/dark red — asymptotic Z.

relative to the size of the sample drawn from the more polarised distribution. In these cases, tests based on the Z statistic can have sizes more than double their nominal levels. Meanwhile, the sizes of tests based on the LR statistic are never more than 20% above their nominal values. Two other exceptions occur: the asymptotic tests are both slightly oversized near the end of the median boundary in the case $(n_x, n_y) = (10, 1000)$, and the bootstrap LR test is slightly oversized near the end of the dominance boundary in the case $(n_x, n_y) = (1000, 10)$.

The rejection rates of all tests drop to zero near the intersection of the median and dominance boundaries, especially when the sample size of the less concentrated distribution is small. The region of the boundary where the rejection rate drops to zero vanishes as the sample sizes increase. The lower rejection rates vis-a-vis those in other points in the boundary are not surprising: for points other than (0.5,0.5), the proportion of neighbouring distributions that belong to the null and alternative space are of equal size, namely 1/2 and 1/2. However, at (0.5,0.5), the proportion of neighbouring distributions that belong to the null space is now equal to 3/4, whereas the proportion of neighbouring distributions that belong to the alternative space is now equal to 1/4. For this reason, the probability of a sample with an empirical distribution in the null space is more likely, leading to fewer rejections of the null hypothesis.

4.2 Power

Besides correct size, the other crucial property for statistical tests is the ability to distinguish between true and false hypotheses. The power of an α-sized test against an alternative distribution $\boldsymbol{\theta_1} \in \Theta_1$ has traditionally been defined to equal the probability of an α-level test rejecting H_0 under θ_1, namely $\mathbb{P}_{\boldsymbol{\theta_1}}[T(\boldsymbol{x}, \boldsymbol{y}) \leq \alpha]$. However, Davidson and MacKinnon (1998) propose the measure:

$$\psi(T; \theta_1, \alpha) \equiv \mathbb{P}_{\boldsymbol{\theta_1}}[T(\boldsymbol{x}, \boldsymbol{y}) \leq \mathbb{P}_{\boldsymbol{\theta_0}}[T(\boldsymbol{x}, \boldsymbol{y}) \leq \alpha]], \tag{2}$$

where $\boldsymbol{\theta_0}$ is the "closest null distribution" to the alternative $\boldsymbol{\theta_1}$. We take $\boldsymbol{\theta_0}$ to be the distribution in the null space which minimises the sample-size-weighted Euclidean distance to the DGP $\boldsymbol{\theta_1}$. In the case $k = 2$, there there are only two candidates for the closest null distribution: the closest distribution in the median boundary, $\boldsymbol{\theta}^M := (\boldsymbol{f}, (\frac{1}{2}, \frac{1}{2}))$; and the closest distribution in the dominance boundary, $\boldsymbol{\theta}^D := (\frac{n_x \boldsymbol{f} + n_y \boldsymbol{g}}{n_x + n_y}, \frac{n_x \boldsymbol{f} + n_y \boldsymbol{g}}{n_x + n_y})$. Figure 1 illustrates both the closest pair of distributions on the median boundary, denoted by red circles, and the closest pair of distributions on the dominance boundary, illustrated by blue circles, to the pair of distributions denoted by black circles, for $k = 5$.

A test T is *uniformly more powerful* than a test T' at level α if $\psi(T; \boldsymbol{\theta_1}, \alpha) \geq \psi(T'; \boldsymbol{\theta_1}, \alpha)$ for all $\boldsymbol{\theta_1} \in \Theta_1$. Typically a uniformly most powerful test does not exist. The best we can do is comparing the power of the tests at different alternative DGPs. Because the boundary separating the null and alternative spaces of the tests introduced in this paper arises as the union of the dominance and median boundaries, we focus on studying power against alternatives that are equidistant from these median and dominance boundaries. We refer to these DGPs as the 'interior locus'. Figure 4 shows the grid of DGPs on the interior locus, connected by a solid blue line, that we use for our experiments with $n_x = n_y$.[6] We also show, for each of these alternative DGPs, the two closest null distributions — one on each boundary — connected to the interior locus by a red dashed line. This interior locus is worth studying because it partitions the alternative space into a set of DGPs which are closest to the median boundary and a set of DGPs closest to the dominance boundary, and every DGP in the alternative space can be uniquely identified with a point on the interior locus which shares the same closest null distribution (be it on the median or the dominance boundary). Moreover, we expect all the tests to have lower power against an arbitrary alternative DGP than against its counterpart in the interior locus, thus the interior locus provides an upper bound on the test's power.

[6]Precise values for these DGPs, and those used for other ratios of n_x to n_y are listed in appendix B.

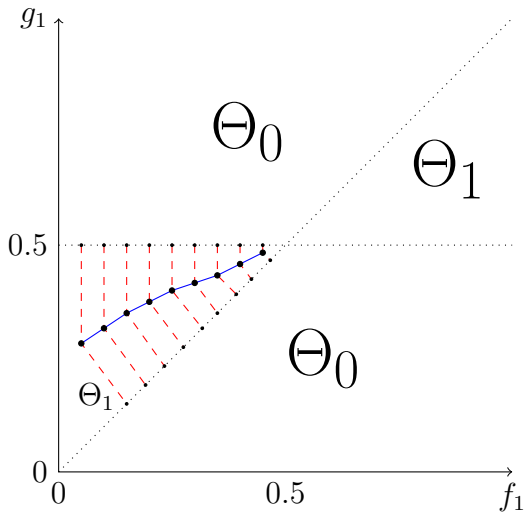

Figure 4: Alternative DGPs used to generate power curves for cases $k = 2$ and $n_x = n_y$.

Results In figure 5 we introduce a novel *power-locus curve* used to investigate power properties of the various tests. By definition, the alternative DGPs in the 'interior locus' have two closest null distributions. Therefore there are two ways to evaluate the expression in equation (2). Each panel of figure 5 illustrates both methods for a different pair of sample sizes. The first half of the horizontal axis, from $f_1 = 0$ to $f_1 = 0.5$, depicts the 'median power curve': the power against each alternative DGP from left to right in terms of figure 4, calculated using the closest null on the *median* boundary. The second half, from $f_1 = 0.5$ to $f_1 = 1$, depicts the reflected 'dominance power curve': power against each alternative calculated using the closest null on the *dominance* boundary and in the reverse order. Such display of results enables us to see how the power varies as the DGP approaches the intersection of the two boundary lines from the 'median direction' and from the 'dominance' direction, respectively. We expect that power against alternatives near the median boundary will behave similarly to the median power curve, and that power against alternatives near the dominance boundary will behave similarly to the dominance power curve.

All the tests are able to perfectly distinguish some alternatives from the null space when-

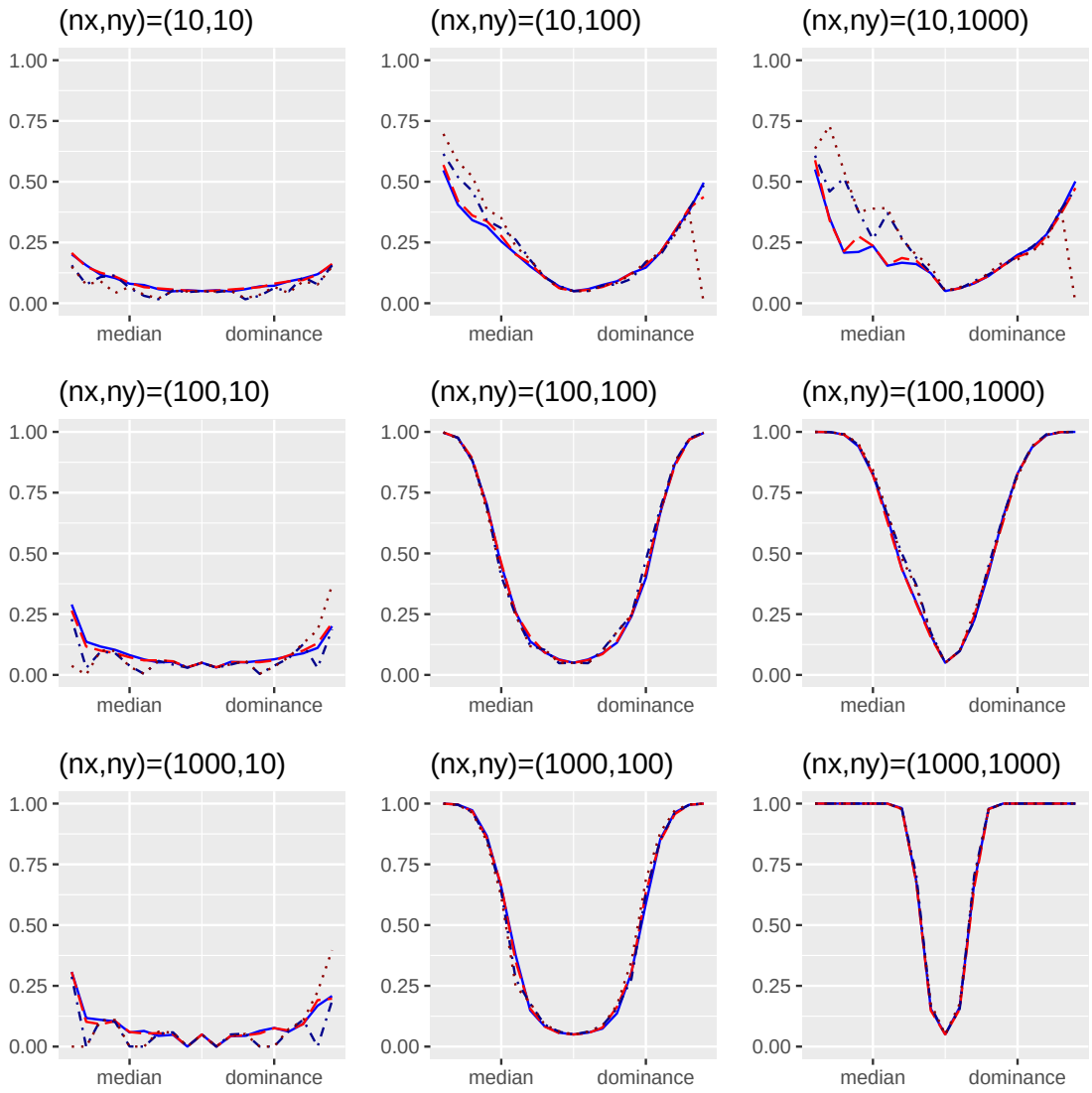

Figure 5: Power-locus curves (for nominal 5% tests). Key: Solid/light blue — bootstrap LR; dashed/light red — bootstrap Z; dotdash/dark blue — asymptotic LR; dotted/dark red — asymptotic Z.

ever both sample sizes are above 100. By the time both sample sizes reach 1000, the tests are able to perfectly distinguish a false hypothesis along distributions pertaining to three quarters of the interior locus. However, power drops rapidly when sample size falls below 100: power halves when n_x falls from 100 to 10, and reduces by a factor of 4 when n_y falls from 100 to 10. All the tests have more or less the same power when sample sizes are both in the order 100 or higher. Even with smaller sample sizes, the choice of test statistic appears to have minimal impact on power. However, there is evidence of systematic disparities between asymptotic and bootstrap inference for smaller sample sizes. When n_x is both small relative to n_y and very small in absolute terms, the asymptotic tests are more powerful against some distributions nearer the median boundary. However, when n_y is very small, the power of asymptotic inference is both erratic and consistently lower than the power of bootstrap inference.

4.3 More than Two Categories

A distribution $(\boldsymbol{f}, \boldsymbol{g})$ with $k > 2$ categories can be decomposed into $k-1$ two-category distributions $(\boldsymbol{f}^i, \boldsymbol{g}^i)$ defined by $(\boldsymbol{F}^i, \boldsymbol{G}^i) = ((F_i, 1), (G_i, 1))$ for any $i \in [k-1]$. The rejection rate of a test at $(\boldsymbol{f}, \boldsymbol{g})$ is therefore a function of the $k-1$ rejection rates at each of the $(\boldsymbol{f}^i, \boldsymbol{g}^i)$. For example, intersection-union tests (Berger, 1982) reject the null that $(\boldsymbol{f}, \boldsymbol{g})$ are not ordered if and only if they reject all the $k-1$ hypotheses that each of the $(\boldsymbol{f}^i, \boldsymbol{g}^i)$ are not ordered. Graphically, this means that we infer that all the coordinates in figure 1 are contained within the two triangles representing the alternative space, if and only if we infer that the coordinate closest to the edge of the triangles is nonetheless inside. The study of real world DGP's characterised by more than two categories in taken up in section 5.

5 Empirical illustrations

We consider three real-world inequality assessments: happiness in the United States, self-reported health in a set of European countries, and sanitation ladders in Pakistan. In each of the three applications we undertake 499 bootstrap replications. In the context of self-reported health and sanitation ladders we present p-value curves constructed from 1000 Monte Carlo simulations.

5.1 Happiness inequality in the United States

We revisit the study of Dutta and Foster (2013) on happiness inequality in the United States. They use data from the U.S. General Social Survey (GSS) between 1972 and 2010 (Dutta and Foster, 2013, table 1, p. 402). The GSS asks the following ordered-response question on wellbeing: "Taken all together, how would you say things are these days — would you say that you are 'very happy', 'pretty happy' or 'not too happy?' " Dutta and Foster (2013) did not test whether the documented ordering of happiness distributions was statistically significant. The family of tests developed in this paper provides the required statistical inference.

Table 1 reports P values of the bootstrap LR test, where an entry in row i of column j is the P value of the sample under the null hypothesis that year j distribution is not an MPS of year i distribution. A blank cell indicates that column j sample is not an MPS of the row i sample. Our results show that most of these inequality comparisons are (individually) statistically significant, with P values close to 0. The inferential exercise broadly supports the underlying pattern identified by Dutta and Foster (2013), namely a fall in happiness inequality across the 70s and 80s, that is reversed in the 90s and 2000s.

There are, however, some noteworthy exceptions. For example consider the finding that 'happiness inequality was lower in 1985 compared with seventeen other years' (Dutta and Foster, 2013, p. 405). The P values reported in Table 1 reveal that the P value of the

Table 1: Bootstrap LR p-values for Dutta and Foster (2013, table 2).

	72	73	74	75	76	77	78	80	82	83	84	85	86	87	88	89	90	91	93	94	96	98	00	02	04	06	08	10
72																												
73	.00		.03																									
74																												
75	.00																											
76	.00	.38	.01																									
77	.00	.00	.00					.00			.00																	
78	.00	.00	.00			.04		.00			.00			.00														
80	.00	.01	.00																									
82	.00	.00	.00		.44																							
83	.00	.05	.00	.00	.09																						.02	
84	.00	.00	.00																									
85	.00	.00	.00	.02	.00	.00		.00	.00		.00		.00	.00		.02			.02			.11	.00		.00	.00		
86	.00	.00	.00			.00		.00			.00			.00														
87	.05		.00																									
88	.00	.00	.00					.03			.09			.00														
89	.00	.00	.00		.02	.00		.00	.03		.00		.00	.00														
90	.00	.00	.00			.45		.00			.00			.00	.03													
91	.00	.00	.00	.00	.00	.00		.00	.00		.00			.02					.01			.00	.05	.02	.00	.00		
93	.00	.00	.00		.00	.00		.00	.00		.00														.14	.14		
94	.00	.00	.00	.00	.00			.00	.00	.00	.00														.00		.00	
96	.00	.00	.00	.00	.00	.00		.00	.00		.00											.00		.00	.00	.27		
98	.00	.00	.00	.16	.00	.10		.00	.00		.00														.02			
00	.00	.00	.00		.00	.00		.00	.00		.00			.27														
02	.00	.00	.00	.00	.00			.00	.00		.00														.00			
04	.00	.00	.00		.00				.50																			
06	.00	.00	.00		.00	.00		.00	.00		.00														.40			
08	.00	[illegible]																										

comparison between 1985 and 1998 equals 0.11. Other comparisons where the P value of the test is 10% or higher include the (1993, 2004) and (1993, 2006) pairs (both with a P value of 0.14), the (1998, 1975) pair (0.16), the (2000, 1987) pair (0.27), the (2004, 1982) pair (0.50), and the (2006, 2004) pair (0.40). If one adopts the standard convention of failing to reject a null hypothesis when the associated P value exceeds the 5% level, one would not find sufficient statistical evidence to support the conclusion that distributions i and j were ordered in the aforementioned comparisons.

5.2 Inequality in self-assessed health in the European Union

Self-assessed health (SAH) measures are increasingly used in health surveys, as such subjective assessments of well-being have shown to be strong predictors of morbidity as well as mortality (Latham and Peek, 2013). The Survey on Incomes and Living Conditions (SILC) conducted by EUROSTAT collects data on five levels of self-assessed health in the European Union. Respondents choose from the following ordered subjective health categories: (1) very bad, (2) bad, (3) fair, (4) good, and (5) very good. In 2017, the multinomial distributions of the Netherlands and Denmark were, respectively, $\boldsymbol{x}/n_x = (0.01, 0.04, 0.19, 0.54, 0.22)$ with a sample size $n_x = 13328$ and $\boldsymbol{y}/n_y = (0.03, 0.06, 0.21, 0.45, 0.25)$ with a sample size $n_y = 5906$. The two samples are ordered: sharing 'good health' as median category and with Denmark's distribution being a MPS of the Netherlands'. As is typical of distributions of self-assessed health, both distributions exhibit some class imbalances, with near-zero probability mass associated with the bottom health categories, and with over 40% mass attached to the median category.

We investigate a Z-test and a likelihood ratio test of the null hypothesis that the Danish distribution is not a MPS of the Dutch distribution. Figure 6 shows that all the tests are able to perfectly distinguish this false hypothesis from the closest (true) null hypothesis (the power curves are vertical). Next, we turn our attention to comparing the four tests in terms

of the P-value curves in the figure. For all relevant nominal test sizes (0 to 10%), all four tests are correctly sized: the asymptotic Z test is undersized, while the actual size of the other three tests coincides with the nominal size. In the context of this application, pertaining to large samples associated with distributions of self-assessed health, it is not possible to infer whether the Z or LR test is preferable in terms of size. Furthermore, this conclusion remains unchanged when we either consider the asymptotic or bootstrap approximation of the related test statistics.

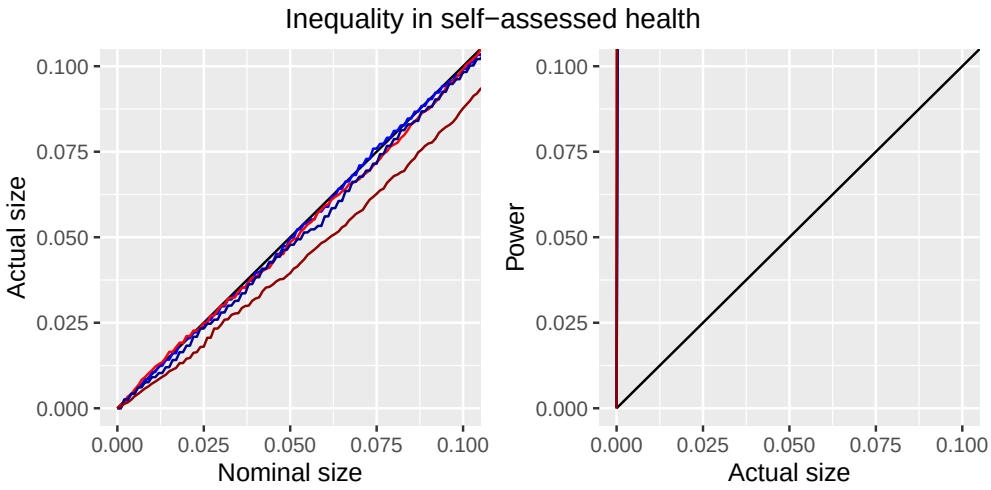

Figure 6

5.3 Inequality in sanitation ladders in Pakistan

Improvements in access to toilet facilities is a measure of living standards related to sanitation in developing countries (Seth and Yalonetzky, 2020). The 2017-8 Demographic and Health Survey of Pakistan collects data on different forms of sanitation facilities which can be grouped into a four-level sanitation ladder following the guidelines of the Joint Monitoring Program by the WHO and UNICEF.[7] The ensuing ordered categories are the following: (1) open defecation, (2) access to an unimproved toilet facility (buckets and latrine toilets

[7]See https://washdata.org/.

that do not flush), (3) shared improved toilet facilities (such as a shared toilet that flushes to piped sewer system) and finally (4) improved toilet facility that is not shared. The probability mass distributions pertaining to Islamabad (the capital city) and Baluchistan are, respectively, $\boldsymbol{x}/n_x = (0.003, 0.001, 0.060, 0.936)$ with a sample size $n_x = 1295$ and $\boldsymbol{y}/n_y = (0.135, 0.039, 0.142, 0.6849$ with a sample size $n_y = 1521$. The two distributions highlight important regional differences in attainment in sanitation, and furthermore in spread. That is, Baluchistan's sample is a MPS of Islamabad's.

There are three interesting properties these distributions exhibit in terms of statistical inference. Firstly, the median state in both distributions is $m = 4 = k$ (the top category). Were it not the case that our proposed tests jointly test that the distributions share an equal median, and are ordered according to the MPS criterion, the inferential exercise here would be equivalent to a test of first-order stochastic dominance (of Islamabad over Baluchistan). However, as we have earlier emphasised, the critical region of the tests in this paper is constrained by the union of the dominance and median boundaries, and in this sense, a simple test of first-order dominance would not provide a valid inferential tool in this application. The second interesting property in the data is that both distributions exhibit severe class imbalances, with the highest sanitation state (improved toilet facility that is not shared) being associated with probability mass in excess of 66%. Finally, the Islamabad distribution has a probability mass of 0.001 (one unique observation) in the second sanitary ladder state.

We investigate a Z-test and a LR test of the null hypothesis that the Baluchistan distribution is not a MPS of Islamabad. As in the EU health application, all the tests are able to perfectly distinguish this false hypothesis from the closest (true) null hypothesis, therefore the reproduce the power curves.[8] We therefore focus our attention on comparing the four tests in terms of the P-value curves, plots of which are provided in Figure 7. The dark red curve presents the P-value plot for the asymptotic approximation of the Z test, while the light red curve pertains to the asymptotic LR test. The dark blue curve is the P-value curve

[8]They are available upon request.

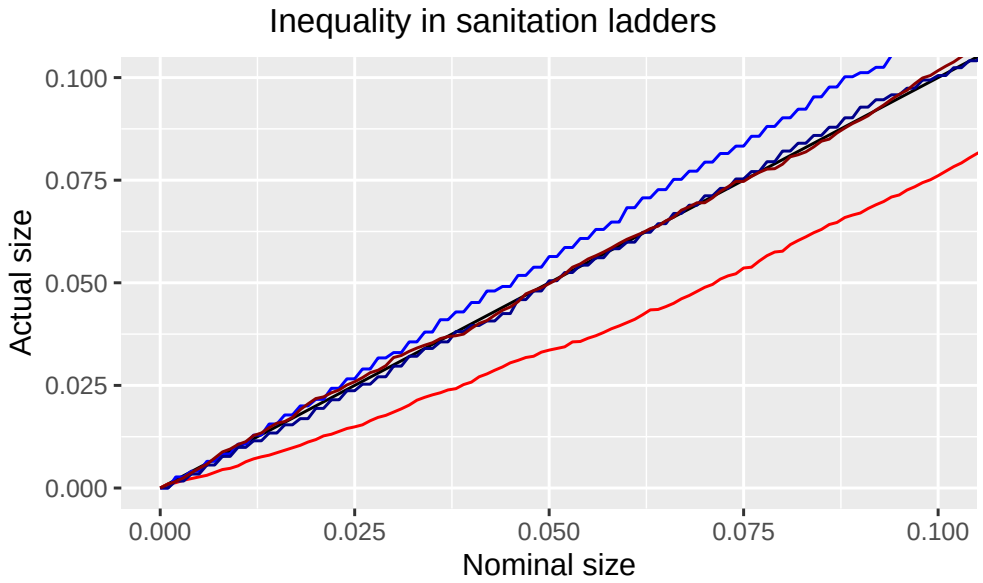

Figure 7

corresponding to the bootstrap approximation of Z statistic's sampling distribution, while finally the light blue curve refers to the bootstrap LR test.

We summarise our findings for all relevant nominal test sizes (0 to 10%) as follows: the asymptotic LR tests is correctly sized (the relevant P-value curve is below the 45 degree line), the asymptotic Z and the bootstrap Z tests have actual size equal to the nominal size (the P-value curves are on the 45 degree line). On the other hand, the bootstrap LR test is moderately oversized (the P-value curve lies above the 45 degree line). In this range, the bootstrap LR test would appear to be somewhat less oversized than the bootstrap Z test. Overall though, all four tests perform satisfactorily in the context of this application, with asymptotic inference being preferable over bootstrap inference. We attribute the better performance of asymptotic inference in the context of this application to the occurrence of severe class imbalances in both distributions.

6 Conclusion

The purpose of this paper was to introduce a family of tests for the hypothesis that an ordered multinomial distribution $\boldsymbol{G}$ is a MPS of $\boldsymbol{F}$. Using Monte Carlo simulations, we found that the choice between Z and LR test statistics does not have a large impact on the tests' properties, but the method used to approximate the sampling distribution of the statistics under the null does. In a wide range of data generating processes, bootstrap inference generally exhibited better size and power properties than asymptotic inference. We have further illustrated the proposed tests in three areas of inequality applications: happiness in the United States, self-assessed health in Europe and sanitation ladders in Pakistan.

The paper can be extended in several directions. For any quantile other than the median, tests of quantile preserving spreads à la Mendelson (1987) can be formulated by replacing the median boundary with an appropriate quantile boundary. Likewise, one may derive tests of hypotheses constructed from linear transformations of the vector of contrasts related to the median preserving spreads ordering; for instance, the bipolarization partial order of Chakravarty and Maharaj (2012). Finally, we mention the need to develop exact inference for tests of median-preserving spreads, yielding the P values of every conceivable sample, as a companion method to the bootstrap and asymptotic methods of inference introduced in this paper.

References

Abul Naga, R. and C. Stapenhurst (2015). Estimation of inequality indices of the cumulative distribution function. *Economics Letters 130*, 109–112.

Abul Naga, R., C. Stapenhurst, and G. Yalonetzky (2020). Asymptotic versus bootstrap inference for inequality indices of the cumulative distribution function. *Econometrics 8*(1).

Abul Naga, R. and T. Yalcin (2008). Inequality measurement for ordered response health data. *Journal of Health Economics 27*, 1614–25.

Allison, R. A. and J. E. Foster (2004). Measuring health inequality using qualitative data. *Journal of Health Economics 23*, 505–24.

Apouey, B. (2007). Measuring health polarisation with self-assessed health data. *Health Economics 16*, 875–94.

Balestra, C. and N. Ruiz (2015). Scale-invariant measurement of inequality and welfare in ordinal achievements: an application to subjective wellbeing and education in oecd countries. *Social Indicators Research 123*.

Berger, R. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics 24*, 295–300.

Chakravarty, S. and B. Maharaj (2012, May). Ethnic polarization orderings and indices. *Journal of Economic Interaction and Coordination 7*(1), 99–123.

Davidson, R. and J.-Y. Duclos (2013). Testing for restricted stochastic dominance. *Econometric Reviews 1*, ?–?

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

Davison, A. and D. Hinkley (1997). *Bootstrap methods and their applications*. Cambridge series on statistical and probabilistic mathematics ; 1. Cambridge: Cambridge University Press.

Dutta, I. and J. Foster (2013). Inequality of happiness in the u.s.: 1972–2010. *The Review of Income and Wealth 59*(3), 393–415.

Kobus, M. (2015). Polarisation measurement for ordinal data. *Journal of Economic Inequality 13*(2), 275–97.

Latham, K. and C. Peek (2013). Self-rated health and morbidity onset among late midlife u.s. adults. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences 68*(1), 107–116.

Lehmann, E. and J. Romano (2005). *Testing statistical hypotheses*. Springer.

MacKinnon, J. G. and R. Davidson (1996, February). The Size And Power Of Bootstrap Tests. Working Paper 932, Economics Department, Queen's University.

Madden, D. (2010). Ordinal and cardinal measures of health inequality: An empirical comparison. *Health Economics 19*(2), 243–250.

Mendelson, H. (1987). Quantile- preserving spread. *Journal of Economic Theory 42*, 334–51.

Mood, A., F. Graybill, and D. Boes (1974). *Introduction to the theory of statistics*. McGraw-Hill series in probability and statistics. McGraw-Hill.

Rothschild, M. and J. Stiglitz (1970). Increasing risk: I. a definition. *Journal of Economic Theory 2*(3), 225–43.

Seth, S. and G. Yalonetzky (2020). Assessing deprivation with an ordinal variable: Theory and application to sanitation deprivation in bangladesh. *World Bank Economic Review*, 1–19.

Silber, J. and G. Yalonetzky (2021). *Handbook of Labor, Human Resources and Population Economics*, Chapter Measuring welfare, inequality and poverty with ordinal variables: the univariate case. Springer.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science, New Series 103*(2684), 677–680.

Yalonetzky, G. (2013). Stochastic dominance with ordinal variables: Conditions and a test. *Econometric Reviews 32*(1), 126–63.

Appendices

A Mathematical appendix

Proof of lemma 1. First we show that every distribution in either the median or the dominance boundaries (or both) is in the boundary. Let $\boldsymbol{\theta} \in \bar{M} \cup \bar{D}$. Since $\boldsymbol{\theta} = (\boldsymbol{F}, \boldsymbol{G})$ is weakly ordered, $\boldsymbol{F}$ and $\boldsymbol{G}$ have at least one common median, m. The common median is unique unless $F_i = G_i = \frac{1}{2}$ for some $i \in [k]$. In this case we let $m = \min\{i \mid F_i = G_i = \frac{1}{2}\}$. Define a sequence $\{\boldsymbol{\theta}^{\mathbf{1}j}\}_{j\in\mathbb{N}}$ by $\boldsymbol{F}^{\mathbf{1}j} := \frac{1}{j}\mathcal{I}(j > m) + \left(1 - \frac{1}{j}\right)\boldsymbol{F}$ and $\boldsymbol{G}^{\mathbf{1}j} := \frac{1}{j}\left(\frac{1}{2}\mathcal{I}(j > m) + \frac{1}{4}\right) + \left(1 - \frac{1}{j}\right)\boldsymbol{G}$, for all $j < k$. This sequence converges to $\boldsymbol{\theta}$ and it is easy to see graphically that each $\boldsymbol{\theta}^{\mathbf{1}j}$ is strictly ordered. Hence every element of $\bar{M} \cup \bar{D}$ is the limit of a sequence in the complement of the null space. Now define a sequence $\{\boldsymbol{\theta}^{\mathbf{0}j}\}_{j\in\mathbb{N}}$ by $\boldsymbol{F}^{\mathbf{0}j} := \frac{1}{j}\frac{1}{2} + \left(1 - \frac{1}{j}\right)\boldsymbol{F}$ and $\boldsymbol{G}^{\mathbf{0}j} := \left(1 - \frac{1}{j}\right)\boldsymbol{G}$. This sequence also converges to $\boldsymbol{\theta}$ and it is easy to see graphically that each $\boldsymbol{\theta}^{\mathbf{1}j}$ is strictly ordered, so long as there exists either an $i \leq m$ such that $F_i^0 = G_i^0$ or else an $i \geq m$ such that $F_i^0 > \frac{1}{2} = G_i^0$. If neither or these conditions hold then, in order for $\boldsymbol{\theta}$ to be in $\bar{M} \cup \bar{D}$, there must exist either an $i > m$ such that $F_i^0 = G_i^0$ or else an $i \leq m$ such that $F_i^0 > \frac{1}{2} = G_i^0$. In this case we use the sequence defined by $\boldsymbol{G}^{\mathbf{0}j} := \frac{1}{j} + \left(1 - \frac{1}{j}\right)\boldsymbol{G}$.

Now we show that every distribution in the boundary is in either the median or the dominance boundaries (or both). If $\boldsymbol{\theta} \in (\bar{M} \cup \bar{D})^c$ is in neither the median nor dominance boundaries, then it must either be strictly ordered, or else unordered. Moreover, $\epsilon = \min\{|\frac{1}{2} - G_i|, |F_i - G_i| \mid i < k\}$ is strictly positive. If $\boldsymbol{\theta}$ it strictly ordered then it strictly satisfies all the inequalities in definition 1 by a margin of at least ϵ. Therefore any distribution $\boldsymbol{\theta'}$ within

a distance ϵ from $\boldsymbol{\theta}$ must also strictly satisfy these inequalities. Thus there cannot exist any sequence of unordered distributions that converges to $\boldsymbol{\theta}$. Similarly, if $\boldsymbol{\theta}$ is unordered then it strictly violates at least one of the inequalities in definition 1 by a margin of at least ϵ. Therefore any distribution $\boldsymbol{\theta'}$ within a distance ϵ from $\boldsymbol{\theta}$ must also violate the same inequality. Thus there cannot exist any sequence of ordered distributions that converges to $\boldsymbol{\theta}$. □

Proof of lemma 2 point 3. The are two ways the null hypothesis can be true: either one of the $k-1$ dominance conditions in [D1] or [D2] of definition 1 can fail, or the distributions do not share the same median and the conditions [M1] or [M2] in definition 1 fail. The easiest way for the former constraint to be satisfied is if $F_i = G_i$ for some $i \in [k-1]$ (which justifies a definition of strict MPS); the easiest way to satisfy the latter is if the median lies between two categories so that $G_{m-1} = \frac{1}{2}$ or $G_m = \frac{1}{2}$. Thus we can restate the problem:

$$\begin{aligned}\tilde{\boldsymbol{\theta}} = &\arg\max_{\boldsymbol{\theta}\in\Theta_0} \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{x}, \boldsymbol{y}] \\ &\text{s.t. } F_i = G_i \text{ for some } i \in [k-1] \\ &\text{or } G_{m-1} = \frac{1}{2} \text{ or } G_m = \frac{1}{2}.\end{aligned}$$

We now break the problem into two steps. We first find the $k+1$ distributions which maximise the likelihood, subject to each of these individual $k+1$ constraints, namely

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}} = \arg\max_{\boldsymbol{\theta}\in\Theta_0} \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{x}, \boldsymbol{y}] \text{ s.t. } F_i = G_i \qquad \forall i < k \tag{3}$$

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{k}} = \arg\max_{\theta\in\Theta_0} \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{x}, \boldsymbol{y}] \text{ s.t. } G_{m-1} = \frac{1}{2} \tag{4}$$

$$\tilde{\boldsymbol{\theta}}_{\boldsymbol{k+1}} = \arg\max_{\boldsymbol{\theta}\in\Theta_0} \mathbb{P}_{\boldsymbol{\theta}}[\boldsymbol{x}, \boldsymbol{y}] \text{ s.t. } G_m = \frac{1}{2} \tag{5}$$

The solution to the original problem is then given by the distribution among these which maximises the sample's likelihood function: $\tilde{\boldsymbol{\theta}} = \arg\max_{\tilde{\boldsymbol{\theta}}=\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}}} \mathbb{P}_{\tilde{\boldsymbol{\theta}}_{\boldsymbol{i}}}[\boldsymbol{x}, \boldsymbol{y}]$.

The solution to the problems in (3) are given in Davidson and Duclos (2013, p. 92) for each $i \in [k-1]$. The solution to (5) is found by noting that the independence of $\boldsymbol{f}$

and $\boldsymbol{g}$ implies that the solution to $\arg\max_{\theta\in\Theta_0}\mathbb{P}_\theta[\boldsymbol{x},\boldsymbol{y}]$ s.t. $G_m=\frac{1}{2}$ is given by the pair $\big(\arg\max_{\boldsymbol{f}}\mathbb{P}_{\boldsymbol{f}}[\boldsymbol{x}],\arg\max_{\boldsymbol{g}}\mathbb{P}_{\boldsymbol{g}}[\boldsymbol{y}]\text{ s.t. }G_m=\frac{1}{2}\big)$. The first of these terms is simply. $\boldsymbol{x}/n_x$. We solve for the second term by taking logarithms of the likelihood function $\mathbb{P}_{\boldsymbol{g}}[\boldsymbol{y}]$ (under the i.i.d. assumption) and by setting up the Lagrangian $\mathcal{L}(\boldsymbol{g},\lambda,\mu)=\sum_{i\in[k]}y_i\log g_i+\lambda(1-\sum_{i\in[k]}g_i)+\mu(\frac{1}{2}-\sum_{i\in[m]}g_i)$. The first order condition requires that

$$\frac{\partial\mathcal{L}}{\partial g_i}=\begin{cases}\frac{y_i}{\tilde{g}_i}-\lambda-\mu & \text{if } i\le m\\ \frac{y_i}{\tilde{g}_i}-\lambda & \text{if } i>m\end{cases}=0$$

which implies

$$y_i=\begin{cases}(\lambda+\mu)\tilde{g}_i & \text{if } i\le m\\ \lambda\tilde{g}_i & \text{if } i>m.\end{cases}$$

This in turn implies that $Y_m=\tilde{G}_m(\lambda+\mu)=\frac{1}{2}(\lambda+\mu)$ and $Y_k-Y_m=(1-\tilde{G}_m)\lambda=\frac{1}{2}\lambda$. Together, these give $(\lambda+\mu)=2Y_m$ and $\lambda=2(Y_k-Y_m)$, and thus

$$\tilde{g}_i=\begin{cases}\frac{y_i}{2Y_m} & \text{if } i\le m\\ \frac{y_i}{2(n_y-Y_m)} & \text{if } i>m.\end{cases}$$

The solution for (4) is found analogously. □

B Size and Power curve grid points

Table 2 lists the first coordinate of the DGP's on the median boundary and table 3 lists the DGPs on the dominance boundary.[9] Our choice of dominance boundary DGPs depends on the sample size; the choice of median DGPs is the same for all sample sizes.

Table 4 lists the first coordinate of the DGPs on the interior locus of the alternative space used to evaluate power. The concentrated distributions $\boldsymbol{f}$ are chosen to be the same for all

[9]Because $k=2$, it is sufficient to note only the mass in the first category; the full distributions can be recovered from f_1 and g_1 by $\boldsymbol{f}=(f_1,1-f_1)$ and $\boldsymbol{g}=(g_1,1-g_1)$.

Table 2: Median boundary DGPs used to construct the size curve in figure 3

	Null DGPs on median boundary (all sample sizes)									
f_1^{med}	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5
g_1^{med}	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5

Table 3: Dominance boundary DGPs used to construct the size curve in figure 3

	Null DGPs on dominance boundary (sample size dependent)										
$(n_x, n_y) = (10, 10), (100, 100), (1000, 1000)$											
$f_1^{\text{dom}} = g_1^{\text{dom}}$	0.46	0.43	0.39	0.36	0.32	0.28	0.25	0.21	0.16	0.10	0.05
$(n_x, n_y) = (10, 100), (100, 1000)$											
$f_1^{\text{dom}} = g_1^{\text{dom}}$	0.47	0.45	0.42	0.39	0.36	0.34	0.30	0.27	0.23	0.15	0.10	0.05
$(n_x, n_y) = (10, 1000)$											
$f_1^{\text{dom}} = g_1^{\text{dom}}$	0.48	0.45	0.43	0.40	0.37	0.35	0.32	0.29	0.25	0.15	0.10	0.05
$(n_x, n_y) = (100, 10), (1000, 100)$											
$f_1^{\text{dom}} = g_1^{\text{dom}}$	0.45		0.40	0.36	0.31	0.26	0.21	0.17	0.12	0.07	0.05
$(n_x, n_y) = (1000, 10)$											
$f_1^{\text{dom}} = g_1^{\text{dom}}$	0.45		0.40	0.35		0.30	0.25	0.20	0.15	0.10	0.05

sample sizes $n_x : n_y$; the precise choice of spread distribution $\boldsymbol{g}$ is then chosen to ensure that it is equidistant from the median and dominance boundaries.

Table 4: Alternative DGPs on the interior locus illustrated in figure 4.

Alternative DGPs (spread distributions depends on sample size)									
f_1^A	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
(nx,ny)=(10,10),(100,100),(1000,1000)									
g_1^A	0.277	0.312	0.342	0.368	0.392	0.415	0.437	0.458	0.479
(nx,ny)=(10,100),(100,1000)									
g_1^A	0.252	0.289	0.32	0.349	0.376	0.402	0.427	0.452	0.476
(nx,ny)=(10,100),(10,1000)									
g_1^A	0.244	0.282	0.314	0.344	0.371	0.398	0.424	0.450	0.475
(nx,ny)=(100,10)									
g_1^A	0.235	0.280	0.315	0.346	0.375	0.401	0.427	0.451	0.476
(nx,ny)=(1000,10)									
g_1^A	0.219	0.268	0.306	0.339	0.369	0.397	0.424	0.449	0.475