Analyzing Treatment Effects on Distributions with Complex Structure

Mark Bebbington

Institute of Fundamental Sciences-Statistics, Massey University, Private Bag 11222, Palmerston North, New Zealand. E-mail: M.Bebbington@massey.ac.nz

Marcel Voia

Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario K1S 5B6, Canada. E-mail: mvoia@connect.carleton.ca

Ričardas Zitikis

Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario N6A 5B7, Canada. E-mail: zitikis@stats.uwo.ca

10 October 2011

Abstract. Comparing treatment effects using, for example, the average treatment value is not particularly informative about specific regions of the treatment and control distributions. For this reason, in the present paper we consider tests which are based on entire distributions and thus provide a more informative analysis of treatments and their effectiveness, especially when the distribution of the outcome variable of interest is complicated in nature. We illustrate the performance of the tests on subpopulations of the most effective treatment from the Pennsylvania Bonus Experiment (PBE), and then use a simulation study based on mixture models fitted to the PBE subpopulations data to examine the power of the suggested tests.

Classification codes: C12, D31, D63.

Key words and phrases: Stochastic dominance, effective treatment, testing for intersection.

1. Introduction and motivating example

Policymakers often rely on the results of experiments that are designed to test the validity of implementing a new policy. The results of these experiments are quantified as treatment effects, and a proper identification of these treatment effects is therefore required in order to decide if an experiment can be generalized as a policy to the overall population Different treatment effects are considered in the literature: average treatment effects (ATE), average treatment effect on the treated (ATT), local average treatment effects (LATE), conditional average treatment effects (CATE) and quantile treatment effects (QTE).While most of the empirical literature has focused on the estimation of the treatment effects less attention was paid to test hypotheses about the presence of treatment effects. If testing was performed, then most of the testing was focused on the null hypothesis that the average effect of interest is zero but a statistical test for the overall distribution, rather than just non-parametric representations, provides more information for the evaluation of a treatment.

Consequently, this paper considers statistical tests that can identify changes in distributions during experimental periods of randomized or non-randomized experiments. Tests of equality of distributions, first-order stochastic dominance, and intersection of distributions are utilized with the aim of identifying the regions of the distributions where a treatment is, or is not, effective.

An noticeably increased interest and, subsequently, extensive literature on testing for stochastic dominance begins with the work by McFadden (1989), who proposed

and analyzed a Kolmogorov-Smirnov-type test statistic for stochastic dominance, followed by a number of authors including Davidson and Duclos (2000), Barrett and Donald (2003), and Linton *et al.* (2005), who developed extensive inferential results for stochastic dominance of any order. Horvath *et al.* (2006) elaborated on these by showing how to modify the statistics in order to test for stochastic dominance over non-compact intervals. In a related paper, Abadie (2002) assessed the distributional consequences of a treatment on some outcome variable of interest when the treatment is nonrandomized. The cumulative distribution functions of the outcome with and without the treatment were compared using tests of equality of distribution, and first-order and second-order stochastic dominance. However, the Kolmogorov-Smirnov type tests suggested by Abadie (2002) to test if the distribution of the treatment group is different than the distribution of the control group uses nonparametric bootstrap critical values that work under very strong assumptions on distributions. This paper shows that when distributions are complex in nature the nonparametric bootstrap critical values cannot be used to make a correct assessment of the treatment effect.

The importance of looking at subpopulations when evaluating treatment effects was acknowledged by Crump et al. (2009) who developed a new estimator of the average and weighted average causal effects for subpopulations, and suggested reporting more precise estimates of the average treatment effect of some subpopulations instead of focusing on an imprecisely estimated average effect for the overall population.

Having this in mind and using the example of the Pennsylvania Bonus Experiment (PBE), this paper shows the relevance of evaluating the impact of this experiment over all the distribution of the targeted population by looking at the effects on different subpopulations. In particular this paper considers evaluating the effects of the experiment on race specific subpopulations and it shows how individuals from

different race groups respond to the incentives induced by the most 'successful' treatment. The results of this analysis can be used to design better experiments and/or policies aimed at reducing the duration of the unemployment of a population.

We should note that many papers have empirically evaluated the treatment effects for the PBE, but they were using quantitative measures of the treatment effect, in other words they were focusing on the estimation side. In particular we mention the paper by Koenker and Bilias (2001), that evaluates the treatment effect on the PBE using quantile regression applied to a duration outcome variable. In this paper we explore tests for dominance and intersection of two cdf's in non-parametric and parametric contexts.

Depending on the outcome variable of interest, the tests discussed in this paper can be used to identify the differences between the outcome distributions of the individuals from the treatment and control groups or possible changes of the heterogeneity distribution of the outcome variable during the treatment period. These methods supplement those of Bebbington *et al.* (2007, 2009), where (sub-)populations were analyzed using certain extremal points of the distributions under consideration. heir results allow for the influence of exogenous (i.e., treatment) effects to be disentangled from the endogenous effects, and estimated.

The rest of the paper is organized as follows. We will next summarize the Pennsylvania Bonus Experiment (PBE) (cf. Decker *et al* 2000), and use it to motivate the design of a number of non-parametric tests, which we will formulate rigorously in Section 3 and apply to one treatment from the PBE. In Section 4 we will fit a family of parametric distributions to the PBE data and use Monte Carlo simulations to validate the test methodology, and examine the power of the tests. Mathematical details of the proofs can be found in the appendix.

2. The Pennsylvania Bonus Experiment (PBE)

A succinct description of the PBE treatment design can be found in Koenker and Bilias (2001, Sections 5.1 and 5.2). A comprehensive description can be found in Corson et al. (1992). In summary, the PBE was a re-employment experiment designed to reduce the duration of unemployment. Individuals that were randomly assigned to one of six treatment groups were given financial incentives if they found a full-time job (of at least 32 hours per week) within a qualification period and if they kept the job for a predetermined period (of at least 16 weeks). To test different treatment schemes, two levels of financial bonuses were offered: one equivalent to 3 weeks of UI benefits and the other equivalent to 6 weeks of UI benefits. Further, two qualification periods were considered: one of 6 weeks duration and the other of 12 weeks. Additionally, a workshop was offered to help individuals in job search. The workshop did not require compliance. The individuals from a comparison group, which were randomly assigned from the same local UI offices as the 6 treatment groups, were subject to the existing rules of UI benefits.

Of the six treatments only one had a significant impact in reducing the number of claimants who exhausted their UI benefits. The present paper is focused on this treatment group (treatment four of the PBE), which combined the highest bonus ($997 on average) with the longest qualification period (12weeks) and a workshop. The overall impact of this treatment was a reduction in unemployment by 0.8 weeks and a reduction in UI benefits by $130 (see Corson et al., 1992).

We ignore, for the purposes of illustrating our methodology, the complications arising from selecting the most significant of a number of experiments. The unemployment duration distribution can be broken down into three subgroups: Black, Hispanic and White. In addition, we also analyze the whole group, denoted 'All'.

The corresponding empirical distribution functions in Figure 1 show apparently differing impacts of the treatment on the distribution of unemployment duration for different subgroups. However, the noticeable differences may of course be due to sampling variation. For this reason, in order to test whether the treatment (dotted) and control (solid) lines coincide, or one dominates the other, or they intersect, we need statistical tests.

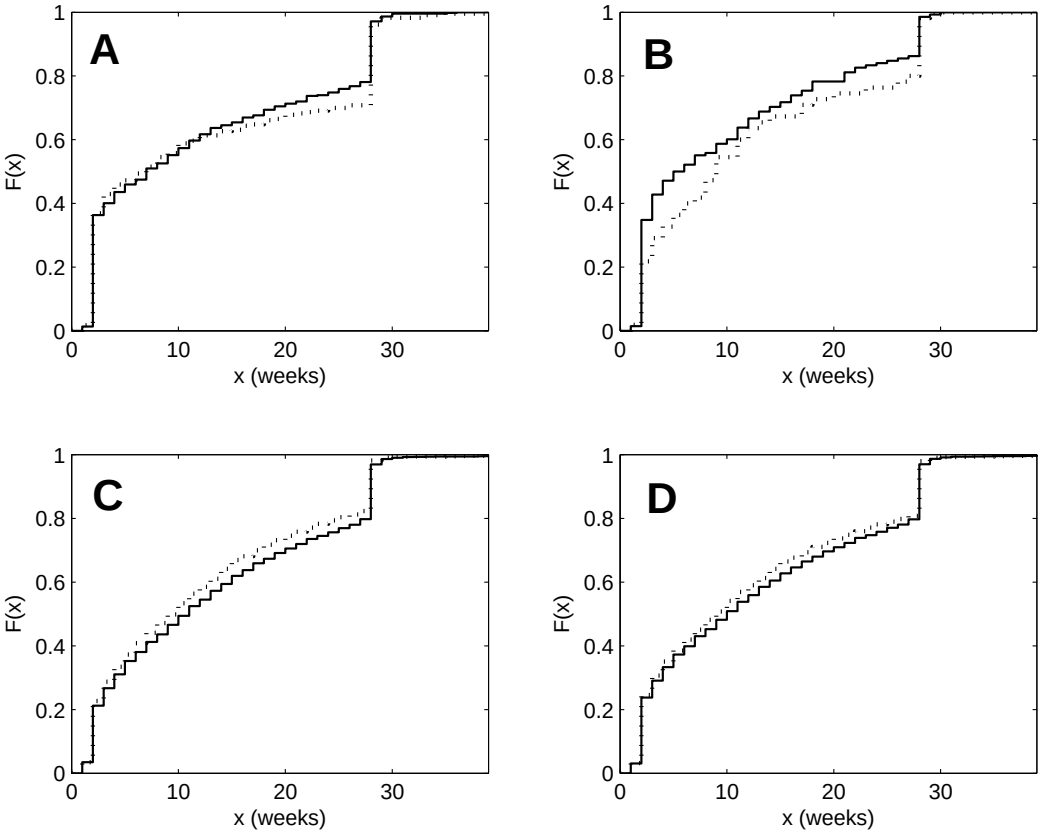

FIGURE 1. The control and fourth-treatment empirical distribution functions by race for the PBE: Black (A), Hispanic (B), White (C), and All (D). In each panel, the treatment is depicted with a dotted line and the control with a solid line.

Mathematically, let G be a random variable taking two values: $G = 0$ if a randomly selected individual is assigned to the control group and $G = 1$ if assigned to

the treatment group. (We use the upper-case G to indicate that this *random* variable assigns individuals to *groups*). There are two time periods. The first one, which we denote by $t = 0$, is the time before a certain treatment policy is introduced. The second period, which we denote by $t = 1$, is the time after the introduction of the treatment policy or, in some circumstances, the time when the effect of the policy is measured. (We use the lower-case t to denote *non*-random *time* periods.)

Hence, we have the random pair (G, t) that takes on one of the four possible values: $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. The variable of interest is $X^{(G,t)}$, which for our motivating examples is the time out-of-work measured in weeks. We are interested in the conditional distribution function $F^{(g,t)}(x) = \mathbf{P}\left[X^{(G,t)} \leq y |\ G = g\right]$ for various choices of $g, t \in \{0, 1\}$.

We assume that at the time of the random assignment (when the treatment has not yet been applied) the control and the treatment groups have the same distribution, which we write as $F^{(0,0)} = F^{(1,0)}$. For our example, given that at the time of the random assignment the individuals have just entered unemployment spell, we use the reported earning distributions to compare the two groups at the baseline. The Kolmogorov-Smirnov two-sample test for the equality of distributions has the P-value 0.4291.

We shall next discuss the possibilities for the two post-treatment distributions $F^{(0,1)}$ and $F^{(1,1)}$. First, we may start by simply testing the null hypothesis H_0 : $F^{(0,1)} = F^{(1,1)}$ (no effect) against the alternative H_1 : $F^{(0,1)} \neq F^{(1,1)}$. Indeed, at the very outset we are naturally interested in whether the distributions of the control and treatment groups differ after the introduction of a new policy.

Looking at Figure 1B, however, we get the impression that the treatment for the Hispanic subpopulation may not have worked as intended which, due to sampling variability, needs to be tested statistically, especially when sample sizes are relatively small (see Table 1).

TABLE 1. Sample sizes for the control and fourth-treatment groups at $t = 1$ for the PBE.

Race	Control	Treatment
Black	457	233
Hispanic	138	55
White	2979	1571
Others	21	18
Total	3595	1877

Table 1 also explains why we do not analyze 'other races', whose control and treatment groups have only 21 and 18 observations, respectively. Hence, the question might be: Is the treatment for the Hispanic subpopulation ineffective (an optimistic statement given the graph) or worse than the control (as suggested by Figure 1B). Hence, we are interested in testing the null hypothesis $H_0 : F^{(0,1)} = F^{(1,1)}$ against the alternative $H_1 : F^{(0,1)} \geq F^{(1,1)}$ with $F^{(0,1)}(x) > F^{(1,1)}(x)$ for at least one x. Rejecting the null hypothesis would imply that the treatment has had a negative effect on the whole treatment group.

Similarly, Figure 1C and Figure 1D suggest testing the null hypothesis $H_0 : F^{(0,1)} = F^{(1,1)}$ against the alternative $H_1 : F^{(0,1)} \leq F^{(1,1)}$ with $F^{(0,1)}(x) < F^{(1,1)}(x)$ for at least one x. Thus, rejecting the null hypothesis would imply that the treatment has had a positive effect on the whole treatment group.

Finally, Figure 1A suggests that there might be an intersection between the control and treatment distribution functions, thus implying that the treatment has been both positive and negative for the treatment group. We formulate this as the null hypothesis $H_0 : F^{(0,1)} = F^{(1,1)}$ against the alternative $H_1 : F^{(0,1)} \bowtie F^{(1,1)}$, which means that there are points x and y such that $F^{(0,1)}(x) < F^{(1,1)}(x)$ and $F^{(0,1)}(x) >$

$F^{(1,1)}(x)$. We note, however, that the treatment line in Figure 1A is just barely above the control line on the left-hand side of the graph and more noticeably below it on the right-hand side. This might indicate dominance, which in turn suggests testing the null hypothesis $H_0 : F^{(0,1)} \geq F^{(1,1)}$ with $F^{(0,1)}(x) > F^{(1,1)}(x)$ for at least one x against the alternative hypothesis $H_1 : F^{(0,1)} \bowtie F^{(1,1)}$. To test these last set of hypotheses, we can use the test for the previous hypotheses (equality vs intersection), although this would be conservative. However, this seems to be the best we can do without specifying how much $F^{(0,1)}$ is above $F^{(1,1)}$.

Another interesting (and similar) problem would be to find out if the behavior of those in the treatment group has changed after the introduction of the new policy, compared to their behavior before the introduction of the policy. For this, the hypotheses are analogous to those formulated above, but with $F^{(1,0)}$ replacing $F^{(0,1)}$.

3. Statistical tests

The hypotheses motivated in the previous section concern relationships between two functions, which can be viewed as infinite dimensional parameters. To simplify the task, we condense the hypothesis into (sometimes simplified) hypotheses concerning the following one-dimensional parameters

$$\delta^- = \sup_x (F^{(0,1)}(x) - F^{(1,1)}(x)), \quad \delta^+ = \sup_x (F^{(1,1)}(x) - F^{(0,1)}(x)),$$

$$\kappa = \max(\delta^-, \delta^+), \quad \text{and} \quad \tau = \min(\delta^-, \delta^+).$$

(Note that $\kappa = \sup_x |F^{(0,1)}(x) - F^{(1,1)}(x)|$.) We have the following relationships:

$$\left.\begin{array}{l} H_0 : F^{(0,1)} = F^{(1,1)} \\ H_1 : F^{(0,1)} \neq F^{(1,1)} \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} H_0 : \kappa = 0 \\ H_1 : \kappa > 0 \end{array}\right. \tag{1}$$

$$\left.\begin{array}{l} H_0 : F^{(0,1)} \leq F^{(1,1)} \text{ or } F^{(0,1)} \geq F^{(1,1)} \\ H_1 : F^{(0,1)} \bowtie F^{(1,1)} \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} H_0 : \tau = 0 \\ H_1 : \tau > 0 \end{array}\right. \tag{2}$$

$$\left.\begin{array}{l} H_0 : F^{(0,1)} \leq F^{(1,1)} \\ H_1 : F^{(0,1)}(x) > F^{(1,1)}(x) \text{ for some } x \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} H_0 : \delta^- = 0 \\ H_1 : \delta^- > 0 \end{array}\right. \tag{3}$$

$$\left.\begin{array}{l} H_0 : F^{(0,1)} \geq F^{(1,1)} \\ H_1 : F^{(0,1)}(x) < F^{(1,1)}(x) \text{ for some } x \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} H_0 : \delta^+ = 0 \\ H_1 : \delta^+ > 0 \end{array}\right. \tag{4}$$

Our next task is to construct empirical estimators for the parameters κ, τ, δ^-, and δ^+. Then, based on the estimators, we shall construct statistics for testing hypotheses (1)–(4). Our data consist of two sets:

$$X_1^{(0,1)}, \ldots, X_n^{(0,1)} \sim F^{(0,1)}, \tag{5}$$

$$X_1^{(1,1)}, \ldots, X_m^{(1,1)} \sim F^{(1,1)}. \tag{6}$$

We assume that the random variables are independent, within and between the two samples, but that they are identically distributed only within each set (5) and (6). We will denote the empirical cdf's corresponding to (5) and (6) by $\widehat{F}^{(0,1)}$ and $\widehat{F}^{(1,1)}$, respectively. Furthermore, we assume that the sample sizes n and m are comparable, which from the mathematical point of view means that there exists a constant $0 < \eta < 1$ such that $m/(n+m) \to \eta$ when $n, m \to \infty$. The estimators of the four parameters κ, τ, δ^-, and δ^+ are constructed by replacing the population cdf's $F^{(0,1)}$ and $F^{(1,1)}$ by the corresponding empirical cdf's $\widehat{F}^{(0,1)}$ and $\widehat{F}^{(1,1)}$, thus producing the statistics $\widehat{\kappa}$, $\widehat{\tau}$, $\widehat{\delta^-}$, and $\widehat{\delta^+}$, respectively. By the classical Glivenko-Cantelli theorem, $\widehat{\kappa}$, $\widehat{\tau}$, $\widehat{\delta^-}$, and $\widehat{\delta^+}$ are strongly consistent estimators of κ, τ, δ^-,

and δ^+, respectively. The corresponding test statistics are

$$\widehat{K} = \sqrt{\frac{nm}{n+m}}\,\widehat{\kappa}, \quad \widehat{T} = \sqrt{\frac{nm}{n+m}}\,\widehat{\tau}, \quad \widehat{D}^- = \sqrt{\frac{nm}{n+m}}\,\widehat{\delta}^-, \quad \widehat{D}^+ = \sqrt{\frac{nm}{n+m}}\,\widehat{\delta}^+.$$

We next discuss calculating critical values for these four test statistics.

3.1. **Testing hypotheses (1).** Under the null hypothesis (i.e., $\kappa = 0$), the statistic $\widehat{K}$ has a non-degenerate limit when $n, m \to \infty$ (see Theorem 1). Under the alternative hypothesis (i.e., $\kappa > 0$), the statistic $\widehat{K}$ tends in probability to $+\infty$, thus implying the asymptotic power 1 of the test (see Theorem 1 in the Appendix). The limit under the null hypothesis $\kappa = 0$ is a non-degenerate random variable, which determines the asymptotic critical value of the test statistic $\widehat{K}$. The random variable is somewhat complex (see Theorem 1), but in the special case when the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ are continuous (and coincide, since $\kappa = 0$), the random variable is the supremum $\sup_t |\mathcal{B}(t)|$ of a standard Brownian bridge $\mathcal{B}$ on $[0, 1]$. The distribution of the supremum is given by the classical Kolmogorov distribution function

$$\mathbf{P}\big[\sup_t |\mathcal{B}(t)| \le x\big] = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2x^2}. \tag{7}$$

Hence, with k_α such that $\mathbf{P}[\sup_t |\mathcal{B}(t)| > k_\alpha] = \alpha$, we have the rejection region

$$\{x \in \mathbf{R} : \, x > k_\alpha\} \tag{8}$$

for the statement $\kappa = 0$ in favor of $\kappa > 0$.

It is useful to contrast this asymptotic rejection region with a bootstrap based rejection region, for which we need to define a bootstrap based critical value k^*_α. For this, we utilize an idea by Horváth *et al.* (2006). Namely, we first combine data sets (5) and (6) into one, thus having $n + m$ observations, whose empirical distribution function we denote by $\widehat{F}^{(\bullet,1)}$. We next draw two sets of i.i.d. observations from the combined set: the first set contains n observations, which we denote by $X_1^{*(0,1)}, \ldots, X_n^{*(0,1)} \sim \widehat{F}^{(\bullet,1)}$, and the second set contains m observations, which

we denote by $X_1^{*(1,1)}, \ldots, X_m^{*(1,1)} \sim \widehat{F}^{(\bullet,1)}$. Using these two sets of observations, we calculate $\widehat{\kappa}$, which we denote by $\widehat{\kappa}^*$, and then calculate

$$\widehat{K}^* = \sqrt{\frac{nm}{n+m}}\, \widehat{\kappa}^*.$$

We repeat the procedure b times and thus obtain b values of $\widehat{K}^*$. Finally, we calculate the $(1-\alpha)$ percentile of the $\widehat{K}^*$ values and denote it by k_α^*. In summary, we have obtained the bootstrap-based rejection region

$$\{x \in \mathbf{R} : \, x > k_\alpha^*\} \tag{9}$$

for the statement $\kappa = 0$ in favor of $\kappa > 0$. Numerical illustration of these (and following) rejection regions will be provided, and discussed, below.

3.2. **Testing hypotheses (2).** Under the null hypothesis (i.e., $\tau = 0$), the statistic $\widehat{T}$ is asymptotically bounded in probability (see Theorem 2 in the Appendix), whereas under the alternative (i.e., $\tau > 0$), the statistic tends in probability to $+\infty$, thus implying the asymptotic power 1 of the test. Next, we need to find a critical value of the test. Since the null hypothesis is composite, we first search for a sub-case of the null hypothesis that would give the most conservative (that is, largest) critical value. The structure of the parameter τ suggests that $\kappa = 0$, which is a 'subset' of $\tau = 0$, is located on the 'boundary' between the null and alternative hypotheses, and this in turn suggests the rejection region

$$\{x \in \mathbf{R} : \, x > k_\alpha\}, \tag{10}$$

for the statement $\tau = 0$ in favor of $\tau > 0$, where k_α is the critical value defined in the previous subsection. Note, however, that the just specified rejection region for $\min(\delta^-, \delta^+) = 0$ is the same as that for $\max(\delta^-, \delta^+) = 0$. In other words the test based on rejection region (10) is naturally expected to be conservative.

We may argue that the rejection region based on a critical value x_α defined by the equation $\mathbf{P}[\min\{\sup_t \mathcal{B}(t), \sup_t(-\mathcal{B}(t))\} > x_\alpha] = \alpha$ would be just what we need. However, we do not know the survival function $\mathbf{P}[\min\{\sup_t \mathcal{B}(t), \sup_t(-\mathcal{B}(t))\} > x]$, unlike $\mathbf{P}[\max\{\sup_t \mathcal{B}(t), \sup_t(-\mathcal{B}(t))\} > x]$ (see equation (7)).

As a way out of the impasse, we employ the earlier used bootstrap idea, which gives the rejection region

$$\{x \in \mathbf{R} : \, x > t^*_\alpha\} \tag{11}$$

for the statement $\tau = 0$ in favor of $\tau > 0$, where the critical value t^*_α is calculated in the same way as in the previous two subsections but now using

$$\widehat{T}^* = \sqrt{\frac{nm}{n+m}}\, \widehat{\tau}^*$$

instead of $\widehat{T}^*$. Naturally, the rejection region is expected to be conservative unless we have $\kappa = 0$, which is a 'subset' of $\tau = 0$ and is on the 'boundary' between the null and alternative hypotheses.

3.3. **Testing hypotheses (3) and (4).** Under the null hypothesis (i.e., $\delta^\pm = 0$), the statistic $\widehat{D}^\pm$ is asymptotically bounded in probability (see Theorem 3 in the Appendix), and when $\delta^\pm > 0$, then the statistic tends in probability to $+\infty$, thus implying that the asymptotic power of the test is 1. When the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ are continuous and coincide, which is a special case of the null hypothesis $\delta^\pm = 0$, then the asymptotic distribution of $\widehat{D}^\pm$ coincides with that of $\sup_t \mathcal{B}(t)$, whose cdf is

$$\mathbf{P}\big[\sup_t \mathcal{B}(t) \le x\big] = 1 - e^{-2x^2}.$$

This formula gives a convenient way to calculate the asymptotic critical value d_α of the test, which satisfies the equation $\mathbf{P}[\sup_t \mathcal{B}(t) > d_\alpha] = \alpha$. We have obtained the rejection region

$$\{x \in \mathbf{R} : \, x > d_\alpha\} \tag{12}$$

TABLE 2. P-values for the PBE data (bootstrap estimates in parentheses)

Race	Test (1)	Test (2)	Test (3)	Test (4)
All	0.064 (0.035)	1.000 (0.918)	0.997 (0.957)	0.032 (0.018)
Black	0.382 (0.233)	0.999 (0.123)	0.192 (0.121)	0.774 (0.527)
Hispanic	0.390 (0.236)	1.000 (0.828)	0.196 (0.119)	0.995 (0.914)
White	0.012 (0.007)	1.000 (0.948)	0.999 (0.972)	0.006 (0.004)

for the statement $\delta^{\pm} = 0$ in favor of $\delta^{\pm} > 0$. Alternatively, we may use the bootstrap based rejection region

$$\{x \in \mathbf{R} : \, x > d^*_\alpha\}, \tag{13}$$

where d^*_α is the critical value calculated in the same way as k^*_α but now using

$$\widehat{D}^* = \sqrt{\frac{nm}{n+m}}\, \widehat{\delta}^*$$

instead of $\widehat{K}^*$. Naturally, the rejection region is expected to be conservative unless we have $\kappa = 0$, which is a 'subset' of $\delta^{\pm} = 0$ and is on the 'boundary' between the null and alternative hypotheses.

3.4. **Example: The PBE data.** Performing the tests suggested above on the PBE data, we obtain the P-values in Table 2. We see, considering the fourth treatment in isolation, that there is evidence that the two distribution functions $F^{(0,1)}$ and $F^{(1,1)}$ are not identical for the White data, and that $F^{(0,1)}$ does not stochastically dominate $F^{(1,1)}$. This result extends to the case when the data is aggregated for all races. The more interesting question, of whether $F^{(1,1)}$ stochastically dominates $F^{(0,1)}$, i.e., the fourth treatment had an effect of reducing unemployment duration for the entire support of the outcome variable, is consistent with the high P-value for Test (3), but not statistically proven. We see that Test (2) is clearly lacking in power due to the weak bound in Theorem 2, as the 'Black' data in Figure 1A clearly has intersecting empirical distributions, but the P-value is anemic, although

the bootstrap estimate alleviates some of the difficulty. Our findings suggest that some of these tests should be used very carefully. To obtain better performances, a more detailed (parametric) analysis of the data is required. This we explore in the next section.

Meanwhile, we are interested to see how much the bootstrap based rejection regions outperform the corresponding asymptotic rejection regions. Figure 2 shows the P-values for each statistic calculated (a) from the asymptotic theory, and (b) by bootstrapping the 'All' data, assuming two equal cdf's $F^{(0,1)}$ and $F^{(1,1)}$. We see

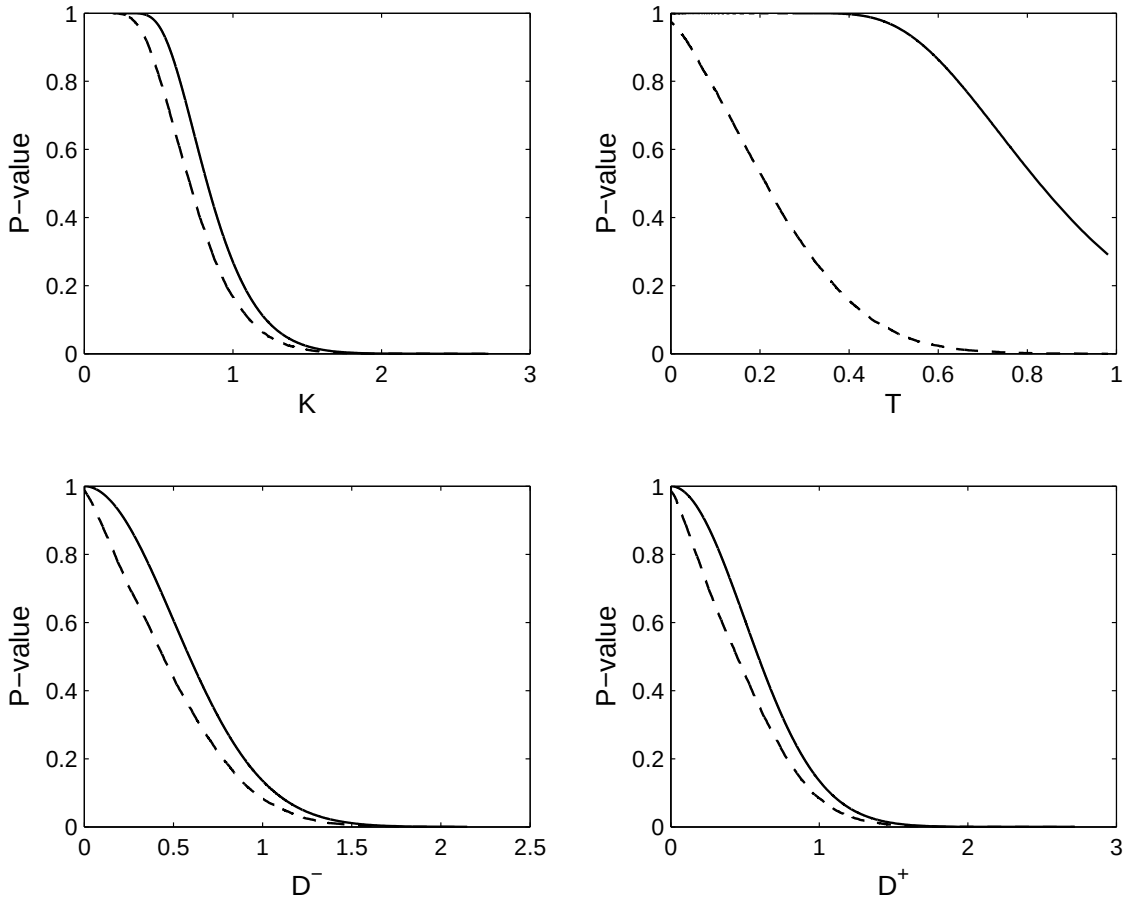

FIGURE 2. P-values estimated from asymptotic distributions (dashed lines) and bootstrapped from 'All' data (solid lines).

that at $\alpha = 0.05$, the two rejection regions for tests (1), (3) and (4) are reasonably

close, as expected. For test (2), the two rejection regions differ considerably, again as expected.

4. Parametric models of unemployment duration in the PBE data

The purpose of this section is to find a good parametric model for our data, which can be used to study the properties of our proposed test statistics. The subsamples of the PBE that are of interest for our analysis are for Whites and Blacks. The former is the most representative subsample from the PBE and is driving the overall positive results of the experiment, while the latter is both sufficiently large to be analyzed statistically, and the only subsample that show mixed results for the treatment effect.

Figure 1 shows a complicated structure for the distribution of the outcome variables for all races and groups. The data in fact is the number of complete weeks unemployed, and hence the actual unemployment duration for a datum x is somewhere between x and $x+1$ weeks. Hence to fit our parametric model we want a continuous distribution, although the data is discretized. There is an obvious discontinuity in the data at 27 weeks. This is due the fact that unemployment insurance is available for a maximum of 26 weeks, and there is a one-week stand-down period. Hence a datum of '27' indicates that the person became employed in the week immediately following the last possible unemployment payment.

Hence we will fit a piecewise model of the form

$$F(t) = \begin{cases} F_1(t), & t \leq 27, \\ F_1(27) + p(t-27), & 27 < t \leq 28, \\ F_1(27) + p + (1 - F_1(27) - p)F_2(t-28), & 28 < t, \end{cases} \tag{14}$$

where p is a parameter, and F_1 and F_2 are appropriately chosen distribution functions, to be discussed next. If we think of the hazard rate $h(t) = F'(t)/[1-F(t)]$

TABLE 3. Estimated parameters and BICs, $G = 0$ 'White' data

Model	Nr. of Param.	BIC	Estimated Parameters
$a_1 = a_2 = 1$	3	1880.2	$p = 0.152, b_1 = 0.064, b_2 = 0.245$
$a_1 = 1$	4	1877.3	$p = 0.152, b_1 = 0.064, a_2 = 0.653, b_2 = 0.349$
$a_2 = 1$	4	1875.0	$p = 0.167, a_1 = 0.876, b_1 = 0.065, b_2 = 0.280$
$a_1 = a_2 = a$	4	1872.8	$p = 0.169, a = 0.861, b_1 = 0.065, b_2 = 0.280$
Free	5	1872.0	$p = 0.167, a_1 = 0.876, b_1 = 0.065, a_2 = 0.653, b_2 = 0.349$

as the 'employment rate', (14) allows for different employment rates before and after the 27 week threshold. We also suppose a uniform distribution for the date of employment obtained during the 27th week.

In order to allow for increasing or decreasing employment rate, we will use the Weibull distribution $F_i(t) = 1 - \exp\left[-(b_i t)^{a_i}\right]$, and we will consider the alternatives $a_1 = a_2 = 1$ (the exponential distribution), $a_1 = 1$, $a_2 = 1$, and $a_1 = a_2$. The parameters are estimated by maximizing the log-likelihood

$$\begin{aligned} LL = \sum_i \big\{ & I_{(0,26]}(t_i) \left[\log a_1 + a_1 \log b_1 + (a_1 - 1) \log t_i - (b_1 t_i)^{a_1}\right] + I_{(26,27]}(t_i) \log p \\ & + I_{(27,\infty]}(t_i) \left[\log a_2 + a_2 \log b_2 + (a_2 - 1) \log(t_i - 27) - (b_2(t_i - 27))^{a_2}\right. \\ & \left. + \log(\exp(-(26 b_1)^{a_1}) - p)\right] \big\}, \end{aligned}$$

where the indicator function $I_A(t)$ is equal to 1 if $t \in A$ and 0 when $t \notin A$. Note that in order to make the discrete data and continuous distribution consistent, we will augment the data by adding a Uniform$(0, 1)$ random variable to each unemployment duration before fitting. The same augmented data is used for all distributions. The results are given in Tables 3 and 4 for the 'White' data in Figure 1C.

We see that the full model is narrowly preferred by BIC from the $a_1 = a_2 = a$ model. That is, the preferred model is a piecewise Weibull-Uniform-Weibull, where the two Weibull's have different shape parameters. These are all less than one,

TABLE 4. Estimated parameters and BICs, $G = 1$ 'White' data

Model	Nr. of Param.	BIC	Estimated Parameters
$a_1 = a_2 = 1$	3	979.4	$p = 0.132, b_1 = 0.071, b_2 = 0.194$
$a_1 = 1$	4	978.9	$p = 0.132, b_1 = 0.071, a_2 = 0.642, b_2 = 0.275$
$a_2 = 1$	4	977.0	$p = 0.148, a_1 = 0.881, b_1 = 0.072, b_2 = 0.194$
$a_1 = a_2 = a$	4	976.4	$p = 0.149, a = 0.872, b_1 = 0.072, b_2 = 0.218$
Free	5	976.6	$p = 0.1480, a_1 = 0.881, b_1 = 0.072, a_2 = 0.642, b_2 = 0.275$

implying that the rate of job-acceptance is declining in the Weibull segments. Most interestingly, the Weibull shape parameters appear to be the same for both the control ($G = 0$) and treatment ($G = 1$) groups. The differences in the b_i parameters seems to imply that the bonus group $G = 1$ are finding employment at a higher rate before the Week 27 cutoff, but at a lower rate afterwards. The fitted distributions are shown in Figure 3. Note that the step function shown is after adding one week to the data, in order to obtain the correct distribution function for the unemployment duration.

We are now in a position to verify the type I error probability of the tests, and to examine their power. The first objective also requires that we fit the model to the combined data. This gives parameter estimates $p = 0.1604, a_1 = 0.8771, b_1 = 0.0671, a_2 = 0.6486, b_2 = 0.3291$. Using these estimates, we will simulate a sample of size $1.5N$, dividing it randomly into a simulated $G = 0$ of size N and a $G = 1$ of size $N/2$. The fractional part of each datum is truncated to conform with the precision in the original data. We then perform the four tests outlined in Section 3. The whole procedure is then repeated 1000 times, and the proportion of P-values less than 0.05 recorded, resulting in the estimated type I error probabilities given in Table 5. We see that the tests appear to be conservative on this 'discretized' data. On continuous data simulated in the same manner, we obtain the estimated

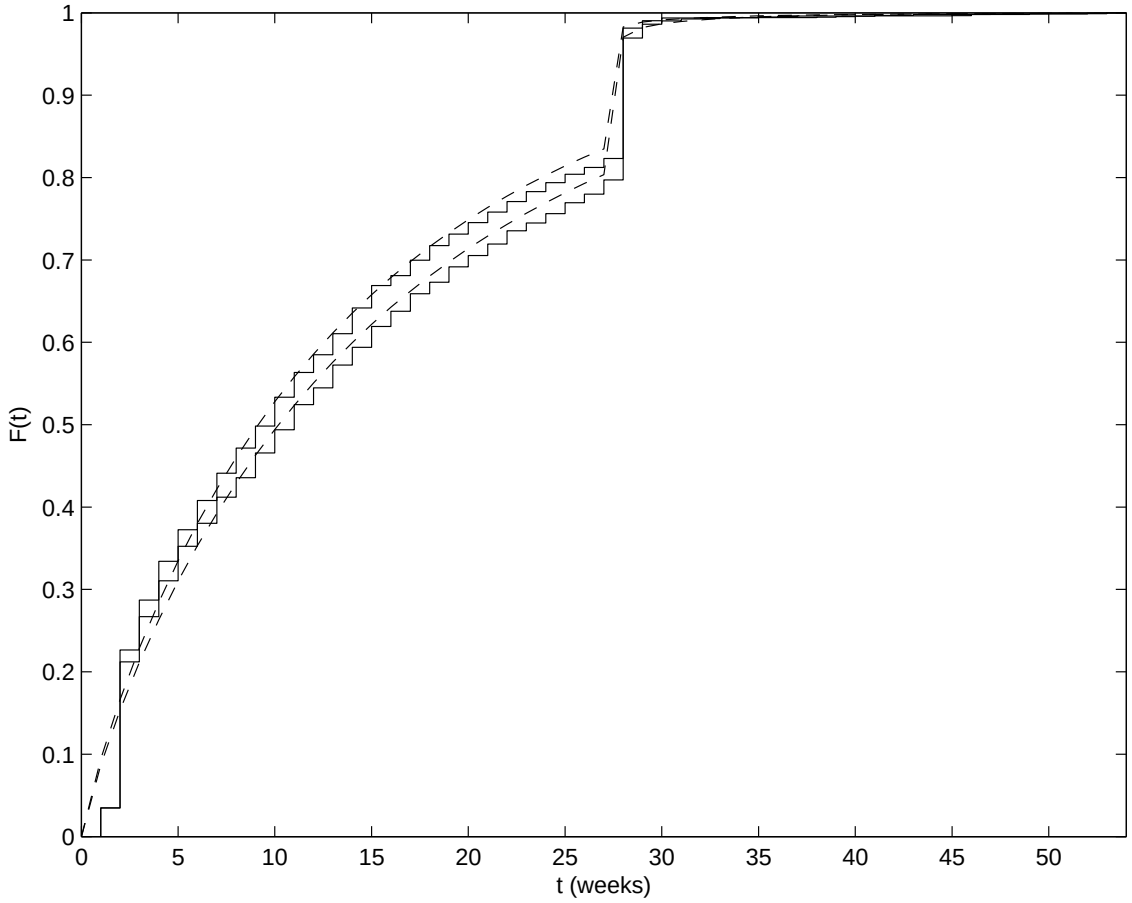

FIGURE 3. Observed (solid) and fitted (dashed) unemployment distributions for 'White' data. $G = 0$ (lower curves) and $G = 1$ (upper curves).

TABLE 5. Type I error probabilities, aggregated PBE ('White') data.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.030	0	0.023	0.029
200	0.030	0	0.031	0.030
500	0.021	0	0.029	0.021
1000	0.023	0	0.025	0.028
2000	0.029	0	0.026	0.029
5000	0.031	0	0.038	0.028

TABLE 6. Type I error probabilities, continuous aggregated PBE ('White') data.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.051	0	0.045	0.044
200	0.041	0	0.039	0.055
500	0.045	0	0.033	0.050
1000	0.039	0	0.043	0.044
2000	0.044	0	0.042	0.040
5000	0.056	0	0.054	0.049

type I error probabilities given in Table 6 which, apart from Test (2), appear to be consistent with the construction. Note that the discretization affects the apparent size of the test. The zero error probability for Test (2) results from the weak bound in Theorem 2.

Using the models for the separate populations as shown in Figure 3, we can examine the power of the tests in the same way, simply simulating N values from the $G = 0$ distribution, and $N/2$ from the $G = 1$ distribution, with the results shown in Table 7. Note that the null hypotheses of Tests (2) and (3) are true in this case. The other tests have false null hypotheses, and we see that the power increases with sample size, as desired.

In order to examine the power of Test (2), we need a situation where the null hypothesis is false. We can best construct such a case by fitting the model (14) to the 'Black' data. The same method of data augmentation is used. The results of the model fitting are given in Tables 8 and 9.

We see that in this case the $a_1 = a_2 = a$ model is preferred by BIC. That is, the preferred model is a piecewise Weibull-Uniform-Weibull, where the two Weibull's have identical shape parameters. These are all less than one, implying that the rate

TABLE 7. Power of test, simulated data from $G = 0$ and $G = 1$ ('White') distributions.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.037	0	0.008	0.076
200	0.056	0	0.009	0.113
500	0.099	0	0.001	0.177
1000	0.208	0	0	0.305
2000	0.421	0	0	0.544
5000	0.807	0	0	0.885

TABLE 8. Estimated parameters and BICs, $G = 0$ 'Black' data

Model	Nr. of Param.	BIC	Estimated Parameters
$a_1 = a_2 = 1$	3	2753.2	$p = 0.133, b_1 = 0.070, b_2 = 0.562$
$a_1 = 1$	4	2758.2	$p = 0.133, b_1 = 0.070, a_2 = 0.809, b_2 = 0.634$
$a_2 = 1$	4	2704.4	$p = 0.168, a_1 = 0.736, b_1 = 0.073, b_2 = 0.562$
$a_1 = a_2 = a$	4	2073.4	$p = 0.167, a = 0.649, b_1 = 0.061, b_2 = 0.280$
Free	5	2709.4	$p = 0.168, a_1 = 0.736, b_1 = 0.073, a_2 = 0.809, b_2 = 0.634$

TABLE 9. Estimated parameters and BICs, $G = 1$ 'Black' data

Model	Nr. of Param.	BIC	Estimated Parameters
$a_1 = a_2 = 1$	3	1370.6	$p = 0.169, b_1 = 0.061, b_2 = 0.207$
$a_1 = 1$	4	1359.9	$p = 0.169, b_1 = 0.061, a_2 = 0.703, b_2 = 0.266$
$a_2 = 1$	4	1322.3	$p = 0.218, a_1 = 0.645, b_1 = 0.061, b_2 = 0.207$
$a_1 = a_2 = a$	4	1320.1	$p = 0.217, a = 0.649, b_1 = 0.061, b_2 = 0.280$
Free	5	1325.4	$p = 0.218, a_1 = 0.645, b_1 = 0.061, a_2 = 0.703, b_2 = 0.266$

of job-acceptance is declining in the Weibull segments. The parameters seems to imply that the bonus group $G = 1$ are finding employment at exactly the same rate

as the control group before and after the week 27 cutoff. The only difference is that a higher proportion of the bonus group find employment at the 27 week mark.

The fitted distributions are shown in Figure 4.

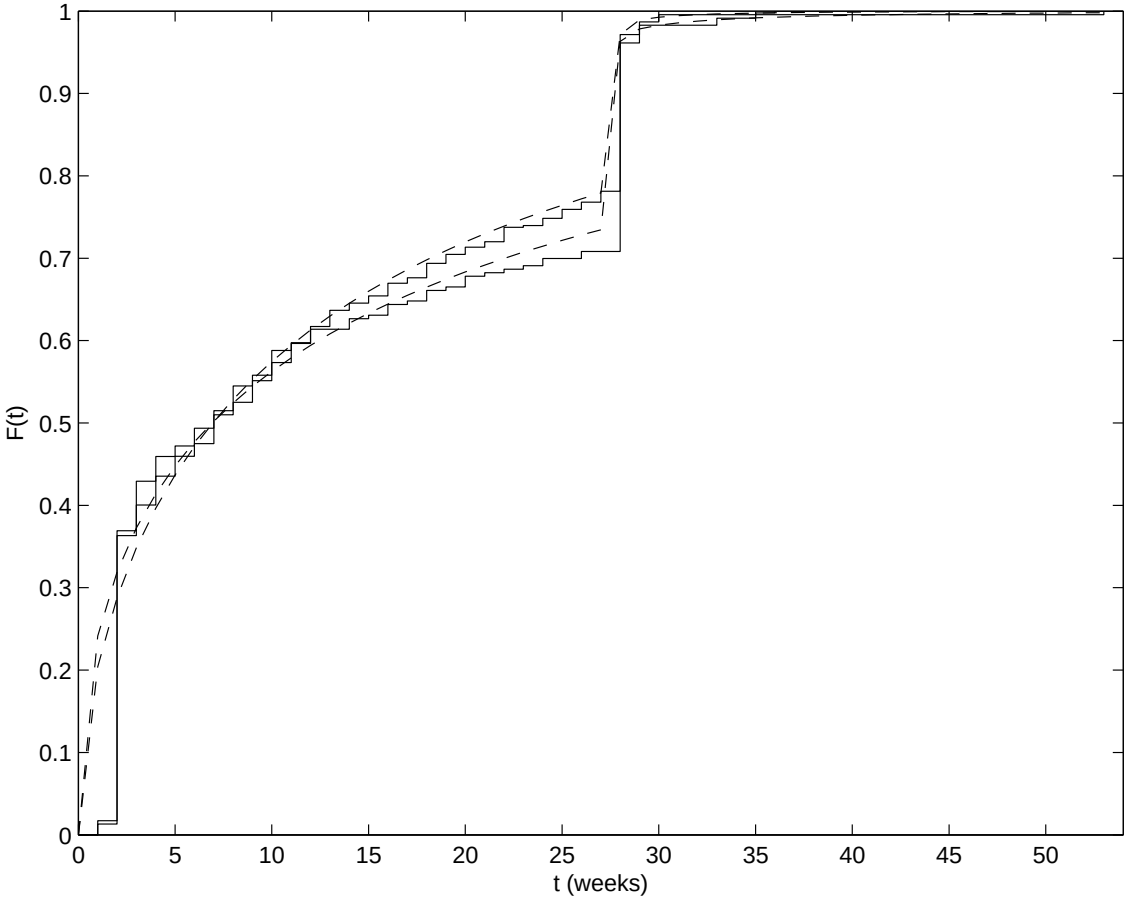

FIGURE 4. Observed (solid) and fitted (dashed) unemployment distributions for 'Black' data. $G = 0$ (lower curves) and $G = 1$ (upper curves).

Fitting the model to the combined 'Black' data we obtain parameter estimates $p = 0.1841, a = 0.7046, b_1 = 0.0691, b_2 = 0.4288$. Using these to repeat, for the 'Black' data, the exercise of estimating the type I error probability we obtain the results given in Table 10 (discretized data) and Table 11 (continuous data). The results are very similar to those from the 'White' data.

TABLE 10. Type I error probabilities, aggregated PBE ('Black') data.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.020	0	0.010	0.030
200	0.025	0	0.035	0.045
500	0.020	0	0.026	0.032
1000	0.029	0	0.029	0.035
2000	0.031	0	0.035	0.033
5000	0.028	0	0.031	0.028

TABLE 11. Type I error probabilities, continuous aggregated PBE ('Black') data.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.050	0	0.040	0.050
200	0.040	0	0.030	0.050
500	0.038	0	0.038	0.046
1000	0.040	0	0.039	0.046
2000	0.041	0	0.051	0.034
5000	0.054	0	0.054	0.053

On continuous data simulated using the parameter estimates from Tables 8 and 9, the rejection rates for our tests of interest are obtained as shown in Table 12. We observe that in this case the power of the test is increasing with the sample size for all tests, as expected.

The information obtained above from the shape parameters of our fitted distributions give another perspective of the differences on take out rates between 'Whites' and 'Blacks'. These results may suggest that while 'Whites' start to look more actively to find a job from the early stages of the unemployment status, 'Blacks' are probably waiting longer for appropriate jobs, as they become more active at the end

TABLE 12. Power of test, continuous aggregated PBE ('Black') data.

N	Test (1)	Test (2)	Test (3)	Test (4)
100	0.033	0	0.041	0.028
200	0.057	0	0.083	0.044
500	0.108	0.001	0.119	0.081
1000	0.237	0.005	0.228	0.148
2000	0.556	0.043	0.527	0.292
5000	0.988	0.584	0.934	0.794

of the bonus period. This difference in attitude towards finding employment even with a significant monetary incentive between the two groups can be actually the root of the relatively poor results of the treatment experiment. Better take out rates of 'Blacks' would have increased significantly the treatment effects. Consequently, policymakers that are designing similar experiments should consider incentives that take into consideration differences in reaction to incentives between different subpopulations. Finding optimum ways to provide incentive to find employment may require a careful analysis of the subpopulations of interest.

5. Conclusion

For a better understanding of the effects of different experiments designed to improve policies that are targeting changes in outcome variables of general economic interest one needs to look at the effects of the experiment on the overall distribution of the outcome variable of interest of the targeted population. The results of such analysis will help the policy makers to identify the risks and the benefits of implementing such policy. Therefore, identifying the subpopulations of the treated group that are either responding or not responding properly to the treatment is of

extreme relevance in assessing the risks of implementing the new policy. Additionally, this type of identification may help the designers of the experiment to design better experiments that are targeting the overall population by understanding why some subpopulations did not respond positively to a specific treatment. Running an alternative experiment that may produce a better response of the targeted population is less costly than implementing a policy that is based on the results of an inefficient experiment. The long-term gains of better response rates of a policy will offset the expenses of running an alternative experiment that will result in a more effective policy.

We have shown in this paper the potential for using entire distributions and, consequently, empirical processes for constructing statistical tests that would statistically justify claims, or reject them, about a variety of inter-relationships between (sub-) populations. In particular, we have explored tests for dominance and intersection of two cdf's in non-parametric and parametric contexts, suggested a parametric model and its appropriate fitting to discrete and continuous data sets, and compared the power of the non-parametric and parametric tests.

These results enable us to make statistically justified conclusions about the (Black, White and Hispanic) sub-populations pertaining to the Pennsylvania Bonus Experiment (PBE). Analyzing the effectiveness of the most successful treatment (treatment 4) while controlling for different race groups, we find that the effect of treatment 4 is significantly different between different race groups, and only a subgroup of individuals benefited from the experiment. This subgroup is also the dominant group in the experiment.

The results of a simulation exercise suggested that when data is complex in nature, for testing purposes, a parametric model and its appropriate fitting to discrete and continuous data sets to be used. The simulation exercise (which is based on matching the distributions of interest from our data) is also providing us with important

information about the characteristics of the data before and after the experimental period. Therefore, the parameter estimates of the shape parameters are used to understand the behavioral dynamics of the unemployed from the treatment group before and after the treatment and also the difference in behavior between the participants in the treatment and control groups.

In particular, for the Whites the experiment generates a complete shift of the distribution of the treated towards a reduction of the unemployment duration. The Blacks had mixed responses to the experiment, those with unemployment durations within the bonus eligibility period responding positively to the treatment only at proximity of the eligibility period termination. Finally, for the Hispanics the experiment generates a complete shift of the unemployment distribution in the opposite direction, implying an adverse treatment effect. This finding shows the relevance of looking at the effect of the experiment on different subpopulations as some subgroups of the population responded differently to the experiment. Hence, to get better results from this sort of experiment, better designs may be required. Such designs should be more flexible to potential response differences of the targeted population so that better policies can be developed.

Acknowledgments

The research of R.Z. has been supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix: Proofs

Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be two (standard) independent Brownian bridges on $[0,1]$, and let Γ be the Gaussian process defined by

$$\Gamma(x) = \sqrt{\eta}\ \mathcal{B}_1(F^{(0,1)}(x)) - \sqrt{1-\eta}\ \mathcal{B}_2(F^{(1,1)}(x)).$$

Theorem 1. *When $\kappa = 0$, then $\lim_{n,m\to\infty} \mathbf{P}[\widehat{K} > x] = \mathbf{P}[\sup_z |\Gamma(z)| > x]$, where the right-most probability does not exceed the probability $\mathbf{P}[\sup_t |\mathcal{B}(t)| > x]$ and is equal to the latter probability when the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ are continuous. When $\kappa > 0$, then the statistic $\widehat{K}$ tends in probability to $+\infty$, thus implying the asymptotic power 1 of the test.*

Proof. We first write $\widehat{K}$ as $\sup_x |\Xi(x)|$, where

$$\Xi(x) = \Delta(x) + \sqrt{\frac{nm}{n+m}}\ (F^{(0,1)}(x) - F^{(1,1)}(x))$$

and

$$\Delta(x) = \sqrt{\frac{nm}{n+m}}\ (\widehat{F}^{(0,1)}(x) - F^{(0,1)}(x)) - \sqrt{\frac{nm}{n+m}}\ (\widehat{F}^{(1,1)}(x) - F^{(1,1)}(x)).$$

When $\kappa = 0$, then $F^{(0,1)} = F^{(1,1)}$ and so $\widehat{K} = \sup_x |\Delta(x)|$. The process Δ converges weakly to the Gaussian process Γ, and thus $\widehat{K}$ converges in distribution to $\sup_z |\Gamma(z)|$, which does not exceed $\sup_t |\mathcal{B}(t)|$ and is equal to it when the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ are continuous. When $\kappa > 0$, then we write the bound

$$\sqrt{\frac{nm}{n+m}}\ |\widehat{\kappa} - \kappa| \le \sup_x |\Delta(x)| \to_d \sup_z |\Gamma(z)|.$$

Since the limiting random variable is non-degenerate and $\sqrt{(nm)/(n+m)}\,\kappa$ converges to infinity, the test statistic $\sqrt{(nm)/(n+m)}\,\widehat{\kappa}$ must converge to infinity in probability. This completes the proof of Theorem 1. □

Theorem 2. *When $\tau = 0$, then $\limsup_{n,m\to\infty} \mathbf{P}[\widehat{T} > y] \le \mathbf{P}[\sup_z |\Gamma(z)| > y]$, where the right-most probability does not exceed $\mathbf{P}[\sup_t |\mathcal{B}(t)| > y]$ and is equal to the latter probability when the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ coincide and are continuous. When $\tau > 0$, then the statistic $\widehat{T}$ tends in probability to $+\infty$, thus implying the asymptotic power 1 of the test.*

Proof. With the earlier defined function $\Xi(x)$, we have that $\widehat{T}$ is the minimum between $\sup_x \Xi(x)$ and $\sup_x(-\Xi(x))$. We have that $\sup_x \Xi(x) \leq \sup_x \Delta(x)$ provided that $F^{(0,1)} \leq F^{(1,1)}$. If, on the other hand, $F^{(0,1)} \geq F^{(1,1)}$, then $\sup_x(-\Xi(x)) \leq \sup_x(-\Delta(x))$. Since we do not know which of the two cases holds (otherwise we would use the previous theorem), we estimate $\widehat{T}$ from above by the maximum of $\sup_x \Delta(x)$ and $\sup_x(-\Delta(x))$. Since the maximum is $\sup_x |\Delta(x)|$, the first half of Theorem 2 follows. To prove the second half of Theorem 2, we note that when $\tau > 0$, then

$$\sqrt{\frac{nm}{n+m}}\,|\widehat{\tau} - \tau| \leq \sup_x |\Delta(x)|.$$

The proof of Theorem 2 is finished. □

Theorem 3. *When $\delta^{\pm} = 0$, then $\lim_{n,m\to\infty} \mathbf{P}[\widehat{D}^{\pm} > y] \leq \mathbf{P}[\sup_z \Gamma(z) > y]$. The above bound becomes equality if we know that $\kappa = 0$, which is a special case of $\delta^{\pm} = 0$. Furthermore, when $\kappa = 0$, then $\mathbf{P}[\sup_z \Gamma(z) > y]$ does not exceed the probability $\mathbf{P}[\sup_t \mathcal{B}(t) > x]$ and is equal to the latter probability when the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$ are continuous. When $\delta^{\pm} > 0$, then both statistics $\widehat{D}^{\pm}$ tend in probability to $+\infty$ thus implying their asymptotic power 1 of the test.*

Proof. We first write the equation $\widehat{D}^{+} = \sup_x \Xi(x)$ with the earlier defined function $\Xi(x)$. When $\delta^{\pm} = 0$, then $\sup_x \Xi(x) \leq \sup_x \Delta(x)$ and the latter converges in distribution to Ψ. Similar arguments are applicable to $\widehat{D}^{-}$ as well; Note that the distributions of $\sup_x \Gamma(x)$ and $\sup_x(-\Gamma(x))$ coincide. To prove the second half of Theorem 3, we first note that when $\delta^{\pm} > 0$, then

$$\sqrt{\frac{nm}{n+m}}\,|\widehat{\delta}^{\pm} - \delta^{\pm}| \leq \sup_x |\Delta(x)|.$$

The second half of Theorem 3 follows. □

References

Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97, 284–292.

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

Bebbington, M., Lai, C.D., and Zitikis, R. (2007), "Modeling Human Mortality Using Mixtures of Bath-tub Shaped Failure Distributions," *Journal of Theoretical Biology*, 245, 528–538.

Bebbington, M., Lai, C.D., and Zitikis, R. (2009), "Identifying Health Inequalities Between Maori and Non-Maori Using Mortality Tables," *Kotuitui: New Zealand Journal of Social Sciences Online*, 4, 103–114.

Corson, W., Decker, P., Shari, D., and Kerachsky, S. (1992), "Pennsylvania Reemployment Bonus Demonstration Final Report," Unemployment Insurance Occasional Paper 92-1, U.S. Department of Labor, Washington, DC.

Crump, R.K., Hotz, V.J., Imbens, G.W. and Mitnik, O.A. (2009) "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96:187–99.

Davidson, R., and Duclos, J.-Y. (2000), "Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality," *Econometrica*, 68, 1435–1464.

Decker, P.T, Olsen, R.B., Freeman, L., and Klepinger, D.H. (2000), "Assisting Unemployment Insurance Claimants: The Long-Term Impacts of the Job Search Assistance Demonstration," W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.

Horváth, L., Kokoszka, P., and Zitikis, R. (2006), "Testing for Stochastic Dominance Using the Weighted McFadden-type Statistic," *Journal of Econometrics*, 133, 191–205.

Koenker, R. and Bilias, Y. (2001), "Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments," *Empirical Economics*, 26: 199-220.

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

McFadden, D. (1989), "Testing for Stochastic Dominance," in Studies in the Economics of Uncertainty, ed. T.B. Fomby and T.K. Seo, New York: Springer-Verlag, pp. 113–134.