The Use Of On-Site Samples In Marketing Research

Christopher D. Azevedo, (E-mail: cazevedo@cmsu1.cmsu.edu), Central Missouri State University

ABSTRACT

On-site sampling procedures are frequently used in demand estimation applications. This results in a sample that is both truncated and endogenously stratified, limiting the inferences that can be drawn from the sample. Shaw (1985) develops a likelihood function that corrects for these problems and uses a Monte Carlo simulation to verify that the likelihood function outperforms models that ignore the problems. However, the simulation is less than ideal. In this paper, I develop an additional method for correcting for the problems associated with on-site samples and conduct a more thorough simulation to examine the performance of the various estimation methods.

Introduction

The objective of statistical inference is to use information drawn from a sample to infer properties about the population from which the sample was drawn. The method of drawing the sample plays an important role in how the statistical inferences are reached as well as how reliable the inferences are. Of course, the best possible scenario is to use a sample that has the same properties as the population of interest, but this can be a challenging and costly objective to achieve.

Sampling is an important part of marketing research. Researchers are interested in a variety of questions such as the selling potential of a new product, how satisfied customers are with an existing product, or what attributes people would like to see in a potential new product. In the first case, the population of interest is likely to be the general population. In the second case, the population of interest is likely to be the population of people who use the product, while in the third case the population of interest is likely to be both the users of the product as well as non-users who might purchase the product if some attribute of the product were changed.

In each of these cases, the ideal situation would be to draw a large random sample from the population of interest and use that sample to generate inferences about the population. In reality, sampling can be quite costly, and it is often a challenge to actually draw from the population of interest, as is likely in the third case mentioned above.

Because of these high costs, researchers often use sampling strategies that are less than optimal. This process is referred to as convenience sampling. For example, if the selling potential of an improved product is of interest, then the relevant population includes both current users of the product as well as potential users who may decide to use the product because of the improvements. Sampling from both users and potential users is likely to be both costly and difficult to carry out in practice. A simpler procedure would be to simply sample from current users. The problem is that the sample is now not representative of the population of interest and any inferences drawn from the sample must be presented with the caveat that they may or may not be representative of the population. These types of convenience samples are often referred to as on-site samples or intercept samples. The phrase 'on-site' is used in the recreation demand literature because respondents are queried (intercepted) while they are on site engaging in recreation. However, the issue is relevant to other demand applications.

The use of on-site samples in demand estimation presents an interesting econometric problem for the researcher. The issue has been investigated in the recreation demand literature where on-site samples are used to investigate the demand for trips to a recreation site, but has obvious extensions to the area of marketing. For example, marketing researchers may be interested in investigating what factors are important in determining the number of trips taken to an amusement park in a season. A natural place to sample respondents is at the park itself. It would also apply to a situation where a product comes with a warranty registration card or other data gathering instrument. Purchasers of the product can choose to fill the card out and return it. In this case, only purchasers of the product would be sampled.

In each of these cases, the data drawn through these sampling procedures can be used to draw inferences about more general populations, but only if the problems associated with on-site samples are accounted for. Specifically, on-site sampling procedures generate samples that are both truncated and endogenously stratified.

For example, suppose a large retail mall is interested in knowing what draws customers to the mall and what changes they might make to draw more people. One possible method of answering this question would be to estimate a mall-trip demand model. The population of interest includes people who currently visit the mall as well as people who might visit the mall if certain changes were made. A cheap and convenient sampling process would be to position interviewers at the entrance to the mall and intercept people as they enter. The problem with this convenience sample is that non-visitors are truncated from the sample. No information is gathered from people who take zero trips. An additional problem is that the sample would be endogenously stratified. People who take a high number of trips to the mall have a higher probability of being sampled than people who take a low number of trips to the mall. The term endogenous stratification is used because the resulting sample is stratified by the endogenous variable, the number of trips taken to the mall. Of course, visitors are sampled randomly within each stratum. However, people who take two trips are twice as likely to be sampled as people who take one trip (but no more likely to be sampled than other two-trip takers). If these problems are not accounted for, then the parameter estimates of the demand model will be biased and any inferences drawn using the sample data will not apply to the population of interest.

This issue has been discussed mainly in the recreation demand literature. Although some authors have dealt with the problems of truncation and endogenous stratification individually, [Amemiya (1973), Manski and McFadden (1982)], Shaw (1985) was the first to account for both problems simultaneously. He recognized that the use of on-site data with likelihood functions designed for random population samples resulted in biased parameter estimates. He estimated a demand model using a likelihood function designed to simultaneously correct for both truncation and endogenous stratification. Shaw (1988) extended the model by developing a Poisson count data model to account for the discrete nature of demand data. Englin and Shonkwiler (1995) extended Shaw's Poisson count data model to the case of the negative binomial, while Laitila (1999) extended it to account for both site choice and trip frequency using on-site samples.

In both his 1985 dissertation work and his 1988 paper, Shaw used Monte Carlo (MC) simulations to evaluate the performance of the likelihood function designed for use with on-site samples against various other estimation procedures. The results of his simulations support his claim that the modification to the likelihood function results in better parameter estimates. However, there are several aspects of his simulation that are less than ideal and may result in a distorted picture of the effectiveness of his likelihood function.

Though Shaw's likelihood function appears to be capable of correcting for the problems created by using on-site samples, it can be a computationally difficult model to implement. In this paper, I develop an additional method of correcting for the problems created by using on-site samples that has the benefit of being computationally very easy to implement. I use a Monte Carlo simulation that is more general than that used by Shaw to test the performance of the various models.

The paper will be organized as follows. In the next section I will discuss the likelihood function developed by Shaw for use with data that is both truncated and endogenously stratified. Shaw's 1985 MC simulation will then be discussed. I will then describe a more general simulation procedure and compare the results to those obtained by Shaw.

SHAW'S LIKELIHOOD FUNCTION

Shaw begins the development of a likelihood function that accounts for truncation and endogenous stratification of on-site samples by specifying a demand model. The good analyzed in Shaw's model is trips to a recreation site, but the basic methodology is appropriate for any type of demand application. Respondent i is assumed to maximize utility, $U_i(y_i, Z_i)$, where y_i is the quantity of the good (recreation trips) and Z_i represents all other goods consumed by the respondent. Utility is maximized subject to the respondent's budget constraint and a boundary constraint that $y_i \geq 0$. Suppose that the solution takes the form

$$y_i^* = X_i\beta + u_i, \tag{1.1}$$

where X_i represents independent variables such as the price of the good (the cost of travel) and income, β represents the parameter vector to be estimated, and $u_i \sim N(0, \sigma^2)$. This model generates observable trip quantities, y_i, that take the form $y_i = y_i^*$ if $y_i^* > 0$ and $y_i = 0$ if $y_i^* \leq 0$.

When estimated with a population-wide random sample this is an example of a censored model and the Tobit likelihood is appropriate. However, the likelihood function must be modified when an on-site sample is used. Shaw notes that the on-site sample can be viewed as being a random sample from a population that is truncated and endogenously stratified. An important assumption he makes is that the probability of selecting an individual in an on-site sample is proportional to the number of visits the individual takes to the site.

The density function of an observation, y_i, given the independent variables, X_i, from this population can be written as

$$h(y_i \mid X_i, sampling\ rule) = \frac{y_i f(y_i \mid X_i)}{\int_0^\infty y_i f(y_i \mid X_i)\,dy}, \tag{1.2}$$

where the sampling rule is defined as the presence of endogenous stratification and truncation in the population. The resulting log-likelihood function takes the form

$$L = -n\ln\left(\sigma^2\sqrt{2\pi}\right) + \sum_{i=1}^{n}\ln y_i - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i - X_i\beta}{\sigma}\right)^2 - \sum_{i=1}^{n}\ln\left[d_i\Phi(d_i) + \phi(d_i)\right], \tag{1.3}$$

where $d_i = X_i\beta/\sigma$, $\Phi(\cdot)$ represents the standard normal cdf, and $\phi(\cdot)$ represents the standard normal pdf. Maximization of this likelihood function with respect to the parameters of the model (β and σ) provides estimates of the parameters of the demand model described in (1.1).

SHAW'S MONTE CARLO SIMULATION

To test the performance of the likelihood function shown in equation (1.3), Shaw designed a Monte Carlo (MC) simulation. Shaw sets up the simulation by assuming the following specification for the demand model

$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i\ i = 1, \ldots, N && (1.4)\\ y_i &= y_i^* && \text{if } y_i^* > 0\\ &= 0, && \text{otherwise} \end{aligned}$$

where x_{1i} are travel costs, x_{2i} are wages, x_{3i} are incomes, y_i^* are the desired number of trips to the recreation site, y_i are the observable number of trips, N is the sample size, and $u_i \sim N(0, \sigma^2)$.

The demand model (1.4) is used to generate data sets for different levels of σ^2 and correlation between wage, x_{2i}, and income, x_{3i}. The specification with $\sigma^2 = 25$ and $corr(x_{2i}, x_{3i}) = 0$ is representative of the results and will be discussed here. Shaw sets the parameters at the following values: $\beta_0 = 25$, $\beta_1 = -0.3$, $\beta_2 = -0.1$, and $\beta_3 = 0.0001$. The assumed distributions for the independent variables are: $x_1 \sim U(0,300)$, $x_2 \sim U(0,30)$, and $x_3 \sim U(0,100000)$. Random number generators are used to generate a population sample of size $N = 1,000$. This specification generates a data set with approximately 30% of the observations taking a positive number of trips.

Shaw uses the following process to create the data set used in the estimation: First, the population data set, called the "P data set," is created by calculating y_i^* for each observation. The P data set is then truncated by rounding off each y_i to the nearest integer and truncating at zero. This new truncated data set is called the "T data set." Endogenous stratification of the T data set is achieved by replicating each observation, i, y_i times. For example, observations for which $y_i = 2$ each appear twice in the new truncated, endogenously stratified data set, called the "T&S data set." Table 1 shows the summary statistics for the P, T, and T&S data sets for Shaw's basic specification ($\sigma^2 = 25$, $corr(x_2, x_3) = 0$).

Shaw compares several estimations in the MC simulation. In order to provide a basis of comparison, he conducts ordinary least squares (OLS) using the P data set. Since neither truncation nor endogenous stratification is present in the P data set, OLS generates the best estimates possible. OLS is also conducted using the T&S data set to show the effect of ignoring the problems of truncation and endogenous stratification. In addition, he tests his derived likelihood function, equation (1.3), with a T&S data set.

Table 2 shows the results of Shaw's MC simulation, which involved estimating each model a single time. The results shown are for the case of $\sigma^2 = 25$, $corr(x_2, x_3) = 0$, but the results are consistent across the specifications considered by Shaw. As expected, OLS with the P data set outperforms the other methods, and the use of OLS with a T&S data set results in biased parameter estimates. The performance of the Shaw likelihood is good, producing parameter estimates that are very close to the true values. Table 3 shows the effect of increasing the error variance, σ^2. In general, increasing the error variance increases the standard errors of the parameter estimates and makes it difficult to distinguish which model performs best.

The approach taken by Shaw in this MC simulation can be improved in a couple of respects. The first is in the creation of the T&S data set. Though the method used by Shaw is capable of creating a data set that exhibits the properties of truncation and endogenous stratification, the process of rounding the trip quantities seems somehow unsatisfying. He acknowledges that rounding could create some biases in data generation, but states that he expects the bias to be small and to cancel out.[1] Though it is certainly possible that the bias could cancel out, it is also possible that the process of endogenously stratifying the sample could exacerbate any bias introduced by rounding.

Since the desire is to compare estimation methods using a continuous demand specification, a more straightforward approach is to use quantities that have not been rounded. This presents no problem in either the creation of the data set or the estimations, and eliminates the possibility that the rounding has an effect on the outcome of the simulation. Additionally, Shaw estimates each model a single time. The power of the Monte Carlo approach lies in the fact that the estimation can be performed a large number of times. Repeated estimation allows for a higher degree of confidence in the results.

In the next section, I will discuss a MC simulation that uses a more general method of constructing the truncated and endogenously stratified data set than that used by Shaw. This method of construction should eliminate any artifact of the data set creation process in the MC comparison of estimation methods. An estimation method not considered by Shaw is also evaluated.

A MORE GENERAL APPROACH TO THE MONTE CARLO SIMULATION

Consider another recreation demand model that takes the following form

$$
\begin{aligned}
y_i^* &= \delta_0 + \delta_1 \left(x_{1i} + \delta_2 x_{2i} x_{3i} \right) + \delta_3 x_{4i} + u_i \quad i = 1, \ldots, N && (1.5)\\
y_i &= y_i^* \qquad \text{if } y_i^* > 0 \\
&= 0 \qquad \text{otherwise}
\end{aligned}
$$

where x_{1i} are respondent i 's out-of-pocket travel costs, x_{2i} are wages, x_{3i} are travel times, x_{4i} are incomes, δ_j ($j = 0, \ldots, 3$) are parameters to be estimated, y_i^* are the desired number of trips to the recreation site, y_i are the observable number of trips, N is the sample size, and $u_i \sim N(0, \sigma^2)$. This specification of trip demand is similar to that used by Shaw, but a slightly different approach is taken with regard to the specification of the price. The price (i.e. travel cost in this example) is made up of explicit costs, x_{1i}, plus a term, $\delta_2 x_{2i} x_{3i}$, that represents the respondent's time cost. The parameter δ_2 represents the fraction of the respondent's full wage rate, x_{2i}, at which they will be compensated for the time, x_{3i}, spent traveling to and from the recreation site.

Though this is a slightly more complicated demand model than that used by Shaw, it is important to keep in mind that they are both simply different specifications of demand functions. The conclusions generated apply to other demand models irregardless of the particular application.

Parameters will be set at the following values: $\delta_0 = 1$, $\delta_1 = -2.5$, $\delta_2 = 0.3$, and $\delta_3 = 0.8$. The distributions for the independent variables are: $x_1 \sim U(10{,}115)$, $x_2 \sim U(0{,}30)$, and $x_3 \sim U(0.4{,}9.6)$, and $x_4 \sim U(0{,}100000)$. I will also assume that $corr(x_{2i}, x_{4i}) = 0$ and $\sigma^2 = 25$. This model specification generates a population data set with approximately 2% of the observations taking a positive number of trips.[2] Random number generators were used to generate a population sample of size $N = 10{,}000$.

Recall that Shaw generated the truncated data set by rounding each trip quantity to the nearest integer and dropping all observations for which $y_i^* \leq 0$. That truncated data set was then endogenously stratified by replicating the observations in each stratum y_i^* times. A more general method of creating the T&S data set is to avoid the rounding step.

The population sample is truncated by eliminating all observations for which $y_i^* \leq 0$, but all trip quantities are left as rational numbers. The truncated data set is then endogenously stratified by randomly sampling observations according to a sampling weight, sw_i, defined as

$$sw_i \equiv \frac{y_i}{\sum_j y_j}. \tag{1.6}$$

Respondents with a high number of trips will have a proportionally higher probability of being selected. This allows for the creation of a T&S data set of any size, and avoids rounding of the trip quantities.

Table 4 shows the summary statistics of the data sets for the basic specification: $\sigma^2 = 25$, $corr(x_2, x_4) = 0$. Because of the smaller proportion of trip takers in the population for this simulation, it was necessary to generate larger P data sets than were generated by Shaw.

The MC simulation proceeds as follows: (1) generate P data set, (2) truncate the P data set to created a T data set, (3) create a T&S data set by randomly sampling 1000 observation from the T data set according to the sampling weights, sw_i, (4) estimate the parameters of the models, (5) go back to step (3), etc. A total of 1000 separate T&S data sets are drawn from the T data set, and a new set of parameter estimates is generated for each iteration. Means and mean squared errors are reported for each model's parameters.

Three separate estimation methods will be examined. The first is the use of the truncated normal likelihood function, shown in equation (1.6), on the T&S data set. This serves to highlight the bias that would be introduced by treating the on-site sample as if it were a random sample from the population. The second method is the use of the likelihood function developed by Shaw, shown in equation (1.3). The final estimation method I will consider is the use of inverse sample weights in the likelihood function. This is a computationally simpler approach than that Shaw likelihood function. The next section will describe how the inverse sample weights are derived.

INVERSE SAMPLE WEIGHTING

The intuition behind inverse sample weighting is that each observation should be given a weight in the likelihood function that is inversely related to the probability that the observation appears in the endogenously stratified sample. Observations that are over-represented in the sample are given a smaller weight than observations that are under-represented.

A weighting mechanism with this characteristic is

$$\omega(t) \equiv \frac{P(t)}{S(t)} \tag{1.7}$$

where $\omega(t)$ represents the weight used for respondents taking t trips per season, $P(t)$ is the percentage of the population taking t trips per season, and $S(t)$ is the percentage of the on-site sample taking t trips per season. Suppose that 5% of the population took three trips per year, but the fraction of respondents in the on-site sample who took three trips per year was 20%. Inverse sample weighting would give each three-trip taker a weight of 0.25 in the likelihood function, thus diminishing their representation in the on-site sample. The problem is that the population fractions, $P(t)$, are not known.

By assuming, as does Shaw, that the fraction of total trips made up of individuals in the on-site sample taking t trips per season, $S(t)$, is the same as their fraction of total trips in the population, the inverse sample weights can be derived from the information contained in the sample, and are given by

$$\omega(t) = \frac{\theta}{t},$$

where

$$\theta = \frac{1}{\left[\sum_{t=1}^{T} \frac{S(t)}{t}\right]}.$$

Derivation of this result is shown in Appendix A.

These three estimation methods were compared in the MC simulation. Table 5 presents the results of the simulation. Mean squared errors are shown for each parameter estimate. In general, the Shaw likelihood function performs well, outperforming, as expected, the model that ignores endogenous stratification as well as the inverse sample weighting model. The parameter estimates generated using the Shaw likelihood function a very similar to the true values. Additionally, the parameter estimates are more precise, as indicated by the smaller mean squared errors.

Unexpectedly, the inverse sample weighting model does not seem to outperform the model that completely ignores endogenous sample weighting. The inverse sample weighting model and the model that completely ignores endogenous stratification generate similar parameter estimates, with the inverse sample weighting model getting closer to the true parameter values in a minority of cases.

Table 6 shows the results of a Monte Carlo simulation with a higher error variance, $\sigma^2 = 625$.[3] Similar to the results found by Shaw, increasing the error variance resulted in a poorer performance for all models. However, the Shaw likelihood function continues to outperform the other models at the higher level of error variance. Again, the model that ignores endogenous stratification and the inverse sample weighting models generate very similar results.

CONCLUSIONS

The results of Shaw's 1985 MC simulation as well as the results for the simulations discussed here indicate that the Shaw likelihood function is capable of correcting the problems created by using data that is both truncated and endogenously stratified. This is an important conclusion for researchers conducting demand research because financial and time considerations often dictate that cheaper on-site sampling procedures be used. By using the Shaw likelihood function to generate the demand parameter estimates, the researcher can use the on-site sample to infer properties of the more general population and draw much stronger conclusion from the data. Unfortunately, the computationally simpler inverse sample weighting method does not appear to be able to significantly correct for the problems created by using on-site samples.

REFERENCES

1. Amemiya, T. Regression Analysis When the Dependent Variable is Truncated Normal, *Econometrica*, 41(1973): 997-1016.
2. Englin, Jeffrey and J. S. Shonkwiler. Estimating Social Welfare Using Count Data Models: An Application to Long-Run Recreation Demand Under Condition of Endogenous Stratification and Truncation, *The Review of Economics and Statistics*, 77(1995): 104-112.
3. Laitila, Thomas. Estimation of Combined Site-Choice and Trip-Frequency Models of Recreational Demand Using Choice-Based and On-Site Sample, *Economics Letters*, 64(1999): 17-23.
4. Manski, C. and D. McFadden. Structural analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge, MA: 1982.
5. Shaw, Daigee. Three Essays in the Economics of Recreation Demand, Ph.D. Dissertation, University of Michigan (1985).
6. Shaw, Daigee. On-site Samples' Regression: Problems of Non-negative Integers, Truncation, and Endogenous Stratification, *Journal of Econometrics*, 37(1988): 211-223.

Table 1: Statistics of data sets for Shaw's basic specification: $\sigma^2 = 25$, $corr(x_2, x_3) = 0$

Data Set	Variable	Number of Observations	Mean	Min	Max	Standard Deviation
P	y^*	1,000	-15.59	-74.00	38.00	26.51
	x_1	1,000	147.62	0.55	298.87	85.94
	x_2	1,000	15.03	0.03	29.99	8.78
	x_3	1,000	51007.50	122.43	99982.20	29459.90
T	$y^* = y$	324	15.75	1.00	38.00	9.04
	x_1	324	49.99	0.55	116.38	30.36
	x_2	324	14.11	0.03	29.95	8.63
	x_3	324	53694.10	308.77	99982.20	28282.80
T&S	$y^* = y$	5,102	20.92	1.00	38.00	7.88
	x_1	5,102	36.02	0.55	116.38	25.87
	x_2	5,102	13.60	0.03	29.95	8.82
	x_3	5,102	56551.70	308.77	99982.20	28576.00

Table 2: Monte Carlo results for Shaw's basic specification: $\sigma^2 = 25$, $corr(x_2, x_3) = 0$

Parameter	True value	OLS of P data	OLS of T&S data	Shaw Likelihood T&S data
β_0	25	25.30 (0.489)	26.38 (0.186)	25.60 (0.211)
β_1	-0.3	-0.30 (0.002)	-0.24 (0.003)	-0.27 (0.003)
β_2	-0.1	-0.11 (0.018)	-0.11 (0.007)	-0.11 (0.008)
β_3	0.0001	0.0001 (0.000005)	0.00008 (0.000002)	0.00009 (0.000003)
σ^2	25	24.662	21.486	24.93 (0.500)

[a] Standard errors are shown in parentheses

Table 3: Monte Carlo results for Shaw's basic specification: $\sigma^2 = 400$, $corr(x_2, x_3) = 0$

Parameter	True value	OLS of P data	OLS of T&S data	Shaw Likelihood T&S data
β_0	25	26.22 (1.96)	41.59 (0.48)	32.59 (0.77)
β_1	-0.3	-0.30 (0.007)	-0.14 (0.004)	-0.24 (0.01)
β_2	-0.1	-0.13 (0.07)	-0.18 (0.02)	-0.27 (0.03)
β_3	0.0001	0.0001 (0.00002)	0.00002 (0.000006)	0.00004 (0.00001)
σ^2	400	396.01	208.87	340.80 (9.00)

[a] Standard errors are shown in parentheses

Table 4: Statistics of data sets for new Monte Carlo simulation

Data Set	Variable	Number of Observations	Mean	Min	Max	Standard Deviation
P	y^*	10,000	-186.31	-487.70	57.13	98.58
	x_1	10,000	67.56	10.03	125.00	32.89
	x_2	10,000	14.87	0.002	30.00	8.65
	x_3	10,000	5.20	0.40	10.00	2.78
	x_4	10,000	49716.41	1.72	99992.06	28770.02
T	$y^* = y$	163	14.50	0.04	44.41	11.19
	x_1	163	16.10	10.01	29.45	4.56
	x_2	163	8.25	0.02	29.57	7.66
	x_3	163	3.30	0.44	9.95	2.65
	x_4	163	80600.00	40039.27	99943.45	14353.11
T&S	$y^* = y$	1000	23.15	1.02	44.41	10.62
	x_1	1000	14.82	10.08	29.45	3.67
	x_2	1000	7.47	0.02	29.52	7.83
	x_3	1000	3.46	0.44	9.94	2.71
	x_4	1000	84248.09	40039.27	99943.45	12669.93

Table 5: Monte Carlo results for basic specification: $\sigma^2 = 25$, $corr(x_2, x_3) = 0$

Parameter	True value	Shaw likelihood T&S data	Ignoring endogenous stratification	Inverse sample weighting
δ_0	1	2.33 (3.41)	6.35 (29.72)	5.20 (18.08)
δ_1	-2.5	-2.40 (0.01)	-2.09 (0.17)	-2.03 (0.22)
δ_2	0.3	0.29 (0.0002)	0.29 (0.0002)	0.29 (0.00007)
δ_3	0.8	0.76 (0.002)	0.66 (0.02)	0.64 (0.02)
σ	5	5.07 (0.02)	4.78 (0.06)	4.73 (0.08)

[a] Mean squared errors are shown in parentheses

Table 6: Monte Carlo results for basic specification: $\sigma^2 = 625$, $corr(x_2, x_3) = 0$

Parameter	True value	Shaw likelihood T&S data	Ignoring endogenous stratification	Inverse sample weighting
δ_0	25	26.33 (15.99)	47.16 (495.37)	35.87 (119.80)
δ_1	-0.90	-0.91 (0.003)	-0.55 (0.12)	-0.45 (0.20)
δ_2	0.30	0.29 (0.0008)	0.31 (0.0008)	0.30 (0.0004)
δ_3	0.80	0.80 (0.002)	0.50 (0.09)	0.43 (0.14)
σ	25	24.42 (377.62)	19.73 (217.15)	18.37 (178.97)

[a] Mean squared errors are shown in parentheses.

ENDNOTES

[1] Shaw states, "The rounding-off of the y_i to the nearest integer may create some biases in data generation. However, we expect the bias to be small and negligible after the differences cancel each other out." (Shaw 1988, p. 220).

[1] The parameter values were chosen in order to provide a higher proportion of zeros in the population sample than the design used by Shaw. It was felt that the higher proportion of non-visitors was closer to reality for a vast majority of recreation sites.

[1] Some parameter values had to be changed in the simulation in order to roughly maintain the proportion of non-visitors in the population sample.

APPENDIX A

Assume that the fraction of total trips made up of individuals in the on-site sample taking t trips per season, $S(t)$, is the same as their fraction of total trips in the population. This implies that

$$\begin{aligned} S(t) &= \frac{tN(t)}{\sum_{s=1}^{T} sN(s)} \\ &= \frac{tP(t)}{\sum_{s=1}^{T} sP(s)} \end{aligned} \tag{1.8}$$

where $N(t)$ indicates the number of individuals in the population taking t trips per season, T represents the maximum number of trips taken per season by anyone in the population, and $P(t) \equiv N(t) \Big/ \sum_{s=1}^{T} N(s)$ denotes the percentage of the population taking t trips per season. Let

$$\theta \equiv \sum_{t=1}^{T} tP(t) \tag{1.9}$$

denote the average number of trips in the population. Then (1.8) implies that

$$P(t)=\frac{\theta S(t)}{t}. \tag{1.10}$$

But since

$$\begin{aligned}1&=\sum_{t=1}^{T}P(t)\\&=\theta\sum_{t=1}^{T}S(t)\\&\Rightarrow\\\theta&=\frac{1}{\left[\sum_{t=1}^{T}\frac{S(t)}{t}\right]},\end{aligned}$$

θ can be calculated using information from the on-site sample, and we have that:

$$P(t)=\frac{S(t)}{t}\left[\sum_{s=1}^{T}\frac{S(s)}{s}\right]^{-1}=\frac{\theta S(t)}{t}.$$

Inverse sample weighting would then be given by:

$$\begin{aligned}\varpi(t)&\equiv\frac{P(t)}{S(t)}\\&=\frac{\theta}{t}.\end{aligned}$$

NOTES