Nonparametric estimation for classic and interval open-ended data in contingent valuation

Magnus Ekström

Research Report
Centre of Biostochastics

Swedish University of Agricultural Sciences

Report 2010:07
ISSN 1651-8543

Nonparametric estimation for classic and interval open-ended data in contingent valuation

Magnus Ekström[1]

Centre of Biostochastics
Swedish University of Agricultural Sciences, Umeå, Sweden

Abstract

Assume that a valuation survey is conducted for estimating the benefits of a policy change. In order to estimate the willingness to pay (WTP) for the policy change, a classic and interval open-ended question is used. This question permits respondents to choose between two types of answer, either an exact WTP amount or a self-selected interval. Accidently, a rather large proportion of the respondents give exact WTP values as well as intervals. In this set up, we derive a nonparametric maximum likelihood estimator of the distribution of WTP. Monte Carlo simulations are performed to study the behavior of the proposed estimator.

Keywords: Classic and interval open-ended question, contingent valuation, interval-censored data, middle-censored data, nonparametric maximum likelihood estimation, self-censored data, willingness to pay.

[1]E-mail address: Magnus.Ekstrom@sekon.slu.se

1 Introduction

Contingent valuation is a survey-based method for the valuation of non-market goods, such as biodiversity, air quality improvements or environmental preservation. Typically the survey asks people what they would be willing to pay for a hypothetical good. There are different types of questions that can be put forth in order to characterize willingness to pay (WTP). In Håkansson (2008), a new approach to quantitative information elicitation in surveys is introduced. More precisely, she develops a classic and interval open-ended elicitation format that permits respondents to choose between two types of answer, either an exact amount or a self-selected interval, as illustrated in the following example from her article:

> **Example 1**.
> Question: Try to state what you are willing to pay, either as an interval between two amounts or as an exact amount.
> Answer: (fill in one of the options below)
> Option 1:
> I am willing to pay between ..... and ..... this year as a lump sum.
> Option 2:
> I am willing to pay ..... this year as a lump sum.

In Parkkila (2009), a valuation survey is described that was conducted for estimating the benefits of a policy change in Baltic salmon fisheries regulation. The population of interest were anglers in the Torne river, who will benefit from the policy change. In order to estimate the angler's WTP for the new regulation programme, the respondents were asked to answer a question of the same format as in Example 1. Accidently, a rather large proportion of the respondents gave exact WTP values as well as intervals. Thus, the observed data are of the following form: For n_1 respondents, only the exact WTP values are observed, for n_2 respondents, both WTP values and intervals are observed, and for the remaining n_3 respondents, the WTP value of interest is observed only to belong to an interval instead of being exactly known. We denote exact WTP values by X_j and intervals by $\boldsymbol{I}_j = (L_j, U_j]$. After re-ordering the data as necessary, we can assume without loss of generality that we have the

following observed data:

$$\{X_1, ..., X_{n_1}, (X_{n_1+1}, \boldsymbol{I}_{n_1+1}), ..., (X_{n_1+n_2}, \boldsymbol{I}_{n_1+n_2}), \boldsymbol{I}_{n_1+n_2+1}, ..., \boldsymbol{I}_n\},$$

where $n = n_1 + n_2 + n_3$. We assume that the $\{X_i\}$ are independent and identically distributed random variables with unknown distribution F, and that the $\{(L_j, U_j)\}$ are independent and identically distributed random vectors. Based on the observations from all n respondents, the problem is to estimate the distribution function F of the X_j's.

In survival analysis, a value X_j is often said to be interval-censored or middle-censored if the only information we have about X_j is that X_j lies in the interval $\boldsymbol{I}_j$; see, e.g., Peto (1973) and Jammalamadaka & Mangalam (2003). However, standard methods for analyzing interval-censored and middle-censored data assume, implicitly or explicitly, that the censoring intervals $\{\boldsymbol{I}_j\}$ are independent of $\{X_j\}$. In our case, this is not a reasonable assumption.

Recall that a respondent may choose not to state the exact location of the WTP value, but instead choose to place it within an interval. We will refer to this as *self*-censoring, rather than interval-censoring or middle-censoring, to stress that it is the respondent him*self*/her*self* who selects the interval (containing the exact WTP value). (Thus, in this context, self-censoring means something quite different than the act of censoring one's own work out of fear or deference to the sensibilities of others.) If $n_1 + n_2 > 0$ and $n_3 > 0$, then the observed data will be referred to as partly self-censored data, and if $n_1 + n_2 = 0$ and $n_3 > 0$, as self-censored data. In Belyaev & Kriström (2010), nonparametric likelihood estimation for self-censored WTP data is studied, and in the current article we will consider partly self-censored data (the case when $n_1 \geq 0$, $n_2 > 0$, and $n_3 > 0$).

If $n_1 \geq 0$, $n_2 > 0$, and $n_3 > 0$, then for some respondents we observe both WTP values and intervals. Here a natural question arises: do we gain anything at all by knowing the interval that contains the exact WTP value if we already know the value in question? In the current paper we will argue that by using this extra information we can construct a valid estimator of the distribution of the WTP values.

In Section 2, we provide a nonparametric maximum likelihood estimator of F. Simulation results are presented in Section 3. Finally we conclude the paper in Section 4.

2 A maximum likelihood estimator of the distribution of WTP

Henceforth we assume that the vectors $\{(L_i, R_i)\}$ defining the censoring intervals $\{\boldsymbol{I}_j\}$ follow a discrete bivariate distribution with finitely many support points, cf. Huang (1999) and Belyaev & Kriström (2010). This is not unreasonable, because the empirical self-censored data considered by Belyaev & Kriström suggest that respondents tend to state rounded intervals from a finite set.

If we observe a point value $X = x_j$ (but no interval), then the contribution to the likelihood is

$$F(x_j) - F(x_j-).$$

If we observe both a point value $X = x_j$ and an interval $\boldsymbol{I} = \boldsymbol{i}_j$, where $\boldsymbol{i}_j = (l_j, r_j]$, then the contribution to the likelihood is the right-hand side of

$$\begin{aligned} P(\boldsymbol{I} = \boldsymbol{i}_j, X = x_j) &= [F(x_j) - F(x_j-)]P(\boldsymbol{I} = \boldsymbol{i}_j | X = x_j) \\ &\propto [F(x_j) - F(x_j-)]. \end{aligned}$$

For the case when we observe an interval (but no point), we define a partition of the positive real line. That is, we determine a set of values, $0 = s_0 < s_1 < ... < s_m < s_{m+1} = +\infty$, and define $\boldsymbol{v}_k = (s_{k-1}, s_k]$ for $k = 1, ..., m+1$. By the law of total probability,

$$\begin{aligned} P(\boldsymbol{I} = \boldsymbol{i}_j) &= \sum_k \alpha_{jk} P(\boldsymbol{I} = \boldsymbol{i}_j | X \in \boldsymbol{v}_k) P(X \in \boldsymbol{v}_k) \\ &= \sum_k \alpha_{jk} w_{jk} [F(s_k) - F(s_{k-1})], \end{aligned}$$

where $w_{jk} = P(\boldsymbol{I} = \boldsymbol{i}_j | X \in \boldsymbol{v}_k)$, and $\alpha_{jk} = 1$ if $\boldsymbol{v}_k \subseteq \boldsymbol{i}_j$ and 0 otherwise. Thus, if we observe an interval $\boldsymbol{I} = \boldsymbol{i}_j$ (but no point), then the contribution to the likelihood is

$$\sum_k w_{jk} \alpha_{jk} [F(s_k) - F(s_{k-1})]$$

(cf. Belyaev & Kriström, 2010). Based on the above, the likelihood function of the observed data is proportional to

$$L(F, \boldsymbol{w}) = \prod_{j=1}^{n_1+n_2} [F(x_j) - F(x_j-)] \prod_{j=n_1+n_2+1}^{n} \left[\sum_k w_{jk} \alpha_{jk} [F(s_k) - F(s_{k-1})] \right],$$

where $\boldsymbol{w} = \{w_{jk}\}$ and $n = n_1 + n_2 + n_3$. Note that the expression of the likelihood depends on the unknown conditional probabilities $w_{jk} = P(\boldsymbol{I} = \boldsymbol{i}_j | X \in \boldsymbol{v}_k)$. However, by plugging in estimates $\hat{\boldsymbol{w}} = \{\hat{w}_{jk}\}$, e.g. the estimates suggested below, we obtain a function $L(F, \hat{\boldsymbol{w}})$ which may be maximized to obtain an estimate of the distribution function F.

Remark 1. If we determine the set of values, $0 = s_0 < s_1 < ... < s_m < s_{m+1} = +\infty$, such that each L_j and R_j are contained in the set, and if for each $j = n_1 + n_2 + 1, ..., n$ there exists w_j such that $w_j = w_{jk}$ for all k such that $\boldsymbol{v}_k \subseteq \boldsymbol{i}_j$, then

$$\begin{aligned} \sum_k w_{jk}\alpha_{jk}[F(s_k) - F(s_{k-1})] &\propto \sum_k \alpha_{jk}[F(s_k) - F(s_{k-1})] \\ &= F(R_j) - F(L_j) \end{aligned}$$

and the likelihood of the observed data is proportional to

$$L(F) = \prod_{j=1}^{n_1+n_2} [F(x_j) - F(x_j-)] \prod_{j=n_1+n_2+1}^{n} [F(R_j) - F(L_j)]. \tag{1}$$

This is essentially the same likelihood as used by Peto (1973) and Jammalamadaka & Mangalam (2003), and a special case of the likelihood defined by Turnbull (1976). The difference lies in the choice of convention for defining the censoring intervals. Peto and Turnbull assume that the intervals are closed rather than half-open, while Jammalamadaka & Mangalam use open intervals. As mentioned by Lindsey & Ryan (1998), good arguments can be made for and against almost any convention for defining the censoring intervals, and in practice, the choice will have little impact and any reasonable convention can be adopted. Further, note that the assumption that $w_j = w_{jk}$, for all k such that $\boldsymbol{v}_k \subseteq \boldsymbol{i}_j$, essentially assumes that the point value X_j and the interval $\boldsymbol{I}_j$ are independent. This assumption, however, is not reasonable if it is the respondent who chooses the interval.

Remark 2. If $n_1 + n_2 = 0$, $n_3 > 0$, and if we determine the set of values, $0 = s_0 < s_1 < ... < s_m < s_{m+1} = +\infty$, such that each L_j and R_j are contained in the set, then $L(F, \boldsymbol{w})$ reduces to the likelihood considered in Belyaev & Kriström (2010). Under this setting, Belyaev & Kriström propose estimators of the conditional probabilities $\{w_{jk}\}$, and under the assumption

that the conditional probabilities $\{w_{jk}\}$ are known (or can be consistently estimated), they show that the obtained nonparametric maximum likelihood estimator of the distribution F is consistent.

In order to estimate $\boldsymbol{w} = \{w_{jk}\}$, we assume that the distribution of the relative position of the point X in the interval $\boldsymbol{I} = \boldsymbol{i} = (l, r]$ does not depend on the interval $\boldsymbol{i}$. That is,

$$P\left(\frac{X-l}{r-l} \leq x \Big| \boldsymbol{I} = \boldsymbol{i}\right) = H(x) \quad \text{for all } x. \tag{2}$$

By Bayes' theorem we have

$$w_{jk} = P(\boldsymbol{I} = \boldsymbol{i}_j | X \in \boldsymbol{v}_k) = \frac{p_{kj} p_j}{\sum_l \alpha_{lk} p_{kl} p_l},$$

where

$$\begin{aligned} p_{kj} &= P(X \in \boldsymbol{v}_k | \boldsymbol{I} = \boldsymbol{i}_j), \\ p_j &= P(\boldsymbol{I} = \boldsymbol{i}_j), \\ \alpha_{jk} &= 1 \text{ if } \boldsymbol{v}_k \subseteq \boldsymbol{i}_j, \text{ and 0 otherwise.} \end{aligned}$$

We estimate $p_j = P(\boldsymbol{I} = \boldsymbol{i}_j)$ by the proportion $\hat{p}_j$ of observed intervals that equals $\boldsymbol{i}_j$.

If respondent j has given both a point x_j and an interval $i_j = (l_j, r_j]$, then we compute x_j's relative position in the interval, i.e.

$$x'_j = \frac{x_j - l_j}{r_j - l_j}, \qquad j = n_1 + 1, ..., n_1 + n_2,$$

and H is estimated by $\hat{H}$, the empirical distribution of $\{x'_j\}_{j=n_1+1}^{n_1+n_2}$. For each $\boldsymbol{v}_k = (s_{k-1}, s_k] \subseteq \boldsymbol{i}_j$, $j = n_1 + 1, ..., n$, compute the relative positions of the endpoints s_{k-1} and s_k in the interval $\boldsymbol{i}_j = (l_j, r_j]$, i.e.

$$s'_{j,k-1} = \frac{s_{k-1} - l_j}{r_j - l_j} \quad \text{and} \quad s'_{j,k} = \frac{s_k - l_j}{r_j - l_j}.$$

The conditional probability $p_{kj} = P(X \in \boldsymbol{v}_k | \boldsymbol{I} = \boldsymbol{i}_j)$ may now be estimated by

$$\hat{p}_{kj} = \hat{H}_{n_2}(s'_{j,k}) - \hat{H}_{n_2}(s'_{j,k-1}),$$

and by Bayes' theorem we obtain the estimator of w_{jk} as

$$\hat{w}_{jk} = \frac{\hat{p}_{kj}\hat{p}_j}{\sum_l \alpha_{lk}\hat{p}_{kl}\hat{p}_l}.$$

Thus, under assumption (2), an estimator, $\hat{F}$, of the distribution F of X is found by maximizing

$$L(F, \hat{\boldsymbol{w}}) = \prod_{j=1}^{n_1+n_2} [F(X_j) - F(X_j-)] \prod_{j=n_1+n_2+1}^{n} \left[\sum_k \hat{w}_{jk}\alpha_{jk}[F(s_k) - F(s_{k-1})] \right],$$

with respect to F.

It should be noted that assumption (2) is rather restrictive. If necessary, the assumption can be relaxed to

$$P\left(\frac{X-l}{r-l} \leq x \middle| \boldsymbol{I} = \boldsymbol{i} \right) = H(x|\theta) \quad \text{for all } x, \tag{3}$$

where $\theta = \theta(l, r)$ is a real-valued (or even vector-valued) function of the interval limits l and r. Let $\theta_j = \theta(l_j, r_j)$, $j = n_1 + 1, ..., n_1 + n_2$. The conditional distribution function $H(x|\theta)$ may be estimated by

$$\hat{H}(x|\theta) = \frac{n_2^{-1} \sum_{j=n_1+1}^{n_1+n_2} I_{\{x'_j \leq x\}} W_h(\theta_j, \theta)}{\hat{\mu}(\theta)},$$

where

$$\hat{\mu}(\theta) = \frac{1}{n_2} \sum_{j=n_1+1}^{n_1+n_2} W_h(\theta_j, \theta)$$

is a kernel estimator of the marginal density $\mu(\theta)$ of θ,

$$W_h(\theta_j, \theta) = h^{-1} K\left(\frac{\theta_j - \theta}{h} \right), \tag{4}$$

and $K(\cdot)$ is a univariate kernel function (Li & Racine, 2007). Under assumption (3), the estimator of w_{jk} is obtained as

$$\tilde{w}_{jk} = \frac{\tilde{p}_{kj}\hat{p}_j}{\sum_l \alpha_{lk}\tilde{p}_{kl}\hat{p}_l}$$

where

$$\tilde{p}_{kj} = \hat{H}(s'_{j,k}|\theta_j) - \hat{H}(s'_{j,k-1}|\theta_j).$$

Thus, under assumption (3), an estimator, $\tilde{F}$, of the distribution F of X is found by maximizing

$$L(F, \tilde{\boldsymbol{w}}) = \prod_{j=1}^{n_1+n_2} [F(X_j) - F(X_j-)] \prod_{j=n_1+n_2+1}^{n} \left[\sum_k \tilde{w}_{jk} \alpha_{jk} [F(s_k) - F(s_{k-1})] \right],$$

with respect to F.

3 Monte Carlo simulations

Four different simulations have been conducted. In each one of them, we determined the set of values, $0 = s_0 < s_1 < ... < s_m < s_{m+1} = +\infty$, such that each L_j and R_j are contained in the set. A motivation for this choice is that this partition of the positive real line needs to be defined in any case, for finding the maximizer of the likelihood (cf. Peto, 1973, and Turnbull, 1976). Furthermore, our results show that this choice gives adequate estimates of the distribution F, and Belyaev & Kriström (2010) use the same partition in the definition of their nonparametric maximum likelihood estimator for self-censored data.

In each simulation, $L_1, ..., L_n$ are independent and exponentially distributed random variables, with mean 100, rounded downwards to the nearest multiple of 10. The upper boundaries of the intervals are defined as $R_j = L_j + Y_j$, $j = 1, ..., n$, where $Y_1, ..., Y_n$ are independent and exponentially distributed random variables, with mean 40, rounded upwards to the nearest multiple of 10. Thus, each interval $(L_j, R_j]$ is at least 10 units wide. The exact values are generated as $X_j = L_j + Y_j Z_j$, $j = 1, ..., n$, where $Z_1, ..., Z_n$ are independent and beta distributed random variables, with parameters α_j and β_j, $j = 1, ..., n$. The number of respondents is $n = 200$. All the respondents give an interval $(L_j, R_j]$, and 20 % of them give an exact value X_j as well.

In simulation 1, we used $\alpha = 5$ and $\beta = 1$, i.e. the beta density function is strictly convex and increasing on its support. This means that the respondents tend to have the exact values X_j placed near the upper boundaries of the intervals $\boldsymbol{I}_j$. In simulation 2, $\alpha = 1$ and $\beta = 5$, which means that the respondents tend to have the exact values placed near the lower boundaries of the intervals. In simulation 3, we used $\alpha_j = 5 + 10(R_j - L_j)/R_j$ and $\beta = 1$. Thus, this simulation is similar to simulation 1, but in this case the tendency

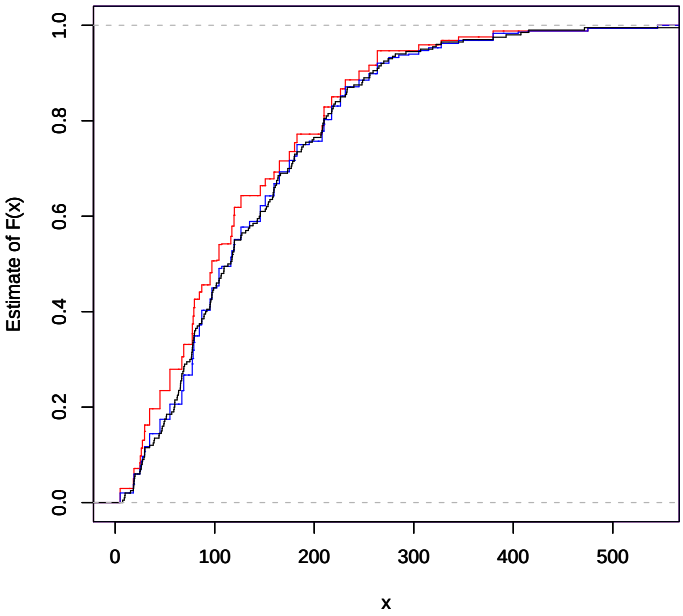

Figure 1: Estimates of F from Simulation 1: blue $= \hat{F}$, red $= F^*$ and black $= F_n$.

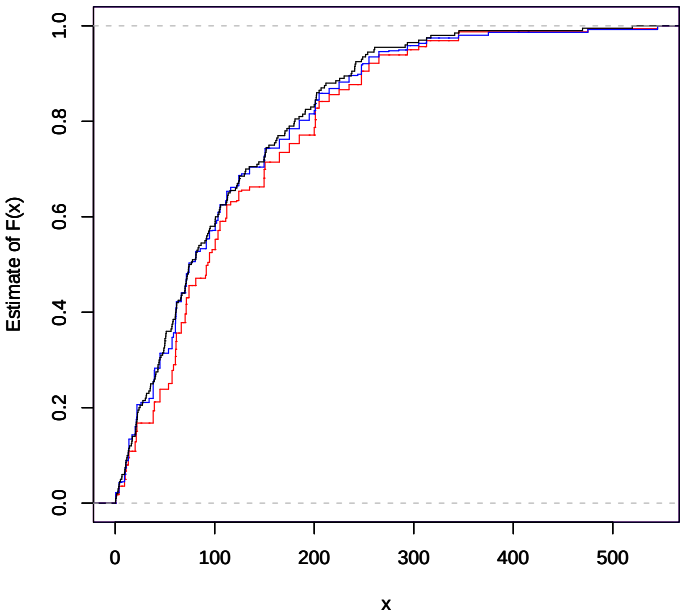

Figure 2: Estimates of F from Simulation 2: blue $= \hat{F}$, red $= F^*$ and black $= F_n$.

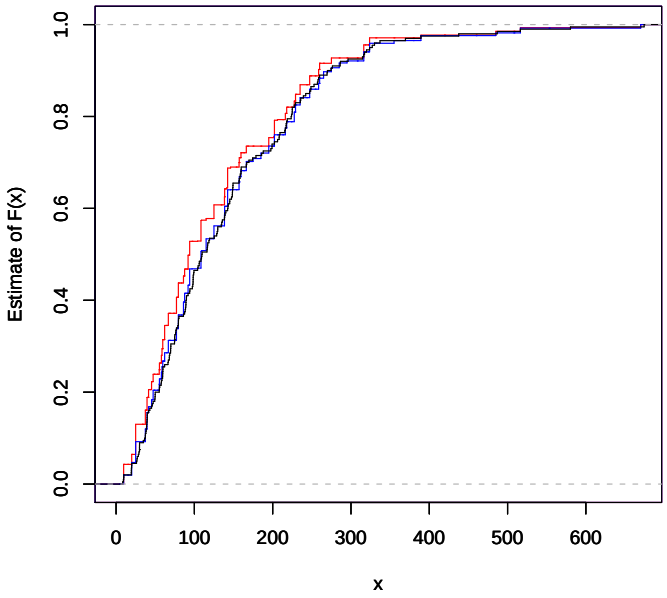

Figure 3: Estimates of F from Simulation 3: blue = $\tilde{F}$, red = F^* and black = F_n.

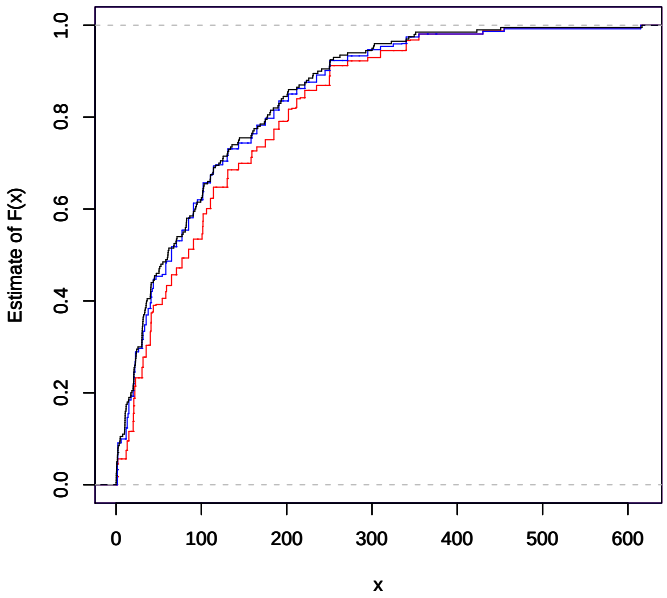

Figure 4: Estimates of F from Simulation 4: blue = $\tilde{F}$, red = F^* and black = F_n.

to have exact values placed near the upper boundaries of the intervals is more pronounced among respondents with large values of $(R_j - L_j)/R_j$. The last simulation have $\alpha = 1$ and $\beta_j = 5+10(R_j-L_j)/R_j$ and is similar to simulation 3, but the tendency to have exact values placed near the lower boundaries of the intervals is more pronounced among respondents with large values of $(R_j - L_j)/R_j$.

In simulations 1 and 2, F is estimated by computing a maximizer of $L(F, \hat{\boldsymbol{w}})$, and in simulations 3 and 4, the estimates were computed by maximizing $L(F, \tilde{\boldsymbol{w}})$. For comparison, we computed the empirical cumulative distribution function (ECDF) F_n of $X_1, ..., X_n$ and the maximizer F^* of (1) in each simulation. That is, F^* is Jammalamadaka & Mangalam's (2003) nonparametric maximum likelihood estimator for middle-censored data.

In the last two simulations we used $\theta_j = (R_j - L_j)/R_j$. The support of this random variable is not the whole real line but the interval $[0, 1]$. To overcome the boundary (bias) problem in the kernel estimation, the following simple boundary corrected kernel (Li & Racine, 2007, p. 31)

$$W_h(\theta_j, \theta) = \begin{cases} h^{-1}K\left(\frac{\theta_j-\theta}{h}\right) / \int_{-\theta/h}^{\infty} K(x)dx & \text{if } \theta \in [0, h) \\ h^{-1}K\left(\frac{\theta_j-\theta}{h}\right) & \text{if } \theta \in [h, 1-h] \\ h^{-1}K\left(\frac{\theta_j-\theta}{h}\right) / \int_{-\infty}^{(1-\theta)/h} K(x)dx & \text{if } \theta \in (1-h, 1], \end{cases}$$

was used instead of (4), where $K(\cdot)$ is the standard normal density function. The bandwidth h was determined by least-squares cross-validation.

The results from the four simulations are shown in Figures 1-4. In Figures 1 and 3, the results from simulations 1 and 3 are presented. In both these simulations the respondents tended to have the exact values placed near the upper boundaries of the intervals. We see that the estimator F^* fails to recognize this behavior in the data. That is, F^* lies above the corresponding unbiased ECDFs in Figures 1 and 3, and this implies that F^* will tend to overestimate the distribution function F. The estimator $\hat{F}$ in simulation 1 and the estimator $\tilde{F}$ in simulation 3, on the other hand, are close to the corresponding ECDFs, and appears to give valid estimates of the distribution function F. The implications of Figures 2 and 4 are that F^* will tend to underestimate F if the respondents tend to have the exact values placed near the lower boundaries of the intervals. The estimators $\hat{F}$ and $\tilde{F}$ do not seem to have this problem.

4 Conclusions

In this paper we have considered estimation for partly self-censored WTP data, and we have developed nonparametric maximum likelihood estimators $\hat{F}$ and $\tilde{F}$ of the distribution, F, of WTP. The simulations performed in this paper imply that if partly self-censored WTP data are handled by standard estimation methods for (partly) interval-censored or middle-censored data, then the resulting estimator of F will tend to be biased. The proposed estimators, $\hat{F}$ and $\tilde{F}$, do not seem to share this problem.

Acknowledgement

The author acknowledges support from the research program "Hydropower - Environmental impacts, mitigation measures and costs in regulated waters," an R&D program established and financed by Elforsk, the Swedish Energy Agency, the National Board of Fisheries and the Swedish Environmental Protection Agency.

References

Belyaev & Kriström (2010). Approach to analysis of self-selected interval data. CERE Working Paper, 2010:2, Centre for Environmental and Resource Economics, Umeå University and the Swedish University of Agricultural Sciences, Sweden.

Håkansson (2008). A new valuation question: analysis of and insights from interval open-ended data in contingent valuation. *Environmental and Resource Economics*, **39**, 175-188.

Huang (1999). Asymptotic properties of nonparametric estimation based on partly interval-censored data. *Statistica Sinica*, **9**, 501-519.

Jammalamadaka & Mangalam (2003). Nonparametric estimation for middle-censored data. *Nonparametric Statistics*, **15**, 253-265.

Li & Racine (2007). *Nonparametric Econometrics - Theory and Practice*, Princeton University Press, Princeton and Oxford.

Lindsey & Ryan (1998). Tutorial in biostatistics - Methods for interval-censored data. *Statistics in Medicine*, **17**, 219-238.

Parkkila (2009). Estimating benefits of new Baltic salmon fisheries management program. 16th Ulvön Conference on Environmental Economics, Ulvön, Sweden.

Peto (1973). Experimental survival curves for interval-censored data. *Applied Statistics*, **22**, 86-91.

Turnbull (1976).The empirical distribution with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, **38**, 290-295.