How Monte Carlo Sampling Contributes to Data Analysis

Eric Slud, Mathematics Department, UMCP

Objective: to explain an "experimental" approach to Probability & Statistics via Simulation

Outline

I. Definition of Probability & Simulation

II. Simulation-based estimation of Probabilities

III. Simulation in relation to Data: Histograms and Densities

IV. Resampling from Data: Why Do It ?

What is Probability ?

- A rule for assigning numbers between 0 and 1 to *Events*

- Obeys combination rules same as Relative Frequencies

  1 means "certain to occur"

  Prob's add for unions of disjoint events

- Definition of Relative Frequency:

  for Random Experiment repeated N times,
  independently (*with mutual non-interference*)
  by same mechanism, and event E occurs $n(E)$ times,

  its **relative frequency** of occurrence is $n(E)/N$.

Example: Dice-Throwing

'Experiment' : tossing pair of dice independently

On 1 toss, 36 outcomes $(1,1),(2,1),\ldots,(6,1),(1,2),\ldots,(6,6)$

'Event' $E = A \cup B$: sum of dots 7 or 11

7 dots : $A = \{(1,6),(2,5),\ldots,(5,2),(6,1)\}$

11 dots : $B = \{(5,6),\ (6,5)\}$

$$P(7 \text{ or } 11) = P(A \cup B) = P(A) + P(B)$$
$$= 6/36 + 2/36$$

Use independently repeated experiments, to reach clear predictions.

Probability as Limiting Relative Frequency

Probability axioms are obeyed by relative frequencies.

Formal mathematics definition of Probability as Set-Function plus def'n of **independent identical-mechanism replications**

$$E_1, E_2, \ldots, E_N \; : \quad \begin{cases} P(E_i) \text{ same for all } i \text{ , and for j's distinct} \\ P(E_{j_1} \cap E_{j_2} \cap \cdots \cap E_{j_k}) \;=\; P(E_{j_1}) \cdots P(E_{j_k}) \end{cases}$$

leads to mathematical theorem **Law of Large Numbers** saying:

$$\text{as} \quad N \to \infty, \qquad \frac{1}{N} \sum_{j=1}^{N} I[E_j] \;\to\; P(E_1)$$

We want to implement this computationally !

What is Monte Carlo Simulation ?

Ingredient #1: **Dynamical Random Number Generator**

- Recursive rule $x_{n+1} = f(x_n)$ operating on fixed-length vectors x_n of integers, plus simple mapping $g : x_n \mapsto U_n$ so that $U_1, U_2, \ldots, U_N$ behaves like independent identically distributed random variables Uniformly distributed in $(0, 1)$

Classic example: $x_n = 0, \ldots, 2^{31} - 1$ $\qquad U_n = x_n / 2^{31}$

- *Linear Congruential:* $x_{n+1} = a \cdot x_n + b \bmod m$

$$a = 7^5 \,,\ b = 0 \,,\ m = 2^{31} - 1$$

(Park & Miller, *Trans. ACM. 1988*)

Demo 1A & B & C

Defining 'Simulation', cont'd

Ingredient #2: Expression of desired data structure:
Data as function of Building Block Uniform(0,1) r.v.'s

Examples: (a) Drawing from a list $1 \ldots 23$ *with replacement*

`Uvec = runif(100)` gives 100-vector Uvec which can be treated as indep. Unif[0,1]

```
Xvec = trunc(23*Uvec) + 1
X = 1 + greatest integer <= 23*U
```

(b) How would you code 100 independent random selections from $1 \ldots 230$ *with* replacement ?

(c) Selections of 100 from $1 \ldots 230$ **without replacement** ?

More on Defining ‘Simulation’

Ingredient 2, cont’d: coding ‘data’ from indep. U_n

(d) **5-card poker hands:** 5 w.o. replacement from $1 \ldots 52$

```
Xvec = trunc(52*Uvec)+1          > Poker
Xnew = unique(Xvec)[1:5]          Clubs   Diam    Heart   Spad
Cards = 1+(Xvec-1) %% 13         "2.Cl"  "2.Di"  "2.He"  "2.Sp"
Hand = Poker[Xnew]               "3.Cl"  "3.Di"  "3.He"  "3.Sp"
Pairs = sum(table(Cards)==2)     "4.Cl"  "4.Di"  "4.He"  "4.Sp"
                                  ...
```

In Hand of 5 cards, tabulate # pairs among card values 2. . .A

‘Simulation’, Ingredient 3

Question or *Event* specification or *Variable* to Average.

(e) **Geometric Probability:** what fraction of random points in the Unit Square fall in Inscribed Circle ?

Coding: `Uvec, Vvec` vectors of X and Y coordinates

Variable: `DistSq` $= (\text{Uvec} - .5)^2 + (\text{Vvec} - .5)^2$

Question: `InCirc` $=$ (DistSq $< 1/4$)

Proportion in Circle $=$ Area $= \pi/4 = .78540$

Average $Dist^2 = \int_0^1 \int_0^1 \left\{(u - .5)^2 + (v - .5)^2\right\} du\, dv = 1/6$

Demo 2

Data from Examples

Poker: Question is prob of 2 pairs, `xxyyz`

Combinatorial answer is: $\frac{1}{\binom{52}{5}}\binom{13}{2}\binom{4}{2}\binom{4}{2}44 = 0.047539$

```
 3 Runs, each with 10^5 simulated hands:

Run 1:  4793  of  1e5  had 2 pairs: estimated prob = .04793

Run 2:  Tally of # pairs is :      0      1      2
                                52669  42504   4827
Run 3:  Tally of # pairs is :      0      1      2
                                52880  42341   4779
```

Data from Geom. Prob Example

In successive runs of N randomly generated points in Unit Square:

Run#	N	Radius	InCirc	AvDistSq
1	1e5	.5	0.7871300	0.1662735
2	1e5	.5	0.7855700	0.1664821
3	1e6	.5	0.7849080	0.1668259
4	1e6	1/3	0.3491230	0.1667787
5	1e6	1/3	0.349002	0.166720

Worksheet Questions. #1. Find a single best estimate from these Data for the probability of a random point falling in the Inscribed Circle, of radius 1/2 about (1/2,1/2) ?

#2. Can you account for the relative frequencies with which random points fall in the circle of radius 1/3 about (1/2, 1/2) ?

Conditional Prob's via Simulation

Conditional questions come up naturally:
condition determines denominator !

Example. Conditional prob. $X \in (.2, .6)$ given $Y \in (.3, .8)$:

(A) if (X,Y) random in the square

(B) if (X,Y) random in the circle $(X - .5)^2 + (Y - .5)^2 < 0.25$

(C) if (X,Y) random in the triangle $X < Y$

Simulations show the difference ! `CondProb Demo`

Further Worksheet Questions

#3. What is the exact conditional probability of
(or relative area of region with)
$X \in (.2, .6)$ given $Y \in (.3, .8)$ for a random point
(X, Y) in the triangular region $0 \le X < Y \le 1$?

#4. Since all of these simulations must be programmed:
how might one tell that there are errors in the program,
or that the random number generator is not behaving
properly ?

This is a probability related question: but we have not touched on the theoretical idea yet: that comes next.

Law of Large Numbers

If $X_1, X_2, \ldots, X_N$ *are bounded random variables, independent and identically distributed, then*

$$P\left(|(X_1 + \cdots + X_N)/N - E(X_1)| > \epsilon\right) \rightarrow 0$$

as $N \rightarrow \infty$, *for each* $\epsilon > 0$.

Key example: $X_i = \{0, 1\}$ indicator that event E occurs in i'th replicated dataset. Then $E(X_1) = P(E)$, *Avg* $=$ *Prob*.

So the LLN lets us make a prediction: if we think a simulation is erratic because of inadequate sample size, then it ought to settle down to stable results with larger N.

Large N Behavior of Estimate $\widehat{p}$

Picture in `CumPoker Demo`

shows estimated fraction of points falling within circle of radius 1/9 about (1/2, 1/2) as number of points N in unit square grows.

To get quantitative idea of errors & variability in simulation averages for a particular N, we next appeal to the **Central Limit Theorem**.

Central Limit Theorem (CLT)

With N indep. repetitions and true probability $p = P(E)$

$S_N = n(E) =$ # occurrences of E

has **Binomial**(N, p) distribution

mean Np, and 'standard deviation' $\sqrt{Np(1-p)}$

The CLT says $(S_N - Np)/\sqrt{Np(1-p)}$ behaves for large N like a 'standard normal' $\mathcal{N}(0, 1)$ distributed random variable,

falling between ± 1 w.p. .68, ± 2 w.p. .95,

± 2.58 w.p. .99, ± 3.29 w.p. .999

Precision Bounds for Relative Frequencies

So if we simulate N replications $E_1, \ldots, E_n$ of event E

and use relative frequency $\widehat{p}_N = \frac{1}{N} \sum_{i=1}^{N} I[E_i \text{ occurs }]$

to estimate $p = P(E)$, then

$$\frac{|\widehat{p}_N - p|}{\sqrt{p(1-p)}} \quad \text{is bounded by} \quad \begin{cases} 1.96/\sqrt{N} & w.p.\ 0.95 \\ 2.576/\sqrt{N} & w.p.\ 0.99 \\ 3.291/\sqrt{N} & w.p.\ 0.999 \end{cases}$$

Even when true p is unknown, w.p. $\geq .999$, successive $\widehat{p}_N$ from separate simulation batches of size N cannot be farther apart than $\sqrt{4p(1-p)} \cdot 3.291/\sqrt{N} \leq 3.291/\sqrt{N}$

(*This relates to Worksheet Question #4 above.*)

Application of Precision Bounds

Recall data from 3 runs of 10^5 simulated Poker Hands:

Run 1 $\widehat{p} = .04793$; **Run 2** $\widehat{p} = .04827$; **Run 3** $\widehat{p} = .04779$

With true $p \approx .048$, find 99% precision bounds

$$2.576\sqrt{(.048)(.952)/1e5} = 0.00174$$

(Multiply by $\sqrt{2}$ to bracket pairwise differences.)

Combine all three runs `(N=3e5)` by averaging, to get .04800 with .999 precision bound $3.291\sqrt{(.048)(.952)/3e5} = .00128$.

Exact 2-pair prob. $= 0.047539$, well within bounds.

```
(.04800-.047539)/sqrt((.048)*(.952)/3e5) = 1.181
```

is a perfectly unexceptional normal deviate.

Definitions: Density & Histogram

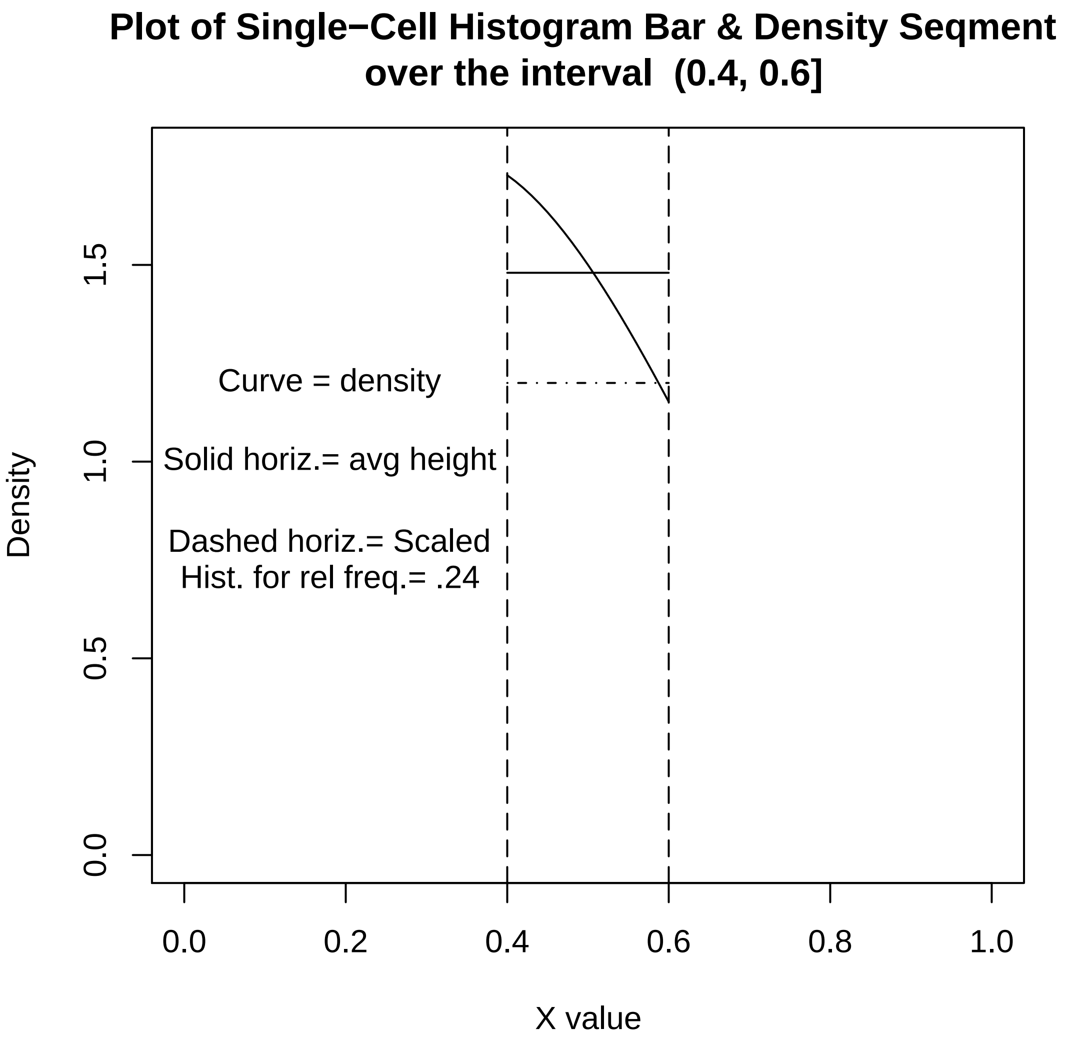
Probability Density: function $f \geq 0$, with $\int_{-\infty}^{\infty} f(x)\,dx = 1$
With random variable following density f

$$\texttt{Area under f over (a,b]} = \int_a^b f(x)\,dx = P(a < X \leq b)$$

Scaled Rel. Freq. Histogram: based on counts $n_1, n_2, \ldots, n_L$ of **numbers of** variable values $X_1, X_2, \ldots, X_N$ resp. falling into (equal-length) intervals $(jh, (j+1)h]$.

Histogram: $$g(x) = \frac{n_j}{Nh} \quad \text{for} \quad jh < x \leq (j+1)h$$

(*Scaling makes total area under* g *equal to* 1.)



Plot of Single−Cell Histogram Bar & Density Seqment
over the interval (0.4, 0.6]
Curve = density
Solid horiz.= avg height
Dashed horiz.= Scaled
Hist. for rel freq.= .24
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
X value

Relationship: Density vs. Histogram

Suppose $X_1, \ldots, X_N$ data points, tallied for histogram,
with n_j values falling between $jh, (j+1)h$.

If f is true density for the X's, then LLN says for large N :

$$\frac{n_j}{N} \approx P(jh < X_1 \leq (j+1)h) = \int_{jh}^{(j+1)h} f(x)dx$$

But the j'th Scaled Histogram Bar is then

$$\frac{n_j}{Nh} \approx \frac{1}{h} \int_{jh}^{(j+1)h} f(x)dx = \text{Avg.Density Height in Cell}$$

which is close to $f(jh)$ when h is small !

HistDemo Illustrations

Further Worksheet Problems

#5. Suppose we do a simulation with $N = 2000$ iterations to evaluate a probability p which (an initial few simulations show) is in the neighborhood of 0.2. What is the 99% precision bound for the estimate (*i.e., the upper bound on* $\widehat{p} - p$ *which holds with approximate probability* 0.99) ?

#6. A certain type of density g is positive only on the interval $[0,1]$ and has a constant value $g_j \le 3$ on each of the intervals $(j/20, (j+1)/20]$. Random variable values $Y_1, \ldots, Y_N$ are observed, with $N = 1000$. How accurate are the histogram bar heights as estimates of g_j, if you can tolerate a probability of error of 0.01 in your precision bounds ?

Sampling From Data

Consider pictured data values $(X_1, Y_1), (X_2, Y_2) \ldots, (X_{299}, Y_{299})$ (measured waiting times between and durations of ‘Old Faithful’ geyser eruptions).

Next suppose we generate $N = 100$ batches of size up to 299 by independently sampling **with replacement** from the observed dataset. Can study typical behavior of various statistics, e.g.

`median` X, `inter-quartile range of` Y ,
`‘best-fitting line slopes’`

GeysPlot, GeysLines, MedSamp Demos

Further Data Analysis via Resampling

Geyser Data: compare mean duration within

groups defined by : $\begin{cases} \texttt{Duration}>3 \\ \texttt{Duration}\leq 3 \end{cases}$

Also compare `Duration` $\times$ `Wait` lines within these groups !

`GeysPlot` `Picture`

References

Google **Random Number Generation**

http://en.wikipedia.org/wiki/Random_number_generator

Diaconis, P. & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American* May, 116-130.

Lecture slides at: http://www.math.umd.edu/∼evs/MMIslid09.pdf .

Visit the **R project** website http://www.r-project.org/ for freely downloadable software !

Scripts for R code in demos at:

http://www.math.umd.edu/ evs/MMIscriptR.txt

More on Central Limit Theorem

Previously discussed CLT for (relative) freq. counts (binomial random variables). Suppose we estimate parameter ϑ like mean or median or best-fitting line slopes from 'statistic' T.

Recalculate statistic values $T_1, \ldots, T_N$ from indep. batches of data. Then sample mean $\bar{T} = (T_1 + \cdots + T_N)/N$ accurately estimates $E(T)$ (may be different from ϑ !),

and $s_T^2 = \frac{1}{N-1} \sum_{i=1}^{N} (T_i - \bar{T})^2$ estimates $\text{Var}(T)$

CLT says $\bar{T} \approx E(T_1) + \frac{Z}{\sqrt{N}} s_T \quad , \quad Z \sim \mathcal{N}(0,1)$