CEP 10-02

Stochastic Dominance, Estimation, and Inference for Censored Distributions with Nuisance Parameters

Kim P. Huynh,* Luke Ignaczak, and Marcel-Cristian Voia**

*Indiana University; **Carleton University

January 2010

CARLETON ECONOMIC PAPERS

Department of Economics
1125 Colonel By Drive
Ottawa, Ontario, Canada
K1S 5B6

Stochastic Dominance, Estimation and Inference for Censored Distributions with Nuisance Parameters*

Kim P. Huynh† Luke Ignaczak‡ Marcel C. Voia§

December 18, 2009

Abstract

This note investigates the behavior of stochastic dominance tests of censored distributions which are dependent on nuisance parameters. In particular, we consider finite mixture distributions that are subject to exogenous censoring. To deal with this potential problem, critical values of the proposed tests statistics are calculated using a parametric bootstrap. The tests are then applied to compare differences between distributions of incomplete employment spells with different levels of censoring obtained from Canadian General Social Survey data. The size of the proposed test statistics is computed using fitted GSS data.

JEL codes: C14, C12, C16, C41.
Key words and phrases: Stochastic Dominance Tests, Parametric Bootstrap, Censored Distributions, Finite Mixtures.

*The data used for this study were provided by Statistics Canada. We thank David Jacho-Chávez for his comments and suggestions. We also acknowledge the use of the Quarry High Performance Cluster at Indiana University where all the simulations were performed. All remaining errors are the responsibility of the authors.

†Department of Economics, Indiana University, 105 Wylie Hall, 100 S Woodlawn, Bloomington, IN 47405–7104, USA. Phone: +1 (812) 855 2288. Fax: +1 (812) 855 3736. E-mail: kphuynh@indiana.edu

‡Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, CANADA. Phone: +1 (613) 520-2600 x3773. Fax: (613) 520-3906. Email: lignacza@connect.carleton.ca

§Department of Economics, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, CANADA. Phone: +1 (613) 520-2600 x3546. Fax: (613) 520-3906. Email: mvoia@connect.carleton.ca

1 Introduction

Stochastic dominance (hereafter, SD) tests are used in economic applications related to poverty, inequality and social welfare. Many of these applications compare censored distributions. The standard method used to deal with censored data is to replace the empirical distribution functions by the corresponding Kaplan-Meier estimators. In some particular cases these distributions are nuisance parameter dependent and the Kaplan-Meier estimator is not consistent. This note evaluates the merits of a parametric bootstrap for the proposed SD tests that account for censored distributions with finite mixtures. We investigate the validity of SD tests proposed by Linton, Maasoumi, and Whang (2005) (hereafter, LMW) when they are applied to censored distributions which are nuisance parameter dependent. If the distribution of the variable of interest is censored and/or a function of finite mixtures the null distributions of commonly used SD test statistics are nuisance parameter dependent, even asymptotically, and asymptotic critical values cannot be used. The bootstrap is often suggested to obtain test-specific critical values, as in Barrett and Donald (2003) or Dufour (2006). To deal with the problem of restricted support and the presence of nuisance parameters we employ a parametric bootstrap.

This note is organized as follows: Section 2 presents the methodology while the Monte Carlo exercise is described in Section 3. Finally, Section 4 concludes.

2 Methodology

2.1 Finite mixture decomposition of censored distributions

To estimate censored distributions with nuisance parameters one needs to identify the type of censoring applied to the data. Depending on the type of censoring, different methods can be used. In the case of exogenous censoring the censored data can be ignored in the estimation as the censoring does not affect the data generating process (hereafter, DGP) of the uncensored data. However, in cases of endogenous censoring where the loss of information is due both to the mixture model and censoring, more complex methods are required, see Chauveau (1995) for a description of a stochastic expectations maximization (SEM) algorithm for mixtures with censored data.

In this note, the data is subject to exogenous censoring. The reason of using the simpler censoring problem is for illustrative purposes but its restrictiveness should not limit the results of the parametric bootstrap application which is the focus of this note. To perform this exercise we simulate censored mixtures by fitting a duration outcome variable obtained from the Canadian General Social Survey. In the survey individuals answered questions related to their work history on a particular survey date. In this specific case the censoring applied to the outcome variable will not affect its DGP, as the survey date is an exogenous process. To verify this hypothesis we perform an informal test for the exogeneity of the censoring. In particular, we check whether censoring the observable data affects our parameter estimates. We choose to do the test on data for those who began working in the 1950s because this data is considered to have complete employment spells. A mixture of three distributions were found in the data that was truncated at the censoring point and a mixture of four distributions for the uncensored data. The first three distributions (ordered according to their mean values), estimated from the uncensored data, were not significantly different than the three estimated distributions obtained on the truncated data. This result suggests that our hypothesis about the exogeneity of censoring is valid. As a result, censoring was not explicitly incorporated into the likelihood. Rather we condition on the censoring and the method could be viewed as a partial likelihood approach.

Using these results we proceeded with estimation for only the restricted support of the outcome variable. Univariate mixture models were estimated following McLachlan and Peel (2000). For a random variable $Y^{(G,t)}$ (our outcome variable of interest) which has positive support and is bounded above at y_c, the finite mixture model decomposes a probability density $f(y|y < y_c)$ into the sum of K class parameter-specific probability density functions. The objective is to generate a mixture model with exogenous censoring. In our example we consider a duration outcome variable. The simulated data is generated in two parts. For the censored part of the distribution, we account for the proportion of censored values. For the uncensored part of the distribution, the class probability densities are assumed to come from log-normal mixtures which are often used to fit financial and duration data, and can easily be transformed into normal mixtures. The parameters of such mixtures are estimated by maximum likelihood and are guided by model selection methods. To fit the censored data

without getting values that exceed the censoring point we need to estimate a density that has a bounded support. In our case we consider estimating mixtures of truncated log-normals using the following likelihood function:

$$f(y,\theta|y<y_c) = \frac{\sum_{k=1}^{K} p_k \frac{1}{y\sigma_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\mu_k)^2}{2\sigma_k^2}\right)}{\int_0^{y_c}\sum_{k=1}^{K} p_k \frac{1}{y\sigma_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\mu_k)^2}{2\sigma_k^2}\right)dy} \times \mathbf{1}\{y<y_c\}. \quad (1)$$

The parameters of interest are: $\theta = \{K, p_k, \mu_k, \sigma_k\}$ with $k = 1, ..., K$ and $\sum_{k=1}^{K} p_k = 1$, where p_k is the proportion of a given type with $\sum_{k=1}^{K} p_k = 1$, μ_k is the mean of given type and σ_k is the standard deviation of a given type. All the parameters of interest with the exception of the number of types are estimated by the likelihood. The number of types are estimated using model selection based on the AIC criteria. The following AIC criteria is minimized:

$$AIC_k = -2\log l(\theta|y) + 2d_k, \quad (2)$$

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

2.2 Stochastic Dominance Testing

To test changes in nuisance parameter dependent censored distributions we use an extension of the stochastic dominance tests introduced by LMW. Consider that we observe the outcome variable of interest of different cohorts at different points in time. The outcome variable of interest is $Y^{(C_t)}$, where $Y^{(C_t)}$ is the measure of the outcome variable for cohort C_t, with t the starting period of a given cohort. Define the associated cumulative distribution functions as $F^{(C_t)}$. We are interested in various properties of the conditional distribution functions

$$F^{(C_t)}(y|y \le y_c) = \mathbf{P}\left[Y^{(C_t)} \le y | y \le y_c\right],$$

where y_c is the value of the outcome at the censoring point c. Three possibilities for $F^{(C_t)}$ are considered:

EoD	$H_0^{(1)}$:	$F^{(C_{t_i})}(y\|y \le y_c) \equiv F^{(C_{t_j})}(y\|y \le y_c)$
FOSD	$H_0^{(2)}$:	$F^{(C_{t_i})}(y\|y \le y_c) \le F^{(C_{t_j})}(y\|y \le y_c)$
SOSD	$H_0^{(3)}$:	$\int_0^{y_c}(y-x)dF^{(C_{t_i})}(x) \ge \int_0^{y_c}(y-x)dF^{(C_{t_j})}(x)$

[a] EoD: Equality of Distributions.
[b] FOSD: First Order Stochastic Dominance.
[c] SOSD: Second Order Stochastic Dominance.

The proposed tests are designed to control for the significance level for each of the above tests.

2.2.1 Testing $H_0^{(1)}$ vs $H_1^{(1)}$.

Considering the parameter

$$\kappa = \sup_{y \leq y_c} \left| F^{(C_{t_i})}(y) - F^{(C_{t_j})}(y) \right|,$$

we rewrite the null and the alternative hypotheses under consideration as follows:

$$H_0^{(1)} : \ \kappa = 0 \quad \text{vs} \quad H_1^{(1)} : \quad \kappa > 0. \tag{3}$$

An estimator of κ can be defined by

$$\widehat{\kappa} = \sup_{y \leq y_c} \left| \widehat{F^{(C_{t_i})}}(y) - \widehat{F^{(C_{t_j})}}(y) \right|,$$

where $\widehat{F^{(C_{t_i})}}(y) = \frac{1}{n} \sum_{i=1}^{n} 1\left\{Y^{(C_{t_i})} \leq y\right\}$ and $\widehat{F^{(C_{t_j})}}(y) = \frac{1}{m} \sum_{j=1}^{m} 1\left\{Y^{(C_{t_j})} \leq y\right\}$ are the corresponding empirical distribution functions. The estimator $\widehat{\kappa}$ is consistent. Based on its asymptotic distribution we obtain that

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

is an appropriate statistic for testing the null hypothesis of equality of distributions (EoD) $H_0^{(1)}$ against the alternative $H_1^{(1)}$ of first order stochastic dominance (FOSD). Here n and m and sample sizes for the two distributions. The corresponding rejection (i.e., critical) region is $R : \ \widehat{K} > k_\alpha$ and the acceptance region is $A : \ \widehat{K} \leq k_\alpha$, where k_α is the critical value. Under the presence of censoring and nuisance parameters, the k_α-critical value is not distribution free, and is estimated using a parametric bootstrap (see Section 2.3).

2.2.2 Testing $H_0^{(2)}$ vs $H_1^{(2)}$.

Now, considering the parameter

$$\delta = \sup_{y \leq y_c} \left(F^{(C_{t_i})}(y) - F^{(C_{t_j})}(y) \right),$$

we rewrite the hypotheses $H_0^{(2)}$ and $H_1^{(2)}$ as follows:

$$H_0^{(2)} : \ \delta = 0 \quad \text{vs} \quad H_1^{(2)} : \quad \delta > 0. \tag{4}$$

The empirical estimator of δ is given by

$$\widehat{\delta} = \sup_{y \leq y_c} \left(\widehat{F^{(C_{t_i})}}(y) - \widehat{F^{(C_{t_j})}}(y) \right).$$

The estimator $\widehat{\delta}$ is consistent. Therefore,

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

is an appropriate statistic for testing the null hypothesis $H_0^{(1)}$ against the alternative $H_1^{(1)}$. The corresponding rejection region is $R: \ \widehat{D} > d_\alpha$ and the acceptance region is $A: \ \widehat{D} \leq d_\alpha$. Since the distributions are not, in general, identical, the critical value d_α is not distribution free and has to be estimated. Again, a parametric bootstrap is used.

2.2.3 Testing $H_0^{(3)}$ vs $H_1^{(3)}$.

In this case, the parameter

$$\tau = \sup_{y \leq y_c} \left(G_2^{(C_{t_i})}(y) - G_2^{(C_{t_j})}(y) \right)$$

is strictly positive under the null. Therefore, to test the SOSD hypothesis we have under the null we have

$$H_0^{(3)}: \ \tau = 0 \quad \text{vs} \quad H_1^{(3)}: \quad \tau > 0. \tag{5}$$

that one of the distributions SOSD the other under the alternative.

Define an estimator of τ by

$$\widehat{\tau} = \sup_{y \leq y_c} \left(\widehat{G_2^{(C_{t_i})}}(y) - \widehat{G_2^{(C_{t_j})}}(y) \right),$$

The estimator $\widehat{\tau}$ is consistent and we have that

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

The corresponding rejection region is $R: \ \widehat{T} > \theta_\alpha$ and the acceptance region is $A: \ \widehat{T} \leq \theta_\alpha$, where θ_α is the α-critical value of a distribution that depends on a transformation of $F^{(C_{t_i})}$ and $F^{(C_{t_j})}(x)$. Hence, θ_α is not distribution free and has to be estimated as in the previous cases.

2.3 Parametric Bootstrap

By transforming the parametric bootstrap proposed by Huynh and Voia (2008) to account for censored distributions we simulate the critical values in the following way:

1. Sample $n_{uncensored}$ values from $Y_1^{(C_{t_i})}, \ldots, Y_n^{(C_{t_i})}$ from the estimated distributions obtained using the data:

$$\int_0^{y_c} \widehat{f}_{Duration}(s)ds \;=\; \int_0^{y_c} \frac{\sum_{k=1}^{K} \widehat{p}_k \frac{1}{y\widehat{\sigma}_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\widehat{\mu}_k)^2}{2\widehat{\sigma}_k^2}\right)}{\int_0^{y_c} \sum_{k=1}^{K} \widehat{p}_k \frac{1}{y\widehat{\sigma}_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\widehat{\mu}_k)^2}{2\widehat{\sigma}_k^2}\right) dy} ds.$$

   Note that this DGP will generate bounded log-normal mixtures up to the censoring point y_c.

2. Then sample $n_{censored}$ observations using the proportion of censored data at the value of the censoring point. Define $n = n_{uncensored} + n_{censored}$.

3. Sample from cohort j the estimated distributions $m = m_{uncensored} + m_{censored}$ values from $Y_1^{(C_{t_j})}, \ldots, Y_m^{(C_{t_j})}$.

4. Adjust the distributions to be stochastically equal under the null hypothesis. This is done in one side by pooling the estimated mixtures obtained for the two distributions of interest and estimating the resulting mixture distribution, and in the other side by estimating the mixture distribution obtained from pooling the data of the two distributions.

5. With the use of the resulting empirical distribution functions, $\widehat{F}^{(C_{t_j})^*}(y)$ and $\widehat{F}^{(C_{t_j})^*}(y)$, we define

$$\widehat{K}^* = \sup_{y \le y_c} \sqrt{\frac{nm}{n+m}} \left| \widehat{F}^{(C_{t_i})^*}(y) - \widehat{F}^{(C_{t_j})^*}(y) \right|.$$

6. Repeat steps 1-5 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 $\widehat{D}^*$ are at or below y.

7. The rejection region is $\widehat{K}^* > k_\alpha^*$.

To estimate the critical values for the FOSD test the same steps are followed as in the EoD case, but by constructing the estimator

$$\widehat{D}^* = \sup_{y \leq y_c} \sqrt{\frac{nm}{n+m}} \left(\widehat{F}^{(C_{t_i})^*}(y) - \widehat{F}^{(C_{t_j})^*}(y) \right).$$

The critical value d^*_α is defined as the smallest value of y subject to at least $100(1-\alpha)\%$ of the obtained B values of $\widehat{D}^*$ are at or below y. The rejection region is $\widehat{D}^* > d^*_\alpha$. The critical values of the SOSD test are calculated similarly to the EoD and FOSD cases. First, define $\widehat{G}_2^{(C_{t_i})^*}(y)$ and $\widehat{G}_2^{(C_{t_j})^*}(y)$ to construct

$$\widehat{T}^* = \sup_{y \leq y_c} \sqrt{\frac{nm}{n+m}} \left(\widehat{G_2}^{(C_{t_i})^*}(y) - \widehat{G_2}^{(C_{t_j})^*}(y) \right).$$

The critical value θ^*_α is the smallest value of y s.t. at least $100(1-\alpha)\%$ of the obtained B values of $\widehat{T}^*$ are at or below y. In this case, the rejection region is $\widehat{T}^* > \theta^*_\alpha$.

3 Monte Carlo Design and Simulations

This simulation exercise was designed to evaluate the effect of censoring and finite mixtures on different sample sizes, at $n = \{100, 500, 1000, 3000\}$. The simulated data was generated in such a way to mimic a large sample of data, which allows for different levels of censoring and a different number of mixtures. The tests of stochastic dominance are used to compare the distributions of two different cohorts. The cohort specific distributions were generated so that cohort 1 FOSDs cohort 2 using the same distance between the distributions as in the actual data.[1]

Table 1 presents simulated data using finite mixtures of log-normal distributions that are generated by fitting the duration of employment for males recorded by the Canadian General Social Survey (GSS) over two cohorts. It is assumed that the two cohorts are censored at a duration of 23, and the mixture decomposition fits the uncensored data (Figure 1 presents the true and fitted data for the 1970s cohort, i.e Cohort 2). We perform a Welch two sample t-test on the null hypothesis of "the true difference in means is equal to 0" and obtained a p-value of 0.2948. Furthermore, a two-sample Kolmogorov-Smirnov test with a null of equality of distributions resulted in p-value of 0.1317. Both tests confirm that the simulated

[1] In the data we found that test 2 rejects the null in favor of the FOSD at levels smaller than 0.01

data fit reasonably well as the nulls are not rejected in other tests. Mixtures of two or, at most, three distributions are estimated for the uncensored data. Table 2 presents the shares of ongoing spells (censored data) for the two cohorts. The levels of censoring in the data vary from 26 percent to 82 percent with cohort 1 having more censoring than cohort 2. We show graphically (see Figure 2) that not accounting for the censored data produces totally different distributions and results than when censored data is properly accounted for. If the censored data is ignored the tests employed are biased.

3.1 Results

Table 2 presents the results of the level of the analyzed test statistics. The level exercise uses the data of the two cohorts from both groups to create stochastically equal distributions. We also check the power of the test statistics. The power exercise uses the fact that the two cohorts have stochastically different distributions. In our specific case the power is equal to 1 and is independent of the sample sizes we chose. The results suggest that the power of the tests is driven by the distance between the two analyzed distributions, therefore we do not report the power in our tables as the distance between the two distributions is large. The results from Table 2 suggest that some tests perform better than other in finite samples, but as the sample size increases the test statistics are very close to their desired levels. A ranking of the three test statistics in terms of finite sample level performances suggest that the FOSD test performs best, it is followed by the SOSD test and then by the EoD test.

4 Conclusions

This note presents a simulation exercise that is data driven and that explores the effect of censoring when two distributions drawn from a large sample are compared. Employing a parametric bootstrap that accounts for both finite mixtures and censoring generates critical values for tests statistics developed by LMW that minimize the effect of type I and type II errors.

References

Barrett, G. F., and S. G. Donald (2003): "Consistent Tests for Stochastic Dominance," *Econometrica*, 71(1), 71–104.

Chauveau, D. (1995): "A stochastic EM algorithm for mixtures with censored data," *Journal of Statistical Planning and Inference*, 46(1), 1 – 25.

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.

Linton, O., E. Maasoumi, and Y.-J. Whang (2005): "Consistent Testing for Stochastic Dominance under General Sampling Schemes," *Review of Economic Studies*, 72(3), 735–765.

McLachlan, G., and D. Peel (2000): *Finite Mixture Models*. Wiley, New York.

Table 1: Simulated mixture distributions calibrated to GSS duration data

Cohort 1 - Males			
	Observations	Uncensored	Censored
	3000	549	2451
Type	Share	μ	σ
I	0.68	4.54	4.96
II	0.32	16.79	3.73
Cohort 2 - Males			
	Observations	Uncensored	Censored
	3000	1446	1554
Type	Share	μ	σ
I	0.68	3.9	3.75
II	0.22	13.33	3.36
III	0.10	19.87	1.59

Notes: The finite mixture results are obtained using uncensored data. The parameters of the finite mixtures: number of types (k), mean of type k (μ_k), and standard deviation of type k (σ_k) were obtained by maximizing the log-likelihood associated to the following density function:

$$f(y,\theta|y<y_c) = \frac{\sum_{k=1}^{K} p_k \frac{1}{y\sigma_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\mu_k)^2}{2\sigma_k^2}\right)}{\int_0^{y_c} \sum_{k=1}^{K} p_k \frac{1}{y\sigma_k\sqrt{2\pi}} \exp\left(\frac{-(\log y-\mu_k)^2}{2\sigma_k^2}\right) dy} \times \mathbf{1}\{y<y_c\}. \quad (6)$$

The number of types were chosen using the AIC criteria.

Table 2: Monte Carlo Simulations

	EoD			FOSD			SOSD		
Observations	0.010	0.050	0.100	0.010	0.050	0.100	0.010	0.050	0.100
100	0.014	0.056	0.095	0.015	0.054	0.113	0.013	0.063	0.115
500	0.015	0.046	0.095	0.011	0.051	0.093	0.013	0.058	0.117
1000	0.018	0.053	0.101	0.010	0.048	0.102	0.017	0.047	0.096
3000	0.016	0.048	0.098	0.010	0.051	0.097	0.011	0.049	0.103

Notes: This table presents the Monte Carlo results of the level of the test. The DGP is constructed using the fitted GSS data presented in Table 1. The level exercise simulates the distributions of the two groups under the null.

Figure 1: Empirical Density Function for the 1970s Cohort

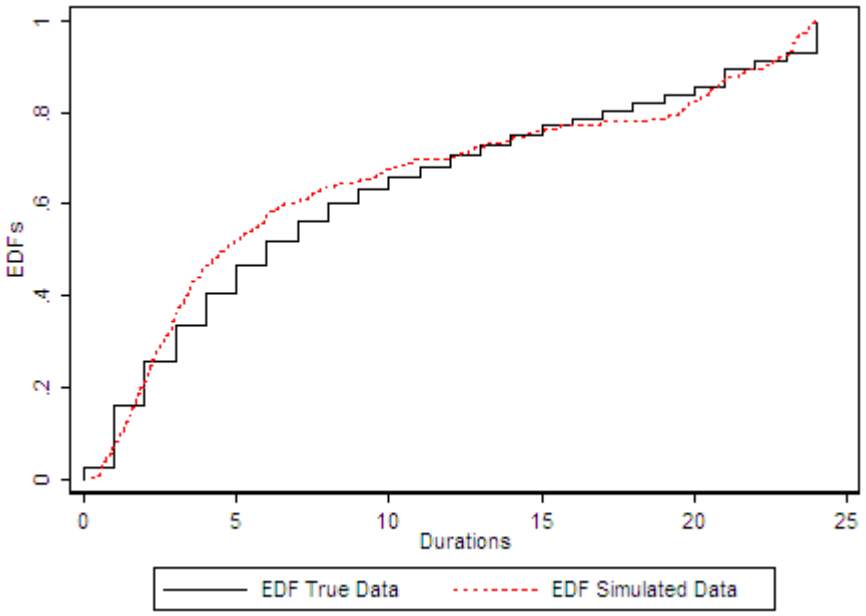

Figure 2: Empirical Density Functions for Uncensored and Censored Data