Tests of stochastic dominance with repeated measurements data

Angel G. Angelov[1] · Magnus Ekström[1,2]

Abstract

The paper explores a testing problem which involves four hypotheses, that is, based on observations of two random variables X and Y, we wish to discriminate between four possibilities: identical survival functions, stochastic dominance of X over Y, stochastic dominance of Y over X, or crossing survival functions. Four-decision testing procedures for repeated measurements data are proposed. The tests are based on a permutation approach and do not rely on distributional assumptions. One-sided versions of the Cramér–von Mises, Anderson–Darling, and Kolmogorov–Smirnov statistics are utilized. The consistency of the tests is proven. A simulation study shows good power properties and control of false-detection errors. The suggested tests are applied to data from a psychophysical experiment.

1 Introduction

Many research questions give rise to a two-sample problem of comparing distributions (or survival functions). Often researchers are interested in one-sided alternative hypotheses and the notion of stochastic dominance (stochastic ordering) is needed in order to formulate rigorously such hypotheses. Let X and Y be random variables with survival functions $S_X(t) = \mathbb{P}(X > t)$ and $S_Y(t) = \mathbb{P}(Y > t)$. We say that X stochastically dominates Y if

✉ Angel G. Angelov
  agangelov@gmail.com

  Magnus Ekström
  magnus.ekstrom@umu.se

1 Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden

2 Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden

$$S_X(t) \geq S_Y(t) \text{ for all } t \text{ with strict inequality for some } t.$$

We will sometimes skip the word stochastically and we will just say that X dominates Y. If X dominates Y, we write $X \succ Y$, or equivalently, $Y \prec X$. If there exist values c_1 and c_2 such that $S_X(c_1) > S_Y(c_1)$ and $S_X(c_2) < S_Y(c_2)$, we say that the survival functions of X and Y cross one another. Stochastic dominance induces four possible hypotheses: (i) X and Y have identical survival functions, (ii) X dominates Y, (iii) Y dominates X, and (iv) the survival functions of X and Y cross one another. The concept of stochastic dominance has been employed in many areas such as economics, psychology, and medicine (see, e.g., Davidson and Duclos 2000; Donald and Hsu 2016; Levy 2016; Ashby et al. 1993; Heck and Erdfelder 2016; Petroni and Wolfe 1994; Ledwina and Wyłupek 2012). As noted by Townsend (1990), stochastic dominance implies but is not implied by the same ordering of the means (if the means exist).

If we are interested in classifying the stochastic dominance relation (into the four cases specified above) based on observations of two random variables, a common procedure is the following (Whang 2019). First, perform two separate tests

$$H_{01} : S_X(t) \geq S_Y(t) \text{ for all } t, \text{ against its negation and}$$
$$H_{02} : S_Y(t) \geq S_X(t) \text{ for all } t, \text{ against its negation.}$$

Then,

(a)   if neither H_{01} nor H_{02} are rejected, decide that X and Y have identical survival functions;
(b)   if H_{01} is rejected and H_{02} is not rejected, decide that X dominates Y;
(c)   if H_{01} is not rejected and H_{02} is rejected, decide that Y dominates X;
(d)   if both H_{01} are H_{02} are rejected, decide that the survival functions of X and Y cross.

However, with this procedure it is difficult to control the possible classification errors, e.g., inferring dominance when in fact the survival functions cross (see, e.g., Whang 2019, p. 106). Bennett (2013) proposed a four-hypothesis testing procedure which allows maintaining (asymptotic) control over the various error probabilities.

Most of the existing tests of stochastic dominance assume independent observations (see, e.g., the recent monograph of Whang 2019 and the references therein). However, many experiments involve repeated measurements from each subject and such observations are not independent. For these designs, appropriate statistical methods that account for the dependence structure of the data are needed. Reducing the repeated measurements to single observations by taking their means or medians is not advisable because the available data are not efficiently used (see, e.g., Roy et al. 2019). Such transformations of the observed data will result in different estimates of the survival functions; in particular, the estimated survival function will have fewer jumps and some information will be lost.

We are not aware of a dominance test with four hypotheses which is suitable for data with repeated measurements. Building upon the ideas of Bennett (2013) and Angelov

et al. (2019b), we suggest four-decision testing procedures for repeated measurements data. In Sect. 2, we introduce the testing procedures. Section 3 reports a simulation study. In Sect. 4, the suggested procedures are applied to data from an experiment concerning the willingness to pay for a certain environmental improvement. Proofs and auxiliary results are given in the Appendix.

2 Testing procedures

Let us consider the following mutually exclusive hypotheses about the random variables X and Y:

$$
\begin{aligned}
H_0 &: \quad X \text{ and } Y \text{ have identical survival functions,} \\
H_\succ &: \quad X \text{ dominates } Y, \\
H_\prec &: \quad Y \text{ dominates } X, \\
H_{\mathrm{cr}} &: \quad \text{the survival functions of } X \text{ and } Y \text{ cross one another.}
\end{aligned}
$$

We explore a four-hypothesis testing problem with null hypothesis H_0 and three alternative hypotheses: $H_\succ$, $H_\prec$, and H_{cr}.

The survival and distribution functions of X and Y are denoted

$$
\begin{aligned}
S_X(t) &= \mathbb{P}\,(X > t), \quad F_X(t) = 1 - S_X(t), \\
S_Y(t) &= \mathbb{P}\,(Y > t), \quad F_Y(t) = 1 - S_Y(t).
\end{aligned}
$$

Throughout the paper it is assumed that we have observations

$$
\{x_{ij}, y_{ij}\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k,
$$

where x_{ij} is the observed value of X for individual/subject i at occasion j and y_{ij} is the observed value of Y for individual/subject i at occasion j, i.e., $\{x_{ij}\}$ are observations from S_X and $\{y_{ij}\}$ are observations from S_Y. We will also use the notation $\mathbf{z}_1, \ldots, \mathbf{z}_n$, where $\mathbf{z}_i = (x_{i1}, \ldots, x_{ik}, y_{i1}, \ldots, y_{ik})$, i.e., $\mathbf{z}_i$ is the vector of observations for individual/subject i. For simplicity, the observations $\{x_{ij}, y_{ij}\}$ denote random variables or values of random variables, depending on the context. Note that the vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent (and identically distributed) but the observations $(x_{i1}, \ldots, x_{ik}, y_{i1}, \ldots, y_{ik})$ within each subject i can be correlated.

The empirical distribution function based on the observations $\{x_{ij}\}$ is

$$
\widehat{F}_X(t) = (1/kn) \sum_{i,j} \mathbb{1}\{x_{ij} \le t\}
$$

and the empirical survival function is $\widehat{S}_X(t) = 1 - \widehat{F}_X(t)$. The functions $\widehat{F}_Y(t)$ and $\widehat{S}_Y(t)$ based on $\{y_{ij}\}$ are defined analogously. Let us denote $m = 2kn$, $\{t_1, \ldots, t_m\} = \{x_{ij}, y_{ij}\}$, $t_1 \le t_2 \le \ldots \le t_m$, and $z^{(+)} = \max\{z, 0\}$ for any real number z. Let $\widehat{G}(t) = (1/m) \sum_l \mathbb{1}\{t_l \le t\}$,

$$\psi_\gamma^\bullet(t_l) = \begin{cases} (\widehat{G}(t_l)[1 - \widehat{G}(t_l)])^{-1/\gamma} & \text{if } \widehat{G}(t_l) \in (0,1) \\ 0 & \text{otherwise} \end{cases}$$

$$\psi_\gamma(t_l) = \frac{\psi_\gamma^\bullet(t_l)}{\sum_l \psi_\gamma^\bullet(t_l)},$$

where $\gamma > 1$ is some real number.

We utilize the following test statistics:

- Modified one-sided Cramér–von Mises statistics

$$W_{X>Y} = \frac{1}{m} \sum_{l=1}^m \left(\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)\right)^{(+)},$$

$$W_{X<Y} = \frac{1}{m} \sum_{l=1}^m \left(\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)\right)^{(+)};$$

- Modified one-sided Anderson–Darling statistics

$$A_{X>Y}^\gamma = \sum_{l=1}^m \psi_\gamma(t_l)\left(\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)\right)^{(+)},$$

$$A_{X<Y}^\gamma = \sum_{l=1}^m \psi_\gamma(t_l)\left(\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)\right)^{(+)};$$
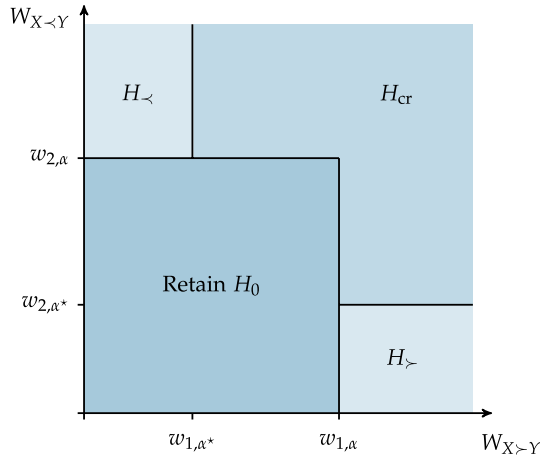
- Modified one-sided Kolmogorov–Smirnov statistics

$$D_{X>Y} = \sup_t \left(\widehat{S}_X(t) - \widehat{S}_Y(t)\right),$$

$$D_{X<Y} = \sup_t \left(\widehat{S}_Y(t) - \widehat{S}_X(t)\right).$$

Unlike the classical Cramér–von Mises statistic (see Anderson 1962), we use modified versions which do not take the squares of the differences. Our statistics are in fact one-sided versions of the statistic considered by Schmid and Trede (1995), which has shown quite similar performance as the classical Cramér–von Mises statistic for two-sample tests against general alternatives (see also Schmid and Trede 1996). The classical Anderson–Darling statistic (see Pettitt 1976) is a weighted version of the Cramér–von Mises statistic with weight $(\widehat{G}(t_l)[1 - \widehat{G}(t_l)])^{-1}$. We consider some modifications in the weight $\psi_\gamma(t_l)$ of the Anderson–Darling statistics (in Sect. 3, we investigate the properties of the corresponding tests for $\gamma = 2$ and $\gamma = 3$). We define the statistics without any normalizing factor because such factor is not needed for applying our tests.

We will describe in detail the testing procedure with the statistics $(W_{X>Y}, W_{X<Y})$; the procedures with the other test statistics are analogous. A four-hypothesis testing problem implies four decision regions defined by four critical values (see Bennett 2013; Heathcote et al. 2010). Let $w_{1,\alpha}$ and $w_{2,\alpha}$ be defined so

Fig. 1 Decision rule

that $\mathbb{P}\,(W_{X\succ Y} \geq w_{1,\alpha}\,|\,H_0) = \alpha$ and $\mathbb{P}\,(W_{X\prec Y} \geq w_{2,\alpha}\,|\,H_0) = \alpha$. Similarly, $w_{1,\alpha^\star}$ and $w_{2,\alpha^\star}$ are such that $\mathbb{P}\,(W_{X\succ Y} \geq w_{1,\alpha^\star}\,|\,H_0) = \alpha^\star$ and $\mathbb{P}\,(W_{X\prec Y} \geq w_{2,\alpha^\star}\,|\,H_0) = \alpha^\star$, where $\alpha^\star > \alpha$. We adopt the following decision rule (cf. Bennett 2013; Angelov et al. 2019b).

Decision rule 1

(a) If $W_{X\succ Y} < w_{1,\alpha}$ and $W_{X\prec Y} < w_{2,\alpha}$, then retain H_0.
(b) If $W_{X\succ Y} \geq w_{1,\alpha}$ or $W_{X\prec Y} \geq w_{2,\alpha}$, then

    (i)   if $W_{X\succ Y} \geq w_{1,\alpha}$ and $W_{X\prec Y} < w_{2,\alpha^\star}$, then accept $H_\succ$;
    (ii) if $W_{X\succ Y} < w_{1,\alpha^\star}$ and $W_{X\prec Y} \geq w_{2,\alpha}$, then accept $H_\prec$;
    (iii) if $W_{X\succ Y} \geq w_{1,\alpha^\star}$ and $W_{X\prec Y} \geq w_{2,\alpha^\star}$, then accept H_{cr}.

The decision rule is illustrated in Fig. 1. Essentially, $H_\succ$ is accepted if $W_{X\succ Y}$ is large enough and $W_{X\prec Y}$ is small enough; similarly, $H_\prec$ is accepted if $W_{X\prec Y}$ is large enough and $W_{X\succ Y}$ is small enough. The value of $\alpha^\star$ controls the discrimination between $H_\succ$, $H_\prec$, and H_{cr}. Increasing the value of $\alpha^\star$ results in larger acceptance region for H_{cr} and smaller acceptance regions for $H_\succ$ and $H_\prec$.

To obtain the critical values or the corresponding *p*-values, we employ a permutation-based approach (sometimes called randomization test approach, see Hemerik and Goeman 2021). That is, we generate random permutations of the data $\mathbf{z}_1, \ldots, \mathbf{z}_n$, calculate the value of the test statistic for each generated permutation, and then use the resulting empirical distribution of the test statistic as an approximation of the null distribution (see Hemerik and Goeman 2018; Lehmann and Romano 2005, Ch. 15; Romano 1989). A random permutation of the data is generated by randomly choosing (with probability 1/2) between $(x_{i1}, \ldots, x_{ik}, y_{i1}, \ldots, y_{ik})$ and $(y_{i1}, \ldots, y_{ik}, x_{i1}, \ldots, x_{ik})$ for each i. The algorithm is given below. Let $(w_1, w_2) = \mathrm{TS}\,(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ denote the value of $(W_{X\succ Y}, W_{X\prec Y})$ for the observed data $\mathbf{z}_1, \ldots, \mathbf{z}_n$. Similarly, $(w_1^{[r]}, w_2^{[r]}) = \mathrm{TS}\,(\mathbf{z}_1^{[r]}, \ldots, \mathbf{z}_n^{[r]})$ is the value of $(W_{X\succ Y}, W_{X\prec Y})$ for the dataset $\mathbf{z}_1^{[r]}, \ldots, \mathbf{z}_n^{[r]}$.

Algorithm 1

Input: $\mathbf{z}_1, \ldots, \mathbf{z}_n$.

for $r = 1, \ldots, R$

    for $i = 1, \ldots, n$

        Generate b_i from Bernoulli(1/2);

        if $b_i = 1$ **then** set $\mathbf{z}_i^{[r]} = (x_{i1}, \ldots, x_{ik}, y_{i1}, \ldots, y_{ik})$;

        else set $\mathbf{z}_i^{[r]} = (y_{i1}, \ldots, y_{ik}, x_{i1}, \ldots, x_{ik})$;

    end for

    Compute $(w_1^{[r]}, w_2^{[r]}) = \text{TS}(\mathbf{z}_1^{[r]}, \ldots, \mathbf{z}_n^{[r]})$;

end for

Output: $w_1^{[1]}, \ldots, w_1^{[R]}, w_2^{[1]}, \ldots, w_2^{[R]}$.

Let us define

$$p_1 = \mathbb{P}(W_{X > Y} \geq w_1 \mid H_0), \quad p_2 = \mathbb{P}(W_{X < Y} \geq w_2 \mid H_0),$$

which we call marginal *p*-values. They can be estimated as follows:

$$\widetilde{p}_1 = \frac{1 + \sum_{r=1}^{R} \mathbb{1}\{w_1^{[r]} \geq w_1\}}{R + 1}, \quad \widetilde{p}_2 = \frac{1 + \sum_{r=1}^{R} \mathbb{1}\{w_2^{[r]} \geq w_2\}}{R + 1}.$$

Then, Decision rule 1 can be expressed in terms of $\widetilde{p}_1$ and $\widetilde{p}_2$:

Decision rule 1'

(a)   If $\widetilde{p}_1 > \alpha$ and $\widetilde{p}_2 > \alpha$, then retain H_0.

(b)   If $\widetilde{p}_1 \leq \alpha$ or $\widetilde{p}_2 \leq \alpha$, then

        (i)   if $\widetilde{p}_1 \leq \alpha$ and $\widetilde{p}_2 > \alpha^\star$, then accept $H_>$;

        (ii)   if $\widetilde{p}_1 > \alpha^\star$ and $\widetilde{p}_2 \leq \alpha$, then accept $H_<$;

        (iii)   if $\widetilde{p}_1 \leq \alpha^\star$ and $\widetilde{p}_2 \leq \alpha^\star$, then accept H_{cr}.

It should be noted that borderline cases may occur when the test statistic is close to the border of the decision region (respectively, a marginal *p*-value is close to one of the thresholds α and $\alpha^\star$). Therefore, it is advisable to report the conclusion of the test together with the marginal *p*-values $\widetilde{p}_1, \widetilde{p}_2$ and the thresholds $\alpha, \alpha^\star$ (see Angelov et al. 2019b).

In a testing problem involving just a null hypothesis (the hypothesis of no difference) and an alternative hypothesis (the hypothesis of interest), the event of wrongly accepting the alternative hypothesis is called Type I error, while the event of not accepting the alternative when it is true is called Type II error. In our setting, if $H_>$ is the hypothesis of interest, false detection of $H_>$ (wrongly accepting $H_>$) and non-detection of $H_>$ (not accepting the true $H_>$) can be viewed as analogues of Type I error and Type II error, respectively.

Let FDP be the probability of a false detection of dominance ($H_>$) and let NDP be the probability of a non-detection of dominance ($H_>$). These probabilities can be expressed as follows:

$$\text{FDP} = \mathbb{P}(\text{ accept } H_> \mid H_0) + \mathbb{P}(\text{ accept } H_> \mid H_{\text{cr}}) + \mathbb{P}(\text{ accept } H_> \mid H_<)$$
$$= \text{FDP}_1 + \text{FDP}_2 + \text{FDP}_3,$$
$$\text{NDP} = \mathbb{P}(\text{ retain } H_0 \mid H_>) + \mathbb{P}(\text{ accept } H_{\text{cr}} \mid H_>) + \mathbb{P}(\text{ accept } H_< \mid H_>)$$
$$= \text{NDP}_1 + \text{NDP}_2 + \text{NDP}_3.$$

The power to detect dominance ($H_>$) is defined as $\mathbb{P}(\text{ accept } H_> \mid H_>) = 1 - \text{NDP}$.

Let U_n be a generic notation for the test statistics defined above.

Assumption 1 There exist a nonrandom sequence τ_n and a nondegenerate random variable U such that $\tau_n \longrightarrow \infty$ and under the null hypothesis $\tau_n U_n$ converges in distribution to U as $n \longrightarrow \infty$.

Assumption 2 The distribution function of U is continuous and strictly increasing at u_α, where $\mathbb{P}(U \geq u_\alpha \mid H_0) = \alpha$.

Assumption 3 The distribution functions F_X and F_Y are continuous.

Assumptions similar to Assumption 1 are common in the literature on subsampling (see, e.g., Politis et al. 1999). Assumption 3 is needed only for the Anderson–Darling test, where it is used for showing that $m\psi_\gamma(t_l)$, $l < m$, is asymptotically bounded away from zero (almost surely).

Some results concerning the error probabilities FDP and NDP are established in the following theorems.

Theorem 1 *Suppose that Assumptions 1 and 2 are satisfied. Then the following are true for the proposed Cramér–von Mises test.*

(a) $\text{FDP}_1 \leq \alpha$.
(b) $\text{FDP}_2 + \text{FDP}_3 \longrightarrow 0$ *as* $n \longrightarrow \infty$.
(c) $\text{NDP}_1 + \text{NDP}_2 + \text{NDP}_3 \longrightarrow 0$ *as* $n \longrightarrow \infty$.

Theorem 2 *Suppose that Assumptions 1, 2, and 3 are satisfied. Then the following are true for the proposed Anderson–Darling test.*

(a) $\text{FDP}_1 \leq \alpha$.
(b) $\text{FDP}_2 + \text{FDP}_3 \longrightarrow 0$ *as* $n \longrightarrow \infty$.
(c) $\text{NDP}_1 + \text{NDP}_2 + \text{NDP}_3 \longrightarrow 0$ *as* $n \longrightarrow \infty$.

Theorem 3 *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then the following are true for the proposed Kolmogorov–Smirnov test.*

(*a*) $\text{FDP}_1 \leq \alpha$.
(*b*) $\text{FDP}_2 + \text{FDP}_3 \longrightarrow 0 \text{ as } n \longrightarrow \infty$.
(*c*) $\text{NDP}_1 + \text{NDP}_2 + \text{NDP}_3 \longrightarrow 0 \text{ as } n \longrightarrow \infty$.

3 Simulation study

3.1 Setup

We conducted simulations to examine the behavior of the suggested tests in terms of false detection of dominance and power to detect dominance.

Let $\mathbf{Z} = (X_1, \ldots, X_k, Y_1, \ldots, Y_k)$, where each X_j, $j = 1, \ldots, k$, is distributed like X and each Y_j, $j = 1, \ldots, k$, is distributed like Y. Let $\boldsymbol{\mu} = (\mu_X, \ldots, \mu_X, \mu_Y, \ldots, \mu_Y)$, $\boldsymbol{\Sigma}_{XY}$ be a $k \times k$ matrix with entries $\rho_{XY}\,\sigma_X\sigma_Y$,

$$\boldsymbol{\Sigma}_X = \begin{pmatrix} \sigma_X^2 & \rho_{12}\,\sigma_X^2 & \cdots & \rho_{1k}\,\sigma_X^2 \\ \rho_{12}\,\sigma_X^2 & \sigma_X^2 & \cdots & \rho_{2k}\,\sigma_X^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\,\sigma_X^2 & \rho_{2k}\,\sigma_X^2 & \cdots & \sigma_X^2 \end{pmatrix}, \quad \boldsymbol{\Sigma}_Y = \begin{pmatrix} \sigma_Y^2 & \rho_{12}\,\sigma_Y^2 & \cdots & \rho_{1k}\,\sigma_Y^2 \\ \rho_{12}\,\sigma_Y^2 & \sigma_Y^2 & \cdots & \rho_{2k}\,\sigma_Y^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1k}\,\sigma_Y^2 & \rho_{2k}\,\sigma_Y^2 & \cdots & \sigma_Y^2 \end{pmatrix},$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_X & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_Y \end{pmatrix}.$$

Let $N(\mu, \sigma)$ and $La(\mu, \sigma)$ denote, respectively, normal distribution and Laplace distribution with mean μ and standard deviation σ, while $LN(\mu, \sigma)$ denotes lognormal distribution with parameters μ and σ such that $X \sim LN(\mu, \sigma) \Longleftrightarrow \log(X) \sim N(\mu, \sigma)$.

We generated data from the following distributions:

(a) Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{Z} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
(b) Multivariate lognormal distribution, $\mathbf{Z} \sim \text{MVLN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longleftrightarrow \log(\mathbf{Z}) \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
(c) Multivariate Laplace distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\mathbf{Z} \sim \text{MVLa}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, see, e.g., Kotz et al. (2001).

Figure 2 depicts survival functions corresponding to some of the scenarios in the simulations. For generating random numbers from the multivariate normal and the multivariate lognormal, the R package MASS was used (see Venables and Ripley 2002), while for the multivariate Laplace distribution, we used the R package LaplacesDemon (see Statisticat 2018).

All computations were performed with (see R Core Team 2019). The R code can be obtained from the corresponding author upon request. The results are based on 3000 simulated datasets under each setting; the number of generated random
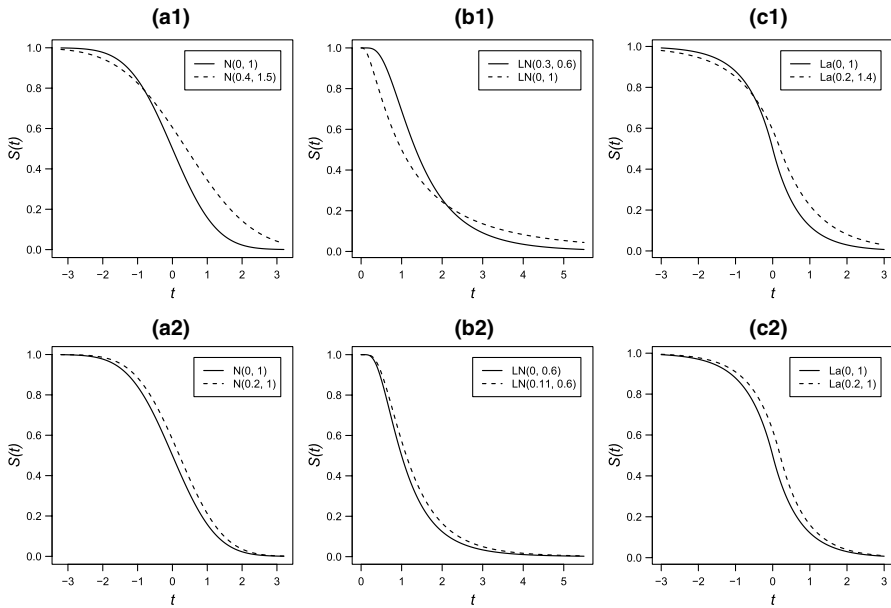
Fig. 2 Survival functions

permutations for each dataset is $R = 4000$, $\alpha = 0.05$, and $\alpha^\star = 0.96$ (cf. Angelov et al. 2019b).

3.2 Results

Let CvM, AD2, AD3, and KS denote, respectively, the Cramér–von Mises test, the Anderson–Darling test with $\gamma = 2$, the Anderson–Darling test with $\gamma = 3$, and the Kolmogorov–Smirnov test.

Simulation results concerning false detection of dominance when the truth is H_0 are presented in Fig. 3. For all tests and all sample sizes, the probability of a false detection (FDP$_1$) is less than $\alpha = 0.05$ (in most cases, it is even not greater than $\alpha/2 = 0.025$). One should not forget that under H_0 three types of erroneous decisions may occur: accepting $H_>$, accepting $H_<$, and accepting H_{cr}. The corresponding error probabilities add up to $\mathbb{P}$ (reject $H_0 \mid H_0$), which is not greater than 2α.

Figure 4 depicts the probability of a false detection of dominance when the truth is H_{cr}. The probability of a false detection (FDP$_2$) tends to zero as the sample size increases. For smaller sample sizes, FDP$_2$ is smallest for the Anderson–Darling test with $\gamma = 2$, followed by the Anderson–Darling test with $\gamma = 3$, the Cramér–von Mises test, and the Kolmogorov–Smirnov test.

Power curves for $n = 70$ are shown in Fig. 5, where the power to detect dominance is plotted against $\delta = \mu_Y - \mu_X$. We see that the power gets closer to one as δ increases. Overall, the Cramér–von Mises test is the most powerful. For $\rho_{XY} = 0.8$,
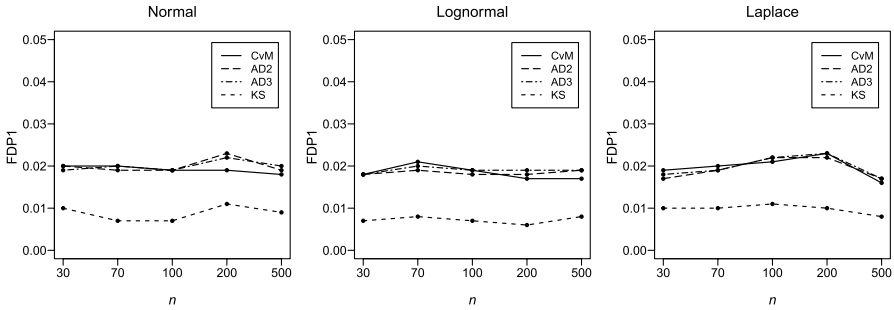
Fig. 3 Probability of a false detection of dominance when the truth is H_0. Results for different sample sizes with $k = 3$, $\rho_{XY} = \rho_{12} = \rho_{13} = \rho_{23} = 0.8$
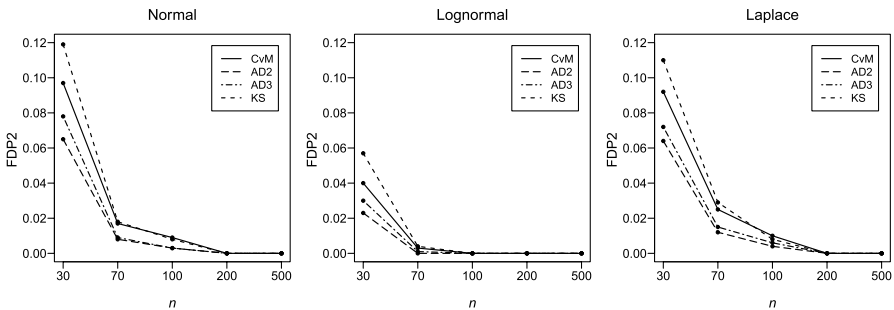

Fig. 4 Probability of a false detection of dominance when the truth is H_{cr}. Results for different sample sizes with $k = 3$, $\rho_{XY} = \rho_{12} = \rho_{13} = \rho_{23} = 0.8$, under the settings depicted in Fig. 2 (first row)

the Kolmogorov–Smirnov test has the lowest power and Anderson–Darling tests have quite similar performance as the Cramér–von Mises test. For $\rho_{XY} = 0.2$, the four tests do not differ that much in terms of power.

Let us consider the following scenarios for the correlation structure:

Scenario (3e) $k = 3$, $\rho_{12} = \rho_{23} = 0.5$, $\rho_{13} = 0.5$, $\rho_{XY} = 0.5$;
Scenario (3ar) $k = 3$, $\rho_{12} = \rho_{23} = 0.62$, $\rho_{13} = 0.38$, $\rho_{XY} = 0.5$;
Scenario (2e) $k = 2$, $\rho_{12} = 0.5$, $\rho_{XY} = 0.5$;
Scenario (2ar) $k = 2$, $\rho_{12} = 0.62$, $\rho_{XY} = 0.5$.

In Scenario (3e), all correlations are equal to 0.5, while in Scenario (3ar), ρ_{12}, ρ_{23}, and ρ_{13} are in accordance with an autoregressive process of order one. Scenarios (2e) and (2ar) are defined in analogy to (3e) and (3ar) but with $k = 2$. We performed simulations to investigate the power to accept a fixed hypothesis of dominance for different sample sizes, under the scenarios specified above. The results are illustrated in Fig. 6. The power approaches one as the sample size increases. The Cramér–von Mises test has the highest power. The Anderson–Darling test with $\gamma = 3$ is slightly less powerful. For the normal and the lognormal
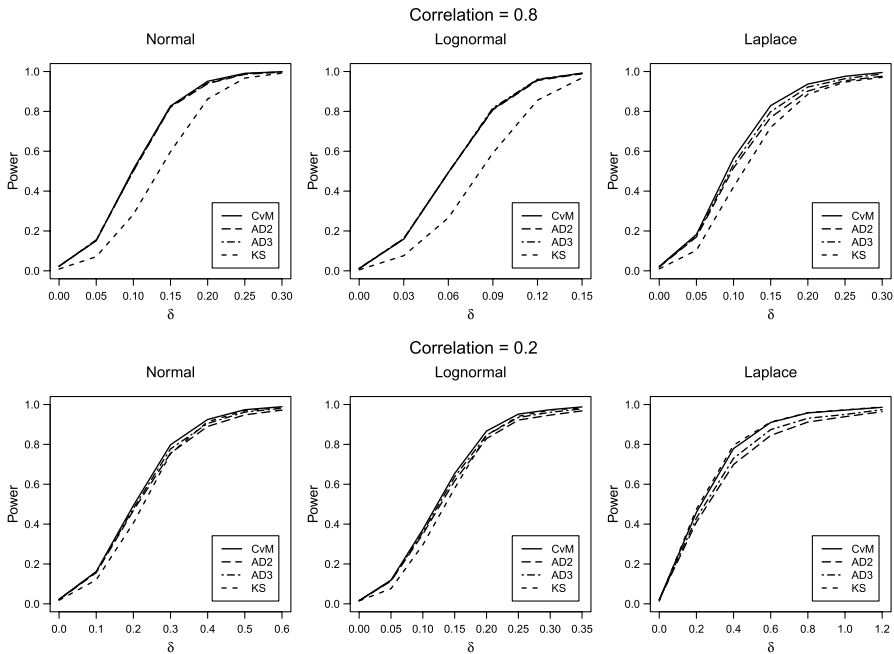
Fig. 5 Power to detect dominance. Results for different values of δ with $n = 70$, $k = 3$; for all distributions $\mu_X = 0$, for the normal and the Laplace $\sigma_X = \sigma_Y = 1$, for the lognormal $\sigma_X = \sigma_Y = 0.6$. In the first row, $\rho_{XY} = \rho_{12} = \rho_{13} = \rho_{23} = 0.8$, while in the second row, $\rho_{XY} = \rho_{12} = \rho_{13} = \rho_{23} = 0.2$

distributions, the Kolmogorov–Smirnov test has the lowest power. For the Laplace distribution, the Anderson–Darling test with $\gamma = 2$ is the least powerful.
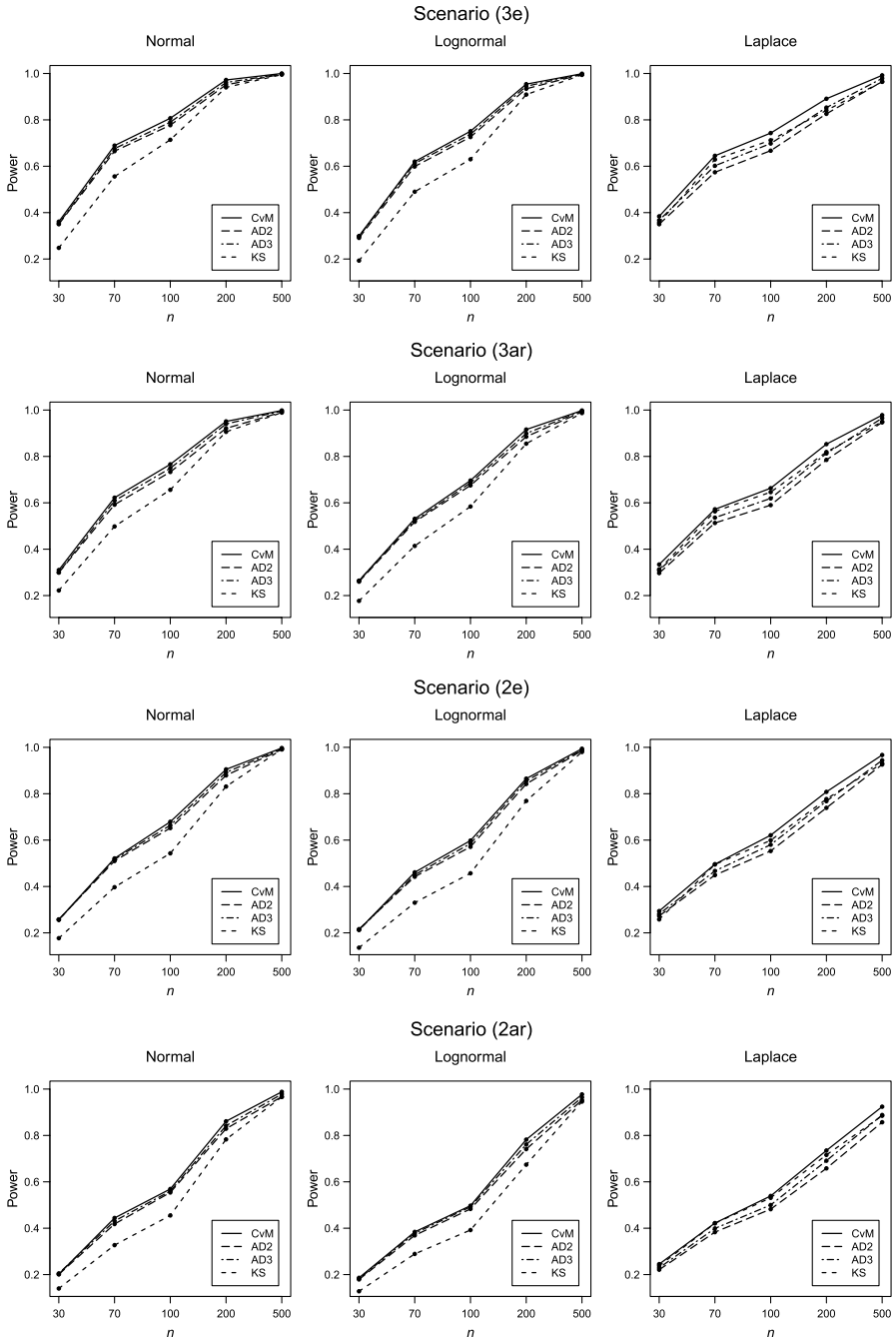
The results for the Cramér–von Mises test under the four scenarios are presented in Fig. 7. We see that the power is higher for $k = 3$ than for $k = 2$. Also, for each k, the scenario where all correlations are equal to 0.5 leads to higher power than the other scenario. In order to further investigate how the correlations between $X_1, \ldots, X_k$ (respectively, $Y_1, \ldots, Y_k$) affect power, we considered the following scenarios:

**Scenario (3w)** $k = 3$, $\rho_{12} = \rho_{23} = 0.38$, $\rho_{13} = 0.38$, $\rho_{XY} = 0.5$;
Scenario (3s) $k = 3$, $\rho_{12} = \rho_{23} = 0.62$, $\rho_{13} = 0.62$, $\rho_{XY} = 0.5$.

The correlations ρ_{12}, ρ_{23}, and ρ_{13} are 'weak' in Scenario (3w) and 'strong' in Scenario (3s). The results show that the power is higher when the correlations between $X_1, \ldots, X_k$ (respectively, $Y_1, \ldots, Y_k$) are weaker (see Figs. 8 and 9).

In summary, the Cramér–von Mises test is the most powerful. The Anderson–Darling test with $\gamma = 3$ is slightly less powerful but has lower probability of a false detection of dominance for small sample sizes compared with the Cramér–von Mises test.

Fig. 6 Power to detect dominance. Results for different sample sizes under Scenarios (3e), (3ar), (2e), (2ar). The underlying survival functions are shown in Figure 2 (second row)
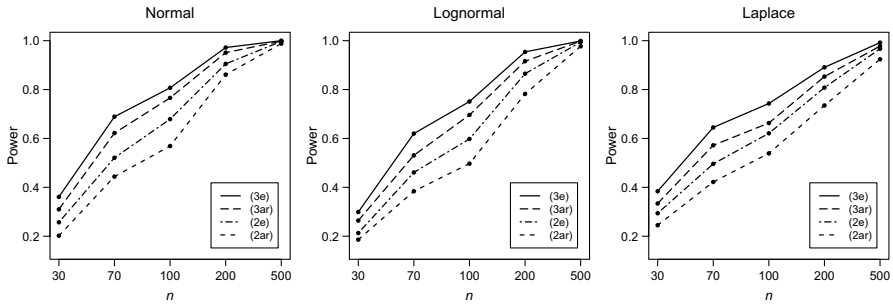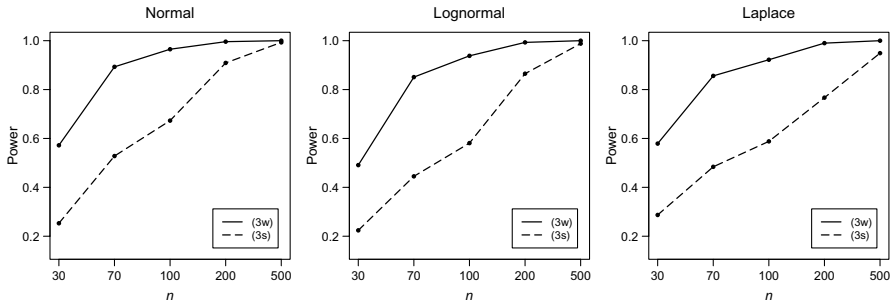
Fig. 7 Power to detect dominance. Results for the Cramér–von Mises test under Scenarios (3e), (3ar), (2e), (2ar)
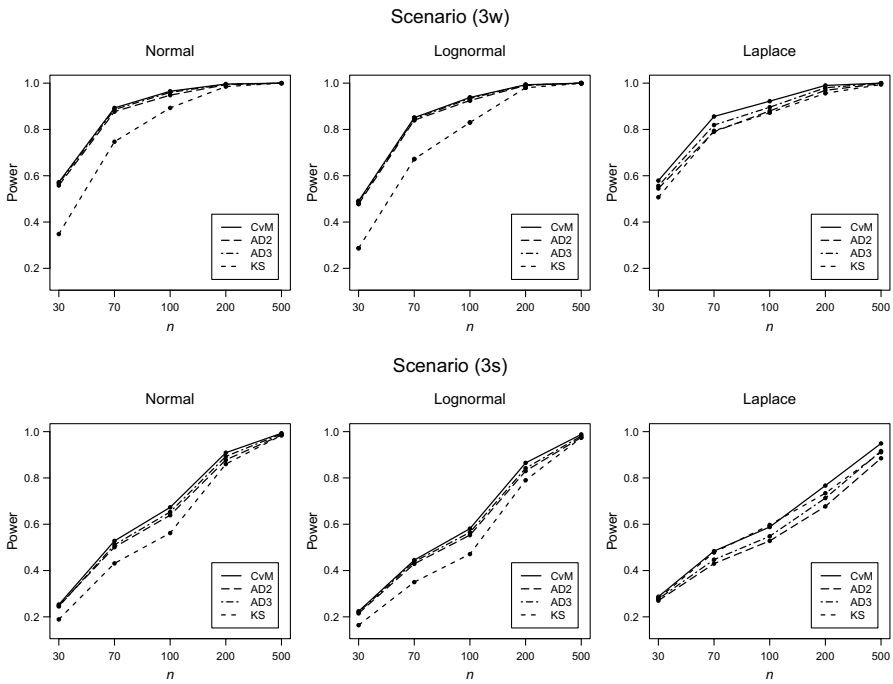
Fig. 8 Power to detect dominance. Results for different sample sizes under Scenarios (3w), (3s). The underlying survival functions are shown in Figure 2 (second row)

4 Real data example

We apply the proposed tests to data from an experiment where participants were asked about their willingness to pay for an improved outdoor sound environment. The dataset is available at Mendeley Data (Angelov et al. 2019a). In a sound

Fig. 9 Power to detect dominance. Results for the Cramér–von Mises test under Scenarios (3w), (3s)

laboratory, the participants listened to recordings of outdoor sound environments and had to imagine that each recording was the noise they hear while sitting on their balcony. They were asked how much they would be willing to pay for a noise reduction that would change a given sound environment with road-traffic noise to an environment without the road traffic noise. Each participant was requested to answer by means of: (i) a self-selected point (SSP), i.e., the amount in Swedish kronor he/she would be willing to pay per month for the improvement, and (ii) a self-selected interval (SSI), i.e., the lowest and highest amounts he/she would be willing to pay. The experiment included five main scenarios (referred to as Scenario 1, 2, 3, 4, and 5) with systematically increasing noise levels: Scenario 1 corresponds to the smallest noise reduction, while Scenario 5 corresponds to the largest. Each participant gave answers for each scenario four times: at two SSP sessions and two SSI sessions.

The following variables are of main interest:

- `pt` is the point answer at the first or the second SSP session.
- `low` and `upp` are, respectively, the lower bound and the upper bound of the interval answered at the first or the second SSI session.
- `mid` is the midpoint of the interval answered at the first or the second SSI session.

Each variable was observed under the five scenarios and these are denoted, e.g., $pt[1], \ldots, pt[5]$.

Our analysis is based on $n = 59$ participants (just as in Angelov et al. 2019b), $\alpha = 0.05$, $\alpha^\star = 0.96$, and $R = 20000$.

We are interested in whether the survival function of SSP lies between the survival functions of the lower and the upper bounds of SSI. The conducted dominance tests confirm this in most cases (see Table 1).

We also want to find out whether the respondents are willing to pay more for higher levels of noise reduction. This implies that the willingness to pay under Scenario 2 stochastically dominates the willingness to pay under Scenario 1; similarly, the willingness to pay under Scenario 3 stochastically dominates the willingness to pay under Scenario 2, and so on. The empirical survival functions for each consecutive pair of scenarios are displayed in Fig. 10. We conducted dominance tests (see Table 2) and in

Table 1 Comparison of self-selected points and self-selected intervals. Conclusion of the test together with the marginal p-values $\widetilde{p}_1$ and $\widetilde{p}_2$

X	Y	CvM	AD2	AD3	KS
low[1]	pt[1]	Retain H_0	Accept $H_<$	Retain H_0	Retain H_0
		1.0000 0.0665	1.0000 0.0427	1.0000 0.0515	1.0000 0.1210
pt[1]	upp[1]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0000	1.0000 0.0000	1.0000 0.0000	1.0000 0.0006
low[2]	pt[2]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0213	1.0000 0.0112	1.0000 0.0140	1.0000 0.0220
pt[2]	upp[2]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0002	1.0000 0.0002	1.0000 0.0002	1.0000 0.0000
low[3]	pt[3]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0004	1.0000 0.0001	1.0000 0.0002	1.0000 0.0019
pt[3]	upp[3]	Accept H_{cr}	Accept H_{cr}	Accept H_{cr}	Accept H_{cr}
		0.9288 0.0005	0.8938 0.0006	0.9116 0.0006	0.9311 0.0002
low[4]	pt[4]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0003	1.0000 0.0001	1.0000 0.0001	1.0000 0.0001
pt[4]	upp[4]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		0.9855 0.0000	0.9808 0.0000	0.9827 0.0000	0.9932 0.0001
low[5]	pt[5]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0001	1.0000 0.0000	1.0000 0.0000	1.0000 0.0157
pt[5]	upp[5]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		0.9805 0.0000	0.9707 0.0000	0.9757 0.0000	0.9921 0.0015

Table 2 Comparison of willingness to pay for different levels of noise reduction. Conclusion of the test together with the marginal p-values $\widetilde{p}_1$ and $\widetilde{p}_2$

| X	Y	CvM	AD2	AD3	KS
mid[1]	mid[2]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0000	1.0000 0.0000	1.0000 0.0000	1.0000 0.0030
mid[2]	mid[3]	Accept H_{cr}	Accept H_{cr}	Accept H_{cr}	Accept $H_<$
		0.9480 0.0000	0.9225 0.0000	0.9360 0.0000	0.9876 0.0000
mid[3]	mid[4]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0000	1.0000 0.0000	1.0000 0.0000	1.0000 0.0000
mid[4]	mid[5]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		0.9923 0.0002	0.9915 0.0001	0.9917 0.0001	1.0000 0.0131
pt[1]	pt[2]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0000	1.0000 0.0000	1.0000 0.0000	1.0000 0.0003
pt[2]	pt[3]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0000	1.0000 0.0000	1.0000 0.0000	1.0000 0.0000
pt[3]	pt[4]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		1.0000 0.0002	1.0000 0.0002	1.0000 0.0002	1.0000 0.0007
pt[4]	pt[5]	Accept $H_<$	Accept $H_<$	Accept $H_<$	Accept $H_<$
		0.9844 0.0000	0.9859 0.0000	0.9852 0.0000	0.9953 0.0001
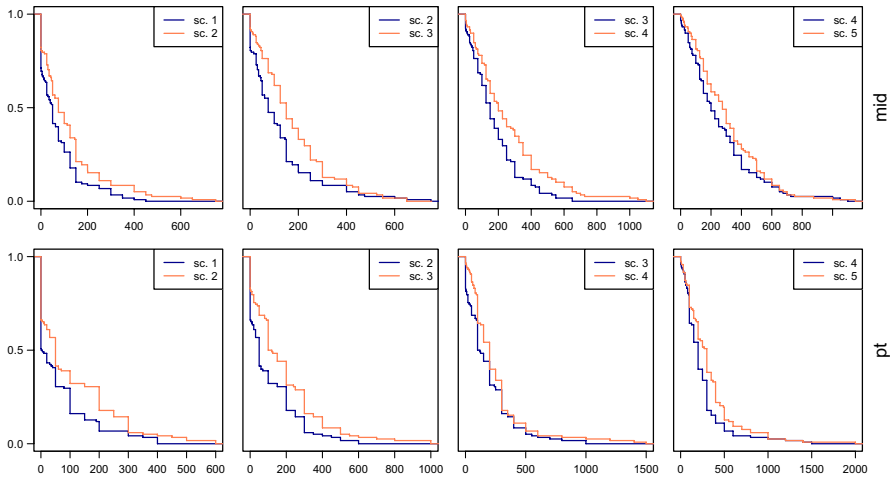
Fig. 10 Empirical survival functions

most cases the tests conclude that the willingness to pay for the higher level of noise reduction dominates the willingness to pay for the lower level.

The four tests in most cases led to the same conclusion. In Angelov et al. (2019b), analogous tests were performed but separately for the first and the second session, while here we apply the new tests for repeated measurements. If we compare the results, we see that the hypothesis of dominance is accepted slightly more often with the new testing procedures, but overall the conclusions are to a great extent similar.

5 Concluding remarks

We proposed permutation-based four-decision tests of stochastic dominance for repeated measurements data. We proved under certain regularity conditions that as the sample size increases, the probability to detect dominance tends to one and the probability of a false detection of dominance does not exceed a pre-specified level. Our simulations indicated good performance of the testing procedures for a range of sample sizes.

Appendix

Lemma 1 $\sup_t |\widehat{F}_X(t) - F_X(t)| \xrightarrow{\text{a.s.}} 0$ *as* $n \longrightarrow \infty$.

Proof Follows directly from the result of Tucker (1959).   □

Lemma 2 $\sup_t(\widehat{S}_X(t) - \widehat{S}_Y(t)) \xrightarrow{\text{a.s.}} \sup_t \left(S_X(t) - S_Y(t)\right)$ *as* $n \longrightarrow \infty$.

Proof The claim is equivalent to $\sup_t(\widehat{F}_Y(t) - \widehat{F}_X(t)) \xrightarrow{a.s.} \sup_t(F_Y(t) - F_X(t))$. Using some trivial inequalities and Lemma 1, we obtain

$$\left| \sup_t(\widehat{F}_Y(t) - \widehat{F}_X(t)) - \sup_t(F_Y(t) - F_X(t)) \right|$$

$$\leq \sup_t \left| \widehat{F}_Y(t) - \widehat{F}_X(t) - F_Y(t) + F_X(t) \right|$$

$$\leq \sup_t \left| \widehat{F}_X(t) - F_X(t) \right| + \sup_t \left| \widehat{F}_Y(t) - F_Y(t) \right| \xrightarrow{a.s.} 0 \quad \text{as } n \longrightarrow \infty.$$

$\square$

Lemma 3 *The following hold true.*

(a) *If $H_>$ is true, then, with probability one,*

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)} > 0$$

(b) *If H_{cr} is true, then, with probability one,*

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)} > 0$$

and

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)]^{(+)} > 0.$$

Proof Part (a). Let $\varepsilon > 0$ and $\mathcal{A}_\varepsilon = \{t : S_X(t) - S_Y(t) > \varepsilon\}$. If $H_>$ is true, $\mathcal{A}_\varepsilon$ is non-empty for small enough ε and $\mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0$. We have

$$\frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)}$$

$$= \frac{1}{m} \sum_{l=1}^{m} \left[[(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))] + S_X(t_l) - S_Y(t_l) \right]^{(+)}.$$

By Lemma 1, $\sup_t |(\widehat{S}_X(t) - S_X(t)) - (\widehat{S}_Y(t) - S_Y(t))| \longrightarrow 0$ as $n \longrightarrow \infty$. Thus, for every $\varepsilon > 0$, there exists a sufficiently large (random) N such that with probability one,

$$\max_l |(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))| < \frac{\varepsilon}{2} \quad \text{for all } n > N.$$

That is, for small enough ε, for all $n > N$, and for each $t_l \in \mathcal{A}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))] + S_X(t_l) - S_Y(t_l) \right] > \frac{\varepsilon}{2}.$$

Recall that $m = 2kn$. Then, by the strong law of large numbers, with probability one,

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)}$$

$$\geq \liminf_{n\to\infty} \frac{1}{2k} \frac{1}{n} \sum_{i=1}^{n} [\widehat{S}_X(y_{i1}) - \widehat{S}_Y(y_{i1})]^{(+)}$$

$$\geq \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{1}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_{i1} \in \mathcal{A}_\varepsilon\} = \frac{\varepsilon}{4k} \mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0,$$

which was to be shown.

Part (b). Let $\varepsilon > 0$, $\mathcal{A}_\varepsilon = \{t : S_X(t) - S_Y(t) > \varepsilon\}$ and $\mathcal{B}_\varepsilon = \{t : S_Y(t) - S_X(t) > \varepsilon\}$. If H_{cr} is true, $\mathcal{A}_\varepsilon$ and $\mathcal{B}_\varepsilon$ are nonempty for small enough ε and $\mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0$, $\mathbb{P}(x_{i1} \in \mathcal{B}_\varepsilon) > 0$. Using similar arguments as in Part (a), we get that for small enough ε, there exists a sufficiently large (random) N_1 such that for all $n > N_1$ and for each $t_l \in \mathcal{A}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))] + S_X(t_l) - S_Y(t_l) \right] > \frac{\varepsilon}{2}.$$

Similarly, for small enough ε, there exists a sufficiently large (random) N_2 such that for all $n > N_2$ and for each $t_l \in \mathcal{B}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_Y(t_l) - S_Y(t_l)) - (\widehat{S}_X(t_l) - S_X(t_l))] + S_Y(t_l) - S_X(t_l) \right] > \frac{\varepsilon}{2}.$$

Then, by the strong law of large numbers, with probability one,

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)}$$

$$\geq \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{1}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_{i1} \in \mathcal{A}_\varepsilon\} = \frac{\varepsilon}{4k} \mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0$$

and

$$\liminf_{n\to\infty} \frac{1}{m} \sum_{l=1}^{m} [\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)]^{(+)}$$

$$\geq \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{1}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_{i1} \in \mathcal{B}_\varepsilon\} = \frac{\varepsilon}{4k} \mathbb{P}(x_{i1} \in \mathcal{B}_\varepsilon) > 0,$$

which completes the proof. $\qquad\square$

Lemma 4 *Suppose that Assumption* 3 *is satisfied. Then the following hold true.*

(a) *If $H_>$ is true, then, with probability one,*

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)} > 0.$$

(b) *If H_{cr} is true, then, with probability one,*

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)} > 0$$

and

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)]^{(+)} > 0.$$

Proof Part (a). Using similar arguments as in the proof of Lemma 3, we get that for small enough ε, for all $n > N$, and for each $t_l \in \mathcal{A}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))] + S_X(t_l) - S_Y(t_l) \right] > \frac{\varepsilon}{2}.$$

For $\gamma > 1$, we have, almost surely,

$$\lim_{m\to\infty} \min_{1\le l\le m-1} m\psi_\gamma(t_l) = \frac{4^{1/\gamma}\Gamma(2 - 2/\gamma)}{(\Gamma(1 - 1/\gamma))^2} = C_\gamma > 0,$$

where $\Gamma(\cdot)$ is the gamma function. Then, by the strong law of large numbers, with probability one,

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)}$$

$$\ge \liminf_{n\to\infty} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} [\widehat{S}_X(y_{i1}) - \widehat{S}_Y(y_{i1})]^{(+)}$$

$$\ge \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_{i1} \in \mathcal{A}_\varepsilon\} = \frac{\varepsilon C_\gamma}{4k} \mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0,$$

which was to be shown.

Part (b). Using similar arguments as above, we get that for small enough ε, there exists a sufficiently large (random) N_1 such that for all $n > N_1$ and for each $t_l \in \mathcal{A}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_X(t_l) - S_X(t_l)) - (\widehat{S}_Y(t_l) - S_Y(t_l))] + S_X(t_l) - S_Y(t_l) \right] > \frac{\varepsilon}{2}.$$

Also, for small enough ε, there exists a sufficiently large (random) N_2 such that for all $n > N_2$ and for each $t_l \in \mathcal{B}_\varepsilon$ we have, with probability one,

$$\left[[(\widehat{S}_Y(t_l) - S_Y(t_l)) - (\widehat{S}_X(t_l) - S_X(t_l))] + S_Y(t_l) - S_X(t_l) \right] > \frac{\varepsilon}{2}.$$

Then, by the strong law of large numbers, with probability one,

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_X(t_l) - \widehat{S}_Y(t_l)]^{(+)}$$

$$\geq \liminf_{n\to\infty} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} [\widehat{S}_X(y_{i1}) - \widehat{S}_Y(y_{i1})]^{(+)}$$

$$\geq \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_{i1} \in \mathcal{A}_\varepsilon\} = \frac{\varepsilon C_\gamma}{4k} \mathbb{P}(y_{i1} \in \mathcal{A}_\varepsilon) > 0$$

and

$$\liminf_{n\to\infty} \sum_{l=1}^{m} \psi_\gamma(t_l)[\widehat{S}_Y(t_l) - \widehat{S}_X(t_l)]^{(+)}$$

$$\geq \liminf_{n\to\infty} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} [\widehat{S}_Y(x_{i1}) - \widehat{S}_X(x_{i1})]^{(+)}$$

$$\geq \lim_{n\to\infty} \frac{\varepsilon}{2} \frac{C_\gamma}{2k} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_{i1} \in \mathcal{B}_\varepsilon\} = \frac{\varepsilon C_\gamma}{4k} \mathbb{P}(x_{i1} \in \mathcal{B}_\varepsilon) > 0,$$

which completes the proof. □

Lemma 5 *If $H_>$ is true, then*

(a) $W_{X<Y} \xrightarrow{\text{a.s.}} 0$ *as* $n \longrightarrow \infty$;
(b) $A^\gamma_{X<Y} \xrightarrow{\text{a.s.}} 0$ *as* $n \longrightarrow \infty$;
(c) $D_{X<Y} \xrightarrow{\text{a.s.}} 0$ *as* $n \longrightarrow \infty$.

Proof Part (a) and Part (b) follow using similar reasoning as in the proofs of Lemma 3 and Lemma 4. Part (c) follows directly from Lemma 2. □

Let $a_{1,\alpha}$, $a_{2,\alpha}$, $a_{1,\alpha^\star}$, and $a_{2,\alpha^\star}$ be defined so that

$$\mathbb{P}(A_{X>Y} \geq a_{1,\alpha} \mid H_0) = \alpha, \qquad \mathbb{P}(A_{X<Y} \geq a_{2,\alpha} \mid H_0) = \alpha,$$

$$\mathbb{P}(A_{X>Y} \geq a_{1,\alpha^\star} \mid H_0) = \alpha^\star, \qquad \mathbb{P}(A_{X<Y} \geq a_{2,\alpha^\star} \mid H_0) = \alpha^\star.$$

Let $d_{1,\alpha}$, $d_{2,\alpha}$, $d_{1,\alpha^\star}$, and $d_{2,\alpha^\star}$ be defined so that

$$\mathbb{P}(D_{X>Y} \geq d_{1,\alpha} \mid H_0) = \alpha, \qquad \mathbb{P}(D_{X<Y} \geq d_{2,\alpha} \mid H_0) = \alpha,$$

$$\mathbb{P}(D_{X>Y} \geq d_{1,\alpha^\star} \mid H_0) = \alpha^\star, \qquad \mathbb{P}(D_{X<Y} \geq d_{2,\alpha^\star} \mid H_0) = \alpha^\star.$$

Let $\widetilde{w}_{1,\alpha}$, $\widetilde{w}_{2,\alpha}$, $\widetilde{w}_{1,\alpha^\star}$, and $\widetilde{w}_{2,\alpha^\star}$ denote the estimated quantiles based on the values $w_1^{[1]}, \ldots, w_1^{[R]}, w_2^{[1]}, \ldots, w_2^{[R]}$ generated by Algorithm 1. The estimated quantiles $\widetilde{a}_{1,\alpha}$, $\widetilde{a}_{2,\alpha}, \ldots, \widetilde{d}_{1,\alpha}, \widetilde{d}_{2,\alpha}, \ldots$ are defined similarly.

Let $\widetilde{q}_\alpha$ be a generic notation for the estimated quantile of U_n (under the null hypothesis) based on the permutation procedure. Recall that u_α is the quantile of U.

The following result will be of use.

Lemma 6 *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then under the null hypothesis* $\tau_n \widetilde{q}_\alpha \overset{\mathbb{P}}{\longrightarrow} u_\alpha$ *as* $n \longrightarrow \infty$.

Proof Follows from Theorem 15.2.3 in Lehmann and Romano (2005). □

Lemma 7 *Let* W_n *and* $\widetilde{w}_n$ *be nonnegative random variables such that* $W_n \overset{a.s.}{\longrightarrow} \infty$ *and* $\widetilde{w}_n$ *is bounded in probability as* $n \longrightarrow \infty$. *Then* $\mathbb{P}(W_n < \widetilde{w}_n) \longrightarrow 0$.

Proof Boundedness in probability implies that for every $\varepsilon > 0$, there exists a positive constant B and an integer N such that $\mathbb{P}(\widetilde{w}_n \leq B) \geq 1 - \varepsilon$ for all $n > N$. We have

$$\mathbb{P}(W_n < \widetilde{w}_n) = \mathbb{P}(W_n < \widetilde{w}_n \mid \widetilde{w}_n \leq B)\,\mathbb{P}(\widetilde{w}_n \leq B) + \mathbb{P}(W_n < \widetilde{w}_n \mid \widetilde{w}_n > B)\,\mathbb{P}(\widetilde{w}_n > B),$$

where $\mathbb{P}(W_n < \widetilde{w}_n \mid \widetilde{w}_n \leq B) \longrightarrow 0$ and $\mathbb{P}(\widetilde{w}_n > B) \leq \varepsilon$. Because ε is arbitrary, the claim follows. □

Proposition 1 *Suppose that Assumptions* 1 *and* 2 *are satisfied. Then the following hold true.*

(a) *If* $H_>$ *is true, then* $\lim_{n \to \infty} \mathbb{P}(W_{X>Y} < \widetilde{w}_{1,\alpha^\star}, \ W_{X<Y} \geq \widetilde{w}_{2,\alpha}) = 0$.
(b) *If* $H_>$ *is true, then* $\lim_{n \to \infty} \mathbb{P}(W_{X>Y} < \widetilde{w}_{1,\alpha}, \ W_{X<Y} < \widetilde{w}_{2,\alpha}) = 0$.
(c) *If* $H_>$ *is true, then* $\lim_{n \to \infty} \mathbb{P}(W_{X>Y} \geq \widetilde{w}_{1,\alpha^\star}, \ W_{X<Y} \geq \widetilde{w}_{2,\alpha^\star}) = 0$.
(d) *If* H_{cr} *is true, then* $\lim_{n \to \infty} \mathbb{P}(W_{X>Y} \geq \widetilde{w}_{1,\alpha^\star}, \ W_{X<Y} \geq \widetilde{w}_{2,\alpha^\star}) = 1$.

Proof

(a)   Using Lemma 3, we get that if $H_>$ is true, then $\tau_n W_{X>Y} \overset{a.s.}{\longrightarrow} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{w}_{1,\alpha^\star}$ is bounded in probability. Then, using Lemma 7, $\mathbb{P}(W_{X>Y} < \widetilde{w}_{1,\alpha^\star}) = \mathbb{P}(\tau_n W_{X>Y} < \tau_n \widetilde{w}_{1,\alpha^\star}) \longrightarrow 0$ and $\mathbb{P}(W_{X>Y} < \widetilde{w}_{1,\alpha^\star}, \ W_{X<Y} \geq \widetilde{w}_{2,\alpha}) \longrightarrow 0$.
(b)   Follows using similar arguments as in (a).
(c)   From Lemma 5, if $H_>$ is true, then $W_{X<Y} \overset{a.s.}{\longrightarrow} 0$ as $n \longrightarrow \infty$. Taking into account that $\widetilde{w}_{2,\alpha^\star} > 0$ for large n, we get $\mathbb{P}(W_{X>Y} \geq \widetilde{w}_{1,\alpha^\star}, \ W_{X<Y} \geq \widetilde{w}_{2,\alpha^\star}) \longrightarrow 0$.

(d)   Using Lemma 3, we get that if H_{cr} is true, then $\tau_n W_{X>Y} \xrightarrow{\text{a.s.}} \infty$ and $\tau_n W_{X<Y} \xrightarrow{\text{a.s.}} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{w}_{1,\alpha^\star}$ and $\tau_n \widetilde{w}_{2,\alpha^\star}$ are bounded in probability. Then, using Lemma 7, $\mathbb{P}(W_{X>Y} < \widetilde{w}_{1,\alpha^\star} \cup W_{X<Y} < \widetilde{w}_{2,\alpha^\star})$ $\leq \mathbb{P}(\tau_n W_{X>Y} < \tau_n \widetilde{w}_{1,\alpha^\star}) + \mathbb{P}(\tau_n W_{X<Y} < \tau_n \widetilde{w}_{2,\alpha^\star}) \longrightarrow 0$. $\qquad\square$

Proposition 2 *Suppose that Assumptions 1, 2, and 3 are satisfied. Then the following hold true.*

(a) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(A_{X>Y} < \widetilde{a}_{1,\alpha^\star}, A_{X<Y} \geq \widetilde{a}_{2,\alpha}) = 0$.*
(b) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(A_{X>Y} < \widetilde{a}_{1,\alpha}, A_{X<Y} < \widetilde{a}_{2,\alpha}) = 0$.*
(c) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(A_{X>Y} \geq \widetilde{a}_{1,\alpha^\star}, A_{X<Y} \geq \widetilde{a}_{2,\alpha^\star}) = 0$.*
(d) *If H_{cr} is true, then $\lim_{n\to\infty} \mathbb{P}(A_{X>Y} \geq \widetilde{a}_{1,\alpha^\star}, A_{X<Y} \geq \widetilde{a}_{2,\alpha^\star}) = 1$.*

Proof

(a)   Using Lemma 4, we get that if $H_>$ is true, then $\tau_n A_{X>Y} \xrightarrow{\text{a.s.}} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{a}_{1,\alpha^\star}$ is bounded in probability. Then, using Lemma 7, $\mathbb{P}(A_{X>Y} < \widetilde{a}_{1,\alpha^\star}) = \mathbb{P}(\tau_n A_{X>Y} < \tau_n \widetilde{a}_{1,\alpha^\star}) \longrightarrow 0$ and $\mathbb{P}(A_{X>Y} < \widetilde{a}_{1,\alpha^\star}, A_{X<Y} \geq \widetilde{a}_{2,\alpha}) \longrightarrow 0$.
(b)   Follows using similar arguments as in (a).
(c)   From Lemma 5, if $H_>$ is true, then $A_{X<Y} \xrightarrow{\text{a.s.}} 0$ as $n \longrightarrow \infty$. Taking into account that $\widetilde{a}_{2,\alpha^\star} > 0$ for large n, we get $\mathbb{P}(A_{X>Y} \geq \widetilde{a}_{1,\alpha^\star}, A_{X<Y} \geq \widetilde{a}_{2,\alpha^\star}) \longrightarrow 0$.
(d)   Using Lemma 4, we get that if H_{cr} is true, then $\tau_n A_{X>Y} \xrightarrow{\text{a.s.}} \infty$ and $\tau_n A_{X<Y} \xrightarrow{\text{a.s.}} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{a}_{1,\alpha^\star}$ and $\tau_n \widetilde{a}_{2,\alpha^\star}$ are bounded in probability. Then, using Lemma 7, $\mathbb{P}(A_{X>Y} < \widetilde{a}_{1,\alpha^\star} \cup A_{X<Y} < \widetilde{a}_{2,\alpha^\star})$ $\leq \mathbb{P}(\tau_n A_{X>Y} < \tau_n \widetilde{a}_{1,\alpha^\star}) + \mathbb{P}(\tau_n A_{X<Y} < \tau_n \widetilde{a}_{2,\alpha^\star}) \longrightarrow 0$. $\qquad\square$

Proposition 3 *Suppose that Assumptions 1 and 2 are satisfied. Then the following hold true.*

(a) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(D_{X>Y} < \widetilde{d}_{1,\alpha^\star}, D_{X<Y} \geq \widetilde{d}_{2,\alpha}) = 0$.*

(b) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(D_{X>Y} < \widetilde{d}_{1,\alpha}, D_{X<Y} < \widetilde{d}_{2,\alpha}) = 0$.*

(c) *If $H_>$ is true, then $\lim_{n\to\infty} \mathbb{P}(D_{X>Y} \geq \widetilde{d}_{1,\alpha^\star}, D_{X<Y} \geq \widetilde{d}_{2,\alpha^\star}) = 0$.*

(d) *If H_{cr} is true, then $\lim_{n\to\infty} \mathbb{P}(D_{X>Y} \geq \widetilde{d}_{1,\alpha^\star}, D_{X<Y} \geq \widetilde{d}_{2,\alpha^\star}) = 1$.*

Proof

(a)   Using Lemma 2, we get that if $H_>$ is true, then $\tau_n D_{X>Y} \xrightarrow{\text{a.s.}} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{d}_{1,\alpha^\star}$ is bounded in probability. Then, using Lemma 7, $\mathbb{P}(D_{X>Y} < \widetilde{d}_{1,\alpha^\star}) = \mathbb{P}(\tau_n D_{X>Y} < \tau_n \widetilde{d}_{1,\alpha^\star}) \longrightarrow 0$ and $\mathbb{P}(D_{X>Y} < \widetilde{d}_{1,\alpha^\star}, D_{X<Y} \geq \widetilde{d}_{2,\alpha}) \longrightarrow 0$.
(b)   Follows using similar arguments as in (a).

(c)   From Lemma 5, if $H_\succ$ is true, then $D_{X<Y} \xrightarrow{\text{a.s.}} 0$ as $n \longrightarrow \infty$. Taking into account that $\widetilde{d}_{2,\alpha^\star} > 0$ for large n, we get $\mathbb{P}(D_{X>Y} \geq \widetilde{d}_{1,\alpha^\star}, D_{X<Y} \geq \widetilde{d}_{2,\alpha^\star}) \longrightarrow 0$.

(d)   Using Lemma 2, we get that if H_{cr} is true, then $\tau_n D_{X>Y} \xrightarrow{\text{a.s.}} \infty$ and $\tau_n D_{X<Y} \xrightarrow{\text{a.s.}} \infty$ as $n \longrightarrow \infty$. From Lemma 6, $\tau_n \widetilde{d}_{1,\alpha^\star}$ and $\tau_n \widetilde{d}_{2,\alpha^\star}$ are bounded in probability. Then, using Lemma 7, $\mathbb{P}(D_{X>Y} < \widetilde{d}_{1,\alpha^\star} \cup D_{X<Y} < \widetilde{d}_{2,\alpha^\star})$ $\leq \mathbb{P}(\tau_n D_{X>Y} < \tau_n \widetilde{d}_{1,\alpha^\star}) + \mathbb{P}(\tau_n D_{X<Y} < \tau_n \widetilde{d}_{2,\alpha^\star}) \longrightarrow 0$. $\qquad\square$

Proof of Theorem 1

(a)   We have $\mathbb{P}(W_{X>Y} \geq \widetilde{w}_{1,\alpha}, W_{X<Y} < \widetilde{w}_{2,\alpha^\star}) \leq \mathbb{P}(W_{X>Y} \geq \widetilde{w}_{1,\alpha}) \leq \alpha$, where the last inequality follows from Theorem 2 in Hemerik and Goeman (2018).

(b)   From Proposition 1 (d), (a), it follows that $\text{FDP}_2 \longrightarrow 0$ and $\text{FDP}_3 \longrightarrow 0$ as $n \longrightarrow \infty$.

(c)   Proposition 1 (b), (c), (a) imply that NDP_1, NDP_2, and NDP_3 tend to zero as $n \longrightarrow \infty$.

$\qquad\square$

Proof of Theorem 2

(a)   We have $\mathbb{P}(A_{X>Y} \geq \widetilde{a}_{1,\alpha}, A_{X<Y} < \widetilde{a}_{2,\alpha^\star}) \leq \mathbb{P}(A_{X>Y} \geq \widetilde{a}_{1,\alpha}) \leq \alpha$, where the last inequality follows from Theorem 2 in Hemerik and Goeman (2018).

(b)   From Proposition 2 (d), (a), it follows that $\text{FDP}_2 \longrightarrow 0$ and $\text{FDP}_3 \longrightarrow 0$ as $n \longrightarrow \infty$.

(c)   Proposition 2 (b), (c), (a) imply that NDP_1, NDP_2, and NDP_3 tend to zero as $n \longrightarrow \infty$.

$\qquad\square$

Proof of Theorem 3

(a)   We have $\mathbb{P}(D_{X>Y} \geq \widetilde{d}_{1,\alpha}, D_{X<Y} < \widetilde{d}_{2,\alpha^\star}) \leq \mathbb{P}(D_{X>Y} \geq \widetilde{d}_{1,\alpha}) \leq \alpha$, where the last inequality follows from Theorem 2 in Hemerik and Goeman (2018).

(b)   From Proposition 3 (d), (a), it follows that $\text{FDP}_2 \longrightarrow 0$ and $\text{FDP}_3 \longrightarrow 0$ as $n \longrightarrow \infty$.

(c)   Proposition 3 (b), (c), (a) imply that NDP_1, NDP_2, and NDP_3 tend to zero as $n \longrightarrow \infty$.

$\qquad\square$

Declarations

Conflicts of interest No.

Availability of data and material The real data is available online.

Code availability Available upon request.

References

Anderson, T.W.: On the distribution of the two-sample Cramer-von Mises criterion. Annal. Math. Statistics **33**(3), 1148–1159 (1962)

Angelov, A.G., Ekström, M., Kriström, B., and Nilsson, M.E.: Data for: Four-decision tests for stochastic dominance, with an application to environmental psychophysics. Mendeley Data, V1 (2019a). https://doi.org/10.17632/mvnwgvht9y.1 [dataset]

Angelov, A.G., Ekström, M., Kriström, B., Nilsson, M.E.: Four-decision tests for stochastic dominance, with an application to environmental psychophysics. J. Math. Psychol. **93**, 102281 (2019b)

Ashby, F.G., Tein, J.-Y., Balakrishnan, J.D.: Response time distributions in memory scanning. J. Math. Psychol. **37**(4), 526–555 (1993)

Bennett, C.J.: Inference for dominance relations. Int. Econ. Rev. **54**(4), 1309–1328 (2013)

Davidson, R., Duclos, J.-Y.: Statistical inference for stochastic dominance and for the measurement of poverty and inequality. Econometrica **68**(6), 1435–1464 (2000)

Donald, S.G., Hsu, Y.-C.: Improving the power of tests of stochastic dominance. Econometric Rev. **35**(4), 553–585 (2016)

Heathcote, A., Brown, S., Wagenmakers, E.J., Eidels, A.: Distribution-free tests of stochastic dominance for small samples. J. Math. Psychol. **54**(5), 454–463 (2010)

Heck, D.W., Erdfelder, E.: Extending multinomial processing tree models to measure the relative speed of cognitive processes. Psychonomic Bulletin Rev. **23**(5), 1440–1465 (2016)

Hemerik, J., Goeman, J.: Exact testing with random permutations. TEST **27**(4), 811–825 (2018)

Hemerik, J., Goeman, J.: Another look at the Lady Tasting Tea and differences between permutation tests and randomisation tests. Int. Statistical Rev. **89**(2), 367–381 (2021)

Kotz, S., Kozubowski, T.J., Podgórski, K.: The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance. Springer, Boston (2001)

Ledwina, T., Wyłupek, G.: Two-sample test against one-sided alternatives. Scand. J. Statistics **39**(2), 358–381 (2012)

Lehmann, E.L., Romano, J.P.: Testing Statistical Hypotheses. Springer, New York, 3 edition (2005)

Levy, H.: Stochastic Dominance: Investment Decision Making under Uncertainty. Springer, New York, 3 edition (2016)

Petroni, G.R., Wolfe, R.A.: A two-sample test for stochastic ordering with interval-censored data. Biometrics **50**(1), 77–87 (1994)

Pettitt, A.N.: A two-sample Anderson-Darling rank statistic. Biometrika **63**(1), 161–168 (1976)

Politis, D.N., Romano, J.P., Wolf, M.: Subsampling. Springer, New York (1999)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2019) https://www.R-project.org

Romano, J.P.: Bootstrap and randomization tests of some nonparametric hypotheses. Annal. Statistics **17**(1), 141–159 (1989)

Roy, A., Harrar, S.W., Konietschke, F.: The nonparametric Behrens-Fisher problem with dependent replicates. Statistics Med. **38**(25), 4939–4962 (2019)

Schmid, F., Trede, M.: A distribution free test for the two sample problem for general alternatives. Comput. Stat. Data Anal. **20**(4), 409–419 (1995)

Schmid, F., Trede, M.: An L_1-variant of the Cramér-von Mises test. Stat. Prob. Lett. **26**(1), 91–96 (1996)

Statisticat LLC: LaplacesDemon: complete environment for Bayesian inference. R package version 16.1.1 (2018). https://CRAN.R-project.org/package=LaplacesDemon

Townsend, J.T.: Truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. Psychol. Bull. **108**(3), 551–567 (1990)

Tucker, H.G.: A generalization of the Glivenko-Cantelli theorem. Annal. Math. Statistics **30**(3), 828–830 (1959)

Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S. Springer, New York, 4 edition (2002)

Whang, Y.-J.: Econometric Analysis of Stochastic Dominance: Concepts, Methods, Tools, and Applications. Cambridge University Press, Cambridge (2019)