Four-decision tests for stochastic dominance, with an application to environmental psychophysics

Angel G. Angelov[a,*], Magnus Ekström[a,b], Bengt Kriström[c], Mats E. Nilsson[d]

[a]*Department of Statistics, Umeå School of Business, Economics and Statistics, Umeå University, Umeå, Sweden*
[b]*Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden*
[c]*Department of Forest Economics, Swedish University of Agricultural Sciences, Umeå, Sweden*
[d]*Gösta Ekman Laboratory, Department of Psychology, Stockholm University, Stockholm, Sweden*

Abstract

If the survival function of a random variable X lies to the right of the survival function of a random variable Y, then X is said to stochastically dominate Y. Inferring stochastic dominance is particularly complicated because comparing survival functions raises four possible hypotheses: identical survival functions, dominance of X over Y, dominance of Y over X, or crossing survival functions. In this paper, we suggest four-decision tests for stochastic dominance suitable for paired samples. The tests are permutation-based and do not rely on distributional assumptions. One-sided Cramér–von Mises and Kolmogorov–Smirnov statistics are employed but the general idea may be utilized with other test statistics. The power to detect dominance and the different types of wrong decisions are investigated in an extensive simulation study. The proposed tests are applied to data from an experiment concerning the individual's willingness to pay for a given environmental improvement.

Key words: Stochastic dominance; Stochastic ordering; Four-hypothesis test; Permutation test; Nonparametric approach; Environmental psychology.

1. Introduction

The concept of stochastic dominance is useful when comparing two things that can be modeled by random variables. It has been employed in many contexts, e.g., for comparing income distributions (Davidson & Duclos, 2000), comparing investment assets (Levy, 2016), comparing medical treatments (Petroni & Wolfe, 1994), and for assessing distributional treatment effects (Abadie, 2002). In psychology, statistical tests for stochastic dominance are used, e.g., when comparing response time distributions in the study of perceptual processes (Ashby et al., 1993; Fitousi & Algom, 2018; Heck & Erdfelder, 2016; Houpt & Townsend, 2010; Yang et al., 2018). As discussed by Townsend (1990), stochastic dominance implies (but is not implied by) the same ordering of the means. The stochastic dominance paradigm is prevalent in behavioral economics where it was introduced mainly in search for a robust alternative to classical mean-variance analysis which has some well-known shortcomings (Hadar & Russell, 1969; Hanoch & Levy, 1969; Rothschild & Stiglitz, 1970; Whitmore, 1970). Mean-variance analysis is used for ranking of investment assets and it posits that an investor has positive preferences over the mean return of an asset and negative preferences over its variance. If two assets has the same mean return, but different variances, the rule picks the one with the smaller variance. A classical example, which we borrow from Levy (2016), considers two assets. Asset A returns 1 or 2 with equal probability, while asset B returns 2 or 4 also with equal probability. A simple calculation shows that asset B has a larger mean return but also a larger variance; therefore the mean-variance criterion is not very helpful. Still, most people would agree that B is to be preferred; after all, B gives at least as high return as A. In this case, the random variable corresponding to B stochastically dominates the one corresponding to A, i.e., the survival function of the former variable lies to the right of the survival function of the latter, implying that a criterion based on stochastic dominance will overcome the issue with the mean-variance criterion.

In this paper, we focus on statistical tests for stochastic dominance. 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

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

The stochastic dominance defined above is sometimes called first order stochastic dominance or stochastic ordering. It can be equivalently defined using the distribution functions of X and Y. Hereafter, we will skip the words stochastic/stochastically and we will just say that X dominates Y. If X dominates Y, we write $X \succ Y$, or equivalently, we can say that Y is dominated by X and denote this $Y \prec X$. If there exist values a and b such that $S_X(a) > S_Y(a)$ and

[*] Corresponding author
Email address: agangelov@gmail.com (Angel G. Angelov)

$S_X(b) < S_Y(b)$, we say that the survival functions of X and Y cross one another. Four possible hypotheses about X and Y can be distinguished: (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.

Two main types of test statistics can be found in the literature on dominance testing. The first type are statistics based on the differences between the empirical distribution functions at a fixed number of points in the support (see Anderson, 1996; Davidson & Duclos, 2000). However, the choice of optimal points is not obvious and these tests might possibly be inconsistent (see Barrett & Donald, 2003). The second type of statistics are based on some real-valued functional whose value is zero when the distributions are identical and strictly positive under dominance. For example, McFadden (1989), Barrett & Donald (2003), Linton et al. (2005), and Donald & Hsu (2016) used one-sided Kolmogorov–Smirnov statistics or their modifications, while Schmid & Trede (1996), Bennett (2008), and Linton et al. (2010) employed one-sided versions of the Cramér–von Mises statistic. The one-sided Kolmogorov–Smirnov statistics actually date back to Smirnov (see, e.g., Darling, 1957; Hodges, 1958 and the references therein). A different approach was suggested by Ledwina & Wyłupek (2012a,b) who deal with the testing problem by formulating the hypotheses in terms of Fourier coefficients in some system of functions. To our knowledge, the existing asymptotic results for the Kolmogorov–Smirnov statistics and the Cramér–von Mises statistics (cf. Durbin, 1973) assume that the two samples are independent and are not applicable to paired samples.

Some test statistics (e.g., the one-sided Kolmogorov–Smirnov) take positive values not only under dominance but also when the two survival functions cross one another, which may too often lead to a false detection of dominance, i.e., the test chooses a hypothesis of dominance when the truth is that the survival functions cross (see Bennett, 2008, 2013). This issue can be tackled using a decision rule that involves all four hypotheses. A testing procedure with four hypothesis (four possible decisions) was proposed by Bishop et al. (1989); see also Bishop & Formby (1999), Tse & Zhang (2004), Knight & Satchell (2008), Heathcote et al. (2010). The procedure, however, does not provide adequate control over the error probabilities related to the different types of wrong decisions in a four-hypothesis setting. Bennett (2013) modified the decision rule of Bishop et al. (1989) and suggested a test that has better power to detect crossings and allows finer control over the different error probabilities based on asymptotic properties of the one-sided Kolmogorov–Smirnov statistics.

To the best of our knowledge, only the tests of Davidson & Duclos (2000), Linton et al. (2005), and Donald & Hsu (2016) are applicable to paired samples and we are not aware of a test with four hypotheses for paired samples. Employing the four-decision rule of Bennett (2013), we suggest dominance testing procedures suitable for paired samples. The procedures are based on a permutation test approach, which allows computing p-values without relying on large-sample results for the test statistic. In Section 2, we introduce the testing procedures. Section 3 presents a simulation study. In Section 4, the suggested procedures are applied to data from an experiment concerning the individual's willingness to pay for an environmental improvement (traffic noise reduction). In this case, dominance tests are used for comparing participants' responses under different scenarios for noise reduction and under different formats of the willingness-to-pay question.

2. Testing procedures

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

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

We explore the problem of testing for stochastic dominance with null hypothesis H_0 and three "alternative" hypotheses: $H_\succ$, $H_\prec$, and H_{cr}.

Now, we define the test statistics under a general setup before restricting the discussion to the case of paired samples. Let $x_1, \ldots, x_n$ be observations from S_X and $y_1, \ldots, y_m$ be observations from S_Y. The empirical distribution function based on the observations $x_1, \ldots, x_n$ is $\widehat{F}_X(t) = (1/n)\sum_i \mathbb{1}\{x_i \le t\}$ 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 $y_1, \ldots, y_m$ are defined analogously. Let us denote $(t_1, \ldots, t_{n+m}) = (x_1, \ldots, x_n, y_1, \ldots, y_m)$, $a_{n,m} = (nm)^{1/2}(n+m)^{-1/2}$, and $z^{(+)} = \max\{z, 0\}$ for any real number z. For simplicity, the observations (e.g., $x_1, \ldots, x_n$) denote random variables or values of random variables, depending on the context. We consider the following test statistics:

- One-sided Cramér–von Mises statistics

$$W_{X\succ Y} = \frac{a_{n,m}}{n+m}\sum_{k=1}^{n+m}\left(\widehat{S}_X(t_k) - \widehat{S}_Y(t_k)\right)^{(+)},$$

$$W_{X\prec Y} = \frac{a_{n,m}}{n+m}\sum_{k=1}^{n+m}\left(\widehat{S}_Y(t_k) - \widehat{S}_X(t_k)\right)^{(+)};$$

- One-sided Kolmogorov–Smirnov statistics

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

$$D_{X\prec Y} = a_{n,m}\sup_t\left(\widehat{S}_Y(t) - \widehat{S}_X(t)\right).$$

If we want to be more precise, $W_{X\succ Y}$ and $W_{X\prec Y}$ should be called modified Cramér–von Mises statistics as they

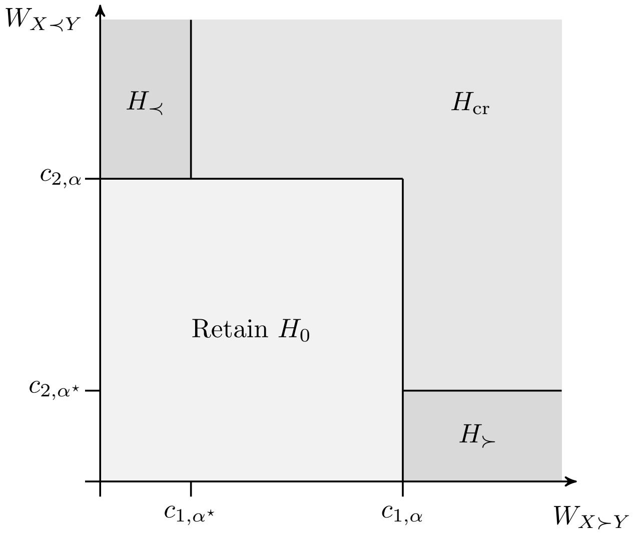

Figure 1: Decision rule.

are not based on squared differences (cf. Schmid & Trede, 1995). We will focus on paired samples, in which case $m = n$ and the observed data can be written $(x_1, y_1), \ldots, (x_n, y_n)$.

Hereafter, we will describe the testing procedure in terms of the statistics $(W_{X\succ Y}, W_{X\prec Y})$; the procedure with $(D_{X\succ Y}, D_{X\prec Y})$ is essentially the same. The testing problem involves four hypotheses; hence there are four decision regions which are determined by four critical values. Let $c_{1,\alpha}$ and $c_{2,\alpha}$ be defined so that $\mathbb{P}\left(W_{X\succ Y} \geq c_{1,\alpha} \,|\, H_0\right) = \alpha$ and $\mathbb{P}\left(W_{X\prec Y} \geq c_{2,\alpha} \,|\, H_0\right) = \alpha$. Similarly, $c_{1,\alpha^\star}$ and $c_{2,\alpha^\star}$ are such that $\mathbb{P}\left(W_{X\succ Y} \geq c_{1,\alpha^\star} \,|\, H_0\right) = \alpha^\star$ and $\mathbb{P}\left(W_{X\prec Y} \geq c_{2,\alpha^\star} \,|\, H_0\right) = \alpha^\star$, where $\alpha^\star > \alpha$. We employ the following decision rule, which is equivalent to the rule studied by Bennett (2013).

Decision rule 1.

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

(b) If $W_{X\succ Y} \geq c_{1,\alpha}$ or $W_{X\prec Y} \geq c_{2,\alpha}$, then

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

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

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

The decision rule is depicted in Figure 1. The main idea is to accept the hypothesis of dominance $H_{\succ}$ 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.

Larger values of $\alpha^\star$ result in smaller values of $c_{1,\alpha^\star}$ and $c_{2,\alpha^\star}$. Thus, if we increase $\alpha^\star$, the acceptance region for crossing is increased, while the acceptance regions for dominance are decreased. In this way, we can control the discrimination between stochastic dominance and crossing survival functions.

For computing the quantiles of the test statistics under the null hypothesis or the corresponding p-values, we adopt a permutation test approach (sometimes called randomization test). That is, we generate random permutations of the data $(x_1, y_1), \ldots, (x_n, y_n)$, compute 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 & Goeman, 2018; Lehmann & Romano, 2005, Ch. 15; Romano, 1989). Generating a random permutation of $(x_1, y_1), \ldots, (x_n, y_n)$ corresponds to switching the places of x_i and y_i in each pair (x_i, y_i) with probability 1/2. The detailed algorithm is presented below.

Let $\texttt{TS}\,(data)$ denote the value of the bivariate test statistic $(W_{X\succ Y}, W_{X\prec Y})$ calculated for the dataset *data*. For example, $(w_1, w_2) = \texttt{TS}\,(x_1, y_1, \ldots, x_n, y_n)$ denotes the value of $(W_{X\succ Y}, W_{X\prec Y})$ for the observed data $(x_1, y_1), \ldots, (x_n, y_n)$.

Algorithm 1

```
Input: $(x_1, y_1), \ldots, (x_n, y_n)$.
for $r = 1, \ldots, R$
    for $i = 1, \ldots, n$
        Generate $b_i$ from Bernoulli(1/2);
        if $b_i = 1$ then set $\mathbf{z}_i^{[r]} = (x_i, y_i)$;
        else set $\mathbf{z}_i^{[r]} = (y_i, x_i)$;
    end for
    Compute $(w_1^{[r]}, w_2^{[r]}) = \texttt{TS}\,(\mathbf{z}_1^{[r]}, \ldots, \mathbf{z}_n^{[r]})$;
end for
Output: $w_1^{[1]}, \ldots, w_1^{[R]}, w_2^{[1]}, \ldots, w_2^{[R]}$.
```

Let us denote

$$p_1 = \mathbb{P}\left(W_{X\succ Y} \geq w_1 \,|\, H_0\right), \quad p_2 = \mathbb{P}\left(W_{X\prec Y} \geq w_2 \,|\, H_0\right),$$

which will be referred to as marginal p-values.

Using $w_1^{[1]}, \ldots, w_1^{[R]}, w_2^{[1]}, \ldots, w_2^{[R]}$ obtained through Algorithm 1, the marginal p-values can be estimated as follows:

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

Decision rule 1 can be reformulated like this:

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 $H_{\succ}$;

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

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

The described procedures can be utilized with any suitable test statistic which takes positives values under dominance and equals zero when the survival functions are identical. Similar procedures can be used in the case of two independent samples; the only difference will be in the generation of random permutations of the data vector $(x_1, \dots, x_n, y_1, \dots, y_m)$.

Similarly to Bennett (2013), the testing problem involves four hypotheses, one of which is the null hypothesis. Still, this problem has much in common with standard hypothesis testing. The critical values (respectively, the marginal p-values) are obtained assuming the null hypothesis is true. Thus, retaining the null hypothesis does not mean that we have proved that there is no difference between the survival functions; it only means that the data do not provide enough evidence against H_0.

There can be some borderline cases 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 $\alpha^\star$). For example, if $\min\{\widetilde{p}_1, \widetilde{p}_2\}$ is slightly below α, the evidence against H_0 is not very strong, but the smaller the value of $\min\{\widetilde{p}_1, \widetilde{p}_2\}$, the greater the statistical incompatibility of the data with H_0 (cf. Wasserstein & Lazar, 2016). Also, if $\widetilde{p}_1 \le \alpha$ and $\widetilde{p}_2$ is near the threshold $\alpha^\star$, then the null hypothesis is rejected, but the support for $H_\succ$ over H_{cr} (or vice versa) is not strong. The case when $\widetilde{p}_2 \le \alpha$ and $\widetilde{p}_1$ is near $\alpha^\star$ is similar. In practice, it is recommended not only to report whether a certain hypothesis is accepted/retained, but to provide also the marginal p-values $\widetilde{p}_1$, $\widetilde{p}_2$ and the thresholds α, $\alpha^\star$.

Because the considered tests involve four hypotheses, the possible errors are more complicated compared to classical hypothesis testing where there are only two hypotheses. With regard to making a wrong decision, two events are of main interest:

(a) false detection of dominance: accepting the hypothesis $H_\succ$ when it is not true;

(b) non-detection of dominance: not accepting the hypothesis $H_\succ$ when it is true.

In classical hypothesis testing (e.g., testing H_0 against $H_\succ$), a false detection may happen only if we accept $H_\succ$ when H_0 is true. In a test with four hypotheses, a false detection may happen in three different ways (there are three other hypotheses). Let FDP be the probability of a false detection of dominance ($H_\succ$) and let NDP be the probability of a non-detection of dominance ($H_\succ$). These probabilities can be expressed as follows:

$$\begin{aligned} \text{FDP} &= \mathbb{P}\,(\text{accept } H_\succ \mid H_0) + \mathbb{P}\,(\text{accept } H_\succ \mid H_{\text{cr}}) \\ &\quad + \mathbb{P}\,(\text{accept } H_\succ \mid H_\prec), \\ \text{NDP} &= \mathbb{P}\,(\text{do not accept } H_\succ \mid H_\succ). \end{aligned}$$

The power to detect dominance ($H_\succ$) is defined as $\mathbb{P}\,(\text{accept } H_\succ \mid H_\succ) = 1 - \text{NDP}$. Note that we consider the probability of a false detection (respectively, non-detection) of either $H_\succ$ or $H_\prec$, depending on what we are interested in.

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 context, if $H_\succ$ is the hypothesis of interest, false detection of $H_\succ$ and non-detection of $H_\succ$ can be viewed as analogues of Type I error and Type II error, respectively. As usual, we want to have a test such that the probability of a false detection is small. From the decision rule we can deduce that $\mathbb{P}\,(\text{accept } H_\succ \mid H_0) \le \alpha$. The probabilities $\mathbb{P}\,(\text{accept } H_\succ \mid H_{\text{cr}})$ and $\mathbb{P}\,(\text{accept } H_\succ \mid H_\prec)$ are hard to assess theoretically and they will be explored in our simulation study (Section 3).

3. Simulation study

3.1. Setup

In order to investigate the different types of wrong decisions and the power to detect dominance, we conducted simulations under H_0, H_{cr}, and $H_\prec$, with sample sizes $n = 50$, 100, 200, and 500.

Let us denote:

$\mathcal{N}(\mu, \sigma)$ normal distribution with mean μ and standard deviation σ;

$\text{LN}(\mu, \sigma)$ lognormal distribution with parameters μ and σ such that $X \sim \text{LN}(\mu, \sigma) \iff \log(X) \sim \mathcal{N}(\mu, \sigma)$;

$\text{La}(\mu, \sigma)$ Laplace distribution with mean μ and standard deviation σ;

$\text{Gu}(\mu, \sigma)$ Gumbel distribution with location parameter μ and scale parameter σ.

Also, let $\boldsymbol{\mu} = (\mu_1, \mu_2)$, $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$.

We simulated data from the following models:

(A) Bivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $(X, Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The R package MASS was used (see Venables & Ripley, 2002).

(B) Bivariate lognormal distribution: $(X, Y) \sim \text{LN}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \iff (\log X, \log Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The R package MASS (Venables & Ripley, 2002) was used for generating $(\log X, \log Y)$.

(C) Bivariate Laplace distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (see, e.g., Kotz et al., 2001). The R package LaplacesDemon was used (see Statisticat LLC, 2018).

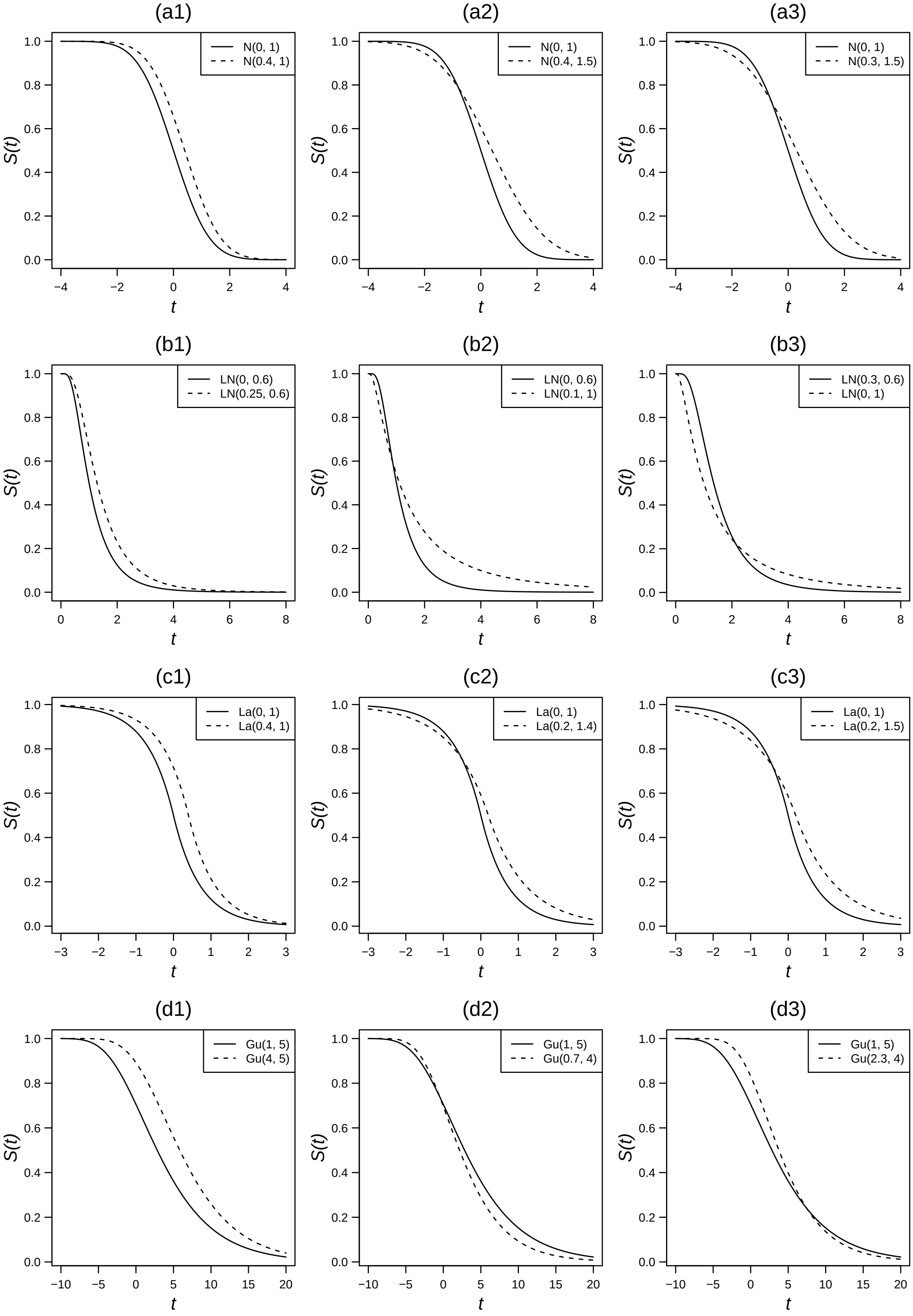

Figure 2: Survival curves: first row – normal distribution, second row – lognormal distribution, third row – Laplace distribution, fourth row – Gumbel distribution.

(D) Bivariate Gumbel distribution with location parameters μ_1, μ_2, scale parameters σ_1, σ_2, and correlation ρ, based on the so-called logistic model (see Tawn, 1988, p. 401). The R package `evd` was used (see Stephenson, 2002).

The following will be of use:

- If $X \sim \mathcal{N}(\mu_1, \sigma)$, $Y \sim \mathcal{N}(\mu_2, \sigma)$ and $\mu_2 > \mu_1$, then $Y \succ X$.
- If $X \sim \mathrm{LN}(\mu_1, \sigma)$, $Y \sim \mathrm{LN}(\mu_2, \sigma)$ and $\mu_2 > \mu_1$, then $Y \succ X$.
- If $X \sim \mathrm{La}(\mu_1, \sigma)$, $Y \sim \mathrm{La}(\mu_2, \sigma)$ and $\mu_2 > \mu_1$, then $Y \succ X$.
- If $X \sim \mathrm{Gu}(\mu_1, \sigma)$, $Y \sim \mathrm{Gu}(\mu_2, \sigma)$ and $\mu_2 > \mu_1$, then $Y \succ X$.

Figure 2 illustrates various pairs of survival functions under some of the settings used in the simulations. All computations were performed with R (see R Core Team, 2018). The R code can be obtained from the first author upon request. The results in this section are based on 3000 simulated datasets under each setting; the number of generated random permutations for each dataset is $R = 5000$, and $\alpha = 0.05$. CvM stands for Cramér–von Mises and KS stands for Kolmogorov–Smirnov.

3.2. Error probabilities against $\alpha^\star$

The value of $\alpha^\star$ controls the acceptance of $H_\succ$, $H_\prec$, and H_{cr}. Thus, it affects the probability of a non-detection of dominance (NDP) and the probability of a false detection of dominance given that the truth is crossing survival curves (FDP$_2$). In order to find a reasonable value of $\alpha^\star$, we examined NDP and FDP$_2$ using different values of $\alpha^\star$. We conducted simulations for $n = 100$ with the settings specified below:

(a) $H_\prec$: bivariate normal with $\rho = 0.8$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0.25$, $\sigma_2 = 1$;

H_{cr} : bivariate normal with $\rho = 0.8$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0.4$, $\sigma_2 = 1.5$;

(b) $H_\prec$: bivariate lognormal with $\rho = 0.8$, $\mu_1 = 0$, $\sigma_1 = 0.6$, $\mu_2 = 0.15$, $\sigma_2 = 0.6$;

H_{cr} : bivariate lognormal with $\rho = 0.8$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 0.3$, $\sigma_2 = 0.6$.

The results with the Cramér–von Mises statistics are presented in Figure 3. We see that as $\alpha^\star$ gets larger, FDP$_2$ decreases, while NDP increases. Different values of $\alpha^\star$ can be chosen depending on how conservative one wants to be. We decide to take $\alpha^\star = 0.96$ as it gives some balance between the two errors. For the Kolmogorov–Smirnov statistics we also take $\alpha^\star = 0.96$ based on similar reasoning. This value of $\alpha^\star$ is used throughout the remainder of the paper.

3.3. False detection of dominance

We first look at the probability of a false detection of dominance when the truth is H_0. Simulation results are shown in Table 1. The probability to wrongly accept $H_\prec$ is approximately the same as the probability to wrongly accept $H_\succ$ and both are not greater than $\alpha/2 = 0.025$. Analogous simulations with the lognormal, the Laplace, and the Gumbel distributions gave similar results (not presented here).

To investigate the false detection of dominance when the truth is H_{cr}, we performed simulations under the settings illustrated in Figure 2, second and third columns. The results are presented in Table 2 and Figure 4. We see that as the sample size increases, the probability of a false detection approaches zero and the probability to detect crossings approaches one. For smaller sample sizes ($n = 50, 100$) the probability to detect crossings is overall higher with the Cramér–von Mises statistics, while for large sample sizes ($n = 200, 500$) the two tests have quite similar performance (see Table 2).

With regard to the probability to accept $H_\succ$ given $H_\prec$, in our simulations under $H_\prec$ (Section 3.4) the hypothesis of dominance in the opposite direction ($H_\succ$) was never accepted. This type of false detection may happen for small n when the two survival curves are close to one another; yet, under such scenarios the tests are not expected to be powerful.

To sum up, under the settings of our study, the results indicate that the total probability of a false detection of dominance in a certain direction is less than α for large sample sizes.

3.4. Power to detect dominance

Power curves for $n = 100$ are illustrated in Figure 5. The power to detect dominance is calculated for a grid of values of $\delta = \mu_2 - \mu_1$, under the following settings, where $\sigma_1 = \sigma_2 = \sigma$:

(a) Bivariate normal with $\mu_1 = 0$, $\sigma = 1$;

(b) Bivariate lognormal with $\mu_1 = 0$, $\sigma = 0.6$;

(c) Bivariate Laplace with $\mu_1 = 0$, $\sigma = 1$;

(d) Bivariate Gumbel with $\mu_1 = 1$, $\sigma = 5$.

The test with Cramér–von Mises statistics displays overall better power compared to the test with Kolmogorov–Smirnov statistics (cf. Schmid & Trede, 1995). However, when the correlation is low, the differences in power are smaller and in the cases of the Laplace and the Gumbel distributions, the Kolmogorov–Smirnov test is slightly better.

Simulation results showing the power to accept a given hypothesis of dominance for different sample sizes are depicted in Figure 6. The settings are as follows: $\rho = 0.5$, μ_1 and σ are as above, and for the normal $\mu_2 = 0.35$, for the lognormal $\mu_2 = 0.2$, for the Laplace $\mu_2 = 0.35$, for the

Table 1: Simulation results with bivariate normal distribution, true hypothesis H_0

ρ	Decision	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		CvM	KS	CvM	KS	CvM	KS	CvM	KS
0.8	Retain H_0	0.905	0.933	0.894	0.925	0.892	0.911	0.898	0.915
	Accept $H_{\succ}$	0.014	0.004	0.021	0.008	0.019	0.009	0.022	0.010
	Accept $H_{\prec}$	0.014	0.008	0.020	0.008	0.021	0.009	0.025	0.008
	Accept H_{cr}	0.066	0.055	0.066	0.059	0.068	0.071	0.055	0.067
0.2	Retain H_0	0.907	0.922	0.899	0.918	0.898	0.923	0.915	0.918
	Accept $H_{\succ}$	0.019	0.011	0.019	0.013	0.022	0.015	0.018	0.012
	Accept $H_{\prec}$	0.019	0.012	0.022	0.016	0.022	0.014	0.014	0.011
	Accept H_{cr}	0.056	0.055	0.060	0.053	0.058	0.048	0.052	0.059

Note. $\mu_1 = \mu_2 = 0$, $\sigma_1 = \sigma_2 = 1$. The table reports the empirical probabilities of making each of the four decisions.

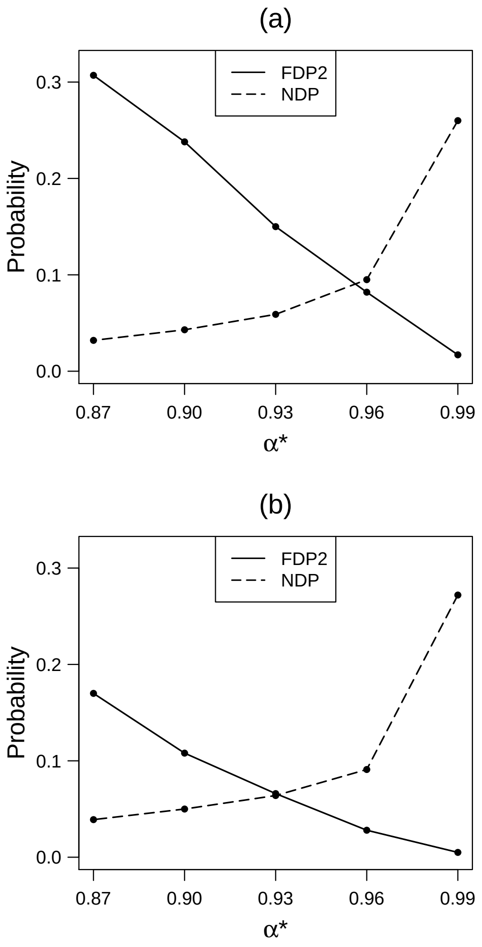

Figure 3: Error probabilities against $\alpha^{\star}$. Panel (a) for normal distribution, panel (b) for lognormal distribution. Solid curves denote FDP$_2$, dashed curves denote NDP. $\rho = 0.8$, $n = 100$.

Gumbel $\mu_2 = 3$. The power increases from 0.40–0.55 for $n = 50$ to approximately one for $n = 500$.

Further simulations concerning the power to detect dominance are reported in Table 3. In Setup 1 (the upper half of the table), the parameter μ_2 is chosen so that the power with the Cramér–von Mises statistics is approximately 0.9. In Setup 2 (the lower half of the table), the choice of parameters allows comparing the power for high and low correlation. As observed before, the Cramér–von Mises statistics lead to better power than the Kolmogorov–Smirnov statistics. This is more noticeable for higher values of the correlation. Overall, it is easier to detect dominance when the correlation between X and Y is higher.

Because the Kolmogorov–Smirnov statistics are based on maximum differences and the Cramér–von Mises statistics are based on average differences, the former may be more powerful when the differences are large only in a small part of the support of the survival functions. In order to illustrate this, we conducted simulations with bivariate normal mixture:

$$X = B\,U_1 + (1 - B)\,V_1,$$
$$Y = B\,U_2 + (1 - B)\,V_2,$$

where $B \sim$ Bernoulli(0.8), (U_1, U_2) is bivariate normal with $(\mu_1, \mu_2) = (0, 0.1)$, $(\sigma_1, \sigma_2) = (1, 1)$, $\rho = 0$, (V_1, V_2) is bivariate normal with $(\mu_1, \mu_2) = (4, 5)$, $(\sigma_1, \sigma_2) = (0.4, 0.4)$, $\rho = 0$. Thus, the correlation between X and Y is approximately 0.8. In this setting, the difference between the two survival functions is more pronounced in the tail (see Figure 7). The simulations with $n = 200$ show that the power of the Kolmogorov–Smirnov test is higher compared to the Cramér–von Mises test (0.378 vs. 0.324).

4. Application

The procedures developed above were applied to data from an experiment conducted in Stockholm. In this experiment, respondents gave answers to a question about

Table 2: Simulation results, true hypothesis H_{cr}

Setting	Decision	$n = 50$		$n = 100$		$n = 200$		$n = 500$	
		CvM	KS	CvM	KS	CvM	KS	CvM	KS
Normal distribution									
(a2)	Retain H_0	0.046	0.105	0.001	0.002	0.000	0.000	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.145	0.175	0.069	0.080	0.017	0.014	0.001	0.000
	Accept H_{cr}	0.809	0.720	0.930	0.918	0.983	0.986	0.999	1.000
(a3)	Retain H_0	0.134	0.202	0.011	0.014	0.000	0.000	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.070	0.084	0.018	0.024	0.002	0.001	0.000	0.000
	Accept H_{cr}	0.796	0.714	0.971	0.962	0.998	0.999	1.000	1.000
Lognormal distribution									
(b2)	Retain H_0	0.289	0.243	0.040	0.019	0.000	0.000	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.003	0.005	0.000	0.000	0.000	0.000	0.000	0.000
	Accept H_{cr}	0.707	0.752	0.960	0.981	1.000	1.000	1.000	1.000
(b3)	Retain H_0	0.013	0.027	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\succ$	0.085	0.126	0.031	0.036	0.003	0.002	0.000	0.000
	Accept $H_\prec$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept H_{cr}	0.902	0.847	0.969	0.964	0.997	0.998	1.000	1.000
Laplace distribution									
(c2)	Retain H_0	0.268	0.384	0.056	0.097	0.000	0.003	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.103	0.119	0.073	0.093	0.025	0.021	0.001	0.001
	Accept H_{cr}	0.630	0.497	0.871	0.810	0.975	0.976	0.999	0.999
(c3)	Retain H_0	0.277	0.360	0.054	0.072	0.000	0.001	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.065	0.085	0.028	0.039	0.004	0.004	0.000	0.000
	Accept H_{cr}	0.659	0.555	0.918	0.888	0.996	0.995	1.000	1.000
Gumbel distribution									
(d2)	Retain H_0	0.679	0.777	0.461	0.577	0.168	0.231	0.002	0.003
	Accept $H_\succ$	0.091	0.060	0.088	0.060	0.060	0.043	0.006	0.005
	Accept $H_\prec$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept H_{cr}	0.230	0.162	0.451	0.363	0.772	0.725	0.992	0.992
(d3)	Retain H_0	0.350	0.464	0.111	0.153	0.006	0.009	0.000	0.000
	Accept $H_\succ$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Accept $H_\prec$	0.137	0.144	0.112	0.133	0.050	0.056	0.007	0.008
	Accept H_{cr}	0.514	0.391	0.777	0.714	0.943	0.935	0.993	0.992

Note. The table reports the empirical probabilities of making each of the four decisions. The results are under the settings depicted in Figure 2 (see the panel titles) with $\rho = 0.8$.

Table 3: Power under $H_{\prec}$

n	Normal				Lognormal			
	μ_2	ρ	CvM	KS	μ_2	ρ	CvM	KS
				Setup 1				
50	0.35	0.8	0.890	0.726	0.21	0.8	0.887	0.711
50	0.60	0.5	0.913	0.864	0.35	0.5	0.903	0.851
50	0.75	0.2	0.900	0.900	0.44	0.2	0.892	0.888
100	0.25	0.8	0.897	0.743	0.15	0.8	0.902	0.747
100	0.40	0.5	0.895	0.826	0.25	0.5	0.906	0.852
100	0.50	0.2	0.885	0.877	0.32	0.2	0.908	0.903
200	0.18	0.8	0.913	0.753	0.11	0.8	0.931	0.785
200	0.28	0.5	0.910	0.835	0.17	0.5	0.912	0.840
200	0.38	0.2	0.921	0.902	0.22	0.2	0.906	0.887
500	0.11	0.8	0.902	0.727	0.07	0.8	0.938	0.789
500	0.18	0.5	0.919	0.849	0.11	0.5	0.923	0.857
500	0.22	0.2	0.894	0.851	0.14	0.2	0.914	0.880
				Setup 2				
50	0.60	0.8	0.997	0.994	0.35	0.8	0.993	0.989
50	0.60	0.2	0.784	0.731	0.35	0.2	0.763	0.727
100	0.40	0.8	0.996	0.989	0.25	0.8	0.996	0.992
100	0.40	0.2	0.754	0.712	0.25	0.2	0.771	0.747
200	0.28	0.8	0.998	0.985	0.17	0.8	0.997	0.983
200	0.28	0.2	0.749	0.695	0.17	0.2	0.743	0.695
500	0.18	0.8	0.999	0.988	0.11	0.8	0.998	0.990
500	0.18	0.2	0.761	0.697	0.11	0.2	0.774	0.706

Note. Normal: $\mu_1 = 0,\ \sigma_1 = \sigma_2 = 1$. Lognormal: $\mu_1 = 0,\ \sigma_1 = \sigma_2 = 0.6$.

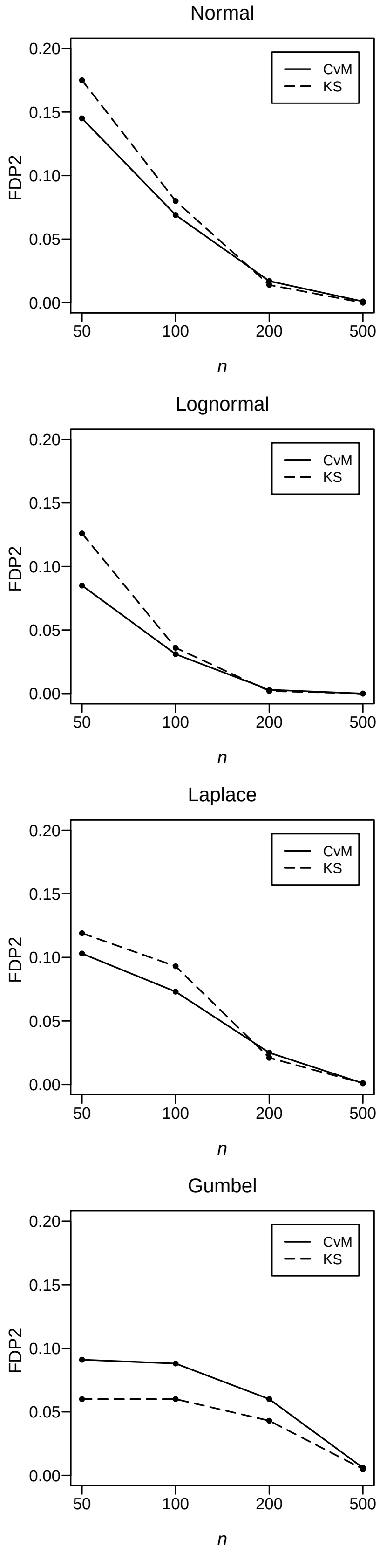

Figure 4: Probability of a false detection of dominance when the truth is H_{cr}. Results for different sample sizes under the settings depicted in Figure 2 (second column) with $\rho = 0.8$.

the willingness to pay for an improved residential sound environment. Each participant was requested to answer the question by means of: (i) a self-selected point (SSP), i.e., the amount in Swedish kronor he/she finds reasonable to pay per month for the improvement, and (ii) a self-selected interval (SSI), i.e., the lowest and highest amounts he/she finds reasonable to pay. Allowing interval answers accounts for the fact that the individual might not know exactly how much he/she would be willing to pay for a given environmental improvement. The purpose of the experiment was twofold: (i) to study whether there is a consistency between SSP and SSI, and (ii) to study whether the answers exhibited an expected dose-response relationship with the degree of noise-exposure reduction.

The scenario was a planned traffic noise reduction that would improve the sound environment in an outdoor living space (e.g., a balcony). The participants of the experiment were seated in a sound laboratory and listened to recordings of outdoor sound environments. Their task was to decide how much they would be willing to pay for a noise reduction that would change a given sound environment with road-traffic noise (quiet-plus-noisy) to an environment without the road traffic noise (quiet). The quiet environment was a recording in a quiet area with no audible noise sources and an average sound pressure level of 45 dB(A). The quiet-plus-noisy environment was a mix of the quiet environment with a recording of distant road-traffic noise. Five scenarios were creates as follows: first, the original road-traffic noise was set to five different levels: 40, 45, 50, 55, and 60 dB(A); then, each of these was mixed with the quiet environment. Thus, five quiet-plus-noisy environments were obtained, with average sound pressure levels of 46, 48, 51, 55, and 60 dB(A), respectively. Hereafter, the five quiet versus quiet-plus-noisy environments are denoted scenario 1, 2, 3, 4, and 5, corresponding to the rank order from smallest to largest noise reduction. In addition to these five scenarios (targets), the experiment involved 26 other quiet and quiet-plus-noisy scenarios (fillers) created from selected outdoor recordings. These were included to mask the experiment's focus on the five scenarios with a systematically increasing noise level.

Each participant was exposed to the 31 environments four times (at four separate sessions); at each session the environments were presented in a random order. The first two sessions were performed on a different day from the other two sessions, where the two days were separated by at least a week. Point answers were given on one day and interval answers on the other day, with the order of answer type counterbalanced across participants, i.e., half of the participants gave point answers on the first day and half of the participants gave point answers on the second day. The dataset contains responses from 60 participants that conducted all four sessions (40 females, 20 males, mean age 29 years). The participants were recruited among students from universities in the Stockholm area.

In each trial of the experiment, the participant could

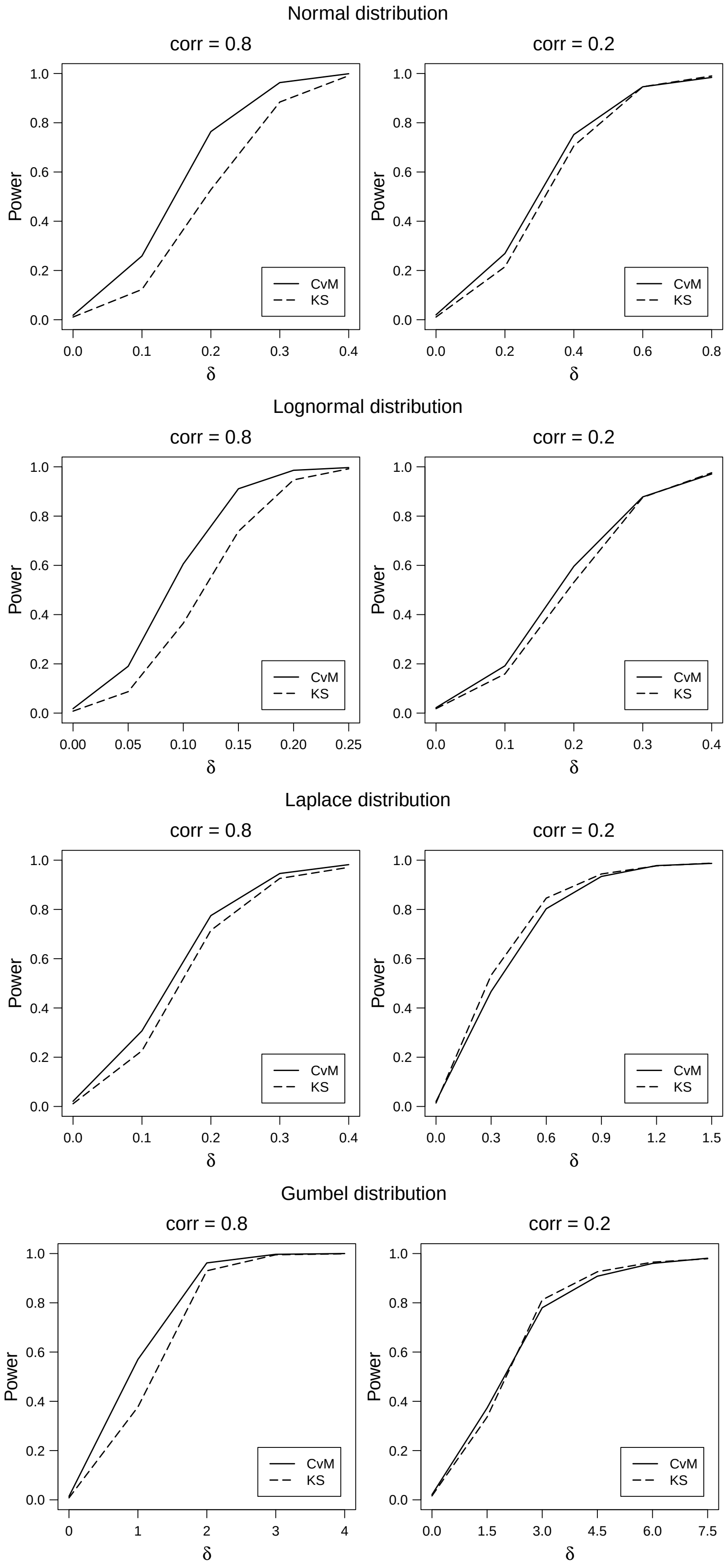

Figure 5: Power as a function of $\delta = \mu_2 - \mu_1$, $n = 100$.

switch between listening to the quiet and the quiet-plus-noisy environments as many times as they liked till they were ready to give an answer. The sounds were presented in a sound laboratory using an ambisonic system with 25 loudspeakers to assure a highly realistic and immersive sound experience. There were no visual stimuli; the participants were asked to imagine that the reproduced sound environment was heard while seated in an outdoor living space at home.

The dataset includes the following variables:

- `pt1` is the point answer at the first SSP session.
- `pt2` is the point answer at the second SSP session.
- `low1` and `upp1` are respectively the lower bound and the upper bound of the interval answered at the first SSI session.
- `low2` and `upp2` are respectively the lower bound and the upper bound of the interval answered at the second SSI session.
- `mid1` is the midpoint of the interval answered at the first SSI session.
- `mid2` is the midpoint of the interval answered at the second SSI session.

Each variable was observed under five scenarios and these are denoted, e.g., $\texttt{pt1}[1], \dots, \texttt{pt1}[5]$. One participant was excluded because of extreme answers: in most cases he/she responded with amounts greater than 4000, while the other participants responded with amounts not greater than 2000. Moreover, in some cases, his/her responses were with a negative dose-response trend, suggesting that he/she had probably misunderstood the instructions. Thus, the analysis is based on 59 participants. Results based on all 60 participants are included in an online supplementary material to this paper.

We focused on comparisons at the distribution level, rather than comparisons at the individual level. Therefore, we utilized the proposed tests for stochastic dominance. The reported results are based on $\alpha = 0.05$, $\alpha^{\star} = 0.96$, and $R = 10000$.

To evaluate consistency between SSP and SSI, we initially plotted the median SSP and the median lower and upper bounds of SSI versus the noise reduction levels (scenarios), see Figure 8. For each level of noise reduction, the median SSP falls between the median lower and upper bounds, i.e., with respect to medians there is a consistency between the two formats of answers. Further, if we compare not just the medians but the survival functions, consistency between SSP and SSI implies that the survival function of SSP lies between the survival functions of the lower and the upper bounds of SSI. In most cases, the performed dominance tests confirm this (see Table 4). In a few cases, however, the conclusion of the test is that the survival functions cross one another, i.e., in these cases there is an inconsistency between SSP and SSI.

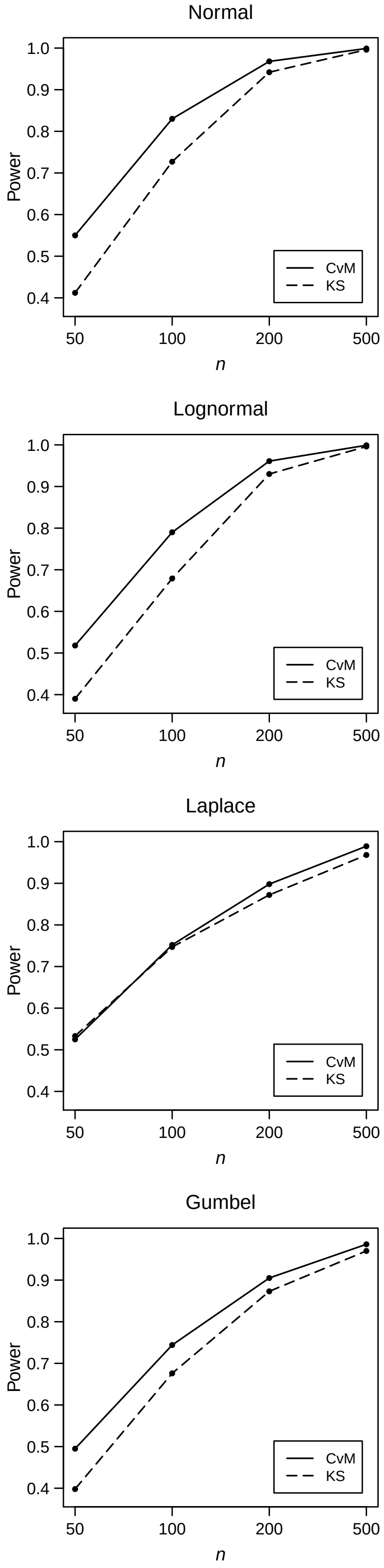

Figure 6: Power for different sample sizes, $\rho = 0.5$.

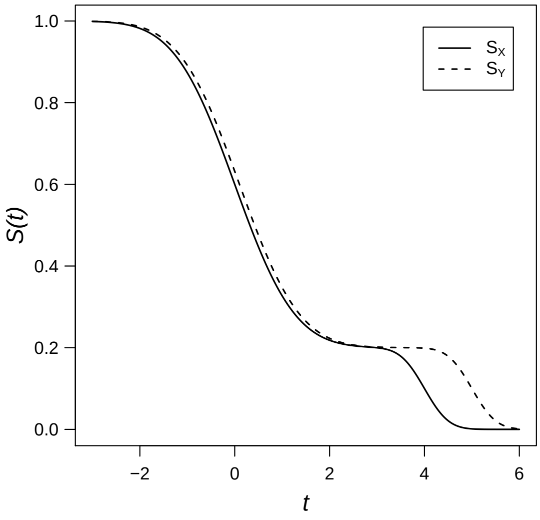

Figure 7: Survival curves: normal mixture.

We expect that the respondents are willing to pay more for higher levels of noise reduction. Thus, the survival function of willingness to pay under scenario 2 should dominate the survival function under scenario 1. Similarly, the survival function under scenario 3 should dominate the survival function under scenario 2, and so on. We performed tests for stochastic dominance based on the observed SSP and the midpoints of SSI (see Table 5). The corresponding empirical survival functions are illustrated in Figure 9 (there is one plot for each pair of variables that are tested for dominance). In most cases, the willingness to pay for the higher level of noise reduction dominates the willingness to pay for the lower level. There are a few cases where the conclusion of the test is that the survival functions cross one another, e.g., `mid1`[2] and `mid1`[3]. However, additional tests show that `mid1`[3] dominates `mid1`[1] and `mid1`[4] dominates `mid1`[2], implying that in this case there is a weaker dose-response relationship. Similarly, we got that `pt1`[1] is dominated by `pt1`[3], which is dominated by `pt1`[5].

In summary, for the most part our stochastic dominance analysis suggests response consistency, with SSP being between the lower and upper bounds of SSI, and a monotonic increase in willingness to pay with amount of noise reduction.

5. Concluding remarks

We suggested permutation-based paired-sample tests for stochastic dominance that involve four hypotheses. Our simulations indicated good power properties and control of false-detection errors. Both the probability to detect dominance and the probability to detect crossings approach one as the sample size increases, implying that the probability to retain a false null hypothesis tends to zero. The Cramér–von Mises statistics provided overall better power than the commonly used Kolmogorov–Smirnov statistics. It would be of interest in future research to develop analogous testing procedures for two independent samples as well as for the case of more than two samples (more than two repeated measurements). The proposed testing procedures rely on less assumptions than the existing asymptotic tests for two independent samples, which assume, e.g., continuous survival functions (cf. Bennett, 2013; Donald & Hsu, 2016). With even more assumptions, e.g., $S_X(t) = S_Y(t - \Delta)$ for some Δ, the problem can be substantially simplified; however, we consider a more general problem.

The empirical example presented here is, to the best of our knowledge, the first use of self-selected intervals in a psychophysical experiment involving a large set of stimuli tested on a sample of participants. The results of our stochastic dominance analysis demonstrated a large degree of group-level consistency in how participants used this question format. This motivates further development of question formats based on intervals as a complement to conventional approaches based on point assessments.

Acknowledgments

The authors would like to thank Xavier de Luna, the consulting editor, and two reviewers for their valuable comments and suggestions.

The work was partially supported by the Jan Wallander and Tom Hedelius Foundation, the Tore Browaldh Foundation (P2015:0196:1) and the Marianne and Marcus Wallenberg Foundation (Project MMW 2017.0075).

Declaration of interest

The authors declare that there are no conflicts of interest related to this paper.

References

Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, *97*, 284–292.

Anderson, G. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica*, *64*, 1183–1193.

Ashby, F. G., Tein, J.-Y., & Balakrishnan, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology*, *37*, 526–555.

Barrett, G. F., & Donald, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica*, *71*, 71–104.

Bennett, C. J. (2008). *Consistent Integral-Type Tests for Stochastic Dominance*. Working paper Vanderbilt University. `https://www.researchgate.net/publication/228799141_Consistent_integral-type_tests_for_stochastic_dominance`.

Bennett, C. J. (2013). Inference for dominance relations. *International Economic Review*, *54*, 1309–1328.

Bishop, J. A., & Formby, J. P. (1999). Tests of significance for Lorenz partial orders. In J. Silber (Ed.), *Handbook of Income Inequality Measurement* (pp. 315–339). Dordrecht: Springer.

Bishop, J. A., Formby, J. P., & Thistle, P. D. (1989). Statistical inference, income distributions, and social welfare. *Research on Economic Inequality*, *1*, 49–82.

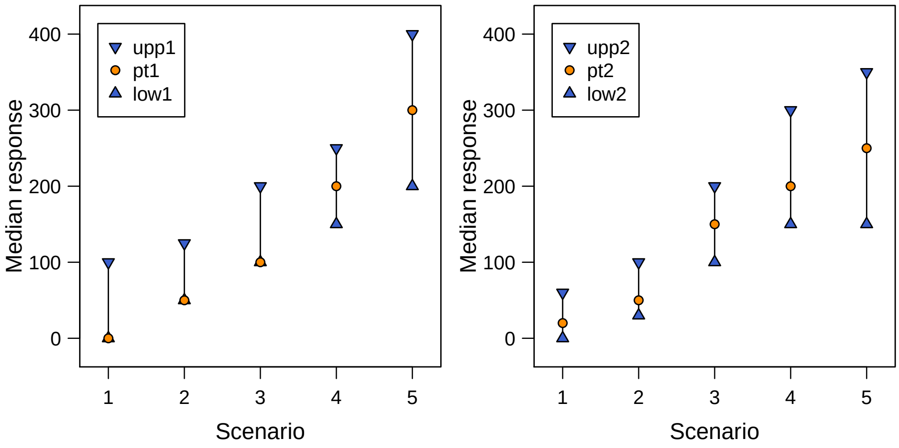

Figure 8: Median response against scenario.

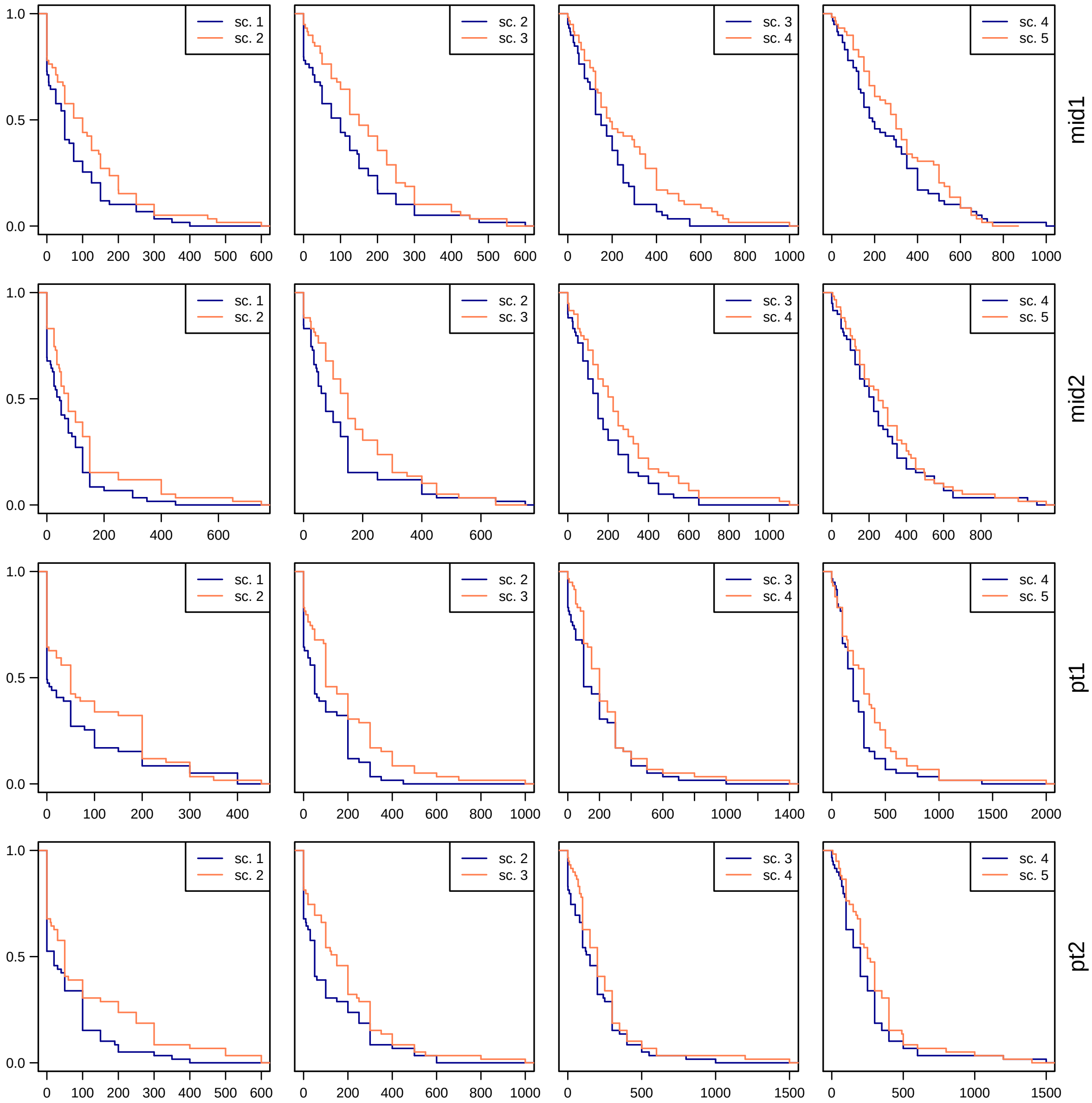

Figure 9: Empirical survival functions.

Table 4: Results: comparison of self-selected points and self-selected intervals

X	Y	CvM			KS		
		Decision	$\widetilde{p}_1$	$\widetilde{p}_2$	Decision	$\widetilde{p}_1$	$\widetilde{p}_2$
low1[1]	pt1[1]	Retain H_0	1.0000	0.0737	Retain H_0	1.0000	0.2548
pt1[1]	upp1[1]	Accept $H_{\prec}$	1.0000	0.0004	Accept $H_{\prec}$	1.0000	0.0022
low1[2]	pt1[2]	Accept $H_{\prec}$	1.0000	0.0437	Accept $H_{\prec}$	1.0000	0.0445
pt1[2]	upp1[2]	Accept $H_{\prec}$	1.0000	0.0004	Accept $H_{\prec}$	1.0000	0.0014
low1[3]	pt1[3]	Accept $H_{\prec}$	1.0000	0.0032	Accept $H_{\prec}$	1.0000	0.0033
pt1[3]	upp1[3]	Accept H_{cr}	0.9461	0.0009	Accept H_{cr}	0.9594	0.0024
low1[4]	pt1[4]	Accept $H_{\prec}$	1.0000	0.0075	Accept $H_{\prec}$	1.0000	0.0066
pt1[4]	upp1[4]	Accept $H_{\prec}$	0.9873	0.0000	Accept $H_{\prec}$	0.9873	0.0001
low1[5]	pt1[5]	Accept $H_{\prec}$	0.9614	0.0030	Accept $H_{\prec}$	0.9745	0.0131
pt1[5]	upp1[5]	Accept H_{cr}	0.9050	0.0001	Accept $H_{\prec}$	0.9723	0.0108
low2[1]	pt2[1]	Retain H_0	1.0000	0.1263	Retain H_0	1.0000	0.1196
pt2[1]	upp2[1]	Accept $H_{\prec}$	1.0000	0.0001	Accept $H_{\prec}$	1.0000	0.0038
low2[2]	pt2[2]	Accept $H_{\prec}$	1.0000	0.0285	Accept $H_{\prec}$	1.0000	0.0372
pt2[2]	upp2[2]	Accept H_{cr}	0.7800	0.0000	Accept H_{cr}	0.7783	0.0000
low2[3]	pt2[3]	Accept $H_{\prec}$	1.0000	0.0007	Accept $H_{\prec}$	1.0000	0.0279
pt2[3]	upp2[3]	Accept $H_{\prec}$	0.9665	0.0054	Accept $H_{\prec}$	0.9821	0.0192
low2[4]	pt2[4]	Accept $H_{\prec}$	1.0000	0.0016	Accept $H_{\prec}$	1.0000	0.0017
pt2[4]	upp2[4]	Accept H_{cr}	0.9239	0.0000	Accept $H_{\prec}$	0.9921	0.0094
low2[5]	pt2[5]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0085
pt2[5]	upp2[5]	Accept $H_{\prec}$	0.9972	0.0000	Accept $H_{\prec}$	0.9989	0.0092

Table 5: Results: comparison of willingness to pay for different levels of noise reduction

X	Y	CvM			KS		
		Decision	$\widetilde{p}_1$	$\widetilde{p}_2$	Decision	$\widetilde{p}_1$	$\widetilde{p}_2$
mid1[1]	mid1[2]	Accept $H_{\prec}$	1.0000	0.0005	Accept $H_{\prec}$	1.0000	0.0050
mid1[2]	mid1[3]	Accept H_{cr}	0.9454	0.0000	Accept H_{cr}	0.9561	0.0061
mid1[3]	mid1[4]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0001
mid1[4]	mid1[5]	Accept $H_{\prec}$	0.9649	0.0009	Accept $H_{\prec}$	0.9975	0.0241
mid2[1]	mid2[2]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0006
mid2[2]	mid2[3]	Accept $H_{\prec}$	0.9768	0.0000	Accept $H_{\prec}$	0.9806	0.0002
mid2[3]	mid2[4]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0021
mid2[4]	mid2[5]	Accept $H_{\prec}$	0.9902	0.0186	Retain H_0	0.9993	0.1941
pt1[1]	pt1[2]	Accept H_{cr}	0.8702	0.0007	Accept H_{cr}	0.8364	0.0076
pt1[2]	pt1[3]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0000
pt1[3]	pt1[4]	Accept $H_{\prec}$	1.0000	0.0010	Accept $H_{\prec}$	1.0000	0.0022
pt1[4]	pt1[5]	Accept H_{cr}	0.8332	0.0001	Accept H_{cr}	0.7775	0.0000
pt2[1]	pt2[2]	Accept $H_{\prec}$	1.0000	0.0005	Accept $H_{\prec}$	1.0000	0.0015
pt2[2]	pt2[3]	Accept $H_{\prec}$	1.0000	0.0000	Accept $H_{\prec}$	1.0000	0.0000
pt2[3]	pt2[4]	Accept $H_{\prec}$	1.0000	0.0043	Accept $H_{\prec}$	1.0000	0.0116
pt2[4]	pt2[5]	Accept $H_{\prec}$	0.9972	0.0000	Accept $H_{\prec}$	0.9972	0.0158

Darling, D. A. (1957). The Kolmogorov–Smirnov, Cramer–von Mises tests. *The Annals of Mathematical Statistics*, *28*, 823–838.

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

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

Durbin, J. (1973). *Distribution Theory for Tests Based on the Sample Distribution Function*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Fitousi, D., & Algom, D. (2018). A system factorial technology analysis of the size congruity effect: Implications for numerical cognition and stochastic modeling. *Journal of Mathematical Psychology*, *84*, 57–73.

Hadar, J., & Russell, W. R. (1969). Rules for ordering uncertain prospects. *The American Economic Review*, *59*, 25–34.

Hanoch, G., & Levy, H. (1969). The efficiency analysis of choices involving risk. *The Review of Economic Studies*, *36*, 335–346.

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology*, *54*, 454–463.

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

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

Hodges, J. L. (1958). The significance probability of the Smirnov two-sample test. *Arkiv för Matematik*, *3*, 469–486.

Houpt, J. W., & Townsend, J. T. (2010). The statistical properties of the Survivor Interaction Contrast. *Journal of Mathematical Psychology*, *54*, 446–453.

Knight, J., & Satchell, S. (2008). Testing for infinite order stochastic dominance with applications to finance, risk and income inequality. *Journal of Economics and Finance*, *32*, 35–46.

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

Ledwina, T., & Wyłupek, G. (2012a). Nonparametric tests for stochastic ordering. *TEST*, *21*, 730–756.

Ledwina, T., & Wyłupek, G. (2012b). Two-sample test against one-sided alternatives. *Scandinavian Journal of Statistics*, *39*, 358–381.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses*. (3rd ed.). New York: Springer.

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

Linton, O., Maasoumi, E., & Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *The Review of Economic Studies*, *72*, 735–765.

Linton, O., Song, K., & Whang, Y.-J. (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics*, *154*, 186–202.

McFadden, D. (1989). Testing for stochastic dominance. In T. B. Fomby, & T. K. Seo (Eds.), *Studies in the Economics of Uncertainty* (pp. 113–134). New York: Springer.

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org.

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

Rothschild, M., & Stiglitz, J. E. (1970). Increasing risk: I. A definition. *Journal of Economic Theory*, *2*, 225–243.

Schmid, F., & Trede, M. (1995). A distribution free test for the two sample problem for general alternatives. *Computational Statistics & Data Analysis*, *20*, 409–419.

Schmid, F., & Trede, M. (1996). Testing for first-order stochastic dominance: A new distribution-free test. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *45*, 371–380.

Statisticat LLC (2018). *LaplacesDemon: Complete Environment for Bayesian Inference*. R package version 16.1.1, https://CRAN.R-project.org/package=LaplacesDemon.

Stephenson, A. G. (2002). evd: Extreme Value Distributions. *R News*, *2*, 31–32.

Tawn, J. A. (1988). Bivariate extreme value theory: Models and estimation. *Biometrika*, *75*, 397–415.

Townsend, J. T. (1990). Truth and consequences of ordinal differences in statistical distributions: Toward a theory of hierarchical inference. *Psychological Bulletin*, *108*, 551–567.

Tse, Y. K., & Zhang, X. (2004). A Monte Carlo investigation of some tests for stochastic dominance. *Journal of Statistical Computation and Simulation*, *74*, 361–378.

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

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*, 129–133.

Whitmore, G. A. (1970). Third-degree stochastic dominance. *The American Economic Review*, *60*, 457–459.

Yang, C.-T., Altieri, N., & Little, D. R. (2018). An examination of parallel versus coactive processing accounts of redundant-target audiovisual signal processing. *Journal of Mathematical Psychology*, *82*, 138–158.