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

Ramses H. Abul Naga[*], Christopher Stapenhurst[†] and Gaston Yalonetzky[‡]

February 2023

Abstract

The median-preserving spread (MPS) ordering for ordinal variables (Allison and Foster, 2004) has become ubiquitous in the inequality literature. We devise statistical tests of the hypothesis that a distribution G is *not* an MPS of a distribution F. Rejecting this hypothesis enables the conclusion that G *is* more unequal than F according to the MPS criterion. Monte Carlo simulations and novel graphical techniques show that a simple, asymptotic Z test is sufficient for most applications. We illustrate our tests with two applications: happiness inequality in the US and self-assessed health in Europe.

Keywords: hypothesis testing; inequality measurement; median-preserving spread; ordinal data.
JEL codes: C12, D63, I14, I32.

[*] University of Malaga, Pan African Scientific Research Council.

[†] Quantitative Social and Management Sciences Research Centre, Faculty of Economic and Social Sciences, Budapest University of Technology and Economics, Budapest, Hungary. c.stapenhurst@edu.bme.hu. ORCiD 0000-0002-1376-6481

[‡] University of Leeds. ORCiD 0000-0003-2438-0223

1 Introduction

Health and wellbeing, educational qualifications, standards of sanitation, credit ratings, and perceived corruption are all examples of ordinal variables studied by social scientists. These variables take values which can be ordered, but not quantified. For example, the EUROSTAT Survey on Incomes and Living Conditions asks respondents to rate their happiness on a five-point scale from (1) very bad, to (5) very good. Movements up the scale correspond to improvement in happiness, but the size of the improvement in moving from, say, 'very bad' to 'bad' may not be the same as that in moving from 'good' to 'very good'. We know that some standard summary statistics, such as the mean and variance, cannot be meaningfully applied to such variables (Stevens, 1946). For instance, Allison and Foster (2004) show that the choice of scale can determine which of two ordered multinomial distributions has the higher variance.

A number of authors have responded to this problem by developing purpose-made inequality indices for ordinal variables which are not sensitive to arbitrary scale choices (see Silber and Yalonetzky, 2021, for a review). Many of these indices (Apouey, 2007; Abul Naga and Yalcin, 2008; Kobus and Milos, 2012; Chakravarty and Maharaj, 2015; Lazar and Silber, 2013; Reardon, 2009) respect the "median-preserving spread" partial ordering of Allison and Foster (2004).[1] A distribution G is a *median-preserving spread* (MPS) of a distribution F ('F and G are ordered') if F and G share a common median and if the probability mass of G lies further away from the median category than F's (i.e. G has 'thicker tails' than F). Therefore, if (and only if) G is an MPS of F, then G is deemed more unequal than F, according to *all* of these inequality indices (Kobus, 2015), in much the same way that two cardinal distributions being ordered by *mean*-preserving spread implies that one is in every sense riskier than the other (Rothschild and Stiglitz, 1970).

Accordingly, the MPS ordering has become popular in its own right for inequality comparisons in the empirical literature (e.g. Dutta and Foster (2013), Balestra and Ruiz (2015), Madden (2010)). However, these studies draw their conclusions by observing MPS-ordered samples of ordinal variables, without carrying out formal statistical inference. Thus, it is unclear whether the populations underlying these samples are really ordered, or the observed orderings are merely a

[1] The MPS partial ordering is itself a special case of Mendelson (1987)'s "quantile-preserving spread"; see section 6.

result of random sampling.

We help to improve on this uncertainty by devising a family of four statistical tests of the null hypothesis that G is *not* an MPS of F. We phrase the null hypothesis in this way because researchers are usually interested in the finding that one distribution *is* more unequal than another. By rejecting the hypothesis that G is *not* an MPS of F, the researcher is able to conclude that G *is* an MPS of F, and therefore that G is more unequal than F in a very robust sense.[2] Each of our four tests is characterised by a test statistic (Likelihood Ratio (LR) or Standardised (Z)), and a method of approximating the distribution of the test statistic under the null (asymptotic or bootstrap). Thus, our family includes both an easy-to-implement test (the asymptotic Z test), and a theoretically-most-attractive test (the bootstrap LR test, see e.g. Davidson and Duclos (2013)), as well as the two intermediate tests.

A natural question arises as to whether and when to prefer the theoretically-most-attractive test over the easy-to-implement test. We answer this by conducting a series of Monte Carlo experiments to compare our tests' performance. We find that all the tests are correctly sized in most cases, although, asymptotic inference can be somewhat oversized when sample sizes are very small or unbalanced. In most other cases, we find little practical difference between the tests in terms of size or power. This finding justifies using the easy-to-implement test in most applications.

Thirdly, we develop new graphical tools for illustrating the null hypothesis of no-ordering, and for comparing the size and power properties of our tests. The set of null distributions is at first hard to visualise because it is defined by the negation of a partial ordering of two ordered multinomial variables. But we show that it can be easily portrayed in the unit square. Moreover, the boundary of our null hypothesis is the union of two sets: one where the so-called median condition of the MPS fails, and one where the MPS' so-called dominance conditions fail. But by focusing on binomial distributions, we are able to give a representative graphical depiction of the size and power of our tests in and around the boundary.

The closest prior research to ours is Gunawan et al. (2018). They conduct Bayesian inference by assuming a uniform prior over all possible distributions and using the observed sample to calculate

[2] See Davidson and Duclos (2013) for a similar framing of the null and alternative hypotheses.

the respective posterior probabilities that a pair of distributions is or is not ordered. We, on the other hand, conduct frequentist inference which relies exclusively on the data to assess whether there is evidence that the distributions are ordered. Yalonetzky (2013) devises an asymptotic test for first-order stochastic dominance with ordinal variables, which we extend by both testing for equality of the medians, and by reversing the order of the dominance relation above and below the common median. 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 formally defines the MPS partial ordering, motivates our null hypothesis, and introduces a novel graphical representation of the parameter space. Section 3 develops our proposed statistical tests. Section 4 compares the size and power properties of the four tests with a series of Monte Carlo experiments. Section 5 demonstrates the broad usefulness of our tests in two applications covering happiness in the United States and self-assessed health in Europe. Finally section 6 offers some concluding remarks.

2 The Null Hypothesis

After presenting the 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 hypotheses.

2.1 *Notation*

Let $k \in \mathbb{N}$ denote the number of ordered categories and $\{1,\ldots,k\}$ denote the set of categories. We focus on a pair of samples (x,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 $x = (x_1,\ldots,x_k) \in \mathbb{N}_0^k$ and $\sum_{i=1}^k x_i = n_x$.[3] Since the states are ordered we can define the cumulative sums $X = (\sum_{i=1}^1 x_i,\ldots,\sum_{i=1}^k x_i = n_x)$ of x; with Y

[3] $\mathbb{N}_0 = \{0,1,2,...\}$.

defined analogously for y. We use $X_{[i]} := (X_1, \ldots, X_i)$ to denote the first i cumulative sums of X. The sample space is then $\mathscr{X}(k, n_x, n_y) = \{(x,y) \in \mathbb{N}_0^k \times \mathbb{N}_0^k \mid X_k = n_x \text{ and } Y_k = n_y\}$. The combined sample is denoted by $W = X + Y$ with combined sample size $n_x + n_y$.

Our ultimate goal is to perform inference on the pair of distributions (f, g) underlying the pair (x, y). Specifically, $x \sim f$ and $y \sim g$ where f_i (respectively g_i) denotes the probability that any particular observation from population f (respectively g) falls into category i. We denote their cumulative distribution functions (henceforth CDF) by $F = (\sum_{i=1}^{1} f_i, \ldots, \sum_{i=1}^{k} f_i = 1)$ and $G = (\sum_{i=1}^{1} g_i, \ldots, \sum_{i=1}^{k} g_i = 1)$ respectively. Let $L = (L_1, \ldots, L_k) := W/(n_x + n_y)$ be the sample-weighted average empirical distribution function (EDF).

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

$$\Theta := \{(f,g) \in [0,1]^k \times [0,1]^k \mid F_k = G_k = 1\}.^{4}$$

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

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

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

2.2 *Median Preserving Spreads*

We define the median of F to be a category $m \in \{1, \ldots, k\}$ 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.[5] 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 $\mathscr{X}$ which ranks distributions according to their spread. Here we define the analogous partial ordering for a pair of distributions:

[4] 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 fixed (e.g. by survey design).

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

Definition 1 (Median Preserving Spread)**.** Let $(f,g) \in \Theta$. We say that g is a strict *median preserving spread* (MPS) of f, or that f and g are ordered, and write $g \succ f$, 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 $i \in \{1,\ldots,m-1\}$

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

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

Note that these four conditions together imply that F has has the same median, m, as G.

2.3 *The 'No Ordering' Hypothesis*

We propose tests of the null hypothesis that g is not a strict MPS of f because we are mainly interested in situations where x and 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 set of null distributions is the subset of all parameter values such that g is not a strict MPS of f:

$$\Theta_0 = \{(f,g) \in \Theta \mid g \not\succ f\}.$$ [6]

A generic null distribution is denoted by $\theta_0 \in \Theta_0$.

[6] 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.

The set of alternative distributions is the complement of the set of null distributions, which is equivalent to the set of all ordered pairs:

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

A generic element of Θ_1 is denoted by θ_1. When the context is clear, we sometimes refer to Θ_0 as the 'null hypothesis', and to Θ_1 as the 'alternative hypothesis'.

We can graphically depict a two dimensional projection of the null and alternative hypotheses relative to the whole parameter space. Specifically, a pair of distributions $(f,g) \in \Theta$ can be written as a set of k pairs of cumulative frequencies (F_i, G_i). Figure 1 plots the pairs of coordinates of distributions $F = (3/24, 9/24, 17/24, 21/24, 1)$ and $G = (7/24, 11/24, 15/24, 17/24, 1)$ in the unit square. The median category of F (respectively G) is given by the state corresponding to the first coordinate to the right of the vertical (respectively 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 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 f first-order dominates g below the median and g first-order dominates 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 g is an MPS of f. In general, $(f,g) \in \Theta_1$ if and only if their coordinates are *all* contained in the triangles labelled Θ_1. Conversely $(f,g) \in \Theta_0$ 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 near here]

The boundary of the null hypothesis plays an important role in the proposed tests. Firstly, the Constrained Maximum Likelihood Estimator (CMLE) lies in the boundary of the null hypothesis. We will use the CMLE 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 hypothesis in order to provide an upper bound on the size of the tests of distributions in the interior of the null hypothesis. Thirdly, when we study the power of the tests, we will compare the rejection rates of distributions in the alternative hypothesis with rejection rates of the corresponding ‘closest null’ distributions, which lie in the boundary of Θ_0.

We characterise the boundary of the null hypothesis 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:

$$\bar{M} = \{(F,G) \in \Theta \mid F \succeq G \text{ and } G_i = \frac{1}{2} \text{ for some } i \in \{1,\ldots,k\}\}.$$

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:

$$\bar{D} = \{(F,G) \in \Theta \mid F \succeq G \text{ and } F_i = G_i \text{ for some } i \in \{1,\ldots,k\}\}.$$

The boundary can now be characterised:

Lemma 1. *The* boundary *of the null hypothesis 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. Examples of distributions on the median and dominance boundaries appear in figure 1.

3 Statistical Tests

A statistical test can be regarded as a function $p : \mathscr{X} \to [0,1]$ returning a p-value for every sample in the sample space $\mathscr{X}$. The p-value describes the probability of observing a sample 'as extreme' as (x,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 $100\alpha\%$ level.'

A test statistic is a function $\mathscr{S} : \mathscr{X} \to \mathbb{R}$ which formalises what it means for one sample to be 'as extreme' as another by associating each sample with a real number: sample (x',y') is more extreme than (x,y) if $\mathscr{S}(x',y') \geq \mathscr{S}(x,y)$. Thus we consider four tests of the form $T(x,y) = \mathbb{P}_{\Theta_0}[\mathscr{S}(x',y') \geq \mathscr{S}(x,y)]$. 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 (x,y) is the ratio of its unconstrained maximum likelihood function to the CMLE, its constrained counterpart under the null.

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

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

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

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

Lemma 2.

1. *The MLE of a sample (x,y) is given by*

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

2. *If x is not a strict MPS of y, then the CMLE is given by*

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

and $LR(x,y) = 0$*.*

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

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

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

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

The likelihood ratio statistic is then given by

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

Proof. See appendix A □

In lemma 2, the multiple cases arise from a Karush-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 tests of stochastic dominance for multivariate distributions of ordinal variables (e.g. Yalonetzky, 2013). Let $\hat{\sigma}_i^Y = \sqrt{[(Y_i/n_y)(1-Y_i/n_y)]/n_y}$ be the standard error of the sample cumulative frequency Y_i/n_y (with an analogous definition for the standard error of X_i/n_x) and $\hat{\sigma}_i^L = \sqrt{\frac{X_i(1-X_i/n_x)+Y_i(1-Y_i/n_y)}{(n_x+n_y)^2}}$ be the standard error of the pooled sample's cumulative frequency L_i. Then consider the Z statistic in definition 4:

Definition 4. The Z statistic for a multinomial sample (x,y) is given by:

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

where:[7]

$$Z_D^{<} := \min\left\{\frac{\frac{Y_i}{n_y}-\frac{X_i}{n_x}}{\hat{\sigma}_i^L(\frac{n_x+n_y}{\sqrt{n_x n_y}})}\ :\ i < m_y\right\},$$

$$Z_D^{\geq} := \min\left\{\frac{\frac{X_i}{n_x}-\frac{Y_i}{n_y}}{\hat{\sigma}_i^L(\frac{n_x+n_y}{\sqrt{n_x n_y}})} : m_y \leq i < k\right\}$$

and

$$Z_M := \min\left\{\frac{0.5-\frac{Y_{m_x-1}}{n_y}}{\hat{\sigma}^Y_{m_x-1}}, \frac{\frac{Y_{m_x}}{n_y}-0.5}{\hat{\sigma}^Y_{m_x}}\right\}.$$

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 Z_M is 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 standardised distances of sample cumulative frequencies Y-X below m_y; whereas $Z_D^{\geq}$ is the minimum among all the standardised distances of sample cumulative frequencies X-Y at and above m_y. Note the similarities with their (unstandardised) 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* $Y \succ X$.

By way of numerical illustration, consider the following ordered samples of happiness distributions in the US: $x = (154, 765, 450)$ for 2002 and $y = (218, 789, 599)$ for 1972 (Dutta and Foster, 2013, table 1).[8] In order to compute the statistics in definition 4, we calculate $X = (154, 919, 1369)$ (with $n_x = 1369$), $Y = (218, 1007, 1606)$ (with $n_y = 1606$), $L = (0.125, 0.6474, 1)$. From these statistics we can compute $X/n_x = (0.113, 0.671, 1)$, $Y/n_y = (0.136, 0.627, 1)$, $\hat{\sigma}_1^L = 0.0061$, $\hat{\sigma}_2^L = 0.0088$, $\hat{\sigma}_1^Y = 0.0085$, $\hat{\sigma}_2^Y = 0.0121$.

[7] $\hat{\sigma}_i^L\left(\frac{n_x+n_y}{\sqrt{n_x n_y}}\right)$ is the pooled-sample formula for the standard errors of $Y_i/n_y - X_i/n_x$ and $X_i/n_x - Y_i/n_y$ under the null hypothesis that $(Y_i/n_y = X_i/n_x)$.

[8] For more details on this application see section 5.

Then, because $m_y = m_x = 2$, we must compute the following Z-statistics: $Z_D^{\lessgtr} = \frac{0.136-0.113}{0.0061(\frac{1369+1606}{\sqrt{1369*1606}})} = 1.9121$; $Z_D^{\gtrless} = \frac{0.671-0.627}{0.0088(\frac{1369+1606}{\sqrt{1369*1606}})} = 2.5216$; $Z_M = \min\left\{\frac{0.5-0.136}{0.0085}, \frac{0.627-0.5}{0.0121}\right\} = 10.5263$. Finally, we get: $Z(x,y) = \min\left\{Z_D^{\lessgtr}, Z_D^{\gtrless}, Z_M\right\} = 1.9121$.

3.2 *Inference*

The ideal choice of null distribution (Lehmann and Romano, 2005) is that which maximises the probability of the upper contour set $\{x', y' \mid |\mathscr{S}(x', y')| \geq |\mathscr{S}(x,y)|\}$, namely

$$\theta_0 \in \underset{\theta \in \Theta_0}{\arg\max}\, \mathbb{P}_\theta[|\mathscr{S}(x', y')| \geq |\mathscr{S}(x,y)|],$$

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 hypothesis, and numerical solutions are computationally intensive. Instead, we follow the standard approach (e.g. Davidson and Duclos (2013)) of using the CMLE of the observed sample $\theta_0 = \tilde{\theta}$, characterised in lemma 2.

We approximate the probability $\mathbb{P}_{\theta_0}[|\mathscr{S}(x', y')| \geq |\mathscr{S}(x,y)|]$ 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 characterised in lemma 1, so that $(x,y) \sim \theta_0 \in \partial\Theta_0$, then:*

1. *$LR(x,y) \xrightarrow{d} \chi^2(1)$, and*

2. *$Z(x,y) \xrightarrow{d} \mathscr{N}(0,1)$.*

Hence if θ_0 lies in the boundary of the null hypothesis, point 1 of the theorem states that the LR statistic 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).[9] In the

[9] See lemma 2 in appendix A.

case where θ_0 lies in the interior of the null hypothesis, 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 θ_0 lies in $\partial\Theta_0$, but otherwise is bounded by $\mathscr{N}(0,1)$. We suggest in practice to approximate the p-value of a sample (x,y) by $1-\Phi[\mathscr{S}(x,y)]$, where Φ denotes the CDF of $\mathscr{N}(0,1)$ 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 set of B samples $\{(x^i,y^i)\}_{i\in\{1,\dots,B\}}$, each independently drawn from (x,y). The p-value of a sample (x,y) is then approximated by $\#\{(x^i,y^i) \mid \mathscr{S}(x^i,y^i) \geq \mathscr{S}(x,y), i\in\{1,\dots,B\}\}/B$.

Proposition 1 (Bootstrap p-values under the null; Davison and Hinkley (1997))**.** *Suppose the true distribution pair lies in the set of null distributions, so that $(x,y)\sim\theta_0\in\Theta_0$ as well as $(x^i,y^i)\sim\theta_0\in\Theta_0$ for all $i\in\{1,\dots,B\}$ where $B\in\mathbb{N}_0$, then the bootstrap p-values for a given statistic $\mathscr{S}(x,y)$ are:*

$$T_{B\mathscr{S}}(x,y) = \#\{(x^i,y^i) \mid |\mathscr{S}(x^i,y^i)| \geq |\mathscr{S}(x,y)|, i\in\{1,\dots,B\}\}/B.$$

Combining the two test statistics with these two methods of approximation gives a family of four tests and respective p-values:[10]

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

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

[10] Code implementing all four tests in R is available at https://cstapenhurst.academic.ws/projects/1551.

4 Size and Power

In this section we introduce novel graphical tools, namely the *size-boundary curve* for the study of test size, and the *power-locus curve* for the study of test power. We adopt the standard practice of using Monte Carlo experiments to construct the empirical distribution of p-values produced by the tests.[11] Specifically, we draw $M = 100,000$ independent samples (x^i, y^i) from sets of judiciously chosen data generating processes (DGPs) $\theta \in \Theta$, and calculate all the p-values, $\{T(x^i, y^i)\}_{i \in \{1,\dots,M\}}$ for each test T. The rejection rate of a nominal size α test at θ is then estimated by $\#\{(x^i, y^i) \mid T(x^i, y^i) \leq \alpha\}/M$.

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 equivalent to the horizontal line intersecting the vertical axis at $(0, 0.5)$ and the dominance boundary is 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 hypothesis. We argue in section 4.3 that the behaviour of the tests in the $k = 2$ case is indicative of behaviour in higher dimensions.

4.1 *Size*

If a test T satisfies the inequality $\mathbb{P}_{\theta_0}[T(x,y) < \alpha] \leq \alpha$ for all null distributions $\theta \in \Theta_0$, then we say that it is *correctly sized* at level α; otherwise it is *oversized*. 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 hypothesis are difficult to capture with a small selection of null distributions. Instead, we build a comprehensive picture of behaviour in the boundary, by holding the nominal size constant at the 5% level and plotting the empirical size of our tests against a grid of different DGPs in the boundary of the $k = 2$ null hypothesis, for a range of sample sizes. Our

[11] Code replicating these experiments in R is available at https://cstapenhurst.academic.ws/projects/1551.

tests will have higher rejection rates on the boundary than anywhere else in the null hypothesis, so this procedure gives an upper bound on the actual size of the tests. Figure 2 illustrates the DGP's used for the cases $n_x = n_y = 10, 100, 1000$. Our interest in small sample sizes, and more specifically in small ratios of n_x to n_y is three fold: (a) to investigate the relative merits of the bootstrap versus asymptotic 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 near here]

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 $i = 1, 2, 3$.

The tests are correctly sized in most cases. The main exceptions arise when sample sizes are highly asymmetric (i.e. (nx,ny)=(10,1000) or (1000,10)) and when there are severe class imbalances (i.e. at the endpoints of the boundary). In such instances, bootstrap tests can reject as much as 6.5% of the time.

Other than this, the test results are rarely substantially different. The most pronounced difference occurs when n_x is very small (around 10). In this case, the size of the asymptotic LR test can be more than double its nominal size on the dominance boundary. Finally, we note that the asymptotic Z test can be slightly oversized when n_y is smaller than n_x and, in a few cases, the bootstrap tests are less oversized than their asymptotic counterparts.

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 hypotheses 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 set of null distributions is now equal to 3/4, whereas the proportion of neighbouring distributions that belong to the alternative hypothesis is now equal to 1/4. For this reason, the probability of a sample with an empirical distribution in the set of null distributions 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, known as the 'power' of the test. Davidson and MacKinnon (1998) propose to assess power by plotting 'size-adjusted' size-power curves. These curves are constructed by plotting the rejection rate for a distribution in the alternative hypothesis against the rejection rate for the closest corresponding null distribution.

For a given alternative distribution $\theta = (f,g) \in \Theta_1$ and sample size (n_x,n_y) specified for the power assessment, we define the closest null distribution by

$$\underset{(f',g')\in\Theta_0}{\arg\max} \sum_{i=1}^{k} [n_x f_i \log(f_i') + n_y g_i \log(g_i')].$$

This expression is derived from the formula for the CMLE of a sample[12] (x,y) in definition 3, but noting that the sample (x,y) is replaced by a 'pseudo sample' $(n_x f, n_y g)$. This pseudo sample has empirical distribution equal to the alternative distribution of interest and the same sample sizes specified for the power assessment. We add the 'pseudo' qualification because, unlike naturally drawn samples, $n_x f$ and $n_y g$ may not necessarily have integer values. This is not problematic because, even though the factorial terms required for calculating likelihoods ($\frac{n_x!}{\prod_i x_i!}$ and $\frac{n_y!}{\prod_i y_i!}$) are

[12] We thank an anonymous referee for this suggestion.

not defined for non-integers of x and y, we can factor them out of the formula as they do not depend on f' or g', in turn not affecting the value of the maximising arguments. Finally, we follow standard procedure in the literature (Kass and Voss, 1997) and opt for the log-likelihood functional form which renders the objective function conveniently concave and does not alter the value of the maximising argument.

In the case $k = 2$, there are only two candidates for the closest null distribution: the closest distribution in the median boundary, $\theta^M := (f, (\frac{1}{2}, \frac{1}{2}))$; and the closest distribution in the dominance boundary, $\theta^D := (\frac{n_x f + n_y g}{n_x + n_y}, \frac{n_x f + n_y g}{n_x + n_y})$. 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 $k = 5$.

As with size, we face a choice over which alternative distributions to study. Because the boundary separating the null and alternative hypotheses 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 $n_x = n_y$. 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 hypothesis 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 hypothesis 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.

[Figure 4 near here]

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 power against these DGPs. 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.

[Figure 5 near here.]

All the tests are able to perfectly distinguish some alternatives distributions from the set of null distributions whenever 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 *Three or More Categories*

A distribution (f, g) with $k > 2$ categories can be decomposed into $k-1$ two-category distributions (f^i, g^i) defined by $(F^i, G^i) = ((F_i, 1), (G_i, 1))$ for any $i \in \{1, \ldots, k-1\}$. The rejection rate of a test at (f,g) is therefore a function of the $k-1$ rejection rates at each of the (f^i,g^i). For example, intersection-union tests (Berger, 1982) reject the null that (f,g) are not ordered if and only if they reject all the $k-1$ hypotheses that each of the (f^i,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 hypothesis, if and only if we infer that the coordinate closest to the edge of the triangles is nonetheless inside. Additional Monte Carlo experiments with more than two categories (not reported) show that the tests do indeed exhibit the same properties as in the two-category case. Specifically, size and power are both low when there are few observations drawn from the more spread out distribution, and when both distributions are evenly divided (i.e. there exists a category i such that (f_i,g_i) is close to (1/2,1/2)). The study of DGP's arising from sample surveys characterised by more than two categories is taken up in section 5.

5 Empirical Illustrations

We consider two real-world inequality assessments: happiness in the United States and self-reported health in a set of European countries. In each of the two applications we undertake 499 bootstrap replications. For self-reported health, we present p-value curves constructed from 100,000 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. The authors 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 can be summarized as follows. Out of the 171 pairs of distributions that are ordered by the median preserving spreads criterion, 16 out of the 177 comparisons are significant at the 1% (less than one in ten), 40 comparisons are significant at the 5% level (less than one in four) and 57 are significant at the 10% level (less than one in three). Like Dutta and Foster (2013), we conclude with strong evidence that most years after 1980 were less unequal than the early 70s. Unlike them, we find little evidence that the 90s were less unequal than the early 2000s. In the light of these findings, we recommend that empirical investigators take the extra step of validating their findings using statistical inference.

[Table 1 near here.]

5.2 *Inequality in Self-assessed Health in Europe*

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 Europe. 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, $x/n_x = (0.01, 0.04, 0.19, 0.54, 0.22)$ with a sample size $n_x = 13328$ and $y/n_y = (0.03, 0.06, 0.21, 0.45, 0.25)$ 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 the actual size of all the four tests coincides with the nominal size for all relevant values (0 to 10%). Moreover, all the tests perfectly distinguish this false hypothesis from the closest (true) null hypothesis (the power curves are vertical). 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 near here.]

6 Conclusion

The purpose of this paper was to introduce a family of tests for the hypothesis that an ordered multinomial distribution G is a MPS of 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 two areas of inequality applications: happiness in the United States and self-assessed health in Europe.

The paper can be extended in several directions. MPS is a special case of Mendelson (1987)'s "quantile-preserving spread" partial ordering: a distribution G is a *quantile-preserving spread* (QPS) of a distribution F ('F and G are ordered') if there exists a category q such that the mass of G lies further away from the state q than does the mass of F (G has a thicker tail than F around q). If this is true for q equal to the lowest (respectively highest) category then G first order dominates F (respectively F first order dominates G), so QPS orders the distributions according to their relative locations. In intermediate cases where q is equal to neither the highest nor the lowest category then QPS is sensitive to both the location and variability of the distributions. Tests of quantile preserving spreads 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.

Acknowledgment

We would like to thank James MacKinnon, Ashraful Mahfuze, Andy Snell, seminar participants at Heriot-Watt and the 9th ECINEQ Meeting, and three anonymous referees and the editor for helpful comments and suggestions. Christopher Stapenhurst gratefully acknowledges financial support from the Economic and Social Research Council through an ESRC 1+3 scholarship.

Declaration of interest

The authors report there are no competing interests to declare.

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), 8.

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 well-being and education in oecd countries. *Social Indicators Research 123*(2), 479–500.

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.

Chakravarty, S. R. and B. Maharaj (2015). Generalized gini polarization indices for an ordinal dimension of human well-being. *International Journal of Economic Theory 11*(2), 231–246.

Davidson, R. and J.-Y. Duclos (2013). Testing for restricted stochastic dominance. *Econometric Reviews 32*(1), 84–125.

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.

Gunawan, D., W. E. Griffiths, and D. Chotikapanich (2018). Bayesian inference for health inequality and welfare using qualitative data. *Economics Letters 162*, 76–80.

Kass, R. and P. Voss (1997). *Geometrical Foundations of Asymptotic Inference*. John Wiley & Sons.

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

Kobus, M. and P. Milos (2012). Inequality decomposition by population subgroups for ordinal data. *Journal of Health Economics 31*, 15–21.

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.

Lazar, A. and J. Silber (2013). On the cardinal measurement of health inequality when only ordinal information is available on individual health status. *Health Economics 22*(1), 106–13.

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

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.

Reardon, S. (2009). Measures of ordinal segregation. In Y. Fluckiger, S. Reardon, and J. Silber (Eds.), *Occupational and Residential Segregation*, Volume 17 of *Research on Economic Inequality*.

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

Silber, J. and G. Yalonetzky (2021, July). Measuring welfare, inequality and poverty with ordinal variables. In K. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. 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 $\theta \in \bar{M} \cup \bar{D}$. Since $\theta = (F, G)$ is weakly ordered, F and G have at least one common median, m. Let I^m denote the CDF of the degenerate distribution with all its mass in the median state, i.e. $I^m_i := \begin{cases} 0 \text{ if } i < m \\ 1 \text{ o/w} \end{cases}$. Define a sequence $\{\theta^{1j}\}_{j \in \mathbb{N}}$ by $F^{1j} := \frac{1}{j} I^m + \left(1 - \frac{1}{j}\right) F$ and $G^{1j} := \frac{1}{j}\left(\frac{1}{2} I^m + \frac{1}{4}\right) + \left(1 - \frac{1}{j}\right) G$, for all $j < k$. This sequence converges to θ and each θ^{1j} is strictly ordered. Hence every element of $\bar{M} \cup \bar{D}$ is the limit of a sequence in the complement of the set of null distributions. Now define a sequence $\{\theta^{0j}\}_{j \in \mathbb{N}}$ by $F^{0j} := \frac{1}{j}\frac{1}{2} + \left(1 - \frac{1}{j}\right) F$ and $G^{0j} := \left(1 - \frac{1}{j}\right) G$. This sequence also converges[13] to θ and it is easy to see graphically that each θ^{1j} is in the set of null distributions, so long as there exists either an $i \leq m$ such that $F^0_i = G^0_i$ or else an $i \geq m$ such that $F^0_i > \frac{1}{2} = G^0_i$. If neither or these conditions hold then, in order for θ to be in $\bar{M} \cup \bar{D}$, there must exist either an $i > m$ such that $F^0_i = G^0_i$ or else an $i \leq m$ such that $F^0_i > \frac{1}{2} = G^0_i$. In this case, the sequence defined by $G^{0j} := \frac{1}{j} + \left(1 - \frac{1}{j}\right) G$ converges to θ and is in the set of null distributions. Thus every element of $\bar{M} \cup \bar{D}$ is the limit of both a sequence of null distributions and a sequence of non-null distributions, therefore it is in the boundary of the null hypothesis.

[13] With respect to any Euclidean metric.

Now we show that every distribution in the boundary is in either the median or the dominance boundaries (or both). Equivalently, we prove the contrapositive, namely if $\theta = (F,G) \in (\bar{M} \cup \bar{D})^c$ is in neither the median nor dominance boundaries, then it must either be strictly ordered, or else unordered, i.e. $\theta \in (\partial\Theta_0)^c$. Let $\varepsilon = \min\{|\frac{1}{2} - G_i|, |F_i - G_i| \mid i < k\}$. If θ is strictly ordered then it strictly satisfies all the inequalities in definition 1 by a margin of at least $\varepsilon > 0$. Therefore any distribution θ' within a distance ε from θ must also strictly satisfy these inequalities. Thus there cannot exist any sequence of unordered distributions that converges to θ, so θ cannot be in the boundary. Similarly, if θ 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 θ' within a distance ε from θ must also violate the same inequality. Thus there cannot exist any sequence of ordered distributions that converges to θ. □

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 $i \in \{1,\dots,k-1\}$ (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:

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

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

$$\tilde{\theta}_i = \arg\max_{\theta\in\Theta_0} \mathbb{P}_\theta[x,y] \text{ s.t. } F_i = G_i \qquad \text{for some } i < k \tag{2}$$

$$\tilde{\theta}_k = \arg\max_{\theta\in\Theta_0} \mathbb{P}_\theta[x,y] \text{ s.t. } G_{m-1} = \frac{1}{2} \tag{3}$$

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

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

The solution to the problems in (2) are given in Davidson and Duclos (2013, p. 92) for each $i \in \{1,\ldots,k-1\}$. The solution to (4) is found by noting that the independence of f and g implies that the solution to $\arg\max_{\theta\in\Theta_0} \mathbb{P}_\theta[x,y]$ s.t. $G_m = \frac{1}{2}$ is given by the pair $\big(\arg\max_f \mathbb{P}_f[x], \arg\max_g \mathbb{P}_g[y]$ s.t. $G_m = \frac{1}{2}\big)$. The first of these terms is simply x/n_x. We solve for the second term by taking logarithms of the likelihood function $\mathbb{P}_g[y]$ (under the i.i.d. assumption) and by setting up the Lagrangian $\mathscr{L}(g,\lambda,\mu) = \sum_{i\in\{1,\ldots,k\}} y_i \log g_i + \lambda(1 - \sum_{i\in\{1,\ldots,k\}} g_i) + \mu(\frac{1}{2} - \sum_{i\in\{1,\ldots,m\}} g_i)$. The first order condition requires that

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

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 equation (3) is found analogously. □

Tables

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	0.27		0.30																									
74																												
75	0.27																											
76	0.07	0.37	0.36																									
77	0.16	0.17	0.06					0.15			0.20																	
78	0.12	0.21	0.04			0.42		0.28			0.30			0.15														
80	0.28	0.21	0.11																									
82	0.04	0.19	0.19		0.30																							
83	0.10	0.48	0.40	0.15	0.44																						0.33	
84	0.19	0.21	0.13																									
85	0.00	0.01	0.00	0.46	0.15	0.04		0.03	0.16		0.03		0.17	0.14		0.3			0.31			0.42	0.17		0.40	0.31		
86	0.12	0.17	0.03			0.31		0.19			0.20			0.23														
87	0.41		0.25																									
88	0.25	0.34	0.12					0.43			0.42			0.22														
89	0.04	0.09	0.01		0.39	0.20		0.14	0.41		0.16		0.17	0.13														
90	0.21	0.28	0.09			0.48		0.38			0.42			0.19	0.26													
91	0.00	0.01	0.00	0.31	0.08	0.05		0.02	0.08		0.02			0.42					0.24			0.23	0.32	0.39	0.28	0.16		
93	0.01	0.01	0.01		0.18	0.08		0.02	0.17		0.03														0.49	0.33		
94	0.01	0.15	0.12	0.03	0.19			0.34	0.31	0.08	0.41														0.32		0.14	
96	0.00	0.04	0.03	0.20	0.03	0.31		0.13	0.06		0.13											0.18		0.17	0.10	0.39		
98	0.00	0.06	0.04	0.41	0.06	0.41		0.23	0.14		0.20														0.27			
00	0.01	0.02	0.00		0.26	0.06		0.04	0.24		0.04			0.47														
02	0.02	0.17	0.14	0.31	0.10			0.34	0.24		0.37														0.18			
04	0.05	0.27	0.26		0.17				0.33																			
06	0.01	0.03	0.02		0.15	0.27		0.10	0.12		0.10														0.38			
08	0.38																											
10																												

Figures

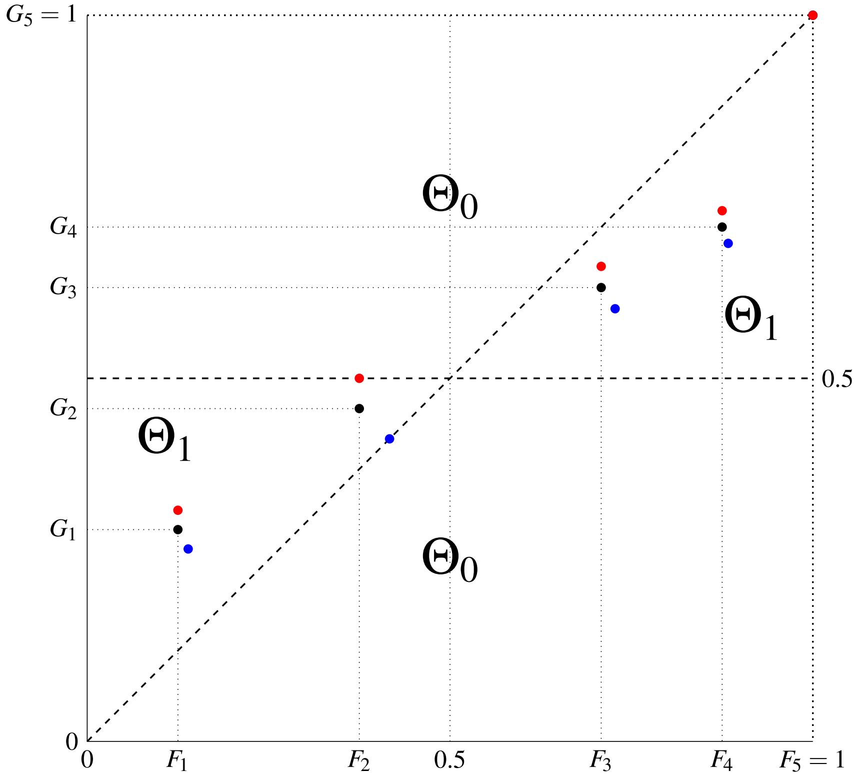

Figure 1: The parameter space, the set of null parameters and the set of alternative parameters projected onto the unit square. The red and blue dots illustrate, respectively, the closest distributions on the median and dominance boundaries to the distribution represented by black dots.

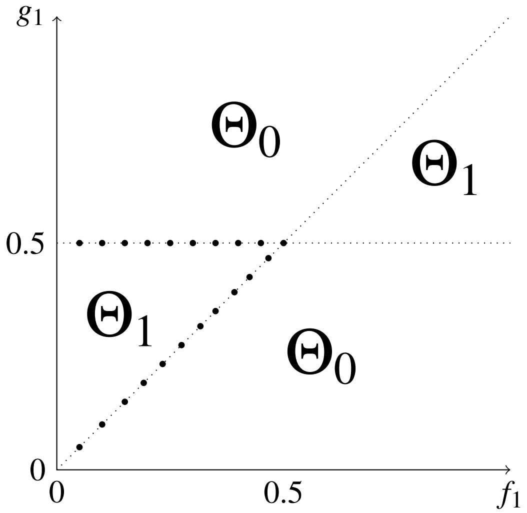

Figure 2: Boundary DGPs used to generate size curves for cases $k = 2$ and $n_x = n_y$.

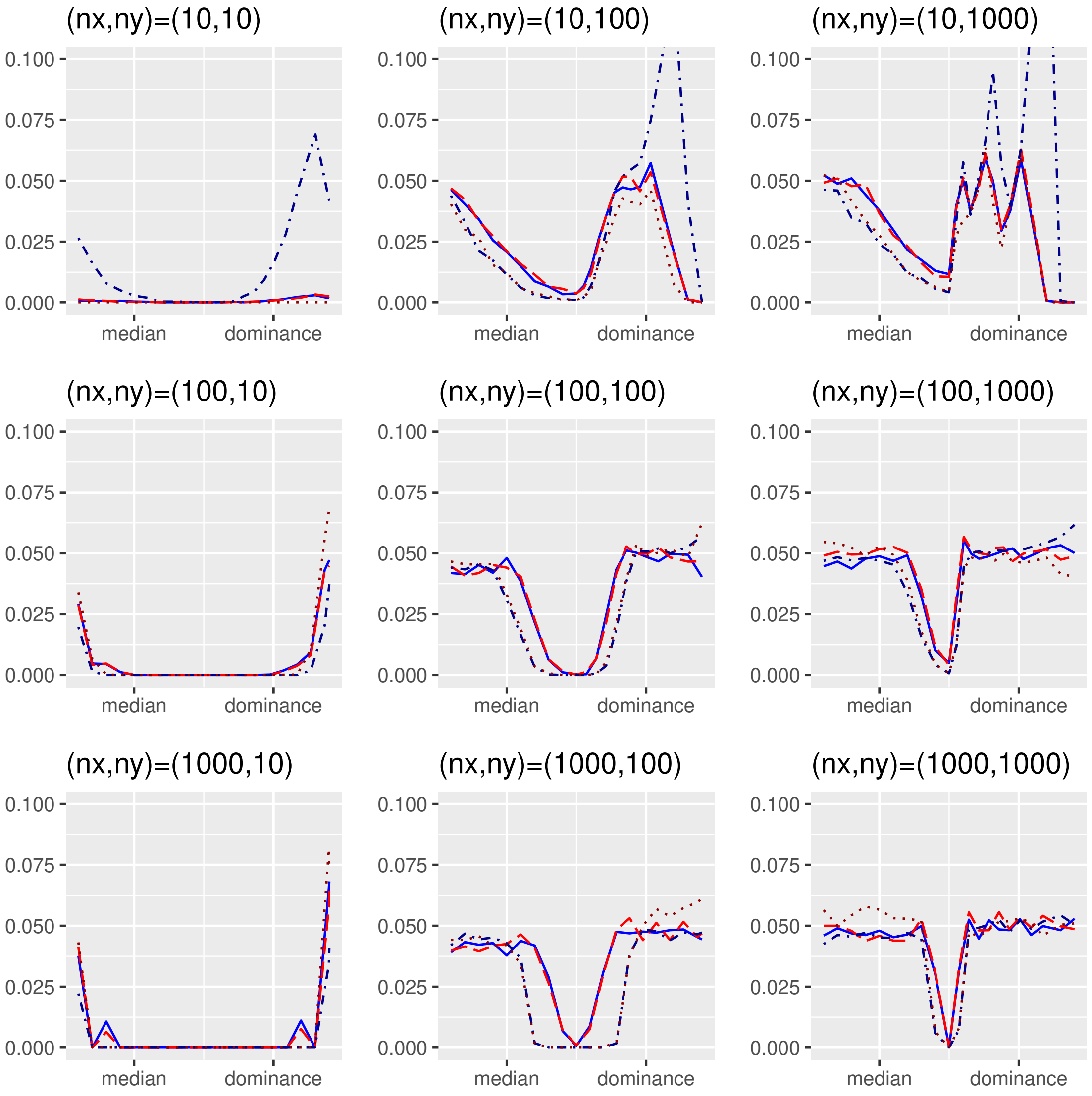

Figure 3: Size-boundary curves (for nominal 5% tests). Key: Solid/light blue — bootstrap LR; dashed/light red — bootstrap Z; dotdash/dark blue — asymptotic LR; dotted/dark red — asymptotic Z.

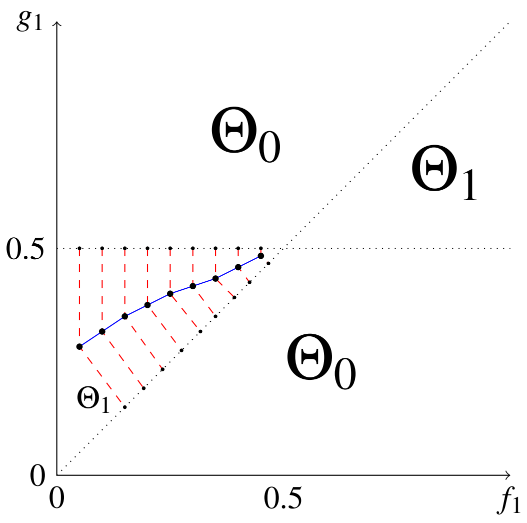

Figure 4: Alternative DGPs used to generate power curves for cases $k = 2$ and $n_x = n_y$.

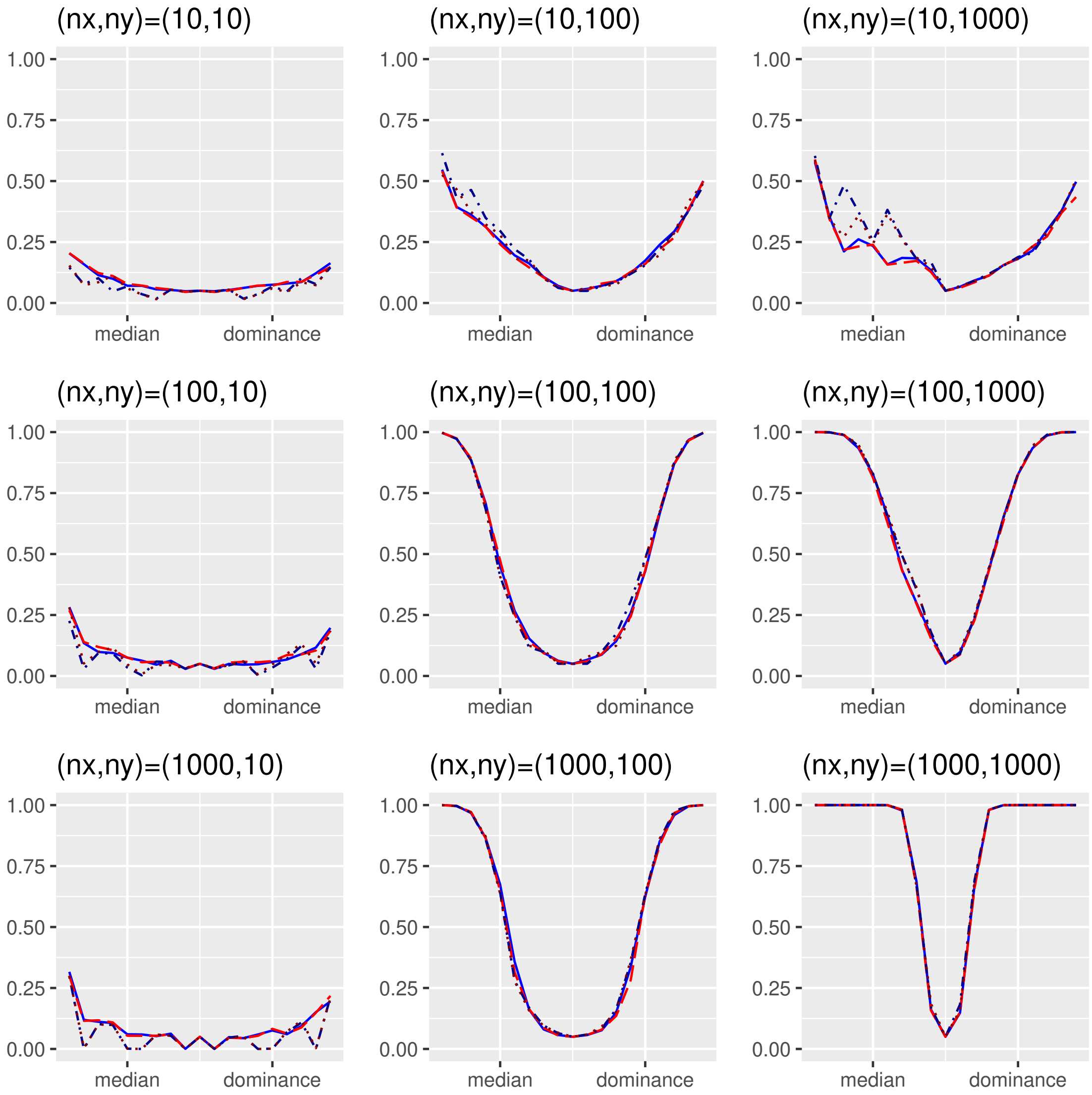

Figure 5: Power-locus curves (for nominal 5% tests). Key: Solid/light blue — bootstrap LR; dashed/light red — bootstrap Z; dotdash/dark blue — asymptotic LR; dotted/dark red — asymptotic Z.

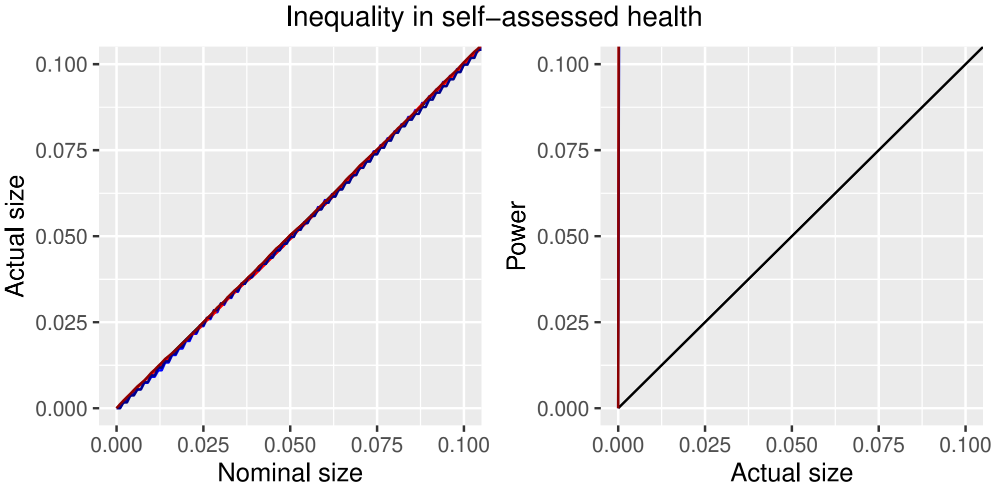

Figure 6: Left: Size curves for the distribution

$(F,G) = ((0.02, 0.06, 0.24, 0.78, 1.00), (0.02, 0.08, 0.29, 0.75, 1.00))$. Right: Power curves for the distribution

$(F,G) = ((0.01, 0.05, 0.24, 0.78, 1.00), (0.03, 0.09, 0.30, 0.75, 1.00))$.

Figure Captions

Figure 1: The parameter space, the set of null parameters and the set of alternative parameters 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.

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

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.

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

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.

Figure 6: Left: Size curves for the distribution

$(F,G) = ((0.02, 0.06, 0.24, 0.78, 1.00), (0.02, 0.08, 0.29, 0.75, 1.00))$. Right: Power curves for the

distribution

$(F,G) = ((0.01, 0.05, 0.24, 0.78, 1.00), (0.03, 0.09, 0.30, 0.75, 1.00)).$