TRANSPORTATION ENGINEERING PROGRAM

Lecture 3
Optimization methods for econometrics models

Cinzia Cirillo
University of Maryland
Department of Civil and Environmental Engineering
06/29/2016

Summer Seminar
June 27-30, 2016
Zinal, CH

Overview

1. Numerical Maximization
2. Algorithms
    1. Newton-Raphson
        1. Quadratics
        2. Step size
        3. Concavity
    2. BHHH
    3. BFGS
3. Convergence Criterion
4. Local vs Global Maximum
5. Variance of the Estimates

1. Numerical Maximization

- Most estimation involves maximization of some function (e.g. likelihood, simulated likelihood, or squared moment conditions).
- The log-likelihood function takes the form:

$$LL(\beta) = \sum_{n=1}^{N} \ln P_n(\beta)/N$$

Where: $P_n(\beta)$ is the probability of the observed outcome for decision maker n.

N is the sample size.

β is a K x 1 vector of parameters.

1. Numerical Maximization (cont.)

- Things to note:
  - The LL is divided by N so that it is the average LL in the sample.
  - N is fixed for a given sample, so it does not affect the location of the maximum.
  - We want to find the β that maximizes LL.
  - LL is always negative. This is because a likelihood is a probability, and probabilities are always between 0 and 1. The log of a value between these numbers is always negative.

1. Numerical Maximization (cont.)

- To maximize LL, we start from β_0.
- At each iteration we move to a new β value at which LL(β) is higher than the previous value.
- How to choose the step?
  - First, we calculate the gradient, the first derivative of LL(β) evaluated at β_t

$$g_t = \left(\frac{\partial LL(\beta)}{\partial \beta}\right)_{\beta_t} \qquad K\ x\ 1$$

  - Then, we define the hessian

$$H_t = \left(\frac{\partial g_t}{\partial \beta'}\right)_{\beta_t} = \left(\frac{\partial^2 LL(\beta)}{\partial \beta\ \partial \beta'}\right)_{\beta_t} \qquad \text{K x K}$$

  - The gradient tells us about the direction of the step, while the hessian tells us how far to step.

2. Algorithms

2.1 Newton-Raphson

- To determine the best value of β_{t+1}, we take a second order Taylor's approximation of LL(β_{t+1}) around LL(β_t)

$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t + \frac{1}{2}(\beta_{t+1} - \beta_t)' H_t (\beta_{t+1} - \beta_t)'$$

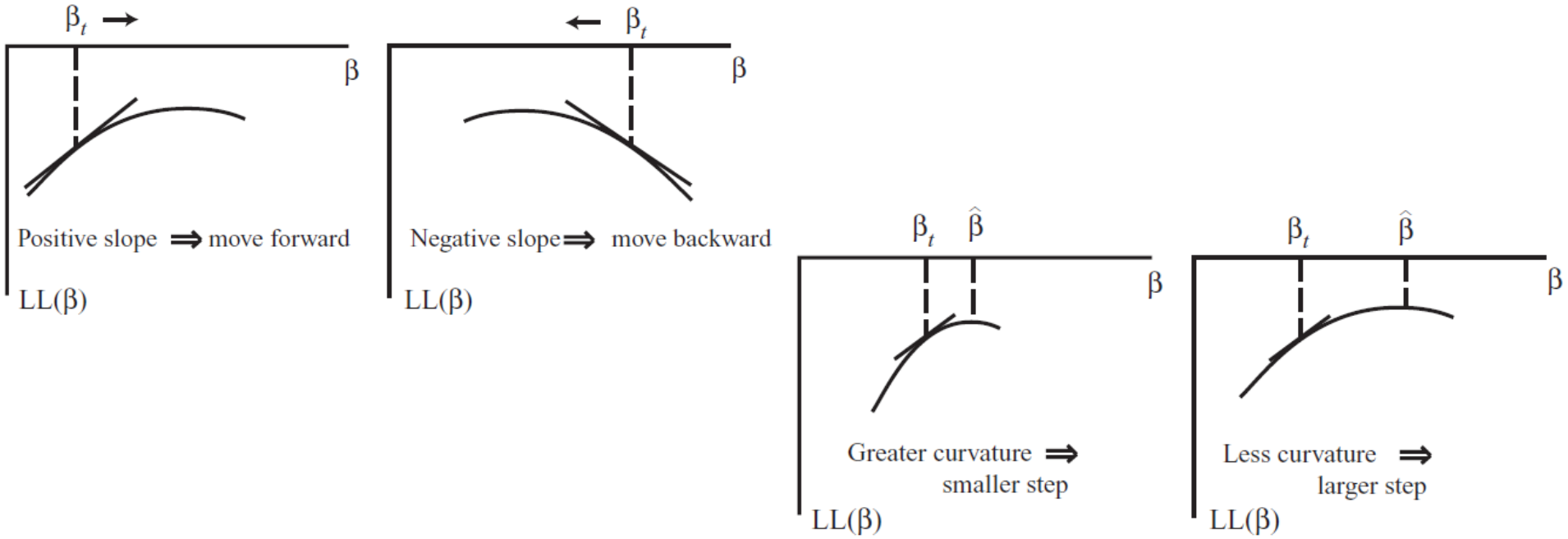
- The value β_{t+1} is found that maximizes this approximation to LL(β_{t+1}):

$$\frac{\partial LL(\beta_{t+1})}{\partial \beta_{t+1}} = g_t + H_t(\beta_{t+1} - \beta_t) \quad = 0$$

$$\beta_{t+1} = \beta_t + (-H_t^{-1}) g_t$$

2.1 Newton-Raphson (cont.)

- The direction of the step follows the slope (left figure).
- Step size is inversely related to the curvature (right figure).



2.1.1 Quadratics

- If LL(β) were exactly quadratic in β, then the Newton-Raphson procedure would reach the maximum in one step from any starting value.

$$LL(\beta) = a + b\beta + c\beta^2$$

$$\frac{\partial LL(\beta)}{\partial \beta} = b + 2c\beta = 0 \quad \rightarrow \; g_t$$

$$\hat{\beta} = -\frac{b}{2c} \qquad \text{and} \qquad H_t = 2c$$

2.1.1 Quadratics (cont.)

- However, LL(β) is NOT quadratic in β, therefore Newton-Raphson takes more than one step.

$$\beta_{t+1} = \beta_t + (-H_t^{-1})g_t$$

$$\beta_{t+1} = \beta_t - \frac{1}{2c}(\mathrm{b} + 2\mathrm{c}\beta_t)$$

$$\beta_{t+1} = -\frac{b}{2c} = \hat{\beta}$$

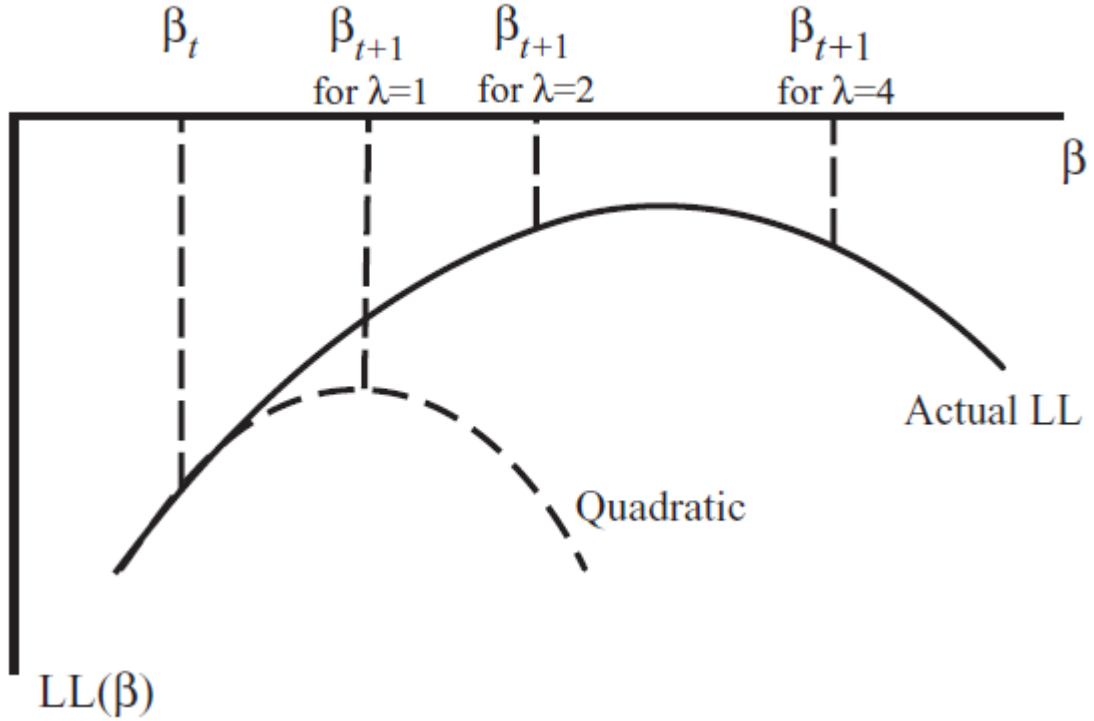
2.1.2 Step Size

- It is possible for the NR procedure to step past the maximum and move to a lower LL(β).
- To avoid this possibility the step is multiplied by a scalar λ.

Step size

$$\beta_{t+1} = \beta_t + \lambda(-H_t^{-1})g_t$$

direction



2.1.3 Concavity

- If the LL function is globally concave, then the NR procedure provides an increase in the LL at each iteration.
- If LL is concave then the Hessian is negative definite at all values of β. The slope is declining and the second derivative is negative.
- A symmetric matrix *M* is positive definite if *x'Mx* > 0 for any *x* = 0. Consider a first-order Taylor's approximation of LL(β_{t+1}) around LL(β_t):

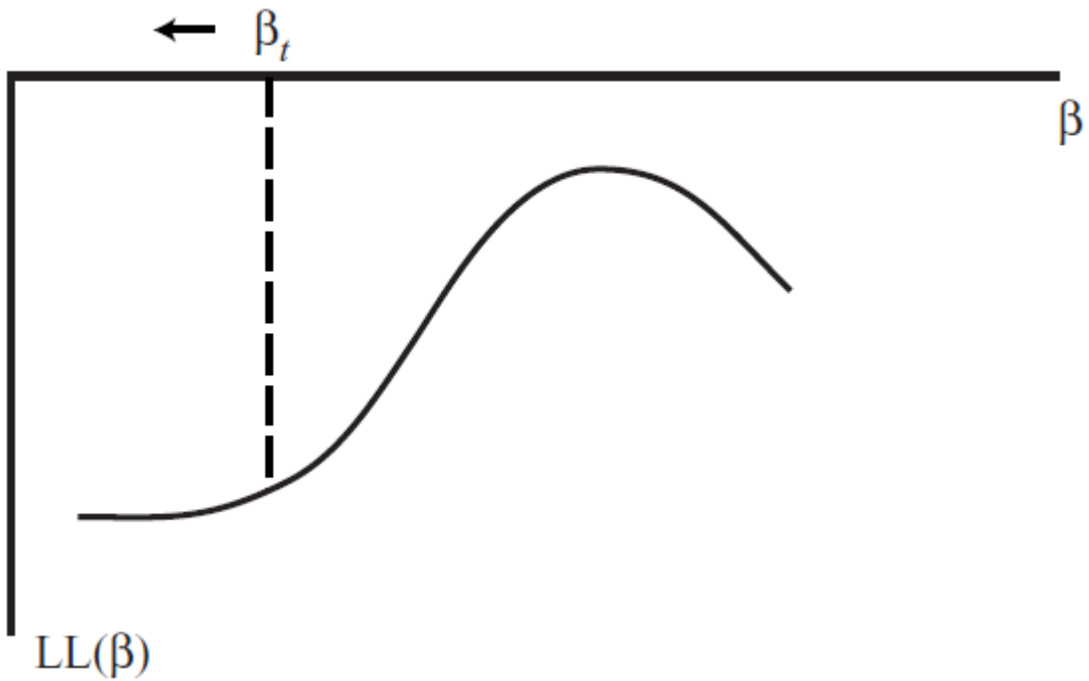
$$LL(\beta_{t+1}) = LL(\beta_t) + (\beta_{t+1} - \beta_t)' g_t \rightarrow \beta_{t+1} - \beta_t = \lambda(-H_t^{-1}) g_t$$

$$LL(\beta_{t+1}) = LL(\beta_t) + \lambda g_t'(-H_t^{-1}) g_t$$

>0

2.1.3 Concavity (cont.)

- If the LL function is NOT concave, then the NR procedure fails to find an increase because the step is in the opposite direction of the slope.



- Other methodologies exists that deal with convex regions.

2.2 Berndt, Hall, Hall, and Hausman (BHHH)

- Maximization can be faster if we utilize the fact that the function being maximized is a sum of terms in a sample.
- The *score* of an observation is the derivative of that observation's log likelihood with respect to the parameters:

$$s_n(\beta_t) = \frac{\partial P_n(\beta)}{\partial \beta} \qquad \text{evaluated at } \beta_t$$

- The gradient is the average score:

$$g_t = \frac{\sum_n s_n(\beta_n)}{N}$$

2.2 BHHH (cont.)

- The outer product of observation n's score is the K x K matrix:

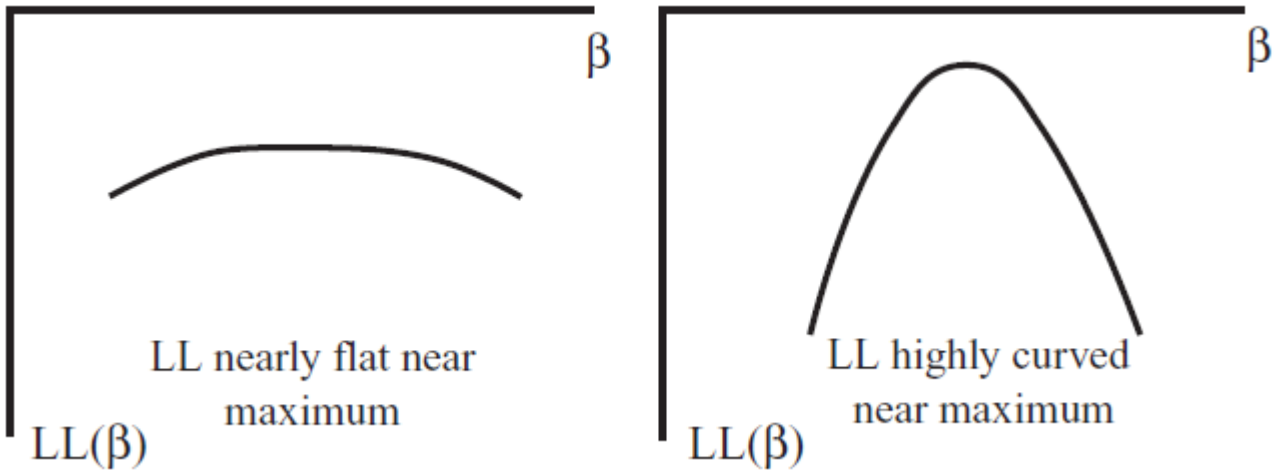
$$s_n(\beta_t)s_n(\beta_t)' = \begin{pmatrix} s_n^1 s_n^1 & s_n^1 s_n^2 & \cdots & s_n^1 s_n^K \\ s_n^1 s_n^2 & s_n^2 s_n^2 & \cdots & s_n^2 s_n^K \\ \vdots & \vdots & & \vdots \\ s_n^1 s_n^K & s_n^2 s_n^K & \cdots & s_n^K s_n^K \end{pmatrix}$$

- The average outer product in the sample is related to the covariance matrix.

$$\beta_t = \frac{\sum_n s_n(\beta_t)\, s_n(\beta_t)'}{N}$$

2.2 BHHH (cont.)

- The maximum occurs where the slope is zero, which means that the gradient (i.e., the average score) is zero and β_t is the variance of scores in the sample.
- The variance of the scores provides a measure of the curvature of the log-likelihood function, similar to the Hessian.



2.2 BHHH (cont.)

- Information Identity:
  - The covariance of the scores at the true parameters is equal to the negative of the expected Hessian.
- BHHH uses β_t in the optimization routine in place of $-H_t$.

$$\beta_{t+1} = \beta_t + \lambda \beta_t^{-1} g_t$$

- This yields two advantages over NR:
  - β_t is faster to calculate than H_t. We don't need to calculate the 2nd derivative.
  - β_t is necessarily positive definite. The BHHH procedure is therefore guaranteed to provide an increase in LL(β) in each iteration, even in convex portions of the function.

2.3 Broyden, Fletcher, Goldfarb, Shanno (BFGS)

- NR and BHHH use information at β_t only to determine the step. This works if the function is quadratic.
- BFGS calculate the approximate Hessian in a way that uses information at more than one point on the LL function to obtain a sense of its curvature.
- An *arc* Hessian can be defined on the basis of how the gradient changes from one point to another.
  - For example, for function *f* (*x*), suppose the slope at *x* = 3 is 25 and at *x* = 4 the slope is 19. The change in slope for a one unit change in *x* is −6. In this case, the arc Hessian is −6, representing the change in the slope as a step is taken from *x* = 3 to *x* = 4.

3. Convergence Criterion

- The maximum of LL(β) occurs when the gradient vector is zero.
- This never happens, but the gradient can be close to zero.
- The static $m_t = g_t'(-H_t^{-1})g_t$ is used to measure convergence. In general, we want ṁ=0.0001.
- The statistic m_t is the test statistic for the hypothesis that all elements of the gradient vector are zero. The statistic is distributed chi-squared with *K* degrees of freedom.
- However, the convergence criterion ṁ is usually set far more lower than the critical value of a chi-squared at standard levels of significance, so as to assure that the estimated parameters are very close to the maximizing values.

3. Convergence Criterion (cont.)

- Small changes in β_t and LL(β_t) accompanied by a gradient vector that is not close to zero indicate that the numerical routine is not effective at finding the maximum.
- Convergence is sometimes assessed on the basis of the gradient vector itself.
- There are two procedures:
  1. Determine whether each element of the gradient vector is smaller in magnitude than some value that the researcher specifies.
  2. Divide each element of the gradient vector by the corresponding element of β, and determine whether each of these quotients is smaller in magnitude than some value specified by the researcher.
- The second approach normalizes for the units of the parameters, which are determined by the units of the variables that enter the model.

4. Local vs Global Maximum

- All of the methods that we have discussed converge at a local maximum that is not the global maximum.
- When the LL function is globally concave, as for logit with linear-in-parameters utility, then there is only one maximum.
- However, most discrete choice models are not globally concave.
- A way to investigate the issue is to use a variety of starting values and observe whether convergence occurs at the same parameter values.

5. Variance of the Estimates

- For correctly specified models:

$$\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, (-\mathbf{H})^{-1}) \quad \text{as N} \rightarrow \infty$$

- Where:
  - β^* is the true parameter vector.
  - $\hat{\beta}$ is the maximum likelihood estimator.
  - **H** is the expected hessian in the population. The negative of the expected Hessian, **–H**, is often called the information matrix.
- The difference between the estimator and the true value, normalized for sample size, converges asymptotically to a normal distribution centered on zero and with covariance equal to the inverse of the information matrix, $-\mathbf{H}^{-1}$.

5. Variance of the Estimates (cont.)

- The asymptotic covariance of $\hat{\beta}$ is $-\mathbf{H}^{-1}/N$.
- For correctly specified models, any of these three are an approximation of **H**:
  - *H* is the average Hessian in the sample.
  - *W* is the covariance of the scores in the sample.
  - *B* is the covariance of the scores, but at the maximizing values of β.
- For non-correctly specified models $\sqrt{N}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1})$
  - **V** is the variance of score in the population.
  - **H** is calculated just at the final iteration.

Drawing from Densities

Part II

Overview

1. Introduction
2. Random Draws
    1. Standard normal and uniform
    2. Transformation of standard normal
    3. Inverse cumulative for univariate densities
    4. Truncated univariate densities
    5. Choleski transformation
3. Variance reduction
    1. Antithetics
    2. Systematic sampling
    3. Halton sequences
    4. Random Halton draws
    5. Scrambled Halton draws

1. Introduction

- Simulation consists of drawing from a density, calculating a statistic for each draw, and averaging the results.
- For example:

$$E = \int t(\varepsilon) * f(\varepsilon)\, d\varepsilon$$

where: t() is the statistic of interest

f() is the density

2. Random Draws

2.1 Standard Normal and Uniform

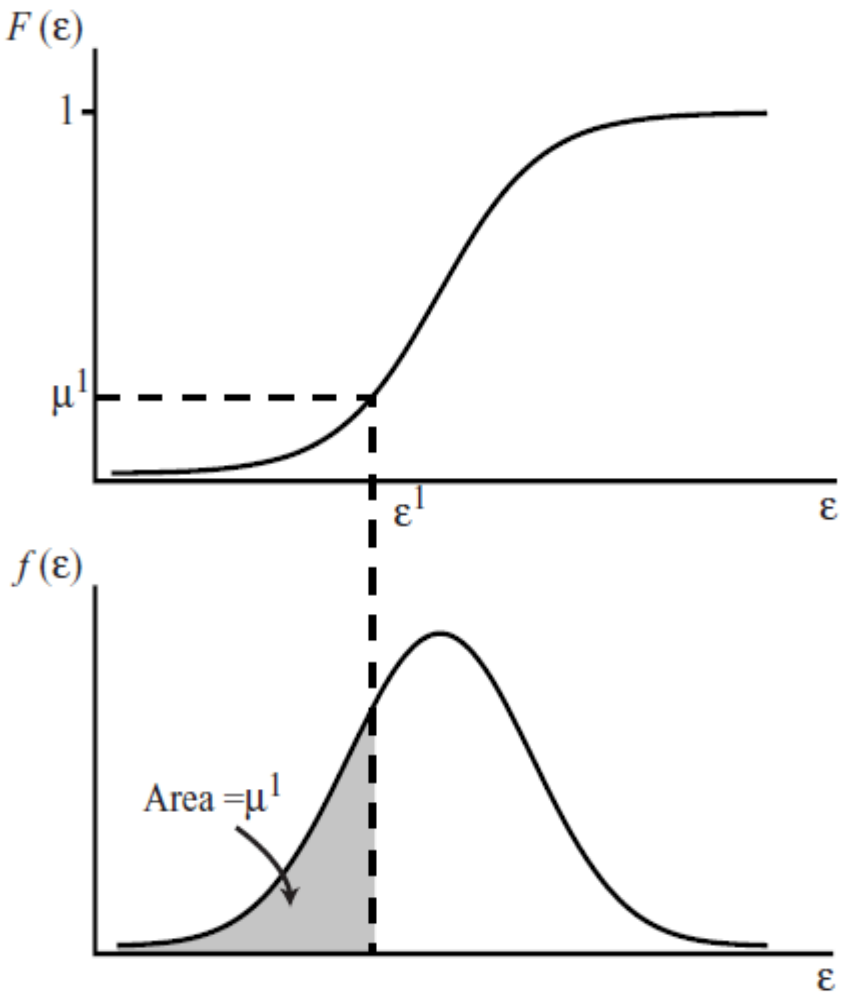
- The researcher can take a draw from a standard normal density η (i.e., a normal with zero mean and unit variance) or a standard uniform density μ (uniform between 0 and 1).
- Most statistical packages contain random number generators for these densities.
- The draws from these routines are actually pseudo-random numbers, because nothing that a computer does is truly random.
- Note that randomness is a theoretical concept that has no operational counterpart in the real world.

2.2 Transformation of Standard Normal

- Some random variables are just transformation of a normal.
- For example:
  - A draw from a normal density with mean b and variance s^2 is: $\varepsilon = b + s\eta$.
  - A draw from a lognormal density is obtained by exponentiating a draw from a normal density: $\varepsilon = e^{b+s\eta}$.
- The moments of the lognormal are functions of the mean and variance of the normal that is exponentiated.
  - Mean = exp[b + (s^2/2)]
  - Variance = exp[2b + s^2]*exp[s^2 - 1]

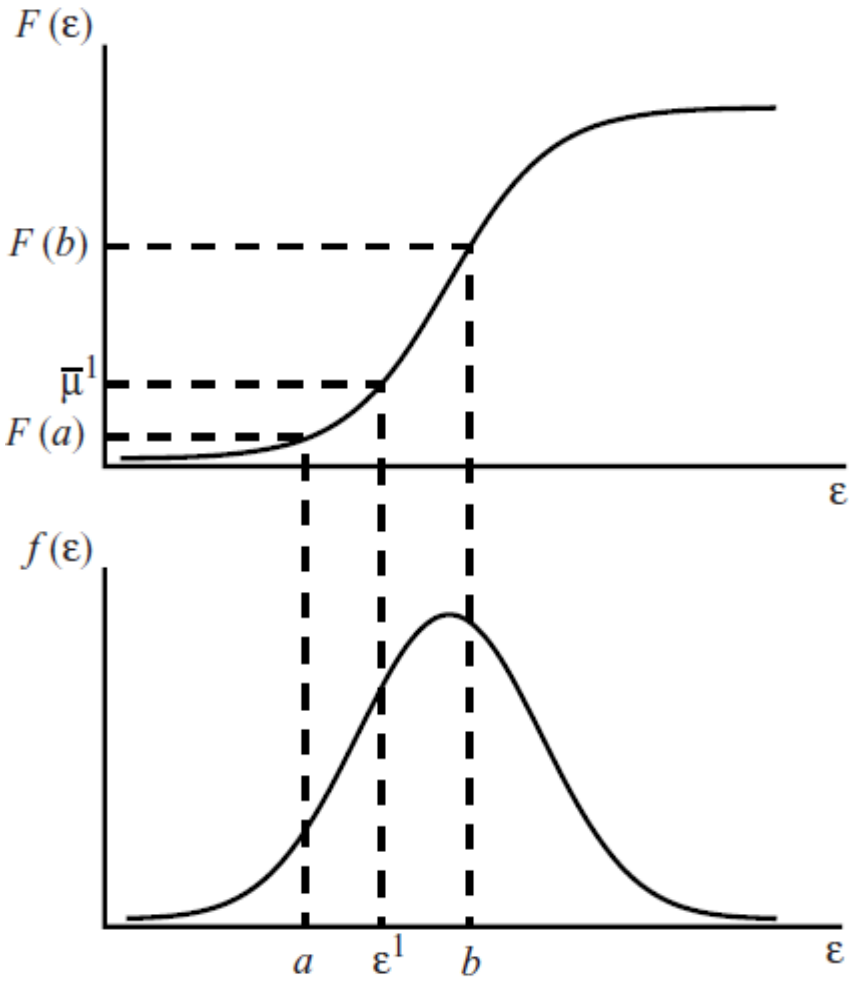
2.3 Inverse Cumulative for Univariate Density

- Consider a random variable with density f(ε) and corresponding cumulative distribution F(ε).
- If F is invertible, F^{-1}, then draws of ε can be obtained from draws of a standard uniform, $\varepsilon = F^{-1}(\mu)$.
- The cumulative distribution of the draws is equal to F, such that the draws are equivalent to draws directly from F.



2.4 Truncated Univariate Densities

- Consider a random variable that ranges from a to b with density proportional to f(ε) within this range.
- The density is (1/k)*f(ε) for a ≤ ε ≤ b, and 0 otherwise.
  - Where $k = \int_a^b f(\varepsilon)d\varepsilon = F(b) - F(a)$
- Draw μ from a standard uniform density. Calculate the weighted average of F(a) and F(b) as $\bar{\mu}$ = (1 − μ)F(a) + μF(b). Then calculate ε = $F^{-1}(\bar{\mu})$.
- Since $\bar{\mu}$ is between F(a) and F(b), ε is necessarily between a and b.



2.5 Choleski Transformation

- It is used to draw from a multivariate normal.
- Let ε be a vector with K elements distributed N(b,Ω).
- The Choleski factor of is defined as a lower-triangular matrix L such that LL'= Ω.
- With K = 1 and variance s^2, the Choleski factor is s (i.e., the standard deviation of ε).
- A draw of ε from N(b, Ω) is obtained as follows. Take K draws from a standard normal, and label the vector of these draws η =[η_1, . . . , η_K]'. Calculate ε = b + Lη.

2.5 Choleski Transformation (cont.)

- Suppose that we want a draw from a three dimensional ε with zero mean.

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \begin{pmatrix} s_{11} & 0 & 0 \\ s_{21} & s_{22} & 0 \\ s_{31} & s_{32} & s_{33} \end{pmatrix} \begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix}$$

- Or simply:
  - $\varepsilon_1 = s_{11}\eta_1$,
  - $\varepsilon_2 = s_{21}\eta_1 + s_{22}\eta_2$,
  - $\varepsilon_3 = s_{31}\eta_1 + s_{32}\eta_2 + s_{33}\eta_3$
- The elements ε_1 and ε_2 are correlated because of the common influence of η_1 on both of them. They are not perfectly correlated because η_2 enters ε_2 without affecting ε_1. Similar analysis applies to ε_1 and ε_3, and ε_2 and ε_3.

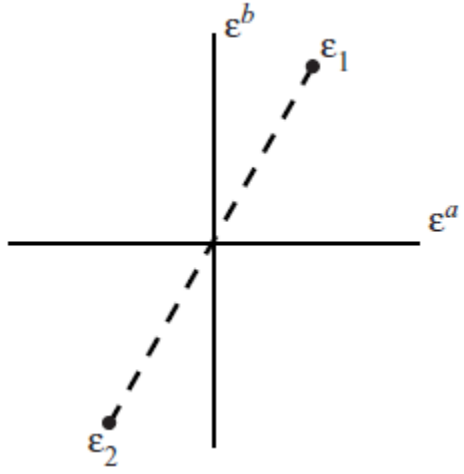
3. Variance Reduction

3. Variance Reduction

- Independent random draws are appealing because it is conceptually straightforward and the statistical properties of the resulting simulator are easy to derive.
- However, there are other ways to take draws that can provide greater accuracy for a given number of draws.
- When solving the integral $\int t(\varepsilon) * f(\varepsilon)\, d\varepsilon$, we want:
  - Coverage: random draws that are spread throughout the domain of f.
  - Covariance: when draws are independent, the covariance over draws is zero. The variance of a simulator based on R independent draws is therefore the variance based on one draw divided by R. If the draws are negatively correlated instead of independent, then the variance of the simulator is lower.

3.1 Antithetics

- Antithetic draws are obtained by creating various types of mirror images of a random draw.
- Suppose a random draw is taken from f (ε) and the value $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$ is obtained.
- The second "draw," which is called the antithetic of the first draw, is created as $\varepsilon_2 = \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'$
- Each draw from f creates a pair of "draws."



3.1 Antithetics (cont.)

- The correlation between a draw and its antithetic variate is exactly −1, then the variance of their sum is zero: $V(\varepsilon 1 + \varepsilon 2) = V(\varepsilon 1) + V(\varepsilon 2) + 2Cov(\varepsilon 1, \varepsilon 2) = 0$.
- This fact does not mean that there is no variance in the simulated probability. The simulated probability is a nonlinear function of the random terms, and so the correlation between $P(\varepsilon 1)$ and $P(\varepsilon 2)$ is less than one.

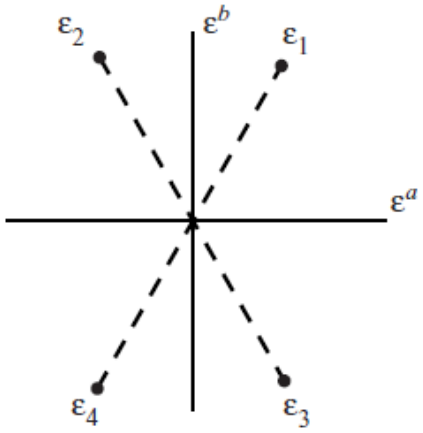
3.1 Antithetics (cont.)

- Reversing the sign of a draw gives evaluation points in opposite quadrants. The concept can be extended to obtain draws in each quadrant by reversing the sign of each element alone (left Figure).
- Better coverage and higher negative correlation can be obtained by shifting the position of each element as well as reversing their signs (right Figure).
- For $\varepsilon_1 = \langle \varepsilon_1^a, \varepsilon_1^b \rangle'$ the antithetic draws are:

$-\varepsilon_2 = \langle -\varepsilon_1^a, \varepsilon_1^b \rangle'$

$-\varepsilon_3 = \langle \varepsilon_1^a, -\varepsilon_1^b \rangle'$

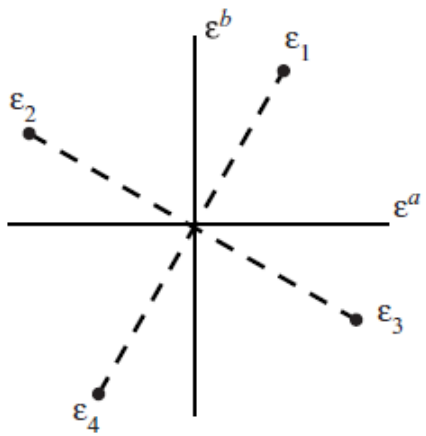
$-\varepsilon_4 = \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'$



$\varepsilon_2 = \langle -\varepsilon_1^b, \varepsilon_1^a \rangle'$

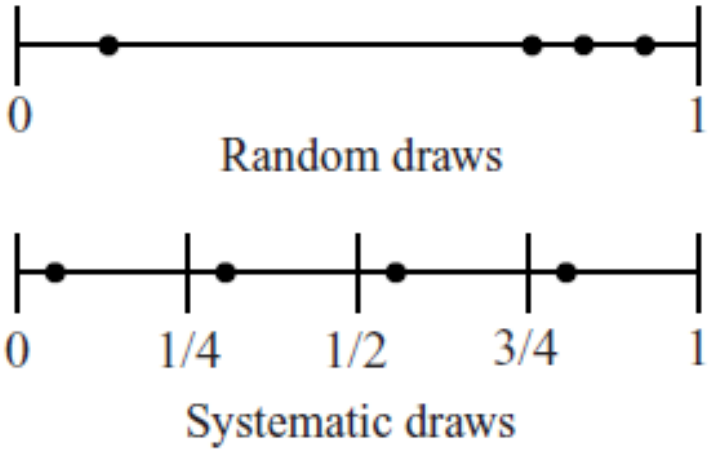
$\varepsilon_3 = \langle \varepsilon_1^b, -\varepsilon_1^a \rangle'$

$\varepsilon_4 = \langle -\varepsilon_1^a, -\varepsilon_1^b \rangle'$



3.2 Systematic Sampling

- Coverage can also be improved through systematic sampling, by dividing the interval into four segments and draws taken in a way that assures one draw in each segment with equal distance between the draws.
- In this case, we take a draw from an uniform between 0 and 0.25, *ε1*, and the others are created as:
  - $\varepsilon 2 = 0.25 + \varepsilon 1$,
  - $\varepsilon 3 = 0.50 + \varepsilon 1$,
  - $\varepsilon 4 = 0.75 + \varepsilon 1$.

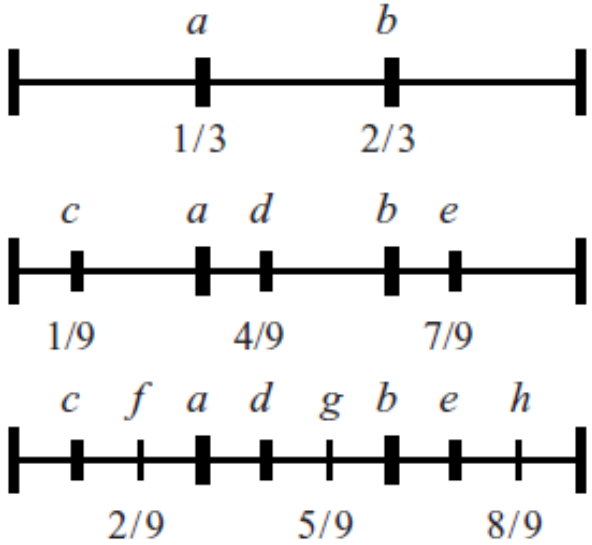


3.2 Systematic Sampling (cont.)

- The issue then arises of how many can I segment the interval?
- To obtain 100 draws, we take one draw *ε1,* between 0 and 0.01, and 99 draws are created from it.
- More segments provide more coverage.
- However, fewer segments provide more randomness.

3.3 Halton Sequences

- Halton sequences provide coverage and induce a negative correlation over observations.
- A Halton sequence is defined in terms of a given number, usually a prime.
- Consider the prime 3.
  - The unit interval is divided into 3 segments.
  - The sequence starts with 1/3 and 2/3.
  - The three segments are divided into thirds.
  - The sequence becomes 1/3, 2/3, 1/9, 4/9, 7/9, 2/9, 5/9, 8/9.



Lower Higher

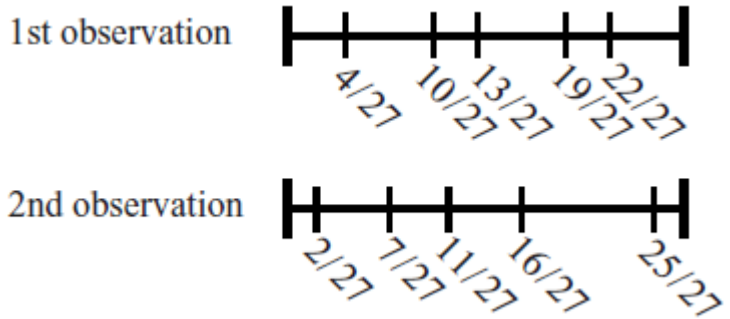
3.3 Halton Sequences (cont.)

- It is easy to create a Halton sequence.
- The sequence is created iteratively. At each iteration t, the sequence is denoted s_t , which is a series of numbers.
- The sequence is extended in each iteration with the new sequence being $s_{t+1} = \{s_t , s_t + 1/3^t , s_t + 2/3^t\}$.
- Starting from zero, the first iteration is:

$$0 = 0,$$
$$1/3 = 1/3,$$
$$2/3 = 2/3,$$
$$0 + 1/9 = 1/9,$$
$$1/3 + 1/9 = 4/9,$$
$$2/3 + 1/9 = 7/9,$$
$$0 + 2/9 = 2/9,$$
$$1/3 + 2/9 = 5/9,$$
$$2/3 + 2/9 = 8/9.$$

3.3 Halton Sequences (cont.)

- One long Halton sequence is usually created and then part of the sequence is used for each observation.
- Suppose there are two observations, and the researcher wants $R = 5$ draws for each.
- We create the sequence: 0, 1/3, 2/3, 1/9, 4/9, 7/9, 2/9, 5/9, 8/9, 1/27, 10/27, 19/27, 4/27, 13/27, 22/27, 7/27, 16/27, 25/27, 2/27, 11/27.
- We eliminate the first 10 elements.
- Then divide the sequence in two.
- The gap left by the first observation is covered by the second.

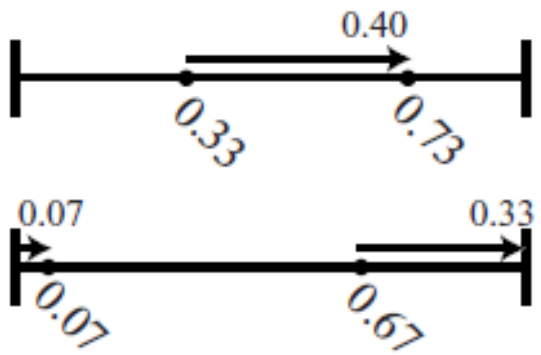


3.3 Halton Sequences (cont.)

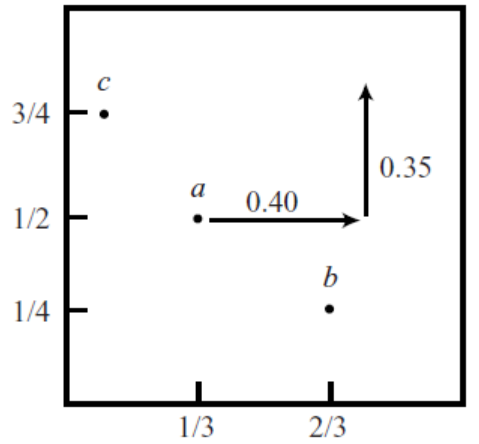
- Halton draws are for a uniform density.
- To obtain a sequence of points for other univariate densities, the inverse cumulative distribution is evaluated at each element of the Halton sequence:
  - $\phi^{-1}(1/3) = -0.43$; $\phi^{-1}(2/3) = 0.43$; $\phi^{-1}(1/9) = -1.2$; $\phi^{-1}(4/9) = -0.14$…
- Halton sequences in multiple dimensions are obtained by using a different prime for each dimension, $\varepsilon_1 = \left\langle \frac{1}{2}, \frac{1}{3} \right\rangle$.
  - It is customary to eliminate the initial part of the series.
  - The initial terms of two Halton sequences are highly correlated, through at least the first cycle of each sequence.

3.4 Random Halton Draws

- Halton sequences can be transformed in a way that makes them random, at least in the same way that pseudorandom numbers are random.
  1. Take a draw from a standard uniform density, μ.
  2. Add μ to each element of the Halton sequence. If the resulting element exceeds 1, subtract 1 from it.
- For μ = 0.40:

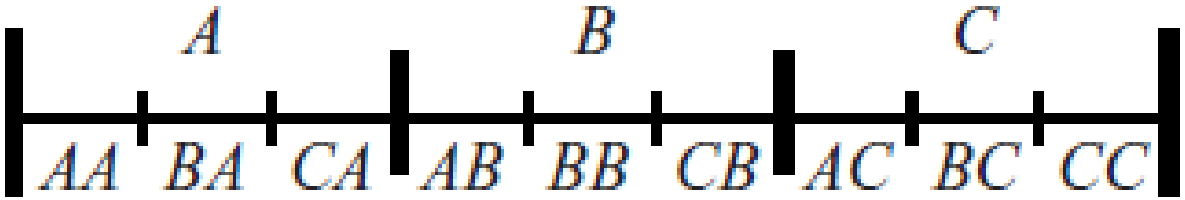


For 2 dimensions:



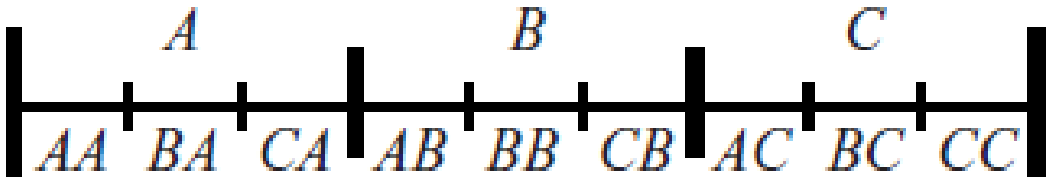
3.5 Scrambled Halton Draws

- Halton draws defined by large primes can be highly correlated with each other over large portions of the sequence.
- This problem can not be eliminated by discarding the initials elements of the sequence, as done earlier.
- This correlation can be removed by scrambling the digits of each element of the sequence.



3.5 Scrambled Halton Draws (cont.)

- Consider the Halton sequence for prime 3: 1/3, 2/3, 1/9, 4/9, 7/9, 2/9, 5/9, 8/9,…



- The Halton sequence is the starting point of each segment arranged alphabetically and ignoring *A* (i.e., ignore *A*, 1/3 for *B*, 2/3 for *C*), followed by the starting point of each subsegment arranged alphabetically and ignoring *A* (i.e., ignore *AA*, *AB*, and *AC*, 1/9 for *BA*, 4/9 for *BB*, 7/9 for *BC*, 2/9 for *CA*, 5/9 for *CB*, and 8/9 for *CC*.)
- The scrambled sequence is obtained by reversing B and C, that is, by considering C to be before B in the alphabet. The alphabetical listing is now: segments A C B, subsegments AA AC AB CA CC CB BA BC BB.

Simulated Assisted Estimation

Part III

Overview

1. Introduction
2. Definition of Estimators
    1. Maximum Simulated Likelihood
    2. Method of Simulated Scores
3. The Central Limit Theorem
4. Properties of Traditional Estimators
5. Properties of Simulation-Based Estimators

1. Introduction

- So far we have examined how to simulate choice probabilities but have not investigated the properties of the parameter estimators that are based on these simulated probabilities.
- The resulting estimator should have desirable properties, such as consistency, asymptotic normality, or efficiency.
- We will examine various methods of estimation in the context of simulation, derive their properties and show the conditions under which each estimator is consistent and asymptotically equivalent to the estimator that would arise with exact values.
- These conditions provide guidance to the researcher on how the simulation needs to be performed to obtain desirable properties of the resultant estimator.

1. Introduction (cont.)

We consider three methods of estimation:

1. Maximum Simulated Likelihood (MSL)

   Same as maximum likelihood (ML) except that simulated probabilities are used in lieu of the exact probabilities. See Gourieroux and Monfort,(1993), Lee(1995), and Hajivassiliou and Ruud (1994).

2. Method of Simulated Moments (MSM)

   Traditionally, residuals are the 0–1 dependent variable that identifies the chosen alternative and the probability of the alternative. The estimates are the parameters values that make the variables and residuals uncorrelated in the sample. Simulated moments just replace probabilities with simulated probabilities.

3. Method of Simulated Scores (MSS)

   The score is the gradient of the LL of an observation. The method of scores finds the parameter values that set the average score to zero. When exact probabilities are used, the method of scores is the same as maximum likelihood. Depending on how the scores are simulated, MSS can differ from MSL and, importantly, can attain consistency and efficiency under more relaxed conditions

2. Definition of Estimators

2.1 Maximum Simulated Likelihood

- The log-likelihood function is $LL(\theta) = \sum_{n=1}^{N} \ln P_n(\theta)$
  - Where: θ is the vector of parameters.
    Pn is the exact probability of the observed choice.
    N is the number of independent observations.
- The ML estimator is the value of θ that maximizes LL(θ). Since the gradient of LL(θ) is zero at the maximum, the ML estimator can also be defined as the value of θ at which $\sum_n s_n(\theta) = 0$
  - Where $s_n(\theta) = \frac{\partial \ln P_n(\theta)}{\partial \theta}$ is the score for observation *n*.

2.1 Maximum Simulated Likelihood (cont.)

- Let $\check{P}_n(\theta)$ be a simulated approximation to $P_n(\theta)$.
- The simulated likelihood function is $SLL(\theta) = \sum_{n=1} \ln \check{P}_n(\theta)$.
- The MSL estimator is the value of θ at which $\sum_n \check{s}_n(\theta) = 0$, where $\check{s}_n(\theta) = \frac{\partial \ln \check{P}_n(\theta)}{\partial \theta}$.
- The problem is the log transformation.
- If $\check{P}_n(\theta)$ is an unbiased estimator of $P_n(\theta)$, then the expectation over draws used in the simulation is $E_r \check{P}_n(\theta) = P_n(\theta)$.
- Given that the log operation is a nonlinear transformation, $\ln \check{P}_n(\theta)$ is not unbiased for $\ln P_n(\theta)$.
- This bias diminishes as more draws are used in the simulation.

2.1 Maximum Simulated Likelihood (cont.)

- We want to determine the asymptotic properties of the MSL estimator when the sample size rises.
- This depends on the relationship between the number of draws used in the simulation, *R*, and the sample size, *N*.
- If R is fixed, the MSL estimator does not converge to the true parameters.
- If R rises with N, then the bias disappears as N (and R) rises without bound.
- If R rises faster than $\sqrt{N}$, then the MSL is estimator is also efficient.

2.2 Method of Simulated Scores

- MSS provides consistency without a loss of efficiency. However, MSS have poor numerical properties, which makes it difficult to calculate the estimator.
- When exact probabilities are used, the method of scores is the same as maximum likelihood. Now, if $\check{s}_n(\theta)$ is calculated as the derivative of the log of the simulated probability, then MSL = MSS.
- If we can construct an unbiased simulator of the score, then when we define $\sum_n \check{s}_n(\theta) = 0$, it does not have any bias because this is a linear equation.
- MSS is consistent with a fixed R, whereas the simulation noise decreases as R rises.
- MSS is asymptotically efficient when R rises with N, instead of $\sqrt{N}$.

2.2 Method of Simulated Scores (cont.)

- How do we find an unbiased score simulator?

$$s_n(\theta) = \frac{\partial \ln P_{nj}(\theta)}{\partial \theta} = \frac{1}{P_{nj}(\theta)} * \frac{\partial P_{nj}(\theta)}{\partial \theta}$$

- Since differentiation is a linear operation, an unbiased simulator of $\frac{\partial P_{nj}(\theta)}{\partial \theta}$ is easily obtained by taking the derivative of the simulated probability.
- The problem is then $\frac{1}{P_{nj}(\theta)}$ since an inverse introduces bias.

2.2 Method of Simulated Scores (cont.)

- $P_{nj}(\theta)$ is the probability that a draw of the random terms of the model will result in alternative j having the highest utility.
- The inverse $1/P_{nj}(\theta)$ can be simulated as follows:
  1. Take a draw of the random terms from their density.
  2. Calculate the utility of each alternative with this draw.
  3. Determine whether alternative j has the highest utility.
  4. If so, call the draw an accept. If not, then call the draw a reject and repeat steps 1 to 3 with a new draw. Define B^r as the number of draws that are taken until the first accept is obtained.
  5. Perform steps 1 to 4 R times, obtaining B^r for r = 1, . . . , R. The simulator of $1/P_{nj}(\theta)$ is $\frac{1}{\mathrm{R}} * \sum_{r=1}^{R} B^r$.

3. Central Limit Theorem

3. Central Limit Theorem

- If we take draws from a distribution with mean μ and variance σ, the mean of these draws will be normally distributed with mean μ and variance σ/N, where N is a large number of draws.
- Suppose t_n is a draw from a distribution with mean μ and variance σ. Then, the sample mean is $\mathrm{t} = \frac{1}{N}\sum_n t_n$.
- We want to derive the sampling distribution of t. In some cases, as the sample sizes rises, the sampling distribution of statistic t converges to a fixed distribution (e.g., normal).
- In the case of converging close to normal, we say that $t \xrightarrow{d} N(t^*, \sigma)$, which is the limiting distribution of t.
- In many cases, a statistic will not have a limiting distribution. As *N* rises, the sampling distribution keeps changing.

3. Central Limit Theorem (cont.)

- If our statistic does not have a limiting distribution we can transform it in such a way that it has a limiting distribution.
- Consider $\sqrt{N}(t-\mu)$ and suppose this statistic has a limiting distribution $\sqrt{N}(t-\mu) \xrightarrow{d} N(0,\sigma)$.
- Recall that for fixed a and b, if a(t − b) is distributed normal with zero mean and variance σ, then t itself is distributed normal with mean b and variance σ/a^2.
- If $\sqrt{N}(t-\mu)$ is distributed N(0, σ) for large N, then t is distributed $t \sim N(\mu, \frac{\sigma}{N})$. Note that this is not the limiting distribution of t; rather, it is called the asymptotic distribution of t.

4. Properties of Traditional Estimators

4. Properties of Traditional Estimators

- The true value of the parameters is denoted θ^*. The ML estimators are root of an equation that takes the form $\sum_n g_n(\hat{\theta})/N = 0$, where $g_n(\theta)$ is the score $\partial \ln P_n(\theta)/\partial\theta$.
- The parameters that solve the equation are the estimators.
- We are interested in the sample mean and variance of $g_n(\theta)$ at the true parameters. Let's label them as $g(\theta)$ and $W(\theta)$, respectively.
- Label the mean of $g_n(\theta^*)$ in the population as **g** and its variance in the population as **W**, and assume that **g** = 0.
- Then, $\hat{\theta}$ is the value of the parameters at which the sample average of $g_n(\theta)$ equals zero.

4. Properties of Traditional Estimators (cont.)

- The information identity states that **V** = −**H**.
  - where:

    $-\boldsymbol{H} = -E\left(\frac{\partial^2 \ln P_n(\theta^*)}{\partial\theta\,\partial\theta'}\right)$ is the information matrix

    V= Var(∂ ln P_n(θ*)/∂θ) is the variance of the scores evaluated at the true parameters.
- When g_n(θ) is the score, we have **W** = **V** by definition and hence **W** = −**H** by the information identity.

4. Properties of Traditional Estimators (cont.)

- **Step 1:**
  - Recall that $g_n(\theta^*)$ varies over decision makers.
  - When taking a sample, the researcher is drawing values of $g_n(\theta^*)$. This distribution has zero mean by assumption and variance denoted **W**.
  - By the central limit theorem: $\sqrt{N}(g_n(\theta^*) - 0) \xrightarrow{d} N(0, \boldsymbol{W})$.
  - The sample mean has distribution $g_n(\theta^*) \sim N(\mu, \frac{\boldsymbol{W}}{N})$.

4. Properties of Traditional Estimators (cont.)

- **Step 2:**
  - Take a first-order Taylor's expansion of $g_n(\hat{\theta})$ around $g_n(\theta^*)$
    - $g_n(\hat{\theta}) = g_n(\theta^*) + D[\hat{\theta} - \theta^*]$ where D = ∂g(θ^*)/∂θ'
  - g($\hat{\theta}$) = 0 so that the left-hand side of this expansion is 0.
    - $0 = g_n(\theta^*) + D[\hat{\theta} - \theta^*]$
    - $\hat{\theta} - \theta^* = -D^{-1} g_n(\theta^*)$
    - $\sqrt{N}(\hat{\theta} - \theta^*) = \sqrt{N}(-D^{-1}) g_n(\theta^*)$ where D is the mean of ∂g(θ^*)/∂θ'
    - We know from Step 1 that $\sqrt{N} g_n(\theta^*) \xrightarrow{d} N(0, \boldsymbol{W})$, hence $\sqrt{N}(\hat{\theta} - \theta^*) = \sqrt{N}(0, D^{-1}\boldsymbol{W}D^{-1})$
    - The limiting distribution tells us that $\hat{\theta} \sim N(\theta^*, D^{-1}\boldsymbol{W}D^{-1}/\mathrm{N})$

4. Properties of Traditional Estimators (cont.)

- **Step 2 (cont.):**
  - The asymptotic distribution of $\hat{\theta}$ is centered on the true value, and its variance decreases as the sample size rises.
  - $\hat{\theta}$ converges in probability to θ^* as the sample size rises without bound: $\hat{\theta} \xrightarrow{p} \theta$.
  - The estimator is therefore consistent.
  - The estimator is asymptotically normal.
  - And its variance is $D^{-1}WD^{-1}/N$, which can be compared with the lowest possible variance, $-H^{-1}/N$, to determine whether it is efficient.

5. Properties of Simulation-Based Estimators

5. Properties of Simulation-Based Estimators

- The properties of MSL can be summarized as follows:
  1. If R is fixed, MSL is inconsistent.
  2. If R rises slower than $\sqrt{N}$, MSL is consistent but not asymptotically normal.
  3. If R rises faster than $\sqrt{N}$, MSL is consistent, asymptotically normal and efficient, and equivalent to ML.