TECHNICAL WORKING PAPER SERIES

BOOTSTRAP TEST FOR THE EFFECT OF A TREATMENT ON THE DISTRIBUTION OF AN OUTCOME VARIABLE

Alberto Abadie

Technical Working Paper 261
http://www.nber.org/papers/T0261

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2000

I would like to thank Joshua Angrist for comments and for providing me with the data and Jinyong Hahn, Kei Hirano, Guido Imbens, Guido Kuersteiner, Whitney Newey, Emmanuel Saez, Jim Stock and seminar participants at the 2000 Econometric Society North American Winter Meeting for comments and discussions. The views expressed herein are those of the author and not necessarily those of the National Bureau of Economic Research.

Bootstrap Tests for the Effect of a Treatment on the Distribution of
an Outcome Variable
Alberto Abadie
NBER Technical Working Paper No. 261
September 2000
JEL No. C14, C21, D31, J31

ABSTRACT

This paper considers the problem of assessing the distributional consequences of a treatment on some outcome variable of interest when treatment intake is (possibly) non-randomized but there is a binary instrument available for the researcher. Such scenario is common in observational studies and in randomized experiments with imperfect compliance. One possible approach to this problem is to compare the counterfactual cumulative distribution functions of the outcome with and without the treatment. Here, it is shown how to estimate these distributions using instrumental variable methods and a simple bootstrap procedure is proposed to test distributional hypotheses, such as equality of distributions, first-order stochastic dominance and second order stochastic dominance. These tests and estimators are applied to the study of the effects of veteran status on the distribution of civilian earnings. The results show a negative effect of military service in Vietnam that appears to be concentrated on the lower tail of the distribution of earnings. First order stochastic dominance cannot be rejected by the data.

Alberto Abadie
John F. Kennedy School of Government
Harvard University
79 John F. Kennedy Street
Cambridge, MA 02138
and NBER
alberto_abadie@harvard.edu

1. Introduction

Although most empirical research on treatment effects focuses on the estimation of differences in mean outcomes, analysts have long been interested in methods for estimating the impact of a treatment on the entire distribution of outcomes. This is especially true in economics, where social welfare comparisons may require integration of utility functions under alternative distributions of income. Following Atkinson (1970), consider the class of symmetric utilitarian social welfare functions:

$$W(P,u)=\int u(y)\; dP(y),$$

where P is an income distribution and $u:\mathbb{R}\mapsto\mathbb{R}$. Let $P_{(1)}$ and $P_{(0)}$ denote the (potential) distributions that income would follow if the population were exposed to the treatment in one case, and excluded from the treatment in the other case. For a given $u=\bar{u}$, we rank $P_{(1)}$ and $P_{(0)}$, by comparing $W(P_{(1)},\bar{u})$ and $W(P_{(0)},\bar{u})$.

Typically, u is not fixed by the analyst but is restricted to belong to some particular classes of functions. Then, stochastic dominance can be used to establish a partial ordering on the distributions of income. If two income distributions can be ranked by first order stochastic dominance, these distributions will be ranked in the same way by any monotonic utilitarian social welfare function ($u'>0$). If two income distributions can be ranked by second order stochastic dominance, these distributions will be ranked in the same way by any concave monotonic utilitarian social welfare function ($u'>0$, $u''<0$) (see Foster and Shorrocks (1988) for details). Therefore, stochastic dominance can be used evaluate the distributional consequences of treatments under mild assumptions about social preferences. Another possible question is whether the treatment has any effect on the distribution of the outcome, that is, whether or not the two distributions $P_{(1)}$ and $P_{(0)}$ are the same.

In general, the assessment of the distributional consequences of treatments may be carried on by estimating $P_{(1)}$ and $P_{(0)}$. Estimation of the potential income distributions, $P_{(1)}$ and $P_{(0)}$, is straightforward when the treatment is randomly assigned in the population. However, this type of analysis becomes difficult in observational studies (or in randomized experiments with imperfect compliance) when treatment intake is not randomly determined. Recently, Imbens and Rubin (1997) have shown that, when there is a binary instrumental variable available for the researcher, the potential distributions of the outcome variable are identified for the subpopulation potentially affected in their treatment status by variation in the instrument (the so-called *compliers*). However, this last feature has never been used to compare the entire potential outcome distributions under different treatments in a statistically rigorous way, that is, by performing hypotheses testing. This paper proposes a bootstrap strategy to perform this kind of comparisons. In particular, equality in distributions, first order stochastic dominance and second order stochastic dominance hypotheses, all important for social welfare comparisons, are considered.

The proposed method is applied to the study of the effects of Vietnam veteran status on the distribution of civilian earnings. Following Angrist (1990), random variation in enrollment induced by the Vietnam era draft lottery is used to identify the effects of veteran status on civilian earnings. However, the focus of the present paper is not restricted to the average treatment effect for compliers. The entire marginal distributions of potential earnings for veterans and non-veterans are described for this subgroup of the population. These distributions differ in a notable way from the corresponding distributions of realized earnings. Veteran status appears to reduce lower quantiles of the earnings distribution, leaving higher quantiles unaffected. Although the data show a fair amount of evidence against equality in potential income distributions for veterans and non-veterans, statistical testing falls short of rejecting this hypothesis at conventional significance levels. First and second order stochastic dominance of the potential income distribution for non-veterans are not rejected by the data.

The rest of the paper is structured as follows. In section 2, I briefly review a framework for identification of treatment effects in instrumental variable models and show how to estimate the distributions of potential outcomes for compliers. In contrast with Imbens and Rubin (1997) who report histogram estimates of these distributions, here a simple method is shown to estimate the cumulative distribution functions (cdf) of the same variables.

The estimation of cdfs has some advantages over the histogram estimates. First, there is no need for making an arbitrary choice of width for the bins of the histogram. The cdf, estimated by instrumental variable methods, can be evaluated at each observation in our sample, just as for the conventional empirical distribution function. In addition, this strategy allows us to implement nonparametric tests based directly on differences in the cdfs (see Darling (1957) for a review of this class of tests). Often, it is easier to define and test some distributional hypotheses of interest in economics, such as first or second order stochastic dominance, using cdfs rather than histograms (see Anderson (1996) for an approach based on histograms). Finally, a complete description of the bootstrap strategy is provided along with a proposition which states the asymptotic validity of the bootstrap for the tests proposed in this paper. Section 3 describes the data and presents the empirical results. Section 4 concludes.

2. Statistical Methods

Let $Y_i(0)$ be the potential outcome for individual i without treatment, and $Y_i(1)$ the potential outcome for the same individual under treatment. Define D_i to be the treatment participation indicator (that is, D_i equals one when individual i has been exposed to the treatment, D_i equals zero otherwise.) In practice, the analyst does not observe both $Y_i(0)$ and $Y_i(1)$ for any individual i, since one of these outcomes is counterfactual. Instead, the realized outcome, $Y_i = Y_i(1) \cdot D_i + Y_i(0) \cdot (1 - D_i)$, is observed. Let Z_i be a binary variable that is independent of the responses $Y_i(0)$ and $Y_i(1)$ but that is correlated with D_i in the population (an *instrument*). Denote $D_i(0)$ the value that D_i would have taken if $Z_i = 0$; $D_i(1)$ has the same meaning for $Z_i = 1$. Again, for any particular individual the analyst does not observe both potential treatment indicators $D_i(0)$ and $D_i(1)$; instead the realized treatment $D_i = D_i(1) \cdot Z_i + D_i(0) \cdot (1 - Z_i)$ is observed. In the analysis of randomized experiments with imperfect compliance, Z_i usually represents treatment assignment (randomized) while D_i represents treatment intake (non-randomized). In observational studies instruments are often provided by the so-called "natural experiments" or "quasi-experiments".

For rest of the paper I will use the following identifying assumption:

Assumption 2.1:

(i) Independence of the Instrument : $(Y_i(0), Y_i(1), D_i(0), D_i(1))$ *is independent of* Z_i.

(ii) First Stage : $0 < P(Z_i = 1) < 1$ *and* $P(D_i(1) = 1) > P(D_i(0) = 1)$.

(iii) Monotonicity : $P(D_i(1) \geq D_i(0)) = 1$.

Assumption 2.1 contains a set of nonparametric restrictions under which instrumental variable models identify the causal effect of the treatment for the subpopulation potentially affected in their treatment status by variation in the instrument: $D_i(1) = 1$ and $D_i(0) = 0$ (see Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996)). This subpopulation is sometimes called *compliers*. When the treatment intake, D_i, is itself randomized, Assumption 2.1 holds for $Z_i = D_i$ and every individual is a complier.

Notice that there are some important exclusion restrictions implicit in the notation. First, for each individual i, potential outcomes do not depend on other individuals' treatment intakes. This restriction is called Stable-Unit-Treatment-Value-Assumption (SUTVA) and is frequently used in statistical models of causal inference (see Rubin (1990)). In addition, potential outcomes do not depend on Z_i. This last restriction, commonly invoked in instrumental variable models, allows us to attribute correlation between the instrument and the outcome variables to the effect of the treatment alone.

In this paper, I study distributional effects of possibly non-randomized treatments by comparing the distributions of potential outcomes $Y_i(1)$ and $Y_i(0)$ with and without the treatment. The first step is to show that the identification conditions in Assumption 2.1 allow us to estimate these distributions for the subpopulation of compliers. To estimate the cdfs of potential outcomes for compliers, the following lemma will be useful.

Lemma 2.1: *Let* $h(.)$ *be a measurable function on the real line such that* $E|h(Y_i)| < \infty$. *If Assumption 2.1 holds, then*

$$\frac{E[h(Y_i)D_i|Z_i = 1] - E[h(Y_i)D_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = E[h(Y_i(1))|D_i(0) = 0, D_i(1) = 1] \qquad (1)$$

and,

$$\frac{E[h(Y_i)(1-D_i)|Z_i=1]-E[h(Y_i)(1-D_i)|Z_i=0]}{E[(1-D_i)|Z_i=1]-E[(1-D_i)|Z_i=0]} = E[h(Y_i(0))|D_i(0)=0, D_i(1)=1]. \tag{2}$$

PROOF: By Lemma 4.2 in Dawid (1979), we have that $(h(Y_i(0))\cdot D_i(0), h(Y_i(1))\cdot D_i(1), D_i(0), D_i(1))$ is independent of Z_i. Then by Theorem 1 in Imbens and Angrist (1994), we have that

$$E[h(Y_i(1))\cdot D_i(1) - h(Y_i(0))\cdot D_i(0)|D_i(0)=0, D_i(1)=1] = \frac{E[h(Y_i)\cdot D_i|Z_i=1]-E[h(Y_i)\cdot D_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}.$$

Finally, notice that $E[h(Y_i(1))\cdot D_i(1) - h(Y_i(0))\cdot D_i(0)|D_i(0)=0, D_i(1)=1] = E[h(Y_i(1))| D_i(0)=0, D_i(1)=1]$, which proves the first part of the lemma. The second part of the lemma follows from an analogous argument. □

Lemma 2.1 provides us with a simple way to estimate the cumulative distribution functions of the potential outcomes for compliers. Define $F_{C_1}(y) = E[1\{Y_i(1)\leq y\}|D_i(1)=1, D_i(0)=0]$ and $F_{C_0}(y) = E[1\{Y_i(0)\leq y\}|D_i(1)=1, D_i(0)=0]$. Apply Lemma 2.1 with $h(Y_i) = 1\{Y_i \leq y\}$ to get

$$F_{C_1}(y) = \frac{E[1\{Y_i\leq y\}D_i|Z_i=1]-E[1\{Y_i\leq y\}D_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} \tag{3}$$

and,

$$F_{C_0}(y) = \frac{E[1\{Y_i\leq y\}(1-D_i)|Z_i=1]-E[1\{Y_i\leq y\}(1-D_i)|Z_i=0]}{E[(1-D_i)|Z_i=1]-E[(1-D_i)|Z_i=0]}. \tag{4}$$

Suppose that we have a random sample, $\{(Y_i, D_i, Z_i)\}_{i=1}^n$, drawn from the studied population. The sample counterparts of equations (3) and (4) can be used to estimate $F_{C_1}(y)$ and $F_{C_0}(y)$ for $y = \{Y_1, .., Y_n\}$. We can compare the distributions of potential outcomes by plotting the estimates of F_{C_1} and F_{C_0}. This comparison tells us how the treatment affects different parts of the distribution of the outcome variable, at least for the subpopulation of compliers.

Researchers often want to formalize this type of comparisons using statistical hypothesis testing. In particular, a researcher may want to compare F_{C_1} and F_{C_0} by testing the hypotheses of equality in distributions, first order stochastic dominance and second order stochastic dominance. For two distributions functions F_A and F_B, the hypotheses of interest can be formulated as follows.

Equality of Distributions:

$$F_A(y) = F_B(y) \qquad \forall y \in \mathbb{R} \tag{H.1}$$

First Order Stochastic Dominance: F_A dominates F_B if

$$F_A(y) \leq F_B(y) \qquad \forall y \in \mathbb{R} \tag{H.2}$$

Second Order Stochastic Dominance: F_A dominates F_B if

$$\int_{-\infty}^{y} F_A(x)\ dx \leq \int_{-\infty}^{y} F_B(x)\ dx \qquad \forall y \in \mathbb{R} \tag{H.3}$$

One possible way to carry on these tests for the distributions of potential outcomes for compliers is to use statistics directly based on the comparison between the estimates for F_{C_1} and F_{C_0}. However, it is easier to test the implications of these hypotheses on the two conditional distributions of the outcome variable given $Z_i = 1$ and $Z_i = 0$. Denote F_1 the cdf of the outcome variable conditional on $Z_i = 1$, and define F_0 in the same way for $Z_i = 0$. That is, $F_1(y) = E[1\{Y_i \leq y\}|Z_i = 1]$ and $F_0(y) = E[1\{Y_i \leq y\}|Z_i = 0]$.

PROPOSITION 2.1: *Under Assumption 2.1, hypotheses (H.1)-(H.3) hold for* $(F_A, F_B) = (F_{C_1}, F_{C_0})$ *if and only if they hold for* $(F_A, F_B) = (F_1, F_0)$.

PROOF: From equations (3) and (4), we have

$$F_{C_1}(y) - F_{C_0}(y) = \frac{E[1\{Y_i \leq y\}|Z_i = 1] - E[1\{Y_i \leq y\}|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}.$$

Therefore $F_{C_1} - F_{C_0} = K \cdot (F_1 - F_0)$ for $K = 1/(E[D_i|Z_i = 1] - E[D_i|Z_i = 0]) < \infty$, and the result of the proposition holds. □

Of course, F_1 and F_0 can be easily estimated by the empirical distribution of Y_i for $Z_i = 1$ and $Z_i = 0$ respectively. Divide $(Y_1, ..., Y_n)$ into two subsamples given by different values for the instrument, $(Y_{1,1},...,Y_{1,n_1})$ are those observations with $Z_i = 1$ and $(Y_{0,1},...,Y_{0,n_0})$ are those with $Z_i = 0$. Consider the empirical distribution functions

$$\mathbb{F}_{1,n_1}(y) = \frac{1}{n_1}\sum_{i=1}^{n_1} 1\{Y_{1,i} \leq y\}, \qquad \mathbb{F}_{0,n_0}(y) = \frac{1}{n_0}\sum_{j=1}^{n_0} 1\{Y_{0,j} \leq y\}.$$

Then, the Kolmogorov-Smirnov statistic provides a natural way to measure the discrepancy in the data from the hypothesis of equality in distributions. A two-sample Kolmogorov-Smirnov statistic can be defined as

$$T_{eq} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y\in\mathbb{R}} \left|\mathbb{F}_{1,n_1}(y) - \mathbb{F}_{0,n_0}(y)\right|. \tag{5}$$

Following McFadden (1989), the Kolmogorov-Smirnov statistic can be modified to tests the hypotheses of first order stochastic dominance (for F_1 dominating F_0)

$$T_{fsd} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y\in\mathbb{R}} \left(\mathbb{F}_{1,n_1}(y) - \mathbb{F}_{0,n_0}(y)\right), \tag{6}$$

and second order stochastic dominance

$$T_{ssd} = \left(\frac{n_1 n_0}{n}\right)^{1/2} \sup_{y\in\mathbb{R}} \int_{-\infty}^{y} \left(\mathbb{F}_{1,n_1}(x) - \mathbb{F}_{0,n_0}(x)\right)\, dx. \tag{7}$$

This kind of nonparametric distance tests have in general good power properties. Unfortunately, the asymptotic distributions of the test statistics under the null hypotheses is, in general, unknown, since it depends on the underlying distribution of the data (see e.g., Romano (1988)). In this paper, I use a bootstrap strategy to overcome such a problem. This strategy is described by the following 4 steps:

STEP 1: In what follows, let T be a generic notation for T_{eq}, T_{fsd} or T_{ssd}. Compute the statistic T for the original samples $(Y_{1,1},...,Y_{1,n_1})$ and $(Y_{0,1},...,Y_{0,n_0})$.

STEP 2: Resample n observations $(Y_1^*,...,Y_n^*)$ from $(Y_1,...,Y_n)$ with replacement. Divide $(Y_1^*,...,Y_n^*)$ into two samples: $(Y_{1,1}^*,...,Y_{1,n_1}^*)$ given by the n_1 first elements of $(Y_1^*,...,Y_n^*)$, and $(Y_{0,1}^*,...,Y_{0,n_0}^*)$ given by the n_0 last elements of $(Y_1^*,...,Y_n^*)$. Use these two generated samples to compute the test statistic $T_{(b)}^*$.

Step 3: Repeat Step 2 B times.

Step 4: Calculate the p-values of the tests with p-value $= \frac{1}{B}\sum_{b=1}^{B} 1\{T^*_{(b)} > T\}$. Reject the null hypotheses if p-value is smaller than some significance level α.

By resampling from the pooled data set $(Y_1, ..., Y_n)$ we approximate the distribution of our test statistics when $F_1 = F_0$. Note that for (H.2) and (H.3), $F_1 = F_0$ represents the least favorable case for the null hypotheses. This strategy allows us to estimate the supremum of the probability of rejection under the composite null hypotheses, which is the conventional definition of test size. Justification of the asymptotic validity of this procedure is provided by the following proposition.

Proposition 2.2: *The procedure described in Steps 1 to 4, for T equal to the test statistics in equations (5)-(7) and hypotheses (H.1)-(H.3), (i) provides correct asymptotic level, (ii) is consistent against any fixed alternative, (iii) has power (greater or equal to size) against contiguous alternatives.*

This proposition is proven in Appendix A. The results of a simulation study to assess the small sample performance of the tests proposed in this paper are reported in Appendix B. This simulation study suggests that the bootstrap distribution of the tests provides a good approximation to the nominal level even in fairly small samples.

The idea of using resampling techniques to obtain critical values for Kolmogorov-Smirnov type statistics is probably due to Bickel (1969) and has also be used by Romano (1988), McFadden (1989), Klecan *et al.* (1991), Præstgaard (1995) and Andrews (1997) among others. A related approach based on simulation of p-values can be found in Barrett and Donald (1999).

Note that Proposition 2.2 naturally applies to tests based on perfectly randomized experiments (in which $Z_i = D_i$ for all i).

3. Empirical Example

The data used in this study consist of a sample of 11,637 white men, born in 1950-1953, from the March Current Population Surveys of 1979 and 1981 to 1985. Annual labor earnings, weekly wages, Vietnam veteran status and an indicator of draft-eligibility based on the Vietnam draft lottery outcome are provided for each individual in the sample. Additional information about the data can be found in Appendix C.

Figure 1 shows the empirical distribution of realized annual labor earnings (from now on, annual earnings) for veterans and non-veterans. We can observe that the distribution of earnings for veterans has higher low quantiles and lower high quantiles than that for non-veterans. A naive reasoning would lead us to conclude that military service in Vietnam reduced the probability of extreme earnings without a strong effect on average earnings. The difference in means is indeed quite small. On average, veterans earn only $264 less than non-veterans and this difference is not significant at conventional confidence levels. However, this analysis does not take into account the non-random nature of veteran status. Veteran status was not assigned randomly in the population. The selection process in the military service was influenced by variables associated to the potential earnings (like educational attainment). Therefore, we cannot draw causal inferences by simply comparing the distributions of realized earnings between veterans and non-veterans.

If draft eligibility is a valid instrument, the marginal distributions of potential outcomes for compliers are consistently estimated by using equations (3) and (4). Figure 2 is the result of applying our data to those equations. Note that, in finite samples, the instrumental variables estimates of the potential cdfs for compliers may not be increasingly monotonic functions (see Imbens and Rubin (1997) for a related discussion). The most remarkable feature of Figure 2 is the change in the estimated distributional effect of veteran status on earnings with respect to the naive analysis. The average effect of military service for compliers can be easily estimated using the techniques in Imbens and Angrist (1994). On average, veteran status is estimated to have a negative impact of $1,278 on earnings for compliers, although this effect is far from being statistically different from zero. Now, veteran status seems to reduce low quantiles of the income distribution, leaving high quantiles unaffected. If this characterization is true, the potential outcome for non-veterans would dominate that for veterans in the first order stochastic sense. The hypothesis of equality in distributions seems less likely.

Following the strategy described in section 2, hypotheses testing is performed. Table I reports p-values for the tests of equality in distributions, first order and second order stochastic dominance. Notice that, for this example, the stochastic dominance tests are for earnings for non-veterans dominating earnings for veterans, so the signs of the statistics T_{fsd} and T_{ssd} are reversed. The first row in Table I contains the results for annual earnings as the outcome variable. In the second row the analysis is repeated for weekly wages. Bootstrap resampling was performed 2,000 times ($B = 2,000$).

Consider first the results for annual earnings. The Kolmogorov-Smirnov statistic for equality in distributions is revealed to take an unlikely high value under the null hypothesis. However, we cannot reject equality in distributions at conventional confidence levels. The lack of evidence against the null hypothesis increases as we go from equality in distributions to first order stochastic dominance, and from first order stochastic dominance to second order stochastic dominance. The results for weekly wages are slightly different. For weekly wages we fall far from rejecting equality in distributions at conventional confidence levels.

This example illustrates how useful can be to think in terms of distributional effects, and not merely average effects, when formulating the null hypotheses to test. Once we consider distributional effects, the belief that military service in Vietnam has a negative effect on civilian earnings can naturally be incorporated in the null hypothesis by first or second order stochastic dominance.

4. Conclusions

When treatment intake is not randomized, instrumental variable models allow us to identify the effects of treatments on some outcome variable, for the group of the population affected in the treatment status by variation in the instrument. For such a group of the population, called *compliers*, the entire marginal distribution of the outcome under different treatments can be estimated. In this paper, a strategy to test for distributional effects of treatments within the population of compliers has been proposed. In particular, I focused on the equality in distributions, first order stochastic dominance and second order stochastic dominance hypotheses. First, it is explained a way to estimate the distributions of potential outcomes. Then, bootstrap resampling is used to approximate the null distribution of our test statistics.

This method is illustrated with an application to the study of the effects of veteran status on civilian earnings. Following Angrist (1990), variation in veteran status induced by randomly assigned draft eligibility is used to identify the effects of interest. Estimates of cumulative distribution functions of potential outcomes for compliers show an adverse effect of military experience on the lower tail of the distribution of annual earnings. However, equality in distributions cannot be rejected at conventional confidence levels. First and second order stochastic dominance are not rejected by the data. Results are more favorable to equality in distributions when we use weekly wages as the outcome variable.

Appendix A: Asymptotic Validity of the Bootstrap

Proof of Proposition 2.2: Part (i) can be proven by extending the argument in van der Vaart and Wellner (1996) chapter 3.7 to tests for first and second order stochastic dominance. Let P_1, P_0 be the probability laws of Y conditional on $Z = 1$ and $Z = 0$ respectively. Let Q be the probability law of Z which is Bernoulli with parameter π. Define the empirical measures

$$\mathbb{P}_{1,n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\delta_{Y_{1,i}} \qquad \mathbb{P}_{0,n_0} = \frac{1}{n_0}\sum_{j=1}^{n_0}\delta_{Y_{0,j}},$$

where δ_Y indicates a probability mass point at Y. Let $\mathcal{F} = \{1\{(-\infty, y]\} : y \in \mathbb{R}\}$, that is, the class of indicators of all lower half lines in $\mathbb{R}$. Since $\mathcal{F}$ is known to be universally Donsker, for $n_0, n_1 \to \infty$ we have

$$G_{1,n_1} = n_1^{1/2}(\mathbb{P}_{1,n_1} - P_1) \Rightarrow G_{P_1} \qquad G_{0,n_0} = n_0^{1/2}(\mathbb{P}_{0,n_0} - P_0) \Rightarrow G_{P_0}$$

in $l^\infty(\mathcal{F})$, where "$\Rightarrow$" denotes weak convergence, $l^\infty(\mathcal{F})$ is the set of all uniformly bounded real functions on $\mathcal{F}$ and G_P is a P-Brownian bridge. Let

$$D_n = \left(\frac{n_1 n_0}{n}\right)^{1/2}(\mathbb{P}_{1,n_1} - \mathbb{P}_{0,n_0}),$$

where $n = n_0 + n_1$. If $n \to \infty$, $\pi_n = n_1/n \to \pi \in (0,1)$ almost surely. Then, if $P_1 = P_0 = P$, $D_n \Rightarrow (1-\pi)^{1/2} \cdot G_P - \pi^{1/2} \cdot G'_P$, where G_P and G'_P are independent versions of a P-Brownian bridge. Since $(1-\pi)^{1/2} \cdot G_P - \pi^{1/2} \cdot G'_P$ is also a P-Brownian bridge, we have that $D_n \Rightarrow G_P$ (see also Dudley (1998), Theorem 11.1.1).

For $t \in \mathbb{R}$, let $h(t) = 1\{(-\infty, t]\} \in \mathcal{F}$ and λ the Lebesgue measure on $\mathbb{R}$. For any $z \in l^\infty(\mathcal{F})$, define the following maps: $T_{eq}(z) = \sup_{f \in \mathcal{F}} |z(f)|$, $T_{fsd}(z) = \sup_{f \in \mathcal{F}} z(f)$ and $T_{ssd}(z) = \sup_{f \in \mathcal{F}} \int_{\{g \in \mathcal{F}: g \le f\}} z(g)\, d\mu(g)$ where $\mu = \lambda \circ h^{-1}$. Our test statistics are $T_{eq}(D_n)$, $T_{fsd}(D_n)$ and $T_{ssd}(D_n)$. As before, let T be a generic notation for T_{eq}, T_{fsd} or T_{ssd}. Notice that, for $z_n, z \in l^\infty(\mathcal{F})$, $T(z_n) \le T(z) + T(z_n - z)$. Since T_{eq} is equal to the norm in $l^\infty(\mathcal{F})$, trivially T_{eq} is continuous. T_{fsd} is also continuous because $T_{fsd}(z_n - z) \le T_{eq}(z_n - z)$. Finally, if we restrict ourselves to functions $z_n, z \in C(u, l) = \{x(f) \in l^\infty(\mathcal{F}) : x(h(t)) = 0 \text{ for } t \in (-\infty, l) \cup (u, \infty)\}$, then it is easy to see that, for some finite K, $T_{ssd}(z_n - z) \le K \cdot T_{fsd}(z_n - z)$, so T_{ssd} is continuous. This restriction is innocuous if P_1 and P_0 have bounded support. For the stochastic dominance tests we will use the least favorable case ($P_1 = P_0$) to derive the null asymptotic distribution. Under the least favorable null hypotheses, by continuity, the tests statistics converge in distribution to $T_{eq}(G_P)$, $T_{fsd}(G_P)$ and $T_{ssd}(G_P)$ respectively. Note that, in general, the asymptotic distribution of our test statistics under the least favorable null hypotheses depends on the underlying probability P. It can be easily seen that our test statistics tend to infinity under any fixed alternative.

Consider a test that rejects the null hypothesis if $T(D_n) > c_n$. This test has asymptotic level α if $\liminf c_n \ge c_P(\alpha) = \inf\{c : P(T(G_P) > c) \le \alpha\}$.

Since $c_P(\alpha)$ depends on P, the sequence $\{c_n\}$ is determined by a resampling method. Consider the pooled sample $(Y_1, ..., Y_n) = (Y_{1,1}, ..., Y_{1,n_1}, Y_{0,1}, ..., Y_{0,n_0})$, and define the pooled empirical measure

$$\mathbb{H}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i},$$

then $\mathbb{P}_{1,n_1} - \mathbb{H}_n = (1-\pi_n)(\mathbb{P}_{1,n_1} - \mathbb{P}_{0,n_0})$. Let $(Y_1^*, ... Y_n^*)$ be a random sample from the pooled empirical measure. Define the bootstrap empirical measures:

$$\widehat{\mathbb{P}}_{1,n_1} = \frac{1}{n_1}\sum_{i=1}^{n_1}\delta_{Y_i^*} \qquad \widehat{\mathbb{P}}_{0,n_0} = \frac{1}{n_0}\sum_{j=n_1+1}^{n}\delta_{Y_j^*}.$$

By Theorem 3.7.7 in van der Vaart and Wellner (1996), if $n \to \infty$, then $n_1^{1/2}(\widehat{\mathbb{P}}_{1,n_1} - \mathbb{H}_n) \Rightarrow G_H$ given almost every sequence $(Y_{1,1}, ..., Y_{1,n_1})$, $(Y_{0,1}, ..., Y_{0,n_0})$, where $H = \pi \cdot P_1 + (1-\pi) \cdot P_0$. The same result holds for $n_0^{1/2}(\widehat{\mathbb{P}}_{0,n_0} - \mathbb{H}_n)$. Let

$$\widehat{D}_n = \left(\frac{n_1 n_0}{n}\right)^{1/2} (\widehat{\mathbb{P}}_{1,n_1} - \widehat{\mathbb{P}}_{0,n_0}).$$

Note that $T(\widehat{D}_n) = T((1-\pi_n)^{1/2} n_1^{1/2}(\widehat{\mathbb{P}}_{1,n_1} - \mathbb{H}_n) - \pi_n^{1/2} n_0^{1/2}(\widehat{\mathbb{P}}_{0,n_0} - \mathbb{H}_n))$. Therefore, $T(\widehat{D}_n)$ converges in distribution to $T((1-\pi)^{1/2} G_H - \pi^{1/2} G'_H)$, where G_H and G'_H are independent H-Brownian bridges. Since $(1-\pi)^{1/2} G_H - \pi^{1/2} G'_H$ is also a H-Brownian bridge, we have that, if $P_1 = P_0 = P$, then for

$$c_n = \inf\{c : P(T(\widehat{D}_n) > c) \leq \alpha\},$$

$\liminf c_n \geq c_P(\alpha) = \inf\{c : P(T(G_P) > c) \leq \alpha\}$ almost surely. The reason is that we can always find continuity points of the null distribution of $T(G_P)$ arbitrarily close to $c_P(\alpha)$ but smaller than $c_P(\alpha)$ (otherwise the set of discontinuity points would be uncountable which is not possible by Theorem 4.30 in Rudin (1976)). By tightness of the limiting process, c_n is bounded in probability and the tests are consistent against any fixed alternative. This proves (i) and (ii).

To prove (iii), consider sequences $\{P_{0,n}\}$, $\{P_{1,n}\}$ of probability measures approaching a common limit P in the following sense:

$$\int \left[n^{1/2}(dP_{z,n}^{1/2} - dP^{1/2}) - \frac{1}{2} x_z \, dP^{1/2} \right]^2 \to 0 \qquad \text{for } z = 0, 1, \tag{A.1}$$

where x_1, x_0 are measurable real functions. It can be shown (van der Vaart and Wellner (1996), Lemma 3.10.11) that the sequences of product measures $\{P_{z,n}^n\}$ and $\{P^n\}$ are contiguous, $Px_z = 0$ and $Px_z^2 < \infty$ for $z = 0, 1$ and

$$A_{z,n} = \sum_{i=1}^{n} \log \frac{dP_{z,n}}{dP}(Y_i) = -\frac{1}{2} P x_z^2 + \frac{1}{n^{1/2}} \sum_{i=1}^{n} x_z(Y_i) + o_p(1) \qquad \text{for } z = 0, 1$$

under $\{P\}$. In addition, $\sup_{f \in \mathcal{F}} |n^{1/2}(P_{z,n} - P)f - Px_z f| \to 0$ for $z = 0, 1$.

To assess the asymptotic power of our tests in this scenario, we first need to study the asymptotic behavior of $n_z^{1/2}(\mathbb{P}_{z,n_z}(f) - P(f))$ under sequences of local alternatives which follow (A.1). Note that under the constant sequence $\{P, Q\}$, we have

$$B_{1,n}(f) = \frac{n_1}{n^{1/2}} (\mathbb{P}_{1,n_1}(f) - P(f)) = \frac{1}{n^{1/2}} \sum_{i=1}^{n} Z_i \cdot (f(Y_i) - Pf),$$

and therefore,

$$\begin{pmatrix} A_{1,n} \\ B_{1,n}(f) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} -\frac{1}{2} P x_1^2 \\ 0 \end{pmatrix}, \begin{pmatrix} P x_1^2 & \pi P x_1 f \\ \pi P x_1 f & \pi P (f - Pf)^2 \end{pmatrix} \right).$$

Applying Le Cam's third lemma, we obtain $B_{1,n}(f) \xrightarrow{d} N(\pi P x_1 f, \pi P(f - Pf)^2)$ under the sequence $\{P_{1,n}, Q\}$. Since $n_1/n \to \pi$ almost surely, then $n_1^{1/2}(\mathbb{P}_{1,n_1} - P)(f) \xrightarrow{d} N(\pi^{1/2} P x_1 f, P(f - Pf)^2)$. Using the Donsker property of $\mathcal{F}$, we obtain the uniform version of last result (see van der Vaart and Wellner (1996), Theorem 3.10.12.) An analogous result holds for $z = 0$. Therefore,

$$D_n \Rightarrow G_P + \pi^{1/2}(1-\pi)^{1/2} \cdot (\nu_1 - \nu_0)$$

where $\nu_z(f) = P x_z f$ (this result corrects an error in Præstgaard (1995)). By contiguity arguments we have that $\widehat{D}_n \Rightarrow G_P$ and therefore $\liminf c_n \geq c_P(\alpha)$ almost surely. Then, using a version of Anderson's lemma for general Banach spaces (see, e.g., van der Vaart and Wellner (1996), Lemma 3.11.4), we obtain the desired result for the test of equality of distributions.

The same result holds for first and second order dominance tests (note that for these tests the sequence of contiguous alternatives should be specified such that $T_{fsd}(\nu_1 - \nu_0) \geq 0$ and $T_{ssd}(\nu_1 - \nu_0) \geq 0$ respectively.)

Appendix B: Small Sample Behavior

To assess the small sample performance of the tests proposed in this paper a Monte Carlo study was conducted. To mimic as closely as possible the actual small sample behavior of these tests in real applications, the data used for the simulation study comes from the empirical example of section 3. In each Monte Carlo iteration, a sample of size n was drawn from the empirical distribution of annual earnings in the data. Each sample was divided into two subsamples following the proportion of draft eligibles / non-eligibles in the original data set. Then, the test statistics in equations (5) to (7) were computed and the bootstrap tests were performed using 2,000 bootstrap iterations. This process was repeated for 4,000 Monte Carlo iteration. Table A.I shows the results of this simulation study for samples sizes equal to 25, 50, 100, 250 and 500 and nominal test levels equal to 0.10, 0.05 and 0.01. Asymptotic standard errors (as the number of Monte Carlo iterations tends to infinity) are reported in the last row of the table. The table shows a highly satisfactory performance of the tests, even in fairly small samples ($n = 25$).

Appendix C: Data Description

The data set was especially prepared for Angrist and Krueger (1995). Both annual earnings and weekly wages are in real terms. Weekly wages are imputed by dividing annual labor earnings by the number of weeks worked. The Vietnam era draft lottery is carefully described in Angrist (1990), where the validity of draft eligibility as an instrument for veteran status is also studied. This lottery was conducted every year between 1970 and 1974 and it used to assign numbers (from 1 to 365) to dates of birth in the cohorts being drafted. Men with lowest numbers were called to serve up to a ceiling determined every year by the Department of Defense. The value of that ceiling varied from 95 to 195 depending on the year. Here, an indicator for lottery numbers lower than 100 is used as an instrument for veteran status. The fact that draft eligibility affected the probability of enrollment along with its random nature makes this variable a good candidate to instrument veteran status.

References

Anderson, G. (1996), "Nonparametric Tests for Stochastic Dominance in Income Distributions," *Econometrica*, vol. 64, 1183-1193.

Andrews, D. W. K. (1997), "A Conditional Kolmogorov Test," *Econometrica*, vol. 65, 1097-1128.

Angrist, J. D. (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *American Economic Review*, vol. 80, 313-336.

Angrist, J. D., G. W. Imbens and D. B. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, vol. 91, 444-472.

Angrist, J. D. and A. B. Krueger (1995), "Split-Sample Instrumental Variables Estimates of the Return to Schooling," *Journal of Business and Economic Statistics*, vol. 13, 225-235.

Atkinson, A. B. (1970), "On the Measurement of Inequality," *Journal of Economic Theory*, vol.2, 244-263.

Barrett G. and S. Donald (1999), "Consistent Tests for Stochastic Dominance," unpublished manuscript.

Bickel P. J. (1969), "A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case," *The Annals of Mathematical Statistics* vol. 40, 1-23.

Darling, D. A. (1957), "The Kolmogorov-Smirnov, Cramer-von Mises Tests," *Annals of Mathematical Statistics*, vol. 28, 823-838.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory," *Journal of the Royal Statistical Society*, vol. 41, 1-31.

Dudley, R. M. (1998), *Uniform Central Limit Theorems.* Unpublished manuscript, MIT.

Foster, J. E. and A. F. Shorrocks (1988), "Poverty Orderings," *Econometrica*, vol. 56, 173-177.

Imbens, G. W. and J. D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, vol. 62, 467-476.

Imbens, G. W. and D. B. Rubin (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economic Studies*, vol. 64, 555-574.

Klecan L., McFadden, R. and D. McFadden (1991), "A Robust Test for Stochastic Dominance," unpublished manuscript. MIT.

McFadden, D. (1989), "Testing for Stochastic Dominance," in *Studies in the Economics of Uncertainty in Honor of Josef Hadar*, ed. by T. B. Fomby and T. K. Seo. New York. Springer-Verlag.

Præstgaard, J. T. (1995), "Permutation and Bootstrap Kolmogorov-Smirnov Tests for the Equality of Two Distributions," *Scandinavian Journal of Statistics*, vol. 22, 305-322.

Romano, J. P. (1988), "A Bootstrap Revival of Some Nonparametric Distance Tests," *Journal of the American Statistical Association*, vol. 83, 698-708.

Rubin, D. B. (1990), "Formal Models of Statistical Inference for Causal Effects," *Journal of Statistical Planning and Inference*, vol. 25, 279-292.

Rudin, W. (1976), *Principles of Mathematical Analysis.* New York: McGraw-Hill.

van der Vaart A. W. and J. A. Wellner, (1996) *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

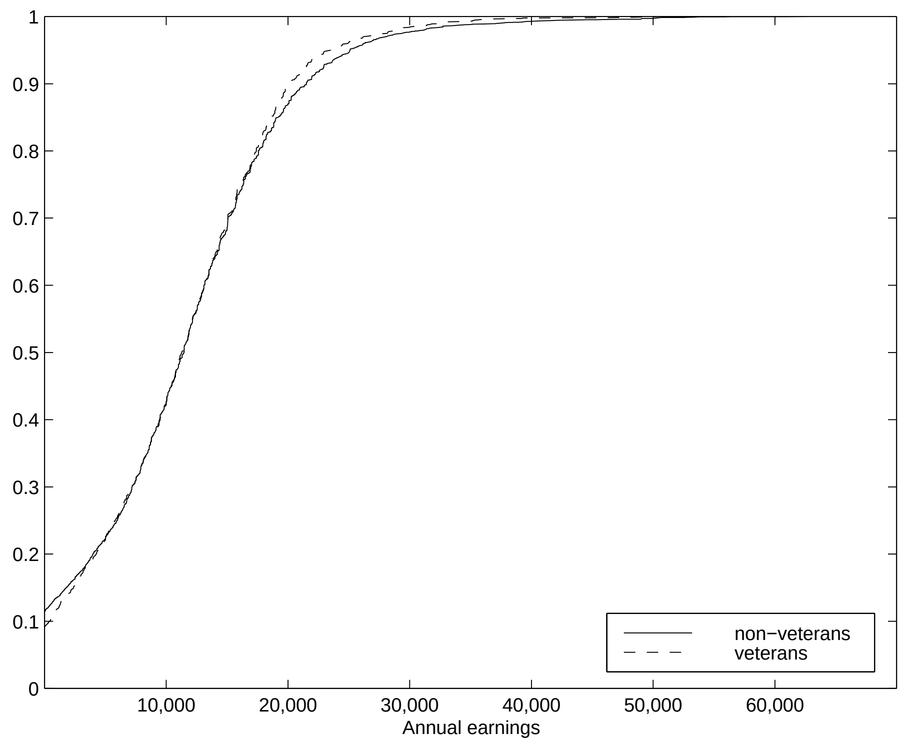

FIGURE 1: Empirical Distributions of Earnings for Veterans and Non-Veterans

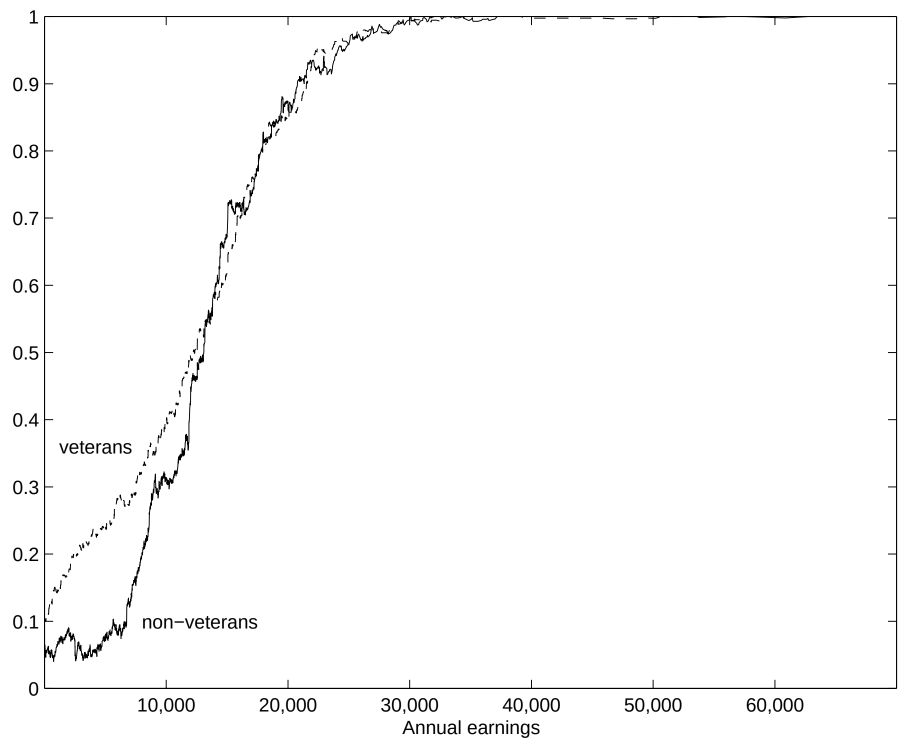

Figure 2: Estimated Distributions of Potential Earnings for Compliers

Table I: Tests on Distributional Effects of Veteran Status on Civilian Earnings, p-values

Outcome variable	Equality in Distributions	First Order Stochastic Dominance	Second Order Stochastic Dominance
Annual Earnings	.1245	.6260	.7415
Weekly Wages	.2330	.6490	.7530

Table A.I: True Test Size in Small Samples, Monte Carlo Simulation

	Sample Size (n)	Nominal Test Level (α) .10	.05	.01
Equality of Distributions	25	.119	.062	.017
	50	.114	.059	.015
	100	.114	.055	.012
	250	.106	.051	.011
	500	.099	.047	.010
First Order Stochastic Dominance	25	.122	.059	.015
	50	.109	.055	.012
	100	.106	.056	.012
	250	.105	.053	.011
	500	.091	.049	.010
Second Order Stochastic Dominance	25	.110	.058	.011
	50	.101	.050	.012
	100	.104	.051	.009
	250	.098	.049	.011
	500	.100	.048	.011
s.e.		.0047	.0034	.0016