A Distributional Analysis of Treatment Effects on Subpopulations of a Socioeconomic Experiment[1]

Liqun Wang

Department of Statistics, University of Manitoba, 332 Machray Hall, Winnipeg, Manitoba R3T 2N2, Canada. E-mail: liqun_wang@umanitoba.ca

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

First draft: March 2009

Revised: February 2010

Abstract. When analyzing treatment effects, the average treatment value is frequently compared to that of the control group. This approach, naturally, is not particularly informative about specific regions of the treatment and control distributions. For this reason and having in view a specific application, in the present paper we consider tests that provide us with more detailed analysis of treatments and their effectiveness. The tests are based on comparing the treatment and control distributions (e.g., whether they are equal, one dominates another, or intersect) over their entire or partial domains of definition. The test of intersection of distributions is introduced in the paper with the scope of pinpointing the region of the tested distribution that is subject to an adverse treatment effect. We illustrate the tests on a simulation study which is based on a matched data and apply them to analyze the Pennsylvania Bonus Experiment.

Classification codes: C12, D31, D63.

Key words and phrases: Stochastic dominance, effective treatment.

[1]We thank Keir Armstrong, William Pouliot and the participants at the LAMES, LACEA, CEA and ESEM meetings for their stimulating comments and suggestions.

1. **Motivation**

Using the example of the Pennsylvania Bonus Experiment (PBE), this paper shows how different subpopulations 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.

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 (12 weeks) 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). In the following sections we provide a detailed analysis of the unemployment duration distribution for three subgroups of the population: Black, Hispanic and White. In addition, we analyze the entire group, called 'All', taking the fourth treatment. Graphical representations of the corresponding empirical distribution functions, provided in Figure 1.1, show apparently different impacts of the treatment on the distribution of unemployment duration for different categories of individuals that are race specific.

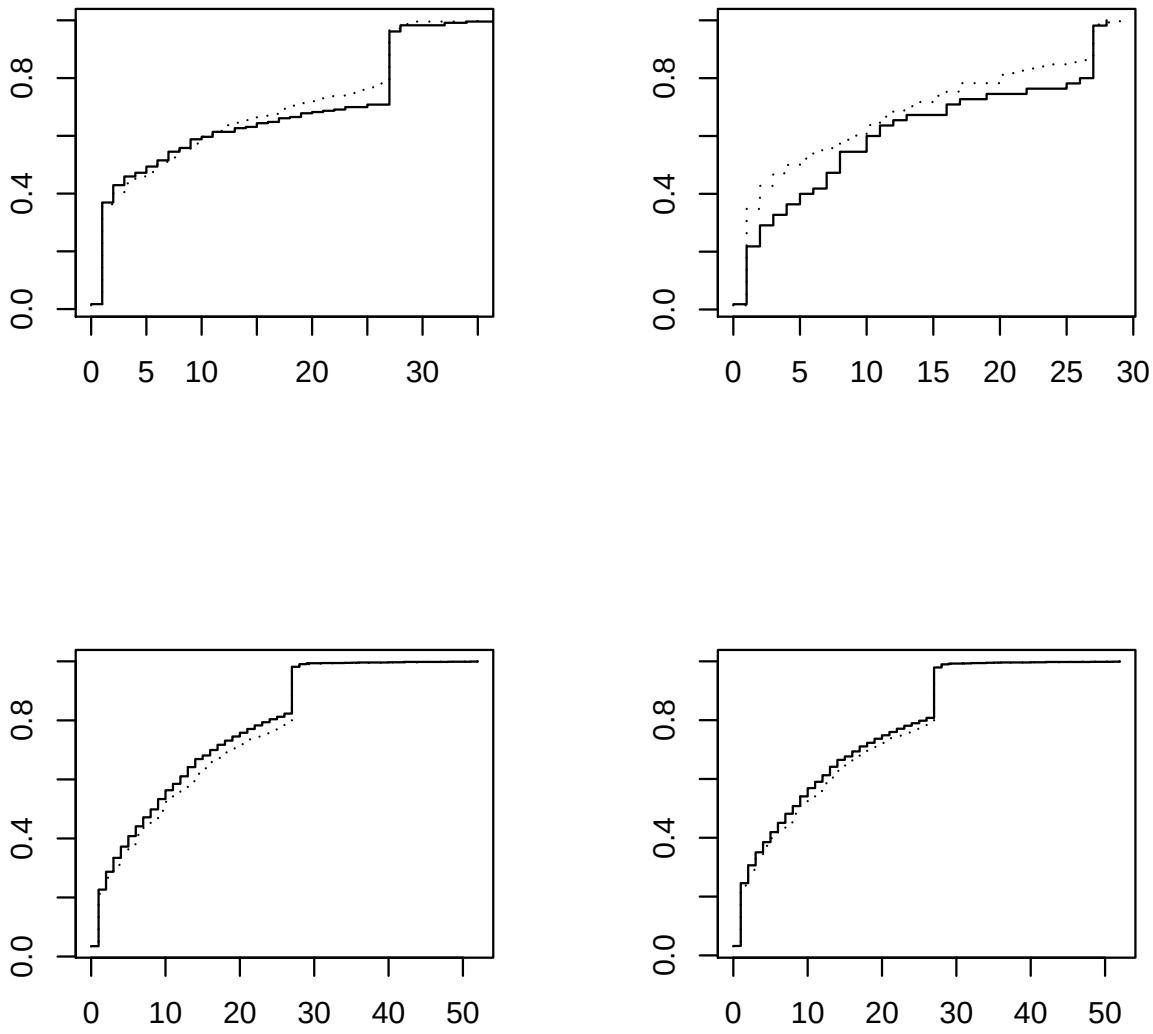

FIGURE 1.1. The control and fourth-treatment empirical distribution functions by race for the Pennsylvania Bonus Experiment: Black (top left), Hispanic (top right), White (bottom left), and All (bottom right). In all four panels, treatment is the solid black line and control is the dotted black line.

It is important to realize, however, that the differences might be due to the sampling variation. To test whether the treatment (solid) and control (dotted) lines coincide, lie one above the other, or intersect, we discuss relevant statistical tests, assess their reliability in a simulation study corresponding to our situation, and then apply

the tests to the original data to draw conclusions about the (most efficient) fourth treatment of the PBE.

The paper is organized as follows. In Section 2 we formulate the problem rigorously. In Section 3 we describe the tests. In Section 4 we apply the tests on different simulation designs. In Section 5 we apply the methodology to analyze the PBE and discuss our findings. Section 6 concludes.

2. **Methodology**

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 at the introduction of a certain treatment policy. The second period, which is denoted by $t = 1$, is the period after the introduction of the treatment policy, or 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 can take on one of the four possible values: $(0, 0)$, $(0, 1)$, $(1, 0)$, and $(1, 1)$. The variable of interest is $Y^{(G,t)}$, which for our motivating examples is the time out-of-work measured in weeks. We are interested in the conditional distribution functions

$$F^{(g,t)}(y) := \mathbf{P}\left[Y^{(G,t)} \leq y |\ G = g\right]$$

for various choices of the parameters $g, t \in \{0, 1\}$.

We assume that at the time of the random assignment (when the treatment was not yet enforced) the control and the treatment groups have the same heterogeneity distribution, which we write as $F^{(0,0)} = F^{(1,0)}$. Given that at the time of the random assignment the individuals just entered the unemployment spell, we use the reported earnings distributions to compare the two groups at the baseline. These corresponding empirical distribution functions for the control and fourth-treatment in the Pennsylvania Bonus Experiment are given in Figure 2.1. The Kolmogorov-Smirnov two-sample test for the equality of distributions has the P-value $P > 0.1$ and thus the test does not reject the null hypothesis $F^{(0,0)} = F^{(1,0)}$ even at 10 percent.

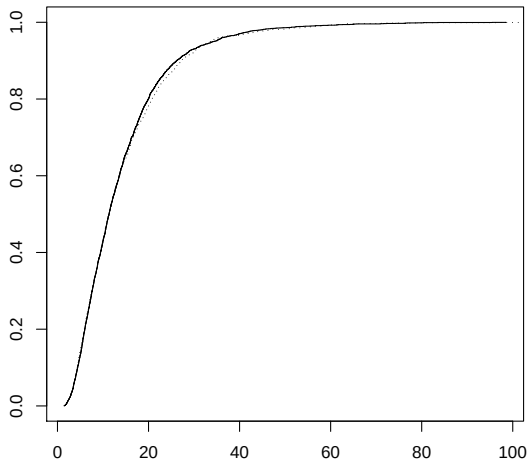

FIGURE 2.1. The control (black line) and fourth-treatment (dashed black line) empirical distribution functions before the introduction of the treatment for the Pennsylvania Bonus Experiment. The baseline earning is used as a proxy outcome variable to compare distributions at the baseline.

As for the two distributions $F^{(0,1)}$ and $F^{(1,1)}$, there are a number of possibilities, which we discuss next. First, we may start by simply testing the following Kolmogorov-Smirnov type hypotheses:

$$\left.\begin{aligned} H_0 &: F^{(0,1)} = F^{(1,1)} \\ H_1 &: F^{(0,1)} \neq F^{(1,1)} \end{aligned}\right\} \tag{2.1}$$

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 1.1, however, the top right panel immediately catches our eyes as it suggests that the treatment for the Hispanic subpopulation does not work, which is indeed so for the particular Hispanic group selected for the experiment but this may not be the case for the entire Hispanic subpopulation in, say, Pennsylvania. The reason is the sampling variability, which depends on the sample size (see Table 4.2). (This also explains why we do not analyze 'other races', whose control and treatment groups have only 21 and 18 observations, respectively.) Hence, our question might be: Is the treatment for the Hispanic subpopulation ineffective (an optimistic statement given the graph) or worse than the control (this is what the top right

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

TABLE 2.1. Sample sizes for the control and fourth-treatment groups at $t = 1$ for the Pennsylvania Bonus Experiment.

panel of Figure 2.1 suggests). Hence, we are interested in testing the hypotheses:

$$\left.\begin{array}{ll} H_0: & F^{(0,1)} = F^{(1,1)} \\ H_1: & F^{(0,1)} \geq F^{(1,1)} \quad \text{with } F^{(0,1)}(x) > F^{(1,1)}(x) \text{ for at least one } x \end{array}\right\} \tag{2.2}$$

Thus, rejecting the null hypothesis would imply that the treatment has not had a positive effect on the whole treatment group. The two bottom panels of Figure 1.1) suggest testing the following hypotheses:

$$\left.\begin{array}{ll} H_0: & F^{(0,1)} = F^{(1,1)} \\ H_1: & F^{(0,1)} \leq F^{(1,1)} \quad \text{with } F^{(0,1)}(x) < F^{(1,1)}(x) \text{ for at least one } x \end{array}\right\} \tag{2.3}$$

Thus, rejecting the null hypothesis would imply that the treatment has had a positive effect on the whole treatment group. The top left panel suggests that there might be an intersection between the control and treatment distributions functions, thus implying that the treatment has *not* been effective for the whole treatment group. We formulate this as the hypotheses:

$$\left.\begin{array}{ll} H_0: & F^{(0,1)} = F^{(1,1)} \\ H_1: & F^{(0,1)} \bowtie F^{(1,1)} \quad \text{meaning intersection, or that there are } x \text{ and } y \text{ such that} \\ & \qquad F^{(0,1)}(x) < F^{(1,1)}(x) \text{ and } F^{(0,1)}(y) > F^{(1,1)}(y) \end{array}\right\} \tag{2.4}$$

Note, however, that the treatment line in the top left panel of Figure 1.1 is just barely above the control line on the left-hand side of the graph and more noticeable below it on the right-hand side. This might indicate dominance, which in turn suggests testing the following hypotheses:

$$\left.\begin{array}{ll} H_0: & F^{(0,1)} \geq F^{(1,1)} \quad \text{with } F^{(0,1)}(x) > F^{(1,1)}(x) \text{ for at least one } x \\ H_1: & F^{(0,1)} \bowtie F^{(1,1)} \end{array}\right\} \tag{2.5}$$

For testing these hypotheses, we can use the test for testing (2.5). The test will be conservative, but 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 is to find out whether the behavior of those in the treatment group has changed or not after the introduction of the new policy when compared to their behavior before the introduction of the policy. For this, we would want to test analogous hypotheses to those formulated in (2.1)-(2.5) but now with $F^{(1,0)}$ instead of $F^{(0,1)}$.

3. **Tests**

Fortunately, there is an extensive literature on testing various hypothesis about two distribution functions, such as the two- and one-sided Kolmogorov-Smirnov tests and their recent extensions and modifications by Linton, Maasoumi and Whang (2005). Hence, we can now concentrate on formulating tests and then, in the following section, on implementing them in a simulation study reflecting the Pennsylvania Bonus Experiment.

The hypotheses are about functions, which can be interpreted as infinite dimensional parameters, which are difficult to deal with. Hence, we shall next find ways (cf., e.g., Linton, Maasoumi and Whang, 2005) of reformulating the hypothesis using one-dimensional parameters. Two quantities will play a crucial role:

$$\theta^- = \sup_y (F^{(0,1)}(y) - F^{(1,1)}(y)) \quad \text{and} \quad \theta^+ = \sup_y (F^{(1,1)}(y) - F^{(0,1)}(y)).$$

Clearly now, we have the equivalence

$$(2.1) \Leftrightarrow \left\{ \begin{array}{l} H_0 : \kappa = \max(\theta^-, \theta^+) = 0 \\ H_1 : \kappa > 0 \end{array} \right\} \tag{3.1}$$

and the following implications:

$$(2.2) \Rightarrow \left\{ \begin{array}{l} H_0 : \theta^- = 0 \\ H_1 : \theta^- > 0 \end{array} \right\} \tag{3.2}$$

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

$$(2.3) \Rightarrow \left\{ \begin{array}{l} H_0 : \theta^+ = 0 \\ H_1 : \theta^+ > 0 \end{array} \right\} \tag{3.3}$$

$$\Leftrightarrow \left\{ \begin{array}{l} H_0 : F^{(0,1)} \geq F^{(1,1)} \\ H_1 : F^{(0,1)}(x) < F^{(1,1)}(x) \text{ for some } x \end{array} \right\},$$

$$(2.4) \Rightarrow \left\{ \begin{array}{l} H_0 : \tau = \min(\theta^-, \theta^+) = 0 \\ H_1 : \tau > 0 \end{array} \right\} \tag{3.4}$$

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

Our next task is to construct empirical estimators for the parameters κ, θ^-, θ^+, and τ. Then, based on the estimators, we shall construct statistics for testing hypotheses (3.1)–(3.4). Our data consist of two sets:

$$Y_1^{(0,1)}, \ldots, Y_n^{(0,1)} \sim F^{(0,1)}, \tag{3.5}$$

$$Y_1^{(1,1)}, \ldots, Y_m^{(1,1)} \sim F^{(1,1)}. \tag{3.6}$$

We assume that all the random variables are independent. However, as specified by the two cdf's $F^{(0,1)}$ and $F^{(1,1)}$, the random variables are identically distributed only within each set (3.5) and (3.6). Denote the corresponding empirical cdf's by $\widehat{F}^{(0,1)}$ and $\widehat{F}^{(1,1)}$. Furthermore, we assume that there is a number $0 < \eta < 1$ such that

$$\frac{m}{n+m} \to \eta \quad \text{when} \quad n, m \to \infty.$$

which is a natural assumption to make the sample sizes n and m comparable. The estimators of the four parameters in (3.1)–(3.4) and the corresponding test statistics are:

$$\widehat{\kappa} = \sup_y |\widehat{F}^{(0,1)}(y) - \widehat{F}^{(1,1)}(y)| \quad \text{and} \quad \widehat{K} = \sqrt{\frac{nm}{n+m}}\, \widehat{\kappa}, \tag{3.7}$$

$$\widehat{\theta^-} = \sup_y (\widehat{F}^{(0,1)}(y) - \widehat{F}^{(1,1)}(y)) \quad \text{and} \quad \widehat{D}^- = \sqrt{\frac{nm}{n+m}}\, \widehat{\theta^-}, \tag{3.8}$$

$$\widehat{\theta^+} = \sup_y (\widehat{F}^{(1,1)}(y) - \widehat{F}^{(0,1)}(y)) \quad \text{and} \quad \widehat{D}^+ = \sqrt{\frac{nm}{n+m}}\, \widehat{\theta^+}, \tag{3.9}$$

$$\widehat{\tau} = \min(\widehat{\theta^-}, \widehat{\theta^+}) \quad \text{and} \quad \widehat{T} = \sqrt{\frac{nm}{n+m}}\, \widehat{\tau}. \tag{3.10}$$

By the classical Glivenko-Cantelli theorem, $\widehat{\kappa}$, $\widehat{\theta^-}$, $\widehat{\theta^+}$, and $\widehat{\tau}$ are strongly consistent estimators of κ, θ^-, θ^+, and τ, respectively. Next we shall discuss critical values

for the test statistics $\widehat{K}$, $\widehat{D}^-$, $\widehat{D}^+$, and $\widehat{T}$. For this, we need auxiliary notation and technical results, formulated as Theorems 3.1–3.3 below.

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

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

Note that when $F^{(0,1)} = F^{(1,1)}$, then $\sup_y |\Gamma(y)| \geq \sup_t |\mathcal{B}(t)|$, and if in addition the cdf's are continuous, then the two supremums are equal. The distribution of $\sup_t |\mathcal{B}(t)|$ is known as the Kolmogorov-Smirnov distribution. We shall use the following notation:

$$\Lambda = \sup_y |\Gamma(y)| \quad \text{and} \quad \Psi = \sup_y(\Gamma(y)).$$

Theorem 3.1. *When $\kappa = 0$, then $\limsup_{n,m\to\infty} \mathbf{P}\big[\widehat{K} > y\big] = \mathbf{P}\big[\Lambda > y\big]$. When $\kappa > 0$, then the statistic $\widehat{K}$ tends in probability to $+\infty$ thus implying its asymptotic power* 1.

Proof. When $\kappa = 0$, then $F^{(0,1)} = F^{(1,1)}$ and so $\widehat{K} = \sup_y |\Delta(y)|$, where

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

Hence, $\widehat{K}$ converges in distribution to Λ. To prove the second half of Theorem 3.1, we first note that under the alternative we have the equality $\widehat{K} = \sup_y |\Xi(y)|$, where

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

Obviously now,

$$\sqrt{\frac{nm}{n+m}}\ |\widehat{\kappa} - \kappa| \leq \sup_y |\Delta(y)| \to_d \Lambda.$$

The second half of Theorem 3.1 follows. □

Using Theorem 3.1 and the notes in the paragraph preceding it, we have that the following rejection region for the statement $\kappa = 0$ in favor of $\kappa > 0$:

$$\widehat{K} > k_\alpha, \tag{3.11}$$

where k_α is the α-critical value of the (classical) Kolmogorov-Smirnov test. The next section shows that the outcome distribution is a mixture of distributions and, therefore, has nuisance parameters. Under the presence of the nuisance parameters,

we cannot use the asymptotic k_α-critical value as it is not distribution free. Therefore, the k_α-critical value is estimated using a bootstrap. Huynh and Voia (2008) show that a parametric bootstrap can achieve the levels for some classes of stochastic dominance tests and a re-centered non-parametric bootstrap is not suitable for distributions that are nuisance parameters dependent. To deal with the mixing distributions, the parametric bootstrap is used to compute the critical values of all test statistics. The steps for implementing the parametric bootstrap are outlined in subsection 4.1 below.

Theorem 3.2. *When $\theta^{\pm} = 0$, then $\limsup_{n,m\to\infty} \mathbf{P}\big[\widehat{D}^{\pm} > y\big] \leq \mathbf{P}\big[\Psi > y\big]$. The above bound becomes equality if we know that $\kappa = 0$, which is a special case of $\theta^{\pm} = 0$. When $\theta^{\pm} > 0$, then both statistics $\widehat{D}^{\pm}$ tend in probability to $+\infty$ thus implying their asymptotic power of 1.*

Proof. We first write the equation $\widehat{D}^{+} = \sup_y(\Xi(y))$ with the earlier defined function $\Xi(y)$. When $\theta^{\pm} = 0$, then $\sup_y(\Xi(y)) \leq \sup_y(\Delta(y))$ and the latter converges in distribution to Ψ. Similar arguments are applicable to $\widehat{D}^{-}$ as well. (Note that the distributions of $\sup_y(\Gamma(y))$ and $\sup_y(-\Gamma(y))$ coincide.) To prove the second half of Theorem 3.2, we first note that when $\theta^{\pm} > 0$, then

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

The second half of the theorem follows. □

The (conservative) critical value of the test depends on the distribution of Ψ, which is not distribution free unless $F^{(0,1)} = F^{(1,1)}$, as well as on the continuity of the cdf's. Hence, due to the nuisance parameters problem, we use parametric bootstrap as in subsection 4.1.

Theorem 3.3. *When $\tau = 0$, then $\limsup_{n,m\to\infty} \mathbf{P}\big[\widehat{T} > y\big] \leq \mathbf{P}\big[\Lambda > y\big]$. When $\tau > 0$, then the statistic $\widehat{T}$ tends in probability to $+\infty$ thus implying its asymptotic power of 1.*

Proof. With the earlier defined function $\Xi(y)$, we have that $\widehat{T}$ is the minimum between $\sup_y(\Xi(y))$ and $\sup_y(-\Xi(y))$. We have that $\sup_y(\Xi(y)) \leq \sup_y(\Delta(y))$ provided that $F^{(0,1)} \leq F^{(1,1)}$. If, on the other hand, $F^{(0,1)} \geq F^{(1,1)}$, then $\sup_y(-\Xi(y)) \leq \sup_y(-\Delta(y))$. If we do not know which of the two cases holds, we estimate $\widehat{T}$ from above by $\max\{\sup_y(\Delta(y)), \sup_y(-\Delta(y))\}$, which is $\sup_y |\Delta(y)|$. The first half of

Theorem 3.3 follows. To prove the second half of the theorem, we note that when $\tau > 0$, then

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

The second half of the theorem follows. □

To construct the (conservative) rejection region, we again employ the parametric bootstrap as in subsection 4.1.

4. **Simulation designs**

An assessment of the performance of the tests discussed in the previous section will be examined in this section. The proposed tests are designed to control for the significance level for each of the above tests. If the distribution of the variable of interest is a function of finite mixtures, the null distributions of our test statistics are nuisance-parameter dependent, even asymptotically. When such problems arise, the bootstrap is often suggested to obtain test-specific critical values; see, for example, Barrett and Donald (2003) or Dufour (2006). Nevertheless, bootstrap and various simulation-based methods may also fail if the nuisance parameter problem is highly irregular; see Dufour (1997). Further, the proposed tests are conservative as we are looking to the least favorable model under the null. This would result in lower power, which can be aggravated by the fact that our distributions have nuisance parameters. Therefore, to avoid such problems, a parametric bootstrap is employed to conduct inference. To construct the parametric bootstrap, the nuisance parameters are estimated using finite mixture distribution decompositions of the outcome variable. The results of the finite mixture decompositions are then used to improve the size and power of the tests employed in the paper.

We did two things to check whether our tests yield good results in relation to our data. First, we fitted the data using a finite mixture model and second, we performed Monte-Carlo simulations on the fitted data. To fit our data we took the following steps:

- we used a histogram to plot the density of our true data which shows that our data is a mixture of distributions,
- we assumed that the true density is a weighted sum of log-normal densities with different expected values (E(y)) and variances (the density plot shows that our data mimic a mixture of log-normal distributions); therefore,

- we estimated the parameters of the mixture by maximum likelihood.

The following likelihood function was used:

$$f(y,\theta)=\sum_{k=1}^{K}p_k\frac{1}{y\sigma_k\sqrt{2\pi}}e^{\frac{-(\ln y-\mu_k)^2}{2\sigma_k^2}},$$

The parameters of interest are: $\theta=\{K,p_k,\mu_k,\sigma_k\}$ with $k=1,...,K$ and $\sum_{k=1}^{K}p_k=1$. All the parameters of interest with the exception of the number of types are estimated by maximum likelihood. The number of types was estimated using model selection based on the AIC criterion. The following AIC criterion was minimized:

$$AIC_k=-2\log l\left(\theta|y\right)+2d_k, \tag{4.1}$$

where d_k is equal to the dimension of the model and acts as a correction term without which one would choose the model that maximizes the unconditional log-likelihood.

Our findings show that our full data is fitted by a mixture of three log-normal distributions with the following parameters:

Type	Proportion	E(y)	Var(y)
I	0.286	1.174	0.179
II	0.470	8.373	19.912
III	0.244	27.113	0.018

TABLE 4.1

The control group data is fitted by a mixture of three log-normal distributions with the parameters as in Table 4.2:

Type	Proportion	E(y)	Var(y)
I	0.286	1.231	0.153
II	0.470	8.873	19.912
III	0.244	27.800	0.077

TABLE 4.2

The treated group data is fitted by a mixture of three log-normal distributions with the parameters as in Table 4.3:

Type	Proportion	E(y)	Var(y)
I	0.286	1.285	0.583
II	0.470	8.331	18.837
III	0.244	27.122	0.536

TABLE 4.3

Using the result of the mixture decomposition, we want to construct a DGP of size n from

$$p_1 f_1(y) + p_2 f_2(y) + p_3 f_3(y).$$

To generate our DGPs of interest the following steps were required:

- divide the interval $[0, 1]$ into three subintervals of lengths p_1, p_2 and p_3 (where $p_1 + p_2 + p_3 = 1$) ,
- generate n random observations uniformly distributed on $[0, 1]$,
- count the number of the uniform random observations in the three subintervals and denote the counts by n_1, n_2, and n_3,
- compute the means and variances of the transformed random variables $ln(y)$ on the three subintervals, and
- draw:

  (1) n_1 observations from the lognormal distribution

$$f_1(y) = \frac{1}{y\sigma_1\sqrt{2\pi}} e^{\frac{-(\ln y - \mu_1)^2}{2\sigma_1^2}}$$

  with

$$\mu_1 = \ln(E_1(y)) - 0.5\ln(1 + \frac{Var_1(y)}{E_1(y)^2}) = 0.1$$

  and

$$\sigma_1^2 = \ln(1 + \frac{Var_1(y)}{E_1(y)^2}) = 0.55,$$

  (2) n_2 observations from the lognormal distribution

$$f_2(y) = \frac{1}{y\sigma_2\sqrt{2\pi}} e^{\frac{-(\ln y - \mu_2)^2}{2\sigma_2^2}}$$

  with

$$\mu_2 = \ln(E_2(y)) - 0.5\ln(1 + \frac{Var_2(y)}{E_2(y)^2}) = 2$$

  and

$$\sigma_2^2 = \ln(1 + \frac{Var_2(y)}{E_2(y)^2}) = 0.49,$$

(3) n_3 observations from the lognormal distribution

$$f_3(y) = \frac{1}{y\sigma_3\sqrt{2\pi}} e^{\frac{-(\ln y - \mu_3)^2}{2\sigma_3^2}}$$

with

$$\mu_2 = \ln(E_3(y)) - 0.5\ln(1 + \frac{Var_3(y)}{E_3(y)^2} = 3.3$$

and

$$\sigma_3^2 = \ln(1 + \frac{Var_3(y)}{E_3(y)^2}) = 0.027.$$

$E_k(y)$ and $Var_k(y)$, $k = \{1, 2, 3\}$ are defined as in Table 4.1.

- Stack together the three data sets to get the random sample with n observations.

Using the above procedure to construct our DGPs of interest, we simulated two data sets $Y_1^{(0,1)}, \ldots, Y_N^{(0,1)}$ and $Y_1^{(1,1)}, \ldots, Y_N^{(1,1)}$. The graphs in figure 4.1 show the fit of the simulated data in relation to the true data. Using the estimated nuisance parameters obtained above, we computed the critical values and the p-values of the tests by employing a parametric bootstrap as follows:

4.1. **Bootstrap.** Huynh and Voia (2008) suggests the use of a parametric bootstrap to control the level of the EoD, FOSD and SOSD tests that are nuisance-parameter dependent. The present paper shows that the parametric bootstrap also controls the level of the new test of intersection introduced herein.

The parametric bootstrap was used to simulate the critical values for the EoD test in the following fashion:

(1) Sample n-values from $Y_1^{(0,1)}, \ldots, Y_n^{(0,1)}$ from the estimated distributions obtained using the control group data:

$$\int_0^y \widehat{f}_{duration}(s)ds \quad = \quad \int_0^y \sum_{k=1}^{K} \widehat{p_k} \frac{1}{s\widehat{\sigma_k}\sqrt{2\pi}} \exp\left(\frac{-(\ln s - \widehat{\mu_k}^2}{2\widehat{\sigma_k}^2}\right) ds,$$

(2) Then sample using the treatment estimated distribution m values from $Y_1^{(1,1)}, \ldots, Y_m^{(1,1)}$.

(3) The distributions are adjusted to be stochastically equal under the null hypothesis.

(4) Using the resulting empirical distribution functions, $\widehat{F}^{(0,1)*}(y)$ and $\widehat{F}^{(1,1)*}(y)$, define

$$\widehat{K}^* = \sup_y \sqrt{\frac{nm}{n+m}} \left| \widehat{F}^{(0,1)^*}(y) - \widehat{F}^{(1,1)^*}(y) \right|.$$

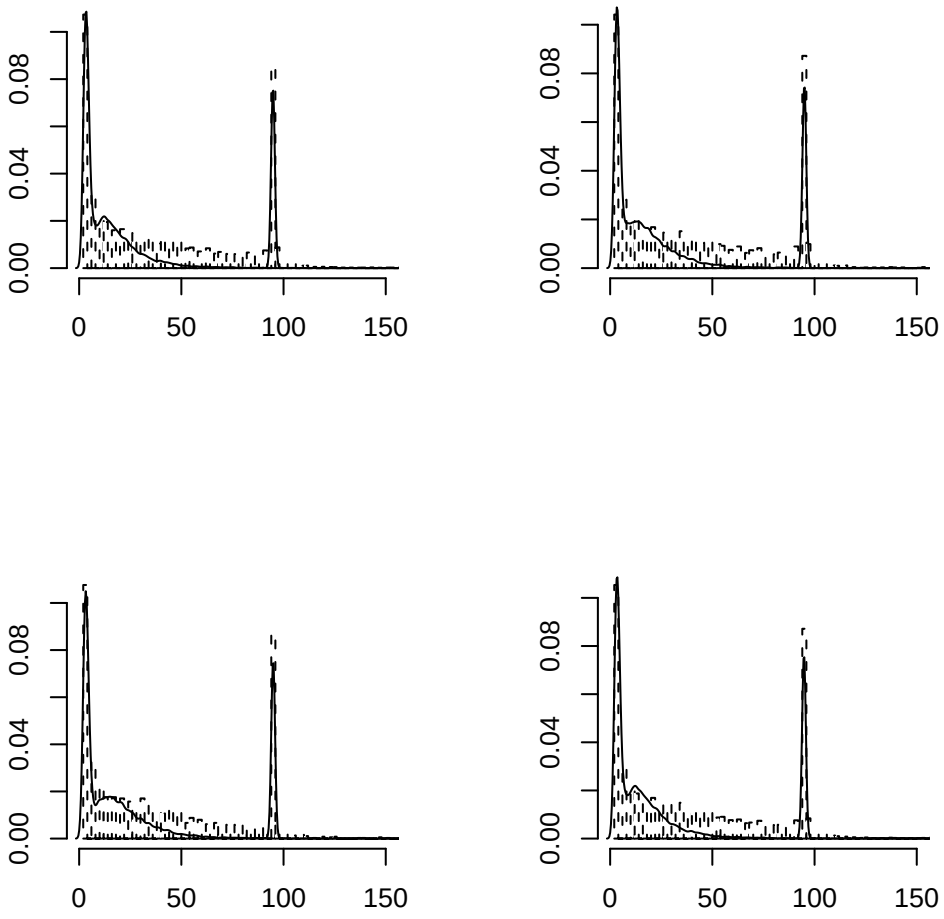

FIGURE 4.1. Top left: histogram of all data (dashed black line) and density of fitted data (solid black line); top right: histogram of control data (dashed black line) and density of fitted control data (black line); bottom left: histogram of treated data (dashed black line) and density of fitted treated data (solid black line); bottom right: histogram of pooled data (solid black line) and density of fitted pooled data (dashed black line).

(5) Repeat steps 1 through 3 B times and define the critical value k_α^* as the smallest value of y subject to at least $100(1-\alpha)\%$ of the obtained B values of $\widehat{K}^*$ that are at or below y.
(6) The rejection region is $\widehat{K} > k_\alpha^*$.

To visualize the distribution of the 1000 p-values for each of the four groups specified above under the null of equality of distributions, we produced the histograms in Figure 4.2.

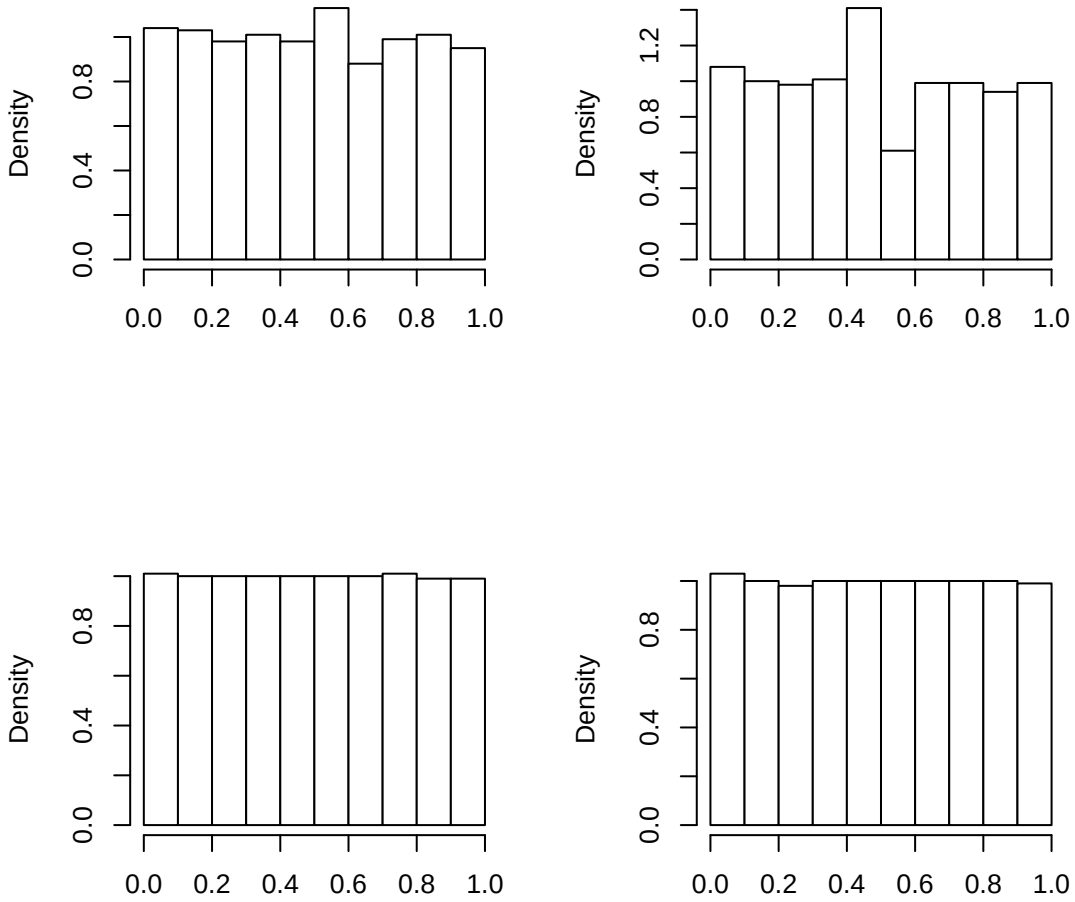

FIGURE 4.2. Histogram of p-values for the subsample of: Blacks (top left), Hispanics (top right), Whites (bottom left), and All (bottom right).

A uniform distribution for the p-values would suggest a good power of the test statistics. The results show that larger samples sizes give better power for the tests than smaller sample sizes.

To estimate the critical values for the FOSD test the same steps were followed as in the EoD case, but with $\Delta^*(y)$ defined as

$$\left\{\sqrt{\frac{nm}{n+m}}\ \left(\widehat{F}^{*(0,1)}(y) - \widehat{F}^{*(1,1)}(y)\right)\right\},$$

and $\widehat{D}^* = \sup_y \Delta^*(y)$. d^*_α is defined as the smallest y such that at least $100(1-\alpha)\%$ of the obtained B values of $\widehat{D}^*$ are at or below y. The (conservative) rejection region is

$$\widehat{D} > d^*_\alpha. \tag{4.2}$$

In estimating the critical values for the Intersection of Distributions test, the same steps were followed as in EoD and FOSD cases, but with the use of the above process $\Delta^*(y)$. Let $\widehat{T}^*$ be the *maximum* of $\sup_y \Delta^*(y)$ and $\sup_y(-\Delta^*(y))$. We define t^*_α to be the smallest y such that at least $100(1-\alpha)\%$ of the obtained B values of $\widehat{T}^*$ are at or below y. The rejection region is

$$\widehat{T} > t^*_\alpha, \tag{4.3}$$

Using the simulated data, we performed the test for the null hypothesis $H_0^{(\text{not } 3)}$: $F^{(0,1)}$ dom $F^{(1,1)}$. Theorem 3.3 says that under the hypothesis $H_0^{(\text{not } 3)}$ the test statistic $\widehat{T}$ is such that, asymptotically, $\mathbf{P}[\widehat{T} > y_\alpha]$ does not exceed the significance level α whenever y_α solves the equation $\mathbf{P}[\max(\Gamma_+, \Gamma_-) > y_\alpha] = \alpha$. The critical value y_α is not distribution free, and so the asymptotic p-value of the test, $\mathbf{P}^*[\max(\Gamma_+, \Gamma_-) > \widehat{T}]$, is not calculable.

$$\mathbf{P}^*[\max(\Gamma_+, \Gamma_-) > \widehat{T}] \approx \mathbf{P}^*[\widehat{T}^* > \widehat{T}],$$

where $\widehat{T}^* := \max(\sup_y(\Delta^*(y)), \sup_y(-\Delta^*(y))$ with the same $\Delta^*(y)$ as above.

For each of the analyzed samples we simulated 1000 sets of random variables and obtained 1000 values of $\widehat{T}$. For each value of $\widehat{T}$, we then calculated $\mathbf{P}^*[\widehat{T}^* > \widehat{T}]$ using 1000 bootstrap iterations. Hence, for each value of $\widehat{T}$ we obtained a value for $\mathbf{P}^*[\widehat{T}^* > \widehat{T}]$, which is an approximate p-value of the test. The results are presented in Table 4.4.

The results suggest that, for Blacks, equality and dominance were rejected while the test for intersection confirms the intersection of the two distributions. The test for intersection can be used to test the intersection on restricted supports, and in this respect can be used to identify multiple intersections. In the case of Blacks, the treatment becomes ineffective around 12 weeks and lasts until about

Group	n	m	Equality p-value	Dominance p-value	Intersection p-value
Black	457	233	0.001	0.135	0.000
Hispanic	138	55	0.000	0.066	0.028
White	2979	1571	0.002	0.014	0.982
All	3595	1877	0.007	0.019	0.967

TABLE 4.4

the 25th week. The subsample of Hispanics is small, and the tests employed were therefore not that precise. However, for this group, we found evidence of ineffective treatment for almost all the support of the outcome variable. There is some evidence of effective treatment during the eligibility period and at the end of the period (the test for intersection does not reject). As Whites represent almost 83% of the sample, this group determines the results for the entire sample. Consequently we found evidence of effective treatment for Whites and for the sample as a whole. To improve the results of the treatment exercise, a careful researcher should also look at the implications of the effectiveness of the treatment on different subgroups of a given sample and target the ones that are responding to treatment in an unpredicted way. Due to their proportion in the sample, Whites were driving the results for the sample as a whole (results also confirmed by the Average Treatment Effect found in the data). In the PBE example, the composition of the race groups was not heterogeneous enough to change the outcome of the experiment. However, different outcomes may be possible if different compositions were found in the data.

4.2. **Simulation of Level and Power for the proposed tests.** Next, we simulated the level and the power of the above test statistics. Given the complications that are due to finite mixture distributions and the fact that we could not find theoretical critical values, we provide an example of a simulated level and power for the test statistics that are used in this example for the subsample of Blacks. With the use of the results obtained using the finite mixture decomposition, the DGPs for the fitted control group data and treatment group data for the four cases $n = \{200,\ 500,\ 1000,\ 2000\}$ were generated. For level analysis, distributions were adjusted to be stochastically equal (compare the pooled simulated mixtures of the treated and control distribution with the simulated mixture distribution obtained by pooling the treated and control group data). For power analysis, the distributions

were transformed so as to be stochastically unequal. Therefore, we compared the treated and control simulated mixtures distributions for which we have rejection of equality.

The following steps were used for the level analysis:

- choose a critical value for the rejection region (i.e., 0.1 , 0.05, 0.01);
- generate the treatment, control and pooled treated and control mixtures for different sample sizes (N=$\{$200, 500, 1000, 2000, 5000$\}$) - the sample sizes were the same for the three groups;
- pool the mixtures of the treatment and control together;
- generate the three tests (equality, dominance, intersection) for the difference between the simulated mixture of the two pooled distributions with the pooled of two mixture distributions;
- do the 3 tests 1000 times for the above-named sample sizes;
- order the results of the three tests;
- choose the cut-off points and the critical values based on the rejection probabilities;
- count the number of rejections and find the proportion of the rejections;
- define the level of the test as the proportion of the rejections.

Tables 4.5 and 4.6 present the results of the level exercise when a non-parametric re-centered bootstrap and a parametric bootstrap are employed. The non-parametric re-centered bootstrap of Barrett and Donald (2003) was used.

Test	Equality			Dominance			Intersection		
α	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
N=200	0.0024	0.0130	0.0301	0.0024	0.0131	0.030	0.0114	0.0452	0.0899
N=400	0.0028	0.0131	0.0598	0.0028	0.0199	0.059	0.0084	0.0472	0.0933
N=1000	0.0051	0.0219	0.0827	0.0051	0.0219	0.0827	0.0103	0.0476	0.0978
N=2000	0.0095	0.0229	0.0918	0.0095	0.0229	0.0920	0.0097	0.0481	0.0999

TABLE 4.5

The simulation exercise shows that, in the presence of finite mixtures, the parametric bootstrap does a better job than the centered non-parametric bootstrap in achieving the desired level, especially for the Equality of Distribution test and FOSD.

Test	Equality			Dominance			Intersection		
α	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
N=200	0.013	0.06	0.07	0.014	0.035	0.06	0.01	0.03	0.065
N=500	0.01	0.045	0.12	0.012	0.051	0.092	0.01	0.054	0.099
N=1000	0.01	0.047	0.1	0.013	0.052	0.011	0.012	0.055	0.103
N=2000	0.012	0.051	0.1	0.011	0.048	0.011	0.0103	0.051	0.101

TABLE 4.6

For the test of Intersection of Distributions, both bootstrap methods perform similarly. Given the results for the level, the parametric bootstrap should be used to obtain the desired critical values for the analyzed test statistics in the presence of nuisance parameters. Therefore, for the power exercise, we present only the results obtained using a parametric bootstrap.

Designing a complete power exercise for test statistics that are subject to nuisance parameters may require separate attention as the power exercise can be very challenging. In the presence of nuisance parameters a complete power exercise requires a mapping on a multidimensional dimensional space (the space generated by the nuisance parameters estimated by the finite mixtures: number of mixtures, share of each distribution in the mixture, mean value of each distribution in the mixture and the variance of each distribution in the mixture). For this reason, in the present paper we show the power of the analyzed test statistics on a subsample of the analyzed data. For this analysis, the distributions were transformed to be stochastically unequal by comparing the simulated mixtures distributions of interest for which we have clear rejection of equality. As for the level exercise, we used the information from the data on Blacks obtained using our finite mixture decomposition. Similar steps as for the level exercise were used to compute the power, but we added another step that computed the actual power by subtracting from one the proportions of rejections. The results for the power exercise are presented in Table 4.7.

The results for the power exercise show that, for our special case, the parametric bootstrap gives very good results for the dominance and intersection tests and reasonable results for the equality of distribution test.

Test	Equality			Dominance			Intersection		
α	0.01	0.05	0.1	0.01	0.05	0.1	0.01	0.05	0.1
N=200	0.89	0.625	0.38	1	1	1	1	0.995	0.985
N=500	0.92	0.79	0.72	1	1	1	1	1	0.99
N=1000	0.935	0.8	0.68	1	1	1	1	1	1
N=2000	0.95	0.8	0.79	1	1	1	1	1	1

TABLE 4.7

5. **Discussion of Results**

Our findings suggest that treatment 4 has significantly different impacts on different race groups, with Whites benefiting the most from this type of treatment. Actually, Whites are the only group that drives the positive results of this treatment. They are also the dominant group in terms of sample size, accounting for about 83% of the data. The dominance test shows that the treatment group duration of the unemployment distribution dominates the control group's distribution at 5% (cf. Table 4.4, row 3). Similar results hold when all data are considered (cf. Table 4.4, row 4). Given the large sample sizes that were used for the Whites group, the power of the employed test statistics should be high when this group and all data are tested (see Figure 4.2.c,d). There are no treatment benefits for the Hispanics group. On the contrary, treatment 4 has a negative impact on unemployment over its entire distribution. The tests of dominance and intersection of the duration of unemployment distribution for the control group over the treatment group's distribution show that dominance exists for almost all of the support of the outcome variable (cf. Table 4.4, row 2). The power of the test statistics is lower than in the previous cases given the number of individuals from the treatment and control groups (cf. Figure 4.2.b), but the strong result of rejecting the null of equality of distributions may overcome the loss of power due to small sample size. For Blacks, there is a marginal benefit only during the bonus eligibility period. We observed a clear intersection of the treatment group duration of unemployment with the control group's distribution around 12 weeks. A null test of dominance of the treatment group duration of unemployment distribution over the control group's distribution is rejected against the alternative of an intersection (cf. Table 4.4, row 1). This result is reinforced by the strong rejection of the Equality of Distributions test, which also

has a reasonably good power (cf. Figure 4.2.a). The results show that the treatment is not effective for Blacks.

Our findings show that there was only a subgroup of individuals that benefited from the PBE. This subgroup was also the dominant group of the experiment. However, the experiment was designed to test the effect of the bonus on an entire population, and this finding shows that some subgroups of the population responded differently to the experiment. Therefore, it is possible that the design of the bonus did not address the race differences sufficiently carefully, which induced an overestimation of the positive response to the bonus of different race groups. To overcome this problem, a more flexible design would be required to address the potential response differences of the targeted population and thereby facilitate the development of better policies.

6. **Conclusions**

Using only the average treatment effect to evaluate specific treatment programs ignores what happens in different regions of the distribution of the outcome variable of interest. In this paper we consider statistical tests that can identify whether a treatment is effective over the entire distribution of a treated group when it is compared with a control group's distribution. Tests of Equality of Distributions, FOSD and a new test for the Intersection of Distributions are considered. The test of Intersection of Distributions can be used as an alternative for the FOSD test as it is more informative than a test of first-order stochastic dominance. There are two reasons for this: first, evidence of intersection of the treatment and control group distributions implies rejection of FSOD; second, evidence of intersection is a clear indication that the treatment is not effective for all individuals from the treatment group. To show how to implement the tests easily, an outline of how to estimate critical values using a parametric bootstrap method is presented when nuisance parameters are present in the employed test statistics. To assess the actual performance of the tests, simulation studies are conducted. The simulation results show the effect of finite samples properties for both the level and power of the test statistics. Larger sample sizes are helpful for all the tests that are considered in the paper.

We apply these tests to analyze the effectiveness of treatment 4 from the Pennsylvania Bonus Experiment. 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. This finding also shows that some subgroups of the population responded differently to the experiment. To get better results for 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.

References

Abadie, A. (2000), Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. Journal of the American Statistical Association, 2002, vol. 97, 284-292.

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

Athey, S. and Imbens, G. (2003), Identification and Inference in Nonlinear Difference-In-Differences Models. NBER Technical Working Paper No. t0280.

Barrett, G.F. and Donald, S.G. (2003), Consistent tests for stochastic dominance, Econometrica 71, 71-104.

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

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.

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.

Dufour, J.-M. (1997), Some Impossibility Theorems in Econometrics with Applications to Structural and Dynamic Models. Econometrica, 65(6), 1365-1388.

Dufour, J.-M. (2006), Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics. Journal of Econometrics, 133(2), 443-477.

Fraker, T. and Maynard, R. (1987), The Adequacy of Comparison Group Designs for Evaluations of Employment Related Programs. Journal of Human Resources, 22, 194-227.

Heckman, J. (1992), Randomization and Social Policy Evaluation. In Charles Manski and Irwin Garfinkle, eds., Evaluating Welfare and Training Programs (Cambridge, Mass.: Harvard University Press), 201-230.

Heckman, J. (1997), Randomization as an Instrumental Variables Estimator: A Study of Implicit Behavioral Assumptions in One Widely-used Estimator. Journal of Human Resources, 32, 442-462.

Heckman, J. and Hotz, J. (1989), Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training. Journal of the American Statistical Association, 84 (408), 862-880.

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

Huynh, Kim P. and Voia, Marcel-C. (2008), Finite Mixtures and Stochastic Dominance: Estimation and Inference, working paper.

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.

Meyer, B.K., Viscusi and Durbin, D. (1995), Workers Compensation and Injury Duration: Evidence from a Natural Experiment. American Economic Review, 85, 322-340.

McFadden, D. (1989), Testing for stochastic dominance. In: Studies in the Economics of Uncertainty (eds. T.B. Fomby and T.K. Seo). Springer-Verlag, New York.

Schmid, F. and Trede, M. (1996a), Testing for first order stochastic dominance in either direction. Comput. Statist. 11, 165-173.

Schmid, F. and Trede, M. (1998), A Kolmogorov-type test for second-order stochastic dominance. Statist. Probab. Lett. 37, 183-193.

Shaked, M. and Shanthikumar, J.G. (1994), Stochastic Orders and their Applications. Academic Press, Boston, MA, 1994.