Distribution-free tests of Stochastic Dominance for small samples

Andrew Heathcote[1], Scott Brown[1], EJ Wagenmakers[2] & Ami Eidels[1]

[1]School of Psychology, University of Newcastle, Australia

[2]Department of Psychological Methods, University of Amsterdam, The Netherlands

Corresponding Author: Prof. Andrew Heathcote

Psychology Building, The University of Newcastle, Callaghan, NSW, 2308, Australia.

Phone: 61-2-49216778; FAX: 61-2-49216906; email: andrew.heathcote@newcastle.edu.au

One variable is said to "stochastically dominate" another if the probability of observations smaller than *x* is greater for one variable than the other, for all *x*. Inferring stochastic dominance from data samples is important for many applications of econometrics and experimental psychology, but little is known about the performance of existing inferential methods. Through simulation, we show that three of the most widely used inferential methods are inadequate for use in small samples of the size commonly encountered in many applications (up to 400 observations from each distribution). We develop two new inferential methods that perform very well in a limited, but practically important, case: where the two variables are guaranteed not to be equal in distribution. We also show that extensions of these new methods, and an improved version of an existing method, perform quite well in the original, unlimited case.

Stochastic dominance denotes an order relationship between cumulative distribution functions. A random variably Y is said to "stochastically dominate" [1] a random variable Z when $F_Y(x) \geq F_Z(x)$ for all *x,* with strict inequality for some *x*. F(*x*) is the random variable's cumulative distribution function (hereafter the distribution function), that is, $F(x)=Pr(X \leq x)$. The concept of stochastic dominance has been extensively employed in a range of scientific disciplines including economics, finance, agriculture, marketing and operations research (see Levy, 1992, for a survey) and in areas of psychology including decision making (e.g., Tversky & Kahneman, 1992) and cognitive modelling (e.g., Townsend & Nozawa, 1995). Townsend (1990) discusses the importance of establishing order relationships induced by experimental manipulations. He describes a "dominance hierarchy", whereby higher order types of dominance logically imply lower order types of dominance distributions (but not vice versa) in an almost distribution free manner. For example, dominance at the level of distribution functions (stochastic dominance) entails the same ordering at the level of means and medians.

In this paper we focus on distribution-free inferential tests of stochastic dominance. Distribution-free statistical tests of dominance have, in the main, been developed for econometric applications where sample sizes are large (e.g., income distributions for nations), and so power is high and asymptotic approximations hold to a good degree. In many practical applications, in contrast, sample sizes can be severely limited. As well as reducing statistical power, small sample sizes force a tradeoff for "histogram" based tests, between the applicability of asymptotic approximations and the resolution with which the distribution function is measured (e.g., Anderson, 1996; Davidson & Duclos, 2000; see Appendix A). These tests characterise distributions by counting observations falling into bins (adjacent ranges).

Narrower bins provide a more accurate characterization of the distribution function, but also reduce the counts in each bin, making the application of asymptotic results questionable. In the first part of this paper we investigate the power in various, relatively small, sample sizes of the two histogram based tests, and a third test based on a finer grained representation based on the empirical cumulative distribution function, the Kolmogorov-Smirnov test.

In the second part of this paper we examine two new tests that apply when the researcher wishes to discriminate between three alternatives: 1) dominance of Y over Z, which we annotate as $Y >_S Z$ (see Figure 1b for an example), 2) dominance of Z over Y ($Z >_S Y$) and 3) non-dominance, that is, the distribution function F_Y is greater than F_Z at some x values and, and less than F_Z for other x values (which we denote $Y <>_s Z$, the lower panels of Figure 1 provide examples of non-dominance). Note that this trichotomy does not include the null case, where Y and Z have exactly the same distribution ($Y =_s Z$). An example from psychology where such three-outcome tests are applicable concerns experiments that identify "mental architecture" (e.g., serial vs. parallel arrangements of cognitive processes). Techniques used in this application assume a stochastically dominant selective influence of some experimental manipulation on the completion time of a cognitive process (e.g., Townsend & Nozawa, 1995; see also Dzhafarov, Schweickert, & Sung, 2004; Eidels, Townsend, & Pomerantz, 2008; Schweickert, Giorgini & Dzhafarov, 2000; Townsend & Thomas, 1994). In any well-designed experiment for this application it seems reasonable to rule out the null case (i.e., completely ineffectual manipulations) *a priori*. The null case is sometimes irrelevant in economic applications as well – for example, stochastic dominance of the outcomes of one investment over another is an important property, but there is no measurable chance that two distinct

investments have identical probabilities for all outcomes. As we show, the benefit of not considering the null case is a substantial increase in power associated with identifying the remaining three cases.

In the final part of this paper we propose extensions of one of the novel three-choice tests and one of the existing tests to the case where the null is not ruled out *a priori*. That is, the extended tests, like the established tests we examine initially, choose between four alternatives: $Y >_s Z$, $Y <_s Z$, $Y <>_s Z$ and $Y =_s Z$. We demonstrate that both extensions have greater power than the existing tests examined in the first part of the paper in identifying non-dominance.

Overview of Test Evaluations

The evaluations of test power reported in this paper were carried out via simulation studies in which we compared samples from two normal distributions in four different conditions, illustrated in Figure 1 (we discuss further simulation studies using other distributional forms below). In each condition, we compared a reference random variable $Y \sim N(0,1)$ with a random variable $Z \sim N(\mu,\sigma^2)$. In the null case (Figure 1a), the comparison distribution was the same as the reference distribution, with $\mu=0$ and $\sigma=1$. In the stochastically dominant case (Figure 1b), the comparison distribution has $\mu=0.5$ and $\sigma=1$, meaning that $Y >_S Z$. There were also two non-dominant cases ($Y <>_S Z$): "central" non-dominance (Figure 1c: $\mu=0$ and $\sigma=1.5$), where the distributions cross once at the mean, and "tail" non-dominance (Figure 1d: $\mu=0.5$ and $\sigma=1.5$), where the distributions cross once in the left tail. For each of the four conditions, we examined test performance with a range of sample sizes, N=50, 100, 200 and 400. In each combination of sample size and comparison distribution, we produced 1024

replicates, sufficient to make Monte Carlo error negligible. Note that we did not consider the converse of the cases in Figure 1b and 1d (i.e., $Y <_S Z$ and crossing in the right tail respectively) as they produce identical results, due to the symmetry of the normal distribution.

As will be shown, tail non-dominance is particularly difficult to identify in data. This happens because, for most of the data range, one distribution function is greater than the other. This means that small samples of data often suggest stochastic dominance, rather than non-dominance, and contrary evidence occurs only in one tail. For example, in Figure 1d the crossover occurs at around $x = -1$, and $F_Y(-1) \approx 0.16$, so contrary evidence is usually available for only 16% of the sample. Central non-dominance is usually easier to identify because the amount of evidence in one direction and the other tend to be balanced. Stochastic dominance is even easier to identify as evidence consistent with the dominant ordering tends to be available over the whole domain, although it is weaker near the ends for unbounded random variables as the two distribution functions must eventually tend to equivalence.

Some important applications of the tests we examine are to distributions which are positively skewed and bounded below; for example, response time distributions, and income or wealth distributions. To investigate whether our results extend to such cases we also evaluated test performance with the Weibull distribution. The results were essentially the same as with normal distributions, so we present these evaluations in Appendix B. We have not formally investigated any non-dominant cases where the distribution functions cross more than once, but limited informal investigation suggested that all tests acted as might be expected; power tends to reduce as more crossings occur because stronger evidence in either direction becomes less common.

Before describing and evaluating the tests we emphasize that a graphical examination of the data is always advisable as a prelude to inferential analysis. Most statistical packages provide routines to plot empirical cumulative distribution function (ECDF) estimates. The ECDF estimator for sample y_1, y_2…y_N is:

$$\hat{F}_Y(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x \le y_i) \tag{1}$$

The function **1**() equals 1 if its argument is true and zero otherwise. Figure 2 illustrates such a plot using samples of 50 observations from normal distributions with parameters corresponding to Figure 1d. The challenging nature of tail non-dominance detection in small samples is clearly illustrated.

Existing Tests

Tse and Zhang (2003) reported a simulation study comparing three stochastic dominance tests, including the highly cited test developed by Anderson (1996), as well as tests proposed by Davidson and Duclos (2000) and Kaur, Rao and Singh (1994). They found that the Davidson-Duclos test performed best, with the Kaur-Rao-Singh test being overly conservative and less powerful than the other tests. In light of their results, and the wide spread use of Anderson's test, we focused on the Anderson and Davidson-Duclos tests (see Appendix A for details of the test calculations).

Anderson and Davidson-Duclos tests

Both of these tests are of the "histogram" type and so require a partition of the data range into K+1 regions ("bins") by specifying K cut points (x_1,…,x_K). As the tests are based on asymptotic normal approximations at least 5 observations are recommended in each bin. In

theory, the cut points should be chosen without reference to the data. In practice it is convenient to use evenly spaced quantiles calculated from the union of the two data samples, so each bin contains approximately the same number of observations. Figure 2 illustrates this approach showing K=4 cut points placed at the 20th, 40th, 60th and 80th percentiles as vertical dotted lines. In investigations not reported here, we did not find any substantive difference in using fixed or data dependent partitions.

The cut points in Figure 2 illustrate a problem caused by wide intervals: inconsistency, that is, a failure to converge on the true answer as sample size increases. A test based on the intervals in Figure 2 would be unlikely to detect the reversal in the lower tail even with large sample sizes. In all of the evaluations reported here we divided the data using semi-deciles (i.e., 5^{th}, 10^{th} … 95^{th} percentiles, with K=19) to minimize such inconsistency. However, we note that in small samples this strategy can lead to very few observations in some bins. We also repeated all our simulations using the widely spaced cut-points from Figure 2 (i.e., the 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles). Those simulations showed that performance was very similar to the semi-decile binning, but with slightly improved accuracy at detecting stochastic dominance in smaller samples, and slightly decreased accuracy at detecting tail non-dominance in all sample sizes.

We adopted the method outlined by Tse and Zhang (2003) to choose between four possible outcomes of the Davidson-Duclos and Anderson tests. Each test produces K test statistics, $T(x_i)$, one for each cut point. The Type 1 error (α) for the overall null is controlled by comparing each to the "studentized maximum modulus statistic" for K and ∞ degrees of freedom, $M^{K}_{\infty,\alpha}$, which was tabulated by Stoline and Ury (1979) for K<20. The overall null hypothesis, which we denote $H_{Y=Z}$, is the logical intersection of the K null hypotheses over x_i.

Similarly, the overall alternative hypotheses for non-dominance ($H_{Y<>Z}$) and for dominance ($H_{Y>Z}$ and $H_{Y>Z}$) are the logical union of the alternative hypotheses. One of these four mutually exclusive hypotheses is chosen as follows:

$$\text{If } |T(x_i)| < M^K_{\infty,\alpha} \text{ for } i = 1 \ldots K \text{ do not reject } \mathrm{H}_{\mathrm{Y=Z}}$$
$$\text{If } -T(x_i) > M^K_{\infty,\alpha} \text{ for some } i \text{ and } T(x_i) < M^K_{\infty,\alpha} \text{ for all } i, \text{ accept } \mathrm{H}_{\mathrm{Y>Z}}$$
$$\text{If } T(x_i) > M^K_{\infty,\alpha} \text{ for some } i \text{ and } -T(x_i) < M^K_{\infty,\alpha} \text{ for all } i, \text{ accept } \mathrm{H}_{\mathrm{Z>Y}}$$
$$\text{If } T(x_i) > M^K_{\infty,\alpha} \text{ for some } i \text{ and } -T(x_i) > M^K_{\infty,\alpha} \text{ for some } i, \text{ accept } \mathrm{H}_{\mathrm{Y<>N}}$$

Kolmogorov-Smirnov test

We also evaluated a third existing test, the Kolmogorov-Smirnov (KS) test, developed by Kolmogorov (1933) for the one sample case and extended by Smirnov (1939) to the two sample case that is relevant here (see Johnson, Blaha, Houpt & Townsend, 2010, for a recent application). The KS test is not subject to consistency problems caused by wide binning, because it is based on the ECDF, and so represents the distribution function with the maximum possible resolution given the sample size. To test $Y >_s Z$ the test uses a statistic proportional to the largest positive difference between the two CDFs:

$$T^{Y>Z} = \sqrt{\frac{n_Y n_Z}{n_Y + n_Z}} \sup_{x \in \Re} \left(\hat{F}_Y(x) - \hat{F}_Z(x) \right) \tag{2}$$

Similarly, to test $Z >_s Y$, (2) is used with $F_Z(x)$ and $F_Y(x)$ swapping roles. If $Z >_s Y$ then $F_Z(x)-F_Y(x)$ will tend to be large, whereas if if $Y >_s Z$ then $F_Y(x)-F_Z(x)$ will tend to be large. In the null case, when $Y =_S Z$, both differences will tend to be small. In the non-dominant case, where $Z <>_S Y$, both differences will tend to be large, although with tail non-dominance one of the differences may be smaller.

The cumulative distribution function of (2) is $\Pr\left(T^{Y>Z} \geq t\right) = e^{-2t^2}$ in the limit as $n_Y, n_Z \to \infty$ (Doob, 1949). We found the limiting distribution to be a very accurate for all of the sample sizes we examined. Hence the limiting distribution was used to obtain a vector of right-tail probabilities, p*, corresponding to observations $t^{Y>Z}$ and $t^{Z>Y}$: $p^* = \exp\left(-2\left(t^{Y>Z}, t^{Z>Y}\right)^2\right)$. Assuming a pre-set Type I error rate, α, the outcome of the test can be decided by comparing the two elements of p* against α as follows:

1. Fail to reject $H_{Y=Z}$ if both elements of p*>α.
2. Accept $H_{Y>Z}$ if only the first element of p*>α (i.e., the element corresponding to $T^{Y>Z}$).
3. Accept $H_{Z>Y}$ if only the second element of p*>α (i.e., the element corresponding to $T^{Z>Y}$).
4. Accept $H_{Y<>Z}$ if both elements of p*>α.

For all simulations, we used α=.05. Figure 3 shows power (i.e., the probability of choosing the data generating model). All tests were calibrated for the null, detecting the truth around 95% of the time, for all sample sizes. All tests also did well detecting dominance in larger sample sizes (N=400) with the KS test clearly best for smaller samples. However, no test ever detected tail non-dominance (in over 12,000 attempts), and all three tests were poor at detecting central non-dominance. Below, we propose a modified version of the KS test that improves its poor performance with the non-dominant cases.

Three-choice tests

In some applications it makes sense to reject the null hypothesis *a priori*. For these situations we propose tests which choose amongst only $H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$. Our first test is based on Klugkist, Kato and Hoijtink's (2005) Bayesian "encompassing prior" approach to

testing hypotheses about orders. We know of one other Bayesian test for stochastic dominance (Chotikapanich & Griffiths, 2006), but we do not examine that test in detail, as it requires parametric assumptions about the distributions (our test is non-parametric in the sense that we model the data by a multinomial distribution at the histogram level). Unlike the histogram tests described earlier, consistency problems in the encompassing prior Bayesian test can be minimized by using a large number of cut points, as the test does not depend on large-sample approximations.

Encompassing Prior Bayesian Test

In overview, the Bayesian test divides the data into a number of bins, and determines the proportion of prior and posterior probability estimates which conform to the joint order constraints dictated by each hypothesis. Denote an observed frequency vector as n_{ij} and corresponding population probabilities p_{ij}, where the index *i*=1…K+1 denotes the bin and *j*=Y,Z denotes the population from which the sample is drawn. We assume the frequencies follow a multinomial distribution, with density:

$$f(n_j \mid p_j) = \frac{N_j!}{\prod_{i=1}^{K+1} n_{ij}!} \prod_{i=1}^{K+1} p_{ij}^{n_{ij}}; \quad N_j = \sum_{i=1}^{K+1} n_{ij} \tag{3}$$

We further assume a Dirichlet prior with parameter vector β, Dir(β), with density (note that Γ is the Gamma function):

$$f(p_j \mid \beta_j) = \frac{1}{B(\beta)} \prod_{i=1}^{K+1} p_{ij}^{\beta_i - 1}; \quad B(\beta) = \frac{\prod_{i=1}^{K+1} \Gamma(\beta_i)}{\Gamma(\sum_{i=1}^{K+1} \beta_i)} \tag{4}$$

The Dirichlet prior is conjugate to the multinomial, which means that the posterior distribution of *p* is also has a Dirichlet distribution; Dir(β+n). Ferguson (1973) made an early application of the Dirichlet process to non-parametric distribution modeling, and argued that it was

appropriate for use with continuous distributions because a continuous distribution can be approximated arbitrarily well by the discrete multinomial distribution.. Finally, we assume that the β_i parameters for each bin are equal and sum to one: $\sum_{i=1}^{K+1}\beta_i = 1$, which causes the prior to have a minimal influence on posterior estimates. In particular, the prior has an influence equal to one observation.

We choose amongst the three hypotheses ($H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$) using Bayes factors. A Bayes factor (BF, see Kass & Raftery, 1995) is the ratio of the marginal probability of the observed data, **D**, given one hypothesis (H_i) divided by that marginal probability of the observed data given another hypothesis (H_k): $BF_{ik}=m(\mathbf{D}|H_i)/m(\mathbf{D}|H_k)$. The marginal probability equals the likelihood of the data given a model with parameters $\boldsymbol{\theta}$ (e.g., the multinomial p_{ij} parameters in our application), $f(\mathbf{D}|M,\boldsymbol{\theta})$, integrated over the prior probability distribution of the parameters, $p(\boldsymbol{\theta}|H)$: $m(\mathbf{D}|H) = \int f(\mathbf{D}|H,\boldsymbol{\theta})\, p(\boldsymbol{\theta}|H)\, d(\boldsymbol{\theta})$. The Bayes factor quantifies the evidence that the data provide for one model versus another, and it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648).

Bayes factors for the hypotheses $H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$ were evaluated relative to an "encompassing" hypothesis, which in our case is simply the unconstrained multinomial. Each of the hypotheses of interest is a special case of the encompassing hypothesis which constrains sums of the estimated prior or posterior multinomial parameters to follow a particular order. For $H_{Y>Z}$ the constraint, for all *i,* is that $\sum_{j=1}^{i}\hat{p}_{jY} > \sum_{j=1}^{i}\hat{p}_{jZ}$, whereas for $H_{Z>Y}$ the constraint, for all *i,* is that $\sum_{j=1}^{i}\hat{p}_{jZ} > \sum_{j=1}^{i}\hat{p}_{jY}$. For $H_{Y<>Z}$ the constraint is that $\sum_{j=1}^{i}\hat{p}_{jY} > \sum_{j=1}^{i}\hat{p}_{jZ}$ for some *i* and $\sum_{j=1}^{i}\hat{p}_{jZ} > \sum_{j=1}^{i}\hat{p}_{jY}$ for the remaining *i*.

Although this characterization makes it clear that, all other things being equal, it is easier to fulfil the non-dominant model's order constraint (i.e., it is a more flexible or complex model), as Myung, Karabatsos and Iverson (2008) state: "Bayes factor-based model selection automatically adjusts for model complexity and avoids overfitting, thereby representing a formal implementation of Occam's razor." (p. 6). Liu and Aitkin (2008) raised concerns about undue influence from the prior when using Bayes factors to select between models with different parameterizations. In the present context, all of the models (hypotheses) have the same parameterization, differing only in the order constraints amongst those parameters. As discussed by Klugkist et al. (2005) this results in a negligible influence of the prior on the Bayes factor for reasonable prior choices. We confirmed this to be true in our application by informal numerical investigations (not reported here).

Estimation of Bayes factors is usually computationally difficult in high dimensional models, such as those considered here, because high-dimensional integration is required. Klugkist et al.'s (2005) method avoids this difficulty by estimating Bayes factors based on simple counts of prior and posterior parameter estimates that conform to the order constraints of a hypothesis. Note that all estimates automatically conform to the encompassing hypothesis, because it was defined to be unconstrained. The algorithm for estimating the BF for each of the constrained hypotheses ($H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$) vs. the unconstrained (encompassing) hypothesis takes the following form:

1. Take samples from the encompassing prior, Dir(β), and posterior, Dir(β+n)
2. Count the proportion of prior (π) and posterior (Π) samples that conform to the order dictated by each of the three constrained hypothesis

3. Calculate $BF_{Y>Z}= \Pi_{Y>Z}/\pi_{Y>Z}$, $BF_{Z>Y}= \Pi_{Z>Y}/\pi_{Z>Y}$ and $BF_{Y<>Z}= \Pi_{Y<>Z}/\pi_{Y<>Z}$ (note that $\Pi=\pi=1$, by definition, for the encompassing hypothesis)

The relative evidence for each hypothesis can be quantified by its posterior model probability. For example, $Pr(H_{Y>Z})= BF_{Y>Z} /(BF_{Y>Z}+BF_{Z>Y}+BF_{Y<>Z})$. In our evaluations we simply chose as the test outcome the hypothesis with the largest BF.

In all of the simulations reported here we used a large number of bins created by K=N-1 cut points, where N is the number of observations in the combined samples. Cut points were placed at the average of each pair of order statistics for the combined sample (i.e., the average of the smallest and 2nd smallest values, the average of the 2nd and 3rd smallest values and so on). This ensures that, for the combined sample, each of the K bins contained exactly one datum, and so each individual sample (Y or Z) has either one or no observations in each bin. Our chosen binning method creates a dependence between the two samples (i.e., in any particular bin, a zero count for Y implies a one count for Z and vice versa). Alternative binning schemes, such as using a fixed number of quantile defined bins or bins defined without reference to the data, can remove this dependence. However, they did not produce different results to those presented below, except when the number of bins was small, which caused consistency problems. We present results for our chosen binning method because it is easy to apply and minimizes consistency problems.

We found the computational cost of obtaining accurate Bayes factor estimates was similar to that of obtaining accurate bootstrap test estimates[2], around 20 seconds for data samples of size N=400, on a standard desktop computer. Note that no Markov Chain Monte Carlo sampling is required; independent samples can be obtained directly from the Dirichlet

distribution through evaluating the Gamma function, which has fast and accurate numerical approximations.

Minimally Dominant (MD) Bootstrap Test

The second three-choice test we develop avoids the null case by making stochastic dominance the null hypothesis – bootstrap samples are drawn from a distribution created by a minimal adjustment of the observed data to fulfil dominance. We call this test the minimally dominant (MD) bootstrap test. The distribution from which bootstrap samples are drawn is constructed to ensure that either $Y >_s Z$ or $Z >_s Y$, according to the following two part algorithm. We assume that each sample is of equal size N, and that ties are broken randomly. The first step in this test is to identify whether the observed samples are closer to fulfilling $Y >_s Z$, or $Z >_s Y$ (the other is rejected):

1. Sort s_Y and s_Z (i.e., get the order statistic vectors o_Y and o_Z)
2. Calculate $O_{Y>Z} = \sum_{i=1}^{N} 1(o_{i,Y} > o_{i,Z})$
3. If $O_{Y>Z} > N-O_{Y>Z}$ reject $H_{Z>Y}$, otherwise reject $H_{Y>Z}$

This results in one dominance hypothesis being selected to play the role of the null, call that hypothesis $H_>$. The second step chooses between the selected dominance hypothesis and non-dominance, $H_{Y<>Z.}$ Denote the count of orders consistent with the selected hypothesis as $O_>$, and let B be the number of bootstrap samples.

1. Create data samples $x^>_Y$ and $x^>_Z$ that accord with $H_>$ by swapping the (minority) of order statistic pairs that violate it. That is, for those *i* where $o_{i,Y}>o_{i,Z}$, swap $o_{i,Y}$ with $o_{i,Z}$ (if $H_>$ is $H_{Y>Z}$, and vice versa otherwise)

2. Set b=1
3. Resample from $x^{>}_{Y}$ and $x^{>}_{Z}$ and use these to calculate $O^{*}_{>,b}$
4. Set b=b+1
5. If (b<B) go to #3.
6. Accept $H_{Y<>Z}$ if $O_{>}$ falls above the (1-α) quantile of the distribution of $\{O^{*}_{>,b}: b=1,..,B\}$ and otherwise accept $H_{>}$

Thus, the MD test first exchanges the minimal number of data values between samples Y and Z until perfect stochastic dominance has been achieved. Next, this perfectly dominant set of samples is used to perform bootstrap draws; for each bootstrap draw, we tally the number of changes needed to again achieve perfect dominance. This yields a distribution of the number of required changes to achieve stochastic dominance, under the null hypothesis that the data are minimally stochastically dominant. The choice of the minimally dominant null hypothesis was motivated by earlier work (Hall & van Keilegom, 2005); the advantage of this null is that it provides a well specified hypothesis about dominance that is as close as possible to the observed data. In our test evaluations the number of bootstrap samples used for a MD test was chosen so that the width of the 99% credible interval for $p^{*} = \frac{1}{B}\sum_{b=1}^{B} 1\left(O^{*}_{>,b} > O_{>}\right)$ was less than .005, assuming a uniform prior. Hence, the credible interval was calculated by taking the difference between the 99.5% and 0.5% points of the Beta(Bp*+1,B(1-p*)+1) posterior distribution of the p* estimate.

We evaluated the performance of both our new three-choice tests in the same way as for the existing tests. Figure 4 shows the results, using the same format as for Figure 3 (except that the upper left panel is now empty, since we no longer test the null hypothesis, $Y=_{S} Z$). Both

three-choice tests are markedly superior to all of the existing tests in detecting both dominance and non-dominance. Even in the extremely difficult case – tail non-dominance (Panel d) – performance is quite good for larger sample sizes. Setting aside any benefit due to the particular tests used, there is clearly a large advantage in power that attends not having to consider the null case.

Note that Figure 4 shows results for the minimally dominant bootstrap tests using α=.2 ("2") and α=.5 ("5"). We examined a range of α values, and found none worked in all situations. As shown in Figure 4, for example, α=.5 worked well in the non-dominant cases and α=.2 in the dominant case; this is to be expected given that α sets a bias towards accepting the hypothesis of non-dominance. Of course, in applications one can never know the truth, so some automatic calibration procedure would be necessary to set α for each test. In the absence of such a procedure, we attempted to extend only the Bayesian test to the four-choice situation (i.e., including the null). We describe this extension, along with a modification that improves the performance of McFadden's (1989) test in detecting non-dominance, in the next section.

New four-choice tests

Our evaluation of existing tests revealed particularly low power to detect non-dominance. To address this issue we propose an adjustment to the KS test, which performed best in the detection of dominance. The adjustment involves adopting one criterion value for rejecting the null hypothesis, and a different criterion for accepting the remaining three alternatives. This was motivated by the observation from Figure 3, that KS test is biased against non-dominance. The second criterion requires knowledge of the quantiles of the distribution of $T^{Y>Z}$ conditional on the value of $T^{Z>Y}$. As we do not know of an analytic result for this conditional

distribution we first estimated the joint distribution using a bootstrap algorithm outlined by Abadie (2002) based on the work of McFadden (1989). Let $T=(T^{Y>Z}, T^{Z>Y})$ denote the pair of test statistics and $s=(s_Y, s_Z)$ the union of the two observed samples from Y and Z. Draw B bootstrap samples using the following algorithm:

1. Calculate the pair of statistics T for the original samples
2. Set b=1
3. Resample s^* (with replacement) from s, divide randomly into s_Y^* and s_Z^* of sizes equal to s_Y and s_Z, and calculate the pair of statistics T_b^*
4. Set b=b+1
5. If (b<B) go to #3.
6. Calculate a pair of *p*-values corresponding to T: $p^* = \sum_{b=1}^{B} 1\left(T_b^* > T\right) \Big/ B$

The number of bootstrap samples was determined in the same way as for the MD test.

Given the set of bootstrap samples we propose the following test, which we describe as the adjusted McFadden test:

1. Test the null as before. Call the critical value corresponding to α, calculated as above, c_1. If the null is not rejected, stop.
2. Otherwise use a new critical value, c_2 to decide which of $H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$ to accept. Select c_2 to equate the probability of $H_{Y>Z}$, $H_{Z>Y}$ and $H_{Y<>Z}$ under the bootstrap null distribution. That is, set c_2 to be the quantile of the bootstrap distribution corresponding to a lower tail probability of $\alpha \Big/ \left(3\sqrt{1-\alpha}\right)$.
3. Accept $H_{Y>Z}$ if $T^{Y>Z} > c_1$ and $T^{Z>Y} < c_2$.

4. Accept $H_{Z>Y}$ if $T^{Z>Y}>c_1$ and $T^{Y>Z}<c_2$.
5. Accept $H_{Y<>Z}$ $T^{Y>Z}>c_2$ and $T^{Z>Y}>c_2$.

We also examined a number of ways of extending the Bayes test to four choices. One possibility is to convert the set of three Bayes Factors to posterior model probabilities, $p_h = BF_h \Big/ \sum_{i=1}^{3} BF_i$, and fail to reject the null if no probability is above a critical probability, α. However, we found that this option performed poorly and that to obtain the best (but still not good) performance, calibration of α was required on a case-by-case basis. A second possibility follows a suggestion by Klugkist, Laudy and Hoijtink (2005), counting prior and posterior samples as favoring an equality hypothesis if they are equal within some tolerance ±δ. We examined a range of δ values and found that detection of dominance was not much affected relative to the performance with the three-choice version. However, detection of the null was poor for smaller values of δ. For larger values, detection of the null improved somewhat but was accompanied by a large degradation in detection of non-dominance. Further, larger values of δ caused detection of non-dominance to become inconsistent, with power decreasing for larger sample sizes. Given these results we did not investigate this approach further, although that does not rule out the possibility that other related approaches may be more effective (e.g., Wetzels, Grasman & Wagenmakers, in press).

Our final approach involved a hybrid test; first applying the KS test of the null, followed by the Bayesian three-choice test in cases where the null was rejected. Figure 5 compares performance of this approach (denoted B-K) and the adjusted McFadden bootstrap test (denoted M-A). As shown in Figure 5, the performance of both tests was almost identical. As would be expected from the results for existing tests in Figure 3, both tests were well calibrated

for the null. The adjusted McFadden test performed similarly to the original for the dominant case, but the performance of the hybrid Bayes test was worse, at least for the two smaller sample sizes. For the central non-dominance case, the adjusted McFadden test performed markedly better than the KS test (see Figure 3), whereas the hybrid Bayes test was clearly worse than the three-choice version for all but the largest sample size (see Figure 4). The results for the hybrid Bayes test illustrate the marked gains in power that can be achieved if it is possible to rule out the null *a priori*.

Finally, the performance of the adjusted McFadden test in the tail non-dominance case was vastly improved on the original KS test (see Figure 3), but remains clearly worse than the hybrid Bayes test for the smallest sample size. The performance of the hybrid Bayes test was equivalent to that of the three choice Bayes test with tail non-dominance, indicating that there was little or no cost associated with testing the null.

General Discussion

The aim of this paper was to evaluate and develop tests of stochastic dominance and non-dominance suited to smaller samples. To that end we compared the performance in samples ranging from 50-400 of three existing tests (Anderson, 1996; Davidson & Duclos, 2000; Kolmogorov, 1933, and Smirnov, 1939), two new three-choice tests (a Minimally Dominant bootstrap test and an encompassing prior Baysian test, Klugkist et al., 2005), and two new four-choice tests (variations of McFadden's, 1989, test and the encompassing Baysian test). The performance of the Anderson and Davidson-Duclos tests was clearly inadequate for all but the largest sample size. The KS test had adequate power (operationalized as around 0.8 or better) in detecting the dominant case for 100 or more observations, and the others for 200 or more

observations, but no existing test had adequate power in detecting either non-dominant case at any sample size. These results indicate that the existing tests are unlikely to be adequate for small sample applications. Some improvements could be gained by using less strict significance criteria, but only at the costs of decreased power in the null case.

Fortunately, the new tests that we proposed fared much better. Two of these tests apply when the null case can be ruled out *a priori*, which is often reasonable in practical applications. For these tests, power was generally more than adequate even at the smallest sample size when detecting dominance and central non-dominance. For the difficult tail non-dominance case, power was generally adequate for sample sizes of 100 or more. For practical application we recommend the Bayesian test over the Minimally Dominant (MD) bootstrap test, as it requires no calibration. In contrast, we found that the performance of the MD bootstrap test was sensitive to the choice of critical value, so some method of choosing the appropriate value is required for general application. A further advantage of the Bayesian test is that it is relatively straightforward to extend to more complicated settings (e.g., hierarchical models and tests of higher-order dominance, see e.g., Levy, 1992)

We also proposed two new tests for the situation in which the null cannot be ruled out *a priori*. One of these tests adjusts the decision procedure proposed by Abadie (2002) for use with McFadden's (1989) test statistic. The other combines the Bayesian three-choice test with a first stage based on a KS test of the of the null. Overall, this hybrid Bayesian test performed a little better, although its advantage was restricted to the tail non-dominance case at the smallest sample size. Both tests combined control of Type 1 error at the nominal level with adequate power for sample sizes of 100 or more for detecting dominance and 100-200 for

detecting both types of non-dominance. These new four-choice tests provided much greater power for the non-dominant cases compared with the existing tests, but still their power was less than that of the three-choice Bayesian test. Hence, we recommend that researchers carefully consider whether it is reasonable to reject the null *a priori* in their application.

Hitherto, we reported the results of test evaluations conducted on data sampled from normal distributions. Economists and psychologists are often interested in random variables which are non-negative and positively skewed (e.g., income and response times). In Appendix B we report the results of a parallel set of test evaluations for one such distribution (the Weibull). These results were generally consistent with those for the normal distribution evaluations, suggesting that our conclusions from the latter case have some generality, and in particular might be applicable to response time and income distributions. Hence, we conclude that the new tests developed in this paper enable testing of stochastic dominance with the sample sizes and types of distributions common in many applications. Implementations in the R statistical environment (R Development Core Team, 2009) of all of the tests examined here can be obtained from the authors.

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.

Bishop, J.A., Formby, J.P. & Thistle, P.D. (1992). Convergence of the south and non-south income distribution, 1969 -1979, *American Economic Review*, *82*, 262-272.

Chotikapanich, D. & Griffiths, W. E. (2006). Bayesian assessment of Lorenz and stochastic dominance in income distributions, *Research Paper 960, Department of Economics*, The University of Melbourne.

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

Doob, J. L. (1949). Heuristic Approach to the Kolmogorov-Smirnov Theorems. *The Annals of Mathematical Statistics*, *20*, 393-403.

Dzhafarov, E., Schweickert, R., & Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology, 48*, 51–64.

Eidels A., Townsend J. T. & Pomerantz J. (2008). Where Similarity Beats Redundancy: The Importance of Context, Higher Order Similarity, and Response Assignment. *Journal of Experimental Psychology: Human Perception & Performance, 34(6),* 1441-1463.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*, 2090-230.

Hall, P., & van Keilegom, I. (2005). Testing for monotone increasing hazard rate. *The Annals of Statistics*, *33*, 1109–1137.

Johnson, S. A., Blaha, L. M., Houpt, J. W. and Townsend, J. T. (2010). Systems Factorial Technology provides new insights on global-local information processing in autism spectrum disorders. *Journal of Mathematical Psychology*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795.

Kaur, A., Rao, B.L.S.P. & Singh, H. (1994). Testing for second-order stochastic dominance of two distributions, *Econometric Theory*, *10*, 849-866.

Klugkist, I., Kato, B. & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Klugkist, I., Laudy, O. & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, *10*, 477-493.

Kolmogorov, A. (1933). Sulla determinazione empirica di une legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, *4*, 83-91.

Levy, H. (1992). Stochastic dominance and expected utility, *Management Science*, *38*, 555-593.

Lewis, S. M., Raftery, A. E., 1997. Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Liu, C.C. & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, *52*, 362-375.

McFadden, D. (1989). Testing for stochastic dominance, *Studies in the Economics of Uncertainty in Honor of Josef Hadar*, T. B. Fomby & T. K. Seo (Eds.), New York. Springer-Verlag.

Myung, J., Karabatsos, G. & Iverson, G.J. (2008). A statisticians view on Bayesian evaluation of informative hypotheses, in H. Hoijtink, I. Klugkist, & P. Boelen (Eds.), *Bayesian Evaluation of Informative Hypotheses*. Springer, Berlin.

R Development Core Team (2007). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Stoline, M.R. & Ury, H.K. (1971). Tables of the studentized maximum modulus distribution and an application to multiple comparison among means, *Technometrics*, *21*, 87- 93.

Schweickert, R., Giorgini, M., & Dzhafarov, E. (2000). Selective Influence and Response Time Cumulative Distribution Functions in Serial-Parallel Task Networks. *Journal of Mathematical Psychology 44,* 504-535.

Smirnov, N. V. (1939). Sur les écarts de la courbe de distribution empirique. *Matematiceskii Sbornik N.S.*, *6*, pp. 3–26, 1939.

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

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321–359.

Townsend, J. T., & Thomas, R. (1994). Stochastic dependencies in parallel and serial models: effects on systems factorial interactions. *Journal of Mathematical Psychology*, *38*, 1–34.

Tse, Y.K. & Zhang, X. (2003). A Monte Carlo investigation of some tests for stochastic dominance, *Working Paper 7/2003, Department of Econometric and Business Statistics*, Monash University, Australia.

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323.

Wetzels, R., Grasman, R.P.P.P., & Wagenmakers, E.-J. (in press). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis*.

Acknowledgements

Thanks to Prof. Murray Aitkin for advice on the Bayesian test, to Trish Van Zandt for pointing out the ideas underpinning the MD test, and the University of Newcastle Academic Research Computing Support Unit for help in running the simulation studies on their grid computing infrastructure, and to an anonymous reviewer for many useful suggestions. This study was supported by a Keats Endowment grant to Ami Eidels and a Vidi grant from the Dutch Organization for Scientific Research to EJ Wagenmakers.

Appendix A

Anderson Test Statistic

Anderson's (1996) method divides the range of Y and Z into i=1…K+1 mutually exclusive regions. Let p_{ij} be the probability of an observation in the *i*'th region for population *j*=Y,Z, and denote $p_j=(p_{j,1}…p_{j,K+1})'$. The two test hypotheses are: H_0: $I_f(p_Y - p_Z)=0$ and H_0: $I_f(p_Y - p_Z)>0$, where I_f is a K×K+1 matrix with unit entries except for zeros above the main diagonal. If H_0 is rejected we conclude that $Y >_s Z$. Suppose we have N_Y observations from Y and N_Z observations from Z with associated frequency vectors $n_j=(n_{1,j}, n_{2,j}, … n_{K+1,j})'$. Under the null $F_Y=F_Z$ and $p_j=p=(p_1,p_2…p_{K+1})$. Denote:

$$v = \frac{n_Y}{N_Y} - \frac{n_Z}{N_Z} \tag{5}$$

$$\Omega = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & . & -p_1p_{K+1} \\ -p_1p_2 & p_2(1-p_2) & . & -p_2p_{K+1} \\ . & . & . & . \\ -p_1p_{K+1} & -p_2p_{K+1} & . & p_{K+1}(1-p_{K+1}) \end{bmatrix} \tag{6}$$

Anderson showed that $n_j/N_j \sim N(p,\Omega/N_j)$ and $v \sim N(0,m\Omega)$ where $m=(N_Y+N_Z)/(N_YN_Z)$. This result holds asymptotically (i.e., $N_j \to \infty$ with $N_j p_i>5$ for *i*=1..K+1). Hence:

$$I_f v \xrightarrow[D]{} N\left(0, mI_f \Omega I_f'\right) \tag{7}$$

To estimate Ω we replace *p* by $\hat{p} = (n_Y + n_Z)/(N_Y + N_Z)$. Denoting the i'th element of $I_f v$ *as* $Iv(i)$ and the i'th diagonal element of $mI_f\hat{\Omega}I_f$ by $mI\hat{\Omega}I(i,i)$ the Anderson test statistic, A~N(0,1) is:

$$A_i = \frac{Iv(i)}{\sqrt{mI\hat{\Omega}I(i,i)}} \tag{8}$$

Davidson-Duclos Test Statistic

Consider the following sample statistics, where $(z)_+=\max(z,0)$.

$$\hat{D}_Y(x)=\frac{1}{N}\sum_{i=1}^{N}(x-y_i)_+ \tag{9}$$

$$\hat{D}_Z(x)=\frac{1}{N}\sum_{i=1}^{N}(x-z_i)_+ \tag{10}$$

$$\hat{V}_Y(x)=\frac{1}{N}\left[\frac{1}{N}\left(\sum_{i=1}^{N}(x-y_i)_+^2\right)-\hat{D}_Y(x)^2\right] \tag{11}$$

$$\hat{V}_Z(x)=\frac{1}{N}\left[\frac{1}{N}\left(\sum_{i=1}^{N}(x-z_i)_+^2\right)-\hat{D}_Z(x)^2\right] \tag{12}$$

$$\hat{V}_{YZ}(x)=\frac{1}{N}\left[\frac{1}{N}\left(\sum_{i=1}^{N}(x-z_i)_+(x-z_i)_+\right)-\hat{D}_Y(x)^2\hat{D}_Z(x)^2\right] \tag{13}$$

Denoting $\hat{V}(x)=\hat{V}_Y(x)+\hat{V}_Z(x)-2\hat{V}_{YZ}(x)$ we obtain the test statistic:

$$DD(x)=\frac{\hat{D}_Y(x)-\hat{D}_Z(x)}{\sqrt{\hat{V}(x)}} \tag{14}$$

In the case where observations from Y and Z are independent the V_{YZ} estimate can be assumed zero. Davidson and Duclos (2000) showed that under H_0: $D_Y(x)=D_Z(z)$ that $DD(x)$ is asymptotically distributed N(0,1).

Appendix B

Figure 6 shows the 9 positively skewed cases which we evaluated using Weibull distributions. Each case used a referent distribution that was a Weibull distribution with shape=2, variance parameter σ=1, and an offset of μ=0.25, that is, with support on the range [0.25,∞). The nine comparison distributions were generated using all combinations of σ=(3/4,1,4/3) – the rows in Figure 6 – and μ=(0,0.25,0.5) – the columns. Figure 7 shows the results of evaluations of the three existing tests. Figure 8 shows the results of the evaluations of the three-choice tests. Figure 9 shows the results of evaluations of the new four-choice tests.

Footnotes

[1] Some authors define dominance the other way, where $F_Y<F_Z$, but this is immaterial.

[2]Equal numbers of samples were taken from the prior and posterior, with the number of samples, *S*, chosen to obtain a numerical accuracy for the posterior model probability estimate corresponding to the hypothesis relative to the encompassing hypothesis (i.e., $BF_h/(1+BF_h) = \Pi_h/(\Pi_h +\pi_h)$), to be less than 0.005. Numerical error was determined by first calculating the 95% credible intervals for the proportion (p^*) estimates corresponding to Π_h and π_h assuming a uniform prior (i.e., for a Beta(Sp^*+1, $S(1-p^*)+1$) distribution). The end points of these interval estimates, $[\Pi_h^L,\Pi_h^U]$ and $[\pi_h^L,\pi_h^U]$, were then combined to calculate a worst-case estimate of numerical error: $[\Pi_h^L /(\pi_h^U+\Pi_h^L), \Pi_h^U /(\pi_h^L + \Pi_h^U)]$.

Figures

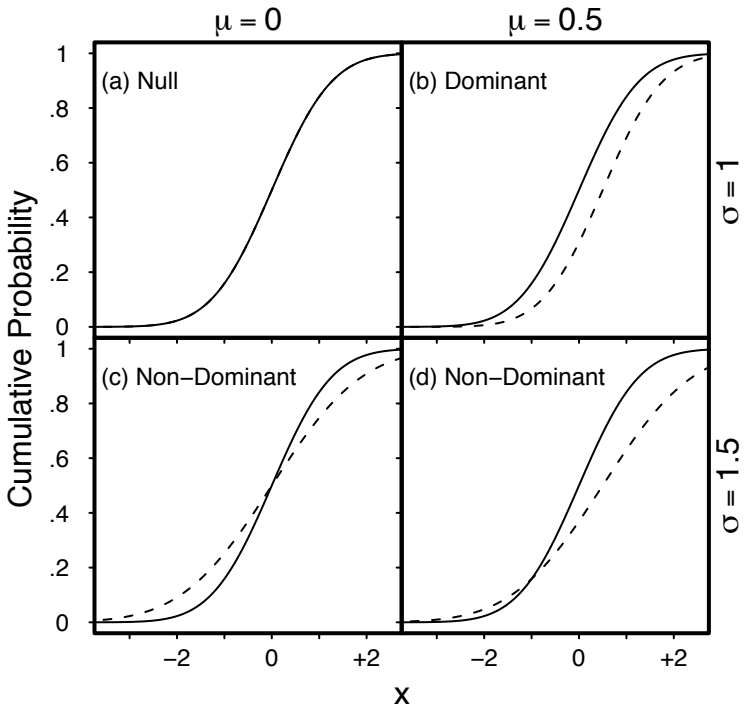

Figure 1. Pairs of normal distribution functions, $N(\mu,\sigma^2)$, which were examined in the simulation studies. In all cases the solid line shows a standard normal distribution, N(0,1). The dashed lines show distributions (a) N(0,1), (b) N(0.5,1), (c) $N(0,1.5^2)$ and (d) $N(0.5,1.5^2)$.

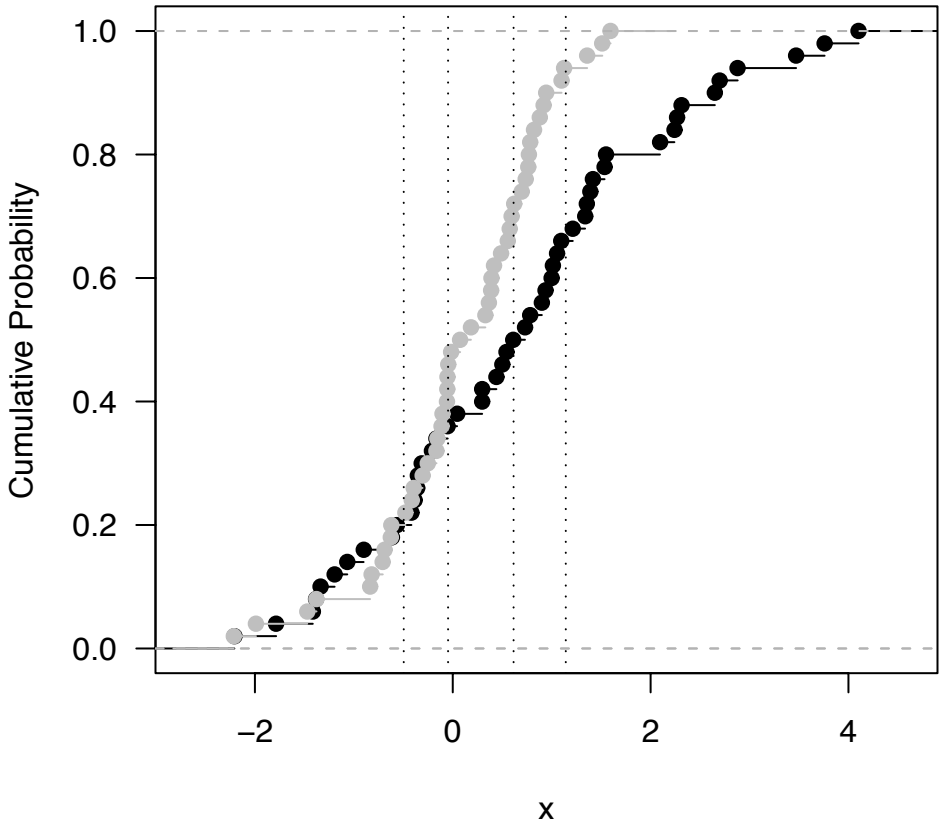

Figure 2. Empirical CDF plots of 50 samples from N(0,1) (grey points and horizontal lines) and N(0.5,1.5^2) (black points and horizontal lines). The dotted vertical lines indicate the 20^{th}, 40^{th}, 60^{th} and 80^{th} percentiles of the combined sample.

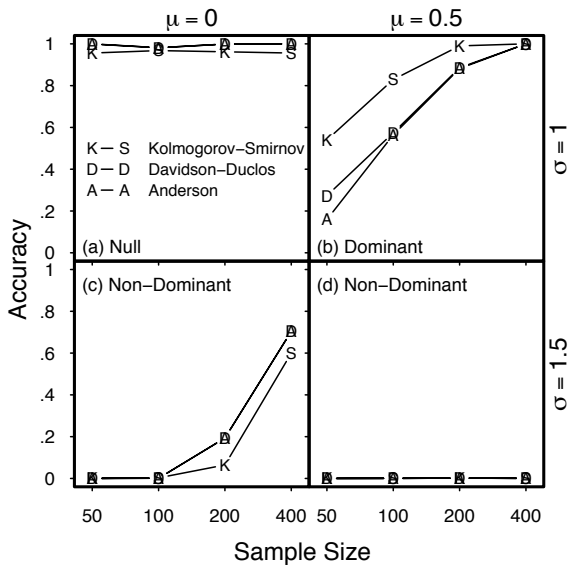

Figure 3. Simulation results for existing tests.

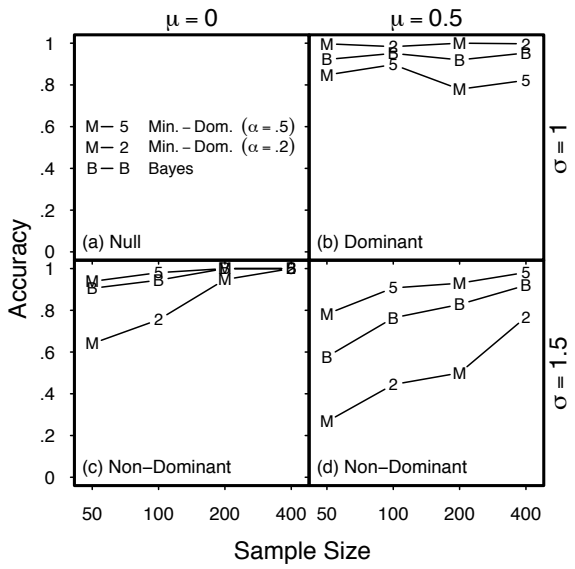

Figure 4. Simulation results for three-choice tests. Note that results are shown for the minimally dominant bootstrap test using α=.2 (“2”) and α=.5 (“5”).

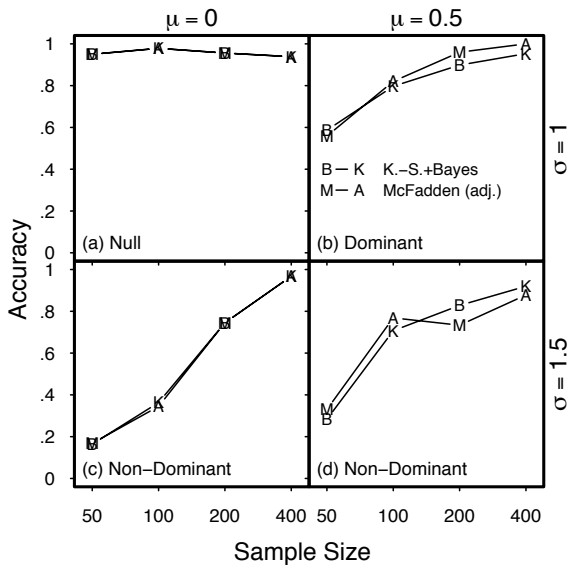

Figure 5. Simulation results for the new four-choice tests.

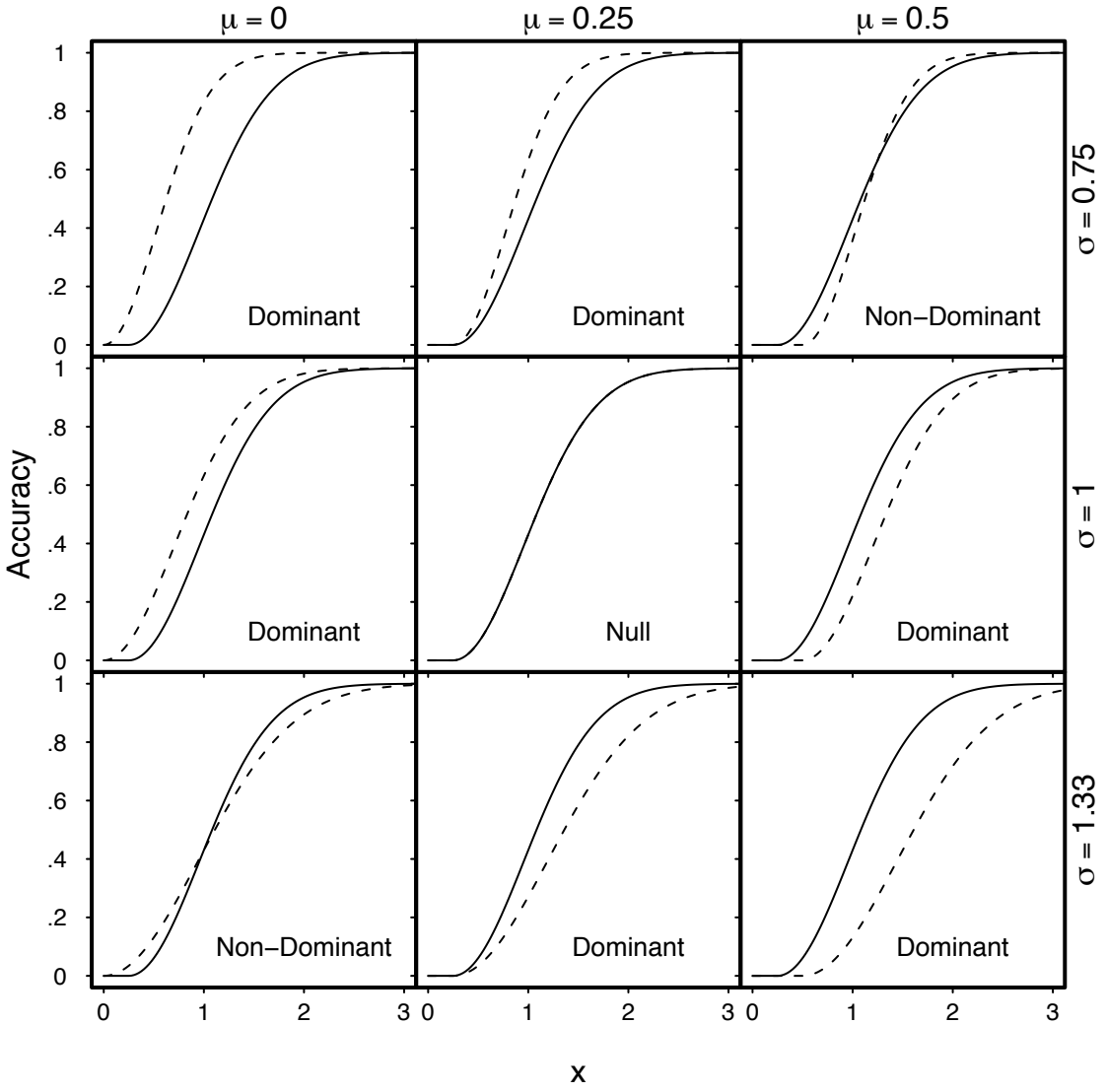

Figure 6. Weibull distribution functions examined in the simulation studies.

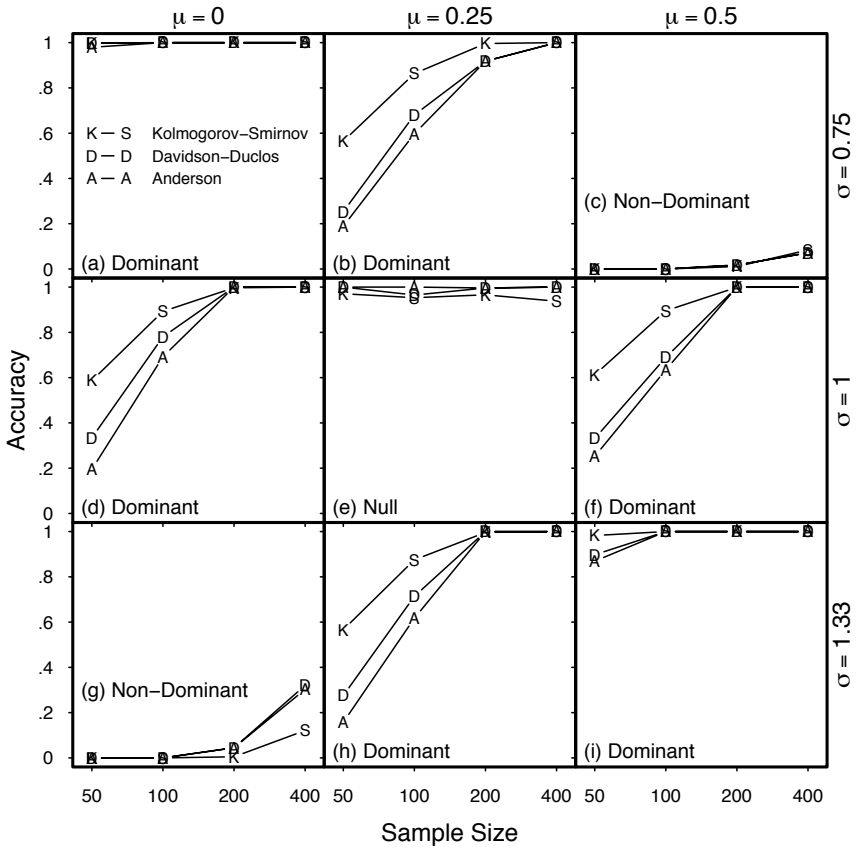

Figure 7. Simulation results for exiting tests.

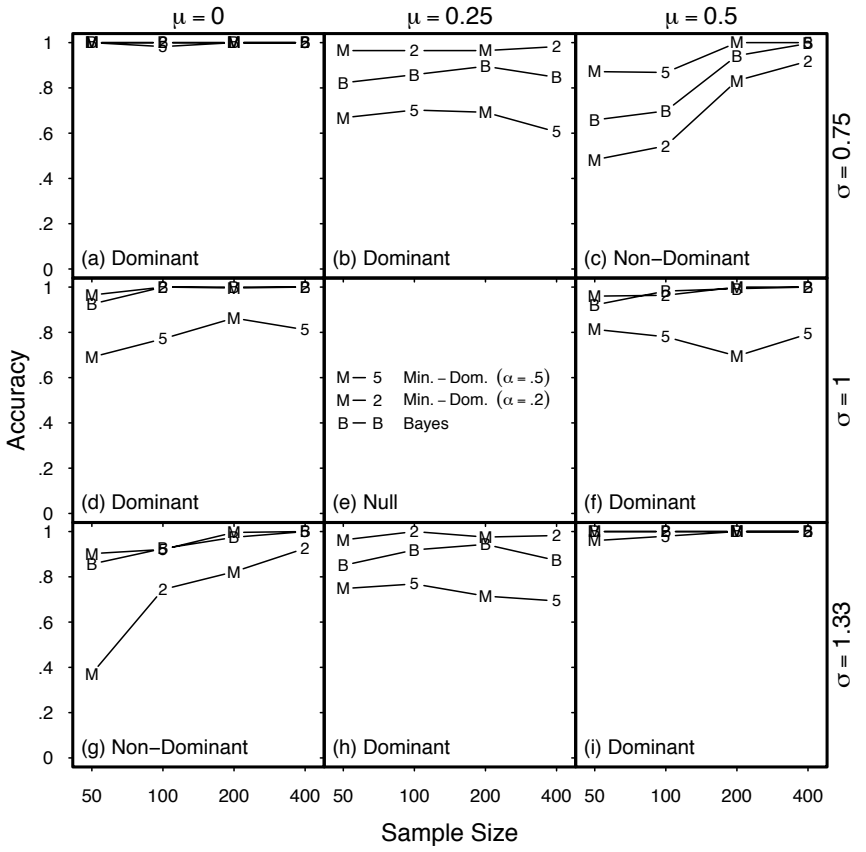

Figure 8. Simulation results for three-choice tests. Note that results are shown for the minimally dominant bootstrap test using α=.2 (“2”) and α=.5 (“5”).

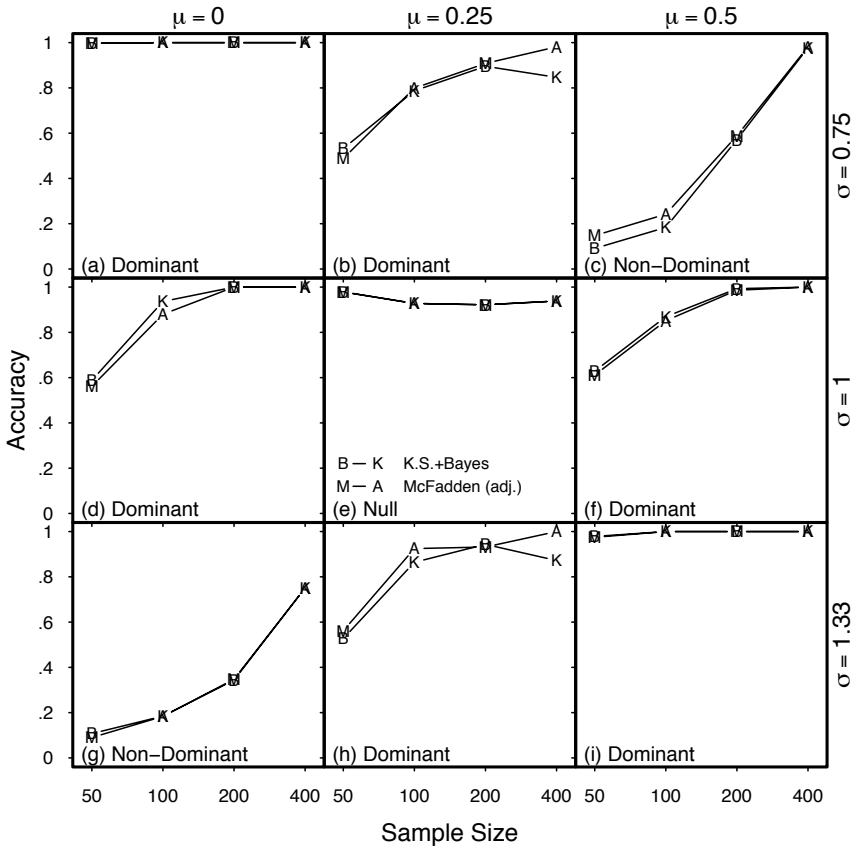

Figure 9. Simulation results for the new four-choice tests.