7 On the Use of Simulated Frequencies to Approximate Choice Probabilities

Steven R. Lerman and Charles F. Manski

7.1 Introduction

A mundane but common problem in probabilistic analysis concerns the calculation of the event probabilities generated by a random process.

To be precise, let α be a real random M-vector whose distribution G has support $\mathbf{A} \subset \mathbf{R}^M$. Let $\mathbf{A}(j), j = 1, \ldots, J$ be a finite Lebesgue measureable partition of $\mathbf{A}$ such that $\alpha \in \mathbf{A}(j)$ if and only if observable event j occurs. Then for any $j = 1, \ldots, J$, the probability that event j occurs is simply

$$P(j) = \text{Prob}[\alpha \in \mathbf{A}(j)] = \int_{\mathbf{A}(j)} dG. \tag{7.1}$$

The generation of event probabilities from an underlying random process in this manner is ubiquitous in probabilistic analysis. Unfortunately unless the distribution G and the subset $\mathbf{A}(j)$ are particularly benign, the integral $\int_{\mathbf{A}(j)} dG$ will not have a closed form. A common problem then is to find a computationally practical method to adequately approximate the event probability.

A simple general solution to problem (7.1) exists. In its most basic form the procedure is to draw a set of pseudorandom realizations of α, observe on each draw whether the realization of α lies within $\mathbf{A}(j)$, and use the frequency of such occurrences over the set of draws as an estimate of the event probability $P(j)$.

This simulated frequency Monte Carlo approach to probability calculation is well known by workers in the area of computer simulation. In particular see Hammersley and Handscomb (1964) and Fishman (1973) for discussions. The potential value of the approach in econometric applications, however, appears not to have been widely recognized.

In this chapter a version of the simulated frequency procedure is applied to solve a specific problem in discrete choice analysis. Let the events $j = 1, \ldots, J$ now be alternatives from which a decision maker must choose one. Assume that behavior is consistent with a random utility model, and let

The work reported here was performed at Cambridge Systematics, Inc., under a Federal Highway Administration contract. We would like to thank Carlos Daganzo, William Eddy, and Daniel McFadden for useful discussions. Responsibility for the contents of this chapter is the authors' alone.

$\alpha = [\alpha(1), \ldots, \alpha(J)]$ be the random utility vector with distribution G. Then as long as G is such that ties have zero probability, the probability that alternative j is chosen is given by

$$P(j) = \text{Prob}[\alpha(j) \geq \alpha(k), k = 1, \ldots, J] = \int_{\mathbf{A}(j)} dG, \tag{7.2}$$

where $\mathbf{A}(j) = [\alpha : \alpha(j) \geq \alpha(k), k = 1, \ldots, J]$.

A long-standing problem of some importance has been to develop practical methods to evaluate the choice probability (7.2) under alternative distributions G. In particular many researchers have been interested in evaluating (7.2) under the assumption that G is multivariate normal, when the choice probabilities have the multinomial probit form. Clearly the simulated frequency procedure offers a solution to the general problem and to that of calculating multinomial probit probabilities specifically.

7.2 The Simulated Frequency Method

This section describes the classical simulated frequency procedure for estimation of a single probability. Section 7.3 sets out a Bayesian version of the procedure. Section 7.4 discusses issues that arise when a collection of probabilities must be calculated and ultimately some function of this collection evaluated.

Let N realizations of α be drawn at random, and let $N(j)$ be the number of such realizations lying in $\mathbf{A}(j)$. Then $F(j, N) = N(j)/N$ is binomially distributed with mean $P(j)$ and variance $P(j)(1 - P(j))/N$. Observe that the distribution of $F(j, N)$ depends only on $P(j)$ and N and not directly on the random process, characterized by G and $\mathbf{A}(j)$, generating event j.

As an estimate of $P(j)$, $F(j, N)$ has well-known statistical properties. It is strongly consistent, minimum variance unbiased, and is the maximum likelihood estimate. Computationally the cost of calculating $F(j, N)$ may reasonably be assumed linear in N. The marginal cost per trial has two additive components. First, there is the cost of drawing a pseudorandom realization of α, and second, the cost of determining whether this realization lies in $\mathbf{A}(j)$. The size of these two marginal cost components are problem specific, the former depending on the nature of the distribution G and the latter on the structure of the set $\mathbf{A}(j)$.

A comment should be made regarding the drawing of pseudorandom realizations of α. If α is univariate, and G is strictly increasing, realizations may be relatively easily drawn by first taking a random variable $\beta \sim U[0, 1]$ and then calculating $\alpha \equiv G^{-1}(\beta)$. See Shreider (1966) for a discussion of methods for drawing the needed uniform random numbers.

On the other hand, if α is multivariate, such a simple general method seems not to be available. A very useful approach is to find M independent random variables $\gamma = [\gamma_1, \ldots, \gamma_M]$ and a function $\mathbf{H}: \mathbf{R}^M \to \mathbf{R}^M$ such that $\mathbf{H}(\gamma)$ is distributed G. Then α-realizations may be generated by drawing values for $\gamma_1, \ldots, \gamma_M$ and computing $\mathbf{H}(\gamma)$. For example, if $\alpha \sim N(\mu, \Sigma)$, let $\gamma_m \sim N(0, 1)$, $m = 1, \ldots, M$, and let Ω be an $M \times M$ triangular matrix such that $\Sigma = \Omega'\Omega$. Then $(\mu + \Omega'\gamma) \sim N(\mu, \Sigma)$ as desired.

It has been speculated that for any multivariate distribution G, there exists some γ and $\mathbf{H}$ such that $\mathbf{H}(\gamma) \sim G$; so the method can be applied generally. However, no proof of this proposition is known to us.

7.3 Bayesian Approach

The classical simulated frequency method appears naive in at least two respects: the procedure provides no means for one to incorporate prior information regarding the value of $P(j)$ into the estimation process; use of the method requires one to fix ahead of time the number of simulation trials N to be performed.

The literature on simulation suggests various sophisticated but ad hoc variants on the basic classical method, including ones utilizing informative stopping rules. See Fishman (1973) for details. A more satisfactory approach is offered by Bayesian statistical decision theory. See DeGroot (1970) for a comprehensive textbook presentation.

The Bayesian approach becomes particularly convenient if it is assumed that a priori, $P(j) \sim B(a, b)$, where B designates the beta distribution. As is well known, the distribution of $P(j)$ after N simulation trials is then $B(a + N(j), b + N - N(j))$.

7.4 Estimation of a Function of a Collection of Probabilities

In many applications the concern is not merely to calculate a single probability but to compute a function of a collection of probabilities. In exploring the properties of simulated estimates of such functions, we will be

concerned with two types of asymptotic behavior. First, we may be interested in the properties of the estimates when the number of probabilities in the collection becomes arbitrarily large, while the simulations per probability is held constant. Second, both the number of draws per probability and the number of probabilities in the collection can become arbitrarily large. This turns out to be a significant distinction in two problems of particular interest that follow.

In the first situation we would like to estimate the share of a population that will choose an alternative. One common approach is to calculate the expected fraction of a random sample of decision makers who will select a given alternative j from some population. Let $t = 1, \ldots, T$ be the sample of decision makers, and let $P_t(j)$ be the probability that decision maker t selects alternative j. Then our concern is to calculate $Q_T(j) = 1/T \sum_{t=1}^{T} P_t(j)$. Observe that $\operatorname{plim}_{T\to\infty} Q_T(j) = Q(j) \equiv E[P(j)]$, the population average probability of selecting j. This fact provides the basis for use of $Q_T(j)$ as an estimate for $Q(j)$ in what has come to be termed the "random sample enumeration" forecasting method.

It is easy to see that the simulated frequency approach is well suited to the task of approximating $Q_T(j)$ and, more important, of estimating $Q(j)$. The classical simulation estimate is $R_T(j, N_t, t = 1, \ldots, T) = 1/T \sum_{t=1}^{T} N_t(j)/N_t$, where $N_t \geq 1$ is the number of trials performed for decision maker t. Conditional on T, this estimate has mean $E_T(R_T) = Q_T(j)$ and variance $V_T(R_T) = 1/T^2 \sum_{t=1}^{T} [P_t(j)(1 - P_t(j))]/N_t$. Letting $T \to \infty$, and recalling that decision makers are drawn at random, one finds that $\operatorname{plim}_{T\to\infty} R_T(j, N_t, t = 1, \ldots, T) = Q(j)$ for any positive values of N_t, $t = 1, \ldots \infty$.[1]

This simple result is quite powerful, as it states that the simulated frequency method consistently (as $T \to \infty$) estimates $Q(j)$ without any requirement for consistent estimation of each $P_t(j)$. That is, we do not require that each $N_t \to \infty$. In the extreme, one simulation trial per decision maker would suffice.

The second application is not nearly so benign. Consider now the problem of calculating a sample log likelihood $L_T = 1/T \sum_{t=1}^{T} \ln P_t(j_t)$, where j_t is the alternative actually selected by the sampled decision maker t.

1. This can be readily demonstrated by noting that $\operatorname{plim}_{T\to\infty} E_T(R_T) = Q(j)$ and $\operatorname{plim}_{T\to\infty} V_T(R_T) = 0$. Chebychev's theorem implies directly that R_T is a consistent estimate for $Q(j)$ as $T \to \infty$.

In this case the classical simulation estimate is $S_T\ (N_t,\ t = 1, \ldots, T) = 1/T\ \Sigma_{t=1}^{T}\ \ln\ [N_t(j_t)/N_t]$.

Observe first that, as long as $P(j) < 1$, there is nonzero probability that alternative j will never have the greatest utility in a finite number of draws. This implies that $\underset{T\to\infty}{\text{plim}}\ S_T = -\infty$. On the other hand, it is the case that, if we let $N_t \to \infty$, then $\ln\ N_t(j_t)/N_t$ is a consistent estimate for $\ln P_t(j_t)$. It follows that for any T the simulation estimate for S_T converges (as $N_t \to \infty$ for each $t = 1, \ldots, T$) to the true sample likelihood L_T. Finally, if both $N_t \to \infty$ for all t and $T \to \infty$, the simulation estimate converges to $E(L)$. Thus the classical simulated frequency method consistently estimates $E(L)$ only if the number of trials per decision maker and the number of decision makers sampled both go to infinity.

The rather extreme result that for fixed N_t values $\underset{T\to\infty}{\text{plim}}\ S_T = -\infty$ can be avoided if each classical simulation estimate is replaced by one of the form $(a_t + N_t(j_t))/(a_t + b_t + N_t)$ for $a_t,\ b_t > 0$. This estimate is interpretable in Bayesian terms as the posterior mean for $P_t(j_t)$ under the assumption that the prior distribution of $P_t(j_t)$ is beta with parameters a_t and b_t. As before, consistency requires that $T \to \infty$ and that for each $t = 1, \ldots, T, N_t \to \infty$.

The foregoing discussion is of some potential consequence for the use of simulation estimates in maximum likelihood estimation. Let $\boldsymbol{\theta}$ be a real parameter vector, and for each t let $P_t(j_t \mid \boldsymbol{\theta})$ be a family of probabilities indexed by $\boldsymbol{\theta}$. Define $N_t(j_t \mid \boldsymbol{\theta})$, $a_t(\boldsymbol{\theta})$, $b_t(\boldsymbol{\theta})$, $L_T(\boldsymbol{\theta})$, and $S_T(\boldsymbol{\theta})$ in the natural way. If we use posterior means as simulation estimates, then for fixed N_t, $t = 1, \ldots, \infty$ the difference $\underset{T\to\infty}{\text{plim}}\ S_T(\boldsymbol{\theta}) - E(L(\boldsymbol{\theta}))$ will in general be a function of $\boldsymbol{\theta}$. Consistency (as $T \to \infty$) of the ideal maximum likelihood estimator $\max_{\theta} L_T(\boldsymbol{\theta})$ therefore does not ensure consistency of the approximate estimator $\max_{\theta} S_T(\boldsymbol{\theta})$. The latter property is guaranteed only as $N_t \to \infty,\ t = 1, \ldots, T$ and $T \to \infty$.

7.5 Application to the Calculation of Multinomial Probit Choice Probabilities

The multinomial probit probabilistic choice model presumes a population of decision makers T each member t of which must select an alternative from a choice set $\mathbf{C}$ consisting of J alternatives. With each $t \in T$ there is associated a utility vector $(U_{tj}, j \in \mathbf{C})$ distributed as multivariate normal.

The probability that t selects some $i \in \mathbf{C}$ is then the normal tail probability $Pr(U_{tj} - U_{ti} \leq 0, \text{ all } j \in \mathbf{C})$.

A quite flexible parametric specification for the utilities $(U_{tj}, j \in \mathbf{C}, t \in T)$ is the random coefficients form $U_{tj} = \mathbf{Z}_{tj} \cdot \boldsymbol{\theta}_t^*$. Here $\mathbf{Z}_{tj}$ is a vector of observed attributes characterizing the decision maker and alternative, and $\boldsymbol{\theta}_t^*$ is an unobserved realization of a random vector $\boldsymbol{\theta}^* \sim N(\bar{\boldsymbol{\theta}}^*, \boldsymbol{\Sigma})$. It is assumed that for any t, $t' \in T$, $t \neq t'$, that $\boldsymbol{\theta}_t^*$ and $\boldsymbol{\theta}_{t'}^*$ are independent realizations of $\boldsymbol{\theta}^*$.

Until recently the multinomial probit model remained a theoretically attractive but empirically unused specification for discrete choice analysis. In particular the random coefficients form of the model drew attention for its great flexibility relative to the widely used conditional logit model. See for example McFadden (1976) for a discussion. The difficulty in applying the model derived from the fact that mathematically a multinomial probit choice probability is a multidimensional integral which has no closed form and in which the domain of integration is unbounded from below. As is well known, classical methods for numerical integration become quite burdensome in multiple integral contexts, with computation times for given accuracy generally increasing with the power of the integral's dimensionality.

The first advances in resolving the computational impasse were due to Dutt (1976) and Hausman and Wise (1978) who investigated series approximations to the multinomial probit choice probability. Dutt's work has never been implemented in any probit estimation program, but Hausman and Wise did succeed in developing a program capable of handling choice sets with three and four alternatives. In personal communications the latter authors have indicated a belief that the series approximation approach may be practical for choice sets of up to but not beyond five alternatives. This judgment, combined with the fact that series evaluation of choice probabilities requires a separate routine for each size choice set, persuaded us that the series approach was too limited in practical scope and too rigid in implementation for use in a general purpose program. Instead our attention turned to the simulated frequency method.

In the abstract the simulation approach appeared attractive for a variety of reasons. First, it is easily programmed and applied to choice sets of any size. Second, the method not only gives the user a measure of the accuracy of the probability calculations after any number of trials but allows him to control this accuracy through his ability to set the number of trials to be

performed. Third, CPU time in the simulation approach goes up relatively slowly with choice set size. In particular for a given number of trials CPU time increases linearly with choice set size.

We have applied the simulated frequency method to the problem of calculating random coefficients—multinomial probit choice probabilities for use in a routine performing maximum likelihood estimation of the parameters $(\bar{\theta}^*, \Sigma)$[2]. Section 7.6 describes the version of the method programmed. While in the midst of this effort we learned of an intriguing alternative probability calculation approach originated by Clark (1961) and unearthed by Daganzo, Bouthelier, and Sheffi (1977). A routine for producing Clark probabilities was subsequently added to our estimation program. Section 7.7 describes the Clark algorithm, and section 7.8 presents numerical tests comparing simulation and Clark probability calculation results. A detailed description of our probit estimation package is not given here. The interested reader should see Albright, Lerman, and Manski (1977a and 1977b).

7.6 The Simulation Routine

A certain amount of informed pragmatism has guided our design of a simulated frequency routine for use in the probit estimation package. The routine programmed has the following features:

1. The quantity $(N_t(j_t) + 1)/(N_t + J)$ is used as the estimate for $P_t(j_t)$. This quantity is interpretable as the posterior mean under the beta prior having mean $1/J$ and variance $(J - 1)/[J^2(J + 1)]$. The assumption of prior mean $1/J$ is natural as the choice set contains J alternatives. The variance assumption imposed has no particular justification.

2. For every realization $\tilde{\theta}_t$ drawn, another one $\tilde{\tilde{\theta}}_t = \bar{\theta} + (\bar{\theta} - \tilde{\theta}_t)$ is used on the following simulation trial. Here $\bar{\theta}$ is the mean of the θ distribution from which draws are made. This use of antithetic variates makes the realizations of pairs of simulation trials negatively correlated, thereby increasing the precision of the simulation estimate relative to that in which independent trials are used. The use of antithetics also halves the number of times the random number generator must be invoked.

3. An informative stopping rule for determining the number of trials N_t is used. Specifically let N_0 and K be positive integers, λ be a positive real,

2. We have recently learned that a similar application has been made by Charles River Associates (1976).

and $F(j, kN_0)$ be the value of the simulation estimate for $P_t(j_t)$ after kN_0 trials, $k = 1, \ldots, K$. Now set $k = 1$, run N_0 trials and stop if $(1 - F(j, kN_0))/[F(j, kN_0)kN_0] < \lambda$ or if $k = K$. Otherwise set $k = k + 1$, run N_0 additional trials, and apply the stopping rule again. Use of this stopping rule was motivated by a concern to estimate ln P with equal precision regardless of the value of P. A more detailed but still heuristic explanation is given in Albright, Lerman, and Manski (1977a).

It should be noted that with these modifications to the classical simulation procedure, the classical variance formula $P(j)(1 - P(j))/N$ no longer is valid. We have nevertheless continued to use this formula to provide a measure of the precision with which $F(j)$ estimates $P(j)$. More precisely, since $P(j)$ is not known, the quantity $F(j)(1 - F(j))/N$ is used. An alternative, and perhaps more justifiable measure of precision is the Bayesian posterior variance $F(j)(1 - F(j))/(N + J + 1)$ obtained under the beta prior introduced above.[3]

7.7 The Clark Method

Clark (1961) suggested an approximation to the distribution of the maximum of M jointly normal random variables. The approximation rests on the fact that if x_1, x_2, and x_3 are jointly normal, the statistics $E(\max(x_1, x_2))$, $\mathrm{VAR}(\max(x_1, x_2))$, and $\mathrm{COV}(\max(x_1, x_2), x_3)$ can be calculated exactly in a straightforward manner. Clark then suggests the approximation that $\max(x_1, x_2)$ is itself normally distributed. Given this demonstrably false assumption, the first two moments of $\max(x_1, x_2, x_3) = \max(\max(x_1, x_2), x_3)$ can be calculated. Repeated application of the approximation allows one to approximate the distribution of the maximum of M jointly normal random variables. As evidence of the success of his approximation, Clark shows that it gives a very close estimate of $E(\max(x_1, \ldots, x_M))$ for M as large as 10. He also suggests error bounds. Clark does not provide evidence as to the suitability of the approximation in estimating higher moments of $\max(x_1, \ldots, x_M)$, or for approximating normal tail probabilities.

Daganzo, Bouthelier, and Sheffi (1977) discovered the Clark paper and applied its method to the estimation of probit choice probabilities. If C is a

3. It should be noted that all stopping rules in which the stopping criterion is monotonic decreasing in $F(j)$ lead to upward biased estimates of $P(j)$. Our rule has this property. However, the bias goes to zero as $\lambda \to 0$ and $KN_0 \to \infty$. See Albright, Lerman, and Manski (1977a).

choice set with utilities ($U_j, j \in \mathbf{C}$) distributed multivariate normal, and if i is the chosen alternative, then the utility differences ($U_j - U_i, j \in \mathbf{C}, j \neq i$) are multivariate normal, and the probability that i is chosen is Prob[(max($U_j - U_i$), $j \in \mathbf{C}$, $j \neq i$) ≤ 0]. Daganzo et al. used the Clark approximation to the distribution of (max($U_j - U_i$), $j \in \mathbf{C}$, $j \neq i$) to estimate this probability.

A priori the Clark method's attractiveness to us lay in its speed. The method requires the evaluation of only univariate normal tail probabilities, a task quite quickly accomplished by series approximation.[4] Moreover it is easy to show that CPU times for the Clark probabilities go up no faster than the square of choice set size.

In contrast to the simulation and series approaches where the user may control the accuracy of his approximated choice probabilities, the Clark method offers only a fixed accuracy level. The crucial question in determining the method's practical usefulness therefore was whether this accuracy was sufficient for probit estimation purposes. In particular it was important to determine whether the deviations of Clark approximation from true choice probabilities systematically vary with choice set size, true probability magnitude, disturbance covariance structure, and so on.[5] Since no analytical approach to determining the Clark properties could be found, we conducted a series of numerical tests.[6]

7.8 Numerical Test Objectives and Design

We have computed high accuracy (large number of trials) simulation probabilities for a broad range of choice problems and the corresponding probabilities approximated by the Clark algorithm. These tests serve two purposes. First, the Clark probabilities may be compared with the accurate

4. For this purpose we have used subroutine NDTR of the IBM scientific subroutine package.

5. Since the Clark method works recursively, the order in which alternatives are treated can also in principle affect the choice probabilities obtained. Daganzo et al. have, however, found that this order effect is quite small and our own tests confirm this. Hence we have not concerned ourselves with alternate orders in our programming.

6. About the only known analytical property of the Clark approach is that it uses a symmetric distribution to approximate one skewed to the right. This fact has led to the unconfirmed belief that the Clark method will overestimate true probabilities at the lower tail. Moreover, since the skewness of the true distribution increases with choice set size, it has been speculated that the accuracy of Clark probabilities decreases with choice set size. Finally, it has been asserted by some that the Clark method is most accurate when the normal covariance matrix contains only positive elements. Again no proof exists.

simulation ones to determine how the Clark accuracy depends on choice set size, the magnitude of the true choice probability and the variance-covariance structure of the utility function. Second, the tests allow us to assess the accuracy obtainable through the simulation approach when the number of trials is set so as to use the same CPU time as does the Clark method. In particular if x_s is the per trial CPU time consumed in the simulation method, and x_c is the CPU time for the Clark algorithm, then $N_c = x_c/x_s$ is the number of simulation trials that can be performed in the time taken by the Clark method.[7] The accuracy of the simulation probabilities obtainable in N_c trials can then be measured through the percent variance formula $(1 - P(j))/(P(j) \cdot N_c)$ and compared with the Clark method's accuracy as previously determined.

The tests reported here assume choice sets containing three or five alternatives.[8] The utility function is assumed to have the form $U_{ti} = \delta_t \cdot Z_{ti} + \varepsilon_{ti}$, where δ_t is a random coefficient and $(\varepsilon_{tj}, j \in \mathbf{C})$ is a set of alternative specific disturbances (random coefficients for alternative specific constants). It is assumed that $\delta_t \sim N(1, \sigma_\delta^2)$, $(\varepsilon_{tj}, j \in \mathbf{C}) \sim N(\mathbf{0}, \mathbf{\Sigma}_\varepsilon)$ and that δ is independent of the ε vector.

Within given choice sets, the attribute differences $(Z_i - Z_j)$ for $i,j \in \mathbf{C}$ lie in the range $[-4, 4]$. Across the set of tests the variances σ_δ^2 of the scalar random taste variable δ lies in the interval [0, 2]. The variance-covariance matrix $\mathbf{\Sigma}_\varepsilon$ of the ε disturbances is in some tests scalar, sometimes diagonal, and sometimes general. The ε variances (diagonal elements of $\mathbf{\Sigma}_\varepsilon$) range over the interval [1/4, 4]. The choice probabilities (as estimated by a large number of simulation drawings) resulting from the various test specifications range over the interval [0.002, 0.718].

For each test specification we calculate simulation and Clark choice probabilities for all alternatives in the relevant choice set.[9] In the simulation runs the stopping rule parameters are set at $N_0 = 1{,}000$, $K = 10$, and $\lambda = 0.0005$, so as to guarantee relatively high accuracy. In general it

7. These calculations are choice set size specific but should not depend on any other aspect of the choice problem.

8. These are a representative selection from our full series of tests, which are documented in Albright, Lerman, and Manski (1977a).

9. The simulation probability for each alternative is calculated independently. Hence over the choice set these probabilities do not in general sum exactly to one. Clark probabilities also do not usually sum to one.

can be expected that the simulation probability estimate $F(j)$ satisfies the conditions $|(F(j) - P(j))/P(j)| < 0.044$ and $|F(j) - P(j)| < 0.030$.[10]

7.9 Test Results and Analysis

Table 7.1 presents simulation and Clark choice probabilities for a set of choice problems, each choice set containing three alternatives. Table 7.2 presents similar results for choice sets containing five alternatives.

The most striking feature of the test results is the relatively high accuracy of the Clark probabilities. Not only are the simulation estimates and Clark approximations generally quite close to one another, but perhaps surprisingly, the accuracy of the Clark probabilities does not appear to systematically vary along the dimensions of concern to us. In particular the size of the choice set, magnitude of the true probability, and covariance structure of the random coefficients δ and ε have no noticeable effects.[11] We do not really understand why Clark's approximation works as well as it does. All we know is that within the domain tested it does work.[12]

The CPU times x_c required for each Clark calculation are quite reasonable, as expected. The reported average values (denoted as $\bar{x}_c$) of 0.007 and 0.009 seconds for three and five alternative choice sets allow the calculation of 8,000 and 6,000 probabilities per minute, a rate sufficiently high for economical probit estimation. The reported values of 14 and 11 for $N_c = \bar{x}_c/\bar{x}_s$ are quite small, implying that only crude simulation probabilities can be obtained in the time used by the Clark method. In fact comparison of the Clark root mean square error estimates and the simulation standard errors reported in tables 7.1 and 7.2 indicate that the

10. Assuming a classical (unmodified) simulation procedure, the standard error of $F(j)$ after N trials is $(P(j)(1 - P(j))/N)^{1/2}$, and the standard error of $F(j)/P(j)$ is $((1 - P(j))/NP(j))^{1/2}$. In our modified procedure a minimum of $N_0 = 1{,}000$ trials is performed, so it seems safe to assume that the standard error of $F(j)$ is always less than $((1/2 \cdot 1/2)/1{,}000) \cong 0.015$. Moreover setting $\lambda \cong 0.0005$ implies that $((1 - F(j)/NF(j))^{1/2} < (0.0005)^{1/2} \cong 0.022$, giving an upper bound on the standard error of $F(j)/P(j)$. Using a conservative two-standard-error criterion, we conclude that generally the fractional error $|(F(j) - P(j))/P(j)|$ will be less than 0.044 and the absolute error $|F(j) - P(j)|$ less than 0.030.

11. Note, however, that none of the tests performed have negative elements in the matrix Σ_ε.

12. Since our simulation probabilities still contain some error even with the large number of trials performed, the Clark method's accuracy may be even better than tables 7.1 and 7.2 indicate. There is a slight tendency for the simulation probabilities to be higher than the Clark ones. This may be a consequence of the simulation stopping rule we used.

Table 7.1
Probability calculation tests using choice sets with three alternatives

Case 1	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = I_3$
Simulation	$P_1 = 0.320$	$P_2 = 0.348$	$P_3 = 0.330$	$\Sigma P = 0.998$	
Clark	0.332	0.332	0.332	0.996	
Case 2	$Z_1 = 1$	$Z_2 = 0$	$Z_3 = 0.75$	$\sigma_\delta = 2$	$\Sigma_\varepsilon = I_3$
Simulation	$P_1 = 0.474$	$P_2 = 0.215$	$P_3 = 0.321$	$\Sigma P = 1.010$	
Clark	0.463	0.210	0.319	0.992	
Case 3	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = \begin{matrix} 0.25 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{matrix}$
Simulation	$P_1 = 0.261$	$P_2 = 0.319$	$P_3 = 0.417$	$\Sigma P = 0.997$	
Clark	0.261	0.300	0.416	0.977	
Case 4	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = \begin{matrix} 1 & 0.75 & 0 \\ 0.75 & 1 & 0 \\ 0 & 0 & 1 \end{matrix}$
Simulation	$P_1 = 0.285$	$P_2 = 0.301$	$P_3 = 0.418$	$\Sigma P = 1.004$	
Clark	0.279	0.279	0.419	0.977	
Case 5	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = \begin{matrix} 0.25 & 0.38 & 0 \\ 0.38 & 1 & 0 \\ 0 & 0 & 4 \end{matrix}$
Simulation	$P_1 = 0.245$	$P_2 = 0.310$	$P_3 = 0.447$	$\Sigma P = 1.002$	
Clark	0.239	0.284	0.447	0.970	

$\bar{x}_c \cong 0.007$ $\quad \bar{x}_s = 0.0005$ seconds $\quad N_c = \dfrac{\bar{x}_c}{\bar{x}_s} \cong 14$

Root mean square value of	(Simulation − Clark) = 0.012	$(P(1-P)/N_c)^{1/2} = 0.126$ for $P = 1/3$
Root mean square value of	$\left(\dfrac{\text{Simulation} - \text{Clark}}{\text{Simulation}}\right) = 0.021$	$((1-P)/PN_c)^{1/2} = 0.379$ for $P = 1/3$

Table 7.2
Probability calculation tests using choice sets with five alternatives

Case 1	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$Z_4 = 0$	$Z_5 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = I_5$
Simulation	$P_1 = 0.198$	$P_2 = 0.196$	$P_3 = 0.189$	$P_4 = 0.203$	$P_5 = 0.200$	$\Sigma P = 0.986$	
Clark	0.198	0.198	0.198	0.198	0.198	0.990	
Case 2	$Z_1 = 2$	$Z_2 = 1$	$Z_3 = 0$	$Z_4 = -1$	$Z_5 = -2$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = I_5$
Simulation	$P_1 = 0.718$	$P_2 = 0.228$	$P_3 = 0.050$	$P_4 = 0.007$	$P_5 = 0.002$	$\Sigma P = 1.005$	
Clark	0.723	0.220	0.048	0.007	0.001	0.999	
Case 3	$Z_1 = 2$	$Z_2 = 1$	$Z_3 = 0$	$Z_4 = -1$	$Z_5 = -2$	$\sigma_\delta = 1$	$\Sigma_\varepsilon = I_5$
Simulation	$P_1 = 0.608$	$P_2 = 0.149$	$P_3 = 0.059$	$P_4 = 0.060$	$P_5 = 0.086$	$\Sigma P = 0.962$	
Clark	0.568	0.134	0.064	0.073	0.093	0.932	
Case 4	$Z_1 = 0$	$Z_2 = 0$	$Z_3 = 0$	$Z_4 = 0$	$Z_5 = 0$	$\sigma_\delta = 0$	$\Sigma_\varepsilon = G$
Simulation	$P_1 = 0.282$	$P_2 = 0.266$	$P_3 = 0.131$	$P_4 = 0.138$	$P_5 = 0.209$	$\Sigma P = 1.026$	
Clark	0.256	0.256	0.118	0.137	0.185	0.952	
Case 5	$Z_1 = 2$	$Z_2 = 1$	$Z_3 = 0$	$Z_4 = -1$	$Z_5 = -2$	$\sigma_\delta = 1$	$\Sigma_\varepsilon = G$
Simulation	$P_1 = 0.667$	$P_2 = 0.289$	$P_3 = 0.050$	$P_4 = 0.014$	$P_5 = 0.003$	$\Sigma P = 1.023$	
Clark	0.639	0.272	0.056	0.019	0.007	0.993	

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 5 & 2 & 2 & 2 \\ 0 & 2 & 2 & 1.75 & 1.75 \\ 0 & 2 & 1.75 & 2.56 & 2.31 \\ 0 & 2 & 1.75 & 2.31 & 3.13 \end{bmatrix}$$

$\bar{x}_c \cong 0.009$ seconds $\quad \bar{x}_s \cong 0.0008$ seconds $\quad N_c = \bar{x}_c / \bar{x}_s \cong 11$

Root mean square value of (Simulation − Clark) $= 0.014$ $\quad (P(1-P)/N_c)^{1/2} = 0.120$ for $P = 1/5$

Root mean square value of $\left(\dfrac{\text{Simulation} - \text{Clark}}{\text{Simulation}}\right) = 0.032$ $\quad ((1-P)/PN_c)^{1/2} = 0.603$ for $P = 1/5$

Clark probabilities are ten to twenty times more accurate than the simulation ones achievable for the same CPU time. Conversely, since simulation accuracy increases as the square root of the number of trials, the simulation method can be seen to require more than one hundred times the Clark CPU time to achieve the same accuracy.

7.10 Conclusions

The above numerical results plus independent corroborating evidence in Daganzo et al. (1977) suggest that the Clark method should provide the preferred means for likelihood evaluation in multinomial probit estimation. While the Clark approximation seems to dominate the simulation approach in this application, our work with the latter method has still been quite valuable. First, without the ability to estimate true choice probabilities using high accuracy simulations, it would have been impossible to assess the accuracy of the Clark probabilities. Second, if some problem with the Clark method should be uncovered, the simulation approach, as programmed in our probit estimation program, will still be available. Thus the simulation approach provides some security to the empirical researcher wishing to estimate a multinomial probit choice model.

More generally the emergence of the Clark method as a successful approach for the calculation of multinomial probit probabilities may be regarded as somewhat fortuitous. Since such a powerful approximation may not exist under other distributions than the normal, it is useful to know that a generally applicable method, the simulation method, is available.

Finally, recall the findings of section 7.4 that in a forecasting situation, the simulation method works well even if the number of trials per observation is quite small. This result suggests that even when the simulation method is dominated by another in an estimation context, the method may still prove cost effective when the estimated model is used in forecasting.

References

Albright, R., S. Lerman, and C. Manski. 1977a. Compendium of Technical Memoranda on the Development and Testing of a Multinomial Probit Estimation Package. Prepared for the Federal Highway Administration.

Albright, R., S. Lerman and C. Manski. 1977b. Report on the Development of an Estimator for the Generalized Multinomial Probit Model. Prepared for the Federal Highway Administration.

Charles River Associates. 1976. Impact of Trade Policies on the U.S. Automobile Market. Report prepared for the Bureau of International Labor Affairs, U.S. Department of Labor, Washington, D.C.

Clark, C. 1961. The Greatest of a Finite Set of Random Variables. *Operations Research.* 9: 145–162.

Daganzo, C., F. Bouthelier, and Y. Sheffi. 1977. Multinomial Probit and Qualitative Choice—A Computationally Efficient Algorithm. *Transportation Science.* 11: 338–358.

DeGroot, M. 1970. *Optimal Statistical Decisions.* New York: McGraw-Hill.

Dutt, J. 1976. Numerical Aspects of Multivariate Normal Probabilities in Econometric Models. *Annals of Economic and Social Measurement.* Vol. 5.

Fishman, G. 1973. *Concepts and Methods in Discrete Event Digital Simulation.* New York: Wiley.

Hammersley, J., and D. Handscomb. 1964. *Monte Carlo Methods.* London: Methuen.

Hausman, J., and D. Wise. 1978. A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences. *Econometrica.* 46: 403–426.

McFadden, D. 1976. Quantal Choice Analysis: A Survey. *Annals of Economic and Social Measurement.* 5: 363–390.

Shreider, Y. ed. 1966. *The Monte Carlo Method.* New York: Pergamon Press.