Metrika (2017) 80:377–399
DOI 10.1007/s00184-017-0610-7

Nonparametric estimation for self-selected interval data collected through a two-stage approach

Angel G. Angelov[1] · **Magnus Ekström**[1]

Received: 12 May 2016 / Published online: 16 January 2017

Abstract Self-selected interval data arise in questionnaire surveys when respondents are free to answer with any interval without having pre-specified ranges. This type of data is a special case of interval-censored data in which the assumption of noninformative censoring is violated, and thus the standard methods for interval-censored data (e.g. Turnbull's estimator) are not appropriate because they can produce biased results. Based on a certain sampling scheme, this paper suggests a nonparametric maximum likelihood estimator of the underlying distribution function. The consistency of the estimator is proven under general assumptions, and an iterative procedure for finding the estimate is proposed. The performance of the method is investigated in a simulation study.

Keywords Informative interval censoring · Self-selected intervals · Nonparameric maximum likelihood estimation · Two-stage data collection · Questionnaire surveys

1 Introduction

When being asked about a quantity, people often answer with an interval if they are not certain. For example, when asked about the distance to a given town, we would say "it is about 60–70 km". This is one of the reasons why in questionnaire surveys respondents are often allowed to give an answer in the form of an interval to a quantitative question. One common question format is the so-called range card, where the respondent is asked

✉ Angel G. Angelov
agangelov@gmail.com

Magnus Ekström
magnus.ekstrom@umu.se

[1] Department of Statistics, USBE, Umeå University, Umeå, Sweden

to select from several pre-specified intervals (called "brackets"). Another approach is known as unfolding brackets. In this case the respondent is asked a sequence of yes-no questions that narrow down the range in which the respondent's true value is. For example, the respondent is first asked "In the past year, did your household spend less than 500 EUR on electrical items?". If the answer is "yes", the next question asks if they spent more than 400 EUR. If the response to the first question is "no", the next question asks if they spent less than 600 EUR and so on. Unfolding brackets can be designed such that they elicit the same information as in a range-card question. These formats are often used for asking sensitive questions, e.g. asking about income, because they allow partial information to be obtained from respondents who are unwilling to provide exact amounts.

However, there are some issues associated with these approaches. Studies have found that the choice of bracket values in range-card questions is likely to influence responses. This is known as the bracketing effect or range bias (see, e.g., McFadden et al. 2005; Whynes et al. 2004). In questions about usage frequency (e.g. "How many hours per day do you spend on the internet?"), respondents might assume that the range of response alternatives represents a range of "expected" behaviors. Thus, they seem reluctant to report behaviors that are "extreme", i.e. the bottom and top brackets (see Schwarz et al. 1985). The unfolding brackets format is susceptible to the so-called anchoring effect (see, e.g., Furnham and Boo 2011; Van Exel et al. 2006), i.e. answers are biased toward the starting value (500 EUR in the example above). Respondents might perceive the initial value as representing a reasonable value of the quantity in question. It serves as an "anchor" or reference point, and respondents adjust their answer to be closer to the anchor than the estimate they had before seeing the question.

It is intuitively plausible that bracketing and anchoring effects would be avoided if the respondent is free to state any interval without having any hints like pre-specified values, in other words, if the question is open-ended. One such format is called respondent-generated intervals, proposed and investigated by Press and Tanur (see, e.g., Press and Tanur 2004a, b and the references therein). In this approach the respondent is asked to provide both a point value (a best guess for the true value) and an interval (a lower and an upper bound) to a question. They used hierarchical Bayesian methods to obtain point estimates and credibility intervals that are based on both the point values and the intervals.

Related to the respondent-generated intervals approach is the self-selected interval (SSI) approach suggested by Belyaev and Kriström (2010), where the respondent is free to provide any interval containing his/her true value. They proposed a maximum likelihood estimator of the underlying distribution based on SSI data. However, this estimator relies on certain restrictive assumptions on some nuisance parameters. To avoid such assumptions, Belyaev and Kriström (2012, 2015) introduced a novel two-stage approach. In the first stage of data collection (we will call it the *pilot stage*), respondents are asked to state single self-selected intervals. In the second stage (the *main stage*), each respondent from a new sample is asked two questions: (i) to provide a SSI and then (ii) to select from several sub-intervals of the SSI the one that most likely contains his/her true value. The sub-intervals in the second question of the main stage are generated from the SSIs collected in the pilot stage. Belyaev and Kriström (2012,

2015) developed a nonparametric maximum likelihood estimator of the underlying distribution for two-stage SSI data.

Data consisting of self-selected intervals or respondent-generated intervals (without the point values) are a special case of interval-censored data. Let X be a random variable of interest. An observation on X is interval-censored if, instead of observing X exactly, only an interval $(L, R]$ is observed, where $L < X \leq R$. Interval censoring also contains right censoring and left censoring as special cases, and if $R = \infty$, the observation is right-censored, while if $L = -\infty$ the observation is left-censored (see, e.g., Zhang and Sun 2010). Interval-censored data are encountered most commonly when the observed variable is the time to some event (known as time-to-event data, failure time data, survival data, or lifetime data). The problem of analyzing time-to-event data appears in many areas such as medicine, epidemiology, engineering, economics, and demography.

With regard to statistical analysis of interval-censored data, Peto (1973) considered nonparametric maximum likelihood estimation and employed a constrained Newton-Raphson algorithm. Turnbull (1976) extended the work of Peto to allow for truncation and suggested a self-consistency algorithm. Considering the case of no truncation, Gentleman and Geyer (1994) provided conditions under which Turnbull's estimator is indeed a maximum likelihood estimator and is unique. All these methods rely on the assumption of noninformative censoring, which implies that the joint distribution of L and R contains no parameters that are involved in the distribution function of X and therefore does not contribute to the likelihood function (see, e.g., Sun 2006). In the sampling schemes considered by Belyaev and Kriström (2010, 2012, 2015) this is not a reasonable assumption, thus the standard methods are not appropriate. The existing methods for analysis of time-to-event data in the presence of informative interval censoring require modeling the censoring process and estimating nuisance parameters (see Finkelstein et al. 2002) or making additional assumptions about the censoring process (see Shardell et al. 2007). These estimators are specific for time-to-event data and are not directly applicable in the context that we are discussing.

In this paper, we extend the work of Belyaev and Kriström (2012, 2015) by considering a sampling scheme where the number of sub-intervals in the second question of the main stage is limited to two or three, which is motivated by the fact that a question with a large number of sub-intervals might be difficult to implement in practice (e.g., in a telephone interview). In Sect. 2, we describe the sampling scheme. Section 3 introduces the statistical model. In Sect. 4, a nonparametric maximum likelihood estimator of the underlying distribution is proposed, and some of its properties are established. In Sect. 5, the results of a simulation study are presented, and Sect. 6 concludes the paper. Proofs and auxiliary results are given in the Appendix.

2 Sampling scheme

We consider the following two-stage scheme for collecting data. In the *pilot stage*, a random sample of n_0 individuals is selected and each individual is asked to state an interval containing his/her value of the quantity of interest. It is assumed that the endpoints of the intervals are rounded, for example, to the nearest integer or to the

nearest multiple of 10. Thus, instead of $(21.3, 47.8]$ respondents will answer with $(21, 48]$ or $(20, 50]$.

Let $d_0 < d_1 < \ldots < d_{k-1} < d_k$ be the endpoints of all observed intervals. The set $\{d_0, \ldots, d_k\}$ can be seen as a set of typical endpoints. The data, collected in the pilot stage are used only for constructing the set $\{d_0, \ldots, d_k\}$, which is then needed for the the main stage. In the case that a similar survey is conducted again, a new pilot stage is not necessary—the data from the previous survey can be used for constructing $\{d_0, \ldots, d_k\}$.

In the *main stage*, a new random sample of individuals is selected and each individual is asked to state an interval containing his/her value of the quantity of interest. We refer to this first question as Qu1. If the interval has endpoints that do not belong to $\{d_0, \ldots, d_k\}$, we exclude the respondent from the collected data. If the endpoints of the stated interval belong to $\{d_0, \ldots, d_k\}$, then the interval is split into two or three sub-intervals with endpoints from $\{d_0, \ldots, d_k\}$ and the respondent is asked to select one of these sub-intervals (the points of split are chosen in some random fashion; for details see Sect. 3). We refer to this second question as Qu2. The respondent may refuse to answer Qu2, and this will be allowed for.

Let us define a set of intervals $\mathcal{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$, where $\mathbf{v}_j = (d_{j-1}, d_j]$, $j = 1, \ldots, k$, and let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$ be the set of all intervals that can be expressed as a union of intervals from $\mathcal{V}$, i.e. $\mathcal{U} = \{(d_l, d_r] : \ d_l < d_r, \ l, r = 0, \ldots, k\}$. For example, if $\mathcal{V} = \{(0, 5], (5, 10], (10, 20]\}$, then $\mathcal{U} = \{(0, 5], (5, 10], (10, 20], (0, 10], (5, 20], (0, 20]\}$. We denote $\mathcal{J}_h$ to be the set of indices of intervals from $\mathcal{V}$ contained in $\mathbf{u}_h$ and $\mathcal{H}_j$ to be the set of indices of intervals from $\mathcal{U}$ containing $\mathbf{v}_j$:

$$\mathcal{J}_h = \{j : \ \mathbf{v}_j \subseteq \mathbf{u}_h\}, \quad h = 1, \ldots, m;$$
$$\mathcal{H}_j = \{h : \ \mathbf{v}_j \subseteq \mathbf{u}_h\}, \quad j = 1, \ldots, k.$$

In the example with $\mathcal{V} = \{(0, 5], (5, 10], (10, 20]\}$, $\mathbf{u}_5 = (5, 20] = \mathbf{v}_2 \cup \mathbf{v}_3$, hence $\mathcal{J}_5 = \{2, 3\}$. Similarly, the interval $\mathbf{v}_3 = (10, 20]$ is contained in $\mathbf{u}_3$, $\mathbf{u}_5$ and $\mathbf{u}_6$, thus $\mathcal{H}_3 = \{3, 5, 6\}$.

We can distinguish three types of answers in the main stage:

type 1. $(\mathbf{u}_h; \text{NA})$, when the respondent stated interval $\mathbf{u}_h$ at Qu1 and refused to answer Qu2;

type 2. $(\mathbf{u}_h; \mathbf{v}_j)$, when the respondent stated interval $\mathbf{u}_h$ at Qu1 and $\mathbf{v}_j$ at Qu2, where $\mathbf{v}_j \subseteq \mathbf{u}_h$;

type 3. $(\mathbf{u}_h; \mathbf{u}_s)$, when the respondent stated interval $\mathbf{u}_h$ at Qu1 and $\mathbf{u}_s$ at Qu2, where $\mathbf{u}_s$ is a union of at least two intervals from $\mathcal{V}$ and $\mathbf{u}_s \subset \mathbf{u}_h$.

In the case when $\mathbf{u}_h \in \mathcal{V}$, Qu2 is not asked, but we input the answer from Qu1, and we consider this as an answer of type 2 : $(\mathbf{u}_h; \mathbf{v}_j = \mathbf{u}_h)$. The number of respondents in the main stage is denoted by n (not counting those who were excluded).

Remark 1 This sampling scheme has two essential differences from the one introduced by Belyaev and Kriström (2012, 2015), namely (i) they include in the data for the main stage only respondents who stated at Qu1 an interval that was observed at the pilot stage, while we allow any interval with endpoints from $\{d_0, \ldots, d_k\}$, and (ii) in their

scheme the interval stated at Qu1 is split into all the sub-intervals $\mathbf{v}_j$ that it contains, while in our scheme it is split into two or three sub-intervals with endpoints from $\{d_0, \ldots, d_k\}$.

Remark 2 A question that arises naturally is: How large should the sample in the pilot stage be so that the proportion of excluded respondents in the main stage is sufficiently small? As noticed by Belyaev and Kriström (2015), this question is related to the problem of estimating the number of species in a population, which dates back to a work by Good (1953) and has been extensively treated in the literature since then. Belyaev and Kriström (2015) suggested a rule for determining the sample size for the pilot stage (stopping the sampling process) based on results by Good (1953). A similar stopping rule can be utilized for our sampling scheme.

3 Statistical model

The unobserved (interval-censored) values $x_1, \ldots, x_n$ of the quantity of interest are considered to be values of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ with distribution function $F(x) = \mathrm{P}(X_i \leq x)$. Our goal is to estimate $F(x)$ by estimating the probability mass placed on each interval $\mathbf{v}_j = (d_{j-1}, d_j]$, i.e. estimating the probabilities

$$q_j = \mathrm{P}(X_i \in \mathbf{v}_j) = F(d_j) - F(d_{j-1}), \quad j = 1, \ldots, k.$$

Thereby, the estimated distribution function will be a step function with jumps only at the points $d_1, \ldots, d_k$. To avoid complicated notation, we assume that $q_j > 0$ for all $j = 1, \ldots, k$. The case when $q_j = 0$ for some j can be treated similarly. Actually, if we have observed at Qu1 an interval $\mathbf{u}_h$ containing $\mathbf{v}_j$, it is plausible to assume that $q_j > 0$. If for some j_0 we have not observed any $\mathbf{u}_h$ containing $\mathbf{v}_{j_0}$, then we can assume that $q_{j_0} = 0$ and proceed by estimating the remaining q_j's.

Let H_i, $i = 1, \ldots, n$, be i.i.d. random variables. If the i-th respondent has stated interval $\mathbf{u}_h$ at Qu1, then $H_i = h$. The event $\{H_i = h\}$ implies $\{X_i \in \mathbf{u}_h\}$. Let us denote

$$w_{h|j} = \mathrm{P}(H_i = h \mid X_i \in \mathbf{v}_j),$$
$$p_{j|h} = \mathrm{P}(X_i \in \mathbf{v}_j \mid H_i = h).$$

The probabilities q_j are the main parameters of interest, while the conditional probabilities $w_{h|j}$ are nuisance parameters. If $w_{h|j}$ does not depend on j, the assumption of noninformative censoring will be satisfied. In our case, there are no grounds for making such assumptions about $w_{h|j}$, and therefore we need the data on Qu2 in order to estimate $w_{h|j}$.

We are considering a sampling scheme where, for the purpose of asking Qu2, the interval stated at Qu1 is split into two or three sub-intervals (we refer to these as 2-split design and 3-split design, respectively). We will now discuss how the points of split are determined. Let $\mathcal{J}_h^\circ$ be the set of indices of points from $\{d_0, \ldots, d_k\}$ that are in the interior of interval $\mathbf{u}_h$, i.e. $\mathcal{J}_h^\circ = \{j : d_{l_h} < d_j < d_{r_h}, (d_{l_h}, d_{r_h}] = \mathbf{u}_h\}$, $h =$

$1, \ldots, m$. In case of a 2-split design, the interval $\mathbf{u}_h$ (stated at Qu1) is split into two sub-intervals: $(d_{l_h}, d_j]$ and $(d_j, d_{r_h}]$, and the respondent is asked to select one of these sub-intervals. The point d_j is chosen with probability δ_{h,d_j}, $\sum_{j\in\mathcal{J}_h^\circ} \delta_{h,d_j} = 1$. In case of a 3-split design, $\mathbf{u}_h$ is split into three sub-intervals: $(d_{l_h}, d_i]$, $(d_i, d_j]$, and $(d_j, d_{r_h}]$. The points d_i and d_j are chosen with probability δ_{h,d_i,d_j}, $\sum_{i,j\in\mathcal{J}_h^\circ,\, i<j} \delta_{h,d_i,d_j} = 1$.

We denote by γ_t the probability that a respondent gives an answer of type t, for $t = 1, 2, 3$, and similarly γ_{ht} denotes the probability that a respondent, who stated $\mathbf{u}_h$ at Qu1, gives an answer of type t for $t = 1, 2, 3$. Later on, we will need to assume that $\gamma_2 > 0$ and $\gamma_{h2} > 0$. Sufficient conditions for this are given by the following proposition.

Proposition 1 (i) *If $\delta_{h,d_j} > 0$ for all $j \in \mathcal{J}_h^\circ$, and $p_{l_h+1|h} > 0$ or $p_{r_h|h} > 0$, then $\gamma_2 > 0$ and $\gamma_{h2} > 0$.*

(ii) *If $\delta_{h,d_i,d_j} > 0$ for all $i, j \in \mathcal{J}_h^\circ$, and $p_{l_h+1|h} > 0$ or $p_{r_h|h} > 0$, then $\gamma_2 > 0$ and $\gamma_{h2} > 0$.*

Let $\delta_{h,j}$ be the probability that $\mathbf{u}_h$ is split so that one of the resulting sub-intervals is $\mathbf{v}_j$, and let δ_{h*s} be the probability that $\mathbf{u}_h$ is split so that one of the resulting sub-intervals is $\mathbf{u}_s$. It is easy to see that the probabilities $\delta_{h,j}$ and δ_{h*s} can be expressed in terms of δ_{h,d_j} in case of a 2-split design, and in terms of δ_{h,d_i,d_j} in case of a 3-split design.

4 Estimation

In this section we discuss the estimation of the distribution function $F(x)$. We prove the consistency of a proposed nonparametric maximum likelihood estimator of the probabilities q_j given that the conditional probabilities $w_{h|j}$ are known. We then show that if we plug in a consistent estimator of $w_{h|j}$, the estimator of q_j is still consistent. Thereafter, we suggest an estimator of $w_{h|j}$ and show its consistency. Iterative procedures are proposed for finding the estimates of q_j and $w_{h|j}$.

4.1 Estimating the probabilities q_j

Henceforth we will need the following frequencies:

$n_{h,\mathrm{NA}}$ = Number of respondents who stated $\mathbf{u}_h$ at Qu1 and NA (no answer) at Qu2;
n_{hj} = Number of respondents who stated $\mathbf{u}_h$ at Qu1 and $\mathbf{v}_j$ at Qu2, where $\mathbf{v}_j \subseteq \mathbf{u}_h$;
n_{h*s} = Number of respondents who stated $\mathbf{u}_h$ at Qu1 and $\mathbf{u}_s$ at Qu2, where $\mathbf{u}_s$ is a union of at least two intervals from $\mathcal{V}$ and $\mathbf{u}_s \subset \mathbf{u}_h$;
$n_{h\bullet}$ = Number of respondents who stated $\mathbf{u}_h$ at Qu1 and any sub-interval at Qu2;
$n_{\bullet j}$ = Number of respondents who stated $\mathbf{v}_j$ at Qu2.

We denote by n', n'', and n''' the number of respondents who gave an answer of type 1, 2, and 3, respectively. The following are satisfied:

$$n' = \sum_h n_{h,\mathrm{NA}}, \qquad n'' = \sum_j n_{\bullet j}, \qquad n''' = \sum_{h,s} n_{h*s}, \qquad n' + n'' + n''' = n.$$

If respondent i has given an answer of type 1, i.e. $\mathbf{u}_h$ at Qu1 and NA at Qu2, then the contribution to the likelihood is $\mathrm{P}(H_i = h) = \sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j$, where the equality follows from the law of total probability. If an answer of type 2 is observed, i.e. $\mathbf{u}_h$ at Qu1 and $\mathbf{v}_j$ at Qu2, then the contribution to the likelihood is $\delta_{h,j}\, w_{h|j}\, q_j$. And in the case that we observe an answer of type 3, i.e. $\mathbf{u}_h$ at Qu1 and $\mathbf{u}_s$ at Qu2, the contribution to the likelihood is $\delta_{h*s} \sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j$. Thus, the log-likelihood function (normed by n) corresponding to the main-stage data is

$$\begin{aligned}\frac{\log L(\mathbf{q})}{n} &= \frac{1}{n}\sum_h n_{h,\mathrm{NA}} \log\Bigg(\sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j\Bigg) + \frac{1}{n}\sum_{h,j} n_{hj} \log(\delta_{h,j}\, w_{h|j}\, q_j)\\ &\quad + \frac{1}{n}\sum_{h,s} n_{h*s} \log\Bigg(\delta_{h*s} \sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j\Bigg) + c_1\\ &= \frac{n'}{n}\sum_h \frac{n_{h,\mathrm{NA}}}{n'} \log\Bigg(\sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j\Bigg) + \frac{n''}{n}\sum_j \frac{n_{\bullet j}}{n''} \log q_j\\ &\quad + \frac{n'''}{n}\sum_{h,s} \frac{n_{h*s}}{n'''} \log\Bigg(\sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j\Bigg) + c_2, \qquad (1)\end{aligned}$$

where c_1 does not depend on $\mathbf{q} = (q_1, \ldots, q_k)$ and

$$c_2 = c_1 + \frac{1}{n}\sum_{h,j} n_{hj} \log(\delta_{h,j}\, w_{h|j}) + \frac{1}{n}\sum_{h,s} n_{h*s} \log \delta_{h*s}.$$

Remark 3 If $n''' = 0$, the log-likelihood (1) has essentially the same form as the one in Belyaev and Kriström (2012).

We say that $\widetilde{\mathbf{q}}$ is an approximate maximum likelihood estimator (see, e.g., Rao 1973 p. 353) of $\mathbf{q}$ if

$$L(\widetilde{\mathbf{q}}) \geq c \sup_{\mathbf{q}\in A} L(\mathbf{q}), \quad 0 < c < 1, \qquad (2)$$

where $L(\mathbf{q})$ is the likelihood function and A is an admissible set of values of $\mathbf{q}$. In our case the admissible set is $A = \{\mathbf{q} :\ 0 < q_j < 1,\ \sum_{j=1}^k q_j = 1\}$.

Theorem 1 *Let $\widetilde{\mathbf{q}}$ be an approximate maximum likelihood estimator of $\mathbf{q}$ and $\mathbf{q}^0$ be the vector of true probabilities. If the conditional probabilities $w_{h|j}$ are known and $\gamma_2 > 0$, then $\widetilde{\mathbf{q}} \xrightarrow{\text{a.s.}} \mathbf{q}^0$ as $n \longrightarrow \infty$.*

In order to find the maximizer of the log-likelihood $\log L(\mathbf{q})$, we will consider the Lagrange function:

$$\mathcal{L}(\mathbf{q}, \lambda) = \frac{\log L(\mathbf{q})}{n} + \lambda(q_1 + \cdots + q_k).$$

If $\mathbf{q} = (q_1, \ldots, q_k)$ is a stationary point of the log-likelihood function $\log L(\mathbf{q})$ in A, then there exists λ such that $(\mathbf{q}, \lambda)$ is a solution of

$$\frac{\partial \mathcal{L}(\mathbf{q}, \lambda)}{\partial q_j} = 0, \quad j = 1, \ldots, k. \tag{3}$$

From the concavity of the log-likelihood function (see Proposition 2 in the Appendix), it follows that it can have no more than one stationary point. It is easy to see that the same is true for $\mathcal{L}(\mathbf{q}, \lambda)$. Therefore, if we find a stationary point of $\mathcal{L}(\mathbf{q}, \lambda)$, it corresponds to the unique stationary point of the log-likelihood, which will be the maximum likelihood estimate.

By taking the derivative of $\mathcal{L}(\mathbf{q}, \lambda)$ with respect to q_j, we can write equations (3) as follows:

$$\begin{aligned} &\frac{n'}{n} \sum_{h \in \mathcal{H}_j} \frac{n_{h,\mathrm{NA}}}{n'} \frac{w_{h|j}}{\sum_{i \in \mathcal{J}_h} w_{h|i}\, q_i} + \frac{n''}{n} \frac{n_{\bullet j}}{n''} \frac{1}{q_j} \\ &\quad + \frac{n'''}{n} \sum_{h,s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''} \frac{w_{h|j}}{\sum_{i \in \mathcal{J}_s} w_{h|i}\, q_i} + \lambda = 0. \end{aligned} \tag{4}$$

By multiplying (4) by q_j, then taking the sum over $j = 1, \ldots, k$ and using the identities

$$\sum_{j=1}^{k} \left(\sum_{h \in \mathcal{H}_j} \frac{n_{h,\mathrm{NA}}}{n'} \frac{w_{h|j}\, q_j}{\sum_{i \in \mathcal{J}_h} w_{h|i}\, q_i} \right) = 1, \quad \sum_{j=1}^{k} \left(\sum_{h,s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''} \frac{w_{h|j}\, q_j}{\sum_{i \in \mathcal{J}_s} w_{h|i}\, q_i} \right) = 1,$$

we get that $\lambda = -1$. Thus, equations (4) can be written as:

$$q_j = \frac{n''}{n} \frac{n_{\bullet j}}{n''} + \frac{n'}{n} \sum_{h \in \mathcal{H}_j} \frac{n_{h,\mathrm{NA}}}{n'} \frac{w_{h|j}\, q_j}{\sum_{i \in \mathcal{J}_h} w_{h|i}\, q_i} + \frac{n'''}{n} \sum_{h,s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''} \frac{w_{h|j}\, q_j}{\sum_{i \in \mathcal{J}_s} w_{h|i}\, q_i}. \tag{5}$$

For finding the solution of (5), we suggest the following iterative process, which is similar to the one proposed by Belyaev and Kriström (2012):

$$\begin{aligned} q_j^{(1)} &= 1/k, \\ q_j^{(r+1)} &= \frac{n''}{n} \frac{n_{\bullet j}}{n''} + \frac{n'}{n} \sum_{h \in \mathcal{H}_j} \frac{n_{h,\mathrm{NA}}}{n'} \frac{w_{h|j}\, q_j^{(r)}}{\sum_{i \in \mathcal{J}_h} w_{h|i}\, q_i^{(r)}} \\ &\quad + \frac{n'''}{n} \sum_{h,s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''} \frac{w_{h|j}\, q_j^{(r)}}{\sum_{i \in \mathcal{J}_s} w_{h|i}\, q_i^{(r)}}, \qquad r = 1, 2, \ldots \end{aligned}$$

When $\mathbf{q}^{(r+1)}$ is close enough to $\mathbf{q}^{(r)}$, the process is stopped. Our simulation experiments showed a very fast convergence of this iterative procedure to the true solution.

Corollary 1 *If we insert a strongly consistent estimator of $w_{h|j}$ into the log-likelihood (1) and $\gamma_2 > 0$, then the approximate maximum likelihood estimator $\widetilde{\mathbf{q}}$ is strongly consistent.*

4.2 Estimating the conditional probabilities $w_{h|j}$

We propose an estimator of the probabilities $p_{j|h},\ j \in \mathcal{J}_h$. Then, an estimator of $w_{h|j}$ can be obtained using the Bayes formula:

$$\widetilde{w}_{h|j} = \frac{\widetilde{p}_{j|h}\,\widehat{w}_h}{\sum_{s\in\mathcal{H}_j}\widetilde{p}_{j|s}\,\widehat{w}_s}, \tag{6}$$

where $\widetilde{p}_{j|h}$ is an estimator of $p_{j|h}$ and

$$\widehat{w}_h = \frac{n_{h\bullet} + n_{h,\mathrm{NA}}}{n}$$

is a strongly consistent estimator of $w_h = \mathrm{P}\,(H_i = h)$. Note that we need to estimate $w_{h|j}$ only for those h that have been observed at Qu1.

Let

$$n''_h = \sum_j n_{hj}, \qquad n'''_h = \sum_s n_{h*s}, \qquad n''_h + n'''_h = n_{h\bullet}.$$

We will consider the estimation of $p_{j|h}$ for a given h. For simplicity, we assume that $p_{j|h} > 0$ for all $j \in \mathcal{J}_h$; the case when some of them are zero can be treated similarly. Let $\mathbf{p}^h$ be the vector of $p_{j|h}$ for $j \in \mathcal{J}_h$. The log-likelihood function (normed by $n_{h\bullet}$), based on the respondents who stated the interval $\mathbf{u}_h$ at Qu1 and any sub-interval at Qu2, will be:

$$\begin{aligned}\frac{\log L_h(\mathbf{p}^h)}{n_{h\bullet}} &= \frac{1}{n_{h\bullet}}\sum_j n_{hj}\log(\delta_{h,j}\,p_{j|h}) + \frac{1}{n_{h\bullet}}\sum_s n_{h*s}\log\Bigg(\delta_{h*s}\sum_{j\in\mathcal{J}_s}p_{j|h}\Bigg) + c_3\\ &= \frac{n''_h}{n_{h\bullet}}\sum_j\frac{n_{hj}}{n''_h}\log p_{j|h} + \frac{n'''_h}{n_{h\bullet}}\sum_s\frac{n_{h*s}}{n'''_h}\log\Bigg(\sum_{j\in\mathcal{J}_s}p_{j|h}\Bigg) + c_4,\end{aligned} \tag{7}$$

where c_3 does not depend on $\mathbf{p}^h$ and

$$c_4 = c_3 + \frac{1}{n_{h\bullet}}\sum_j n_{hj}\log\delta_{h,j} + \frac{1}{n_{h\bullet}}\sum_s n_{h*s}\log\delta_{h*s}.$$

The admissible set is $A_h = \{\mathbf{p}^h :\ 0 < p_{j|h} < 1,\ \sum_{j\in\mathcal{J}_h} p_{j|h} = 1\}$.

Theorem 2 *Let $\widetilde{p}_{j|h}$ be an approximate maximum likelihood estimator of $p_{j|h}$ and $p^0_{j|h}$ be the true probability, $j \in \mathcal{J}_h$. If $\gamma_{h2} > 0$, then $\widetilde{p}_{j|h} \xrightarrow{\text{a.s.}} p^0_{j|h}$ as $n \longrightarrow \infty$.*

Remark 4 From the strong law of large numbers, it follows that $\widehat{w}_h$ is a strongly consistent estimator of w_h. This, together with Theorem 2, implies that the estimator $\widetilde{w}_{h|j}$ is strongly consistent.

The maximizer of the log-likelihood function $\log L_h(\mathbf{p}^h)$ can be found by employing the same method we used for $\log L(\mathbf{q})$. The concavity of $\log L_h(\mathbf{p}^h)$ is shown in Proposition 3 (see the Appendix). The unique stationary point is the solution of:

$$p_{j|h} = \frac{n''_h}{n_{h\bullet}} \frac{n_{hj}}{n''_h} + \frac{n'''_h}{n_{h\bullet}} \sum_{s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''_h} \frac{p_{j|h}}{\sum_{i \in \mathcal{J}_s} p_{i|h}}, \qquad j \in \mathcal{J}_h.$$

Again, we suggest an iterative process for finding the solution:

$$p^{(1)}_{j|h} = \frac{1}{|\mathcal{J}_h|},$$
$$p^{(r+1)}_{j|h} = \frac{n''_h}{n_{h\bullet}} \frac{n_{hj}}{n''_h} + \frac{n'''_h}{n_{h\bullet}} \sum_{s \in \mathcal{H}_j} \frac{n_{h*s}}{n'''_h} \frac{p^{(r)}_{j|h}}{\sum_{i \in \mathcal{J}_s} p^{(r)}_{i|h}}, \qquad r = 1, 2, \ldots$$

Remark 5 If $n_{h\bullet} = 0$, i.e. if the interval $\mathbf{u}_h$ has not been observed in type 2 or in type 3 answers, we do not have any observations in order to estimate the probabilities $p_{j|h}$, $j \in \mathcal{J}_h$. In that presumably rare case, we need to make assumptions about those probabilities. In our simulation experiments, we have assumed that all sub-intervals $\mathbf{v}_j$, $j \in \mathcal{J}_h$, are equally likely, i.e. $p_{j|h} = 1/|\mathcal{J}_h|$

5 Simulation study

We have conducted a simulation study in order to investigate the behavior of the proposed estimator. The data for the pilot stage and for Qu1 at the main stage are generated in the same way. Here we describe it for Qu1 in order to avoid unnecessary notations. In all simulations, the random variables $X_1, \ldots, X_n$ are independent and have a Weibull distribution:

$$F(x) = \mathrm{P}\,(X_i \le x) = 1 - \exp(-(x/\sigma)^a), \quad \text{for } x > 0,$$

where $a = 1.5$ and $\sigma = 80$. Let $U^{\mathrm{L}}_1, \ldots, U^{\mathrm{L}}_n$ and $U^{\mathrm{R}}_1, \ldots, U^{\mathrm{R}}_n$ be sequences of i.i.d. random variables defined below:

$$\begin{aligned} U^{\mathrm{L}}_i &= M_i\, U^{(1)}_i + (1 - M_i)\, U^{(2)}_i, \\ U^{\mathrm{R}}_i &= M_i\, U^{(2)}_i + (1 - M_i)\, U^{(1)}_i, \end{aligned} \tag{8}$$

Table 1 Summary statistics about the length of the interval at Qu1 (sample size is 2000)

Min.	1st quart.	Median	Mean	3rd quart.	Max.
10.0	40.0	50.0	51.9	60.0	80.0

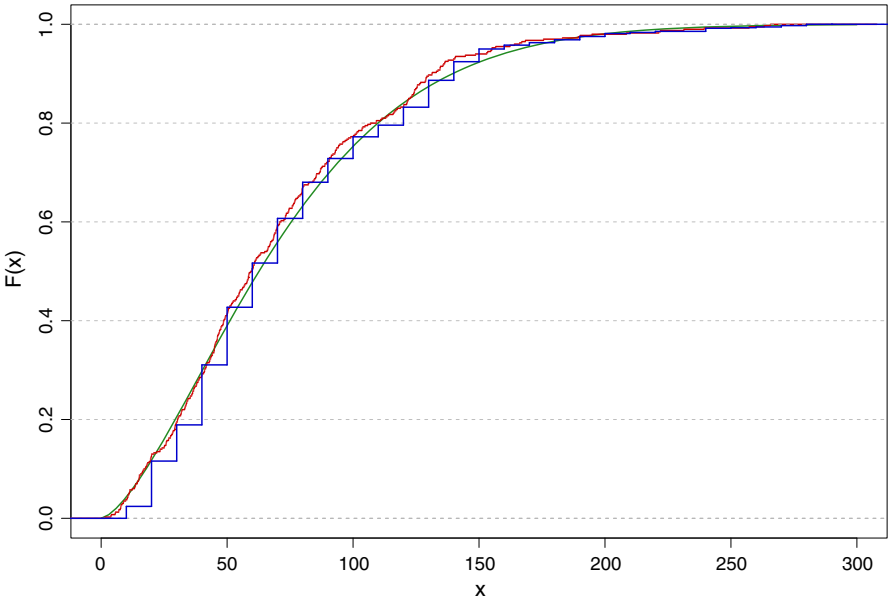

Fig. 1 True c.d.f. (*the smooth curve*), estimated c.d.f. $\widetilde{F}(x)$ using the 2-split design (*the stepwise curve* with jumps at 10, 20, 30, . . .), and empirical c.d.f. $\widehat{F}_n(x)$ of the uncensored observations for sample size $n = 400$

where $M_i \sim$ Bernoulli(1/2), $U_i^{(1)} \sim$ Uniform(0, 20), and $U_i^{(2)} \sim$ Uniform(20, 50). Let $(L_{1i}, R_{1i}]$ be the interval stated by the i-th respondent at Qu1. The left endpoints are generated as $L_{1i} = (X_i - U_i^{\mathrm{L}})\,\mathbb{1}\{X_i - U_i^{\mathrm{L}} > 0\}$ rounded downwards to the nearest multiple of 10. The right endpoints are generated as $R_{1i} = X_i + U_i^{\mathrm{R}}$ rounded upwards to the nearest multiple of 10. For the second question (Qu2) we have considered three different designs: splitting the interval stated at Qu1 into two sub-intervals, into three sub-intervals, and into all sub-intervals $\mathbf{v}_j$ that it contains. The latter corresponds to the sampling scheme explored by Belyaev and Kriström (2012). In case of a 2-split design, the point of split is chosen equally likely from all the possible points d_j that are within the interval. Similarly, in case of a 3-split design, both points of split are chosen equally likely. The probability that a respondent gives no answer to Qu2 is 1/6, and the sample size for the pilot stage is equal to 200 unless stated otherwise. The computations were performed in R (R Core Team 2015).

Some descriptive statistics about the length of the interval at Qu1 for a simulated sample of size 2000 are shown in Table 1.

Figures 1 and 2 illustrate the results of simulations with the 2-split design for sample sizes $n = 400$ and $n = 2000$. The estimated distribution function $\widetilde{F}(x) = \sum_{j:\, d_j \le x} \widetilde{q}_j$ is plotted together with the true distribution function $F(x)$ and the empirical cumulative

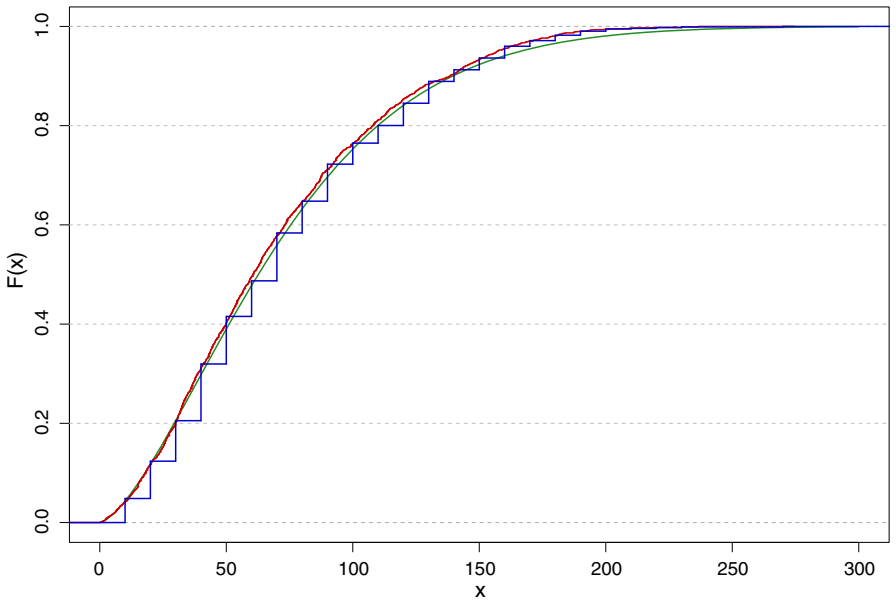

Fig. 2 True c.d.f. (*the smooth curve*), estimated c.d.f. $\widetilde{F}(x)$ using the 2-split design (*the stepwise curve* with jumps at 10, 20, 30, ...), and empirical c.d.f. $\widehat{F}_n(x)$ of the uncensored observations for sample size $n = 2000$

distribution function (e.c.d.f.) of the uncensored observations $x_1, \ldots, x_n$, i.e. $\widehat{F}_n(x) = (1/n)\sum_{i=1}^{n} \mathbb{1}\{x_i \leq x\}$. We can see that the estimate $\widetilde{F}(d_j)$ is very close to true probability $F(d_j)$ for most j, and when $\widetilde{F}(d_j)$ deviates from $F(d_j)$ a similar deviation is observed for $\widehat{F}_n(d_j)$.

It is of interest to compare the mean square error of different estimators of the probabilities q_j, $j = 1, \ldots, k$, based on different sampling schemes. We have generated 5000 samples (only the main stage is repeated 5000 times) according to the three designs described above and calculated the root mean square error (RootMSE) and the root relative mean square error (RootRelMSE). These are compared with the corresponding error when q_j is estimated from the empirical c.d.f. $\widehat{F}_n(x)$ of the uncensored observations. Figure 3 shows the results for sample size $n = 400$ and Fig. 4 shows the results for $n = 2000$. The design, corresponding to the sampling scheme in Belyaev and Kriström (2012), is denoted as "all-split". The error when using the all-split design is fairly close to the error when q_j is estimated using the uncensored observations $x_1, \ldots, x_n$. As we can expect, when using the 2-split or 3-split designs, the errors are a bit larger. We observe similar patterns for $n = 400$ and $n = 2000$, the main difference is that the error decreases with increasing sample size.

In relation to Remark 2, we have performed simulations in order to see what proportion of respondents will be accepted at the main stage when the data are generated according to the model described above. The results are given in Table 2, where n_0 is the number of respondents at the pilot stage and $n + n_{\text{rej}}$ is the number of respondents at the main stage (accepted and rejected). In the third column are the proportions when

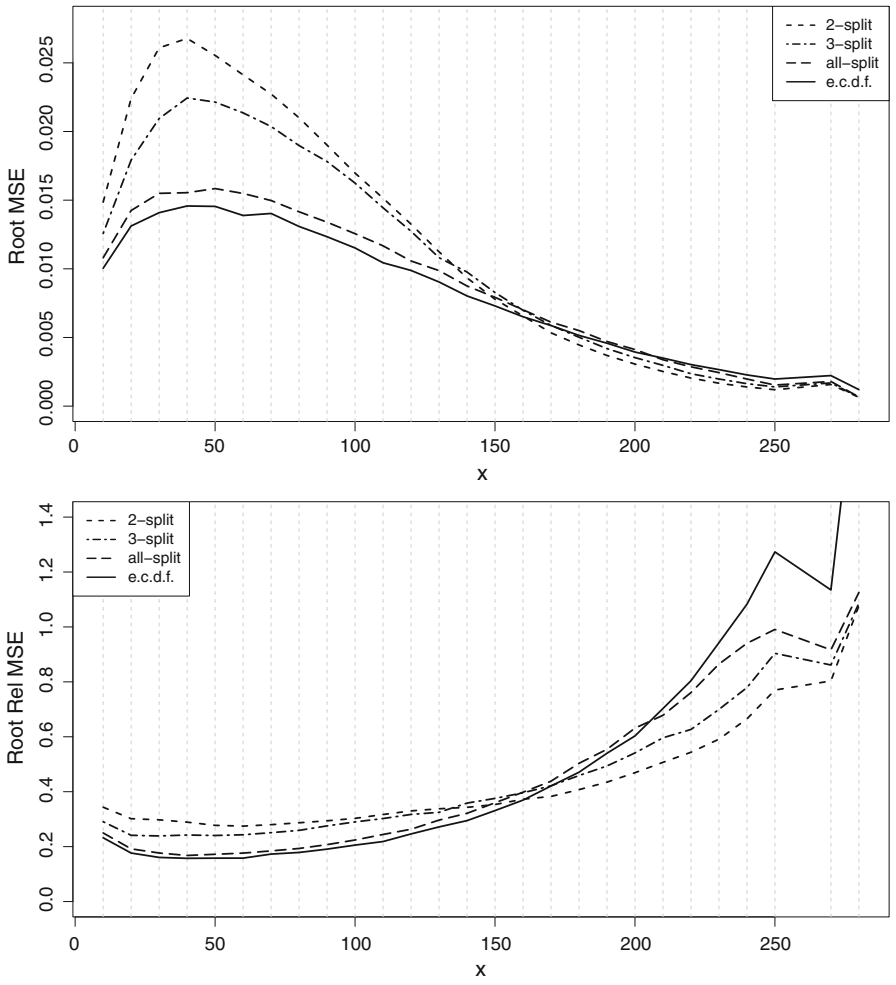

Fig. 3 Root mean square error (*top*) and root relative mean square error (*bottom*) for different estimators of $q_j = F(d_j) - F(d_{j-1}),\ j = 1, \ldots, k$, for $n = 400$. The *vertical dashed lines* correspond to the points $d_0, \ldots, d_k$. The respective error for each estimator of q_j is plotted against x-coordinate d_j

using the sampling scheme of Belyaev and Kriström (2012), and in the fourth column are the proportions when using the sampling scheme suggested in this paper (the average proportion over 3000 replications is reported). As expected, the proportion of accepted is larger for our scheme. For both schemes, the proportion gets close to one with increasing values of n_0.

We have carried out simulations to examine potential bias due to wrongly assuming that $w_{h|j}$ does not depend on j. This assumption implies noninformative censoring and in this case our method is essentially equivalent to the estimator proposed by Turnbull (1976). We compare the estimator suggested in this paper (i.e. estimating both $w_{h|j}$ and q_j from the data) with Turnbull's estimator (i.e. assuming that $w_{h|j}$ does not depend

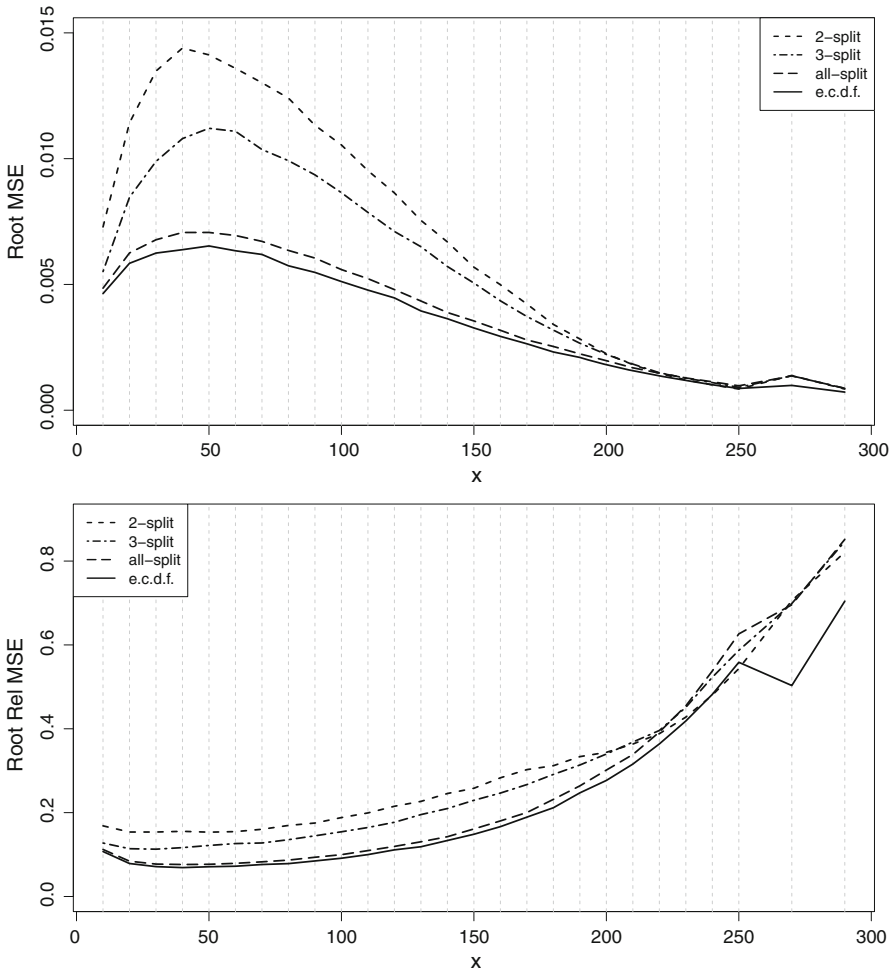

Fig. 4 Root mean square error (*top*) and root relative mean square error (*bottom*) for different estimators of $q_j = F(d_j) - F(d_{j-1}),\ j = 1, \ldots, k$, for $n = 2000$. The *vertical dashed lines* correspond to the points $d_0, \ldots, d_k$. The respective error for each estimator of q_j is plotted against x-coordinate d_j

Table 2 Average proportion of accepted respondents in the main stage (based on 3000 replications)

n_0	$n + n_{\text{rej}}$	BK2012 scheme	Modified scheme
200	400	0.8715	0.9852
200	1000	0.8721	0.9850
200	2000	0.8714	0.9855
500	1000	0.9486	0.9944
500	2500	0.9485	0.9945
500	5000	0.9485	0.9944

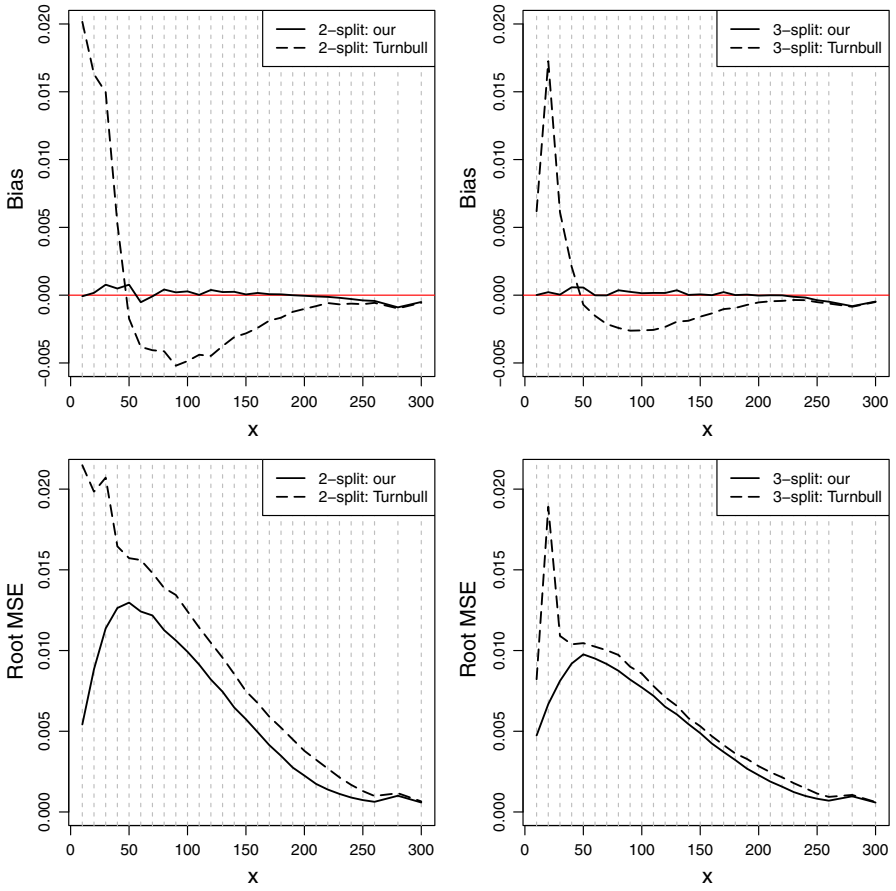

Fig. 5 Bias and root mean square error for our estimator (*solid curve*) and Turnbull's estimator (*dashed curve*), for $n = 2000$. The *vertical dashed lines* correspond to the points $d_0, \ldots, d_k$. The respective bias and error for each estimator of q_j are plotted against x-coordinate d_j

on j). For generating data, we use the model stated above with $M_i \sim$ Bernoulli(0.02) in (8). This model corresponds to a specific behavior of the respondents, that is, at Qu1 they tend to choose an interval in which the true value is located in the right half of the interval. Figure 5 presents the bias and the root mean square error of the two estimators based on 5000 simulated samples (only the main stage is repeated) of size $n = 2000$ for both the 2-split and 3-split designs. The bias of our estimator is negligible, while the bias of Turnbull's estimator is substantially larger. The RootMSE of Turnbull's estimator is larger, as well. We see that Turnbull's method on average overestimates the mass in the left tail because it puts mass uniformly over the observed interval when in fact it should put more mass to the right. It is also of interest to compare Turnbull's estimator applied to Qu1 data with Turnbull's estimator applied to 2-split data. The results, based on 5000 simulated samples of size $n = 2000$, are shown in Fig. 6. As we might expect, the bias is much larger if only the data from Qu1 are used.

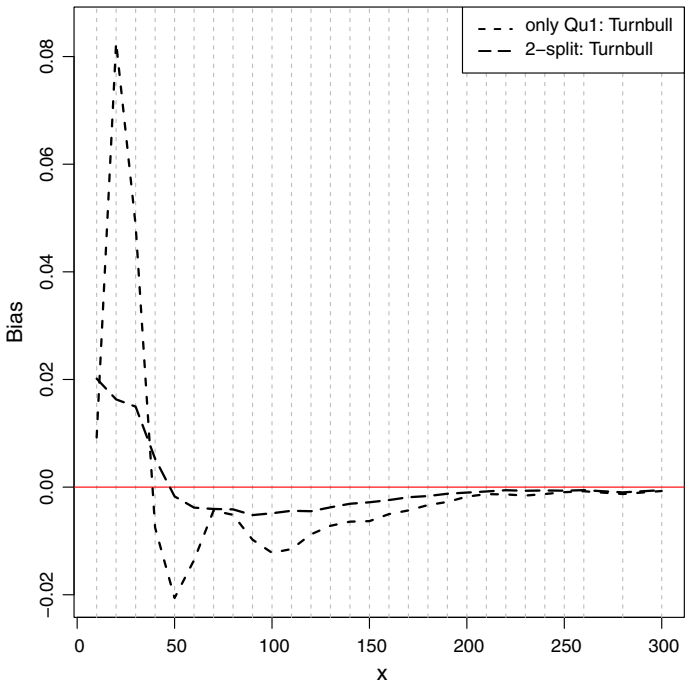

Fig. 6 Bias of Turnbull's estimator applied to Qu1 data (*short-dashed curve*) and applied to 2-split data (*long-dashed curve*), $n = 2000$

6 Concluding comments

In this paper, we considered a two-stage scheme for collecting self-selected interval data in which the number of sub-intervals in the second question of the main stage is limited to two or three. We suggested a nonparametric maximum likelihood estimator of the underlying distribution function and showed its strong consistency under easily verifiable conditions. Our simulations indicated a good performance of the proposed estimator—its error is comparable with the error of the empirical c.d.f. of the uncensored observations. It is important to note that the censoring in this context is imposed by the design of the question. A design allowing uncensored values might introduce bias in the estimation if respondents are forced to give an exact value of a quantity that is hard to evaluate exactly (e.g., number of hours spent on the internet), and consequently they give a rough "best guess". We also showed via simulations that ignoring the informative censoring and thus applying a standard method (Turnbull's estimator) can lead to serious bias.

It would be of interest to investigate the accuracy of the estimator theoretically, but we leave that as a future work.

Acknowledgements The authors would like to thank Maria Karlsson and an anonymous referee for their valuable comments which helped to improve this paper.

Appendix

Proof of Proposition 1 Using the definitions of γ_2 and γ_{h2}, we have that $\gamma_2 = \sum_h \gamma_{h2}\, w_h$. Note that γ_{h2} is defined for h such that $w_h > 0$. Let us consider a 2-split design. Then $\gamma_{h2} = \delta_{h,d_{l_h+1}}\, p_{l_h+1|h} + \delta_{h,d_{r_h-1}}\, p_{r_h|h}$, and (i) is trivial. Now, let us consider a 3-split design. Then

$$\gamma_{h2} = \delta_{h,d_{l_h+1},\bullet}\, p_{l_h+1|h} + \delta_{h,\bullet,d_{r_h-1}}\, p_{r_h|h} + \sum_{j\in\mathcal{J}_h^\circ\setminus\{r_h-1\}} \delta_{h,d_j,d_{j+1}}\, p_{j+1|h},$$

where $\delta_{h,d_{l_h+1},\bullet}$ is the probability to choose d_{l_h+1} and any other point from $\mathcal{J}_h^\circ$, and $\delta_{h,\bullet,d_{r_h-1}}$ defined similarly. From here (ii) follows trivially. □

Proposition 2 *For each $j \in \{1,\dots,k\}$, let at least one of the following be satisfied:*

(a1) there exists h, such that $j \in \mathcal{J}_h$, $n_{h,\mathrm{NA}} > 0$ and $w_{h|j} > 0$;
(a2) $n_{\bullet j} > 0$;
(a3) there exist h, s, such that $j \in \mathcal{J}_s$, $n_{h*s} > 0$ and $w_{h|j} > 0$.

Then the log-likelihood function $\log L(\mathbf{q})$ is strictly concave on A.

Proof of Proposition 2 Let $\mathbf{q}_1$ and $\mathbf{q}_2$ be any two points in A such that $\mathbf{q}_1 \neq \mathbf{q}_2$. The points $\mathbf{q}(t) = (1-t)\mathbf{q}_1 + t\mathbf{q}_2$, $t \in [0,1]$, constitute the segment that connects $\mathbf{q}_1$ and $\mathbf{q}_2$. Because A is a convex set, $\mathbf{q}(t) \in A$.

We will show that the function $\varphi(t) = \log L(\mathbf{q}(t))$, $t \in [0,1]$, is strictly concave. We have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\Bigg(\sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j(t)\Bigg) = -\frac{\left(\sum_{j\in\mathcal{J}_h} w_{h|j}\,(q_{j2}-q_{j1})\right)^2}{\left(\sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j(t)\right)^2},$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log q_j(t) = -\frac{(q_{j2}-q_{j1})^2}{(q_j(t))^2},$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\Bigg(\sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j(t)\Bigg) = -\frac{\left(\sum_{j\in\mathcal{J}_s} w_{h|j}\,(q_{j2}-q_{j1})\right)^2}{\left(\sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j(t)\right)^2}.$$

From the above it follows that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sum_j n_{h,\mathrm{NA}} \log\Big(\sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j(t)\Big)\right) \le 0, \tag{9}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sum_j n_{\bullet j} \log q_j(t)\right) \le 0, \tag{10}$$

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\left(\sum_{h,s} n_{h*s} \log\Big(\sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j(t)\Big)\right) \le 0. \tag{11}$$

If at least one of the conditions (a1)–(a3) is fulfilled, then at least one of the inequalities (9)–(11) will be strict. Therefore the second derivative of $\varphi(t)$ is negative, and the log-likelihood function $\log L(\mathbf{q})$ is strictly concave. □

Lemma 1 *(Information inequalities) Let $\sum_i a_i$ and $\sum_i b_i$ be convergent series of positive numbers such that $\sum_i a_i \ge \sum_i b_i$. Then*

$$\sum_i a_i \log\frac{b_i}{a_i} \le 0. \tag{12}$$

Further, if $a_i \le 1,\ b_i \le 1,\ \forall i$, then

$$-\sum_i a_i \log\frac{b_i}{a_i} \ge \frac{1}{2}\sum_i a_i (b_i - a_i)^2. \tag{13}$$

A proof can be found in Rao (1973, p. 58).

Proof of Theorem 1 Using the notations $\widehat{\gamma}_1 = n'/n,\ \widehat{\gamma}_2 = n''/n,\ \widehat{\gamma}_3 = n'''/n$ and

$$\widehat{w}_{h,\mathrm{NA}} = \frac{n_{h,\mathrm{NA}}}{n'}, \qquad w_h = \sum_{j\in\mathcal{J}_h} w_{h|j}\, q_j, \qquad \widehat{q}_j = \frac{n_{\bullet j}}{n''},$$
$$\widehat{w}_{h*s} = \frac{n_{h*s}}{n'''}, \qquad w_{h*s} = \sum_{j\in\mathcal{J}_s} w_{h|j}\, q_j,$$

we can write the log-likelihood (1) in a more compact way:

$$\frac{\log L(\mathbf{q})}{n} = \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log w_h + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log q_j + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log w_{h*s} + c_2. \tag{14}$$

By convention, we define $0\log 0 = 0$ and $0\log\frac{a}{0} = 0$ on the basis that $\lim_{x\downarrow 0} x\log x = 0$ and $\lim_{x\downarrow 0} x\log\frac{a}{x} = 0$ for $a > 0$. Taking logarithm of (2)

and dividing by n, we get

$$\frac{1}{n}\log L(\widetilde{\mathbf{q}}) \geq \frac{\log c}{n} + \frac{1}{n}\sup \log L(\mathbf{q}) \geq \frac{\log c}{n} + \frac{1}{n}\log L(\mathbf{q}^0).$$

After substituting $\log L(\cdot)$ from (14), the above inequality becomes

$$\begin{aligned}
&\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log \widetilde{w}_h + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \widetilde{q}_j + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \widetilde{w}_{h*s} \\
&\quad \geq \frac{\log c}{n} + \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log w_h^0 + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log q_j^0 + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log w_{h*s}^0, \quad (15)
\end{aligned}$$

where $\widetilde{w}_h = \sum_{j\in\mathcal{J}_h} w_{h|j}\,\widetilde{q}_j$, $w_h^0 = \sum_{j\in\mathcal{J}_h} w_{h|j}\,q_j^0$, and $\widetilde{w}_{h*s}$, w_{h*s}^0 are defined similarly. From inequality (12) the following are true:

$$\begin{aligned}
\sum_h \widehat{w}_{h,\mathrm{NA}} \log \widehat{w}_{h,\mathrm{NA}} &\geq \sum_h \widehat{w}_{h,\mathrm{NA}} \log \widetilde{w}_h, \\
\sum_j \widehat{q}_j \log \widehat{q}_j &\geq \sum_j \widehat{q}_j \log \widetilde{q}_j, \\
\sum_{h,s} \widehat{w}_{h*s} \log \widehat{w}_{h*s} &\geq \sum_{h,s} \widehat{w}_{h*s} \log \widetilde{w}_{h*s}.
\end{aligned}$$

From the above and (15) it follows that

$$\begin{aligned}
&\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log \widehat{w}_{h,\mathrm{NA}} + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \widehat{q}_j + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \widehat{w}_{h*s} \\
&\quad \geq \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log \widetilde{w}_h + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \widetilde{q}_j + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \widetilde{w}_{h*s} \\
&\quad \geq \frac{\log c}{n} + \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log w_h^0 + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log q_j^0 + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log w_{h*s}^0,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
0 &\geq \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log \frac{\widetilde{w}_h}{\widehat{w}_{h,\mathrm{NA}}} + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \frac{\widetilde{q}_j}{\widehat{q}_j} + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \frac{\widetilde{w}_{h*s}}{\widehat{w}_{h*s}} \\
&\geq \frac{\log c}{n} + \widehat{\gamma}_1 \sum_h \widehat{w}_{h,\mathrm{NA}} \log \frac{w_h^0}{\widehat{w}_{h,\mathrm{NA}}} + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \frac{q_j^0}{\widehat{q}_j} + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \frac{w_{h*s}^0}{\widehat{w}_{h*s}}. \quad (16)
\end{aligned}$$

From the strong law of large numbers (SLLN) it follows that

$$\begin{aligned} \widehat{\gamma}_t &\xrightarrow{\text{a.s.}} \gamma_t \\ \widehat{w}_{h,\text{NA}} &\xrightarrow{\text{a.s.}} w_h^0 \\ \widehat{q}_j &\xrightarrow{\text{a.s.}} q_j^0 \\ \widehat{w}_{h*s} &\xrightarrow{\text{a.s.}} w_{h*s}^0 \end{aligned} \tag{17}$$

as $n \longrightarrow \infty$, and therefore

$$\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\text{NA}} \log \frac{\widetilde{w}_h}{\widehat{w}_{h,\text{NA}}} + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \frac{\widetilde{q}_j}{\widehat{q}_j} + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \frac{\widetilde{w}_{h*s}}{\widehat{w}_{h*s}} \xrightarrow{\text{a.s.}} 0 \tag{18}$$

as $n \longrightarrow \infty$.

By applying inequality (13), we have

$$\begin{aligned} &- \left(\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\text{NA}} \log \frac{\widetilde{w}_h}{\widehat{w}_{h,\text{NA}}} + \widehat{\gamma}_2 \sum_j \widehat{q}_j \log \frac{\widetilde{q}_j}{\widehat{q}_j} + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} \log \frac{\widetilde{w}_{h*s}}{\widehat{w}_{h*s}} \right) \\ &\geq \frac{1}{2} \Bigg(\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\text{NA}} (\widetilde{w}_h - \widehat{w}_{h,\text{NA}})^2 + \widehat{\gamma}_2 \sum_j \widehat{q}_j (\widetilde{q}_j - \widehat{q}_j)^2 \\ &\quad + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} (\widetilde{w}_{h*s} - \widehat{w}_{h*s})^2 \Bigg) \geq 0, \end{aligned}$$

which implies that

$$\begin{aligned} &\widehat{\gamma}_1 \sum_h \widehat{w}_{h,\text{NA}} (\widetilde{w}_h - \widehat{w}_{h,\text{NA}})^2 + \widehat{\gamma}_2 \sum_j \widehat{q}_j (\widetilde{q}_j - \widehat{q}_j)^2 \\ &\quad + \widehat{\gamma}_3 \sum_{h,s} \widehat{w}_{h*s} (\widetilde{w}_{h*s} - \widehat{w}_{h*s})^2 \xrightarrow{\text{a.s.}} 0. \end{aligned}$$

Therefore

$$\widehat{\gamma}_2 \sum_j \widehat{q}_j (\widetilde{q}_j - \widehat{q}_j)^2 \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \longrightarrow \infty.$$

Because $\gamma_2 > 0$ from the above and (17) it follows that

$$\widetilde{q}_j \xrightarrow{\text{a.s.}} q_j^0 \quad \text{as } n \longrightarrow \infty.$$

□

Proof of Corollary 1 The proof follows the same lines as that of Theorem 1. Let $\overline{w}_{h|j}$ be a strongly consistent estimator of $w_{h|j}$, i.e. $\overline{w}_{h|j} \xrightarrow{\text{a.s.}} w_{h|j}$ as $n \longrightarrow \infty$.

In (15) and (16), instead of w_h^0 and w_{h*s}^0, we will have $\overline{w}_h^0 = \sum_{j\in\mathcal{J}_h} \overline{w}_{h|j}\, q_j^0$ and $\overline{w}_{h*s}^0 = \sum_{j\in\mathcal{J}_s} \overline{w}_{h|j}\, q_j^0$, respectively. The strong consistency of $\overline{w}_{h|j}$ implies that

$$\overline{w}_h^0 \xrightarrow{\text{a.s.}} w_h^0 \quad \text{and} \quad \overline{w}_{h*s}^0 \xrightarrow{\text{a.s.}} w_{h*s}^0 \quad \text{as } n \longrightarrow \infty.$$

This, together with (17), implies (18), and the rest of the proof is identical. □

Proposition 3 *For each $j \in \mathcal{J}_h$, let at least one of the following be satisfied:*

(b1) $n_{hj} > 0$;
(b2) $n_{h*s} > 0$ *for some s, such that $j \in \mathcal{J}_s$.*

Then the log-likelihood function $\log L_h(\mathbf{p}^h)$ *is strictly concave on* A_h.

Proof of Proposition 3 Because we consider $\log L_h(\mathbf{p}^h)$ for a fixed h, we will write p_j instead of $p_{j|h}$, and $\mathbf{p}$ instead of $\mathbf{p}^h$. Let $\mathbf{p}_1$ and $\mathbf{p}_2$ be any two points in A_h such that $\mathbf{p}_1 \neq \mathbf{p}_2$. The points $\mathbf{p}(t) = (1-t)\mathbf{p}_1 + t\mathbf{p}_2,\ t \in [0,1]$, constitute the segment that connects $\mathbf{p}_1$ and $\mathbf{p}_2$. Because A_h is a convex set, $\mathbf{p}(t) \in A_h$.

We will show that the function $\psi(t) = \log L_h(\mathbf{p}(t)),\ t \in [0,1]$, is strictly concave,

$$\psi(t) = \sum_j n_{hj} \log p_j(t) + \sum_s n_{h*s} \log\Bigg(\sum_{j\in\mathcal{J}_s} p_j(t) \Bigg) + nc_4.$$

We have

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log p_j(t) = -\frac{(p_{j2} - p_{j1})^2}{(p_j(t))^2},$$
$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \log\Bigg(\sum_{j\in\mathcal{J}_s} p_j(t) \Bigg) = -\frac{\left(\sum_{j\in\mathcal{J}_s} (p_{j2} - p_{j1})\right)^2}{\left(\sum_{j\in\mathcal{J}_s}\ p_j(t)\right)^2}.$$

From the above it follows that

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Bigg(\sum_j n_{hj} \log p_j(t) \Bigg) \leq 0 \quad \text{and} \quad \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Bigg(\sum_s n_{h*s} \log\Bigg(\sum_{j\in\mathcal{J}_s} p_j(t) \Bigg)\Bigg) \leq 0. \tag{19}$$

If at least one of the conditions (b1) and (b2) is fulfilled, then at least one of the inequalities in (19) will be strict. Therefore the second derivative of $\psi(t)$ is negative, and the the log-likelihood function $\log L_h(\mathbf{p})$ is strictly concave. □

Proof of Theorem 2 The proof follows the same arguments as that of Theorem 1. Using the notations

$$\widehat{\gamma}_{h2} = \frac{n''_h}{n_{h\bullet}}, \qquad \widehat{\gamma}_{h3} = \frac{n'''_h}{n_{h\bullet}}, \qquad \widehat{p}_{j|h} = \frac{n_{hj}}{n''_h},$$
$$\widehat{p}_{*s|h} = \frac{n_{h*s}}{n'''_h}, \qquad p_{*s|h} = \sum_{j\in\mathcal{J}_s} p_{j|h},$$

we can write the log-likelihood (7) in a more compact way:

$$\frac{\log L_h(\mathbf{p}^h)}{n_{h\bullet}} = \widehat{\gamma}_{h2} \sum_j \widehat{p}_{j|h} \log p_{j|h} + \widehat{\gamma}_{h3} \sum_s \widehat{p}_{*s|h} \log p_{*s|h} + c_4. \tag{20}$$

Using (2) and (12) we get

$$\begin{aligned} 0 &\geq \widehat{\gamma}_{h2} \sum_j \widehat{p}_{j|h} \log \frac{\widetilde{p}_{j|h}}{\widehat{p}_{j|h}} + \widehat{\gamma}_{h3} \sum_s \widehat{p}_{*s|h} \log \frac{\widetilde{p}_{*s|h}}{\widehat{p}_{*s|h}} \\ &\geq \frac{\log c}{n_{h\bullet}} + \widehat{\gamma}_{h2} \sum_j \widehat{p}_{j|h} \log \frac{p^0_{j|h}}{\widehat{p}_{j|h}} + \widehat{\gamma}_{h3} \sum_s \widehat{p}_{*s|h} \log \frac{p^0_{*s|h}}{\widehat{p}_{*s|h}}. \end{aligned}$$

From the SLLN it follows that

$$\begin{aligned} \widehat{\gamma}_{ht} &\xrightarrow{\text{a.s.}} \gamma_{ht} \\ \widehat{p}_{j|h} &\xrightarrow{\text{a.s.}} p^0_{j|h} \\ \widehat{p}_{*s|h} &\xrightarrow{\text{a.s.}} p^0_{*s|h} \end{aligned} \tag{21}$$

as $n \longrightarrow \infty$, and therefore

$$\widehat{\gamma}_{h2} \sum_j \widehat{p}_{j|h} \log \frac{\widetilde{p}_{j|h}}{\widehat{p}_{j|h}} + \widehat{\gamma}_{h3} \sum_s \widehat{p}_{*s|h} \log \frac{\widetilde{p}_{*s|h}}{\widehat{p}_{*s|h}} \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \longrightarrow \infty.$$

Applying inequality (13), we get

$$\widehat{\gamma}_{h2} \sum_j \widehat{p}_{j|h} (\widetilde{p}_{j|h} - \widehat{p}_{j|h})^2 + \widehat{\gamma}_{h3} \sum_s \widehat{p}_{*s|h} (\widetilde{p}_{*s|h} - \widehat{p}_{*s|h})^2 \xrightarrow{\text{a.s.}} 0.$$

Because $\gamma_{h2} > 0$ from the above and (21) it follows that

$$\widetilde{p}_{j|h} \xrightarrow{\text{a.s.}} p^0_{j|h} \quad \text{as } n \longrightarrow \infty.$$

□

References

Belyaev Y, Kriström B (2010) Approach to analysis of self-selected interval data. Working Paper 2010:2, CERE, Umeå University and the Swedish University of Agricultural Sciences, http://ssrn.com/abstract=1582853

Belyaev Y, Kriström B (2012) Two-step approach to self-selected interval data in elicitation surveys. Working Paper 2012:10, CERE, Umeå University and the Swedish University of Agricultural Sciences, http://ssrn.com/abstract=2071077

Belyaev Y, Kriström B (2015) Analysis of survey data containing rounded censoring intervals. Inf Appl 9(3):2–16

Finkelstein DM, Goggins WB, Schoenfeld DA (2002) Analysis of failure time data with dependent interval censoring. Biometrics 58(2):298–304

Furnham A, Boo HC (2011) A literature review of the anchoring effect. J Socio Econ 40(1):35–42

Gentleman R, Geyer CJ (1994) Maximum likelihood for interval censored data: consistency and computation. Biometrika 81(3):618–623

Good IJ (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40(3–4):237–264

McFadden DL, Bemmaor AC, Caro FG, Dominitz J, Jun BH, Lewbel A, Matzkin RL, Molinari F, Schwarz N, Willis RJ, Winter JK (2005) Statistical analysis of choice experiments and surveys. Mark Lett 16(3–4):183–196

Peto R (1973) Experimental survival curves for interval-censored data. J R Stat Soc C Appl 22(1):86–91

Press SJ, Tanur JM (2004a) An overview of the respondent-generated intervals (RGI) approach to sample surveys. J Mod Appl Stat Methods 3(2):288–304

Press SJ, Tanur JM (2004b) Relating respondent-generated intervals questionnaire design to survey accuracy and response rate. J Off Stat 20(2):265–287

R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Rao CR (1973) Linear statistical inference and its applications, 2nd edn. Wiley, New York

Schwarz N, Hippler HJ, Deutsch B, Strack F (1985) Response scales: effects of category range on reported behavior and comparative judgments. Public Opin Q 49(3):388–395

Shardell M, Scharfstein DO, Bozzette SA (2007) Survival curve estimation for informatively coarsened discrete event-time data. Stat Med 26(10):2184–2202

Sun J (2006) The statistical analysis of interval-censored failure time data. Springer, New York

Turnbull BW (1976) The empirical distribution function with arbitrarily grouped, censored and truncated data. J R Stat Soc B (Methodol) 38(3):290–295

Van Exel N, Brouwer W, Van Den Berg B, Koopmanschap M (2006) With a little help from an anchor: Discussion and evidence of anchoring effects in contingent valuation. J Socio Econ 35(5):836–853

Whynes DK, Wolstenholme JL, Frew E (2004) Evidence of range bias in contingent valuation payment scales. Health Econ 13(2):183–190

Zhang Z, Sun J (2010) Interval censoring. Stat Methods Med Res 19(1):53–70