Willingness to pay confidence interval estimation methods: a comparison

Valerio Gatta[*], Edoardo Marcucci[†] and Luisa Scaccia[‡]

May 7, 2014

Number of words including tables: 9875.

Abstract

The present paper proposes a comprehensive comparison of methods to build a confidence interval for willingness to pay/accept measures. The methods are compared on the basis of simulation performed under different scenarios. All the scenarios considered revealed a certain skewness in the estimated willingness to pay/accept distribution, which should be reflected in the confidence interval. The commonly used Delta method, producing intervals which are symmetric around the point estimate, fails to account for such skewness. Other methods are found to deliver more realistic confidence intervals, namely the method based on the *t*-test inversion, as well as the bootstap methods belonging to the "percentile family". The conclusions reached on the basis of the simulations are also illustrated on empirical data.

Keywords: Delta method; willingness to pay; confidence interval; discrete choice models; elasticities; standard errors.

[*] Dipartimento di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre, Via G. Chiabrera, 199, 00145 Roma, Italy, *email*: valerio.gatta@uniroma3.it

[†] Dipartimento di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre, Via G. Chiabrera, 199, 00145 Roma, Italy, *email*: emarcucci@uniroma3.it

[‡] Dipartimento di Istituzioni Economiche e Finanziarie, Università di Macerata, Via Crescimbeni 20, 62100 Macerata, Italy, *email*: scaccia@unimc.it

1 Introduction

In Economics, the willingness to pay (or accept) is the amount of money an agent would pay in order to obtain a desired good or a service (or to accept in compensation to something undesired). We will refer to this concept using the acronym WTP. In a choice modeling framework, WTP for improvement in a certain attribute, can be obtained as the ratio of that attribute and cost coefficients, when the model is linear in the attributes. However, since model estimation yields an estimate of the true coefficients, the computed WTP is also an estimate, coming from a certain probability distribution. When coefficients are estimated by means of maximum likelihood, the distribution of the WTP is the ratio between two correlated, asymptotically normal distributions. The distribution of the ratio of two normal variables has been derived by **?** and **?**, and shown to be approximately normal when the coefficient of variation of the denominator variate is negligible. More recently, **?** study the distribution of the estimate for WTP and give conditions for the finiteness of its moments, under different distributions used for the cost coefficient in random coefficient models. **?** consider, instead, the more general problem of calculating standard errors for measures derived from parameter estimation in a choice modeling framework.

Since the exact distribution of the WTP estimator is not known, different methods have been proposed in literature to build confidence interval for the parameter value. For example, under the Delta method, it is assumed that WTP is normally distributed and its variance is obtained by taking a first order Taylor expansion around the mean value of the variables involved in the ratio and calculating the variance for this expression. **??** suggest, instead, using a sort of parametric bootstrap which is based on taking a large number of draws from a multivariate normal distribution with means given by the estimated attribute coefficients and covariance given by the estimated covariance matrix of the coefficients. From these draws, simulated values of WTP are calculated and used to obtain the percentiles of the simulated distribution reflecting the desired level of confidence. Other alternatives are the Jackknife

(?) and the non parametric bootstrap (??), according to which the simulated distribution of WTP is generated, respectively, by creating subsamples through deletion of observations from the original sample and by generating subsamples randomly (with replacement) from the original sample. Each of these subsamples is used to derive an estimate of WTP. The confidence interval can then be derived in an analogous fashion to the Krinsky and Robb percentile interval. Alternative methods based on generalizing Fieller's theorem (??) have recently been considered. ?, for example, proposes to use Fieller-based confidence regions, formed by inverting a Wald-type test associated with a conveniently specified null hypothesis on WTP. On the same line, ?, derives confidence intervals based on the inversion of the t-test or the likelihood ratio test.

All these methods have advantages and drawbacks and different studies have been done to compare some of them (e.g. ?????). However, the conclusions from these studies are not always in accordance. Moreover, up to our knowledge, a comparison of all these methods does not exist. We propose to fill such a lack by presenting a comprehensive Monte Carlo study of all these procedures. In the study, we will also include methods which have not yet been considered to build WTP confidence intervals in a choice model framework, such as the bootstrap bias-corrected accelerated confidence interval (??) which is expected to be more accurate than basic bootstrap intervals, as well as bootstrap inversion of test statistics (???). In addition to the Monte Carlo study, the different approaches will also be applied to real data coming from two surveys with specific characteristics and goals: measuring service quality in local public transport and studying preference heterogeneity in airport choice behavior.

The paper is organized as follows: Section 2 briefly illustrates random utility models, commonly used in choice modeling, and WTP estimation; Section 3 is devoted to various methods for estimating WTP confidence intervals; Section 4 describes the simulation study to compare the various methods and comments its results; Section 5 applies the various methods to real data; Section 6 gives the conclusions and some guidelines for the choice of the appropriate method.

2 Logit models and WTP estimation

We consider a sample of N decision makers, facing J different alternatives, in T different choice experiments. Usual random utility formulations imply that the choice of individual n, for $n = 1, \ldots, N$, will be:

$$y_{int} = \begin{cases} 1 & \text{if } U_{int} \geq U_{jnt} \text{ for } j = 1, \ldots, J \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$U_{int} = V_{int} + \epsilon_{int} \tag{2}$$

is the unobservable utility that individual n derives from alternative i (for $i = 1, \ldots, J$), in choice experiment t (for $t = 1, \ldots, T$), V_{int} is the observable utility and ϵ_{int} is an error term. Observable utility is generally taken to be linear in the attributes so that

$$V_{int} = X_{int}\beta, \tag{3}$$

where X_{int} is a $(1 \times K)$ vector of attributes and β is a $(K \times 1)$ vector of coefficients. It follows that the choice probability associated with the alternative i chosen by individual n in choice experiment t, is defined as:

$$P_{int} = P\left(U_{int} \geq U_{jnt}, \text{ for } j = 1, \ldots, J\right).$$

Many different sub-models can be obtained from the generic formulation in (2), according to the assumptions made on the error term. For example, assuming that the error vector ϵ_n, obtained by stacking the vectors $\epsilon_{nt} = (\epsilon_{1nt} \cdots \epsilon_{Jnt})$, is independent, identically distributed (i.i.d.) according to a Gumbel density, leads to the well known Multinomial Logit (MNL) model, for which the P_{int} are analytically tractable. Assuming, instead, that ϵ_n is i.i.d. multivariate normal gives the Multinomial Probit (MNP) model. In this case, the computation of P_{int} requires the evaluation of multi-dimensional integrals, which may be analytically intractable for large choice sets. In such situations, the choice probabilities are usually simulated.

From now on, we will drop all the subscripts to lighten notation, unless they are really required, and write the random utility simply as $U = V + \epsilon$. When the observed utility is specified to be linear in the alternative attributes, as in (3), the total derivative of U with respect to changes in the k-th attribute X_k and the cost attribute X_C is given by $dU = \beta_k dX_k + \beta_C dX_C$. Setting this expression equal to zero and solving for $dX_C = dX_k$ yields the change in cost that keeps utility unchanged given a change in X_k:

$$\frac{dX_C}{dX_k} = WTP_k = -\frac{\beta_k}{\beta_C},$$

which is the willingness to pay for an improvement in X_k, WTP_k.

A point estimate of WTP (dropping the subscript for simplicity) can be obtained from sample data as:

$$\widehat{WTP} = -\frac{\hat{\beta}_k}{\hat{\beta}_C}, \tag{4}$$

where $\hat{\beta}_k$ and $\hat{\beta}_C$ are the maximum likelihood estimates of β_k and β_C, respectively. If the sample size is sufficiently large, the maximum likelihood estimates of the coefficients in the model are asymptotically normally distributed. Thus, in large samples, it seems reasonable to assume that the estimator $\widehat{WTP}$ is distributed according to the ratio of two gaussian variables. The distribution of the ratio of two gaussians has been derived by **?** and **?**, who show that the distribution is approximately normal when the coefficient of variation of the denominator variable (the cost coefficient estimate, in our framework), is negligible. The distribution of $\widehat{WTP}$ is required to build interval estimates of the parameter. These are generally built using the Delta method, briefly illustrated in Section 3.1, which relies on the double assumption of normality of the maximum likelihood estimates of the attribute coefficients and normality of their ratio. These two assumptions might not hold in small samples or whenever the coefficient of variation of the cost coefficient is not negligible, making it worth to explore and compare alternative solutions to build confidence intervals for WTP.

3 WTP confidence intervals

In this section we briefly illustrate different approaches to build conference intervals for WTP. Some of them have already been investigated in literature, while others have not been considered so far in the choice modeling framework. The illustration is mainly aimed at underlining the peculiarity of each method, with special regard to the assumptions on which any of them relies.

3.1 The Delta method

Assuming that the maximum likelihood estimates of the attribute coefficients are normally distributed, the Delta method states that also $\widehat{WTP}$, being a function of normal variates, will be asymptotically normal, i.e.

$$\widehat{WTP} \sim N\left(-\frac{\beta_k}{\beta_C}; \mathrm{var}(\widehat{WTP})\right),$$

where

$$\begin{aligned}\mathrm{var}(\widehat{WTP}) &= (\widehat{WTP}_{\beta_k})^2\hat{\sigma}^2_{\hat{\beta}_k} + (\widehat{WTP}_{\beta_C})^2\hat{\sigma}^2_{\hat{\beta}_C} + 2\widehat{WTP}_{\beta_k}\widehat{WTP}_{\beta_C}\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C} = \\ &= (-1/\hat{\beta}_C)^2\hat{\sigma}^2_{\hat{\beta}_k} + (\hat{\beta}_k/\hat{\beta}^2_C)^2\hat{\sigma}^2_{\hat{\beta}_C} + 2(-1/\hat{\beta}_C)(\hat{\beta}_k/\hat{\beta}^2_C)\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C},\end{aligned}$$

with $\widehat{WTP}_{\beta_k}$ and $\widehat{WTP}_{\beta_C}$ being the partial derivatives of $\widehat{WTP}$ with respect to β_k and β_C respectively, evaluated at the maximum likelihood estimates, and with $\hat{\sigma}^2_{\hat{\beta}_k}$, $\hat{\sigma}^2_{\hat{\beta}_C}$ and $\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C}$ represent, respectively, the estimated variances of $\hat{\beta}_k$ and $\hat{\beta}_C$ and the estimated covariance of $\hat{\beta}_k$ and $\hat{\beta}_C$.Then, a confidence interval for WTP at the $(1-\alpha)$-level is simply

$$\widehat{WTP} \pm z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}, \tag{5}$$

where $z_{\alpha/2}$ indicates the $(1-\alpha/2)\%$ percentile of the standard normal density, such that $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, with Φ being the cumulative standard normal density.

The Delta method is very simple and produces narrow confidence intervals. However, it holds only for continuous functions, so that β_C can never take the zero value. **?** show that the effective coverage rate of confidence intervals built with this method deteriorates rapidly as β_C gets closer to 0, no matter how large the sample size is. In addition, it renders symmetric confidence intervals around the point estimates $\widehat{WTP}$, while it has been shown that in practice the distribution of $\widehat{WTP}$ might be non-symmetric (**?**).

3.2 Methods based on test statistic inversion

The methods presented in this section exploit the duality between confidence intervals and hypothesis testing. The advantage of these methods, over the Delta method, is that they do not make assumptions on the distribution of $\widehat{WTP}$ and thus might perform better in case the assumption of normality does not hold for $\widehat{WTP}$. Moreover, they do not present discontinuity points, as does the Delta method in $\beta_C = 0$, and the confidence intervals are defined for all β_C. However, they require some more computational effort to calculate the confidence interval.

3.2.1 Asymptotic t-test inversion

The asymptotic t-test is generally used to test whether a parameter, whose estimator is normally distributed, is significantly different from zero. **?** extend this test to a linear combination of parameters. Recalling (4), we can postulate the following hypothesis:

$$H_0 : \beta_k + WTP\beta_C = 0. \tag{6}$$

The test statistic is (**?**):

$$T(WTP) = \frac{\hat{\beta}_k + WTP\hat{\beta}_C}{\sqrt{WTP^2\hat{\sigma}^2_{\hat{\beta}_C} + 2WTP\hat{\sigma}_{\hat{\beta}_k,\hat{\beta}_C} + \hat{\sigma}^2_{\hat{\beta}_k}}}.$$

The above test statistic is asymptotically standard normal, under the null hypothesis. The corresponding confidence interval is given by the set of WTP values for which it is not possible to reject H_0 at a given significance level. Thus, the $(1-\alpha)$-level confidence interval corresponds to the values WTP_0 such that $|T(WTP_0)| \leq z_{\alpha/2}$ or $T^2(WTP_0) \leq z^2_{\alpha/2}$. **?** derived upper and lower bounds for the confidence interval for WTP and **?** extended the result to simultaneous confidence intervals. Upper and lower bounds are obtained by solving the following second-degree-polynomial inequality for WTP_0:

$$A(WTP_0)^2 + 2B(WTP_0) + C \leq 0,$$

where

$$A = \hat{\beta}_C^2 - z_{\alpha/2}^2 \hat{\sigma}_{\hat{\beta}_C}^2 \quad , \quad B = \hat{\beta}_k \hat{\beta}_C - z_{\alpha/2}^2 \hat{\sigma}_{\hat{\beta}_k, \hat{\beta}_C} \quad , \quad C = \hat{\beta}_k^2 - z_{\alpha/2}^2 \hat{\sigma}_{\hat{\beta}_k}^2 . \tag{7}$$

The following algorithm can be used to compute the confidence interval:

1. fit the model to obtain maximum likelihood estimates of the parameter vector β and variance-covariance matrix of its estimate;

2. compute the quantities A, B and C as in (7) and let $\Delta = B^2 - AC$;

3. calculate the confidence interval as:

$$\begin{array}{rl} [WTP_L \ ; \ WTP_U] & \text{if } \Delta > 0 \text{ and } A > 0 \\ (-\infty \ ; \ WTP_L] \bigcup [WTP_U \ ; \ \infty) & \text{if } \Delta > 0 \text{ and } A < 0 \\ (-\infty \ ; \ \infty) & \text{if } \Delta < 0 \text{ (which implies } A < 0) \end{array} \tag{8}$$

   where $WTP_L = \dfrac{-B - \sqrt{\Delta}}{A}$ and $WTP_U = \dfrac{-B + \sqrt{\Delta}}{A}$.

It is worthwhile noticing that the confidence interval in (8) can be either a bounded or an unbounded interval, with the unbounded solution occurring only when $|\hat{\beta}_C / \hat{\sigma}_{\hat{\beta}_C}| \leq z_{\alpha/2}$, i.e. when the parameter β_C is not significantly different from zero at level α. Thus, the coverage rate provided by the asymptotic t-test method does not deteriorate as β_C approaches zero. Notice, also, that the bounded confidence interval in (8) is not symmetrical with respect to $\widehat{WTP}$, with the interval's mid-point being greater than $\widehat{WTP}$.

3.2.2 Likelihood ratio test inversion

The likelihood ratio test for the null hypothesis in (6) compares the likelihood of the unrestricted model with that of the restricted model, where the restriction is the one imposed under the null hypothesis. The test statistic is:

$$LR = -2[l(\hat{\beta}^R) - l(\hat{\beta})], \tag{9}$$

where $l(\hat{\beta}^R)$ and $l(\hat{\beta})$ represent the logarithm of the likelihood at the maximum likelihood estimates for the restricted and unrestricted model, respectively. Under the null hypothesis, LR is distributed χ^2 with one degree of freedom, corresponding to the single linear restriction $\beta_k + WTP\beta_C = 0$. Inverting the test statistic (9) to obtain a confidence interval for WTP, requires a search for the maximum and minimum values of WTP for which $-2[l(\hat{\beta}^R) - l(\hat{\beta})] \leq \chi^2_{1,\alpha}$. The following algorithm (**?**) can be used to estimate the interval lower bound (a similar one can be set for the upper bound). First of all, fit the model to the unconstrained systematic utility function

$$V = \beta_k X_k + \beta_C X_C + \sum_{h=1}^{K} \beta_h X_h \tag{10}$$

and obtain maximum likelihood estimates $\hat{\beta}$, the corresponding $\widehat{WTP}$ and the unrestricted log-likelihood $l(\hat{\beta})$. Then, initialize the algorithm by letting $g = 1$, $\text{Inf}^{(0)} = \widehat{WTP} - \lambda$, with λ being a sufficiently large positive value, $\text{Sup}^{(0)} = \widehat{WTP}$, $\text{Tol} = 1000$ and ϵ be an arbitrarily small tolerance limit, and perform the following steps until $\text{Tol} > \epsilon$:

1. let $WTP^{(g)} = \dfrac{\text{Inf}^{(g-1)} + \text{Sup}^{(g-1)}}{2}$;

2. fit the constrained model using the constrained utility function

$$V_{\text{con}} = \beta_C(-WTP^{(g)} X_k + X_C) + \sum_{h=1}^{K} \beta_h X_h, \tag{11}$$

   obtain maximum likelihood restricted estimates and the restricted log-likelihood and then calculate $LR^{(g)}$ as in (9);

3. if $LR^{(g)} < \chi^2_{1,\alpha}$ then let $\text{Sup} = WTP^{(g)}$ and $WTP^{(g+1)} = \dfrac{\text{Inf}^{(g)} + \text{Sup}^{(g)}}{2}$, otherwise if $LR^{(g)} > \chi^2_{1,\alpha}$ let $\text{Inf} = WTP^{(g)}$ and $WTP^{(g+1)} = \dfrac{\text{Inf}^{(g)} + \text{Sup}^{(g)}}{2}$;

4. set $\text{Tol} = |LR^{(g)} - \chi^2_{1,\alpha}|$ and $g = g + 1$.

When the algorithm stops, the lower bound of the interval is simply the last value $WTP^{(g)}$.

In addition to the advantages of test statistic inversion methods, the usage of this method is not restricted to linear utility functions. As a drawback, it requires an iterative procedure to obtain each interval limit and is computationally more demanding than the Delta method and the method based on the t-test inversion but much less intensive than any bootstrap method. To our knowledge, the potentiality of this method in computing confidence intervals for WTP have not been investigated through an appropriate simulation study, so far. **?** simply compared it with other methods on the basis of real data.

3.3 Methods based on parametric and nonparametric bootstrap

The bootstrap approach to confidence intervals consists in using a simulated distribution of the parameter estimator in place of its analytic distribution to obtain interval estimates for the parameter of interest (**??**). Before we discuss the various methods for bootstrap confidence interval construction, we give algorithms for non-parametric and parametric simulation, and illustrate these in a regression context, which applies to our study.

3.3.1 Resampling plans

In the following we briefly illustrate different strategies to produce a bootstrap distribution of $\widehat{WTP}$. Such a distribution is obtained on the basis of B bootstrap samples.

Parametric resampling. In parametric resampling we assume that a parametric model for the data is known up to the unknown parameter vector, which is generally replaced by its maximum likelihood estimates. In the regression context 'assuming the model' means treating the assumptions of the regression model as true, that is, assuming that the predictors are known without error and that the error terms follow a specific distribution (for example, the Gumbel distribution if we believe that the data have been generated under a MNL model).

Let $\hat{\beta}$ be the maximum likelihood estimate of β obtained by fitting the logit model (the MNL for example) to the original data. The algorithm to produce the bootstrap distribution of $\widehat{WTP}$, according to the parametric resampling scheme performs the following steps, for $b = 1, \ldots, B$:

1. generate a vector of residuals (equal in size to the number of observations in the original data set) parametrically, drawing each component independently from the same specified distribution (the Gumbel distribution, if we assume a MNL model); let this vector be $e^{\star}_{(b)}$;
2. compute $\hat{U}^{\star}_{int(b)} = X_{int}\hat{\beta} + e^{\star}_{int(b)}$ and, thus, y_{int} according to (1), $\forall i, n, t$, and produce a parametric bootstrap sample, $y^{\star}_{(b)}$;
3. regress the bootstrapped values $y^{\star}_{(b)}$ on the fixed predictors to obtain bootstrap replications of the estimated regression coefficients, $\hat{\beta}^{\star}_{(b)}$, and bootstrap replications of the estimated WTP parameter, $\widehat{WTP}^{\star}_{(b)}$.

We stress that under this resampling plan, the predictors are treated as fixed, making this scheme particularly suitable for designed experiments where the values of the predictors are set by the experimenter. Notice that this is the natural framework of a stated preference study.

Non-parametric resampling. Non-parametric resampling makes no assumptions concerning the model behind the data. Let the original sample of observations be $w_{int} = (y_{int} \;, X_{int})$, for $i = 1, \ldots, J$, $n = 1, \ldots, N$ and $t = 1, \ldots, T$. For $b = 1, \ldots, B$ the sampling algorithm proceeds as follows:

1. resample the observations w_{int} with replacement in order to generate a new sample of observations; let this sample be $w^{\star}_{(b)}$ and have the same number of observations as the original one;

2. fit the logit model to the bootstrap sample $w^{\star}_{(b)}$ and obtain the estimated regression coefficients, $\hat{\beta}^{\star}_{(b)}$, and the estimated WTP parameter, $\widehat{WTP}^{\star}_{(b)}$.

Notice that under this sampling scheme, also the predictors are treated as random. As stated before, thinking of the model matrix as fixed makes more sense in our stated preferences framework. However we also considered this non-parametric random-x resampling plan for the following reason. Fixed-x resampling enforces the assumption that the errors are identically distributed by resampling residuals from a common distribution. Consequently, if the model is incorrectly specified – for example, if there is unmodelled nonlinearity, non-constant error variance, outliers – these characteristics will not carry over into the resampled data sets. For this reason, it may be preferable to perform random-x resampling even when it makes sense to think of the model matrix as fixed. As **?** and **?** stress, non-parametric simulation, in practice, gives results that generally mimic the results obtained under the best fitting, *not* the simplest parametric model.

Krinsky and Robb resampling. We conclude with a sampling scheme proposed by **??** specifically for estimating confidence intervals for elasticities. Under their simulation scheme, many samples of the parameters of the model are generated by taking drawings from a multivariate normal distribution with the mean and variance-covariance matrix of the estimated parameters. Let $\hat{\beta}$ and $\hat{\Sigma}_{\hat{\beta}}$ be respectively the parameter estimates for the original data set and their variance-covariance matrix. Then, for $b = 1, \ldots, B$ the sampling algorithm proceeds as follows:

1. draw a vector $\hat{\beta}^{\star}_{(b)}$ from a $N(\hat{\beta}, \hat{\Sigma}_{\hat{\beta}})$;
2. use the vector $\hat{\beta}^{\star}_{(b)}$ to calculate $\widehat{WTP}^{\star}_{(b)}$.

This sampling scheme obviously relies on the assumptions that maximum likelihood estimates $\hat{\beta}$ are multivariate normal distributed and, thus, can be considered as a parametric sampling scheme. Such an assumption might be inappropriate, particularly for small samples. Moreover, the variance-covariance matrix of parameter estimates might be incorrect in case of model misspecification.

3.3.2 Bootstrap confidence intervals

Once an appropriate bootstrap sample $\widehat{WTP}^{\star}_{(b)}$, for $b = 1, \ldots, B$, as been drawn following one of the sampling strategies illustrated in Section 3.3.1, bootstrap confidence intervals can be obtained in many different ways. In this section, we briefly review nine different strategies. These can be divided into three families. The first family is the pivotal family, in which the confidence interval is constructed in the usual way, using a pivotal function, except that the quantiles of known distributions (normal, Student's-t etc.) are replaced by their bootstrap estimates. In the Normal-theory interval described below, instead of the quantiles, is the standard deviation of the estimator which is replaced by its bootstrap estimate. The second family is that of non-pivotal methods, which all originate from the percentile method as successively more complex analytical corrections for this. The third family is that of test-inversion intervals, which exploits the duality between confidence intervals and tests. Different methods are discussed in technical detail by **?** and **?**. The test-inversion intervals are reviewed by **??**. Practical examples of confidence interval construction are given by **?** and **?**, together with some S-plus software. An alternative viewpoint is given by **?**. Most of the methods presented below have not been used, yet, to build confidence intervals for WTP and their performances have not been explored through simulation studies. We will explicitly mention the methods already studied in the literature.

Basic interval or non-Studentized bootstrap method (B). The most natural way of constructing a confidence interval for WTP is to seek a function of the estimator $\widehat{WTP}$ and the parameter WTP whose distribution is known, and then use the quantiles of this known distribution to construct a confidence interval for the parameter. When the distribution of the population from which the observations are drawn is unknown, it is not clear which function of the parameter and estimator should be chosen. However, since many estimators are asymptotically normally distributed about their mean, it is reasonable to use

$$W = \widehat{WTP} - WTP \tag{12}$$

as such a function. If the distribution of W were known, a $(1-\alpha)$-level confidence interval for WTP would be $\left[\widehat{WTP} - w_{1-\alpha/2} \ ; \ \widehat{WTP} - w_{\alpha/2}\right]$, where w_α is the quantile of W such that $P(W < w_\alpha) = \alpha$. In case the distribution of W is unknown, the bootstrap procedure suggests to replace the quantile, w_α, with the appropriate quantile, $w^\star_\alpha$, of $W^\star$, calculated through the following algorithm:

1. set $W^\star_{(b)} = \widehat{WTP}^\star_{(b)} - \widehat{WTP}$, for $b = 1, \ldots, B$;
2. estimate the α-th quantile of $W^\star$ as $\hat{w}^\star_\alpha$, the ordered value of $\{W^\star_{(b)}, b = 1, \ldots, B\}$ which occupies the position $\alpha(B+1)$.
3. calculate the $(1-\alpha)$-level non-Studentized pivotal interval as:

$$\left[\widehat{WTP} - \hat{w}^\star_{1-\alpha/2} \ ; \ \widehat{WTP} - \hat{w}^\star_{\alpha/2}\right]. \tag{13}$$

This procedure involves two distinct sources of error: the bootstrap error, arising from the replacement of the quantile of W with that of $W^\star$ and the Monte Carlo error, arising from Monte Carlo simulation used to estimate the α-th quantile of $W^\star$. Provided the number of simulations B is sufficiently large, the Monte Carlo error is usually negligible compared to the bootstrap error. Unfortunately, the distributions of W and $W^\star$ might differ markedly, leading to substantial coverage error in the confidence interval in (13). Morover, if there is a parameter constraint (such as $WTP > 0$) then the interval might include invalid parameter values. On the other hand, the advantage of this procedure is to provide simple to calculate confidence intervals, which proved to be particularly accurate for some parameters such as the median (**?**, p. 52).

Bootstrap-t interval or Studentized pivotal (S). This method was first suggested in **?**, but some poor numerical results reduced its appeal. **?** showed the bootstrap-t's good second-order properties, reviving interest in its use. The method tries to overcome the shortcomings of the non-Studentized pivotal method, due to the difference often observed between the distributions of W and $W^\star$. Such a difference derives in most cases from the difference between the variances of the two distributions. By analogy with Student's t-statistic, the bootstrap-t methodology suggests to replace the function (12) with

$$W = \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}}, \tag{14}$$

where $\sqrt{\mathrm{var}(\widehat{WTP})}$ is an estimate of the standard deviation of $\widehat{WTP}$. The endpoints of a $(1-\alpha)$-level two-sided confidence interval for WTP are:

$$\left[\widehat{WTP} - w_{1-\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\ ;\ \widehat{WTP} - w_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\right].$$

In the usual Student's-t case, the percentiles w_α are taken to be those of the Student distribution. However, such a distribution might not be appropriate in the present case, to approximate the distribution of W. The idea of the bootstrap-t is to estimate the percentiles of W by bootstrapping, through the following algorithm

1. set $W^\star_{(b)} = \dfrac{\widehat{WTP}^\star_{(b)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^\star_{(b)})}}$, for $b = 1, \ldots, B$, where $\mathrm{var}(\widehat{WTP}^\star_{(b)})$ must be numerically computed for each bootstrap data set, using for example the Delta method illustrated in Section 3.1;

2. estimate the α-th quantile of $W^\star$ as $\hat{w}^\star_\alpha$, the ordered value of $\{W^\star_{(b)}, b = 1, \ldots, B\}$ which occupies the position $\alpha(B+1)$.

3. calculate the $(1-\alpha)$-level Studentized pivotal interval as:

$$\left[\widehat{WTP} - \hat{w}^\star_{1-\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\ ;\ \widehat{WTP} - \hat{w}^\star_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP})}\right]. \tag{15}$$

In practice, the only difference between the confidence interval in (5) and the Studentized pivotal interval consists in the quantiles used.

Normal-theory interval (N). Assuming that $\widehat{WTP}$ is approximately normally distributed, another type of $(1-\alpha)$-level bootstrap confidence interval can be obtained as in (5), where now $\mathrm{var}(\widehat{WTP})$ is estimated on the bootstrap sample. The following algorithm delivers a normal-theory bootstrap confidence interval:

1. estimate $\mathrm{var}(\widehat{WTP}^{\star}) = \dfrac{1}{B-1}\sum_{b=1}^{B}(\widehat{WTP}^{\star}_{(b)} - \overline{WTP}^{\star})^2$ where $\overline{WTP}^{\star} = \sum_{b=1}^{B} \widehat{WTP}^{\star}_{(b)}/B$ is the mean of the B bootstrap replicates of $\widehat{WTP}$;
2. calculate the $(1-\alpha)$-level bootstrap confidence interval as:

$$\left[\widehat{WTP} - z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}\ ;\ \widehat{WTP} + z_{\alpha/2}\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}\right]. \tag{16}$$

Bootstrap percentile interval (P). The empirical percentiles of the bootstrap distribution of $\widehat{WTP}$ are used to obtain a $(1-\alpha)$-level confidence interval through the following algorithm:

1. let $\widehat{WTP}^{\star}_{[1]}, \ldots, \widehat{WTP}^{\star}_{[B]}$ be the ordered bootstrap replicates of $\widehat{WTP}$;
2. calculate $L = (B+1)\alpha/2$ and $U = (B+1)(1-\alpha/2)$, round them to the nearest integers and build the confidence interval for WTP as:

$$\left[\widehat{WTP}^{\star}_{[L]} \quad ; \quad \widehat{WTP}^{\star}_{[U]}\right]. \tag{17}$$

The rationale for this interval, which is then pushed forward to get bias corrected and bias corrected, accelerated confidence intervals given below, is as follows. Let $g(\cdot)$ be a monotonically increasing function, and write $\phi = g(WTP)$, $\hat{\phi} = g(\widehat{WTP})$ and $\hat{\phi}^{\star} = g(\widehat{WTP}^{\star})$. Choose $g(\cdot)$, such that

$$\hat{\phi} - \phi \sim \hat{\phi}^{\star} - \hat{\phi} \sim N(0, \sigma^2) \tag{18}$$

delivering the following $(1-\alpha)$-level confidence interval for WTP:

$$\left[g^{-1}(\hat{\phi} - \sigma z_{\alpha/2}) \quad ; \quad g^{-1}(\hat{\phi} + \sigma z_{\alpha/2})\right]. \tag{19}$$

However, (18) implies that $\hat{\phi} - \sigma z_{\alpha/2} = F^{-1}_{\hat{\phi}^\star}(\alpha/2)$ and $\hat{\phi} + \sigma z_{\alpha/2} = F^{-1}_{\hat{\phi}^\star}(1-\alpha/2)$, with $F^{-1}_{\hat{\phi}^\star}(\cdot)$ being the inverse of the cumulative distribution of $\hat{\phi}^\star$. Further, since g is monotonically increasing $F^{-1}_{\hat{\phi}^\star}(\alpha/2) = g(F^{-1}_{\widehat{WTP}^\star}(\alpha/2))$ and analogously for $F^{-1}_{\hat{\phi}^\star}(1-\alpha/2)$ where $F^{-1}_{\widehat{WTP}^\star}$ is the bootstrap inverse cumulative distribution of $\widehat{WTP}^\star$. Interval (19) then becomes

$$\left[F^{-1}_{\widehat{WTP}^\star}(\alpha/2) \quad ; \quad F^{-1}_{\widehat{WTP}^\star}(1-\alpha/2)\right], \tag{20}$$

which is exactly the interval in (17).

The simplicity of the percentile method is particularly appealing: neither the estimate of $\text{var}(\widehat{WTP})$ (unlike the bootstrap-t) nor the specification of the function g are required. Further, no invalid parameter values can be included in the interval, an important advantage over the pivotal methods. Unfortunately, the coverage error is often substantial if the distribution of $\widehat{WTP}$ is not nearly symmetric. The reason is that the justification of the method rests on the existence of a $g(\cdot)$ such that (18) holds, and for many problems such a g does not exist.

? proposes a Monte Carlo study to compare the performance of percentile method (based on both a non-parametric bootstrap sample and a parametric Krinsky and Robb bootstrap sample) with that of the Delta method, in computing confidence intervals for WTP.

Bias-corrected bootstrap percentile interval (BC). This method tries to improve over the bootstrap percentile intervals, relaxing the assumption that $\widehat{WTP}$ is symmetric around WTP and considers a monotonically increasing function $g(\cdot)$, chosen in such a way that

$$\hat{\phi} - \phi \sim \hat{\phi}^\star - \hat{\phi} \sim N(-c\sigma, \sigma^2), \tag{21}$$

for some constant c. An analogous argument than that used in the case of the percentile interval determines the interval

$$\left[F^{-1}_{\widehat{WTP}^\star}\left(\Phi(2c - z_{\alpha/2})\right) \quad ; \quad F^{-1}_{\widehat{WTP}^\star}\left(\Phi(2c + z_{\alpha/2})\right)\right], \tag{22}$$

which is a slightly more complex version of the interval in (20), with c being the bias-correction parameter. The parameter c can be estimated as

$$c = \Phi^{-1}\left(\frac{\#\{\widehat{WTP}^\star_{(b)} \leq \widehat{WTP}\}}{B+1}\right) \tag{23}$$

where $\Phi^{-1}(\cdot)$ is the standard-normal quantile function, and $\dfrac{\#\{\widehat{WTP}^\star_{(b)} \leq \widehat{WTP}\}}{B+1}$ is the proportion of bootstrap replicates at or below the original-sample estimate $\widehat{WTP}$ of WTP. If the bootstrap sampling distribution is symmetric, and if $\widehat{WTP}$ is unbiased, then this proportion will be close to 0.5, and the correction factor c will be close to 0, reducing the BC interval (22) to the percentile interval (20).

The algorithm to compute BC intervals is sketched below:

1. estimate c as in (23);
2. calculate $L = (B+1)\Phi(2c - z_{\alpha/2})$ and $U = (B+1)\Phi(2c + z_{\alpha/2})$ and build the confidence interval for WTP as in (17).

As already said, the BC interval represents an improvement over the percentile interval in non-symmetric problems. However, also the validity of this method depends upon the existence of a $g(\cdot)$ such that (21) holds, and for many problems such a g does not exist.

Bias-corrected, accelerated bootstrap percentile interval (BC$_a$). The BC$_a$ method allows not only for the lack of symmetry in the distribution of $\widehat{WTP}$, but also for the fact that its shape, or skewness, might change as WTP varies. It is defined in terms of two numerical parameters: the bias-correction c and the acceleration a. This time, the monotonically increasing function $g(\cdot)$ is chosen so that

$$\hat{\phi} - \phi \sim N(-c\sigma(\phi), \sigma^2(\phi)) \quad \text{and} \quad \hat{\phi}^\star - \hat{\phi} \sim N(-c\sigma(\hat{\phi}), \sigma^2(\hat{\phi})), \tag{24}$$

where $\sigma(x) = 1 + ax$. An analogous argument to that used for the percentile and BC interval gives the BC$_a$ interval

$$\left[F^{-1}_{\widehat{WTP}^\star}\left(\Phi\left(c + \frac{c - z_{\alpha/2}}{1 - a(c - z_{\alpha/2})}\right)\right) \quad ; \quad F^{-1}_{\widehat{WTP}^\star}\left(\Phi\left(c + \frac{c + z_{\alpha/2}}{1 - a(c + z_{\alpha/2})}\right)\right)\right]. \tag{25}$$

The calculation of a depends on whether the simulation is non-parametric or parametric, and in the latter case, whether nuisance parameters are present. In our study we use a simple jackknife estimate of a obtained as:

$$a = \frac{\sum_{h=1}^{NT}(\widehat{WTP}_{(-h)} - \overline{WTP})^3}{6\left[\sum_{h=1}^{NT}(\widehat{WTP}_{(-h)} - \overline{WTP})^2\right]^{\frac{3}{2}}}, \tag{26}$$

where $\widehat{WTP}_{(-h)}$ represents the estimate of WTP produced when the h-th observation is deleted from the original sample (there are NT of these quantities) and $\overline{WTP}$ represents the average of the $\widehat{WTP}_{(-h)}$, that is $\overline{WTP} = \sum_{h=1}^{NT} \widehat{WTP}_{(-h)}/NT$.

The following algorithm can be used to compute the BC$_a$ confidence interval:

1. estimate c as in (23) and a as in (26);
2. calculate $L = (B{+}1)\Phi\left(c + \frac{c - z_{\alpha/2}}{1 - a(c - z_{\alpha/2})}\right)$ and $U = (B{+}1)\Phi\left(c + \frac{c + z_{\alpha/2}}{1 - a(c + z_{\alpha/2})}\right)$, with both L and U rounded to the nearest integer, and build the confidence interval for WTP as in (17).

When the correction factors $a = 0$ and $c = 0.5$, the BC$_a$ interval reduces to the percentile interval. In all other cases, the BC$_a$ method generally has a smaller coverage error than the percentile and BC intervals. However, coverage error of this method increases as $\alpha \to 0$ and confidence intervals should be considered cautiously when $\alpha < 0.025$ (**?**, p. 205, p. 231).

Test-inversion bootstrap method (TIB). The duality between confidence intervals and hypothesis testing means that, if $[WTP_L \; ; \; WTP_U]$ are the correct endpoints of the $(1 - \alpha)$-level interval and a bootstrap sample is drawn after setting $WTP = WTP_L$, then under some natural monotonicity conditions,

$$P\left(\widehat{WTP}^{\star} \geq \widehat{WTP} \mid WTP = WTP_L\right) = \alpha/2. \tag{27}$$

Similarly, if a resample is taken under $WTP = WTP_U$, then

$$P\left(\widehat{WTP}^{\star} \leq \widehat{WTP} \mid WTP = WTP_U\right) = \alpha/2. \tag{28}$$

Solving (27) and (28) with respect to WTP_L and WTP_U respectively, delivers an estimate for the $(1-\alpha)$-level confidence interval for WTP. Obviously, we need to be able to simulate from the bootstrap distribution at different values of WTP and this is clearly possible only within a parametric resampling scheme. Suppose, for example, that the current estimate of the upper bound is $\widehat{WTP}_U^{(g)}$. Then a bootstrap sample can be obtained according to the parametric resampling scheme described in Section 3.3.1, where now the utility function is computed as $\hat{U}^{\star} = V_{\text{con}} + e^{\star}$, with V_{con} being the utility function in the WTP space as given in (11), with $WTP^{(g)}$ replaced by $\widehat{WTP}_U^{(g)}$. In addition to being able to sample at different values of WTP, a stochastic root finding algorithm is needed to solve (27) and (28). Among the various algorithm proposed in the literature at this purpose, the Robbins-Monro algorithm seems to be the most efficient one (**??**). Let $g = 1$ and $\widehat{WTP}_U^{(g)}$ be an initial reasonable estimate of $WTP_U^{(g)}$. Then, the Robbins-Monro algorithm can be described as follows:

1. generate a bootstrap sample with WTP set equal to $\widehat{WTP}_U^{(g)}$ and let $\widehat{WTP}^{(g)}$ be the estimate of WTP from this sample;

2. set
$$\begin{cases} \widehat{WTP}_U^{(g+1)} = \widehat{WTP}_U^{(g)} - \ell\dfrac{\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} > \widehat{WTP} \\ \widehat{WTP}_U^{(g+1)} = \widehat{WTP}_U^{(g)} + \ell\dfrac{1-\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} \leq \widehat{WTP} \end{cases},$$
   where ℓ is the step length constant.

Each step is expected to reduce the distance from WTP_U and the algorithm is iterated a predetermined number G of times, so that $\widehat{WTP}_U^{(G)}$ is taken as an estimate of WTP_U.

Obviously, an independent search is needed for WTP_L. If $\widehat{WTP}_L^{(g)}$ is the estimate of WTP_L after g steps of the algorithm, then $\widehat{WTP}_L^{(g+1)}$ is found as:

$$\begin{cases} \widehat{WTP}_L^{(g+1)} = \widehat{WTP}_L^{(g)} + \ell\dfrac{\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} < \widehat{WTP} \\ \widehat{WTP}_L^{(g+1)} = \widehat{WTP}_L^{(g)} - \ell\dfrac{1-\alpha/2}{g} & \text{if } \widehat{WTP}^{(g)} \geq \widehat{WTP} \end{cases}.$$

Details about choosing the step length constant ℓ, the starting value estimates for WTP_L and WTP_U, and the stopping rule can be found in **?**.

Studentized test-inversion bootstrap method (STIB). This method is aimed at reducing the coverage error of the TIB confidence interval by replacing $\widehat{WTP}$ in (27) and (28) with a Studentized statistic. Then if $[WTP_L\ ;\ WTP_U]$ are the correct endpoints of the $(1-\alpha)$-level interval and a bootstrap sample is drawn after setting $WTP = WTP_L$, we have

$$P\left(\frac{\widehat{WTP}^{\star} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}} \geq \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}} \mid WTP = WTP_L\right) = \alpha/2$$

where the variances are estimated using, for example, the Delta method. Similarly, if a resample is taken under $WTP = WTP_U$, then

$$P\left(\frac{\widehat{WTP}^{\star} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{\star})}} \leq \frac{\widehat{WTP} - WTP}{\sqrt{\mathrm{var}(\widehat{WTP})}} \mid WTP = WTP_U\right) = \alpha/2.$$

The same algorithm used to estimate TIB confidence interval can be used for STIB confidence intervals, where now the estimates of WTP_L and WTP_U are updated at each step g respectively as:

$$\begin{cases} \widehat{WTP}_L^{(g+1)} = \widehat{WTP}_L^{(g)} + \ell\dfrac{\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} < \dfrac{\widehat{WTP} - \widehat{WTP}_L^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \\ \widehat{WTP}_L^{(g+1)} = \widehat{WTP}_L^{(g)} - \ell\dfrac{1-\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\mathrm{var}(\widehat{WTP}^{(g)})}} \geq \dfrac{\widehat{WTP} - \widehat{WTP}_L^{(g)}}{\sqrt{\mathrm{var}(\widehat{WTP})}} \end{cases}$$

and

$$\begin{cases} \widehat{WTP}_U^{(g+1)} = \widehat{WTP}_U^{(g)} - \ell\dfrac{\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\text{var}(\widehat{WTP}^{(g)})}} > \dfrac{\widehat{WTP} - \widehat{WTP}_U^{(g)}}{\sqrt{\text{var}(\widehat{WTP})}} \\ \widehat{WTP}_U^{(g+1)} = \widehat{WTP}_U^{(g)} + \ell\dfrac{1-\alpha/2}{g} & \text{if } \dfrac{\widehat{WTP}^{(g)} - \widehat{WTP}}{\sqrt{\text{var}(\widehat{WTP}^{(g)})}} \leq \dfrac{\widehat{WTP} - \widehat{WTP}_U^{(g)}}{\sqrt{\text{var}(\widehat{WTP})}} \end{cases}.$$

Both the TIB and the STIB methods share the advantages of the inversion test methods in Section 3.2 (no assumptions on the distribution of $\widehat{WTP}$, no discontinuity points, no invalid parameter values included in the intervals) as well as those of the bootstrap methods (no assumptions on the distribution of the test statistic). On the other hand, they are computationally demanding, since different searches are needed for the lower an the upper confidence limits, with a bootstrap sample being required at each search step, and they require careful monitoring to assess convergence to interval limits (**?**).

4 Simulation study

In this section, we compare the different approaches considered in Section 3 on the basis of some simulation studies. We construct a number of data sets which resemble as much as possible an actual choice experiment, on the same line as **?**. A number N of hypothetical subjects is faced by $T = 16$ different scenarios, each one presenting $J = 2$ alternatives characterized by three different attributes, denoted as X_1, X_2 and X_C, with X_C being the attribute 'cost'. As in **?**, X_1 and X_2 have two levels, coded as 1 and 2, while X_C has four different levels, coded as 1, 2, 3 and 4. Dropping all subscripts for simplicity, the observed difference in utility between the two alternatives is:

$$V_1 - V_2 = \beta_0 + \beta_1(X_{11} - X_{12}) + \beta_2(X_{21} - X_{22}) + \beta_C(X_{C1} - X_{C2}),$$

where the values of the parameters are opportunely set.

A single data set can be simulated by drawing, independently for each subject and for each scenario, a value for the error difference $\epsilon_1 - \epsilon_2$ from an appropriate distribution. If this value is less than the difference $V_1 - V_2$, than the first alternative is chosen and the choice variable y is set equal to 1 for the first alternative and to 0 for the second one. Otherwise, the second alternative is chosen.

To assess the performance of the various methods for computing confidence intervals, we particularly explore the case of correct and uncorrect model specification, as well as the case in which β_C approaches zero, determining values of WTP close to its discontinuity point. In all cases, we consider different sample sizes and different confidence levels for the intervals. For each model specification and each sample size, $M = 1000$ different data sets are generated. A MNL model is then fitted to each data set and its parameters are estimated via maximum likelihood. Then, WTP_1 and WTP_2 are calculated as well as the various confidence intervals. These M sample values of the confidence intervals are used to evaluate the coverage rates, the average interval length and the average interval shape attaint by the various methodologies. Let $\widehat{WTP}_L^{(m)}$ and $\widehat{WTP}_U^{(m)}$ represent, respectively, the lower and the upper limits of the confidence interval, calculated with a certain method, for the m-th Monte Carlo data set, for $m = 1, \ldots, M$. Coverage, length and shape of the intervals are evaluated as:

$$\text{Coverage} = \frac{1}{M}\sum_{m=1}^{M} I\left(\widehat{WTP}_L^{(m)} \leq WTP \leq \widehat{WTP}_U^{(m)}\right)$$
$$\text{Length} = \frac{1}{M}\sum_{m=1}^{M} \left(\widehat{WTP}_U^{(m)} - \widehat{WTP}_L^{(m)}\right)$$
$$\text{Shape} = \frac{1}{M}\sum_{m=1}^{M} \frac{\widehat{WTP}_U^{(m)} - WTP}{WTP - \widehat{WTP}_L^{(m)}}$$

where $I(\cdot)$ is the indicator function. In order to better understand the behaviour of the effective coverage, it might be interesting to evaluate also the left rejection probability (LRP) and the right rejection probability (RRP). For each confidence interval which fails to contain the true value WTP, it was checked whether WTP lies on its left or on its right side, in the following way:

$$\text{LRP} = \frac{1}{M}\sum_{m=1}^{M} I\left(WTP \leq \widehat{WTP}_L^{(m)}\right) \quad \text{and} \quad \text{RRP} = \frac{1}{M}\sum_{m=1}^{M} I\left(WTP \geq \widehat{WTP}_U^{(m)}\right).$$

Finally, we also give the Monte Carlo estimates of the confidence limits, derived by calculating the $(\alpha/2)\%$ and $(1-\alpha/2)\%$ percentiles of the M estimates $\widehat{WTP}^{(m)}$. Monte Carlo confidence intervals serve as a benchmark for the accuracy of the other methods. The results are illustrated in the following three sections.

4.1 Correct model specification

The M data sets are simulated from a MNL model, drawing the differences $\epsilon_1 - \epsilon_2$ from a logistic distribution. The scale parameter is set to 1 so that the variance of the error differences is $\pi^2/3$. The correct MNL model is then fitted to the data sets.

Tables 1 and 2 are obtained by letting $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_C = -1$. From these settings it follows that $WTP_1 = 1$ and $WTP_2 = 0.5$. Table 1 gives length, shape, LRP, RRP and coverage rate of the different confidence intervals at the 95% for a sample size $N = 10$. Some interesting results can be noticed.

First of all, with such a small sample size, the coverage rates for some of the intervals are significantly different from the expected levels. In all the tables presented, we choose to use the bold typeface for coverage rates below the nominal level as well as for LRPs and RRPs above the nominal level. Such departures from the nominal levels are considered more serious than coverage rates above the nominal level or LRPs and RRPs below the nominal level, denoted in italics in the tables. Looking at Table 1, most of the methods produce intervals for which the real value of WPT is above the upper limit $\widehat{WTP}_U^{(m)}$ with a frequency significantly higher than $\alpha/2$, i.e. intervals having a too high RRP. At the same time, for most of these methods, the real value of WPT is below the lower limit $\widehat{WTP}_L^{(m)}$ with frequency significantly lower than $\alpha/2$, giving intervals with a too small LRP. This means that most of the methods produce interval limits which are smaller than they should, underestimating the value of WTP. The problem is probably due to the fact that with such a small sample size, the distribution of $\widehat{WTP}$ is not symmetric, and probably shows positive asymmetry.

The asymmetry of $\widehat{WTP}$ can be noticed also looking at the value of the shape index for the Monte Carlo confidence interval which is slightly larger than one, meaning that the distance between the upper limit and the parameter is larger that the distance between the lower limit and the parameter. Methods which produce confidence intervals which are symmetric around the point estimate, such as the Delta method and the normal-theory interval (N) that have a shape index exactly equal to 1, are not able to capture such asymmetry. All the other intervals show a shape index slightly larger than one, reflecting the fact that they catch, at least partly, the positive asymmetry of $\widehat{WTP}$. Only for some of them, though, the asymmetry of the confidence interval translates into a correct coverage rate. Looking at Table 1, these are the intervals based on the likelihood ratio test (L) and the asymptotic t-test inversion (T), on the BC method and on the TIB method. On the contrary, the STIB method and the bootstrap-t interval (S) perform poorly. The reason is probably that both of them require the estimates of $\text{var}(\widehat{WTP}^{\star}_{(b)})$ at each bootstrap sample b, and with such a small sample size, these estimates might have a huge variability.

An other observation concerns the difference between the three sampling schemes. A clear predominance of a sampling scheme over the others cannot be noticed. Perhaps, the parametric sampling performs, overall, slightly better than the others, while the Krinsky and Robb scheme performs somehow worst, notably for the BC$_a$ and the bootstrap-t intervals. The performance of the parametric sampling scheme is probably due to the fact that we are simulating under the correct model specification.

We conclude with a look at the interval length. Obviously, for a given confidence level, shorter intervals are to be preferred as providing more precise information on the parameter. The likelihood ratio test and the asymptotic t-test inversion methods seem to produce confidence intervals with the correct coverage rate and with a length in line with the Monte Carlo golden standard. Moreover, as conjectured in **?**, confidence intervals based on likelihood ratio test inversion are usually contained in those based on t-test inversion, making the first method preferable to the second. A striking result is, on the contrary, the average length of the intervals produced with the Delta method and the bootstrap-t method under the Krinsky and Robb sampling. The standard deviations of the length (in brackets) as well as the observation of the single confidence intervals, reveal that the mean length is affected by some anomalously wide confidence intervals. A deeper analysis of the simulation output explained the problem. For some of the simulated Monte Carlo data sets, the estimates of the parameter β_C were not significantly different from 0, having large standard errors. This determined very large estimates of the variance of $\widehat{WTP}$, estimated through the Delta method, and thus uselessly wide confidence intervals. For these same data sets, the Krinsky and Robb scheme samples values of the parameters from a multivariate normal distribution with very large values of $\hat{\sigma}^2_{\hat{\beta}_C}$ in the variance-covariance matrix. Thus, in the Krinsky and Robb bootstrap sample there will be very few huge values of $\widehat{WTP}^\star_{(b)}$, with huge estimates of $\text{var}(\widehat{WTP}^\star_{(b)})$ (obtained in concurrence with those samples b with $\hat{\beta}^\star_{C(b)}$ close to 0) and a lot of nearly-0 values of $\widehat{WTP}^\star_{(b)}$, with very small estimates of $\text{var}(\widehat{WTP}^\star_{(b)})$ (obtained in concurrence with all those samples b with $\hat{\beta}^\star_{C(b)}$ very large positive or negative value). Therefore, the bootstrap percentiles $\hat{w}^\star_{\alpha/2}$ and $\hat{w}^\star_{1-\alpha/2}$ will have very distant values, producing huge confidence intervals.

A general worsening in the performance of the methods can be seen when moving from a 95% to a 99% confidence level (this last table is not given here for reason of space; it is available from the authors upon request, as all the other tables only mentioned in the text). This is quite natural since higher confidence levels are more sensitive to the exact nature of the tails of the sample distribution of $\widehat{WTP}$. Notably, the t-test inversion method and, to a less extent, the likelihood ratio test inversion method maintain a coverage rate, LRP and RRP non significantly different from the nominal level.

Increasing the sample size to $N = 25$ determines a sensible improvement in the coverage rates, at the 95% (see Table 2) as well as the 99% level (Table not reported here). In Table 2, some of the intervals still show coverage rates above the nominal level, which is though less worrying than having coverage rates below the nominal level. Also the LRP is in some cases significantly smaller than expected, revealing that some asymmetry is still present in the distribution of $\widehat{WTP}$ and not completely caught by some of the methods. The reduction in the asymmetry of $\widehat{WTP}$ when moving from $N = 10$ to $N = 25$ is clearly witnessed by the decrease in the values of the shape index for the Monte Carlo intervals. Finally, an obvious contraction in the average length of all the intervals can be observed, as well as the disappearing of the anomalous behavior sometimes observed for the intervals calculated through the Delta method and the bootstrap-t method, for $N = 10$.

		WTP_1					WTP_2				
	Method	Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
	Monte Carlo	0.9781	1.1045	0.0250	0.0250	0.9500	0.8651	1.1539	0.0250	0.0250	0.9500
	D	1.7486	1.0000	0.0300	**0.0410**	**0.9290**	1.3642	1.0000	*0.0150*	0.0330	0.9520
		(8.4380)	(0.0000)	(0.0054)	(0.0063)	(0.0081)	(5.3775)	(0.0000)	(0.0038)	(0.0056)	(0.0068)
	L	0.9295	1.1395	0.0350	0.0290	0.9360	0.8761	1.1350	0.0240	0.0310	0.9450
		(0.1353)	(0.2320)	(0.0058)	(0.0053)	(0.0077)	(0.1330)	(0.3319)	(0.0048)	(0.0055)	(0.0072)
	T	0.9674	1.1692	0.0313	0.0172	0.9520	0.9064	0.1608	0.0192	0.0262	0.9550
		(0.1174)	(0.1116)	(0.0055)	(0.0041)	(0.0068)	(0.1210)	(0.7869)	(0.0044)	(0.0051)	(0.0066)
Parametric sampling	B	1.0633	1.0446	0.0200	0.0260	0.9540	0.8824	0.9947	*0.0090*	**0.0530**	0.9380
		(0.1734)	(0.3637)	(0.0044)	(0.0050)	(0.0066)	(0.1410)	(0.1228)	(0.0030)	(0.0071)	(0.0076)
	S	0.8858	1.0987	0.0540	**0.0590**	**0.8870**	0.8500	1.3236	0.0280	0.0340	0.9380
		(0.1484)	(0.0140)	(0.0071)	(0.0075)	(0.0100)	(0.2238)	(5.5666)	(0.0052)	(0.0057)	(0.0076)
	N	1.0485	1.0000	0.0190	0.0250	0.9560	0.8895	1.0000	*0.0090*	**0.0480**	0.9430
		(0.1444)	(0.0000)	(0.0043)	(0.0049)	(0.0065)	(0.1305)	(0.0000)	(0.0030)	(0.0068)	(0.0073)
	P	1.0633	1.0275	0.0340	**0.0380**	**0.9280**	0.8824	1.0186	*0.0160*	**0.0440**	0.9400
		(0.1734)	(0.2488)	(0.0057)	(0.0060)	(0.0082)	(0.1410)	(0.1094)	(0.0040)	(0.0065)	(0.0075)
	BC	1.0792	1.0520	0.0280	0.0220	0.9500	0.8980	1.1733	*0.0140*	0.0320	0.9540
		(0.1832)	(0.2286)	(0.0052)	(0.0046)	(0.0069)	(0.1368)	(0.2724)	(0.0037)	(0.0056)	(0.0066)
	BC_a	1.0660	1.0084	0.0270	0.0310	0.9420	0.8913	1.1314	*0.0140*	0.0320	0.9540
		(0.1722)	(0.2347)	(0.0051)	(0.0055)	(0.0074)	(0.1373)	(0.2205)	(0.0037)	(0.0056)	(0.0066)
	TIB	1.0066	1.1847	0.0230	0.0110	0.9660	0.9672	1.3322	0.0190	0.0340	0.9470
		(0.1837)	(0.5528)	(0.0047)	(0.0033)	(0.0057)	(0.1867)	(0.6007)	(0.0043)	(0.0057)	(0.0071)
	STIB	0.9382	1.1134	**0.0420**	**0.0590**	**0.8990**	0.8942	1.1062	0.0230	**0.0490**	**0.9280**
		(0.1974)	(0.2873)	(0.0063)	(0.0075)	(0.0095)	(0.1847)	(0.2595)	(0.0047)	(0.0068)	(0.0082)
Non parametric sampling	B	1.0932	1.0584	0.0180	0.0250	0.9570	0.8982	0.9976	*0.0090*	**0.0480**	0.9430
		(0.1798)	(0.3767)	(0.0042)	(0.0049)	(0.0064)	(0.1500)	(0.1421)	(0.0030)	(0.0068)	(0.0073)
	S	0.8833	1.1020	**0.0550**	**0.0610**	**0.8840**	0.8384	1.1818	0.0270	**0.0470**	**0.9260**
		(0.1531)	(0.1460)	(0.0072)	(0.0075)	(0.0101)	(0.1447)	(1.3064)	(0.0051)	(0.0067)	(0.0083)
	N	1.0786	1.0000	0.0200	0.0240	0.9560	0.9061	1.0000	*0.0090*	**0.0460**	0.9450
		(0.1493)	(0.0000)	(0.0044)	(0.0048)	(0.0065)	(0.1382)	(0.0000)	(0.0030)	(0.0066)	(0.0072)
	P	1.0932	1.0129	0.0340	0.0360	**0.9300**	0.8982	1.0206	*0.0100*	**0.0430**	0.9470
		(0.1798)	(0.2458)	(0.0057)	(0.0059)	(0.0081)	(0.1500)	(0.1317)	(0.0031)	(0.0064)	(0.0071)
	BC	1.1115	1.0329	0.0220	0.0210	0.9570	0.9144	1.1775	*0.0120*	0.0310	0.9570
		(0.1901)	(0.2227)	(0.0046)	(0.0045)	(0.0064)	(0.1462)	(0.2804)	(0.0034)	(0.0055)	(0.0064)
	BC_a	1.0999	0.9878	0.0230	0.0350	0.9420	0.9082	1.1378	*0.0120*	0.0310	0.9570
		(0.1807)	(0.2279)	(0.0047)	(0.0058)	(0.0074)	(0.1472)	(0.2375)	(0.0034)	(0.0055)	(0.0064)
Krinsky and Robb sampling	B	0.9580	0.8392	0.0180	**0.0480**	**0.9340**	0.8974	0.8669	*0.0050*	**0.0380**	0.9570
		(0.1500)	(0.2214)	(0.0042)	(0.0068)	(0.0079)	(0.1491)	(0.2321)	(0.0022)	(0.0060)	(0.0064)
	S	108.2669	-28.6955	**0.0550**	**0.0410**	**0.9040**	61.6306	-37.4626	**0.0460**	**0.0400**	**0.9140**
		(1083.9918)	(386.3830)	(0.0072)	(0.0063)	(0.0093)	(675.9360)	(400.1996)	(0.0066)	(0.0062)	(0.0089)
	N	0.9574	1.0000	0.0260	0.0330	0.9410	0.8960	1.0000	*0.0110*	0.0360	0.9530
		(0.1399)	(0.0000)	(0.0050)	(0.0056)	(0.0075)	(0.1500)	(0.0000)	(0.0033)	(0.0059)	(0.0067)
	P	0.9580	1.1563	0.0350	0.0230	0.9420	0.8974	1.1128	0.0210	0.0350	0.9440
		(0.1500)	(0.2490)	(0.0058)	(0.0047)	(0.0074)	(0.1491)	(0.2075)	(0.0045)	(0.0058)	(0.0073)
	BC	0.9627	1.1501	0.0370	0.0220	0.9410	0.9008	1.1027	0.0250	0.0310	0.9440
		(0.1585)	(0.2828)	(0.0060)	(0.0046)	(0.0075)	(0.1468)	(0.2636)	(0.0049)	(0.0055)	(0.0073)
	Bca	0.9610	1.0885	**0.0390**	0.0320	**0.9290**	0.8999	1.0564	0.0210	0.0340	0.9450
		(0.1567)	(0.2847)	(0.0061)	(0.0056)	(0.0081)	(0.1458)	(0.2637)	(0.0045)	(0.0057)	(0.0072)

Table 1: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals (standard errors given in brackets). Model simulated: MNL model. Sample size: $N = 10$.

		WTP_1					WTP_2				
Method		Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
Monte Carlo		0.5540	1.0351	0.0250	0.0250	0.9500	0.5482	1.0831	0.0250	0.0250	0.9500
	D	0.5715	1.0000	0.0180	0.0270	0.9550	0.5402	1.0000	0.0180	0.0330	0.9490
		(0.0372)	(0.0000)	(0.0042)	(0.0051)	(0.0066)	(0.0419)	(0.0000)	(0.0042)	(0.0056)	(0.0070)
	L	0.5766	1.0852	0.0270	0.0310	0.9420	0.5444	1.0556	0.0230	0.0300	0.9470
		(0.0714)	(0.2369)	(0.0051)	(0.0055)	(0.0074)	(0.0552)	(0.1460)	(0.0047)	(0.0054)	(0.0071)
	T	0.5849	1.0913	0.0210	*0.0150*	*0.9640*	0.5526	1.0686	0.0210	0.0270	0.9520
		(0.0384)	(0.0332)	(0.0045)	(0.0038)	(0.0059)	(0.0430)	(0.0209)	(0.0045)	(0.0051)	(0.0068)
Parametric sampling	B	0.5802	0.9619	*0.0140*	0.0270	0.9590	0.5436	0.9714	*0.0150*	0.0320	0.9530
		(0.0412)	(0.0608)	(0.0037)	(0.0051)	(0.0063)	(0.0433)	(0.0509)	(0.0038)	(0.0056)	(0.0067)
	S	0.5808	1.0740	0.0210	0.0170	0.9620	0.5446	1.0821	0.0200	0.0280	0.9520
		(0.0375)	(0.0586)	(0.0045)	(0.0041)	(0.0152)	(0.0424)	(0.0530)	(0.0530)	(0.0052)	(0.0068)
	N	0.5807	1.0000	0.0170	0.0260	0.9570	0.5432	1.0000	0.0170	0.0310	0.9520
		(0.0373)	(0.0000)	(0.0041)	(0.0050)	(0.0064)	(0.0428)	(0.0000)	(0.0041)	(0.0055)	(0.0068)
	P	0.5802	1.0442	0.0210	0.0220	0.9570	0.5436	1.0322	0.0210	0.0310	0.9480
		(0.0412)	(0.0807)	(0.0045)	(0.0046)	(0.0064)	(0.0433)	(0.0524)	(0.0045)	(0.0055)	(0.0070)
	BC	0.5815	1.0576	0.0190	0.0190	0.9620	0.5452	1.0699	0.0230	0.0310	0.9460
		(0.0415)	(0.0879)	(0.0043)	(0.0043)	(0.0060)	(0.0432)	(0.0710)	(0.0047)	(0.0055)	(0.0071)
	Bca	0.5813	1.0339	*0.0160*	0.0200	*0.9640*	0.5450	1.0589	0.0210	0.0310	0.9480
		(0.0416)	(0.0892)	(0.0040)	(0.0044)	(0.0059)	(0.0432)	(0.0699)	(0.0045)	(0.0055)	(0.0070)
	TIB	0.5960	1.0449	0.0180	0.0310	0.9510	0.5599	1.0686	*0.0160*	0.0280	0.9560
		(0.0713)	(0.2242)	(0.0042)	(0.0055)	(0.0068)	(0.0704)	(0.2186)	(0.0040)	(0.0052)	(0.0065)
	STIB	0.5993	1.1106	*0.0160*	0.0190	*0.9650*	0.5707	1.1106	0.0220	0.0260	0.9520
		(0.0759)	(0.2340)	(0.0040)	(0.0043)	(0.0058)	(0.0735)	(0.2370)	(0.0046)	(0.0050)	(0.0068)
Non parametric sampling	B	0.5851	0.9610	*0.0150*	0.0260	0.9590	0.5467	0.9708	*0.0110*	0.0340	0.9550
		(0.0443)	(0.0647)	(0.0038)	(0.0050)	(0.0063)	(0.0443)	(0.0550)	(0.0033)	(0.0057)	(0.0066)
	S	0.5814	1.0753	0.0200	0.0170	*0.9630*	0.5448	1.0843	0.0240	0.0290	0.9470
		(0.0415)	(0.0572)	(0.0044)	(0.0041)	(0.0060)	(0.0441)	(0.0542)	(0.0048)	(0.0053)	(0.0071)
	N	0.5856	1.0000	*0.0160*	0.0210	*0.9630*	0.5463	1.0000	*0.0150*	0.0340	0.9510
		(0.0408)	(0.0000)	(0.0040)	(0.0045)	(0.0060)	(0.0436)	(0.0000)	(0.0038)	(0.0057)	(0.0068)
	P	0.5851	1.0455	0.0210	0.0200	0.9590	0.5467	1.0333	0.0190	0.0340	0.9470
		(0.0443)	(0.0798)	(0.0045)	(0.0044)	(0.0063)	(0.0443)	(0.0565)	(0.0043)	(0.0057)	(0.0071)
	BC	0.5862	1.0575	0.0200	0.0180	0.9620	0.5483	1.0764	0.0170	0.0300	0.9530
		(0.0448)	(0.0911)	(0.0044)	(0.0042)	(0.0060)	(0.0443)	(0.0759)	(0.0041)	(0.0054)	(0.0067)
	Bca	0.5860	1.0347	0.0200	0.0200	0.9600	0.5482	1.0656	*0.0160*	0.0300	0.9540
		(0.0445)	(0.0923)	(0.0044)	(0.0044)	(0.0062)	(0.0444)	(0.0738)	(0.0040)	(0.0054)	(0.0066)
Krinsky and Robb sampling	B	0.5849	0.9166	*0.0140*	0.0330	0.9530	0.5517	0.9333	*0.0110*	0.0330	0.9560
		(0.0409)	(0.0503)	(0.0037)	(0.0056)	(0.0067)	(0.0451)	(0.0462)	(0.0033)	(0.0056)	(0.0065)
	S	0.5595	1.0898	0.0250	0.0240	0.9510	0.5282	1.0651	0.0220	0.0330	0.9450
		(0.0384)	(0.0553)	(0.0049)	(0.0048)	(0.0068)	(0.0432)	(0.0501)	(0.0046)	(0.0056)	(0.0072)
	N	0.5825	1.0000	*0.0160*	0.0240	0.9600	0.5495	1.0000	*0.0160*	0.0280	0.9560
		(0.0397)	(0.0000)	(0.0040)	(0.0048)	(0.0062)	(0.0439)	(0.0000)	(0.0040)	(0.0052)	(0.0065)
	P	0.5849	1.0943	0.0220	*0.0160*	0.9620	0.5517	1.0741	0.0210	0.0260	0.9530
		(0.0409)	(0.0600)	(0.0046)	(0.0040)	(0.0060)	(0.0451)	(0.0535)	(0.0045)	(0.0050)	(0.0067)
	BC	0.5862	1.0887	0.0210	0.0190	0.9600	0.5530	1.0705	0.0210	0.0270	0.9520
		(0.0410)	(0.0814)	(0.0045)	(0.0043)	(0.0062)	(0.0454)	(0.0733)	(0.0045)	(0.0051)	(0.0068)
	Bca	0.5854	1.0638	0.0190	0.0200	0.9610	0.5527	1.0593	0.0200	0.0270	0.9530
		(0.0409)	(0.0755)	(0.0043)	(0.0044)	(0.0061)	(0.0453)	(0.0719)	(0.0044)	(0.0051)	(0.0067)

Table 2: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals (standard error given in brackets). Model simulated: MNL model. Sample size: $N = 25$.

4.2 Uncorrect model specification

We now assume some form of heteroscedasticity in the data. In particular, we consider an hypothetical population made up of two equal sized groups, such that unobserved factors have different variance for decision makers in the first and in the second group. We accomplished this by letting the scale parameter of the first group being equal to 1, that is the variance of the error term equal to $\pi^2/6$, and the scale parameter of the second group being equal to σ^2, corresponding to a variance of the error term equal to $\sigma^2 \times \pi^2/6$. Different simulations were performed setting $\sigma^2 = 2^2$ and $\sigma^2 = 4^2$, with $N = 10$ and $N = 25$. Obviously, the scale parameter does not affect the ratio of any two coefficients, since it drops out in the ratio, so that WTP and other measures of marginal rates of substitution are not affected. Only the magnitudes of all coefficients is affected. We estimated a MNL model on the simulated data set *without* taking into account such a form of heterogeneity in the data, in order to study the performance of the different methods for confidence intervals in case of model mispecification. The heterogeneity in the data should not affect the correctness of the coefficient estimates, but will affect the correctness of their standard error estimates.

Tables 3 and 4 show that the confidence intervals for WTP are more skewed than those in Table 1 and the skewness increases with σ^2. Looking at Table 3 reveals that the LRP of most intervals are below $\alpha/2$ and, more worryingly, for some of them the RRP is significantly above $\alpha/2$, namely the Delta methods and all the bootstrap methods belonging to the pivotal family. When increasing the value of σ^2 (Table 4), also the method based on likelihood ratio test inversion fails to account for this skewness, determining a confidence interval for WTP_1 with a coverage rate significantly lower than the nominal level. On the other hand, the method based on the inversion of the t-test keeps showing good performances, as well as all the bootstrap methods belonging to the percentile family, no matter the sampling scheme used.

Some of the values were removed from Table 4, since they were too large and made it impossible to fit the table in the page, and were replaced by "–". These values are simply due to few anomalous bootstrap samples which produced huge values of $\widehat{WTP}$, with huge standard deviations.

Also in this case, increasing N (results are not reported) determines a reduction in the skewness of the distribution of $\widehat{WTP}$ and, thus, an amelioration in the performance of the various methods.

4.3 Cost coefficient approaching zero

The M data sets are again simulated from a MNL model, setting the scale parameter equal to 1 for all the decision makers and letting $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_C = -0.5$. From these settings it follows that $WTP_1 = 2$ and $WTP_2 = 1$.

Also in Table 5, some values, due to few anomalous bootstrap samples with huge $\widehat{WTP}$ and standard deviations, were too large and thus replaced by "–". The S and the STIB methods are not affected by these values, since they are both studentized and divide by $\mathrm{var}(\widehat{WTP}^{\star}_{(b)})$. The B and P methods are also not affected, since the huge values of $\widehat{WTP}$ are very few and are far right in the queue of the distribution.

The distribution of $\widehat{WTP}$ in this case is quite skewed, as can be seen from the shape index of the Monte Carlo interval. For the interval based on the Delta method, as well as for other Bootstrap based confidence intervals, this determines a LRP significantly lower and a RRP significantly higher than expected. Overall the coverage rate is not significantly different from $1 - \alpha$, but the examination of LRP and RRP reveals confidence intervals underestimating WTP. On the contrary, likelihood ratio test and t-test inversion methods, as well as all the methods belonging to the bootstrap percentile family, are characterized by reasonable LRP, RRP and coverage rate. Similar results (not reported here) were obtained at the 99% confidence level.

This time, increasing N (results not reported) does not determine any improvement: even if the shape index for the Monte Carlo interval is smaller than that in Table 5, the only intervals with correct LRP, RRP and coverage rate are still those based on likelihood ratio test and t-test inversion methods and those belonging to the bootstrap percentile family.

5 Real data application

In this section we compare WTP confidence intervals computed for two real data set, accordingly to the methods described in Section 3. In both cases we fitted a MNL model to the data. The comparison is driven by the results of the simulation study, and looks at the differences produced by the various methods in monetary terms, for easiness of interpretation. The empirical examples allow a more-in-depth analysis of the implications that arise in practice when choosing the method to construct WTP confidence intervals.

5.1 Data description

The first data set refers to a study conducted in five geographical areas of the Marche, a region in central Italy, for measuring and integrating service quality in local public transport contracts (?). The stated preference exercise was prepared by asking the interviewee to make repeated choices between three alternatives, one representing his current trip (status quo) and two hypothetical trips (different bundles of trip attribute level). We used five attributes as the most appropriate dimensions to characterize service quality from a user's perspective: Cost (bus fare); Delay (amount of delay at bus stop); Trip Length (bus travel time); Frequency (bus frequency as number of buses per hour); Availability (elapse of time between service inception and service closure). The attribute levels were selected as percentage changes from the status quo. A fractional factorial design was constructed, in which profiles for each respondent are formed using the least often previously used attribute levels for that respondent, subject to minimal overlap. Each one-way level frequency within attributes is balanced. Through this formal experimental design, the attribute levels were combined into bus options and we constructed 8 choice sets per interview. One of these choice sets had a control function, formed by three fixed-design alternatives: the best, the worst and the current one. To allow for a rich variation in the combination of attribute levels we use a blocking strategy, preparing 25 different versions of the survey form. Overall for the five geographical service segments, we administered 264 paper-and-pencil interviews either on board or at the bus stops associated with the main routes. We used a random sampling strategy to select the sample. For econometric analysis, we ignored all the interviews in which the agents failed to answer correctly the control choice exercise (14 out of 264).

The second data set refers to a study of airport choice in a multi-airport region with the intent of testing different methods to account for preference heterogeneity (?). Data acquisition was based on a stated preference choice experiment describing a potential choice situation among four regional airports. The interviews were carried out by trained university students as computer assisted personal interviews. Each interview included five hypothetical choice exercises in which respondents were asked to evaluate the four airports and choose the preferred one. The study area considered includes two regions in central Italy, Marche and Emilia-Romagna, and four airports: Ancona, Rimini, Forlì and Bologna which are all located within the same catchment area. A total of 1,419 interviews generating 6,839 observations were gathered at both the four airports and their respective catchment areas. A fractional factorial, full profile, experimental design with complete enumeration was planned. The structural variables used were: A_MIN (continuous variable measuring access time in minutes); P_AIRL (binary variable coded 1 when representing the preferred airline company and 0 otherwise); F_EURO (continuous variable measuring ticket cost in euros); NONSTOP (binary variable coded 1 when the flight is non-stop from origin to destination and 0 otherwise); BAL_M_AV (continuous variable measuring the absolute value of the difference between desired and actual departure time in minutes).

5.2 WTP confidence intervals

A first general consideration concerns the total number of observations. In both empirical examples, sample size is very large and thus all various methods produce very similar WTP confidence intervals. Table 6 reports the upper and lower bounds of the WTP interval for the different attributes considered in the local public transport data set. To facilitate interpretation we calculate the amount of money the interviewees are willing to pay for ameliorative conditions defined by using the appropriate attribute levels adopted in the experimental design. In particular, the WTP point estimates are: 0.18 euro for a 100% reduction in delay; 0.16 euro for a 50% reduction in trip length; 0.27 euro for a 50% increase in frequency; 0.40 euro for a 20% increase in availability. At a first glance, for each attribute, the confidence intervals estimated with the different methods seem quite small. However when turning to a policy perspective, having in mind that average cost of the ticket is about 1 euro, the interval lengths become substantial. In fact, on average, we have the following distances between upper and lower bound: 0.10 euro for delay; 0.10 euro for trip length; 0.13 euro for frequency; 0.16 euro for availability. When comparing the upper and lower bounds obtained under the different methods we find a slight variability (coefficient of variation equal to 1.3% and 2.5%, respectively). Consistently with the findings of the simulation study, the Delta method underestimates both lower and upper bounds producing intervals shifted to the left with respect to those obtained through the t-test inversion method and other reliable methods such as the inversion of the likelihood ratio test and the bootstrap methods of the percentile family P, BC, BCa. Such a shift is due, as already discussed in Section 4 to the fact that the Delta method, by construction, delivers symmetric confidence intervals. However, this is not a desirable characteristic, since the distribution of $\widehat{WTP}$ is often positively skewed, and this seems to hold true also for the data at hand, as witnessed by the confidence intervals obtained with the other methods, for which the shape index is always larger than one. On the basis of the simulation study, if we take the t-test inversion method as benchmark, we can evaluate how far are the intervals obtained through each method from this benchmark, for example by simply summing the distances, in absolute value, between the respective lower and upper bounds. In particular, the methods belonging to the bootstrap percentile family determine the confidence intervals which are the closest to the benchmark. The Delta method intervals are in the middle, between these ones and those which show the most relevant differences, namely the basic bootstrap, TIB and STIB intervals. Notwithstanding the relevance and pertinence of the considerations reported, from a purely policy perspective, these differences are less valuable. In fact, the highest difference is equals to 0.02 euro which represents a small difference in this context.

Table 7 reports the results for the airport choice data set. Here, the interviewees are willing to pay 75 euro to reduce by 60 minutes the access time; 14 euro for the presence of the preferred airline; 92 euro for a nonstop flight; 30 euro to reduce the difference between desired and actual departure times to 120 minutes. Similarly to the first example, the distances between upper and lower bound are, on average, nearly 15 euro which represents a 10% of the average cost. When comparing WTP confidence intervals with the benchmark, as in the first data set, the highest distance is found for TIB, STIB and B methods; while method D is now the one with the lowest distance (about 0.2 euro), which is negligible from a policy perspective. A possible explanation relates to the sample size that is larger than for the first example. Thus in this case the distribution of $\widehat{WTP}$ is closer to symmetry (as witnessed by all the shape indexes being close to one) and the D method performs as fine as methods producing not-necessarily symmetric intervals.

6 Conclusions

In this paper, we compared different methods to compute confidence intervals for the WTP parameter. We used Monte Carlo simulations to draw a large number M of data sets under different scenarios and used these data sets to assess the performances of the different methods. In particular, we considered, as possible scenarios, the case of correct model specification (data were generated under a MNL model), the case of uncorrect model specification (data were generated under a MNL model with two different scale parameter values in the sample) and the case of cost coefficient approaching zero. Our findings are briefly summarized below.

1. All the scenarios considered revealed a certain degree of skewness in the distribution

of the estimate of WTP, which should be translated into an asymmetric confidence interval. The Delta method and the bootstrap normal-theory method N produce, by construction, intervals which are symmetric around the point estimate of WTP. Thus, they fail to take into account the asymmetry in the sample distribution of $\widehat{WTP}$. Since in empirical situation the mean values of the intervals are generally greater than the point estimates, as stressed by **?**, symmetric intervals based on Delta method would undervalue how much the individually are actually willing to pay, if used in project evaluation.

2. The skewness of $\widehat{WTP}$ is more relevant in case of model mispecification and for values of the cost parameter approaching zero. It tends, instead, to decrease as the sample size increases, so that using symmetric confidence intervals becomes less misleading if the sample size is reasonable. As illustrated in **?**, however, very large sample sizes would be required to compensate for very small values of the cost parameter.

3. Bootstrap methods belonging to the pivotal family seem to be not too accurate, giving sometimes confidence intervals with coverage rates significantly lower than the nominal level. On the other side, bootstrap methods belonging to the percentile family, and in particular the BC and the BC$_a$ methods, proved to be more accurate and performed generally well in the situations considered. Both of the families, however, might be affected in their performances by values of the cost parameter close to zero. We could not observe this in our simulation, probably because we did not go further enough in decreasing the value of the cost parameter. Simulations in **?** seem to show, however, that smaller values of the cost parameters are necessary to make coverage rates of bootstrap methods break down, than those sufficient to make the Delta method coverage rate break down. Bootstrap inversion methods are not affected by small values of the cost parameters but show in some cases not completely satisfactory results, probably due to the fact that they require careful convergence monitoring.

4. The scenarios considered did not highlight relevant differences between the parametric, the non-parametric and the Krinsky and Robb sampling scheme. However, this last one, relies on the estimated standard errors of the coefficients, which are generally biased if the model is mispecified, and thus is somehow felt as less reliable than the other two, even if much less time consuming. In theory, the Krinsky and Robb sampling scheme may even determine confidence intervals not containing the WTP point estimate.

5. Methods based on test inversion, such as the likelihood ratio test and the t-test inversion methods have good performances, are not affected by small values of the cost parameter and are very simple and time-saving to calculate. Monte Carlo simulations also allow us to confirm the intuition in **?**, namely that confidence intervals based on likelihood ratio inversion are usually contained in those based on t-test inversion, making the first method preferable to the second. However, we found the method based on the likelihood ratio test inversion more sensitive to departures from homoscedasticity and in general, it might be supposed, to departure from correct model specification, making the other method preferable.

In summary, on the basis of the simulation study, we would suggest always using the t-test inversion method as producing confidence intervals not necessarily symmetric, not affected by small values of the cost parameter and having the correct coverage rate under all the scenarios considered. If we are sure that the model is correctly specified the method based on the inversion of the likelihood ratio test can be considered a valid alternative as producing, on average, shorter confidence intervals with the right coverage rate. Also bootstrap methods of the percentile family could be taken into account, provided the cost parameter is not too small, as producing, as a byproduct, the entire simulated distribution of $\widehat{WTP}$, which might be of interest in policy evaluation, at the cost of a greater computational time required for their implementation.

The conclusions drawn from the simulation study are supported by the results of the empirical application in that all the methods produce fairly similar confidence intervals, due to the large sample size. However, the Delta method, delivering a symmetric interval, produce a confidence interval which is slightly shifted on the left with respect to those produced by the likelihood-ratio test and *t*-test inversion method, as well as those produced by the bootstrap methods belonging to the percentile family. In any case, when translated into monetary terms, the difference between results produced through the various methods appear to be irrelevant, for the data set considered.

References

		WTP_1					WTP_2				
	Method	Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
	Monte Carlo	1,1435	1,1608	0,0250	0,0250	0,9500	1,1199	1,0379	0,0250	0,0250	0,9500
	D	1,1619	1,0000	*0,0090*	**0,0430**	0,9480	1,1014	1,0000	*0,0150*	0,0370	0,9480
		(0,2157)	(0,0000)	(0,0030)	(0,0064)	(0,0070)	(0,1795)	(0,0000)	(0,0038)	(0,0060)	(0,0070)
	L	8,7132	0,9290	*0,0150*	0,0180	*0,9670*	8,6720	0,8735	*0,0120*	*0,0160*	*0,9720*
		(9,2039)	(0,5010)	(0,0038)	(0,0042)	(0,0056)	(9,2362)	(0,4054)	(0,0034)	(0,0040)	(0,0052)
	T	1,2656	1,2960	*0,0140*	0,0210	*0,9650*	1,1909	0,3963	0,0210	0,0240	0,9550
		(0,2737)	(0,1542)	(0,0037)	(0,0045)	(0,0058)	(0,2173)	(2,2048)	(0,0045)	(0,0048)	(0,0066)
Parametric sampling	B	1,2380	0,8619	*0,0000*	**0,0390**	0,9610	1,1457	0,9056	*0,0040*	0,0330	*0,9630*
		(0,2669)	(0,1482)	(0,0000)	(0,0061)	(0,0061)	(0,2165)	(0,0890)	(0,0020)	(0,0056)	(0,0060)
	S	1,1538	1,2553	**0,0290**	**0,0390**	**0,9320**	1,0913	1,1486	0,0330	0,0370	**0,9300**
		(0,1986)	(0,1605)	(0,0053)	(0,0061)	(0,0080)	(0,1661)	(0,1161)	(0,0056)	(0,0060)	(0,0081)
	N	1,2357	1,0000	*0,0040*	0,0370	0,9590	1,1420	1,0000	*0,0100*	0,0350	0,9550
		(0,2851)	(0,0000)	(0,0020)	(0,0060)	(0,0063)	(0,2174)	(0,0000)	(0,0031)	(0,0058)	(0,0066)
	P	1,2380	1,1907	*0,0130*	0,0360	0,9510	1,1457	1,1155	0,0220	0,0300	0,9480
		(0,2669)	(0,1920)	(0,0036)	(0,0059)	(0,0068)	(0,2165)	(0,1165)	(0,0046)	(0,0054)	(0,0070)
	BC	1,2430	1,2179	*0,0140*	0,0320	0,9540	1,1506	1,1485	0,0220	0,0290	0,9490
		(0,2669)	(0,1916)	(0,0037)	(0,0056)	(0,0066)	(0,2149)	(0,1339)	(0,0046)	(0,0053)	(0,0070)
	Bc_a	1,2395	1,1806	*0,0130*	0,0350	0,9520	1,1487	1,1317	0,0210	0,0290	0,9500
		(0,2662)	(0,1931)	(0,0036)	(0,0058)	(0,0068)	(0,2137)	(0,1261)	(0,0045)	(0,0053)	(0,0069)
	TIB	1,2174	1,0806	*0,0100*	0,0290	0,9610	1,1243	1,1298	*0,0100*	*0,0120*	*0,9780*
		(0,2506)	(0,2534)	(0,0031)	(0,0053)	(0,0061)	(0,2056)	(0,1324)	(0,0031)	(0,0034)	(0,0046)
	STIB	1,2746	1,2678	0,0200	0,0310	0,9490	1,1984	1,1134	0,0200	0,0180	0,9620
		(0,3256)	(0,3334)	(0,0044)	(0,0055)	(0,0070)	(0,2214)	(0,1678)	(0,0044)	(0,0042)	(0,0060)
Non parametric sampling	B	1,2591	0,8690	*0,0000*	**0,0400**	0,9600	1,1601	0,9090	*0,0030*	0,0320	*0,9650*
		(0,2767)	(0,1608)	(0,0000)	(0,0062)	(0,0062)	(0,2207)	(0,1026)	(0,0017)	(0,0056)	(0,0058)
	S	1,1483	1,2658	0,0280	**0,0410**	**0,9310**	1,0872	1,1492	0,0300	**0,0420**	**0,9280**
		(0,2025)	(0,1619)	(0,0052)	(0,0063)	(0,0080)	(0,1679)	(0,1159)	(0,0054)	(0,0063)	(0,0082)
	N	1,3125	1,0000	*0,0010*	**0,0380**	0,9610	1,1644	1,0000	*0,0070*	0,0320	0,9610
		(1,5435)	(0,0000)	(0,0010)	(0,0060)	(0,0061)	(0,3012)	(0,0000)	(0,0026)	(0,0056)	(0,0061)
	P	1,2591	1,1871	*0,0150*	0,0340	0,9510	1,1601	1,1146	0,0200	0,0300	0,9500
		(0,2767)	(0,2139)	(0,0038)	(0,0057)	(0,0068)	(0,2207)	(0,1313)	(0,0044)	(0,0054)	(0,0069)
	BC	1,2649	1,2199	*0,0160*	0,0290	0,9550	1,1658	1,1515	0,0210	0,0300	0,9490
		(0,2782)	(0,2156)	(0,0040)	(0,0053)	(0,0066)	(0,2200)	(0,1533)	(0,0045)	(0,0054)	(0,0070)
	Bc_a	1,2610	1,1821	*0,0160*	0,0310	0,9530	1,1643	1,1342	0,0190	0,0300	0,9510
		(0,2769)	(0,2160)	(0,0040)	(0,0055)	(0,0067)	(0,2195)	(0,1451)	(0,0043)	(0,0054)	(0,0068)
Krinsky and Robb sampling	B	1,2687	0,7785	*0,0000*	**0,0460**	0,9540	1,1901	0,8685	*0,0030*	0,0230	*0,9740*
		(0,2808)	(0,0916)	(0,0000)	(0,0066)	(0,0066)	(0,2236)	(0,0901)	(0,0017)	(0,0047)	(0,0050)
	S	1,0928	1,2797	0,0330	**0,0430**	**0,9240**	1,0286	1,1502	**0,0430**	**0,0490**	**0,9080**
		(0,1983)	(0,1513)	(0,0056)	(0,0064)	(0,0084)	(0,1596)	(0,1151)	(0,0064)	(0,0068)	(0,0091)
	N	1,2708	1,0000	*0,0010*	**0,0390**	0,9600	1,1850	1,0000	*0,0090*	0,0240	*0,9670*
		(0,3719)	(0,0000)	(0,0010)	(0,0061)	(0,0062)	(0,2450)	(0,0000)	(0,0030)	(0,0048)	(0,0056)
	P	1,2687	1,3044	0,0170	0,0240	0,9590	1,1901	1,1647	0,0210	0,0240	0,9550
		(0,2808)	(0,1750)	(0,0041)	(0,0048)	(0,0063)	(0,2236)	(0,1310)	(0,0045)	(0,0048)	(0,0066)
	BC	1,2661	1,2608	*0,0160*	0,0250	0,9590	1,1915	1,1451	0,0190	0,0220	0,9590
		(0,2788)	(0,1797)	(0,0040)	(0,0049)	(0,0063)	(0,2229)	(0,1362)	(0,0043)	(0,0046)	(0,0063)
	Bc_a	0,7541	0,3579	*0,0130*	0,0310	0,9560	0,7092	0,2925	0,0130	0,0290	0,9580
		(0,6498)	(1,1163)	(0,0036)	(0,0055)	(0,0065)	(0,6012)	(1,0628)	(0,0036)	(0,0053)	(0,0063)

Table 3: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals (standard error given in brackets). Model simulated: heteroscedastic model ($\sigma^2 = 2^2$). Sample size: $N = 10$.

		WTP_1					WTP_2				
Method		Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
Monte Carlo		1.5186	1.4656	0.0250	0.0250	0.9500	1.3855	1.1408	0.0250	0.0250	0.9500
	D	1.5477	1.0000	*0.0030*	**0.0440**	0.9530	1.4422	1.0000	*0.0120*	0.0340	0.9540
		(0.4631)	(0.0000)	(0.0017)	(0.0065)	(0.0067)	(0.3547)	(0.0000)	(0.0034)	(0.0057)	(0.0066)
	L	1.7777	1.4978	0.0260	**0.0400**	**0.9340**	1.6253	1.2458	0.0260	0.0310	0.9430
		(0.8723)	(0.5970)	(0.0050)	(0.0062)	(0.0079)	(0.8170)	(0.3758)	(0.0050)	(0.0055)	(0.0073)
	T	1.8100	1.4937	0.0210	0.0280	0.9510	1.6526	0.2998	0.0250	0.0280	0.9470
		(0.8008)	(0.4286)	(0.0045)	(0.0052)	(0.0068)	(0.6176)	(9.9119)	(0.0049)	(0.0052)	(0.0071)
Parametric sampling	B	1.7572	0.7412	*0.0000*	**0.0460**	0.9540	1.5928	0.8470	*0.0010*	0.0230	*0.9760*
		(0.8625)	(0.1447)	(0.0000)	(0.0066)	(0.0066)	(0.6099)	(0.1390)	(0.0010)	(0.0047)	(0.0048)
	S	1.4907	1.4354	**0.0550**	**0.0380**	**0.9070**	1.3788	1.2153	**0.0470**	**0.0460**	**0.9070**
		(0.4200)	(0.3348)	(0.0072)	(0.0060)	(0.0092)	(0.2968)	(0.2372)	(0.0067)	(0.0066)	(0.0092)
	N	–	1.0000	*0.0000*	**0.0380**	*0.9620*	–	1.0000	*0.0040*	0.0300	0.9660
		(–)	(0.0000)	(0.0000)	(0.0060)	(0.0060)	(–)	(0.0000)	(0.0020)	(0.0054)	(0.0057)
	P	1.7572	1.4221	0.0230	0.0350	0.9420	1.5928	1.2231	0.0250	0.0350	0.9400
		(0.8625)	(0.4147)	(0.0047)	(0.0058)	(0.0074)	(0.6099)	(0.2869)	(0.0049)	(0.0058)	(0.0075)
	BC	1.7798	1.4558	0.0210	0.0340	0.9450	1.5998	1.2346	0.0250	0.0340	0.9410
		(1.0925)	(0.5546)	(0.0045)	(0.0057)	(0.0072)	(0.6296)	(0.3124)	(0.0049)	(0.0057)	(0.0075)
	Bca	1.7745	1.4275	0.0210	0.0350	0.9440	1.5972	1.2184	0.0250	0.0340	0.9410
		(1.0934)	(0.5652)	(0.0045)	(0.0058)	(0.0073)	(0.6325)	(0.3110)	(0.0049)	(0.0057)	(0.0075)
	TIB	–	–	*0.0040*	**0.0410**	0.9550	–	–	**0.0130**	0.0350	0.9520
		(–)	(–)	(0.0020)	(0.0063)	(0.0066)	(–)	(–)	(0.0036)	(0.0058)	(0.0068)
	STIB	1.7565	1.4517	0.0340	0.0320	**0.9340**	1.6371	1.2336	0.0300	0.0310	0.9390
		(0.6541)	(0.4719)	(0.0057)	(0.0056)	(0.0079)	(0.5257)	(0.3637)	(0.0054)	(0.0055)	(0.0076)
Non parametric sampling	B	1.8005	0.7354	*0.0000*	**0.0440**	0.9560	1.6102	0.8527	*0.0020*	0.0230	*0.9750*
		(1.0420)	(0.1510)	(0.0000)	(0.0065)	(0.0065)	(0.6153)	(0.1491)	(0.0014)	(0.0047)	(0.0049)
	S	1.4820	1.4380	**0.0570**	**0.0370**	**0.9060**	1.3733	1.2183	**0.0490**	**0.0500**	**0.9010**
		(0.4176)	(0.3354)	(0.0073)	(0.0060)	(0.0092)	(0.2951)	(0.2330)	(0.0068)	(0.0069)	(0.0094)
	N	6.3433	1.0000	*0.0000*	0.0360	*0.9640*	4.9632	1.0000	*0.0030*	0.0290	*0.9680*
		(75.2829)	(0.0000)	(0.0000)	(0.0059)	(0.0059)	(57.4517)	(0.0000)	(0.0017)	(0.0053)	(0.0056)
	P	1.8005	1.4400	0.0210	0.0330	0.9460	1.6102	1.2197	0.0260	0.0310	0.9430
		(1.0420)	(0.4282)	(0.0045)	(0.0056)	(0.0071)	(0.6153)	(0.2934)	(0.0050)	(0.0055)	(0.0073)
	BC	1.8149	1.4740	0.0220	0.0320	0.9460	1.6178	1.2284	0.0240	0.0330	0.9430
		(1.0556)	(0.5390)	(0.0046)	(0.0056)	(0.0071)	(0.6311)	(0.3167)	(0.0048)	(0.0056)	(0.0073)
	BC$_a$	1.8110	1.4451	0.0220	0.0330	0.9450	1.6146	1.2126	0.0240	0.0330	0.9430
		(1.0875)	(0.5629)	(0.0046)	(0.0056)	(0.0072)	(0.6307)	(0.3142)	(0.0048)	(0.0056)	(0.0073)
Krinsky and Robb sampling	B	1.8239	0.7010	*0.0000*	**0.0430**	0.9570	1.6590	0.8267	*0.0000*	*0.0170*	*0.9830*
		(0.9442)	(0.1347)	(0.0000)	(0.0064)	(0.0064)	(0.6219)	(0.1445)	(0.0000)	(0.0041)	(0.0041)
	S	1.4270	1.4329	**0.0590**	**0.0450**	**0.8960**	1.3123	1.2163	**0.0490**	**0.0530**	**0.8980**
		(0.4202)	(0.3173)	(0.0075)	(0.0066)	(0.0097)	(0.2929)	(0.2299)	(0.0068)	(0.0071)	(0.0096)
	N	6.5380	1.0000	*0.0000*	0.0340	*0.9660*	4.6809	1.0000	*0.0020*	0.0260	*0.9720*
		(51.9470)	(0.0000)	(0.0000)	(0.0057)	(0.0057)	(34.4329)	(0.0000)	(0.0014)	(0.0050)	(0.0052)
	P	1.8239	1.4984	0.0230	0.0310	0.9460	1.6590	1.2563	0.0250	0.0260	0.9490
		(0.9442)	(0.4059)	(0.0047)	(0.0055)	(0.0071)	(0.6219)	(0.2917)	(0.0049)	(0.0050)	(0.0070)
	BC	1.8233	1.4928	0.0220	0.0290	0.9490	1.6654	1.2550	0.0250	0.0290	0.9460
		(0.9055)	(0.4456)	(0.0046)	(0.0053)	(0.0070)	(0.6376)	(0.3196)	(0.0049)	(0.0053)	(0.0071)
	BC$_a$	1.8200	1.4627	0.0220	0.0300	0.9480	1.6622	1.2385	0.0230	0.0270	0.9500
		(0.9391)	(0.4653)	(0.0046)	(0.0054)	(0.0070)	(0.6382)	(0.3158)	(0.0047)	(0.0051)	(0.0069)

Table 4: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals (standard error given in brackets). Model simulated: heteroscedastic model ($\sigma^2 = 4^2$). Sample size: $N = 10$.

		WTP_1					WTP_2				
Method		Length	Shape	LRP	RRP	Coverage	Length	Shape	LRP	RRP	Coverage
Monte Carlo		2,4038	1,7780	0,0250	0,0250	0,9500	1,8588	1,4622	0,0250	0,0250	0,9500
	D	2,3178	1,0000	*0,0000*	**0,0510**	0,9490	1,8428	1,0000	*0,0040*	0,0300	0,9660
		(3,0024)	(0,0000)	(0,0000)	(0,0070)	(0,0070)	(1,7322)	(0,0000)	(0,0020)	(0,0054)	(0,0057)
	L	2,8793	2,0992	0,0310	0,0230	0,9460	2,2084	1,6822	0,0290	0,0200	0,9510
		(2,0577)	(0,8784)	(0,0055)	(0,0047)	(0,0071)	(1,4798)	(0,7872)	(0,0053)	(0,0044)	(0,0068)
	T	3,3293	2,2810	0,0241	0,0221	0,9510	2,4591	0,4249	0,0251	*0,0120*	0,9620
		(9,2766)	(3,3164)	(0,0049)	(0,0047)	(0,0068)	(6,2061)	(1,8161)	(0,0050)	(0,0035)	(0,0060)
Parametric sampling	B	3,2783	0,5379	*0,0000*	**0,0700**	0,9300	2,3832	0,6695	*0,0000*	0,0330	*0,9670*
		(7,7144)	(0,1520)	(0,0000)	(0,0081)	(0,0081)	(4,7720)	(0,1574)	(0,0000)	(0,0056)	(0,0056)
	S	2,3542	2,0644	**0,0820**	0,0360	**0,8820**	1,8075	1,6161	**0,0770**	0,0270	**0,8960**
		(5,9061)	(1,5462)	(0,0087)	(0,0059)	(0,0102)	(3,1211)	(1,1554)	(0,0084)	(0,0051)	(0,0097)
	N	–	1,0000	*0,0000*	**0,0410**	0,9590	–	1,0000	*0,0000*	0,0260	*0,9740*
		(–)	(0,0000)	(0,0000)	(0,0063)	(0,0063)	(–)	(0,0000)	(0,0000)	(0,0050)	(0,0050)
	P	3,2783	2,0651	0,0230	0,0280	0,9480	2,3832	1,6130	0,0280	0,0190	0,9530
		(7,7144)	(0,8449)	(0,0047)	(0,0052)	(0,0070)	(4,7720)	(0,5789)	(0,0052)	(0,0043)	(0,0067)
	BC	–	–	0,0280	0,0260	0,9460	–	–	0,0310	0,0190	0,9500
		(–)	(–)	(0,0052)	(0,0050)	(0,0071)	(–)	(–)	(0,0055)	(0,0043)	(0,0069)
	Bca	–	–	0,0300	0,0290	0,9410	–	–	0,0310	0,0190	0,9500
		(–)	(–)	(0,0054)	(0,0053)	(0,0075)	(–)	(–)	(0,0055)	(0,0043)	(0,0069)
	TIB	–	–	**0,0020**	**0,0530**	0,9450	–	–	*0,0060*	0,0310	*0,9630*
		(–)	(–)	(0,0014)	(0,0071)	(0,0072)	(–)	(–)	(0,0024)	(0,0055)	(0,0060)
	STIB	2,8019	2,0422	**0,0570**	0,0300	**0,9130**	2,2158	1,5644	**0,0450**	*0,0160*	0,9390
		(4,5397)	(0,7880)	(0,0073)	(0,0054)	(0,0089)	(3,7248)	(0,5708)	(0,0066)	(0,0040)	(0,0076)
Non parametric sampling	B	3,4508	0,5361	*0,0000*	**0,0710**	**0,9290**	2,4945	0,6677	*0,0000*	0,0280	*0,9720*
		(10,6034)	(0,1552)	(0,0000)	(0,0081)	(0,0081)	(6,5799)	(0,1646)	(0,0000)	(0,0052)	(0,0052)
	S	2,3240	2,0689	**0,0850**	0,0340	**0,8810**	1,7920	1,6171	**0,0750**	0,0290	**0,8960**
		(5,0822)	(1,3449)	(0,0088)	(0,0057)	(0,0102)	(2,7050)	(1,0276)	(0,0083)	(0,0053)	(0,0097)
	N	20,0046	1,0000	*0,0000*	**0,0400**	0,9600	12,1190	1,0000	*0,0010*	0,0230	*0,9760*
		(166,1420)	(0,0000)	(0,0000)	(0,0062)	(0,0062)	(99,1889)	(0,0000)	(0,0010)	(0,0047)	(0,0048)
	P	3,4508	2,0811	0,0200	0,0260	0,9540	2,4945	1,6265	0,0260	0,0180	0,9560
		(10,6034)	(0,8797)	(0,0044)	(0,0050)	(0,0066)	(6,5799)	(0,6009)	(0,0050)	(0,0042)	(0,0065)
	BC	5,6093	2,3963	0,0250	0,0260	0,9490	4,2091	1,9539	0,0280	*0,0160*	0,9560
		(70,0261)	(6,6081)	(0,0049)	(0,0050)	(0,0070)	(57,4945)	(8,7702)	(0,0052)	(0,0040)	(0,0065)
	BC_a	5,7454	2,4537	0,0260	0,0280	0,9460	4,2165	1,9597	0,0280	*0,0160*	0,9560
		(70,3736)	(6,7851)	(0,0050)	(0,0052)	(0,0071)	(57,4941)	(9,0044)	(0,0052)	(0,0040)	(0,0065)
Krinsky and Robb sampling	B	3,3315	0,5227	*0,0000*	**0,0730**	**0,9270**	2,4327	0,6466	*0,0000*	0,0300	*0,9700*
		(8,3729)	(0,1419)	(0,0000)	(0,0082)	(0,0082)	(4,4944)	(0,1521)	(0,0000)	(0,0054)	(0,0054)
	S	2,2942	2,0183	**0,0840**	**0,0380**	**0,8780**	1,7325	1,5997	**0,0770**	0,0340	**0,8890**
		(5,6635)	(1,3793)	(0,0088)	(0,0060)	(0,0103)	(3,0300)	(1,1135)	(0,0084)	(0,0057)	(0,0099)
	N	20,0132	1,0000	*0,0000*	**0,0380**	0,9620	8,4012	1,0000	*0,0000*	0,0200	*0,9800*
		(149,7747)	(0,0000)	(0,0000)	(0,0060)	(0,0060)	(45,0835)	(0,0000)	(0,0000)	(0,0044)	(0,0044)
	P	3,3315	2,1030	0,0210	*0,0240*	0,9550	2,4327	1,6615	0,0250	*0,0120*	*0,9630*
		(8,3729)	(0,8179)	(0,0045)	(0,0048)	(0,0066)	(4,4944)	(0,5571)	(0,0049)	(0,0034)	(0,0060)
	BC	7,6728	2,5308	0,0230	0,0250	0,9520	3,8845	1,8997	0,0270	*0,0140*	0,9590
		(141,4618)	(12,0342)	(0,0047)	(0,0049)	(0,0068)	(48,2656)	(6,8794)	(0,0051)	(0,0037)	(0,0063)
	BC_a	7,7294	2,5845	0,0240	0,0280	0,9480	3,8938	1,9041	0,0280	*0,0140*	0,9580
		(141,4576)	(12,5029)	(0,0048)	(0,0052)	(0,0070)	(48,2672)	(7,0942)	(0,0052)	(0,0037)	(0,0063)

Table 5: Length, shape, LRP, RRP and coverage of 95%-level confidence intervals (standard error given in brackets). Model simulated: MNL model with $\beta_C = -0.5$. Sample size: $N = 10$.

		Delay	Trip length	Frequency	Availability
	D	[−0.1140; −0.0658]	[−0.0213; −0.0115]	[0.2118; 0.3363]	[0.0020; 0.0030]
	L	[−0.1161; −0.0670]	[−0.0215; −0.0116]	[0.2057; 0.3412]	[0.0020; 0.0031]
	T	[−0.1154; −0.0668]	[−0.0216; −0.0117]	[0.2150; 0.3406]	[0.0021; 0.0031]
Parametric sampling	B	[−0.1111; −0.0623]	[−0.0213; −0.0113]	[0.2062; 0.3321]	[0.0020; 0.0030]
	S	[−0.1137; −0.0661]	[−0.0216; −0.0117]	[0.2153; 0.3396]	[0.0021; 0.0031]
	N	[−0.1140; −0.0658]	[−0.0213; −0.0116]	[0.2115; 0.3366]	[0.0020; 0.0030]
	P	[−0.1175; −0.0686]	[−0.0216; −0.0116]	[0.2160; 0.3419]	[0.0021; 0.0031]
	BC	[−0.1175; −0.0686]	[−0.0216; −0.0116]	[0.2143; 0.3383]	[0.0021; 0.0031]
	BC$_a$	[−0.1176; −0.0688]	[−0.0216; −0.0116]	[0.2143; 0.3383]	[0.0021; 0.0031]
	TIB	[−0.1155; −0.0638]	[−0.0210; −0.0130]	[0.2038; 0.3349]	[0.0022; 0.0030]
	STIB	[−0.1184; −0.0650]	[−0.0206; −0.0122]	[0.2009; 0.3430]	[0.0021; 0.0032]
Non parametric sampling	B	[−0.1127; −0.0651]	[−0.0214; −0.0113]	[0.2060; 0.3327]	[0.0020; 0.0030]
	S	[−0.1153; −0.0683]	[−0.0217; −0.0115]	[0.2163; 0.3400]	[0.0021; 0.0031]
	N	[−0.1136; −0.0662]	[−0.0214; −0.0115]	[0.2114; 0.3367]	[0.0020; 0.0030]
	P	[−0.1147; −0.0671]	[−0.0215; −0.0115]	[0.2154; 0.3421]	[0.0021; 0.0031]
	BC	[−0.1139; −0.0664]	[−0.0216; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
	BC$_a$	[−0.1140; −0.0665]	[−0.0215; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
Krinsky and Robb sampling	B	[−0.1127; −0.0651]	[−0.0214; −0.0113]	[0.2060; 0.3327]	[0.0020; 0.0030]
	S	[−0.1153; −0.0683]	[−0.0217; −0.0115]	[0.2163; 0.3400]	[0.0021; 0.0031]
	N	[−0.1136; −0.0662]	[−0.0214; −0.0115]	[0.2114; 0.3367]	[0.0020; 0.0030]
	P	[−0.1147; −0.0671]	[−0.0215; −0.0115]	[0.2154; 0.3421]	[0.0021; 0.0031]
	BC	[−0.1139; −0.0664]	[−0.0216; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]
	BC$_a$	[−0.1140; −0.0665]	[−0.0215; −0.0115]	[0.2150; 0.3414]	[0.0021; 0.0031]

Table 6: Local public transport data set: 95% confidence intervals of WPT for the various attributes of the service.

		A_MIN	P_AIRL	NONSTOP	BAL_M_AV
	D	$[-1,3565;-1,1588]$	$[7,2097;20,4778]$	$[84,0203;99,7497]$	$[-0,2847;-0,2233]$
	L	$[-1,3578;-1,1612]$	$[7,1250;20,4844]$	$[84,1506;99,6975]$	$[-0,2854;-0,2238]$
	T	$[-1,3594;-1,1614]$	$[7,2316;20,5176]$	$[84,2062;99,9617]$	$[-0,2853;-0,2239]$
Parametric sampling	B	$[-1,3499;-1,1544]$	$[6,8436;20,5904]$	$[83,5852;99,2945]$	$[-0,2844;-0,2230]$
	S	$[-1,3558;-1,1579]$	$[6,8724;20,7013]$	$[83,9460;99,6783]$	$[-0,2851;-0,2243]$
	N	$[-1,3567;-1,1586]$	$[7,0291;20,6585]$	$[83,8946;99,8753]$	$[-0,2845;-0,2235]$
	P	$[-1,3610;-1,1654]$	$[7,0972;20,8440]$	$[84,4755;100,1848]$	$[-0,2851;-0,2237]$
	BC	$[-1,3666;-1,1672]$	$[6,9333;20,8103]$	$[84,3460;100,1788]$	$[-0,2861;-0,2246]$
	BC$_a$	$[-1,3666;-1,1672]$	$[6,9333;20,8103]$	$[84,3460;100,1788]$	$[-0,2861;-0,2246]$
	TIB	$[-1,3415;-1,1905]$	$[5,8319;20,2506]$	$[81,9890;101,6963]$	$[-0,2873;-0,2204]$
	STIB	$[-1,3597;-1,1675]$	$[7,3181;22,1669]$	$[83,0714;99,3709]$	$[-0,2808;-0,2211]$
Non parametric sampling	B	$[-1,3585;-1,1540]$	$[7,1219;20,2241]$	$[83,9143;99,6879]$	$[-0,2845;-0,2224]$
	S	$[-1,3664;-1,1585]$	$[7,1595;20,3316]$	$[84,2423;100,0542]$	$[-0,2862;-0,2238]$
	N	$[-1,3599;-1,1554]$	$[7,2128;20,4747]$	$[83,9171;99,8528]$	$[-0,2857;-0,2224]$
	P	$[-1,3613;-1,1568]$	$[7,4635;20,5656]$	$[84,0821;99,8557]$	$[-0,2856;-0,2235]$
	BC	$[-1,3610;-1,1542]$	$[7,6215;20,7207]$	$[84,0345;99,8418]$	$[-0,2860;-0,2238]$
	BC$_a$	$[-1,3610;-1,1542]$	$[7,6215;20,6885]$	$[84,0345;99,8418]$	$[-0,2860;-0,2238]$
Krinsky and Robb sampling	B	$[-1,3585;-1,1540]$	$[7,1219;20,2241]$	$[83,9143;99,6879]$	$[-0,2845;-0,2224]$
	S	$[-1,3664;-1,1585]$	$[7,1595;20,3316]$	$[84,2423;100,0542]$	$[-0,2862;-0,2238]$
	N	$[-1,3599;-1,1554]$	$[7,2128;20,4747]$	$[83,9171;99,8528]$	$[-0,2857;-0,2224]$
	P	$[-1,3613;-1,1568]$	$[7,4635;20,5656]$	$[84,0821;99,8557]$	$[-0,2856;-0,2235]$
	BC	$[-1,3610;-1,1542]$	$[7,6215;20,7207]$	$[84,0345;99,8418]$	$[-0,2860;-0,2238]$
	BC$_a$	$[-1,3610;-1,1542]$	$[7,6215;20,6885]$	$[84,0345;99,8418]$	$[-0,2860;-0,2238]$

Table 7: Airport choice data set: 95% confidence intervals of WPT for the various structural variables considered.