Repeated Random Variation in Simulations of Spatial Search

John E. Abraham
HBA Specto Incorporated
April 2010

Abstract

In search algorithms that use random utility theory it is important that random components of utility be stored, so that if elements of behaviour are reconsidered during the search they are evaluated consistently. Directly storing each random component can easily overwhelm computer memory when many individuals, activity patterns, preference-commonality and destinations are considered. This paper describes a system for effectively managing and assigning random components from a pool, to obtain a rich simulation of choice across multiple dimensions of land use and transport behaviour incorporated preference commonality between individuals and alternatives.

In spatial choices, zones are aggregations of potential destinations. To avoid biases due to arbitrary zone boundaries, it is important that individual (sub-zonal) destinations be considered, or that extreme value theory be used to assign appropriate random variation components to zonal aggregations. This paper describes options for treating zonal destinations appropriately in a simulation model given random utility theory.

Introduction

In Random Utility Theory, individuals choose the option that has the highest personal utility, and the utility is represented with a function that includes random variables which follow a certain distribution, and represent variation in preference and perception between individuals.

Choices are interrelated in that choosing an option in one dimension might constrain and influence the options available for another dimension. For instance, in forecasting land use and related travel behaviour dimensions might include choices for:

- home location (residential location),
- how many automobiles to own,
- where to work,
- when to go to work,
- whether to drive to work,
- route if driving to work,

Since these choices are interrelated (routes to work obviously depend on home and work location, but the choice of where to live and work also depends on the attributes of available routes), we wish to have a mathematical formula that describes the attractiveness of the full joint choice of all of these options. Such a utility function might have the following general form:

$$U(a)=\sum_{i\in E}\left(V_i\left(c_{ia}\right)+\varepsilon_i\left(c_{ia}\right)+\sum_{j\in E,j\neq i}\left(V_{ij}\left(c_{ia},c_{ja}\right)+\varepsilon_{ij}\left(c_{ia},c_{ja}\right)+\sum_{k\in E,k\neq i,j}\left(V_{ijk}\left(c_{ia},c_{ja},c_{ka}\right)+\varepsilon_{ijk}\left(c_{ia},c_{ja},c_{ka}\right)+\cdots\right)\right)\right) \quad (1)$$

where:

c_{na} = description of alternative *a* under consideration with regards to dimension *n*

E = set of dimensions,

V_i = a mathematical function measuring the expected utility of one dimension *i* for an individual decision maker,

ε_i = a random variable for dimension *i*, representing the probabilistic nature of the choice model and the degree to which individual situations and individual's perception differ from the deterministic function V_i

V_{ij} = a mathematical function measuring the expected value of the utility of the interdependencies between dimension *i* and *j*

ε_{ij} = a random variable for dimensions *i* and *j*, representing the degree to which individual joint situations across *i* and *j* and individual's perception of joint situations differ from the deterministic function $V_i + V_j + V_{ij}$

V_{ijk} = a mathematical function measuring any interdepencies across three dimensions

ε_{ijk} = a random variable for three dimensions, representing any special non-deterministic dependencies across three dimensions.

In **allocation** algorithms, often the logit model is used which obviates the need for sampling from the underlying distributions: if the non-random components of utility (V_i, V_{ij}, V_{ijk} ...) of all of the alternatives can be calculated, and if the elements can be arranged in a hierarchical nest with a very specific assumptions regarding the random variables (specifically, independent and identical Gumbel distributions for various combinations of ε_i, ε_{ij}, ε_{ijk}, …) then the probability of each option can be calculated directly. These probabilities are used to allocate the total amongst the options[1].

In **simulation** algorithms, where choice options are assigned to individual decision makers using sampling, there are generally two options: 1) assume the specific set of independent and identical Gumbel distrubtions for various combinations of ε_i, ε_{ij}, ε_{ijk}, …, and sample an option from the same resulting closed-form probability distribution used in allocation algorithms, or 2) sample the random components of utility ε_i, ε_{ij}, ε_{ijk}, … from appropriate distributions for *each* option for *each* individual and then assign the highest utility option to the individual.

Methods for dealing with large choice sets

When many dimensions are considered, the number of options can rapidly become large. For instance, in the example joint choice for a medium city there could be a 300,000

[1] Implying either an acceptance of the law of large numbers, or a model that forecasts the expected value and not one realization of the range of possible futures

dwellings, 50,000 jobs for which an individual is qualified, the household could own up to, say, 5 automobiles, the person could depart to work perhaps up to 60 minutes early, could drive or not, and if driving could take perhaps 10 different reasonable routes. This gives 3e5 * 5e4 * 6 * 60 * 2 * 10 = 1.08E14 alternatives for one individual, and if there are one million individuals there could be 1e20 alternatives under consideration by the citizens of a city. When the number of options are very large, considering each option explicitly is unrealistic behaviourally (since in reality people can not evaluate so many options) and computationally infeasible (even computers using simplified representations can not evaluate many billion options for many million decision makers). Three strategies are common for dealing with a large number of options, *sample* from amongst the options, *group/nest* the options and use a series of conditional models, or use a *search algorithm* to explore the set of options.

Sampling of alternatives

In a sampling algorithm, some simpler distribution is used to find a small group of options available to a decision maker. Random utility theory is then used to choose from amongst the smaller set. Behaviourally, this could be interpreted as decision makers using a simple system for determining a few options from amongst the many options, and then choosing from amongst those few options using more complete information as assumed in random utility theory. Sampling is generally applicable to simulation algorithms, but can also be applied to allocation algorithms using sample enumeration approaches.

Grouping/Nesting

When using grouping (also known as nesting) to manage large choice sets, the set of alternatives is partitioned (often repeatedly into a tree structure), with each partition containing groups of similar options. The similar options can have similar deterministic terms (V terms) and/or correlated random terms (ε terms), perhaps because they share the same alternative in one or more dimensions of the choice. The choice between options within a group is simplified because the number of options has been reduced. The probability of choosing a group uses extreme value theory to calculate the expected maximum utility of the underlying alternatives within the group.

One example of this technique occurs when aggregate combinations of individual locations are considered as a "zone". The alternatives within a zone are given the same values for $c_{i,\ i=\text{location choice}}$ and the ε_i is given an independent and identical Gumbel distribution. Individuals choose the best option (for them) within the zone, and the utility of this best option from extreme value theory is:

$$V_i^z = V_i\left(c_i^z\right) + \frac{1}{\lambda_i}\ln S_z + \varepsilon_j \qquad (2)$$

where:

λ = parameter for the Gumbel distribution, inversely related to its standard deviation

S_z = number of alternatives (the "size") of the zone z.

c_i^z = the attributes of the zone *z* influencing its attractiveness with regards to choice dimension *i*.

ε_j = a random variable with the same distribution as ε_i but associated with the zone choice *j* instead of individual location choice *i*.

Grouping is applicable to allocation and simulation.

Search Algorithm

In a search algorithm, individuals explore the options available to them, without an exhaustive search of all options. During the search it is possible to consider specific detailed options, or groups of related options. If groups are being explored the utility of the group should be calculated using extreme value theory. A system of determining when to stop the search process is required, and when the search process is complete the option with the highest utility is chosen. Search algorithms are applicable to simulation but not to allocation.

Sampling random terms in search algorithms

In search algorithms, alternatives may be considered in any order. It is also possible that alternatives will be reconsidered as the algorithm progresses. One alternative in one dimension *i* would be considered early in the search process, then could be discarded, perhaps because it was being considered in conjunction with options in other dimensions, *j* and *k*, that were undesirable. The option for dimension *i* is then likely to be reconsidered later in the process, when it may be accepted as the best alternative. As a concrete example, a destination could be considered along with a transit mode trip, and rejected during the search process because of the difficulty of travel by transit. Later the same destination could be accepted in conjunction with an auto trip.

Each random term can be sampled from its underlying distribution, but once sampled must remain in memory and re-used if that element of the choice is reconsidered, otherwise the notion of a random term ε_i as specific to the individual option and the individual's preferences is not respected.

In the context of activity based travel demand modeling, I have developed certain implementations of search algorithms that did *not* store the random terms, but rather random terms were resampled when required. This led to a destination choice model that always chose nearby large zones. In early parts of the search, a nearby large zone (with a large S_z) could have been rejected because a smaller, farther away zone, had been associated with a larger sampled random component ε_j. But as the search progressed, the nearby large zones would be reconsidered, and when they were reconsidered the random component was resampled. Eventually the random component for the nearby large zone became large, and the smaller, farther away zones were no longer chosen. Thus these implementations of search algorithm with dynamic (re)sampling of random terms was shown to be inappropriate given random utility theory.

Behaviourally, the interpretation in destination choice is that there is a non zero probability that individuals will choose each destination specifically because there is something unique about that destination that makes it most attractive for one individual's preferences and perceptions. The sampled random components represent these preferences and perceptions. Resampling the random components repeatedly, and throwing out the previous sample, was violating the purpose of the random components, and causing a malfunction in the destination choice model.

Storing random components

It is theoretically possible to generate random components as they are required, and to store them associated with the individual and the appropriate portion of the alternative. Thus, for instance, when an individual *y* first considers a particular mode *m* a random compoenent ε^{y}_{m} can be sampled from the appropriate distribution and stored for reuse the next time individual *y* considers mode *m*. This approach becomes infeasible, however, when large numbers of alternatives are considered, as in the example where 1.08e14 alternatives are available. Not all options would be considered for every individual, but still storing each individual's random components for a destination choice model was deemed unrealistic.

The model does not need that much detail in the random variation. 1.08e14 is a large number; we require an appropriate variation in the model results, but we should not need that much variation. What is required is a system for generating and storing a reasonably large quantity of random numbers, and then assigning them to random components with minimal storage.

A pool *R* of R random numbers can be stored in an array. R should be sufficiently large to represent the random variation that is required in model behaviour – for example it could be larger than any one dimension in the choice set (e.g. with 2000 potential destination zones, R should be larger than 2000). Each individual requires a different random component for each alternative for each dimension of the choice E. Let ς_e^{y} be an integer randomly sampled from the uniform distribution [1,R] and skip_e^{y} be an integer randomly sampled in the range [0,Skip], where Skip << R. Then let random component $\varepsilon^{y}_{e,n}$, being the random component associated with the n^{th} option of dimension *e* for individual *y,* be

$$\varepsilon^{y}_{e,n} = \mathrm{G}(R[\mathrm{modulus}(\varsigma_e^{y} + \mathrm{skip}_e^{y} * \mathrm{n},\mathrm{R})]) \tag{3}$$

where:

G() = the appropriate transformation between the distribution used to sample the pool of random numbers *R* and the distribution required for random component $\varepsilon^{y}_{e,n}$.

In other words, the random component for an element of choice $e \in E$ for individual *y* is chosen based on a entry point into the array, ς_e^{y}, and a skip value, skip_e^{y}, that determines how many entries are skipped over while working through the array.

The sampling of random terms when populating the array ensures that each element of each option has an appropriate random component. The sampling of the entry point ς_e^{y} for each individual *y* and choice dimension *e* ensures that random components do not get inappropriately assigned to the same alternative for multiple individuals. The sampling of the skip value skip$_e^y$ ensures that sequences of alternatives do not get inappropriately assigned the same sequence of random components.

Similar techniques can be used for any non-zero random terms which apply to more than one element of the choice, ε_{ij}^{y}.

The flexibility associated with adopting the G() function to transform one distribution to another suggests that the pool R should be sampled from well-used or easy-to-transform distributions. The uniform distribution between 0 and 1, and the standard normal distributions are options. It would also be possible to maintain multiple pools of random numbers, e.g. one pool from the standard normal distribution and one pool from the uniform distribution [0.1).

This technique can be compared to an old technique, no longer in fashion, of assigning random components from a book of random numbers generated by a truly physical process (RAND corporation, 1955). Modulus(ς_e^y+ skip$_e^y$ * n,R) would be the place in the book where the random number is found for element *e* and individual *y.*

Destination choice random component distributions

The system described above requires a function G(), which transforms the random number drawn from the pool *R* into an appropriate random component for element *e* for alternative *a*. Given elemental choice options, with no grouping, the random component should probably be transformed into a normal distribution given the central limit theorem. The variation of the normal distribution could be sampled separately for individual *y*, and for destination activity type α, and applied in the transformation, thus providing a mixed probit model of destination choice.

For the choice of a destination *zone*, it is important to note that zones are aggregations of elemental choices. Zones have different sizes, being the number of elemental choices that they include. Thus the zone is a grouping of elemental alternatives. To avoid biases due to the arbitrary nature of zones, the random component associated with a zone should be calculated using a method that respects the notion that individuals will choose the best option, for them, within the zone.

Direct consideration of group of elemental alternatives

The utility of a zone could be calculated directly by finding the maximum over the set of elemental alternatives within the zone. Thus the destination choice alternatives could be constructed as individual discrete options within a zone (corresponding to individual jobs, homes, or blocks of 100 square feet of retail space). If the deterministic parts of the utility are the same for all alternatives within the zone, the best option will be the one with the largest random component, so the random component of the attractiveness of the zone would simply the largest random component in the set of random components G(*R*[modulus(ς_e^y+ skip$_e^y$ * n,R] for *n* within the zone. This would be not too difficult to calculate, would be direct and clear, and would support a future extension of the model into a sub-zone representation of destinations.

Direct consideration of elemental alternatives

Instead of calculating the maximum over the individual alternatives within the zone when considering a zone as a destination, the search algorithm itself could be modified to consider individual discrete options within the model region. Thus the model would not use zones as destinations, but rather individual destinations within each zone. If done cleverly, this may not require much additional calculation but could support eventual sub-zone interactions with other variables – i.e. the deterministic part of the utility function could also vary across a zone (e.g. due to differences in parking cost or transit walk time for individual destinations within a zone). Further, this supports the behavioural notion that individuals have to search within a zone for the best elemental alternative for them, and may not always find the best sub-zone alternative.

Using extreme value theory

Extreme value theory can be used to assign a distribution to the maximum value from a number of options. Embrechts et al (1997) describe the appropriate distribution to use for the expected maximum of a set of elemental alternatives. If the elemental alternatives are assumed normal, and there are enough elemental alternatives, the cumulative distribution of the expected maximum of the group is given by the Gumbel distribution. Thus a number appropriately drawn from the random number pool can be transformed using a G() function that takes into account the size of the zone and the variation in the underlying distribution, which could be dependent on the individual *i*, the zone *z* and the size of the zone. The Encyclopedia of Statistical Sciences (2004) and Gumbel (1958) explain the distribution of the maximum of a set of normal values. The distributions are quite complicated, but approximations can be used or there are tables that have been published (or can be generated) to interpolate appropriate parameters.

One approximation, shown in Hall (1979) but attributable to Fisher and Tippet (1928) is to let b_n be the solution of the equation:

$$2\pi b_n^2 \exp\left(b_n^2\right) = n^2$$

where *n* is the number of elemental alternatives, let

$$a_n = \frac{1}{b_n}$$

These are approximated by α_n and β_n using

$$\beta_n = \frac{1}{\alpha_n} = (2\log n)^{\frac{1}{2}} - \frac{\frac{1}{2}(\log\log n + \log 4\pi)}{(2\log n)^{\frac{1}{2}}}$$

And then the Gumbel distribution

$$\exp\left(-e^{-x_1}\right)$$

is used with

$$x_1 = \alpha_n x + \beta_n$$

to sample the extreme.

Thus, given a random number ξ drawn from [0,1), an appropriate random term for zone z can be calculated as

$$\beta_n - \ln(-\ln(\xi)) \cdot \alpha_n$$

and then scaled according to the variance of the underlying normal distribution.

For the purpose of activity based travel modeling, it would be computationally efficient and more accurate to use a direct sample of the maximum of individually pre-sampled normal distributions for zones with a small number of jobs, dwellings or other elemental alternatives (i.e use the method described above as *Direct consideration of group of elemental alternatives* when *n* is small). The direct method should, perhaps, be used for zones with less than 50 elemental alternatives, and this Fisher and Tippet/Hall approximation to the Normal Extreme for larger zones. Figure 1 and Figure 2 compare Halls approximation to a randomly generated extreme normal distribution for n=20 and n=2000, from which it can be seen that *n* should be quite high before the approximation is used, and if zones are quite small it would be better to use a direct sample.

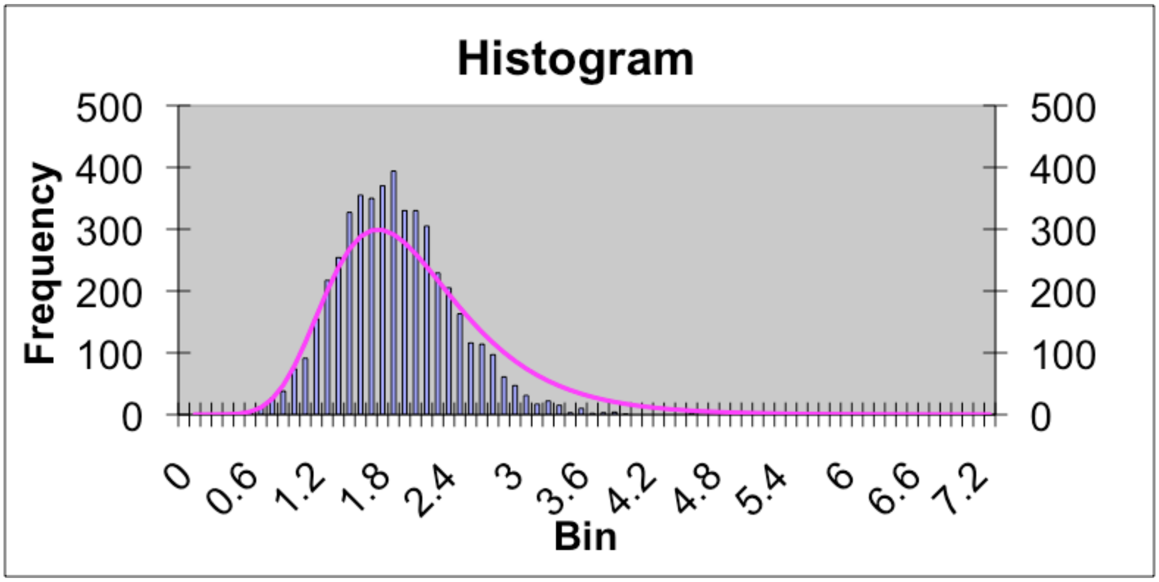

Figure 1: Halls approximation for N=20 (bars are a randomly generated extreme normal distribution)

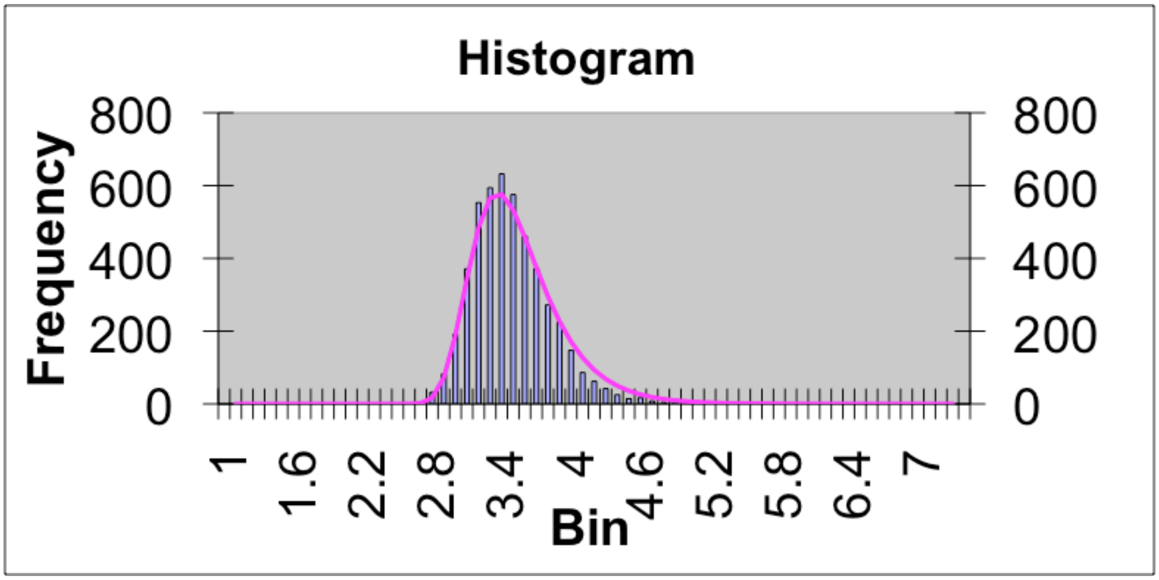

Figure 2: Halls approximation for N=2000, bars are a randomly generated extreme normal distribution

Assume an underlying Gumbel distribution

If the underlying elemental destination alternatives in each zone are themselves given Gumbel distributions instead of Normal distributions, then the standard nested logit approach could be used, as in equation 2. This is a little bit simpler, but is less than satisfactory because there is little reason to believe that the underlying elemental destination parameters should have Gumbel distributed error terms, given the Central Limit Theorum.

In particular, the expected maximum value of underlying Gumbel values is also Gumbel distributed, but with the same error term on the expected maximum as on the individual error term. This is very different than the expected maximum value of Normal values, where the error term on the expected maximum is lower than the error term on the underlying distribution. As groups of alternatives become larger, the expected maximum increases at a slower rate and with more certainty with an underlying normal distribution, as compared to an underlying Gumbel distribution.

This is shown in Figure 3, which shows the calculated average extreme value and standard deviation based on a sample of 1000 extreme values for each group size. Using an underlying Gumbel distribution causes larger alternatives to seem substantially more attractive than using an underlying normal distribution. This calls into question the standard practice of assuming an underlying Gumbel distribution for computational convenience, and in particular for assuming independent and identical distributions for zones (or other groups of alternatives) that have different sizes. Larger options should have utilities that are not proportionally as large, and/or error terms should be smaller on the larger options.

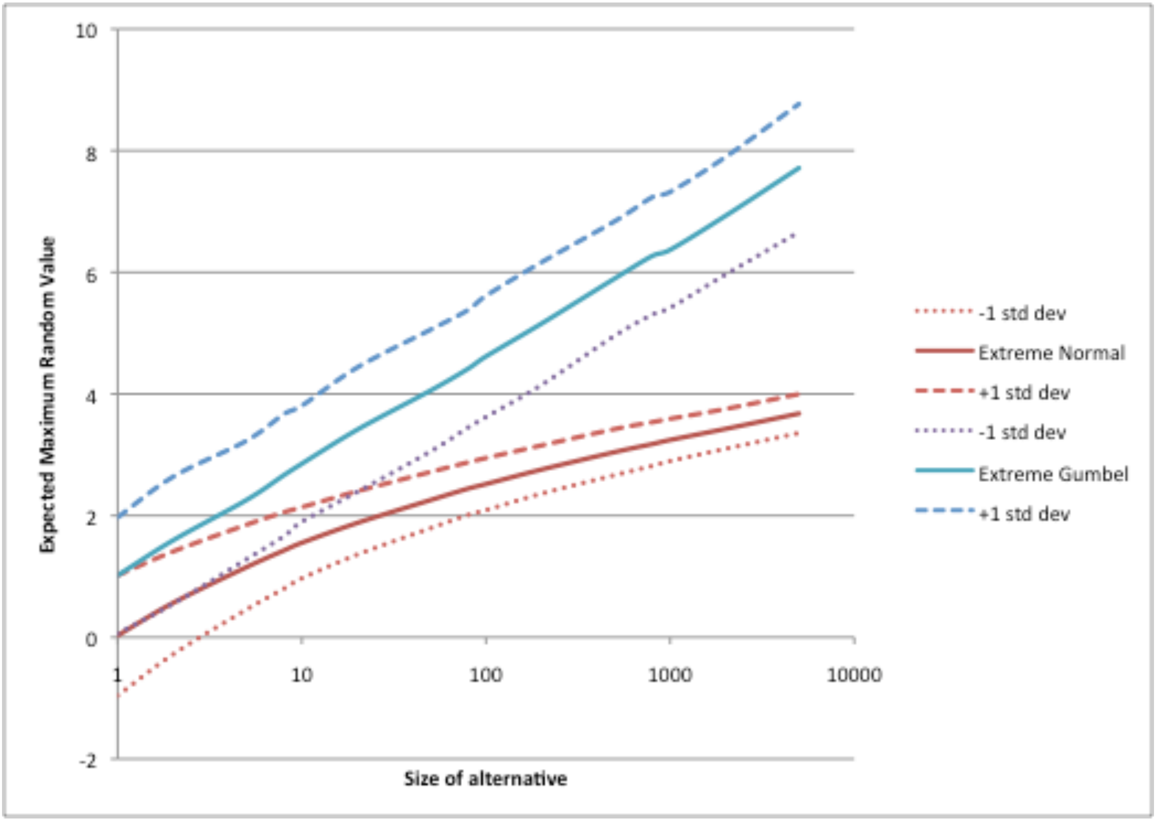

Figure 3: Expected maximum value of Gumbel and Normal distributions, showing ±1 standard deviations bands

Conclusions

Substantial errors can occur in modeling search processes using random utility theory. First, random terms must be stored so that they are not resampled if a dimension of an alternative is revisited. Second, if the random terms are expected to be normal, then extreme normal distributions should be used for the expected maximums of a group of alternatives. The extreme normal distribution can be calculated directly for small groups, or the Fisher and Tippet (a.k.a. Hall) approximation can be used for larger alternatives. The Gumbel distribution could be used for the expected maximum, but it does not seem to have much of an advantage over the Fisher and Tippet Extreme Normal distribution.

The common alternative to search processes, involving nesting structures and aggregate alternatives with size terms, has some large potential pitfalls. First, calculating the expected maximum value of a large group of alternatives is often not computationally feasible. Second, the common approach of assuming independent and identical Gumbel distributions for the best option within different groups of options is not supported when the groups are of different sizes, unless the underlying distribution is thought to be Gumbel. The Gumbel distribution looks like the Normal distribution when only one alternative is viewed, but the Extreme Gumbel distrubution is very different than the Extreme Normal distrubution for large n's. While it may be true that the extreme value of any underlying distribution is Gumbel distributed given a high enough *n*, it is a mistake to assume that the extreme value distribution for a group has the same standard deviation as the underlying distribution, as is commonly done in nested logit modeling.

References

Embrechts, P., C. Klüppelberg, and T. Mikosch, 1997, *Modelling extremal events for insurance and finance*. Spring Verlag, Berlin

Embrechts, P., S.I. Resnick and G. Samorodnitsky, 1999, Extreme value theory as a risk management tool, *North American Actuarial Journal*, Vol. 3, No. 2, (April 1999), pp. 30-41.

Gumbel, EJ, 1958, *Statistics of Extremes*, Columbia University Press, New York and London

Hall, P, 1979, On the Rate of Convergence of Normal Extremes, *Journal of Applied Probability*, 16:433-439

Kotz, S. and S. Nadarajah, 2000, *Extreme value distributions : theory and applications*, Imperial College Press, London

RAND Corporation, 1955, *A Million Random Digits with 100,000 Normal Deviates*, The Free Press, (republished by RAND Corporation in 2001)