Testing for restricted stochastic dominance

Russell Davidson, Jean-Yves Duclos

GREQAM

Groupement de Recherche en Economie
Quantitative d'Aix-Marseille - UMR-CNRS 6579
Ecole des Hautes Etudes en Sciences Sociales
Universités d'Aix-Marseille II et III

Document de Travail
n°2009-39

TESTING FOR RESTRICTED STOCHASTIC DOMINANCE

Russell DAVIDSON
Jean-Yves DUCLOS

July 2009

Testing for Restricted Stochastic Dominance

by

Russell Davidson

Department of Economics and CIREQ
McGill University
Montréal, Qubec, Canada
H3A 2T7

GREQAM
Centre de la Vieille Charité
2 Rue de la Charité
13236 Marseille cedex 02, France

email: **russell.davidson@mcgill.ca**

and

Jean-Yves Duclos

Department of Economics and CIRPÉE
Université Laval
Ste-Foy, Québec, Canada
G1K 7P4
email: **jyves@ecn.ulaval.ca**

Abstract

Asymptotic and bootstrap tests are studied for testing whether there is a relation of stochastic dominance between two distributions. These tests have a null hypothesis of nondominance, with the advantage that, if this null is rejected, then all that is left is dominance. This also leads us to define and focus on *restricted* stochastic dominance, the only empirically useful form of dominance relation that we can seek to infer in many settings. One testing procedure that we consider is based on an empirical likelihood ratio. The computations necessary for obtaining a test statistic also provide estimates of the distributions under study that satisfy the null hypothesis, on the frontier between dominance and nondominance. These estimates can be used to perform dominance tests that can turn out to provide much improved reliability of inference compared with the asymptotic tests so far proposed in the literature.

Keywords: Stochastic dominance, empirical likelihood, bootstrap test

JEL codes: C100, C120, C150, I320

This research was supported by the Canada Research Chair program (Chair in Economics, McGill University) and by grants from the Social Sciences and Humanities Research Council of Canada, the Fonds Québécois de Recherche sur la Société et la Culture, and the PEP Network of the International Development Research Centre. We are grateful to Abdelkrim Araar and Marie-Hélène Godbout for their research assistance and to Olga Cantó, Bernard Salanié, Bram Thuysbaert and two anonymous referees for helpful comments.

July 2008

1. Introduction

Consider two probability distributions, A and B, characterised by cumulative distribution functions (CDFs) F_A and F_B. In practical applications, these distributions might be distributions of income, before or after tax, wealth, or of returns on financial assets. Distribution B is said to dominate distribution A stochastically at first order if, for all z in the union of the supports of the two distributions, $F_A(z) \geq F_B(z)$. If B dominates A, then it is well known that expected utility and social welfare are greater in B than in A for all utility and social welfare functions that are symmetric and monotonically increasing in returns or in incomes, and that all poverty indices that are symmetric and monotonically decreasing in incomes are smaller in B than in A. These are powerful orderings of the two distributions[1] and it is therefore not surprising that a considerable empirical literature has sought to test for stochastic dominance at first and higher orders in recent decades.

Testing for dominance, however, requires leaping over a number of hurdles. First, there is the possibility that population dominance curves may cross, while the sample ones do not. Second, the sample curves may be too close to allow a reliable ranking of the population curves. Third, there may be too little sample information from the tails of the distributions to be able to distinguish dominance curves statistically over their entire theoretical domain. Fourth, testing for dominance typically involves distinguishing curves over an interval of an infinity of points, and therefore should also involve testing differences in curves over an infinity of points. Finally, dominance tests are always performed with finite samples, and this may give rise to concerns when the properties of the procedures that are used are known only asymptotically.

Dominance and nondominance

Until now, the most common approach to test whether there is stochastic dominance, on the basis of samples drawn from the two populations A and B, is to posit a null hypothesis of dominance, and then to study test statistics that may or may not lead to rejection of this hypothesis[2]. This is arguably a matter of convention and convenience: convention in the sense that it follows the usual practice of making the theory of interest the null and seeking evidence contrary to it, and convenience in that the null is then relatively easy to formulate.

Rejection of a null of dominance can, however, sometimes be viewed as an inconclusive outcome since it fails to rank the two populations. Further, in the absence of information on the power of the tests, non-rejection of dominance does not enable

[1] See Levy (1992) for a review of the breadth of these orderings, and Hadar and Russell (1969) and Hanoch and Levy (1969) for early developments.

[2] See, for instance, Beach and Richmond (1985), McFadden (1989), Klecan, McFadden and McFadden (1991), Bishop, Formby and Thistle (1992), Anderson (1996), Davidson and Duclos (2000), Barrett and Donald (2003), Linton, Maasoumi and Whang (2005), and Maasoumi and Heshmati (2005).

one to accept dominance, which is nevertheless often the outcome of interest. Hence, under this first approach, stochastic dominance merely remains either contradicted or uncontradicted, but cannot be established.

From a logical point of view, it may thus seem desirable in some settings to posit instead a null of *non*dominance. If we succeed in rejecting this null, we may indeed then legitimately infer the only other possibility, namely dominance. This is the approach that we develop in this paper.

In order to clarify the above, it may be useful to consider a very simple case with two distributions A and B with the same support, consisting of three points, $y_1 < y_2 < y_3$. Since $F_A(y_3) = F_B(y_3) = 1$, inference on stochastic dominance can be based on just two quantities, $\hat{d}_i \equiv \hat{F}_A(y_i) - \hat{F}_B(y_i)$, for $i = 1, 2$. The hats indicate estimates of the CDFs at the two points. Distribution B dominates distribution A if, in the population, $d_i \geq 0$.

Figure 1 shows a two-dimensional plot of $\hat{d}_1$ and $\hat{d}_2$. The first quadrant corresponds to dominance of A by B in the sample. In order to reject a hypothesis of dominance, therefore, the observed $\hat{d}_1$ and $\hat{d}_2$ must lie significantly far away from the first quadrant, for example, in the area marked as "B does not dominate A" separated from the first quadrant by an L-shaped band. This is essentially the procedure followed by the first approach described above, which is based on testing a null of dominance.

For a rejection of nondominance, on the other hand, the observed sample point must lie "far enough" inside the first quadrant that it is significantly removed from the area of nondominance, as in the area marked "B dominates A". One of this paper's primary objectives is to assess what is "far enough". The zone between the rejection regions for the two possible null hypotheses of dominance and nondominance corresponds to situations in which neither hypothesis can be rejected. We see that this happens when one of the $\hat{d}_i$ is close to zero and the other is positive. Note also from Figure 1 that inferring dominance by rejecting the hypothesis of nondominance is more demanding than failing to reject the hypothesis of dominance, since, for dominance, both statistics must have the same sign and be statistically significant.

The two approaches described above can thus be seen as complementary. Positing a null of dominance cannot be used to infer dominance; it can however serve to infer nondominance. Positing a null of nondominance cannot serve to infer nondominance; it can however lead to inferring dominance.

Objectives of the paper

In this paper, we pursue the approach of testing the null of nondominance. We find that it leads to testing procedures that are actually simpler to implement than conventional procedures in which the null is dominance. A simple procedure for testing nondominance was proposed originally by Kaur, Prakasa Rao, and Singh (1994) (henceforth KPS) for continuous distributions A and B, and a similar test was proposed in an unpublished paper by Howes (1993a) for discrete distributions.

Here, we develop an alternative procedure, based on an empirical likelihood ratio statistic. It turns out that this statistic is always numerically very similar to the KPS statistic in all the cases we consider. However, the empirical likelihood approach produces as a by-product a set of probabilities that can be interpreted as estimates of the population probabilities under the assumption of nondominance.

These probabilities make it possible to set up a bootstrap data-generating process (DGP) which lies on the frontier of the null hypothesis of nondominance. We show that, on this frontier, both the KPS and the empirical likelihood statistics are asymptotically pivotal, by which we mean that they have the same asymptotic distribution for all configurations of the population distributions that lie on the frontier. A major finding of this paper is that bootstrap tests that make use of the bootstrap DGP we define can yield more satisfactory inference than tests based on the asymptotic distributions of the statistics.

The paper also shows that it is not possible with continuous distributions to reject nondominance in favour of dominance over the entire supports of the distributions. Accepting dominance is empirically sensible only over *restricted* ranges of incomes or returns. Indeed, this is implicit in the KPS paper, where the test statistic is computed only over a restricted range. Here we make this point explicit, and explore its consequences. This necessitates a recasting of the usual theoretical links between stochastic dominance relationships and orderings in terms of poverty, social welfare and expected utility. It also highlights better why a failure to reject the usual null hypothesis of unrestricted dominance *cannot* be interpreted as an acceptance of unrestricted dominance, since unrestricted dominance can never be inferred from continuous data.

In Section 2, we investigate the use of empirical likelihood methods for estimation of distributions under the constraint that they lie on the frontier of nondominance, and develop the empirical likelihood ratio statistic. The statistic is a minimum over all the points of the realised samples of an empirical likelihood ratio that can be defined for all points z in the support of the two distributions. In Section 3 we recall the KPS statistic, which is defined as a minimum over z of a t statistic, and show that the two statistics are locally equivalent for all z at which $F_A(z) = F_B(z)$. Section 4 shows why it turns out to be impossible to reject the null of nondominance when the population distributions are continuous in their tails. In Section 5, the concept of restricted stochastic dominance is introduced, and justified by reference to ethical considerations that have been discussed in the literature on poverty. The relevance of the concept to analyses of financial and insurance data is also explored.

In Section 6, we discuss how to test the null hypothesis of restricted stochastic nondominance, developing procedures in which we are obliged to censor the tails of continuous distributions. In that section, we also show that, for configurations of nondominance that are *not* on the frontier where the CDFs touch at exactly one point, the rejection probabilities of tests based on either of our two statistics are no greater than they are for configurations on the frontier. This allows us to restrict attention to the frontier, knowing that, if we can control Type I error there by choice of an appropriate significance level, then the probability of Type I error in the interior of the null hypothesis is no greater than that on the frontier. We are then able to show that the statistics are asymptotically pivotal on the frontier.

Section 7 presents the results of a set of Monte Carlo experiments in which we investigate the rejection probabilities of both asymptotic and bootstrap tests, under the null and under some alternative setups in which there actually is dominance. We find that bootstrapping can lead to very considerable gains in the reliability of inference. Section 8 illustrates the use of the techniques with data drawn from the Luxembourg Income Study surveys, and finds that, even with relatively large sample sizes, asymptotic and bootstrap procedures can lead to different inferences. Possible extensions of the methods of the paper are described in Section 9, and conclusions and some related discussion are presented in Section 10.

2. Stochastic Dominance and Empirical Likelihood

Let two distributions, A and B, be characterised by their cumulative distribution functions (CDFs) F_A and F_B. Distribution B stochastically dominates A at first order if, for all x in the union U of the supports of the two distributions, $F_A(x) \geq F_B(x)$. In much theoretical writing, this definition also includes the condition that there should exist at least one x for which $F_A(x) > F_B(x)$ strictly. Since in this paper we are concerned with statistical issues, we ignore this distinction between weak and strong dominance since no statistical test can possibly distinguish between them. We therefore want to test the null hypothesis of non-dominance of distribution A by B

$$\max_{z \in U} \big(F_B(z) - F_A(z)\big) \geq 0, \tag{1}$$

against the alternative that B dominates A

$$\max_{z \in U} \big(F_B(z) - F_A(z)\big) < 0. \tag{2}$$

Suppose now that we have two samples, one each drawn from the distributions A and B. We assume for simplicity that the two samples are independent. Let N_A and N_B denote the sizes of the samples drawn from distributions A and B respectively. Let Y^A and Y^B denote respectively the sets of (distinct) realisations in samples A and B, and let Y be the union of Y^A and Y^B. If, for $K = A, B$, y_t^K is a point in Y^K, let the positive integer n_t^K be the number of realisations in sample K equal to y_t^K. This setup is general enough for us to be able to handle continuous distributions, for which all the $n_t^K = 1$ with probability 1, and discrete distributions, for which this is not the case. In particular, discrete distributions may arise from a discretisation of continuous distributions. The empirical distribution functions (EDFs) of the samples can then be defined as follows. For any $z \in U$, we have

$$\hat{F}_K(z) = \frac{1}{N_K} \sum_{t=1}^{N_K} \mathrm{I}(y_t^K \leq z),$$

where $\mathrm{I}(\cdot)$ is an indicator function, with value 1 if the condition is true, and 0 if not. If it is the case that $\hat{F}_A(y) \geq \hat{F}_B(y)$ for all $y \in Y$, we say that we have first-order stochastic dominance of A by B in the sample.

In order to conclude that B dominates A with a given degree of confidence, we require that we can reject the null hypothesis of nondominance of A by B with that degree of confidence. Such a rejection may be given by a variety of tests. In this section we develop an empirical likelihood ratio statistic on which a test of the null of nondominance can be based; see Owen (2001) for a survey of empirical likelihood methods. As should become clear, it is relatively straightforward to generalise the approach to second and higher orders of dominance, although solutions such as those obtained analytically here would then need to be obtained numerically.

For a given sample, the "parameters" of the empirical likelihood are the probabilities associated with each point in the sample. The empirical loglikelihood function (ELF) is then the sum of the logarithms of these probabilities. If as above we denote by n_t the multiplicity of a realisation y_t, the ELF is $\sum_{y_t \in Y} n_t \log p_t$, where Y is the set of all realisations, and the p_t are the "parameters". If there are no constraints, the ELF is maximised by solving the problem

$$\max_{p_t} \sum_{y_t \in Y} n_t \log p_t \quad \text{subject to} \quad \sum_{y_t \in Y} p_t = 1. \tag{3}$$

It is easy to see that the solution to this problem is $p_t = n_t/N$ for all t, N being the sample size, and that the maximised ELF is $-N \log N + \sum_t n_t \log n_t$, an expression which has a well-known entropy interpretation.

With two samples, A and B, using the notation given above, we see that the probabilities that solve the problem of the unconstrained maximisation of the total ELF are $p_t^K = n_t^K/N_K$ for $K = A, B$, and that the maximised ELF is

$$-N_A \log N_A - N_B \log N_B + \sum_{y_t^A \in Y^A} n_t^A \log n_t^A + \sum_{y_t^B \in Y^B} n_t^B \log n_t^B. \tag{4}$$

Notice that, in the continuous case, and in general whenever $n_t^K = 1$, the term $n_t^K \log n_t^K$ vanishes.

The null hypothesis we wish to consider is that B does not dominate A, that is, that there exists at least one z in the interior of U such that $F_A(z) \leq F_B(z)$. We need z to be in the interior of U because, at the lower and upper limits of the joint support U, we always have $F_A(z) = F_B(z)$, since both are either 0 or 1. In the samples, we exclude the greatest point in the set Y of realisations, for the same reason. We write Y° for the set Y without its upper extreme point. If there is a $y \in Y^\circ$ such that $\hat{F}_A(y) \leq \hat{F}_B(y)$, there is nondominance in the samples, and, in such cases, we plainly do not wish to reject the null of nondominance. This is clear in likelihood terms, since the unconstrained probability estimates satisfy the constraints of the null hypothesis, and so are also the constrained estimates.

If there is dominance in the samples, then the constrained estimates must be different from the unconstrained ones, and the empirical loglikelihood maximised under the constraints of the null is smaller than the unconstrained maximum value. In order to satisfy the null, we need in general only one z in the interior of U such that $F_A(z) = F_B(z)$. Thus, in order to maximise the ELF under the constraint of the null, we begin by computing the maximum where, for a given $z \in Y^\circ$, we impose the condition that $F_A(z) = F_B(z)$. We then choose for the constrained maximum that value of z which gives the greatest value of the constrained ELF.

For given z, the constraint we wish to impose can be written as

$$\sum_{y_t^A \in Y^A} p_t^A \mathrm{I}(y_t^A \le z) = \sum_{y_s^B \in Y^B} p_s^B \mathrm{I}(y_s^B \le z), \tag{5}$$

where the $\mathrm{I}(\cdot)$ are again indicator functions. If we denote by $F^K(\boldsymbol{p}^K; \cdot)$ the cumulative distribution function with points of support the y_t^K and corresponding probabilities the p_t^K, grouped in the vector $\boldsymbol{p}^K$, then it can be seen that condition (5) imposes the requirement that $F^A(\boldsymbol{p}^A, z) = F^B(\boldsymbol{p}^B, z)$.

The maximisation problem can be stated as follows:

$$\max_{p_t^A, p_s^B} \sum_{y_t^A \in Y^A} n_t^A \log p_t^A + \sum_{y_s^B \in Y^B} n_s^B \log p_s^B$$

$$\text{subject to} \sum_{y_t^A \in Y^A} p_t^A = 1, \sum_{y_s^B \in Y^B} p_s^B = 1, \sum_{y_t^A \in Y^A} p_t^A \mathrm{I}(y_t^A \le z) = \sum_{y_s^B \in Y^B} p_s^B \mathrm{I}(y_s^B \le z).$$

The Lagrangian for this constrained maximisation of the ELF is

$$\sum_t n_t^A \log p_t^A + \sum_s n_s^B \log p_s^B + \lambda_A \left(1 - \sum_t p_t^A\right) + \lambda_B \left(1 - \sum_s p_s^B\right)$$
$$-\mu \left(\sum_t p_t^A \mathrm{I}(y_t^A \le z) - \sum_s p_s^B \mathrm{I}(y_s^B \le z)\right),$$

with obvious notation for sums over all points in Y^A and Y^B, and where we define Lagrange multipliers λ_A, λ_B, and μ for the three constraints.

The first-order conditions are the constraints themselves and the relations

$$p_t^A = \begin{cases} n_t^A/(\lambda_A + \mu) & \text{if } y_t^A \le z \\ n_t^A/\lambda_A & \text{if } y_t^A > z, \end{cases} \quad \text{and} \quad p_s^B = \begin{cases} n_s^B/(\lambda_B - \mu) & \text{if } y_s^B \le z \\ n_s^B/\lambda_B & \text{if } y_s^B > z. \end{cases} \tag{6}$$

Using these relations, we can write the constraints as

$$\frac{N_A(z)}{\lambda_A + \mu} + \frac{M_A(z)}{\lambda_A} = 1 = \frac{N_B(z)}{\lambda_B - \mu} + \frac{M_B(z)}{\lambda_B}, \quad \text{and} \quad \frac{N_A(z)}{\lambda_A + \mu} = \frac{N_B(z)}{\lambda_B - \mu}, \tag{7}$$

where $N_A(z) = \sum_t n_t^A \mathrm{I}(y_t^A \le z)$ and $N_B(z) = \sum_s n_s^B \mathrm{I}(y_s^B \le z)$ are the numbers of points less than or equal to z in samples A and B respectively, and we define $M_A(z) = N_A - N_A(z)$ and $M_B(z) = N_B - N_B(z)$.

On multiplying the first two constraints in (7) by λ_A and λ_B respectively, we see that

$$\lambda_A = N_A - \frac{\mu N_A(z)}{\lambda_A + \mu} \quad \text{and} \quad \lambda_B = N_B + \frac{\mu N_B(z)}{\lambda_B - \mu}.$$

From the third constraint, it then follows that

$$\lambda_A + \lambda_B = N_A + N_B \equiv N. \tag{8}$$

The constraints (7) along with (8) imply that $M_A(z)/\lambda_A = M_B(z)/(N - \lambda_A)$, from which we see that

$$\lambda_A = \frac{N M_A(z)}{M_A(z) + M_B(z)}. \tag{9}$$

If we define $\nu = \lambda_A + \mu$, and note that $\lambda_B - \mu = N - \nu$, the third constraint can be solved for ν, with the result

$$\nu = \frac{N N_A(z)}{N_A(z) + N_B(z)}. \tag{10}$$

The relations (6), along with (9) and (10), allow us to write the probabilities p_t^A and p_s^B explicitly as

$$\begin{aligned} p_t^A &= \frac{n_t^A\big(N_A(z) + N_B(z)\big)}{N N_A(z)} \text{ if } y_t^A \le z \text{ and } \frac{n_t^A\big(M_A(z) + M_B(z)\big)}{N M_A(z)} \text{ if } y_t^A > z \text{ and} \\ p_s^B &= \frac{n_s^B\big(N_A(z) + N_B(z)\big)}{N N_B(z)} \text{ if } y_s^B \le z \text{ and } \frac{n_s^B\big(M_A(z) + M_B(z)\big)}{N M_B(z)} \text{ if } y_s^B > z. \end{aligned} \tag{11}$$

Thus the constrained maximisation of the empirical likelihood leads to four probabilities, assigned to observations according to whether they are from sample A or B, and to whether their value is less or greater than z.

We may use (11) in order to express the value of the ELF maximised under constraint as

$$\begin{aligned} &\sum_t n_t^A \log n_t^A + \sum_s n_s^B \log n_s^B \\ &-N_A(z)\log\nu - M_A(z)\log\lambda_A - N_B(z)\log(N-\nu) - M_B(z)\log(N-\lambda_A). \end{aligned} \tag{12}$$

Twice the difference between the unconstrained maximum (4), which can be written as

$$\sum_t n_t^A \log n_t^A + \sum_s n_s^B \log n_s^B - N_A \log N_A - N_B \log N_B,$$

and the constrained maximum (12) is an empirical likelihood ratio statistic.

Use of (10) and (9) for ν and λ_A shows that the statistic satisfies the relation

$$\begin{aligned}\tfrac{1}{2}\mathrm{LR(z)} = N\log N - N_A\log N_A - N_B\log N_B + N_A(z)\log N_A(z) + N_B(z)\log N_B(z)\\ +M_A(z)\log M_A(z) + M_B(z)\log M_B(z) - \big(N_A(z)+N_B(z)\big)\log\big(N_A(z)+N_B(z)\big)\\ -\big(M_A(z)+M_B(z)\big)\log\big(M_A(z)+M_B(z)\big).\end{aligned} \qquad (13)$$

We will see later how to use the statistic in order to test the hypothesis of nondominance.

The methods and results of this section can be extended to tests for second- and higher-order dominance. The only slight practical difficulty in doing so is that there do not seem to exist closed-form solutions, like (10) and (9), for the Lagrange multipliers needed to solve the problem of maximising the ELF subject to the relevant constraints. An efficient numerical solution can be found by applying Newton's method to a nonlinear system of two equations, with computing times that will of course be somewhat longer compared to those needed for first-order dominance; see Davidson (2007).

3. The Minimum t Statistic

In Kaur, Prakasa Rao, and Singh (1994), a test is proposed based on the minimum of the t statistic for the hypothesis that $F_A(z)-F_B(z)=0$, computed for each value of z in some closed interval contained in the interior of U. The minimum value is used as the test statistic for the null of nondominance against the alternative of dominance. The test can be interpreted as an intersection-union test. It is shown that the probability of rejection of the null when it is true is asymptotically bounded by the nominal level of a test based on the standard normal distribution. Howes (1993a) proposed a very similar intersection-union test, except that the t statistics are calculated only for the predetermined grid of points.

In this section, we show that the empirical likelihood ratio statistic (13) developed in the previous section, where the constraint is that $F_A(z)=F_B(z)$ for some z in the interior of U, is locally equivalent to the square of the t statistic with that constraint as its null, under the assumption that indeed $F_A(z)=F_B(z)$.

Since we have assumed that the two samples are independent, the variance of $\hat F_A(z)-\hat F_B(z)$ is just the sum of the variances of the two terms. The variance of $\hat F_K(z)$, $K=A,B$, is $F_K(z)\big(1-F_K(z)\big)/N_K$, where N_K is as usual the size of the sample from population K, and this variance can be estimated by replacing $F_K(z)$ by $\hat F_K(z)$. Thus the square of the t statistic is

$$t^2(z) = \frac{N_AN_B\big(\hat F_A(z)-\hat F_B(z)\big)^2}{N_B\hat F_A(z)\big(1-\hat F_A(z)\big)+N_A\hat F_B(z)\big(1-\hat F_B(z)\big)}. \qquad (14)$$

Suppose that $F_A(z) = F_B(z)$ and denote their common value by $F(z)$. Also define $\varDelta(z) \equiv \hat{F}_A(z) - \hat{F}_B(z)$. For the purposes of asymptotics, we consider the limit in which, as $N \to \infty$, N_A/N tends to a constant r, $0 < r < 1$. It follows that $\hat{F}_K(z) = F(z) + O_p(N^{-1/2})$ and that $\varDelta(z) = O_p(N^{-1/2})$ as $N \to \infty$.

The statistic (14) can therefore be expressed as the sum of its leading-order asymptotic term and a term that tends to 0 as $N \to \infty$:

$$t^2(z) = \frac{r(1-r)}{F(z)(1-F(z))} N\varDelta^2(z) + O_p(N^{-1/2}). \tag{15}$$

It can now be shown that the statistic $\text{LR}(z)$ given by (13) is also equal to the right-hand side of (15) under the same assumptions as those that led to (15). The algebra is rather messy, and so we state the result as a theorem.

Theorem 1

As the size N of the combined sample tends to infinity in such a way that $N_A/N \to r$, $0 < r < 1$, the difference between the statistic $\text{LR}(z)$ defined by (13) and the statistic $t^2(z)$ of (15) is of order $N^{-1/2}$ for any point z in the interior of U, the union of the supports of populations A and B, such that $F_A(z) = F_B(z)$.

Proof: In Appendix A. ■

Remarks:

- It is important to note that, for the result of the above theorem and for (15) to hold, the point z must be in the *interior* of U. As we will see in the next section, the behaviour of the statistics in the tails of the distributions is not adequately represented by the asymptotic analysis of this section.
- It is clear that both of the two statistics are invariant under monotonically increasing transformations of the measurement units, in the sense that if an income z is transformed into an income z' in a new system of measurement, then $t^2(z)$ in the old system is equal to $t^2(z')$ in the new, and similarly for $\text{LR}(z)$.

Corollary

Under local alternatives to the null hypothesis that $F_A(z) = F_B(z)$, where $F_A(z) - F_B(z)$ is of order $N^{-1/2}$ as $N \to \infty$, the asymptotic equivalence of $t^2(z)$ and $\text{LR}(z)$ continues to hold.

Proof:

Let $F_A(z) = F(z)$ and $F_B(z) = F(z) - N^{-1/2}\delta(z)$, where $\delta(z)$ is $O_p(1)$ as $N \to \infty$. Then $\varDelta(z)$ is still of order $N^{-1/2}$ and the limiting expression on the right-hand side of (15) is unchanged. The common asymptotic distribution of the two statistics now has a positive noncentrality parameter. ■

4. The Tails of the Distribution

Although the null of nondominance has the attractive property that, if it is rejected, all that is left is dominance, this property comes at a cost, which is that it is impossible to infer dominance over the full support of the distributions if these distributions are continuous in the tails. Moreover, and as we shall see in Section 6, the tests of nondominance that we consider can be used to delimit the range over which restricted dominance can be inferred.

The nondominance of distribution A by B can be expressed by the relation

$$\max_{z \in U} \bigl(F_B(z) - F_A(z)\bigr) \geq 0, \tag{16}$$

where U is as usual the joint support of the two distributions. But if z^- denotes the lower limit of U, we must have $F_B(z^-) - F_A(z^-) = 0$, whether or not the null is true. Thus the maximum over the whole of U is never less than 0. Rejecting (16) by a statistical test is therefore impossible. The maximum may well be significantly greater than 0, but it can never be significantly less, as would be required for a rejection of the null.

Of course, an actual test is carried out, not over all of U, but only at the elements of the set Y of points observed in one or other sample. Suppose that A is dominated by B in the sample. Then the smallest element of Y is the smallest observation, y_1^A, in the sample drawn from A. The squared t statistic for the hypothesis that $F_A(y_1^A) - F_B(y_1^A) = 0$ is then

$$t_1^2 \equiv \frac{N_A N_B (\hat{F}_A^1 - \hat{F}_B^1)^2}{N_B \hat{F}_A^1 (1 - \hat{F}_A^1) + N_A \hat{F}_B^1 (1 - \hat{F}_B^1)},$$

where $\hat{F}_K^1 = \hat{F}_K(y_1^A)$, $K = A, B$; recall (14). Now $\hat{F}_B^1 = 0$ and $\hat{F}_A^1 = 1/N_A$, so that

$$t_1^2 = \frac{N_A N_B / N_A^2}{(N_B/N_A)(1 - 1/N_A)} = \frac{N_A}{N_A - 1}.$$

The t statistic itself is thus approximately equal to $1 + 1/(2N_A)$. Since the minimum over Y of the t statistics is no greater than t_1, and since $1 + 1/(2N_A)$ is nowhere near the critical value of the standard normal distribution for any conventional significance level, it follows that rejection of the null of nondominance is impossible. A similar, more complicated, calculation can be performed for the test based on the empirical likelihood ratio, with the same conclusion.

If the data are discrete or censored in the tails, it is no longer impossible to reject the null if there is enough probability mass in the atoms at either end or over the censored areas of the distribution. If the distributions are continuous but are discretised or censored, then it becomes steadily more difficult to reject the null as the censoring becomes less severe, and in the limit once more impossible. The difficulty is clearly that, in the tails of continuous distributions, the amount of information conveyed by the sample tends to zero, and so it becomes impossible to discriminate among different hypotheses about what is going on there. Focussing on *restricted* stochastic dominance is then the only empirically sensible course to follow.

5. Restricted stochastic dominance and distributional rankings

An empirical focus on restricted dominance might seem to demand some motivation, however, since the theoretical normative literature is (almost) exclusively couched in terms of unrestricted dominance.

There nevertheless does exist in welfare economics and in finance a limited strand of the normative literature that is concerned with restricted dominance – see for instance Chen, Datt and Ravallion (1994), Bishop, Chow, and Formby (1991) and Mosler (2004). One reason for this concern is the suspicion formalised above that testing for unrestricted dominance is too statistically demanding, since it forces comparisons of dominance curves over areas where there is too little information (a good example is Howes 1993b). This insight is interestingly also present in Rawls's (1971) practical formulation of his famous "difference" principle (a principle that leads to the "maximin" rule of maximising the welfare of the most deprived), which Rawls defines over the most deprived *group* rather than the most deprived *individual*:

> *In any case, we are to aggregate to some degree over the expectations of the worst off, and the figure selected on which to base these computations is to a certain extent* ad hoc. *Yet we are entitled at some point to plead practical considerations in formulating the difference principle. Sooner or later the capacity of philosophical or other argument to make finer discriminations is bound to run out.* (Rawls 1971, p.98)

As we shall see below, a search for restricted dominance is indeed consistent with a restricted aggregation of the plight of the worst off.

A second motivation for restricted dominance is the normative view that unrestricted dominance does not impose sufficient limits on the ranges over which certain ethical principles must be obeyed. It is often argued, for instance, that the precise value of the living standards of those that are abjectly deprived should not be of concern: the number of such abjectly deprived people should be sufficient information for social welfare analysts. It does not matter for social evaluation purposes what the exact value of one's income is when it is clearly too low. Said differently, the distribution of living standards under some low threshold should not matter: everyone under that threshold should certainly be deemed to be in very difficult circumstances. This comes out strongly in Sen (1983)'s views on capabilities and the shame of being poor:

> *On the space of the capabilities themselves – the direct constituent of the standard of living – escape from poverty has an absolute requirement, to wit, avoidance of this type of shame. Not so much having equal shame as others, but just not being ashamed, absolutely.* (Sen 1983, p.161)

Bourguignon and Fields (1997) interpret this as

> *the idea that a minimum income is needed for an individual to perform 'normally' in a given environment and society. Below that income level some basic function of physical or social life cannot be fulfilled and the individual is somehow excluded from society, either in a physical sense (e.g. the long-run effects of an insufficient*

diet) or in a social sense (e.g. the ostracism against individuals not wearing the proper clothes, or having the proper consumption behavior). (Bourguignon and Fields 1987, p.1657)

Such normative views militate in favour of the use of *restricted* poverty indices, indices that give equal ethical weight to all those who are below a survival poverty line. The same views also suggest an analogous concept of restricted social welfare.

To see this more precisely, consider the case in which we are interested in whether there is more poverty in a distribution A than in a distribution B. Consider for expositional simplicity the case of additive poverty indices:

$$P(z) = \int \pi(y; z)\, \mathrm{d}F(y) \tag{17}$$

where z is a poverty line, y is income, $F(y)$ is the cumulative distribution function, and $\pi(y; z) \geq 0$ is the poverty contribution to total poverty of someone with income y, with $\pi(y; z) = 0$ whenever $y > z$. This definition is general enough to encompass many of the poverty indices that are used in the empirical literature. Also assume that $\pi(y; z)$ is non-increasing in y and let $Z = [z^-, z^+]$, with z^- and z^+ being respectively some lower and upper limits to the range of possible poverty lines. Then denote by $\Pi^1(Z)$ the class of "first-order" poverty indices that contains all of the indices $P(z)$, with $z \in Z$, whose function $\pi(y; z)$ satisfies the conditions

$$\pi(y; z) \quad \begin{cases} \text{equals 0 whenever } y > z, \\ \text{is non-increasing in } y, \text{ and} \\ \text{is constant for } y < z^-. \end{cases} \tag{18}$$

We are then interested in checking whether $\Delta P(z) \equiv P_A(z) - P_B(z) \geq 0$ for all of the poverty indices in $\Pi^1(Z)$. It can be shown that this can be done using the following definition of restricted first-order poverty dominance:

(Restricted first-order poverty dominance)

$$\Delta P(z) > 0 \text{ for all } P(z) \in \Pi^1(Z) \text{ iff } \Delta F(y) > 0 \text{ for all } y \in Z, \tag{19}$$

with $\Delta F(y) \equiv F_A(y) - F_B(y)$. Note that the $\Pi^1(Z)$ class includes discontinuous indices, such as some of those considered in Bourguignon and Fields (1997), as well as the headcount index itself, which would seem important given the popularity of that index in the poverty and in the policy literature. Traditional unrestricted poverty dominance is obtained with $Z = [0, z^+]$.[3] The indices that are members of $\Pi^1(Z)$ are insensitive to changes in incomes when these take place outside of Z. This eliminates concern with the precise living standards of the most deprived – for some, a possibly controversial ethical procedure, but unavoidable from a statistical and empirical point of view.

[3] See for instance Foster and Shorrocks (1988a).

A simple member of $\Pi^1(z)$ is obtained by using in (17) the specification $\pi(y; z) = 1$ for all $y \leq z$. The index so defined is then given by

$$\tilde{P}(z) = \max\bigl(F(z^-), F(z)\bigr) \text{ for } z < z^+.$$

This index, obviously closely related to the conventional headcount ratio, can be supported by a view that a poverty line cannot sensibly lie below z^-: anyone with z^- or less should necessarily be considered as being in equally abject deprivation. The popular FGT (see Foster, Greer and Thorbecke 1984) indices are defined (in their un-normalised form) for $\alpha \geq 0$ as

$$P(z; \alpha) = \int g(y; z)^\alpha \mathrm{d}F(y) \tag{20}$$

where the poverty gap $g(y; z) = \max(z - y, 0)$. If we redefine g so that $g(y, z) = g(z^-, z)$ for all $y < z^-$, the index becomes, for $z \geq z^-$,

$$\tilde{P}(z; \alpha) = (z - z^-)^\alpha F(z^-) + \int_{z^-}^{z} (z - y)^\alpha \, \mathrm{d}F(y).$$

and it now belongs to $\Pi^1(Z)$. It is the same as the traditional FGT index when all incomes below z^- are lowered to 0, again presumably because everyone with z^- or less is deemed to be in complete deprivation.

A setup for restricted social welfare dominance can proceed analogously. For example, welfare W can be defined using weakly increasing utilitarian functions as $W = \int u(y) \, \mathrm{d}F(y)$, where u can be strictly increasing only over some restricted range of income Z. Verifying whether $\Delta F(y) > 0$ for all $y \in Z$ is then the corresponding test for restricted first-order welfare dominance. Fixing $Z = [0, \infty[$ yields traditional first-order welfare dominance.[4]

These ideas can also be applied to the fields of finance and insurance. "Value at risk" (VaR), for instance, is often used by security houses and investment banks to measure the market risk of their asset portfolios[5]. VaR represents the maximum amount to be lost from an investment over a given holding period and with a given probability. A bank might expect, for instance, that, with a probability of 95%, the value of its portfolio will decrease by at most \$1 million during some period. Let F be the distribution of changes in the market value of a portfolio. Then $F^{-1}(p)$ is the VaR at a $1 - p$ confidence level.

A related indicator, "probability of ruin" (PoR), serves to capture the risk of insolvency or ruin when negative positions in some business lines cannot be offset by capital

[4] See for instance Foster and Shorrocks (1988b).

[5] See *inter alia* Holton (2003) and Jorion (2001).

transfers. In insurance, for example, PoR is the probability that claims exceed premiums by some time t. If F is the distribution of the difference between premiums and claims, then PoR is given by $F(0)$, and the probability of premiums not exceeding claims by at least z is $F(z)$. PoR is a key parameter in the theory of risk and insurance and has received considerable attention over the years — see for example Feller (1971), Beard, Pentikainen and Pesonen (1984), Gerber (1979) and Buhlmann (1970).

In the spirit of the poverty and welfare analysis sketched above, VaR and PoR dominance tests can be performed by comparing their values across two distributions of portfolio returns and business positions over *ranges* of p and z, respectively. That could help check whether these distributions can be "robustly" ordered, and could also help address regulatory concerns that VaR and PoR measures can be subject to manipulation by firms. If a distribution shows a lower VaR or a lower PoR over a wide range of p or z, then a firm's financial or insurance position can be shown to be better under that distribution for a broad class of evaluation measures.

As above, however, the immediate difficulty is that, in the tails of such distributions, the amount of available empirical information will almost always inevitably tend to zero as z and p approach their lower and upper bounds. It will therefore be generally impossible to rank two distributions of returns or net positions over their entire unrestricted support. Checking for *restricted* VaR and PoR dominance will then again emerge as the only empirically sensible course to follow.

6. Testing the hypothesis of restricted nondominance

Suppose that we wish to base inference on two random samples drawn from the populations A and B of interest. As before, we denote by y_t^A, $t = 1, \ldots, N_A$, and y_s^B, $s = 1, \ldots, N_B$, the observations in the samples. Empirical distribution functions $\hat{F}_A$ and $\hat{F}_B$, say, can be computed for the two samples. If $\hat{F}_B(y) \leq \hat{F}_A(y)$ for all y, then we say that there is dominance of A by B in the sample. If we are interested in a prespecified interval $[z^-, z^+]$, B dominates A in the sample over that interval if $\hat{F}_B(y) < \hat{F}_A(y)$ for all $y \in [z^-, z^+]$. In what follows, we assume that we have chosen an interval $[z^-, z^+]$ and that our null hypothesis is nondominance restricted to this interval. With continuous distributions, the interval is contained in the interior of the joint support of the distributions. With discrete distributions, it is not always necessary for the interval to be restricted, although it may be. Obviously, it is only when there is restricted dominance in the sample that there is any possible reason to reject the null of restricted nondominance. It is impossible, therefore, to reject restricted nondominance of A by B and also reject restricted nondominance of B by A with one and the same data set.

We have at our disposal two test statistics to test the null hypothesis that distribution B does not dominate distribution A, the two being locally equivalent in some circumstances. Let us redefine the set Y° to be the set of those y_t^A and y_s^B that lie in $[z^-, z^+]$. Then the minimum t statistic of which the square is given by (14) can be found by minimising $t(z)$ over $z \in Y^\circ$. There is no loss of generality in restricting the search for the maximising z to the elements of Y°, since the quantities $N_K(z)$ and $M_K(z)$ on which (12) depends are constant on the intervals between adjacent elements of Y°. Thus the element $\hat{z} \in Y^\circ$ which maximises (12) can be found by a simple search over the elements of Y°.

Since the EDFs are the distributions defined by the probabilities given by the unconstrained maximisation of the empirical loglikelihood function, they define the unconstrained maximum of that function. For the empirical likelihood ratio test statistic, we also require the maximum of the ELF constrained by the requirement of nondominance. This constrained maximum is given by the ELF (12) for the value $\tilde{z}$ that maximises (12). Like $\hat{z}$, $\tilde{z}$ can be found by search over the elements of Y°.

The constrained empirical-likelihood estimates of the CDFs of the two distributions $K = A, B$ can be written as

$$\tilde{F}_K(z) = \sum_{y_t^K \le z} p_t^K n_t^K, \tag{21}$$

where the probabilities p_t^K are given by (11) with $z = \tilde{z}$. Normally, $\tilde{z}$ is the only point in Y° for which $\tilde{F}_A(z)$ and $\tilde{F}_B(z)$ are equal. Certainly, there can be no z for which $\tilde{F}_A(z) < \tilde{F}_B(z)$ with strict inequality, since, if there were, the value of the ELF could be increased by imposing $\tilde{F}_A(z) = \tilde{F}_B(z)$, so that we would have $\text{ELF}(z) > \text{ELF}(\tilde{z})$, contrary to our assumption. Thus the distributions $\tilde{F}_A$ and $\tilde{F}_B$ are on the frontier of the null hypothesis of nondominance, and they represent those distributions contained in the null hypothesis that are closest to the unrestricted EDFs, for which there is dominance, by the criterion of the empirical likelihood.

For the remainder of our discussion, we restrict the null hypothesis to the frontier of nondominance, that is, to distributions such that $F_A(z_0) = F_B(z_0)$ for exactly one point z_0 in $[z^-, z^+]$, and $F_A(z) > F_B(z)$ with strict inequality for all $z \ne z_0$ in that interval. These distributions constitute the least favourable case of the hypothesis of nondominance in the sense that, with either the minimum t statistic or the minimum EL statistic, the probability of rejection of the null is no smaller on the frontier than with any other configuration of nondominance. This result follows from the following theorem.

Theorem 2

Suppose that the distribution F_A is changed so that the new distribution is weakly stochastically dominated by the old at first order. Then, for any z in the interior of the joint support U, the new distribution of the statistic $t(z)$ of which the square is given by (14) and the sign by that of $\hat{F}_A(z) - \hat{F}_B(z)$ weakly stochastically dominates its old distribution at first order. Consequently, the new distribution of the minimum t statistic also weakly stochastically dominates the old at first order. The same is true for the square root of the statistic $\text{LR}(z)$ given by (13) signed in the same way, and its minimum over z. If F_B

is changed so that the new distribution weakly stochastically dominates the old at first order, the same conclusions hold.

Proof: In Appendix A. ■

Remarks:

The changes in the statement of the theorem all tend to move the distributions in the direction of greater dominance of A by B. Thus we expect that they lead to increased probabilities of rejection of the null of nondominance. If, as the theorem states, the new distributions of the test statistics dominate the old, that means that their right-hand tails contain more probability mass, and so they indeed lead to higher rejection probabilities.

We are now ready to state the most useful consequence of restricting the null hypothesis to the frontier of nondominance.

Theorem 3

The minima over z of both the signed asymptotic t statistic $t(z)$ and the signed empirical likelihood ratio statistic $\mathrm{LR}^{1/2}(z)$ are asymptotically pivotal for the null hypothesis that the distributions A and B lie on the frontier of restricted nondominance of A by B. The frontier is such that there exists exactly one $z_0 \in [z^-, z^+]$ for which $F_A(z_0) = F_B(z_0)$, while $F_A(z) > F_B(z)$ strictly for all $z \neq z_0$ in $[z^-, z^+]$.

Proof: In Appendix A. ■

Remarks:

- Theorem 3 shows that we have at our disposal two test statistics suitable for testing the null hypothesis of restricted nondominance of A by B stochastically at first order, namely the minima of $t(z)$ and $\mathrm{LR}^{1/2}(z)$. For configurations that lie on the frontier of this hypothesis, as defined above, the asymptotic distribution of both statistics is $\mathrm{N}(0, 1)$. By Theorem 2, use of the quantiles of this distribution as critical values for the test leads to an asymptotically conservative test when there is nondominance inside the frontier.
- It is clear from the remark following the proof of Theorem 1 that both statistics are invariant under monotonic transformations of the measuring units of income.
- The fact that the statistics are asymptotically pivotal means that we can use the bootstrap to perform tests that should benefit from asymptotic refinements in finite samples; see Beran (1988). Specifically, the difference between the true rejection probability under the null hypothesis and the nominal level of the tests should converge to zero faster when the bootstrapped statistic is asymptotically pivotal than otherwise. We study this possibility by means of simulation experiments in the next section, where the bootstrap DGP generates bootstrap samples

of the same size for each population as in the original sample, using the constrained empirical-likelihood estimates (21).

When there is dominance in the sample, another approach is to seek the longest interval $[\hat{z}^-, \hat{z}^+]$ for which the hypothesis

$$\max_{z\in[\hat{z}^-,\hat{z}^+]} \bigl(F_B(z) - F_A(z)\bigr) \geq 0 \tag{22}$$

can be rejected. For simplicity, we concentrate in what follows on the lower bound $\hat{z}^-$. Since it is estimated from the sample, $\hat{z}^-$ is random. It is useful to conceive of it in much the same way as the limit of a confidence interval. Consider a nested set of null hypotheses, parametrised by z^-, of the form

$$\max_{z\in[z^-,z^+]} \bigl(F_B(z) - F_A(z)\bigr) \geq 0, \tag{23}$$

for some given upper limit z^+. As z^- increases, the hypothesis becomes progressively more constrained, and therefore easier to reject. For a given nominal level α, $\hat{z}^-$ is the smallest value of z^- for which the hypothesis (23) can be rejected at level α. This definition is analogous to that of the upper limit β_+ of a confidence interval for some parameter β. Just as $\hat{z}^-$ is the smallest value of z^- for which (23) can be rejected, so β_+ is the smallest value of β_0 for which the hypothesis $\beta = \beta_0$ can be rejected at (nominal) level α, where $1 - \alpha$ is the desired confidence level for the interval.

The analogy can be pushed a little further. The length of a confidence interval is related to the power of the test on which the confidence interval is based. Similarly, $\hat{z}^-$ is related to the power of the test of nondominance. The closer is $\hat{z}^-$ to the bottom of the joint support of the distributions, the more powerful is our rejection of nondominance.

It is important to realise that the empirically determined $\hat{z}^-$ is quite distinct from any ethically suggested z^- of the sort discussed in Section 5 above. $\hat{z}^-$ depends on, among other things, the sample sizes, and so, if there truly is dominance in the populations, $\hat{z}^-$ will depend negatively on sample size. The value of z^-, however, is unrelated to properties of any given samples. Consequently, $\hat{z}^-$ is determined by how informative the data are, while z^- is determined by *a priori* considerations, which may be of an ethical nature in poverty analysis, and, in analyses of financial or insurance data, may have to do with risk aversion or the regulatory environment.

7. Simulation Experiments

The main purpose of this section is to compare the performance of the bootstrap test we have proposed with that of the asymptotic version, which is to all intents and purposes the KPS test. There are various things that we wish to vary in the simulation experiments discussed in this section and presented in greater detail in Appendix B. First is sample size. Second is the extent of censoring by the choice of z^- and z^+.

Third is the way in which the two populations are configured. In those experiments in which we study the rejection probability of various tests under the null, we wish most of the time to have population A dominated by population B except at one point, where the CDFs of the two distributions are equal, When we wish to investigate the power of the tests, we allow B to dominate A to a greater or lesser extent.

Although we could compare the bootstrap test with tests for which the null is dominance of A by B, there is little interest in doing so, since, as we saw in the introduction, when there is dominance in the underlying distributions, it is easier to fail to reject the null of dominance of A by B than to reject nondominance of A by B. The latter permits a stronger conclusion, and so it is quite normal that it should occur less frequently.

Stochastic dominance to first order is invariant under increasing transformations of the variable z that is the argument of the CDFs F_A and F_B. It is therefore without loss of generality that we define our distributions on the $[0, 1]$ interval. We always let population A be uniformly distributed on this interval: $F_A(z) = z$ for $z \in [0, 1]$. For population B, the interval is split up into eight equal segments, with the CDF being linear on each segment. In the base configuration, the cumulative probabilities at the upper limit of each segment are 0.03, 0.13, 0.20, 0.50, 0.57, 0.67, 0.70, and 1.00. This is contrasted with the evenly increasing cumulative probabilities for A, which are 0.125, 0.25, 0.375, 0.50, 0.625, 0.75, 0.875, and 1.00. Clearly B dominates A everywhere except for $z = 0.5$, where $F_A(0.5) = F_B(0.5) = 0.5$. This base configuration is thus on the frontier of the null hypothesis of nondominance, as discussed in the previous section. In addition, we consider nondominance restricted to the interval $[z^-, z^+] = [0.1, 0.9]$.

In Table 1, we give the rejection probabilities of two asymptotic tests, based on the minimised values of $t(z)$ and $\text{LR}^{1/2}(z)$, as a function of sample size. The samples drawn from A are of sizes $N_A = 16$, 32, 64, 128, 256, 512, 1024, 2048, and 4096. The corresponding samples from B are of sizes $N_B = 7$, 19, 43, 91, 187, 379, 763, 1531, and 3067, the rule being $N_B = 0.75N_A - 5$. The results are based on 10,000 replications. Preliminary experiments showed that, when the samples from the two populations were of the same size, or of sizes with a large greatest common divisor, the possible values of the statistics, which depend on the quantities $N_K(z)$ and $M_K(z)$, $K = A, B$ that may have large common factors, were so restricted that their distributions were lumpy. For our purposes, this lumpiness conceals more than it reveals, and so it seemed preferable to choose sample sizes that were relatively prime.

As is evident in Table 1, the two test statistics turn out to be very close indeed in value when each is minimised over z. It is clear from both Table 1 and Figure 4 in Appendix B that the asymptotic tests have a tendency to underreject, a tendency which disappears only slowly as the sample sizes grow larger. This is hardly surprising. If the point of contact of the two distributions is at $z = z_0$, then the distribution of $t(z_0)$ and $\text{LR}^{1/2}(z_0)$ is approximately standard normal. But minimising with respect to z always yields a statistic that is no greater than one evaluated at z_0. Thus the rejection probability can be expected to be smaller, as we observe.

Table 1

N_A	$\alpha = 0.01$ $t_{\min}$	$\alpha = 0.01$ $\mathrm{LR}_{\min}$	$\alpha = 0.05$ $t_{\min}$	$\alpha = 0.05$ $\mathrm{LR}_{\min}$	$\alpha = 0.10$ $t_{\min}$	$\alpha = 0.10$ $\mathrm{LR}_{\min}$
16	0.001	0.000	0.005	0.005	0.013	0.017
32	0.000	0.000	0.004	0.004	0.017	0.015
64	0.001	0.001	0.009	0.010	0.026	0.030
128	0.003	0.003	0.021	0.021	0.048	0.047
256	0.001	0.006	0.033	0.033	0.070	0.069
512	0.010	0.010	0.039	0.039	0.082	0.082
1024	0.007	0.007	0.042	0.042	0.087	0.087
2048	0.010	0.010	0.043	0.043	0.087	0.087
4096	0.009	0.009	0.044	0.044	0.092	0.092

Rejection probabilities, asymptotic tests, base case, α = nominal level

We now consider bootstrap tests based on the minimised statistics. In bootstrapping, it is essential that the bootstrap samples are generated by a bootstrap DGP that satisfies the null hypothesis, since we wish to use the bootstrap in order to obtain an estimate of the distribution of the statistic being bootstrapped *under the null hypothesis*. Here, our rather artificial null is the frontier of nondominance, on which the statistics we are using are asymptotically pivotal, by Theorem 3.

Since the results we have obtained so far show that the two statistics are very similar even in very small samples, we may well be led to favour the minimum t statistic on the basis of its relative simplicity. But the procedure by which the empirical likelihood ratio statistic is computed also provides a very straightforward way to set up a suitable bootstrap DGP. Once the minimising z is found, the probabilities (11) are evaluated at that z, and these, associated with the realised sample values, the y_t^A and the y_s^B, provide distributions from which bootstrap samples can be drawn.

The bootstrap DGP therefore uses discrete populations, with atoms at the observed values in the two samples. In this, it is like the bootstrap DGP of a typical resampling bootstrap. But, as in Brown and Newey (2002), the probabilities of resampling any particular observation are not equal, but are adjusted, by use of the probabilities (11), so as to satisfy the null hypothesis under test. In our experiments, we used bootstrap DGPs determined in this way and generated bootstrap samples from them. For each bootstrap sample, then, we compute the minimum statistics just as with the original data. Bootstrap P values are then computed as the proportion of the bootstrap statistics that are greater than the statistic from the original data.

In Table 2, we give results like those in Table 1, but for bootstrap tests rather than asymptotic tests. For each replication, 399 bootstrap statistics were computed. Results are given only for the empirical likelihood statistic, since the t statistic gave results that were indistinguishable.

Table 2

N_A	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
32	0.001	0.018	0.051
64	0.003	0.033	0.082
128	0.006	0.049	0.104
256	0.013	0.053	0.106
512	0.010	0.049	0.102
1024	0.010	0.051	0.100

Rejection probabilities, bootstrap tests, base case, α = nominal level

It is not necessary, and it would have taken a good deal of computing time, to give results for sample sizes greater than those shown, since the rejection probabilities are not significantly different from nominal already for $N_A = 128$.

We see that, like the asymptotic tests, the bootstrap test suffers from a tendency to underreject in small samples. However, this tendency disappears much more quickly than with the asymptotic tests. Once sample sizes are around 100, the bootstrap seems to provide very reliable inference. This is presumably related to the fact that the bootstrap distribution, unlike the asymptotic distribution, is that of the *minimum* statistic, rather than of the statistic evaluated at the point of contact of the two distributions.

The property whereby bootstrap tests reject more than asymptotic ones is seen in the results of Appendix B to be quite general. Configurations can be found in which both tests underreject quite severely in small samples, but the bootstrap test underrejects less than the asymptotic test. In configurations in which the alternative of restricted dominance is true, the tendency to underreject leads to very poor power for the asymptotic test in samples of size up to around 500, while the bootstrap test has meaningful power already with a sample size of 64. Details of the simulations leading to these conclusions are in Appendix B.

8. Illustration using LIS data

We now illustrate briefly the application of the above methodology to real data using the Luxembourg Income Study (LIS) data sets[6] of the USA (2000), the Netherlands (1999), the UK (1999), Germany (2000) and Ireland (2000). The raw data are treated in the same manner as in Gottschalk and Smeeding (1997), taking household income to be income after taxes and transfers and using purchasing power parities and price

[6] See `http://lissy.ceps.lu` for detailed information on the structure of these data.

indices drawn from the Penn World Tables[7] to convert national currencies into 2000 US dollars. As in Gottschalk and Smeeding (1997), we divide household income by an adult-equivalence scale defined as $h^{0.5}$, where h is household size. All incomes are therefore transformed into year-2000 adult-equivalent US dollars. All household observations are also weighted by the product of household sample weights and household size. Sample sizes are 49,600 for the US, 5,000 for the Netherlands (NL), 25,000 for the UK, 10,900 for Germany (GE) and 2,500 for Ireland (IE).

This illustration abstracts from important statistical issues, such as the fact that the LIS data, like most survey data, are actually drawn from a complex sampling structure with stratification and clustering. (Note, however, that in any case the sampling design information of the LIS data is not available in the publicly accessible datasets.) Note also that negative incomes have been set to 0 (this affects no more than 0.5% of the observations), and that we ignore the possible presence of measurement errors in the data.

Figure 2 graphs the P values of tests of the null hypothesis that $F_A(z) \leq F_B(z)$ against the alternative that $F_A(z) > F_B(z)$ at various values of z over a range of \$1500 to \$7500, for various pairs of countries, and for both asymptotic and bootstrap tests. (Distribution A is the first country that appears in the legends in the Figure.) In all cases, bootstrap tests were based on 499 bootstrap samples. We set z^- to \$1500 and z^+ to \$7500 since these two bounds seem to be reasonable enough to encompass most of the plausible poverty lines for an adult equivalent. This "ethically suggested" \$1500 is also where we are able to start ranking the UK and the US. The asymptotic and bootstrap P values are very close for the comparisons of the US with either Germany or the UK. The bootstrap P values are slightly lower than the asymptotic ones for the NL-US comparison and somewhat larger for the US-IE one. These slight differences may be due to the smaller NL and IE samples. Although the differences are not enormous, they are significant enough to make bootstrapping worthwhile even if one is interested only in point-wise tests of differences in dominance curves.

Figure 3 presents the results of similar tests but this time over intervals of the form $[\$1500, z^+]$ for various values of z^+. The null hypothesis is therefore that $F_A(z) \leq F_B(z)$ for at least some z in $[\$1500, z^+]$ against the alternative hypothesis that $F_A(z) > F_B(z)$ over the entire range $[\$1500, z^+]$. For the NL-US comparison, note first that $\hat{F}_{\mathrm{US}}(z)$ is always lower than $\hat{F}_{\mathrm{NL}}(z)$ but that the difference between the two empirical distribution functions is small for z between around \$4800 to about \$10000. Although it is therefore difficult to reject the null hypothesis of nondominance for much of the range of z^+ values, the bootstrap P values are significantly lower than their asymptotic counterparts, as is to be expected, given the greater power of the bootstrap test procedure seen in the simulation experiments. A similar result is found for a US-UK nondominance test. The US-GE comparison yields very close asymptotic and bootstrap P values, and both procedures would reject at a level of 5% the null

[7] See Summers and Heston (1991) for the methodology underlying the computation of these parities, and `http://pwt.econ.upenn.edu/` for access to the 1999-2000 figures.

hypothesis of nondominance of Germany over the US for a range of approximately [\$1500, \$6750]. A US-IE test of nondominance generates bootstrap P values that are larger than the asymptotic ones. Again, this is in contrast with the other comparisons, and it also brings back to mind that bootstrap and asymptotic results can differ somewhat with small samples and tests covering the tails of distributions.

Table 3 illustrates how the differences in the power of the asymptotic and bootstrap tests can influence the ranges over which we may reject nondominance of UK over the US. The P values of the first two reported tests are both equal to 5%, but the asymptotic test is over the range $Z = [\$1550, \$5577]$ and the bootstrap test is over the wider range $Z = [\$1550, \$5680]$. Thus, using a bootstrap test extends by about \$100 the range Z of poverty lines over which we can declare – at a level of 5% – the UK to have less poverty than the US for all of the poverty indices that belong to $\Pi^1(Z)$ (recall (19)), and it is therefore more powerful than the asymptotic test. A similar result applies for a test level of 10%: the range over which we can reject that $\Pi^1(Z)$ poverty is no lower in the UK is $Z = [\$1068, \$5698]$ for the asymptotic test and $Z = [\$1068, \$5784]$ for the bootstrap test. Almost as importantly, given the prevalence of the use of the poverty headcount index in policy and poverty analysis circles, the bootstrap procedure extends the range of poverty lines over which we can confidently and jointly declare the headcount to be lower in the UK than in the US.

Table 3

Type of tests	Range of z	P values
Asymptotic	$[\$1550, \$5577]$	5%
Bootstrap	$[\$1550, \$5680]$	5%
Asymptotic	$[\$1068, \$5698]$	10%
Bootstrap	$[\$1068, \$5784]$	10%

P values of four tests of the null hypothesis that the UK does not dominate the US

9. Possible Extensions

In the literature on testing the null of dominance, the paper that deals with the widest set of conditions that we know of is Linton, Maasoumi, and Whang (2005), henceforth LMW. There, the samples may be serially correlated, and, if the observations in two samples can be paired, correlation between the paired observations is allowed for. In addition, this paper considers hypotheses about more than two samples, and also treats samples of "generated" data, where the observations depend on parameters that are estimated using the same data as those used for testing for dominance. Finally, orders of stochastic dominance higher than the first are taken into account. In this section, we briefly discuss how one may test for restricted nondominance in situations of this sort, and how the procedures of this paper can be extended in order to do so.

One of the hypotheses considered by LMW is that, among a set of distributions, or prospects as they call them, there is at least one that stochastically dominates at least one of the others. From the point of view of testing for nondominance, the analogous hypothesis is that no distribution dominates any other. For, if this hypothesis is rejected, all that remains is the null hypothesis of LMW, although, in practice, with continuous distributions, the hypothesis of nondominance must be replaced by one of restricted nondominance. The hypothesis that no distribution dominates any other can be rejected only if one distribution dominates another in the sample. In this case, we can perform a pairwise test of the sort we have developed. If there is more than one pair with restricted dominance in the sample, the intersection-union principle allows us to use the most significant pairwise rejection to determine the P value for the overall hypothesis.

If the samples to be compared are, for instance, generated as residuals from the linear regression of some directly observed variables on other directly observed covariates, or, more generally, if the samples depend on estimated parameters, the empirical likelihood problem can be posed in such a way as to include these parameters in the maximisation problem. Solving this problem simultaneously gives parameter estimates and a set of probabilities that implicitly define weighted EDFs that satisfy the null of restricted nondominance. In this way, the parameter estimates are obtained under the null, and so, when the null is true, are in general more efficient than unrestricted estimates.

Higher orders of stochastic dominance can be dealt with by using an order-s constraint in (5) ($s = 1, 2, ...$) of the form (for given z)

$$\sum_{y_t^A \in Y^A} p_t^A \mathrm{I}(y_t^A \le z)(z - y_t^A)^{s-1} = \sum_{y_s^B \in Y^B} p_s^B \mathrm{I}(y_s^B \le z)(z - y_t^B)^{s-1}. \tag{24}$$

A possible correlation within paired observations can also be taken into account by considering the joint ELF of A and B in the constrained and unconstrained maximisation problem. The details are shown in Davidson (2007), which extends the approach of this paper. For both the higher-order and the dependence extensions, it can *inter alia* be shown that the appropriately redefined t and LR statistics are asymptotically pivotal on the frontier of the null hypothesis of nondominance.

Serially correlated data pose a harder problem, unless one is prepared to model the serial dependence parametrically, but it may be possible to adopt a modified version of the LMW approach and use a subsampling bootstrap for inference. This remains a problem for future work.

10. Discussion and Conclusions

In this paper, we have adopted the point of view that, if we really wish to demonstrate statistically that the distribution of population B stochastically dominates that of population A at first order, then it is appropriate to use a null hypothesis of nondominance, since, if we reject it, all that is left is dominance. However, we show that it is impossible to reject this null at any conventional significance level if we have continuous distributions and test for nondominance at all of the observations in samples drawn from them. With discrete distributions, this problem does not necessarily arise, and indeed, in practice, many investigators explicitly or implicitly discretise their samples by setting up a grid of points and agglomerating observations in the samples on to atoms at the points of the grid.

If we are ready in the case of continuous distributions to censor the tails of the distributions, and thus to seek restricted dominance, then we have seen that it is easy to set up both asymptotic and bootstrap tests for the null of nondominance. Note that such censoring will also protect at least partially against measurement errors and outliers in the tails of the distributions. We consider two seemingly different statistics, one the minimum t statistic of KPS, the other an empirical likelihood ratio statistic. We show that the two statistics typically take on very similar values in practice, and that inference using one of them is indistinguishable from inference using the other. The advantage of the empirical likelihood ratio statistic is that, in order to compute it, we compute a set of probabilities that estimate the probabilities of the populations under the hypothesis that they are at the frontier of nondominance, that is, that they are such that there is dominance of A by B everywhere except at exactly one point in the interior of the common support of the distributions.

This fact makes it possible to use the bootstrap in order to estimate the distributions of either one of the two statistics under data-generating processes that are on the frontier of nondominance. In fact, we show that the statistics are asymptotically pivotal on the frontier, so that we can expect that the bootstrap will provide more reliable inference than the asymptotic distributions of the statistics. This turns out to be the case in a selection of configurations that we study by means of simulation experiments. Our preferred testing procedure is thus a bootstrap procedure, in which the bootstrap samples are generated using the probabilities computed in the process of evaluating the empirical likelihood ratio statistic. It does not seem to matter whether the minimum t statistic or the likelihood ratio statistic is used.

Most of the literature on testing relations between a pair of distributions deals with tests for which the null hypothesis is dominance. It is plausible to suppose that these tests too can be dealt with by the methods of empirical likelihood, but it is less simple to do so. For this sort of test, we do not reject the null of dominance unless there is nondominance in the sample. In that case, we wish to find the distributions that respect the null of dominance and are closest, by the criterion of the empirical likelihood, to the unrestricted estimates that exhibit nondominance. These distributions must of course lie on the frontier of the null hypothesis. In general, however, it is not enough to require that there should be just one point $y \in Y$ at which the restricted estimates coincide. In Wolak (1989), this matter is considered for the case of discrete distributions, and it is shown that locating the pair of distributions on the frontier of the null closest to a pair of sample distributions which display nondominance involves the solution of a quadratic programming problem. Further, the asymptotic distribution of the natural test statistic, under a DGP lying on the frontier, is a mixture of chi-squared distributions that is not as simple to treat as the standard normal asymptotic distributions found in this paper. It remains for future research to see whether empirical likelihood methods, used with continuous distributions, can simplify tests with a null of dominance.

The empirical likelihood methods of this paper could also prove useful for tests of a general intersection-union type, for which, as in this paper, the null hypothesis is formulated as a union of multiple hypotheses and the alternative is the intersection of the contraries of these multiple hypotheses. There are numerous examples of this in economics, such as when we want to check whether a ranking is valid for a range of parameter values (for instance, of equivalence scales, prices indices, behavioural elasticities), or for an array of measurement procedures, or for a set of socio-economic groups or environments.

Appendix A

Proof of Theorem 1:

For $K = A, B$, $N_K(z) = N_K \hat{F}_K(z)$ and $M_K(z) = N_K(1 - \hat{F}_K(z))$. Therefore

$$\begin{aligned} & N_K(z) \log N_K(z) + M_K(z) \log M_K(z) \\ = & N_K \log N_K + N_K\big(\hat{F}_K(z) \log \hat{F}_K(z) + (1 - \hat{F}_K(z)) \log(1 - \hat{F}_K(z))\big). \end{aligned} \tag{25}$$

Further,

$$\begin{aligned} \Big(\sum_{K=A,B} N_K(z)\Big) \log\Big(\sum_{K=A,B} N_K(z)\Big) + \Big(\sum_{K=A,B} M_K(z)\Big) \log\Big(\sum_{K=A,B} M_K(z)\Big) = \\ N \log N + \Big(\sum_{K=A,B} N_K \hat{F}_K(z)\Big) \log\Big(\sum_{K=A,B} \frac{N_K}{N} \hat{F}_K(z)\Big) + \\ \Big(\sum_{K=A,B} N_K(1 - \hat{F}_K(z))\Big) \log\Big(\sum_{K=A,B} \frac{N_K}{N}(1 - \hat{F}_K(z))\Big) \end{aligned} \tag{26}$$

From (13), we see that $\mathrm{LR}(z)$ is equal to twice the expression

$$\begin{aligned} & \sum_{K=A,B} \Big(N_K(z) \log N_K(z) + M_K(z) \log M_K(z) - N_K \log N_K\Big) + N \log N \\ & -\Big(\sum_{K=A,B} N_K(z)\Big) \log\Big(\sum_{K=A,B} N_K(z)\Big) - \Big(\sum_{K=A,B} M_K(z)\Big) \log\Big(\sum_{K=A,B} M_K(z)\Big) \end{aligned}$$

From (25) and (26), this expression can be written as

$$\begin{aligned} & - \sum_{K=A,B} N_K \hat{F}_K(z) \log\Big(\frac{N_A \hat{F}_A(z) + N_B \hat{F}_B(z)}{N \hat{F}_K(z)}\Big) \\ & - \sum_{K=A,B} N_K(1 - \hat{F}_K(z)) \log\Big(\frac{N - (N_A \hat{F}_A(z) + N_B \hat{F}_B(z))}{N(1 - \hat{F}_K(z))}\Big). \end{aligned} \tag{27}$$

Consider now the first sum in the above expression, which can be written as

$$\begin{aligned} & -\big(N_A \hat{F}_A(z) + N_B \hat{F}_B(z)\big) \log\big(N_A \hat{F}_A(z) + N_B \hat{F}_B(z)\big) \\ & + N_A \hat{F}_A(z) \log N \hat{F}_A(z) + N_B \hat{F}_B(z) \log N \hat{F}_B(z). \end{aligned} \tag{28}$$

Define $\Delta(z) \equiv \hat{F}_A(z) - \hat{F}_B(z)$. Then we see that $N_A \hat{F}_A(z) + N_B \hat{F}_B(z) = N \hat{F}_B(z) + N_A \Delta(z)$. Making these substitutions lets us write expression (28) as

$$\begin{aligned} & -\big(N \hat{F}_B(z) + N_A \Delta(z)\big) \Big(\log N \hat{F}_B(z) + \log\Big(1 + \frac{N_A \Delta(z)}{N \hat{F}_B(z)}\Big)\Big) \\ & + N_A(\hat{F}_B(z) + \Delta(z))\big(\log N \hat{F}_B(z) + \log\big(1 + \Delta(z)/\hat{F}_B(z)\big)\big) + N_B \hat{F}_B(z) \log N \hat{F}_B(z). \end{aligned}$$

Taylor expanding up to second order in $\Delta(z)$ then gives

$$(-N + N_A + N_B)\hat{F}_B(z)\log N\hat{F}_B(z) - N_A\Delta(z) + \frac{1}{2}\frac{N_A^2\Delta^2(z)}{N\hat{F}_B(z)} - N_A\Delta(z)\log N\hat{F}_B(z)$$
$$-\frac{N_A^2\Delta^2(z)}{N\hat{F}_B(z)} + N_A\Delta(z) - \frac{1}{2}\frac{N_A\Delta^2(z)}{\hat{F}_B(z)} + N_A\Delta(z)\log N\hat{F}_B(z) + \frac{N_A\Delta^2(z)}{\hat{F}_B(z)} + O_p(N^{-1/2}),$$

since, under our assumptions, $N_K = O_p(N)$ and $\Delta(z) = O_P(N^{-1/2})$. The term independent of $\Delta(z)$ in the above expression and the terms linear in $\Delta(z)$ all cancel, and so what remains is just a term of order unity and a remainder that tends to zero as $N \to \infty$:

$$\frac{1}{2}\frac{N_A(N-N_A)\Delta^2(z)}{N\hat{F}_B(z)} + O_p(N^{-1/2}) = \frac{1}{2}\frac{N_A N_B\Delta^2(z)}{N\hat{F}_B(z)} + O_p(N^{-1/2}).$$

Since $\hat{F}_B(z) = F(z) + O_p(N^{-1/2})$, this expression is equal to $N_A N_B\Delta^2(z)/2NF(z)$ to the same order. An exactly similar calculation for the second line of (27) shows that, to the same order of approximation, it is equal to $N_A N_B\Delta^2(z)/2N(1 - F(z))$. The entire expression (27) is therefore

$$\frac{1}{2}\frac{N_A N_B\Delta^2(z)}{N}\left(\frac{1}{F(z)} + \frac{1}{1-F(z)}\right) = \frac{1}{2}\frac{N_A N_B\Delta^2(z)}{NF(z)(1-F(z))} + O_p(N^{-1/2}). \qquad (29)$$

Finally, since $N_A/N \to r$ as $N \to \infty$ and $N_B/N \to 1-r$, we see that the large-sample limit of $\text{LR}(z)$, which is twice that of (29), is

$$\frac{r(1-r)}{F(z)(1-F(z))}\operatorname*{plim}_{N\to\infty} N\Delta^2(z),$$

which is the leading-order term on the right-hand side of (15), as required. ■

Proof of Theorem 2:

The proof relies on the following construction, based on that in the proof of Lemma 1 on page 84 of Lehmann (1986).

Consider two CDFs F_1 and F_2 defined on the real line such that F_1 weakly stochastically dominates F_2 at first order, and a random variable V distributed uniformly on $[0, 1]$. As in Lehmann, define the quantile functions f_i, $i = 1, 2$, by the relation

$$f_i(y) = \inf\{x \mid F_i(x-) \le y \le F_i(x)\}.$$

Clearly the f_i are weakly increasing and such that $f_i\big(F_i(x)\big) \le x$ and $F_i\big(f_i(y)\big) \ge y$ for all real x and y for which the functions are defined. In addition, the inequalities $f_i(y) \le x$ and $y \le F_i(x)$ are equivalent. Thus

$$\Pr\big(f_i(V) \le x\big) = \Pr\big(V \le F_i(x)\big) = F_i(x),$$

so that the random variable $f_i(V)$ has CDF F_i. Since $F_1(x) \le F_2(x)$ for all real x by the hypothesis of weak stochastic dominance, it follows that $f_1(y) \ge f_2(y)$ for all real y.

Let $\{u_i\}$, $i = 1, \ldots, N$ be a sequence of IID "random numbers", each distributed uniformly on $[0, 1]$. These random numbers can generate two IID random samples, $\mathcal{Y} \equiv \{y_i\}$ and $\mathcal{Z} \equiv \{z_i\}$, $i = 1, \ldots, N$, with $y_i = f_1(u_i)$ and $z_i = f_2(u_i)$. The sample $\{y_i\}$ is a sample drawn from the distribution F_1, while $\{z_i\}$ is drawn from F_2. Since $f_1(u) \ge f_2(u)$ for all u, the EDF of $\mathcal{Y}$ stochastically dominates that of $\mathcal{Z}$ at first order.

Consider now two random samples of N IID draws, generated by the same set of random numbers, the first from distribution F_A, the second from a new distribution $F_{A'}$ that is weakly stochastically dominated by F_A. The above result demonstrates that the EDF $\hat{F}_A$ of the first sample is nowhere greater than the EDF $\hat{F}_{A'}$ of the second.

We show below that the square root statistics $t(z)$ and $\mathrm{LR}(z)$ defined in the statement of the theorem are non-decreasing functions of $\hat{F}_A(z)$ for all z. Thus, for each z, $t(z) \le t'(z)$ where $t(z)$ is the statistic computed using the first sample and $t'(z)$ is that computed using the second sample. It follows that the minimum statistic for the first sample, t_* say, is no greater than the minimum statistic t'_* for the second sample.

Let $\mathcal{U}$ denote the set of random numbers $\{u_i\}$ for which $t'_* \le x$ for a given real value x. Then $t_* \le x$ for all sets of random numbers in $\mathcal{U}$. Thus $\Pr(t_* \le x) \ge \Pr(t'_* \le x)$, which means that the distribution of t'_* weakly stochastically dominates that of t_*, as stated by the theorem.

The same arguments apply to the minimum LR statistic, and also to changes in F_B as described in the statement of the theorem, since, as seen below, $t(z)$ and $\mathrm{LR}(z)$ are non-increasing functions of $\hat{F}_B(z)$.

We compute the derivative with respect to $\hat{F}_A(z)$ of $t(z)$ as given by the square root of expression (14). This square root can be written in the form

$$C \frac{x - y}{\big(x(1-x) + k\big)^{1/2}} \tag{30}$$

where $x = \hat{F}_A(z)$, $y = \hat{F}_B(z)$, $k = (N_A/N_B)\hat{F}_B(z)(1 - \hat{F}_B(z))$, and C is a positive constant. The derivative of expression (30) with respect to x is C times

$$\frac{2x(1-x) + 2k - (x-y)(1-2x)}{2\big(x(1-x) + k\big)^{3/2}}.$$

This expression is certainly positive unless $x-y$ and $1-2x$ have the same sign. Suppose first that $x \le 1/2$ and $x - y > 0$. Then, since $y \ge 0$, $x \ge x - y$ and so

$$2x(1-x) - (x-y)(1-2x) \ge 2x(1-x) - x(1-2x) = x \ge 0.$$

Similarly, if $x \geq 1/2$ and $x - y < 0$, we see that $|x - y| \leq 1 - x$. Then

$$2x(1-x) - (y-x)(2x-1) \geq 2x(1-x) - (1-x)(2x-1) = 1 - x \geq 0.$$

Thus the derivative is positive in all cases. The proof that the derivative of $t(z)$ with respect to $\hat{F}_B(z)$ is negative is exactly similar.

The statistic $\mathrm{LR}(z)$ is given by twice the expression (27). The first line of (27) is in turn equal to (28), of which the derivative with respect to $\hat{F}_A(z)$ is

$$\begin{aligned}-N_A \log\bigl(N_A\hat{F}_A(z) + N_B\hat{F}_B(z)\bigr) - N_A + N_A \log N\hat{F}_A(z) + (N_A/N)N \\ = -N_A \log\Bigl(1 - \frac{N_B\varDelta(z)}{N\hat{F}_A(z)}\Bigr).\end{aligned}$$

Since $N_B/(N\hat{F}_A(z))$ is positive, this expression has the same sign as $\varDelta(z)$. Similarly, the derivative of the second line of (27) with respect to $\hat{F}_A(z)$ is

$$N_A \log\Bigl(1 + \frac{N_B\varDelta(z)}{N(1-\hat{F}_A(z))}\Bigr),$$

of which the sign is also the same as that of $\varDelta(z)$. Since the square root statistic is defined to have the same sign as $\varDelta(z)$, its derivative with respect to $\hat{F}_A(z)$ is everywhere nonnegative. This completes the proof. ■

Proof of Theorem 3:

Under the restricted null hypothesis of the statement of the theorem, the statistic $t(z_0)$ is distributed asymptotically as $\mathrm{N}(0,1)$. The probability that $t(z_0) \leq z_{1-\alpha}$, where $z_{1-\alpha}$ is the $(1-\alpha)$ quantile of $\mathrm{N}(0,1)$, therefore tends to $1-\alpha$ as $N \to \infty$. The probability that the minimum over $z \in Y^\circ$ of $t(z)$ is less than $z_{1-\alpha}$ is therefore no smaller than $1-\alpha$ asymptotically. Thus the probability of rejecting the null of nondominance on the basis of the minimum of $t(z)$ is no greater than α. This is the standard intersection-union argument used to justify the use of the minimum of $t(z)$ as a test statistic.

In Theorem 2.2 of KPS, it is shown that, if the distributions A and B belong to the restricted null hypothesis, then the probability of rejecting the null is actually equal to α asymptotically. We conclude therefore that the asymptotic distribution of the minimum of $t(z)$ is $\mathrm{N}(0,1)$. Since this is a unique distribution, it follows that this statistic is asymptotically pivotal for the restricted null. The local equivalence of $t(z)$ and $\mathrm{LR}^{1/2}(z)$ shown in Theorem 1 then extends the result to the empirical likelihood ratio statistic. ■

Appendix B

Comparison of asymptotic and bootstrap tests

In Figure 4, we graph P value plots for the minimised values of $t(z)$ and $\mathrm{LR}^{1/2}(z)$ over the full range from 0 to 1. A P value plot is of the CDF of the P value of a test. See Davidson and MacKinnon (1998) for a discussion of P value plots.

Two sample sizes are shown: $N_A = 32$ and $N_A = 256$. In the latter case, it is hard to see any difference between the plots for the two statistics, and even for the much smaller sample size, the differences are plainly very minor indeed.

In the experimental setup that gave rise to Figure 2, it was possible to cover the full range of the statistics, since, even when there was nondominance in the sample, we could evaluate the statistics as usual, obtaining negative values. This was for illustrative purposes only. In practice, one would stop as soon as nondominance is observed in the sample, thereby failing to reject the null hypothesis.

In Figure 5, P value plots are given for $N_A = 32$ and 128, for the asymptotic and bootstrap tests based on the empirical likelihood statistic. This time, we show results only for P values less than 0.5. In the bootstrap context, if there is nondominance in the original samples, no bootstrapping is done, and a P value of 1 is assigned. If there is dominance in the original samples, an event which under the null has a probability that tends to one half as the sample sizes tend to infinity, then bootstrapping is undertaken; each time the bootstrap generates a pair of samples without dominance, since the bootstrap test statistic would be negative, and so not greater than the positive statistic from the original samples, this bootstrap replication does not contribute to the P value. Thus a bootstrap DGP that generates many samples without dominance leads to small P values and frequent rejection of the null of nondominance.

Effect of censoring

We now look at the effects of censoring in the tails of the distributions. In Figure 6 are shown P value plots for the base case with $N_A = 128$, for different amounts of censoring. Results for the asymptotic test are in the left panel; for the bootstrap test in the right panel. It can be seen that, for the asymptotic test, the rejection rate diminishes steadily with z^- over the range $[0.01,\ 0.10]$, where the null is restricted nondominance over the interval $[z^-, 1 - z^-]$. This behaviour is entirely as expected, in accord with the discussion in Section 4. For values of z^- in the range 0.10 to 0.16, the P value plots are essentially identical.

With the bootstrap, dependence on the extent of censoring is considerably less: for P values up to around 0.3, and z^- greater than 0.03, the dependence is very slight.

Other configurations that satisfy the null

The base case we have considered so far is one in which B dominates A substantially except at one point in the middle of the distribution. We now consider two other configurations, the first in which the two distributions still touch in the middle, but the dominance by B is less elsewhere. The cumulative probabilities at the upper limits of the eight segments in this case are 0.1, 0.2, 0.3, 0.5, 0.6, 0.7, 0.8, and 1.0. The second configuration has the two distributions touching twice, for values of z equal to 0.25 and 0.75. The cumulative probabilities are 0.10, 0.25, 0.35, 0.45, 0.55, 0.75, 0.85, and 1.00. Results are shown in Figure 7, with z^- set to 0.1, and $N_A = 64$ and $N_B = 43$. The tests are based on the minimum t statistic. As usual, the empirical likelihood statistic gives essentially indistinguishable results.

For both configurations, all the tests are conservative, with rejection probabilities well below nominal in reasonably small samples. In the second configuration, in which the distributions touch twice, the tests are more conservative than in the first configuration. In both cases, it can be seen that the bootstrap test is a good deal less conservative than the asymptotic one. However, in all cases, the P value plots flatten out for larger values of P, because the P value is bounded above by 1 minus the proportion of bootstrap samples in which there is nondominance. In these two configurations, the probability of dominance in the original data, which is the asymptote to which the P value plots tend, is substantially less than a half.

Another configuration that we looked at needs no graphical presentation. If both populations correspond to the uniform distribution on $[0, 1]$, rejection of the null of nondominance simply did not occur in any of our replications. Of course, when the distributions coincide over their whole range, we are far removed from the frontier of the null hypothesis, and so we expect to have conservative tests.

Power

We now turn our attention to considerations of power. We study two configurations in which population B dominates A. In the first, we modify our base configuration slightly, using as cumulative probabilities at the upper limits of the segments the values 0.03, 0.13, 0.20, 0.40, 0.47, 0.57, 0,70, and 1.00. There is therefore clear dominance in the middle of the distribution. The second configuration uses cumulative probabilities of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, and 1.0. This distribution is uniform until the last segment, which has a much greater probability mass than the others.

In Figure 8, various results are given, with those for the first configuration in the upper left-hand panel and for the second in the other two panels. Both asymptotic and bootstrap tests based on the minimum t statistic are considered, and z^- is set to 0.1. There is nothing at all surprising in the left-hand panel. We saw in Figure 4 that, with the base configuration, the asymptotic test underrejects severely for $N_A = 32$ and $N_B = 19$. Here, the rejection rate is still less than the nominal level for those sample sizes. With the base configuration, the bootstrap test also underrejects, but less severely, and here it achieves a rejection rate modestly greater than the significance level. For $N_A = 64$ and $N_B = 43$, the increased power brought by larger samples is manifest. The asymptotic test gives rejection rates modestly greater than the level, but the bootstrap test does much better, with a rejection rate of slightly more than 14% at a 5% level, and nearly 28% at a 10% level.

In the second configuration, power is uniformly much less. If we were to change things so that the null of nondominance was satisfied, say by increasing the cumulative probability in population B for z around 0.25, then the results shown in Figure 5 indicate that the tests would be distinctly conservative. Here we see the expected counterpart when only a modest degree of dominance is introduced, namely low power. Even for $N_A = 128$, the rejection rate of the asymptotic test is always smaller than the significance level. With the larger sample sizes of the right-hand panel, some ability to reject is seen, but it is not at all striking with $N_A = 256$. In contrast, the bootstrap test has some power for all sample sizes except $N_A = 64$, and its rejection rate rises rapidly in larger samples, although rejection rates comparable to those obtained with the first configuration with $N_A = 64$ are attained only for N_A somewhere between 256 and 512.

The possible configurations of the two populations are very diverse indeed, and so the results presented here are merely indicative. However, a pattern that emerges consistently is that bootstrap tests outperform their asymptotic counterparts in terms of both size and power. They are less subject to the severe underrejection displayed by asymptotic tests even when the configuration is on the frontier of the null hypothesis, and they provide substantially better power to reject the null when it is significantly false.

Conventional practice often discretises data, transforming them so that the distributions have atoms at the points of a grid. Essentially, the resulting data are sampled from discrete distributions. A few simulations were run for such data. The results were not markedly different from those obtained for continuous data censored in the tails. The tendency of the asymptotic tests to underreject is slightly less marked, because the discretisation means that the minimising z is equal to the true (discrete) z_0 with high probability. However, the lumpiness observed when the two sample sizes have a large greatest common divisor is very evident indeed, and prevents simulation results from being as informative as those obtained from continuous distributions.

References

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

Barrett, G. and S. Donald (2003). "Consistent Tests for Stochastic Dominance" *Econometrica*, **71**, 71–104.

Beach, C.M. and J. Richmond (1985). "Joint Confidence Intervals for Income Shares and Lorenz Curves" *International Economic Review*, **26**, 439–50.

Beard, R., T. Pentikainen and E. Pesonen (1984). *Risk Theory: The Stochastic Basis of Insurance*, Chapman and Hall, New York.

Buhlmann, H. (1970). *Mathematical Methods in Risk Theory*, Springer-Verlag, Heidelberg.

Beran, R. (1988). "Prepivoting test statistics: A bootstrap view of asymptotic refinements" *Journal of the American Statistical Association*, **83**, 687–697.

Bishop, J., V. Chow, and J. Formby (1991). "A Stochastic Dominance Analysis of Growth, Recessions and the U.S. Income Distribution, 1967-1986" *Southern Economic Journal*, **57**, 936–46.

Bishop, J., J. Formby, and P. Thistle (1992). "Convergence of the South and Non-South Income Distributions, 1969–1979" *American Economic Review*, **82**, 262–272.

Bourguignon, F. and G. Fields (1997). "Discontinuous Losses from Poverty, Generalized P_α Measures, and Optimal Transfers to the Poor" *Journal of Public Economics*, **63**,155–175.

Brown, B. W. and W. Newey (2002). "Generalized method of moments, efficient bootstrapping, and improved inference" *Journal of Business and Economic Statistics*, **20**, 507–517.

Chen, S., G. Datt and M. Ravallion (1994). "Is Poverty Increasing in the Developing World?" *Review of Income and Wealth*, **40**, 359–76.

Davidson, R. (2007). "Testing for Restricted Stochastic Dominance: Some Further Results", working paper, McGill University Department of Economics.

Davidson, R. and J.-Y. Duclos (2000). "Statistical Inference for Stochastic Dominance and the for the Measurement of Poverty and Inequality" *Econometrica*, **68**, 1435–1465.

Davidson, R. and J. G. MacKinnon (1998). "Graphical Methods for Investigating the Size and Power of Hypothesis Tests" *The Manchester School*, **66**, 1–26.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications*, vol. 2, Wiley, New York.

Foster, J., J. Greer and E. Thorbecke (1984). "A Class of Decomposable Poverty Measures" *Econometrica*, **52**, 761–776.

Foster, J.E. and A.F. Shorrocks (1988a). "Poverty Orderings" *Econometrica*, **56**, 173–177.

Foster, J.E. and A.F. Shorrocks (1988b). "Poverty Orderings and Welfare Dominance" *Social Choice Welfare*, **5**, 179–198.

Gerber, H. (1979). *An Introduction to Mathematical Risk Theory*, Irwin Homewood, Illinois.

Gottschalk, P. and T. M. Smeeding (1997). "Cross-National Comparisons of Earnings and Income Inequality" *Journal of Economic Literature*, **35**, 633–687.

Hadar, J. and William Russell (1969). "Rules for Ordering Uncertain Prospects" *The American Economic Review*, **59**, 25–34.

Hanoch, G. and H. Levy (1969). "The Efficiency Analysis of Choices Involving Risk" *Review of Economic Studies*, **36**, 335–346.

Holton, G. A. (2003). *Value-at-Risk: Theory and Practice*, Academic Press.

Howes, S. (1993a). "Asymptotic Properties of Four Fundamental Curves of Distributional Analysis" Unpublished paper, STICERD, London School of Economics.

Howes, S. (1993b). "Restricted Stochastic Dominance: a Feasible Approach to Distributional Analysis" Unpublished paper, STICERD, London School of Economics.

Jorion, P. (2001). *Value at Risk: The New Benchmark for Managing Financial Risk*, 2nd ed., McGraw-Hill.

Kaur, A., B.L. Prakasa Rao, and H. Singh (1994). "Testing for Second-Order Stochastic Dominance of Two Distributions" *Econometric Theory*, **10**, 849–866.

Klecan, L., R. McFadden and D. McFadden (1991), "A robust test for stochastic dominance" Working paper, Department of Economics, MIT.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd edition, Wiley, New York.

Levy, H. (1992). "Stochastic Dominance and Expected Utility: Survey and Analysis" *Management Science*, **38**, 555–593.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005). "Consistent Testing for Stochastic Dominance under General Sampling Schemes" *Review of Economic Studies*, **72**, 735–765.

Maasoumi, E. and A. Heshmati (2005). "Evaluating Dominance Ranking of PSID Incomes by Various Household Attributes" IZA Discussion Paper No. 1727.

McFadden, D. (1989). "Testing for Stochastic Dominance" in *Studies in the Economics of Uncertainty*, T.B. Fomby and T.K. Seo, editors, Springer-Verlag, New York.

Mosler, K. (2004). "Restricted Lorenz dominance of economic inequality in one and many dimensions" *Journal of Economic Inequality*, **2**, 89–103.

Owen, A. B. (2001). *Empirical Likelihood*, Chapman and Hall.

Rawls, J. (1971). *A Theory of Justice*, Harvard University Press, Cambridge, Mass.

Sen, A. (1983). "Poor, relatively speaking" *Oxford Economic Papers*, **35**, 153–169.

Summers, R. and A. Heston (1991). "The Penn World Table (Mark 5): An Expanded Set of International Comparisons, 1950-1988" *The Quarterly Journal of Economics*, **106**, 327–368.

Wolak, F. A. (1989). "Testing Inequality Constraints in Linear Econometric Models" *Journal of Econometrics*, **41**, 205–235.

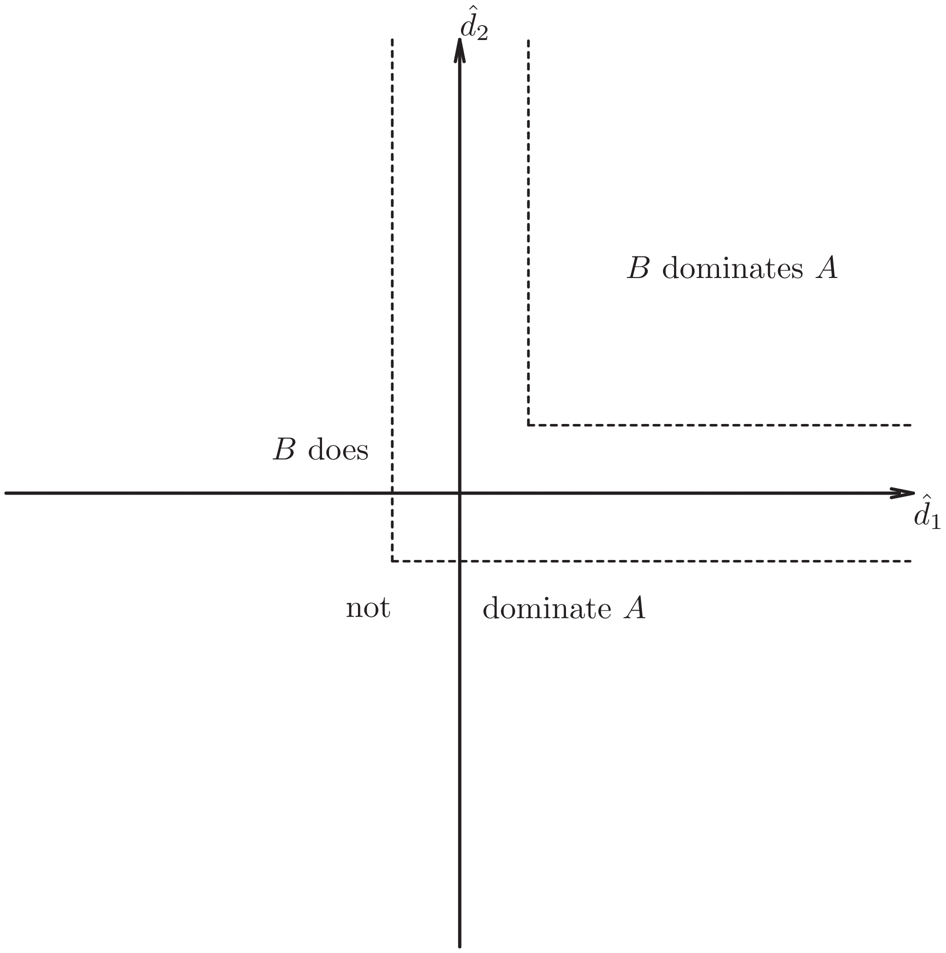


Figure 1: Tests of dominance and non-dominance

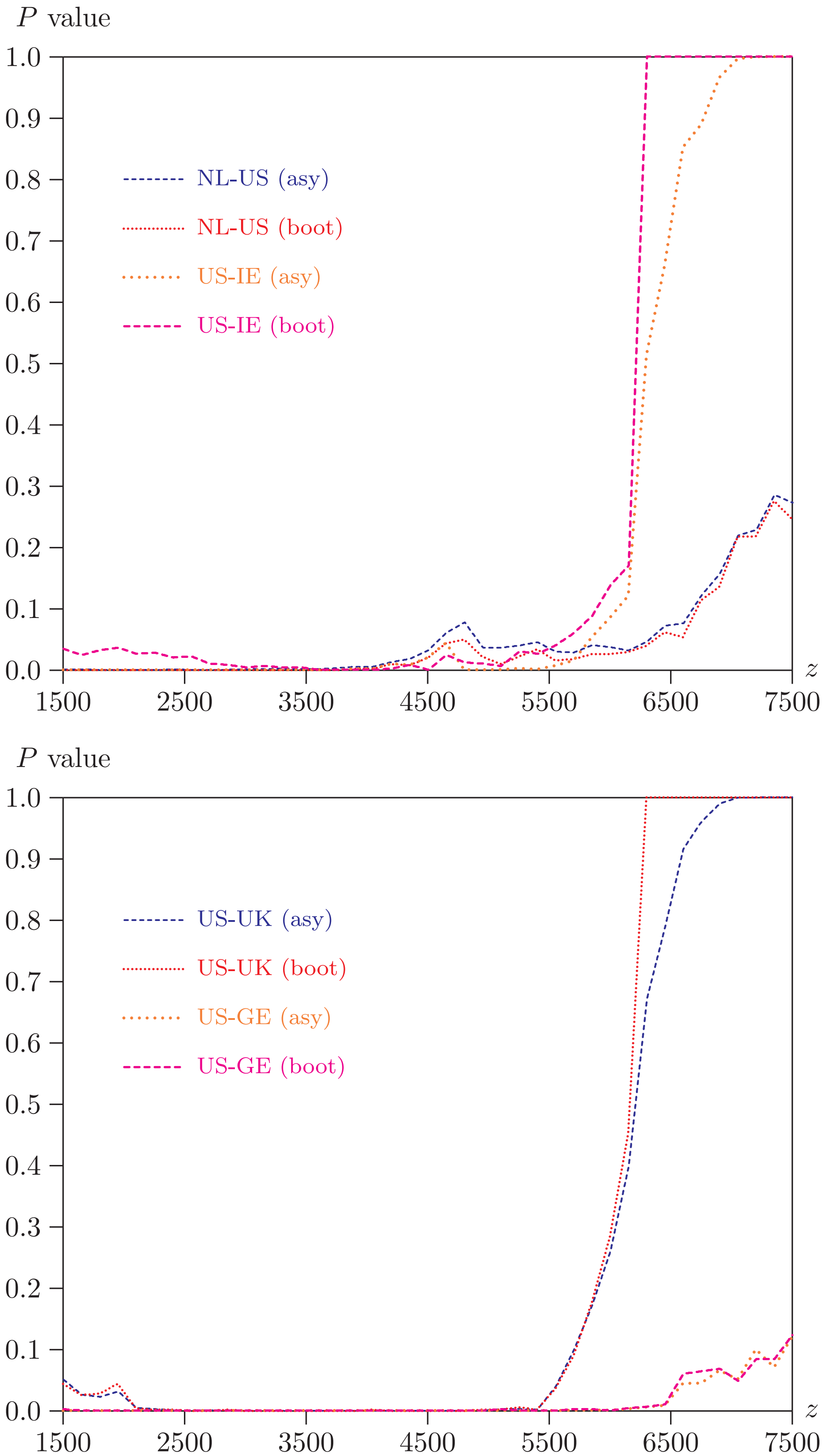


Figure 2: P values for dominance at given points

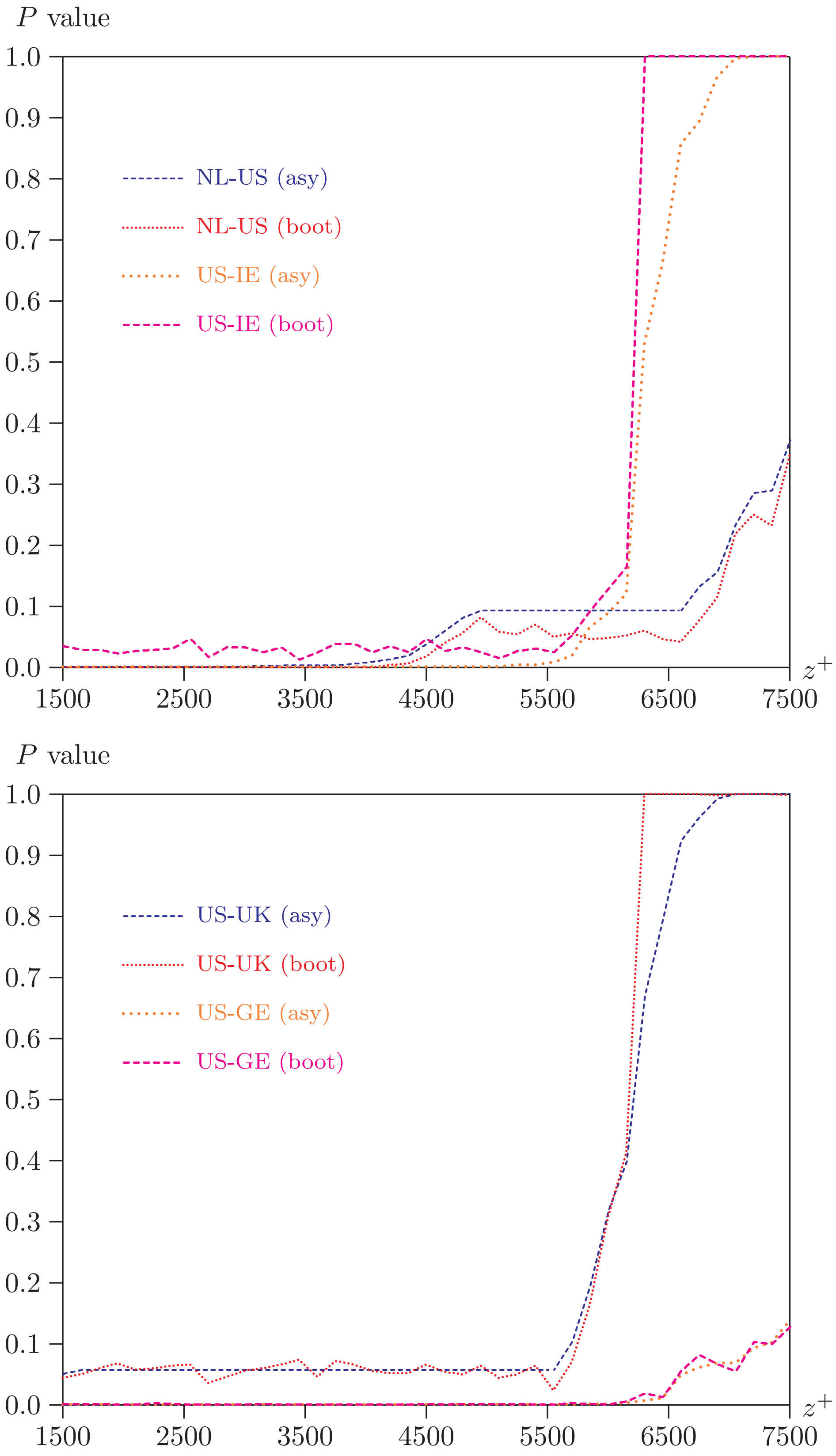


Figure 3: ***P*** **values for restricted dominance over interval**

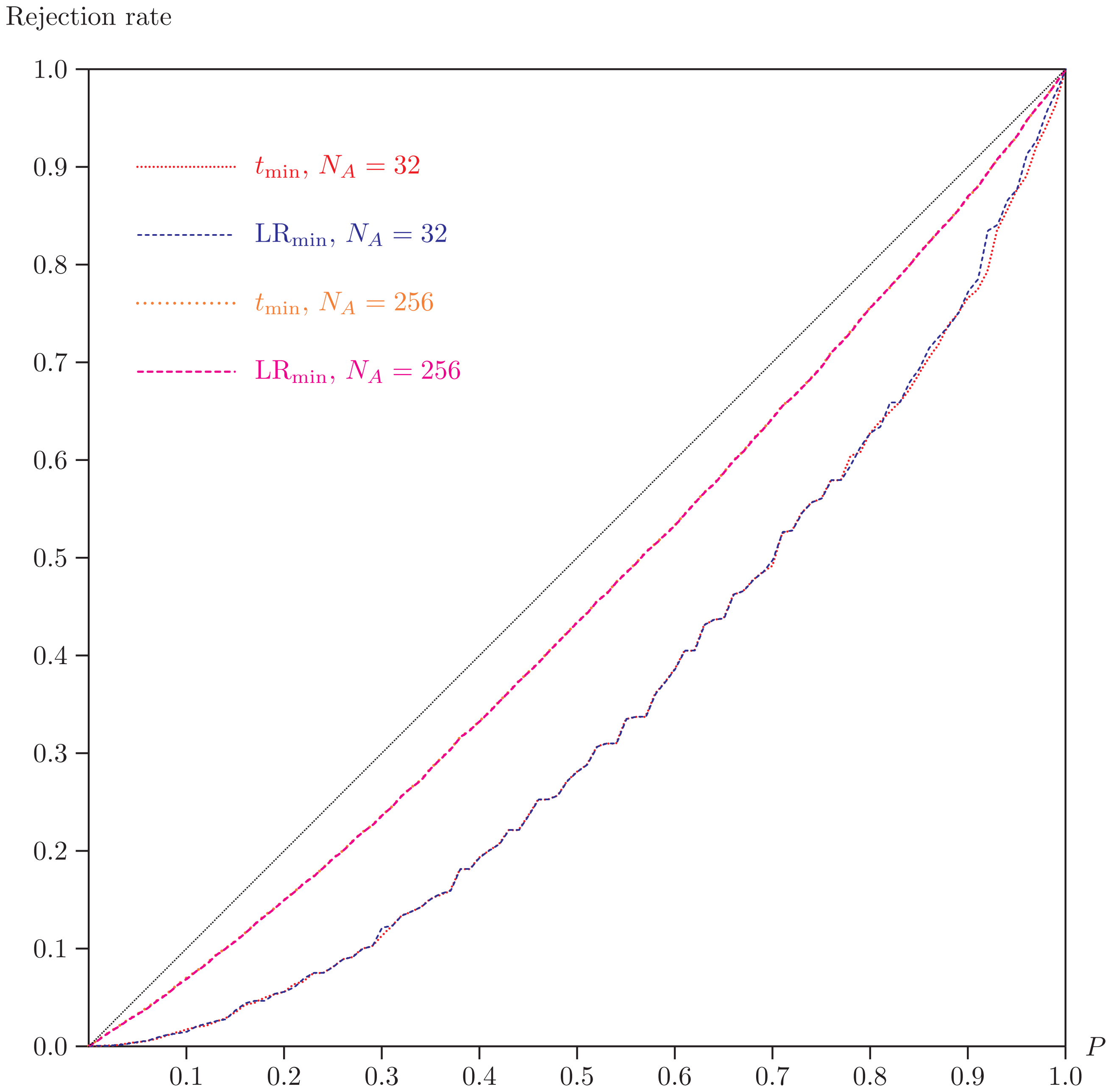


Figure 4: *P* value plots for asymptotic tests

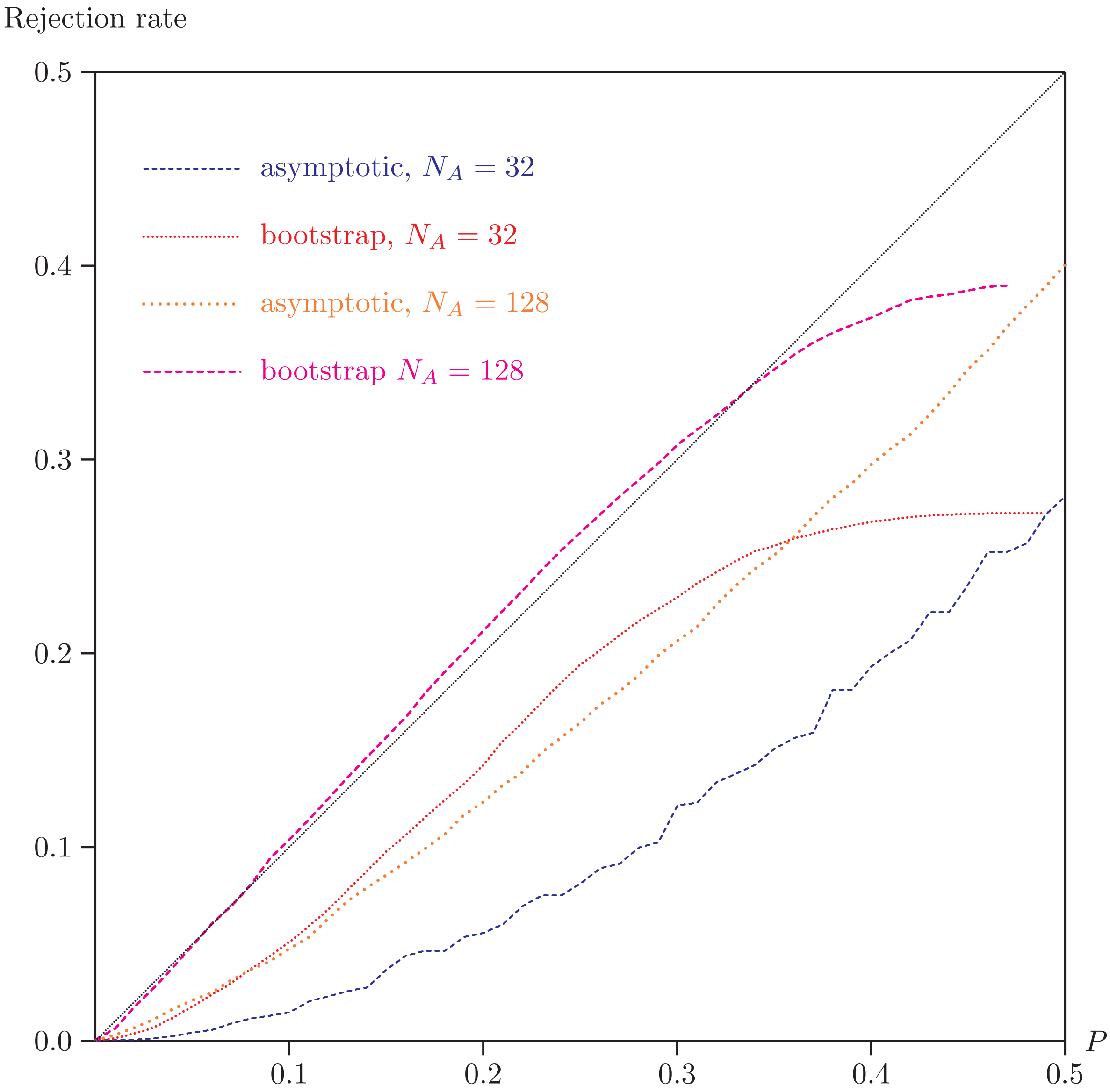


Figure 5: P value plots for asymptotic and bootstrap tests

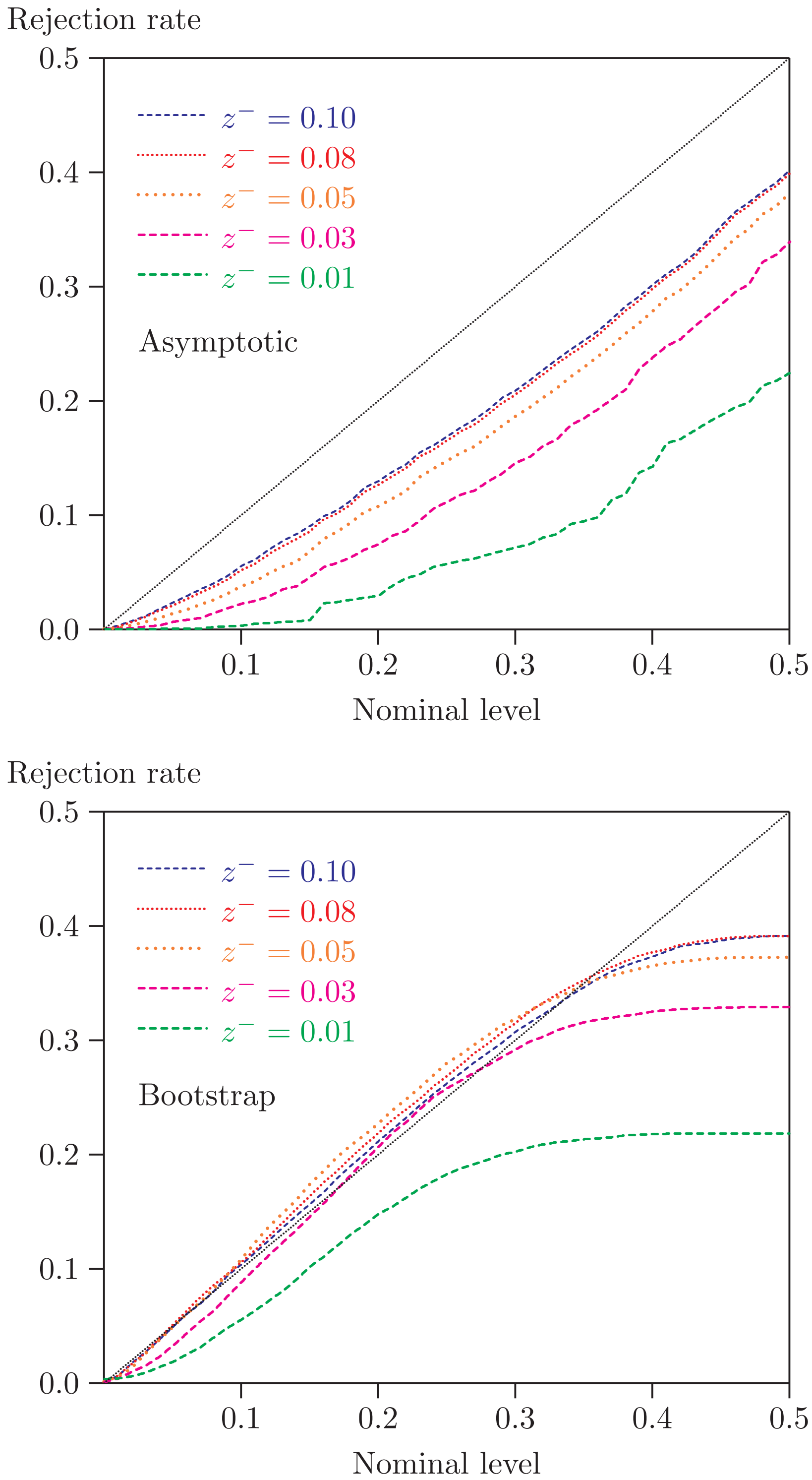


Figure 6: Varying amounts of censoring; base case, $N_A = 128$

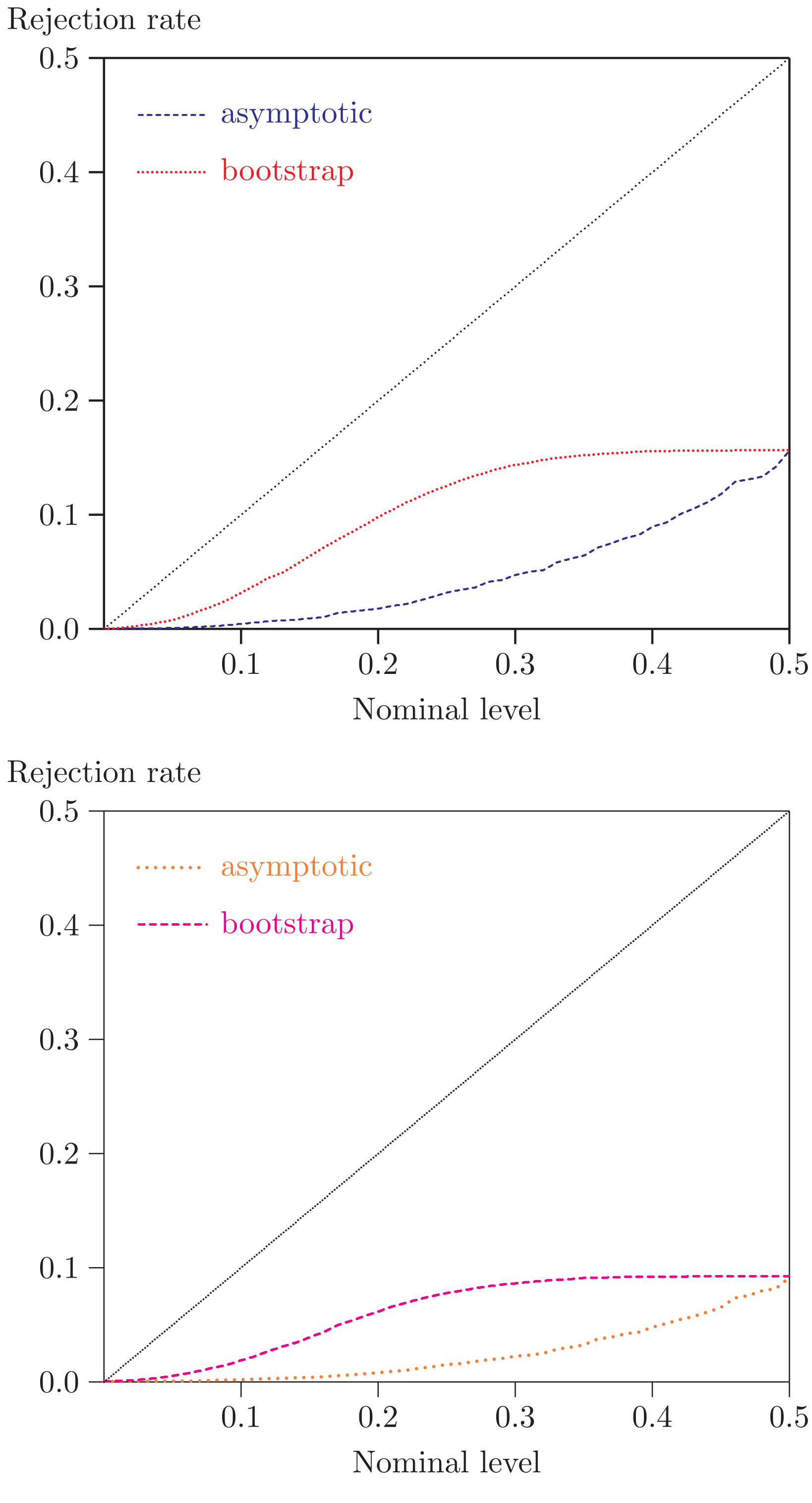
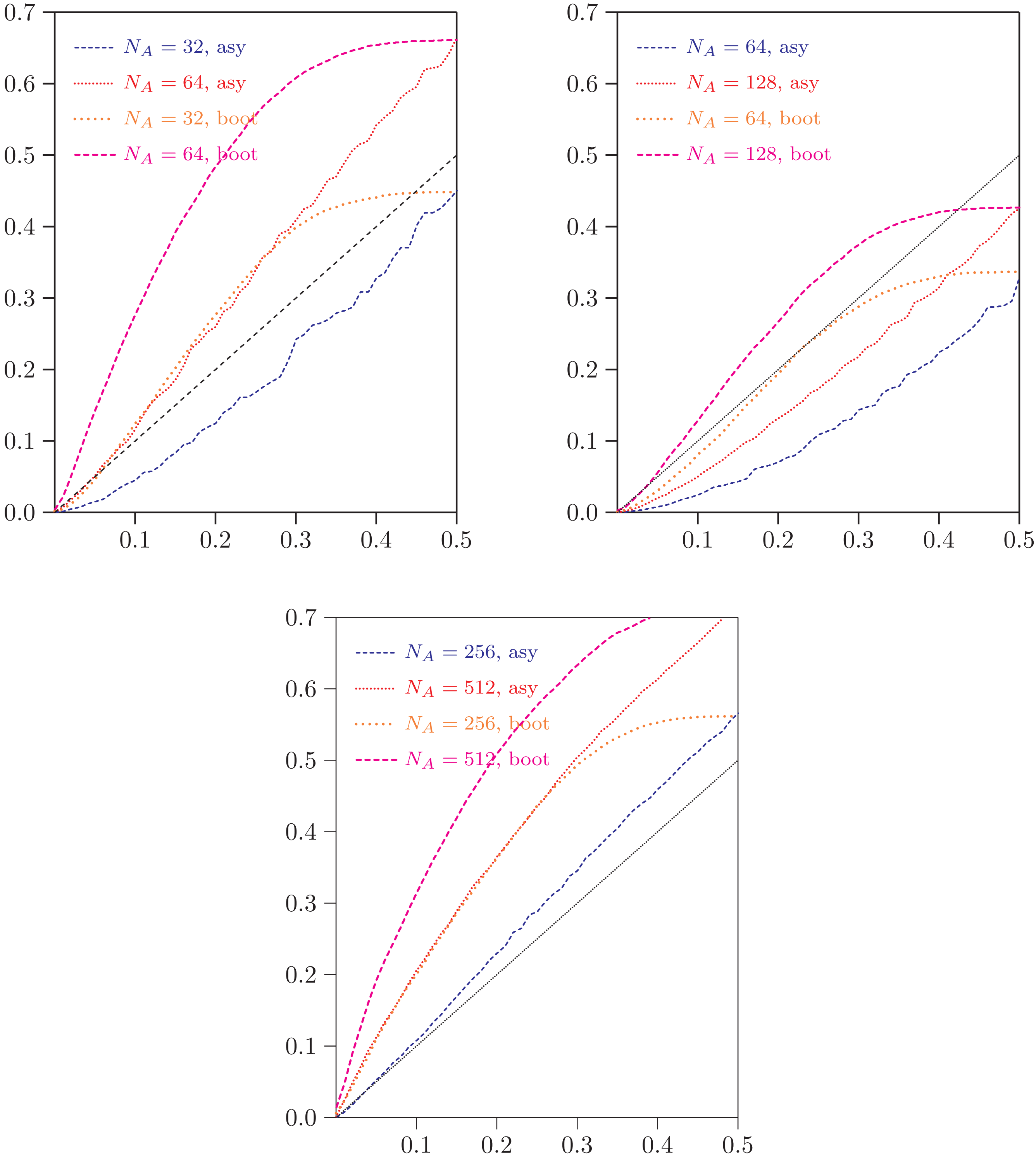


Figure 7: Alternative configurations, $N_A = 64$, $z^- = 0.1$



Rejection rate on vertical axis, nominal level on horizontal

Figure 8: Power curves, $z^{-} = 0.1$.