NON-PARAMETRIC ESTIMATION OF A DISTRIBUTION FUNCTION WITH INTERVAL CENSORED DATA

Samuel D. Zapata
Assistant Professor and Extension Economist, Department of Agricultural Economics, Texas A&M AgriLife Extension Service, Texas A&M University, Weslaco, TX.
samuel.zapata@ag.tamu.edu

Carlos E. Carpio
Associate Professor, Department of Agricultural and Applied Economics, Texas Tech University, Lubbock, TX. carlos.carpio@ttu.edu

Selected Paper prepared for presentation at the Southern Agricultural Economics Association's 2016 Annual Meeting, San Antonio, Texas, February, 6-9 2016

NON-PARAMETRIC ESTIMATION OF A DISTRIBUTION FUNCTION WITH INTERVAL CENSORED DATA

Samuel D. Zapata
Assistant Professor and Extension Economist, Department of Agricultural Economics, Texas A&M AgriLife Extension Service, Texas A&M University, Weslaco, TX.
samuel.zapata@ag.tamu.edu

Carlos E. Carpio
Associate Professor, Department of Agricultural and Applied Economics, Texas Tech University, Lubbock, TX. carlos.carpio@ttu.edu

Abstract

Disjoint interval-censored (DIC) observations are found in a variety of applications including survey responses, contingent valuation studies and grouped data. Despite being a recurrent type of data, little attention has been given to their analysis in the nonparametric literature. In this study, we develop an alternative approach for the estimation of the empirical distribution function of DIC data by optimizing their nonparametric maximum likelihood (ML) function. In contrast to Turnbull's standard nonparametric method, our estimation approach does not require iterative numerical algorithms or the use of advanced statistical software packages. In fact, we demonstrate the existence of a simple closed-form solution to the nonparametric ML problem, where the empirical distribution, its variance, and measures of central tendency can be estimated by using only the frequency distribution of observations. The advantages of our estimation approach are illustrated using two empirical datasets.

Keywords: Empirical likelihood, Mean bounds, Turnbull, Variance-covariance matrix.

Introduction

Censoring is a very common data problem in many economic applications. A special case of censoring is disjoint interval-censored (DIC) data, in which a random variable is observed only as a set of non-overlapping intervals; thus, the true value for each observation is only known to be within a specific interval among several discrete class intervals. DIC data are found on survey responses to closed-category questions, grouped data, and contingent valuation studies using payment card designs.

Closed category questions are commonly used in survey instruments used by economists and other social-scientists. In order to simplify the respondents' task and to encourage a response, researchers tend to gather demographic and sensitive quantitative information – such as income and age – using a discrete number of categories (Bailey, 1994). As a result, the exact value of these variables is unknown, instead the survey responses are observed as DIC data.

DIC data are also generated when a continuous variable is compiled and summarized on the form of a frequency table or grouped data which is the typical case with population-wide surveys (e.g., U.S. Census Bureau, 2012). Further analysis on these data may require access to the individual responses; unfortunately, the desired level of disaggregation may not be possible given confidentiality limitations. Additionally, DIC data are found in contingent valuation studies, particularly in payment card questions, where respondents are asked to select their maximum willingness to pay for nonmarket goods from a set of possible choices (e.g., Cameron and Huppert, 1989).

The analysis of interval-censored variables requires the use of special statistical techniques. Interval-censored observations are commonly analyzed using both parametric and nonparametric maximum likelihood (ML) methods (i.e., Bhat, 1994; Hanemann, Loomis and

Kannimen, 1991; Zapata et *al*., 2011). However, parametric ML estimation relies on *a priori* assumptions about the underlying distribution of the variable of interest. Hence, if the distribution function is misspecified, results may generate inconsistent estimates. For this reason, some authors have advocated to the use of nonparametric ML techniques, in which no distribution function is imposed on the observations. In the nonparametric approach, the empirical distribution function is first estimated and then it is used to calculate statistics of interest (Haab and McConnell, 1997).

Despite being a recurrent type of data, little attention has been given to the specific analysis of DIC data in the nonparametric literature (Li, et al., 2005). In this study, we develop an alternative approach for the estimation of the empirical distribution function of DIC data using the nonparametric maximum likelihood (ML) function. In contrast to the standard nonparametric method, our estimation approach does not require iterative numerical algorithms or the use of advanced statistical software packages. In fact, we demonstrate the existence of a simple closed-form solution to the nonparametric ML problem, where the empirical distribution, its variance, and measures of central tendency can be estimated by using only the frequency distribution of observations. The advantages of the proposed estimation approach are illustrated using two empirical datasets: a population-wide survey where the income variable is available as grouped data, and another from a contingent valuation dataset.

Estimation Methods

Data Generating Process

When information is collected using a close-category elicitation format, every respondent *i* is presented with K disjoint closed intervals of the form $[A_{k-1}, A_k)$, $k = 1, \dots, K$. Denoting the true

(but unobserved) variable of interest for the i[th] individual as y_i, then the i[th] respondent is asked to select the interval that encloses her true value of y_i. Consequently, every y_i is observed to fall into one of the intervals: $[A_0, A_1)$, $[A_1, A_2)$, …, $[A_{K-1}, A_K)$.

The probability that y_i is in the k[th] interval with boundary values of A_{k-1} and A_k is given by:

$$P(A_{k-1} \leq y_i \leq A_k) = F(A_k) - F(A_{k-1}) \qquad i = 1,2, \ldots . N,$$

where $F(\cdot)$ is the underlying cumulative distribution function (CDF) of y.

Given a sample of N individuals, the log-likelihood function can be represented by

(1) $$lnL = \sum_{i=1}^{N} ln \sum_{k=1}^{K} d_{ik}[F(A_k) - F(A_{k-1})],$$

where d_{ik} is a dummy variable that indicates whether the i[th] individual chooses the k[th] interval among K options.

Parametric Estimation

The parametric procedure assumes that y follows a particular statistical distribution with parameter vector $\boldsymbol{\theta}$. Thus, the generic log-likelihood function in (1) can be re-written as a function of $\boldsymbol{\theta}$:

(2) $$lnL(\boldsymbol{\theta}|d) = \sum_{i=1}^{N} ln \sum_{k=1}^{K} d_{ik}[F(A_k; \boldsymbol{\theta}) - F(A_{k-1}; \boldsymbol{\theta})].$$

Then, optimization algorithms are used to find the value of the vector $\boldsymbol{\theta}$ that maximizes the conditional log-likelihood function in (2). For example, if the variable of interest y is assumed to follow a Normal distribution with mean μ and variance σ^2, then the specific log-likelihood function can be written as[1]:

[1] Alternatively, the log-likelihood in expression (3) could be modeled as a Truncated Normal distribution with truncation points at A_0 and A_K, or either of them.

(3) $$lnL(\mu,\sigma|d) = \sum_{i=1}^{N} ln \sum_{k=1}^{K} d_{ik} \left[\Phi\left(\frac{A_k-\mu}{\sigma}\right) - \Phi\left(\frac{A_{k-1}-\mu}{\sigma}\right)\right],$$

where $F(\cdot)$ in equation (2) has been replaced by the cumulative standard normal $\Phi(\cdot)$, and $\hat{\mu}$ and $\hat{\sigma}^2$ are the corresponding estimates of the true parameters (Swan, 1969).

Nonparametric Estimation

The nonparametric ML procedure, on the other hand, does not rely on *a priori* assumptions about the probability distribution of the variable of interest (y). Given that the probability distribution of y (F) is unknown, the nonparametric procedure considers each $F_k = F(A_k)$ in (1) as a parameter to be estimated. Moreover, in order to ensure that the likelihood estimates define a valid CDF, the ML estimation needs to be expressed as a constrained maximization problem of the form:

(4) $$\text{Max}_{\boldsymbol{F}}\ ln\, L(\boldsymbol{F}|d) = \sum_{i=1}^{N} ln \sum_{k=1}^{K} d_{ik}(F_k - F_{k-1})$$

subject to: $0 = F_0 \leq F_1 \ldots \leq F_K = 1.$

Nonparametric ML estimates ($\widehat{\boldsymbol{F}}$) are usually obtained by applying the Turnbull's self-consistent algorithm[2] (Day, 2007; Gomez, Calle, and Oller, 2004; and Turnbull, 1976), where the variance of the $\hat{F}_k$'s is calculated using finite difference approximation techniques.

Nonparametric Closed-Form Solution

Even though some authors have claimed that there is no closed-form solution to the general nonparametric ML problem described in expression (4) (e.g., Haab and McConnell, 1997; Day, 2007), there are some especial cases where a closed-form solution is available. One of these cases occurs when the observed intervals are a set of disjoint closed intervals. In particular, this

[2] Limitations of the Turnbull's self-consistent algorithm, as well as alternative algorithms to solve the nonparametric ML problem are discussed in Day (2007).

is the type of data collected in both close-category survey questions and on contingent valuation studies using a payment card design.

Note that given N individuals and K observed intervals, the unconstrained version of (4) can be rewritten as

(5) $$Max_{\boldsymbol{F}}\ ln\, L(\boldsymbol{F}|N) = \sum_{k=1}^{K} N_k ln(F_k - F_{k-1}),$$

where $F_0 = 0$, $F_K = 1$, and N_k are the number of respondents who chose the k[th] interval. The first order conditions (FOC) for the maximum likelihood problem in (5) are given by

(6) $$\frac{\partial lnL(\boldsymbol{F}|N)}{\partial F_k} = \frac{N_k}{F_k - F_{k-1}} - \frac{N_{k+1}}{F_{k+1} - F_k} = 0 \qquad k = 1,2 \dots K-1$$

or equivalently by

(7) $$N_{k+1}F_{k-1} - (N_k + N_{k+1})F_k + N_k F_{k+1} = 0 \qquad k = 1,2 \dots K-1.$$

Moreover, the FOC in (7) can be expressed in matrix form as a system of $K-1$ linear equations:

(8) $$\boldsymbol{N}\boldsymbol{F} = \boldsymbol{\Theta},$$

where $\boldsymbol{N} = \begin{bmatrix} -(N_1+N_2) & N_1 & 0 & \cdots & 0 \\ N_3 & -(N_2+N_3) & N_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & N_K & -(N_{K-1}+N_K) \end{bmatrix}$ is a $(K-1 \times K-1)$ matrix; and $\boldsymbol{F} = [F_1, F_2, \dots, F_{K-1}]^t$ and $\boldsymbol{\Theta} = [0,0, \dots, -N_{K-1}]^t$ are $(K-1)$ vectors.

Consequently, the vector $\widehat{\boldsymbol{F}} = \boldsymbol{N}^{-1}\,\boldsymbol{\Theta}$ is the solution to system of equations in (8). Furthermore, it can be shown that the k[th] element of $\widehat{\boldsymbol{F}}$ is given by

(9) $$\hat{F}_k = \frac{\sum_{j=1}^{k} N_j}{\sum_{j=1}^{K} N_j} = \frac{\sum_{j=1}^{k} N_j}{N} \qquad k = 1,2 \dots K-1.$$

Note that the unconstrained solution to F_k in (9) ensures that $0 < \hat{F}_k < 1$ and $\hat{F}_k < \hat{F}_{(k+1)}$, implicitly satisfying the constrains imposed to (4). The advantage of the estimates in (9)

compared to those obtained using the standard estimation routine is that the $\hat{F}_k$ values can be estimated simply using the "raw" proportions of observations belonging to each category without the need of any numerical optimization technique.

Although, the data generating process and methods described previously focus on individuals' responses to survey questions and corresponding maximum likelihood methods using individual level data, the result in equation (9) highlights the fact that the proposed method to estimate the empirical distribution can also be used with grouped data since the only data requirements are the intervals, the number of observations within an interval, and total number of observations in the study; or alternatively, the intervals and the proportion of observations on each interval.

The variance-covariance matrix of $\widehat{\boldsymbol{F}}$, $\text{var}(\widehat{\boldsymbol{F}})$, is given by $(-E[H(\boldsymbol{F})])^{-1}$, where $H(\boldsymbol{F})$ is the corresponding Hessian Matrix (i.e., the matrix of partial derivatives of the FOC with respect to the F_k's). In particular, it can be shown that

$$\text{(10)} \quad \text{var}(\widehat{\boldsymbol{F}}) = N^{-2}\begin{bmatrix} N_1^{-1}+N_2^{-1} & -N_2^{-1} & 0 & \cdots & 0 \\ -N_2^{-1} & N_2^{-1}+N_3^{-1} & -N_3^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -N_{K-1}^{-1} & N_{K-1}^{-1}+N_K^{-1} \end{bmatrix}^{-1},$$

where the k[th] diagonal element of $\text{var}(\widehat{\boldsymbol{F}})$ (i.e., the variance of $\hat{F}_k$) is equal to

$$\text{(11)} \qquad \text{var}(\hat{F}_k) = N^{-3}\sum_{j=1}^{k} N_j \sum_{i=k+1}^{K} N_i \qquad k = 1,2 \dots K-1.$$

There may be the case in which no y_i is observed at one or several intermediate intervals[3], if this occurs then intervals with no observations need to be pooled as described in the Appendix.

[3] The 1[st] and K[th] intervals are given by the first and last non-empty intervals (intervals with observations), respectively, regardless if there are additional empty intervals before or after them.

Mean bounds estimation

The empirical distribution estimates can be further used to obtain lower bound (LB) and upper bound (UB) estimates of the underlying mean of y. Particularly, the expected value of y can be written as (Haab and McConnell, 1997):

$$E(y) = \int_{A_0}^{A_K} y\, dF(y) \tag{12}$$

$$= \sum_{k=1}^{K} \int_{A_{k-1}}^{A_k} y dF(y).$$

Under a set of continuous and disjoined intervals, the mean bound estimators proposed by Haab and McConnell (1997) reduce to a simple weighted average, where the lower bound or upper bound of each interval is weighted by its corresponding response relative frequency. Specifically, the lower and upper mean bound estimators are given by

$$LB(y) = \sum_{k=1}^{K} A_{k-1}\left(\hat{F}_k - \hat{F}_{k-1}\right) \tag{13}$$

$$= \boldsymbol{PY_{LB}},$$

and

$$UB(y) = \sum_{k=1}^{K} A_k\left(\hat{F}_k - \hat{F}_{k-1}\right) \tag{14}$$

$$= \boldsymbol{PY_{UB}},$$

respectively. Where $\boldsymbol{P} = \left[\frac{N_1}{N}, \frac{N_2}{N}, \ldots, \frac{N_K}{N}\right]$(response relative frequency), $\boldsymbol{Y_{LB}} = [A_0, A_1, \ldots, A_{K-1}]^t$ (intervals lower bounds) and $\boldsymbol{Y_{UB}} = [A_1, A_2, \ldots, A_K]^t$ (intervals upper bounds).

Income and Welfare Applications

The main attributes of our estimation approach are first illustrated in a simulated dataset. The dataset consists of 1,000 observations generated using a Normal distribution with mean (μ) 50 and standard deviation (σ) of 12. The generated observations were then allocated into 5 class

intervals: [0, 19.99], [20, 39.99], [40, 59.99], [60, 79.99] and [80, 100]. The corresponding distribution of the simulated data set, as well as the nonparametric CDF estimates and their respective standard errors are presented in Table 1. The F_k values can be directly estimated using the "raw" proportions of observations belonging to each category. For example, $\hat{F}_1$=0.004, $\hat{F}_2$= 0.004+0.212= 0.216, and so for. Similarly, the lower and upper mean bounds can be calculated using expressions (13) and 0. The lower and upper mean bounds for the simulated data enclose the true mean and were estimated to be 39.54 and 59.53, respectively. Although not shown in the paper, it is relevant to mention that both the estimated $\hat{F}_k$'s and their variance-covariance matrix were identical to those obtained employing standard numerical optimization algorithms of the Turnbull nonparametric method.

In an attempt to demonstrate the usefulness of our nonparametric modeling approach in real data, the proposed estimation methods are used to analyze two different well-known interval-censored datasets. The first dataset comes from the U.S. Census Bureau (U.S. Census Bureau, 2012), and it contains a summary of the 2011 U.S. household income distribution. Two main characteristics make this dataset very appealing to our illustration purposes. First, this dataset is larger than most datasets analyzed in empirical studies in terms of the number of observations and intervals used. Second, as is the case with most public data containing sensitive information, the specific income level of each household is not available; instead the household income is presented in an aggregated manner. Specifically, this dataset consists of 121,084 observations (in thousand units) grouped in 42 income class intervals (Table 2).

The fact that household income data are only available at an aggregated level is not a limitation to the analysis, because the only information needed to estimate their empirical distribution and nonparametric mean bounds are the number of observations per interval. The

estimated CDF and their corresponding standard errors are presented in Table 2. The lower and upper bounds of the mean household income were estimated to be $63,194 and $86,324, respectively[4]. Our mean bound estimate encloses actual average household income reported by the U.S. Census Bureau ($69,677 with and standard error of $368). Additionally, based on the CDF in Table 2 we can also infer that the median household income in the U.S ranges from $50,000 to $54,999.

The second empirical dataset considered in this study comes from a 1984 - 1985 multipurpose survey of saltwater anglers residing in northern and central California coastal counties near San Francisco Bay. This dataset is used to illustrate a practical application of our nonparametric approach on welfare analysis. Particularly, on one of the contingent valuation questions of the survey, anglers were asked how much they are willing to pay per year to support hatcheries and habitat restoration that would result in a doubling of current fish catch rates. Anglers' willingness to pay (WTP) was gathered using a payment card elicitation approach. Specific results of the survey regarding the welfare analysis appear in Cameron and Huppert (1989), and are reproduced in Table 3. Note that there are several empty intermediate intervals in the data, thus before estimating the empirical CDF we pooled some of the intervals. Specifically, two pooled intervals were created (i.e., $250 - $449.99 and $450 - $749.99) by merging the empty intervals with their appropriate neighboring intervals (see Table 3).

The nonparametric lower and upper mean bounds of anglers' WTP for enhancement of fish stocks were estimated to be $38.83 and $59.64, respectively[5]. The median WTP is expected

[4] The lower bound of the first interval and the upper bound of the last interval were set to be equal to $0 and $1,000,000, respectively.

[5] The maximum anglers' WTP for enhancement of fish stocks was set to be equal to $1,500.

to be enclosed by the interval \$20 - \$24.99. Following Cameron and Huppert (1989), we also estimated a parametric model of anglers' WTP function assuming a Log-normal distribution (i.e., the probability of observing the k^{th} interval in expression (3) is given by $\Phi\left(\frac{\ln(A_k)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(A_{k-1})-\mu}{\sigma}\right)$). Under this assumption, the parametric unconditional mean WTP was estimated to be \$51.96 with standard error of \$4.90. The unconditional parametric WTP median was estimated to be \$23.43[6]. Both parametric mean and median estimates were enclosed by their nonparametric counterparts. Based on these results, we can argue that the Log-normal distribution seems to be a good approximation of the underlying true distribution.

Summary and Conclusions

Interval-censored observations are found in a variety of applications, from survey responses to grouped data presented to report population-wide surveys; thus the necessity of a robust and practical approach to analyze this type of data. In this paper we developed an alternative approach to estimate the empirical distribution function of the variables of interest by optimizing their corresponding nonparametric maximum likelihood function. We focused on the special censoring case where the observed data are a set of disjoint intervals. In contrast to Turnbull's standard nonparametric method, our estimation approach does not require iterative numerical algorithms or the use of advanced statistical software packages. In fact, our estimates are very

[6] A conditional mean WTP (conditional on observable characteristics) of \$61.80 with and standard deviation of \$42.22 is reported in Cameron and Huppert (1989). Additionally, the conditional median WTP was estimated to be \$25.76.

intuitive and easy to compute, and they are identical to those obtained using the standard nonparametric approach.

Empirical distribution estimates were further used to create estimates of central tendency. A practical way to estimate nonparametric lower and upper mean bounds is presented. In fact, we demonstrated that for disjoint intervals the lower and upper bound of the mean are equal to simple weighted averages, where the weights are given by the intervals' relative frequency. Also, based on the estimated empirical distribution function, an interval containing the median of the distribution can be inferred.

The attributes of our estimation approach were illustrated in two well-known datasets. The first dataset consisted of grouped observations and it contained a summary of the 2011 U.S. household income distribution. The second dataset used corresponds to a contingent valuation study, where Californian anglers where asked to state their willingness to pay for enhancement of fish stocks using a payment card elicitation technique. In both empirical applications the nonparametric mean estimates enclosed the reported mean or the estimated parametric mean of the original studies.

The proposed estimation methods can also be used as a model validation tool. Particularly, a comparison between candidate models' estimates of central tendency and the nonparametric mean bounds and median interval could be considered as an initial step to discriminate between potential models. Also, in the analysis of survey responses with grouped data, the nonparametric mean bounds can be presented along with the traditional survey summary statistics (e.g., mean, standard deviation) as a data-driven and robust interval of the true mean.

One caveat of interval-censored data analysis, in general, is that this type of analysis tends to be sensitive to the number of intervals used, as well as to the values of the interval boundaries. In terms of parametric estimation, it has been demonstrated that the intervals chosen and the estimation method employed significantly influence the outcome estimates (e.g., Cameron and Huppert, 1989). Even though we are not aware of any study analyzing the effect of interval's characteristics on the reliability of nonparametric estimates, similar effects as those observed in parametric analysis may be found on nonparametric estimates.

Two particular shortcomings of the nonparametric approaches should also be mentioned. First, in some instances the first and last intervals are open intervals, thus *ad hoc* boundary values need to be used. Finally, the non-parametric approaches limit the exploration of the effects of covariates on the estimated function.

Table 1. Distribution of the Simulated Dataset.

Interval	Observations	Relative Frequency	CDF[a]	CDF Std Error
0 - 19.99	4	0.0040	0.0040	0.0020
20 - 39.99	212	0.2120	0.2160	0.0130
40 - 59.99	599	0.5990	0.8150	0.0123
60 - 79.99	173	0.1730	0.9880	0.0034
80 – 100	12	0.0120		

[a] Estimated $\hat{F}_k$ values and corresponding standard errors are from equation 9 and 11, respectively.

Table 2. 2011 U.S. Household Income Distribution.

Household Income	Observations (1,000)	Relative Frequency	CDF[a]	CDF Std Error
Under $5,000	4,261	0.0352	0.0352	0.0005
$5,000 - $9,999	4,972	0.0411	0.0763	0.0008
$10,000 - $14,999	7,127	0.0589	0.1351	0.0010
$15,000 - $19,999	6,882	0.0568	0.1919	0.0011
$20,000 - $24,999	7,095	0.0586	0.2505	0.0012
$25,000 - $29,999	6,591	0.0544	0.3050	0.0013
$30,000 - $34,999	6,667	0.0551	0.3600	0.0014
$35,000 - $39,999	6,136	0.0507	0.4107	0.0014
$40,000 - $44,999	5,795	0.0479	0.4586	0.0014
$45,000 - $49,999	4,945	0.0408	0.4994	0.0014
$50,000 - $54,999	5,170	0.0427	0.5421	0.0014
$55,000 - $59,999	4,250	0.0351	0.5772	0.0014
$60,000 - $64,999	4,432	0.0366	0.6138	0.0014
$65,000 - $69,999	3,836	0.0317	0.6455	0.0014
$70,000 - $74,999	3,606	0.0298	0.6753	0.0013
$75,000 - $79,999	3,452	0.0285	0.7038	0.0013
$80,000 - $84,999	3,036	0.0251	0.7289	0.0013
$85,000 - $89,999	2,566	0.0212	0.7500	0.0012
$90,000 - $94,999	2,594	0.0214	0.7715	0.0012
$95,000 - $99,999	2,251	0.0186	0.7901	0.0012
$100,000 - $104,999	2,527	0.0209	0.8109	0.0011
$105,000 - $109,999	1,771	0.0146	0.8256	0.0011
$110,000 - $114,999	1,723	0.0142	0.8398	0.0011
$115,000 - $119,999	1,569	0.0130	0.8527	0.0010
$120,000 - $124,999	1,540	0.0127	0.8655	0.0010
$125,000 - $129,999	1,258	0.0104	0.8758	0.0009
$130,000 - $134,999	1,211	0.0100	0.8858	0.0009
$135,000 - $139,999	918	0.0076	0.8934	0.0009
$140,000 - $144,999	1,031	0.0085	0.9019	0.0009
$145,000 - $149,999	893	0.0074	0.9093	0.0008
$150,000 - $154,999	1,166	0.0096	0.9189	0.0008
$155,000 - $159,999	740	0.0061	0.9251	0.0008
$160,000 - $164,999	697	0.0058	0.9308	0.0007
$165,000 - $169,999	610	0.0050	0.9359	0.0007
$170,000 - $174,999	617	0.0051	0.9410	0.0007
$175,000 - $179,999	530	0.0044	0.9453	0.0007
$180,000 - $184,999	460	0.0038	0.9491	0.0006
$185,000 - $189,999	363	0.0030	0.9521	0.0006
$190,000 - $194,999	380	0.0031	0.9553	0.0006
$195,000 - $199,999	312	0.0026	0.9578	0.0006
$200,000 - $249,999	2,297	0.0190	0.9768	0.0004
$250,000 and over	2,808	0.0232		

[a] Estimated $\hat{F}_k$ values and corresponding standard errors are from equation 9 and 11, respectively.

Table 3. Anglers' WTP Function Distribution.

WTP	Observations	Relative Frequency	CDF[a]	CDF Std Error
$0 - $4.99	52	0.1520	0.1520	0.0194
$5 - $9.99	14	0.0409	0.1930	0.0213
$10 - $14.99	38	0.1111	0.3041	0.0249
$15 - $19.99	49	0.1433	0.4474	0.0269
$20 - $24.99	31	0.0906	0.5380	0.0270
$25 - $49.99	49	0.1433	0.6813	0.0252
$50 - $74.99	57	0.1667	0.8480	0.0194
$75 - $99.99	6	0.0175	0.8655	0.0184
$100 - $149.99	28	0.0819	0.9474	0.0121
$150 - $199.99	6	0.0175	0.9649	0.0099
$200 - $249.99	9	0.0263	0.9912	0.0050
$250 - $299.99	1	0.0029		
$300 - $349.99	0			
$350 - $399.99	0			
$400 - $449.99	0		0.9942 (Pooled)	0.0041
$450 - $499.99	1	0.0029		
$500 - $549.99	0			
$550 - $599.99	0			
$600 - $749.99	0		0.9971 (Pooled)	0.0029
$750 and over	1	0.0029		

[a] Estimated $\hat{F}_k$ values and corresponding standard errors are from equation 9 and 11, respectively.

References

Bailey, K. D. 1994. Methods of Social Research. The Free Press, New York.

Bhat, C.R. "Imputing a Continuous Income Variable from Grouped and Missing Income Observations." *Economics Letters* 46(1994):311-319.

Cameron, T.A and Huppert, D.D. 1989. OLS versus ML estimation of non-market resource values with payment card interval data. *Journal of Environmental Economics and Management* 17(3): 230-246.

Day, B. "Distribution-Free Estimation with Interval-Censored Contingent Valuation Data: Troubles with Turnbull?" *Environmental and Resource Economics* 37 (2007):777-95.

Gomez, G., Calle, L.M.and Oller, R. 2004. Frequentist and bayesian approaches for interval-censored data. *Statistical Papers* 45 (2): 139-73.

Haab, T. C. and McConnell, K.E. 1997. Referendum models and negative willingness to pay: alternative solutions. *Journal of Environmental Economics and Management* 32 (2): 251-70.

Hanemann, M., J. Loomis, and B. Kannimen. "Statistical Efficiency of Double-Bounded Dichotomous Choice Contingent Valuation." *American Journal of Agricultural Economics* 73(1991):1255-1263.

Li, G., Li, R. and Zhou, M. 2005. "Empirical likelihood in survival analysis." In *Contemporary Multivariate Analysis and Design of Experiments*, Edited by J. Fan and G. Li. pp. 337-350. The World Scientific Publisher.

Swan, A.V. 1969. "Computing Maximum-Likelihood Estimates for Parameters of the Normal Distribution from Grouped and Censored Data." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 18(1): 65-69

Turnbull, B.W. “The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data.” *Journal of the Royal Statistical Society* Series B 38(1976):290-95.

U.S. Census Bureau. “Current Population Survey.” 2012 Annual Social and Economic Supplement.

Zapata, S.D., C.E. Carpio, O. Isengildina-Massa, and R.D. Lamie. “Do Internet-Based Promotion Efforts Work? Evaluating Marketmaker.” *Journal of Agribusiness* 29(2011):159-180.

Appendix

Pooling Procedure to Merge Empty Intervals.

(i) For $k = 2 \rightarrow K - 1$, identify intervals with no observations.

(ii) If no participant chose the (k+1)th interval then the kth and (k+1)th intervals need to be merged into one interval containing N_k observations with boundary values of $A_{(k-1)}$ and $A_{(k+1)}$.

(iii) Continue until intervals are pooled sufficiently so that all remaining intervals have observations.

(iv) Estimate the resulting F_k's of the pooled intervals using expression (9).