The FULL MONTE CARLO: A LIVE PERFORMANCE (with STARS)

Xiao-Li Meng (孟晓犁)
Department of Statistics
Harvard University

Desired and Feared—What Do We Do Now and Over the Next 50 Years?

Xiao-Li MENG

An intense debate about Harvard University's General Education Curriculum demonstrates that statistics, as a discipline, is now both desired and feared. With this new status comes a set of enormous challenges. We no longer simply enjoy the privilege of playing in or cleaning up everyone's backyard. We are now being invited into everyone's study or living room, and trusted with the task of being their offspring's first quantitative nanny. Are we up to such a nerve-wracking task, given the insignificant size of our profession relative to the sheer number of our hosts and their progeny? Echoing Brown and Kass's "What Is Statistics?" (2009), this article further suggests ways to prepare our profession to meet the ever-increasing demand, in terms of both quantity and quality. Discussed are (1) the need to supplement our graduate curricula with a *professional development curriculum* (*PDC*); (2) the need to develop more *subject oriented statistics* (*SOS*) *courses* and *happy courses* at the undergraduate level; (3) the need to have the most qualified statisticians—in terms of both teaching and research credentials—to teach introductory statistical courses, especially those for other disciplines; (4) the need to deepen our foundation while expanding our horizon in both teaching and research; and (5) the need to greatly increase the general awareness and avoidance of unprincipled data analysis methods, through our practice and teaching, as a way to combat "incentive bias," a main culprit of false discoveries in science, misleading information in media, and misguided policies in society.

KEY WORDS: Communication skills; General education curriculum.

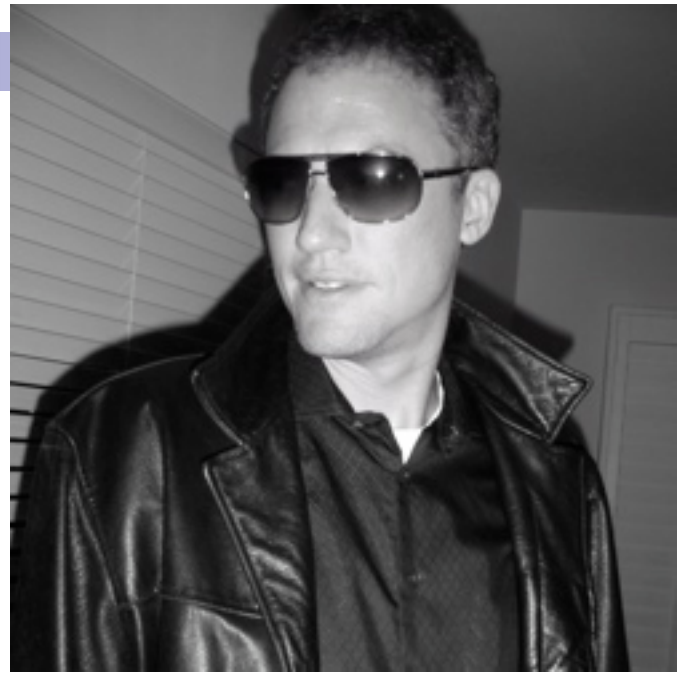
new General Education (Gen Ed) curriculum. One of the initial categories of Gen Ed was *Empirical Reasoning*, with the following proposed requirement. Courses in this category must:

a. teach how to gather and assess empirical data, weigh evidence, understand estimates of probabilities, draw inferences from the data available, and also recognize when an issue cannot be settled on the basis of the available evidence;
b. teach the conceptual and theoretical tools used in reasoning and problem solving, such as statistics, probability theory, mathematics, logic, and decision theory;
c. provide exercises in which students apply these tools to concrete problems in an area of general interest to undergraduates; and
d. where practicable, familiarize students with some of the mistakes human beings typically make in reasoning and problem-solving.

Pleasantly surprised by this proposal, I wanted to know which of my statistical colleagues were involved in drafting it. So did my colleagues, as they thought that I must have had a hand in this, representing our department. Given the language, particularly (a), it is not illogical to infer a statistician's involvement.

No statisticians, at least by the current definition, were involved. It was written by several social and natural scientists. Naturally, my colleagues and I were delighted, at least until the FAS faculty meeting in which it was voted on. With the support from social and natural scientists, surely it would pass with flying colors, right? Quite the contrary—it was defeated! Our academic relatives in mathematics, applied mathematics,

CHASC or CBASC



- **Current CHASC Participants:**
- Paul Baines , Harvard University
- James Chiang, Stanford Linear Accelerator Center
- Alanna Connors, Eureka Scientific
- Paul Edlefsen, Harvard University
- Vinay Kashyap, Harvard-Smithsonian Center for Astrophysics
- Jason Kramer, University of California, Irvine
- Hyunsook Lee, Harvard-Smithsonian Center for Astrophysics
- Thomas Lee, Chinese University of Hong Kong
- Alan Lenarcic, Harvard University
- Xiao Li Meng, Harvard University
- Taeyoung Park, Pittsburgh University
- Aneta Siemiginowska, Harvard-Smithsonian Center for Astrophysics
- Nathan Stein, Harvard University
- David van Dyk, University of California, Irvine
- Alex Young, NASA Goddard Space Flight Center
- Yaming Yu, University of California, Irvine
- Andreas Zezas, Harvard-Smithsonian Center for Astrophysics

MARKOV CHAIN MONTE CARLO:

A Workhorse for Modern Scientific Computation

Markov chain Monte Carlo (MCMC) methods, originated in computational physics more than half a century ago, have seen an enormous range of applications in quantitative scientific investigations. This is mainly due to their ability to simulate from very complex distributions needed by all kinds of statistical models, from bioinformatics to financial engineering to astronomy. This talk provides an introductory tutorial of the two most frequently used MCMC algorithms: the Gibbs sampler and the Metropolis-Hastings algorithm. Using simple yet non-trivial examples, we demonstrate, via live performance, the good, bad, and ugly implementations. Along the way, we reveal the statistical thinking underlying their designs, including the secret behind the greatest statistical magic…

Monte Carlo Applications(应用)

物理 Physics
化学 Chemistry
天文 Astronomy
生物 Biology
环境 Environment
工程 Engineering
交通 Traffic

社会 Sociology
教育 Education
心理 Psychology
人文 Arts
语言 Linguistics
历史 History
医学 Medical Science

经济 Economics
金融 Finance
管理 Management
政策 Policy
军事 Military
政府 Government
商务 Business

…

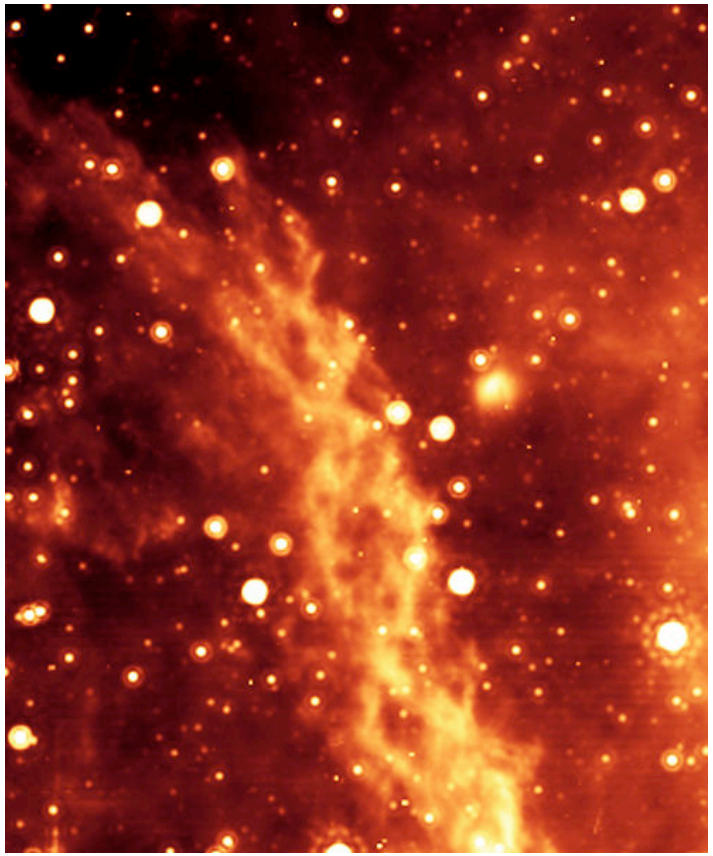
A Recent Thesis …

Markov Chain Monte Carlo Applications in Bioinformatics and Astrophysics

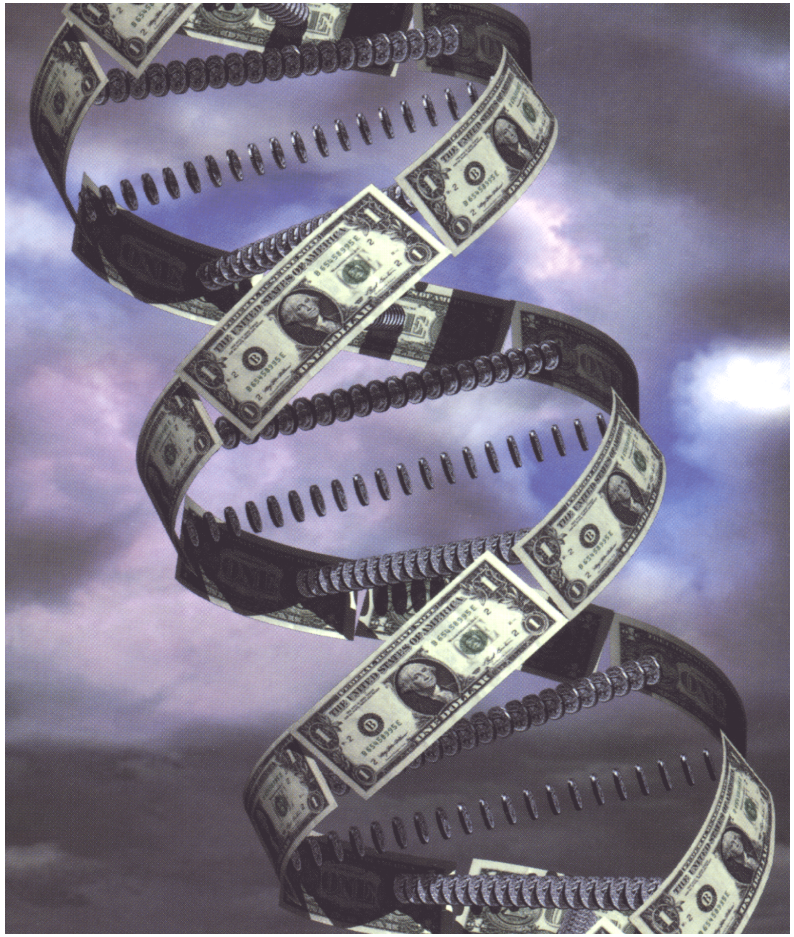
Hosung Kang

May, 2005

So what's the "average scale" between these two extremes?



Apparently it is about 2.5” by 6” …



Monte Carlo Integration(积分)

- Suppose we want to compute

$$I = \int g(x)f(x)dx,$$

where f(x) is a probability density(分布密度). If we have samples(样本) $x_1,\ldots,x_n \sim f(x)$, we can estimate(估计) I by

$$I_n = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$$

Monte Carlo Optimization(优化)

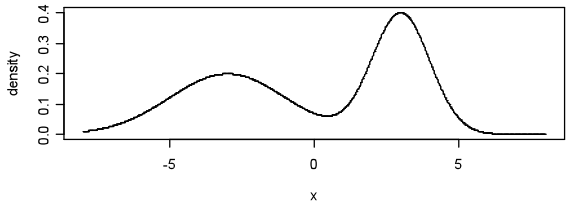
- We want to maximize p(x)
- Simulate from

  $f(x) / p^{\lambda}(x)$.

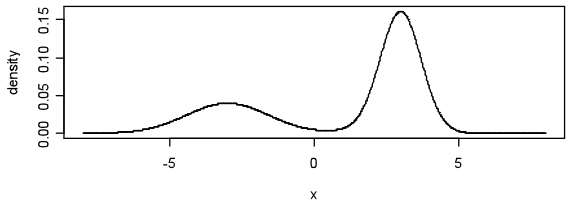
  As $\lambda \to \infty$, the simulated draws will be more and more concentrated around the maximizer of p(x)

  “两极分化”

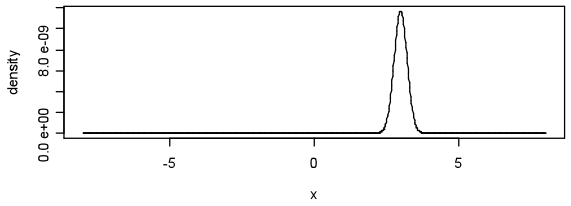
$\lambda = 1$



$\lambda = 2$



$\lambda = 20$



Simulating from a Distribution(抽样)

- What does it mean?

  Suppose a random variable (随机变量) X *can only take two values:*

$$P(X = 0) = \frac{1}{4} \qquad P(X = 1) = \frac{3}{4}$$

  Simulating from the distribution of X *means that we want a collection of 0's and 1's:*

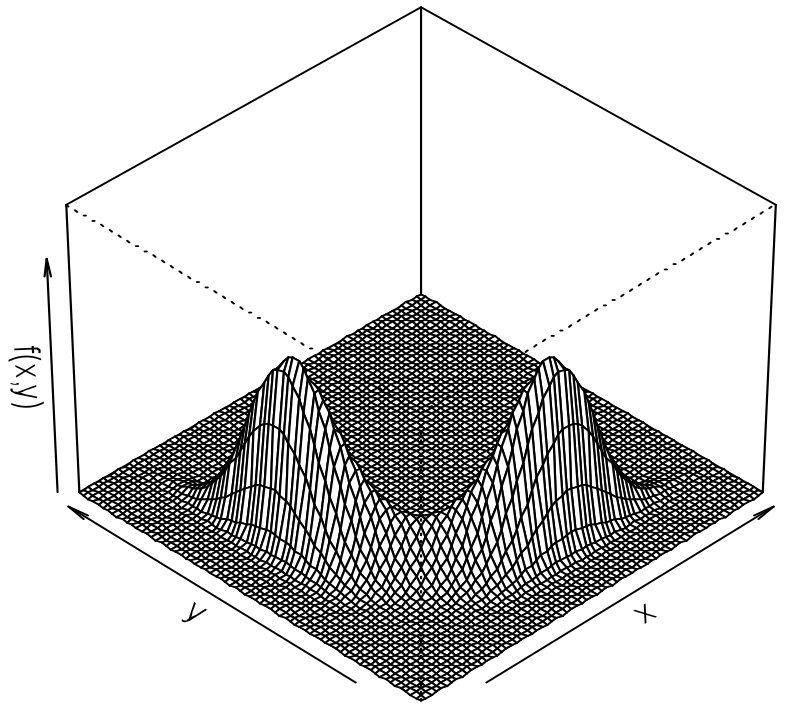
$$x_1, x_2, \ldots, x_n$$

  such that about 25% of them are 0's and about 75%of them are 1's, when n, the simulation size is large.

- The $\{x_i, i = 1,\ldots,n\}$ don't have to be independent(独立)

Simulating from a Complex Distribution

- Continuous variable X, described by a density function f(x)
- Complex:
  - *the form of* f(x)
  - *the dimension of* x



$$f(x,y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y))$$

Markov Chain Monte Carlo

$$x^{(t)} = \varphi(x^{(t-1)}, U^{(t)}),$$

where $\{U^{(t)}, t=1,2,\ldots\}$ are identically and independently distributed (独立同分布).

- Under regularity conditions (正则条件),

$$f(x^{(t)}) \overset{t\to\infty}{\to} f(x)$$

So We can treat $\{x^{(t)}, t= N_0, \ldots, N\}$ as an approximate sample from f(x), the stationary/limiting distribution.

Gibbs Sampler (Gibbs 抽样法)

- Target density: $f(x, y)$
- We know how to simulate form the conditional distributions

$$f(x|y) \quad \text{and} \quad f(y|x)$$

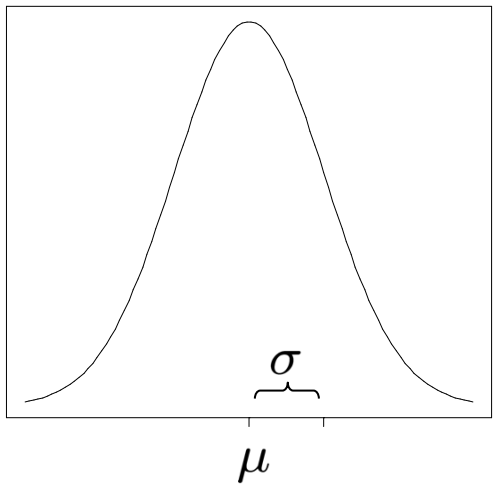
- For the previous example,

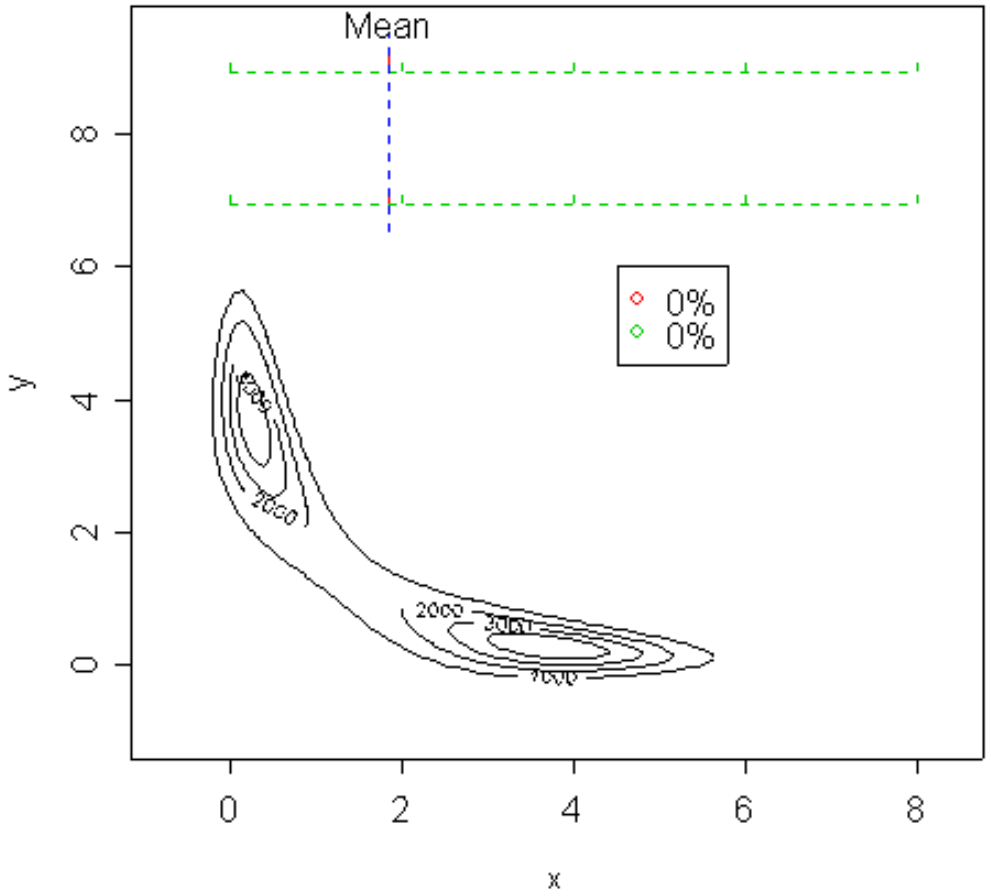
$$f(x, y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y))$$

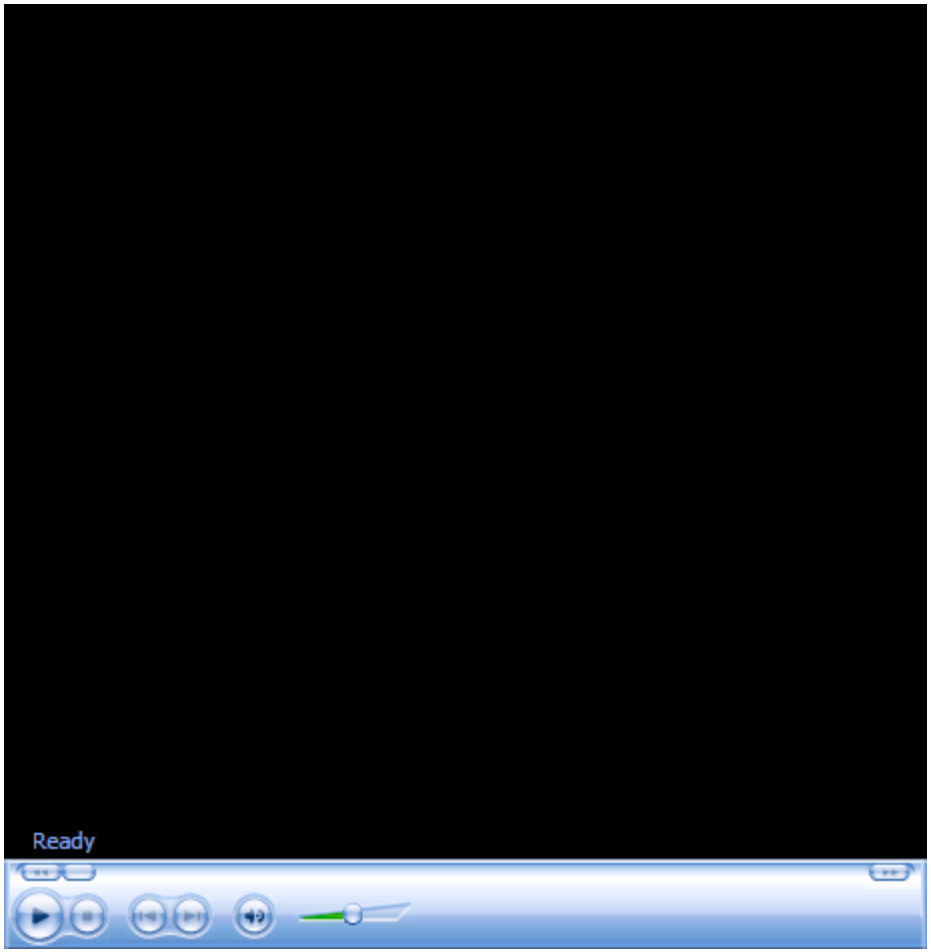
$$f(x|y) = N(\frac{4}{1+y^2}, \frac{1}{1+y^2})$$

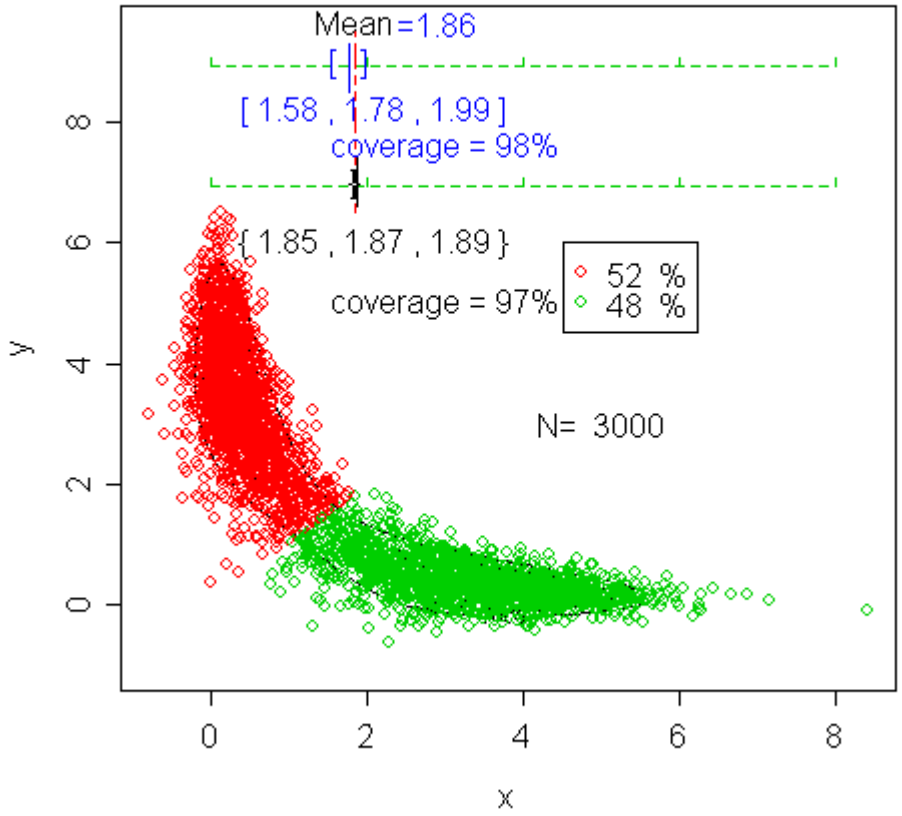
$$f(y|x) = N(\frac{4}{1+x^2}, \frac{1}{1+x^2})$$

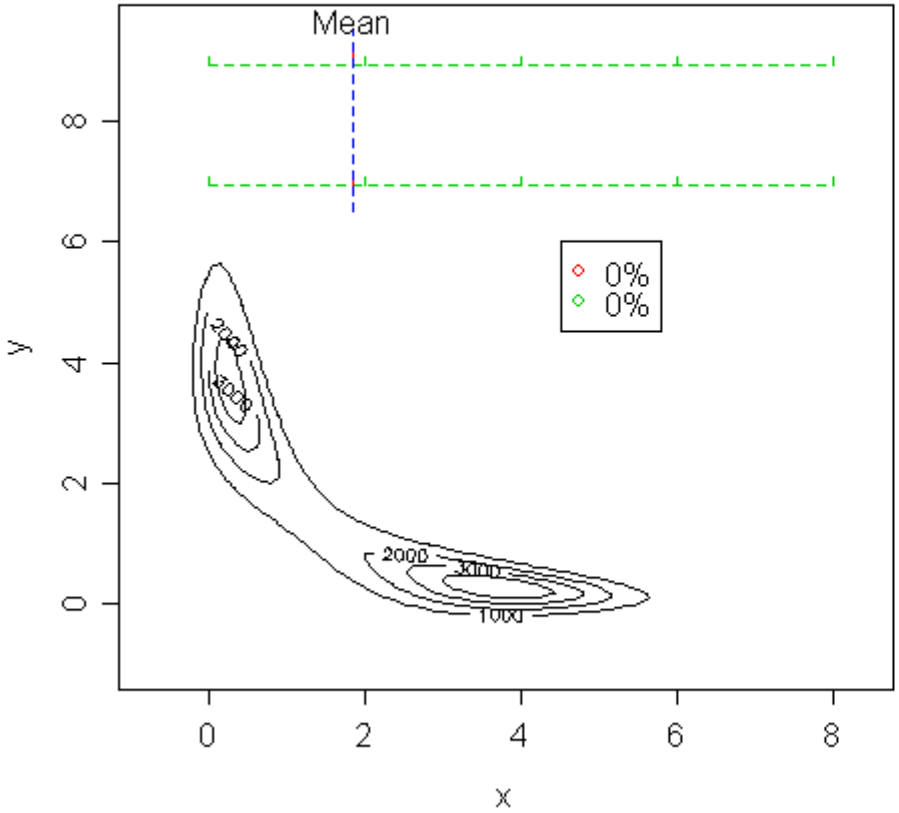
N(μ,σ²)
Normal Distribution
(正态分布)
"Bell Curve"

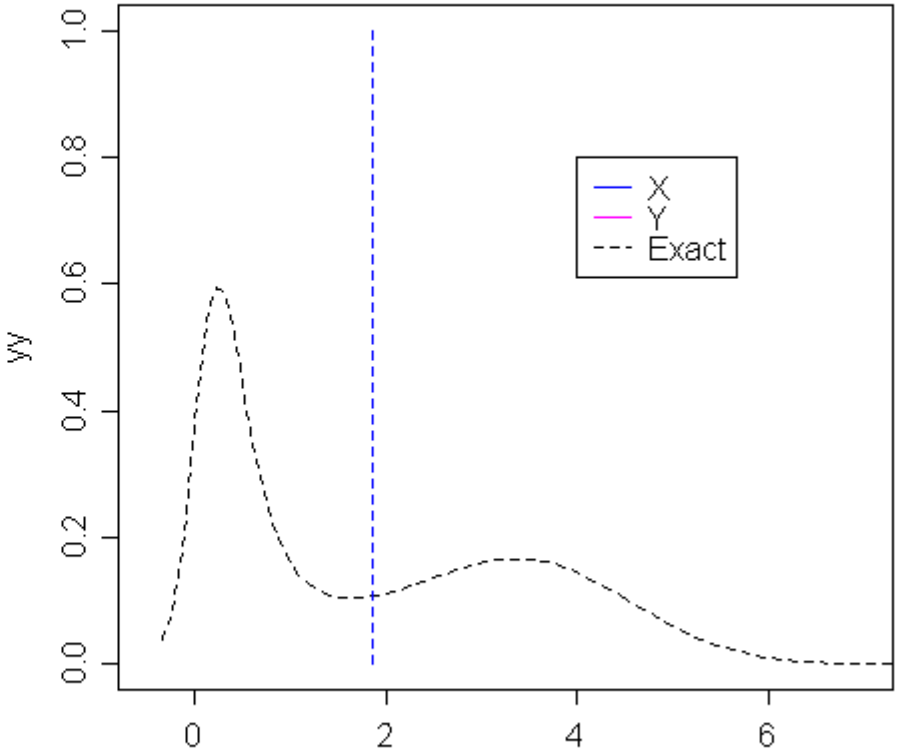


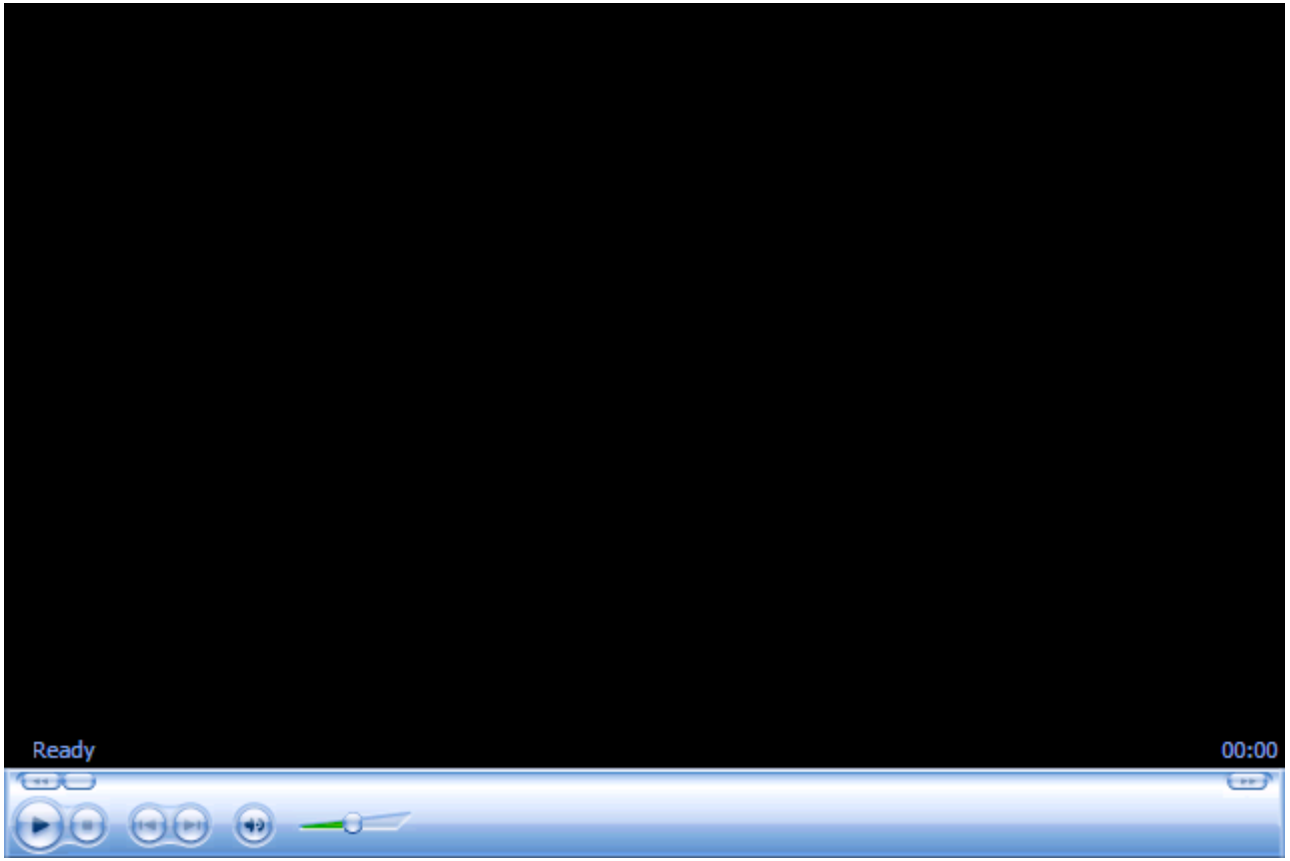

Mean
0%
0%
y
x
0
2
4
6
8
2000
3000
7000

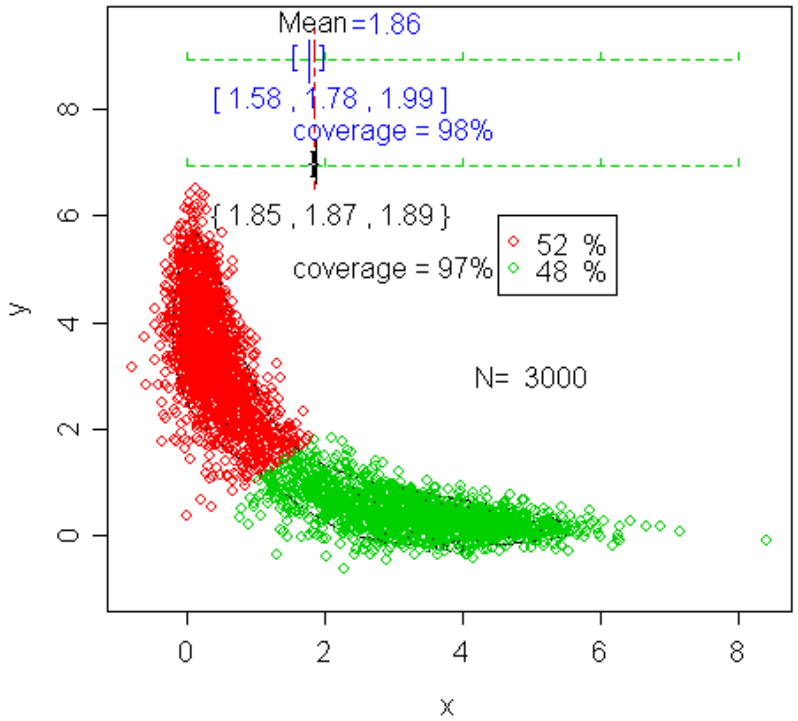

Ready

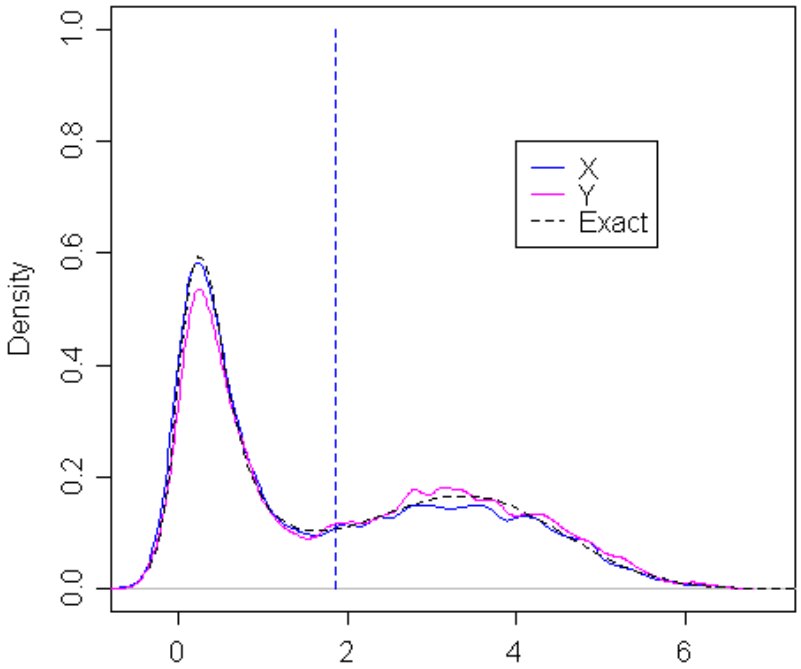

Mean=1.86
[1.58 , 1.78 , 1.99]
coverage = 98%
{ 1.85 , 1.87 , 1.89 }
coverage = 97%
52 %
48 %
N= 3000
y
x
0
2
4
6
8

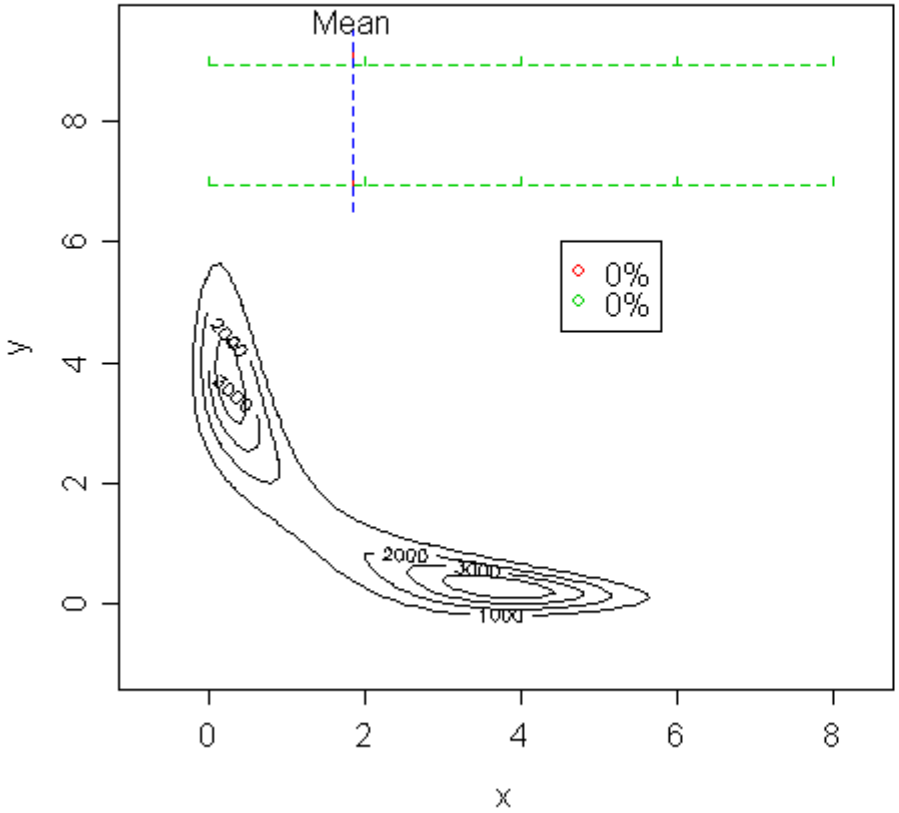

Mean
0%
0%
3000
4000
2000
3000
1000
y
x
0
2
4
6
8

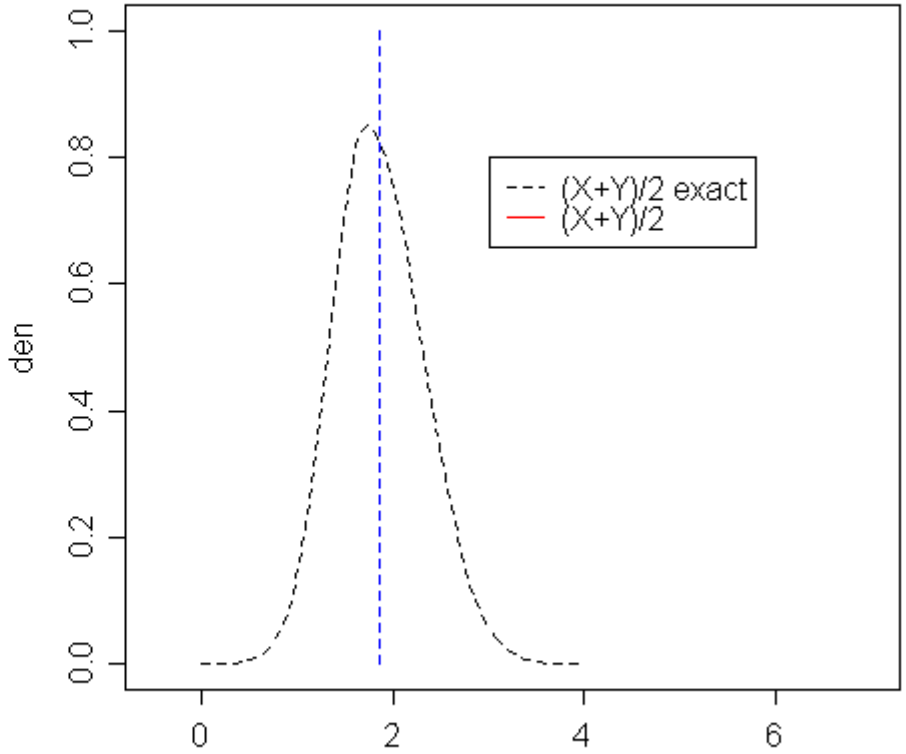

X
Y
Exact
yy
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

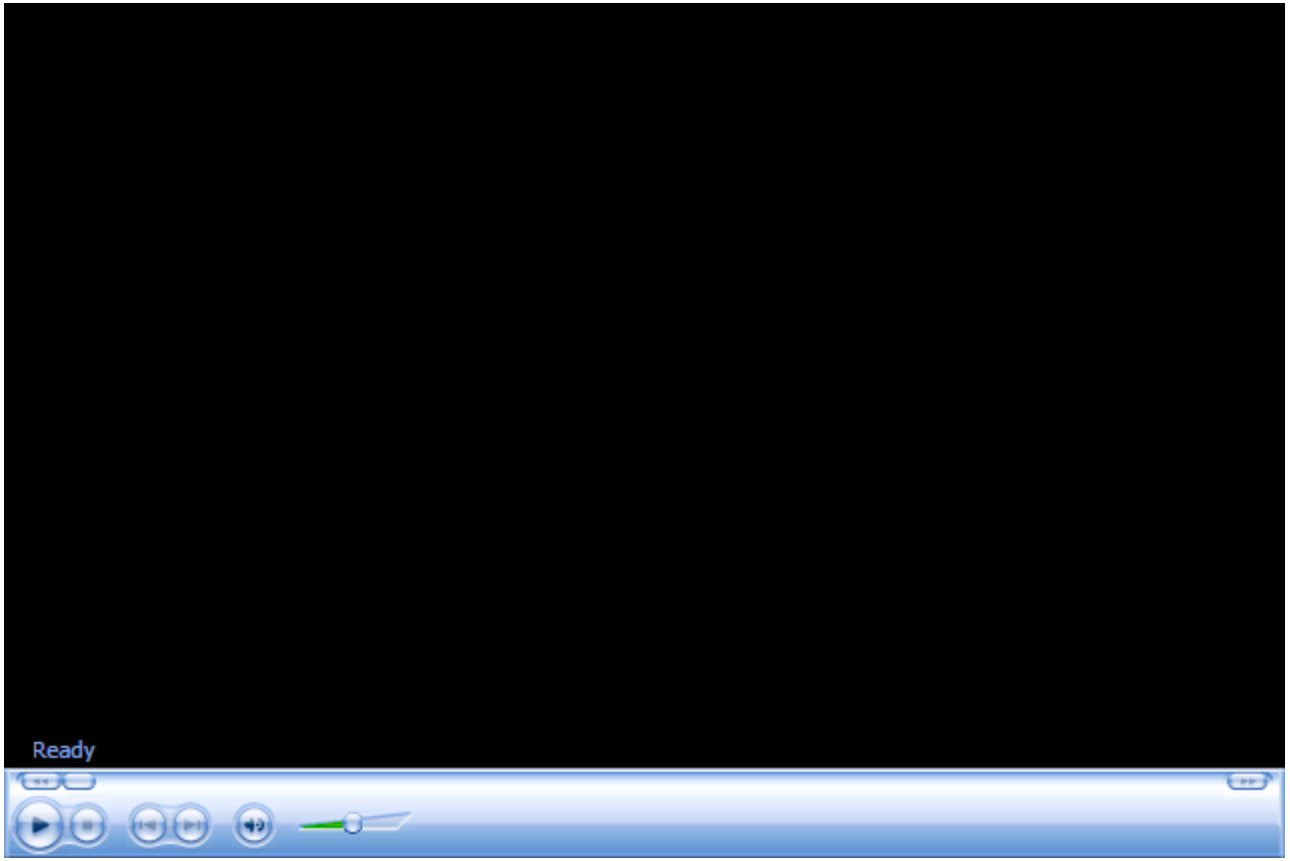

Ready
00:00

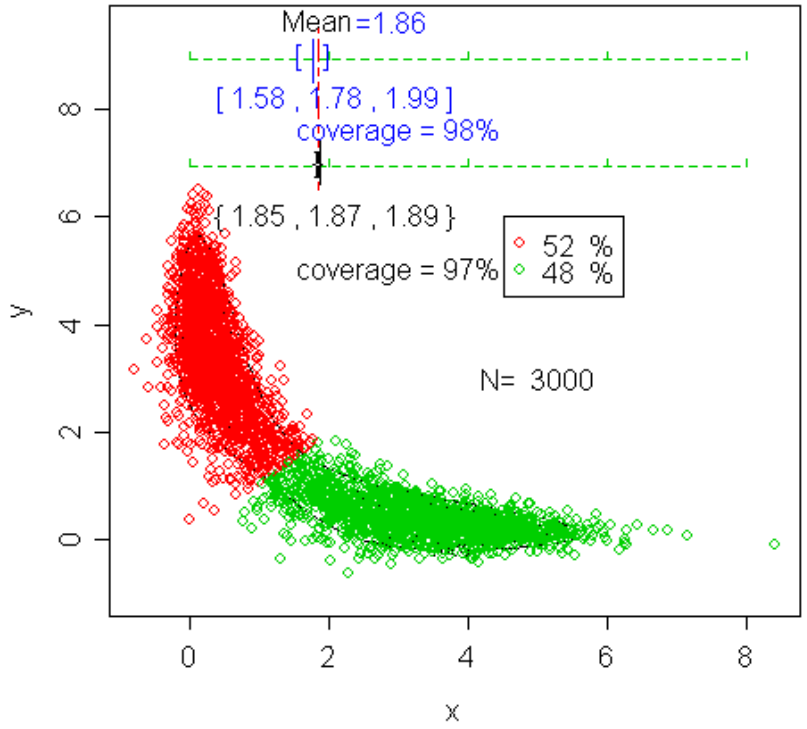

Mean=1.86
[1.58 , 1.78 , 1.99]
coverage = 98%
{ 1.85 , 1.87 , 1.89 }
coverage = 97%
52 %
48 %
N= 3000
y
x
0
2
4
6
8

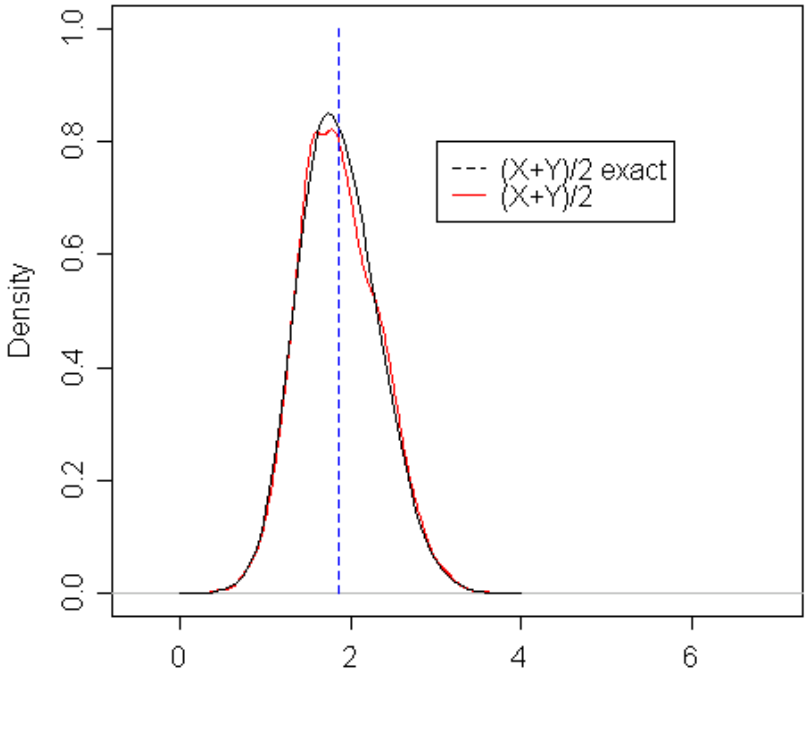

X
Y
Exact
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

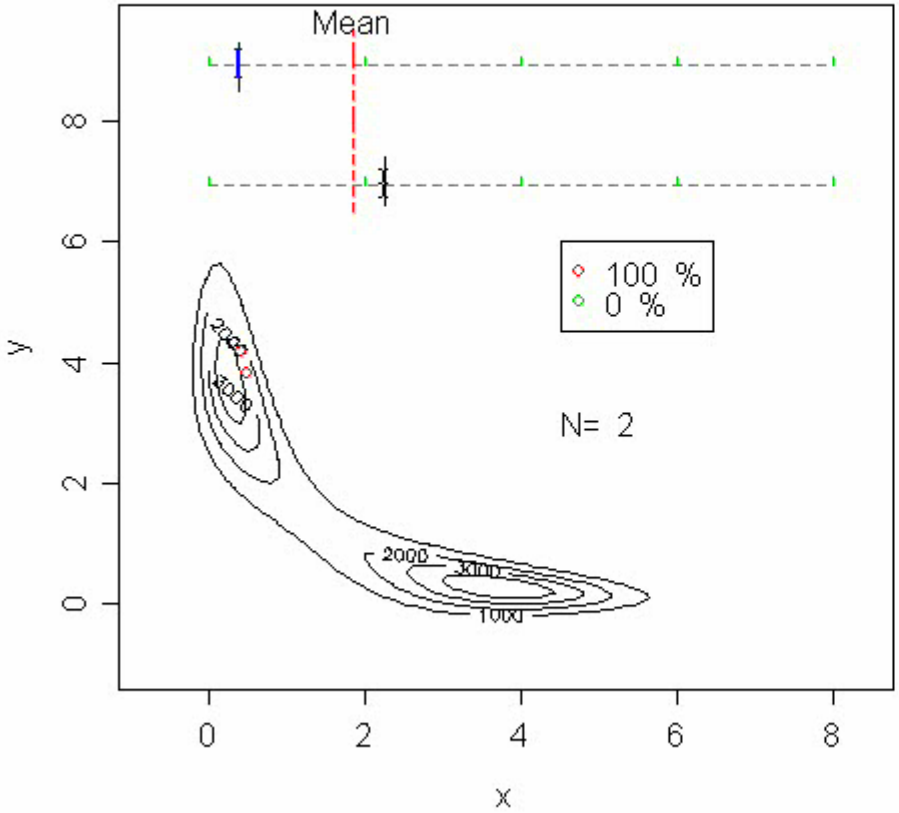

Mean
0%
0%
y
x
0
2
4
6
8
2000
3000
1000

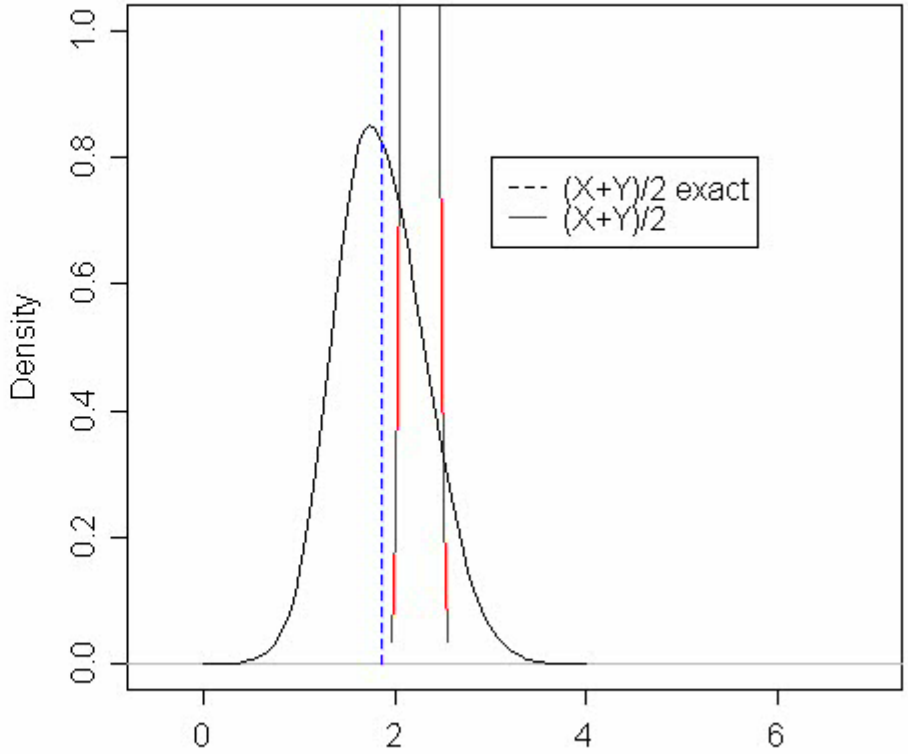

(X+Y)/2 exact
(X+Y)/2
den
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

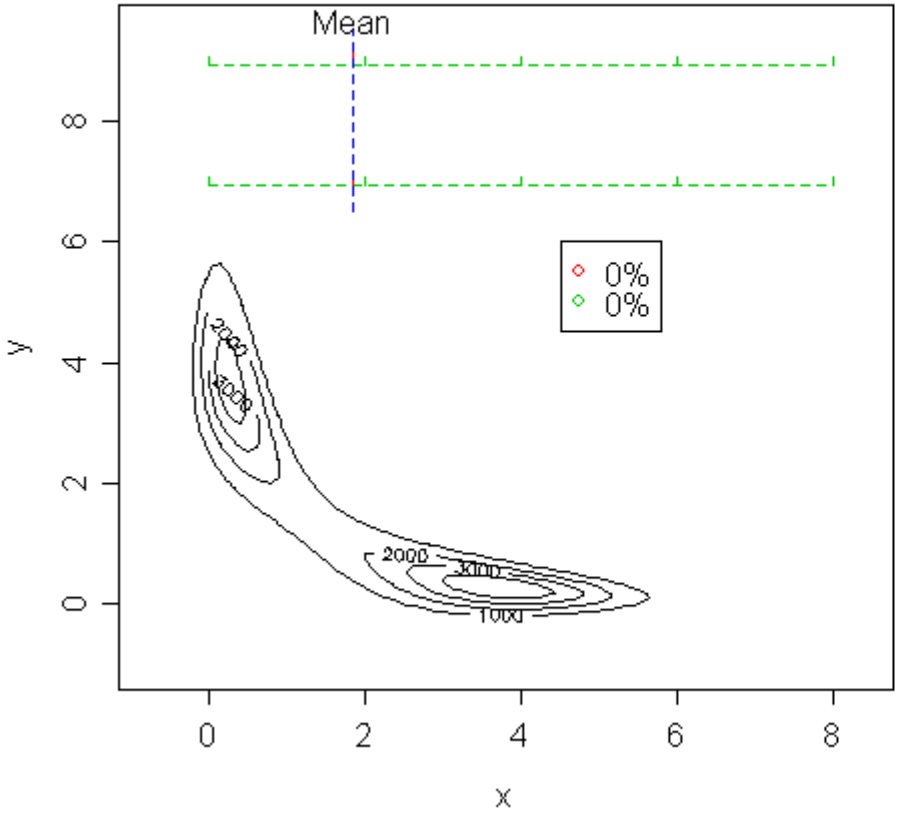

Ready

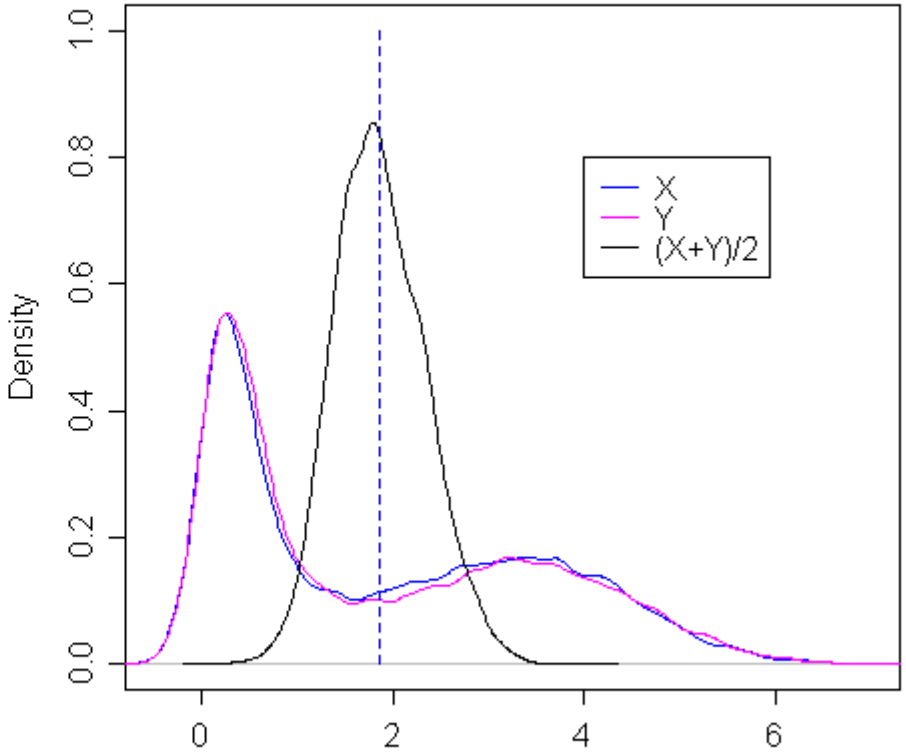

Mean=1.86
[1.58 , 1.78 , 1.99]
coverage = 98%
{ 1.85 , 1.87 , 1.89 }
coverage = 97%
52 %
48 %
N= 3000
y
x
0
2
4
6
8

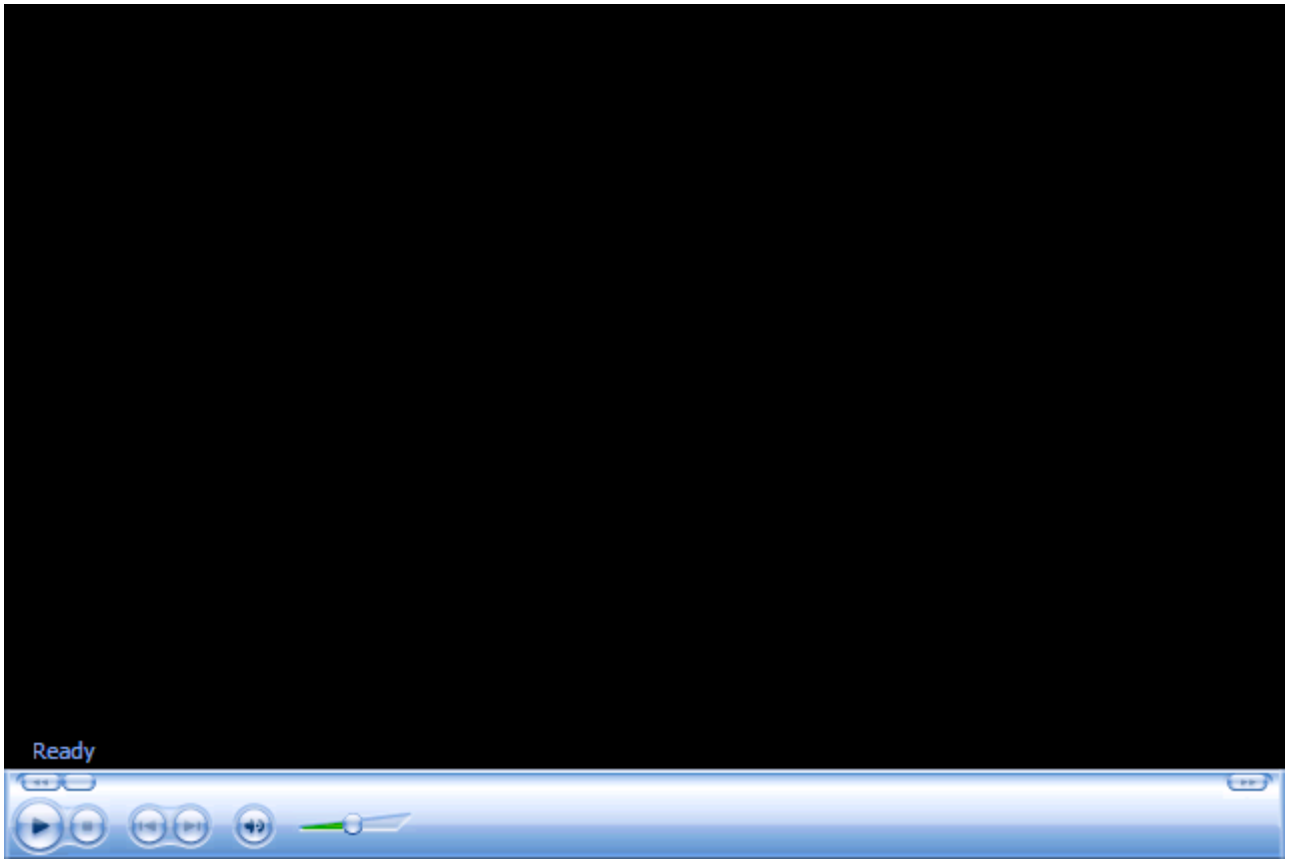

(X+Y)/2 exact
(X+Y)/2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

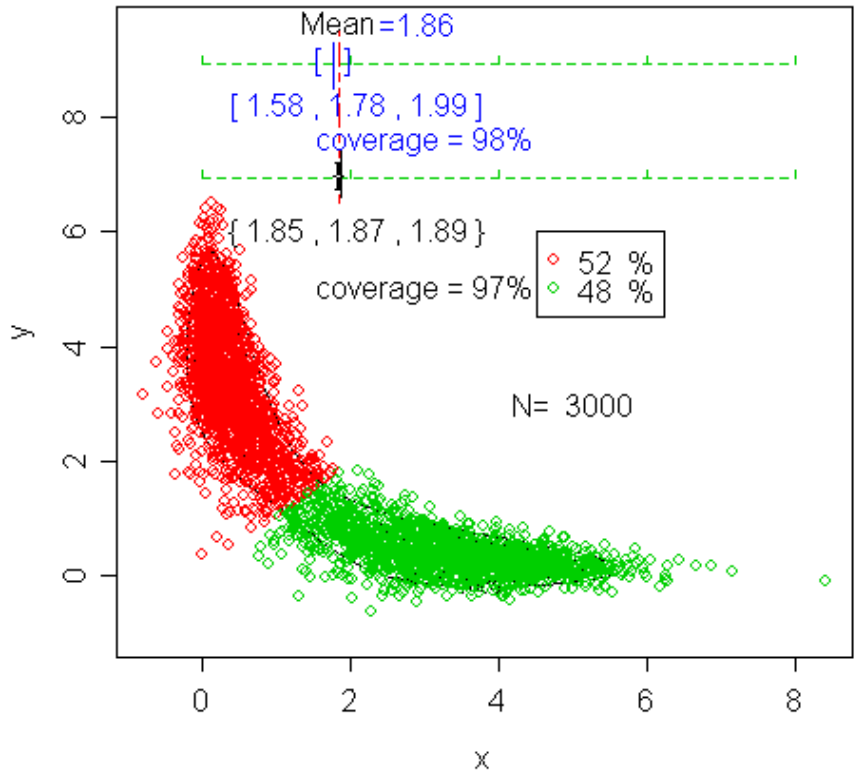

Mean
100 %
0 %
N= 2
2000
3000
1000
y
x
0
2
4
6
8

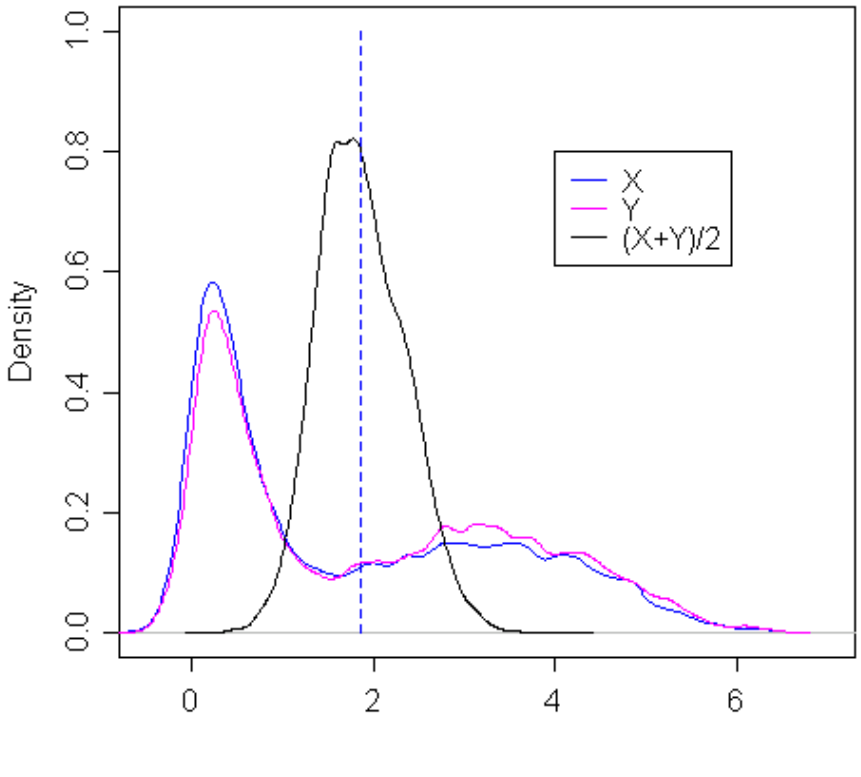

(X+Y)/2 exact
(X+Y)/2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

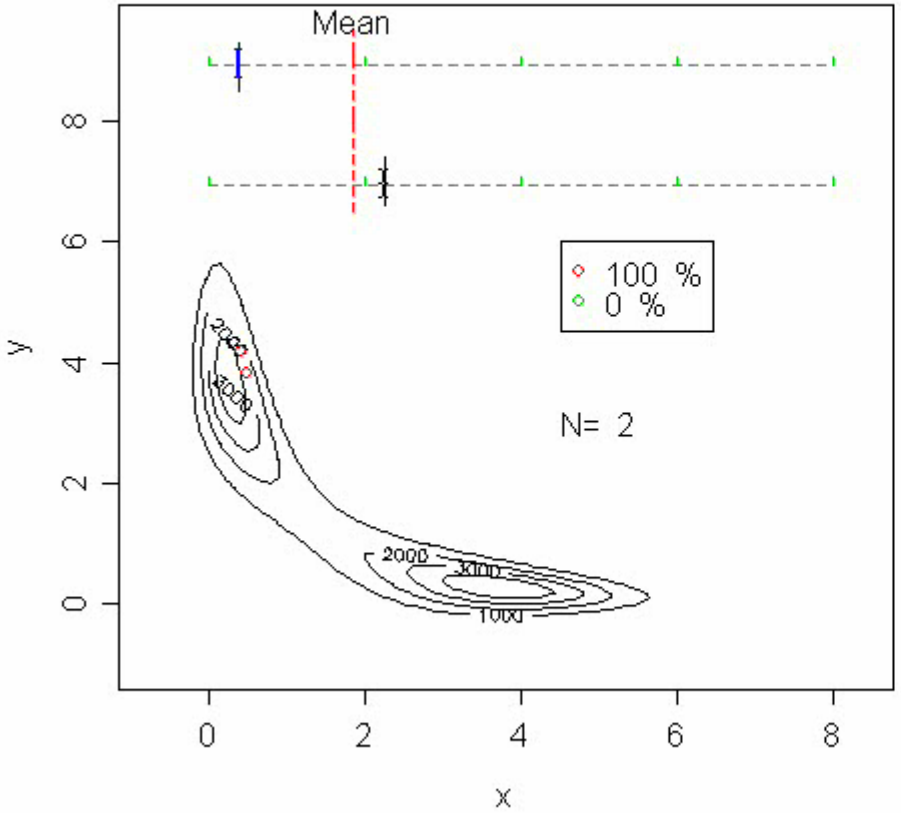

Mean
0%
0%
3000
4000
2000
3000
1000
y
x
0
2
4
6
8

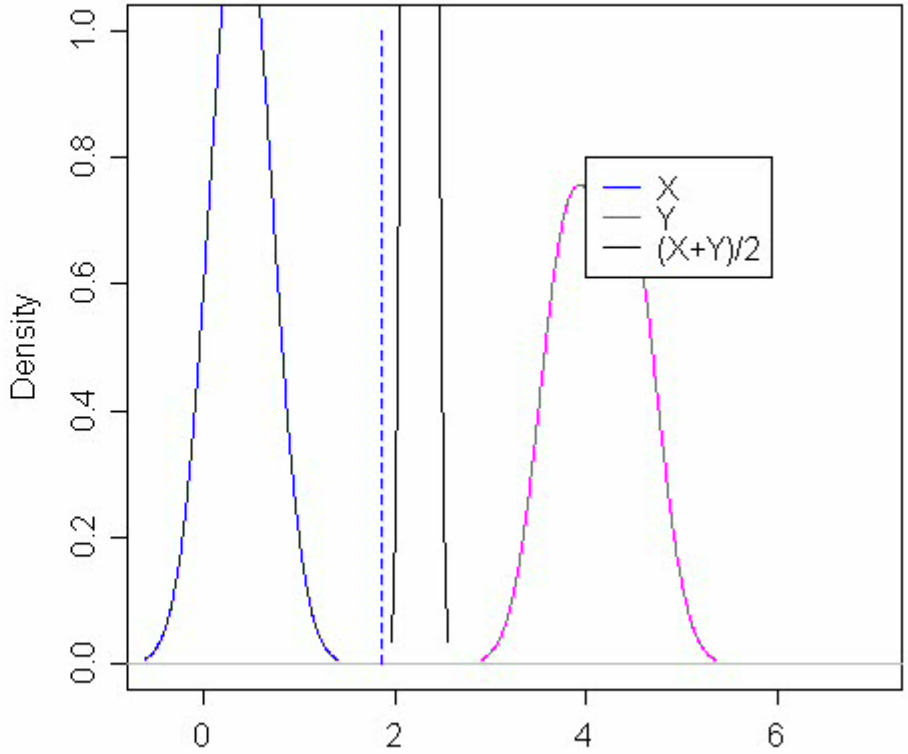

X
Y
(X+Y)/2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6


Ready


Mean=1.86
[1.58 , 1.78 , 1.99]
coverage = 98%
{ 1.85 , 1.87 , 1.89 }
coverage = 97%
52 %
48 %
N= 3000
y
x
0
2
4
6
8

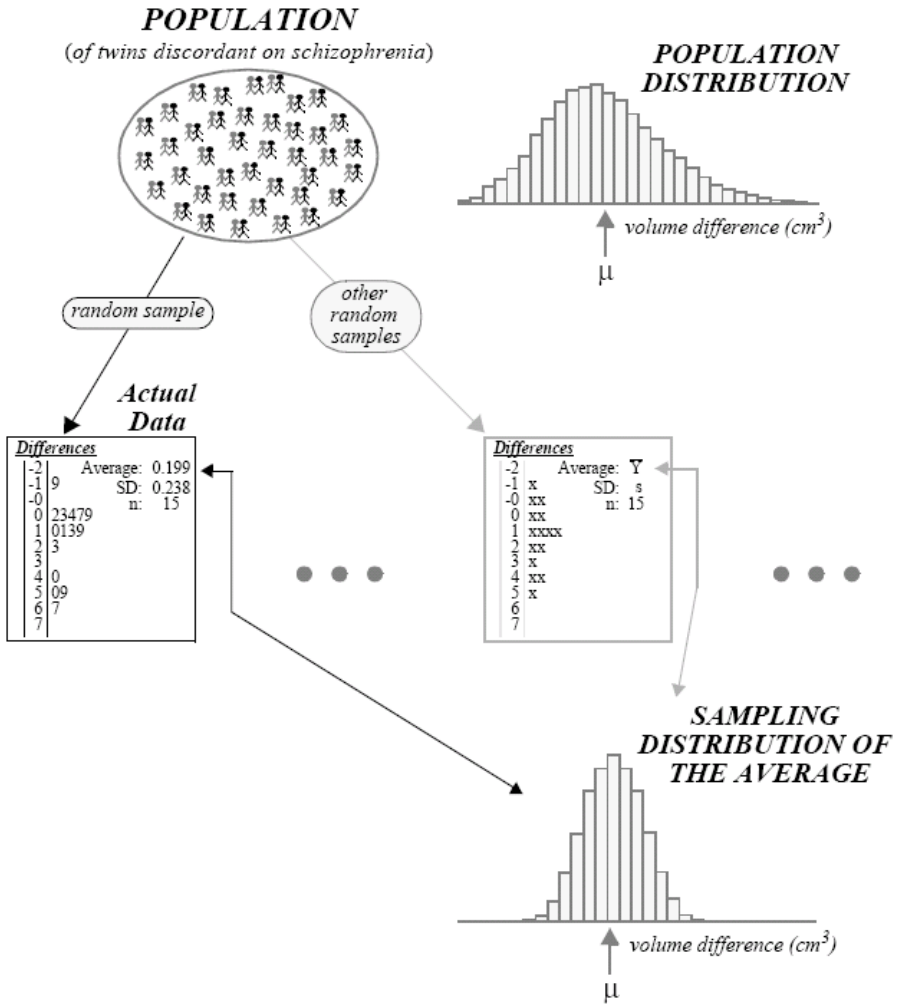

Density
X
Y
(X+Y)/2
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6


Mean
100 %
0 %
N= 2
2000
3000
1000
y
x
0
2
4
6
8


X
Y
(X+Y)/2
Density
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6

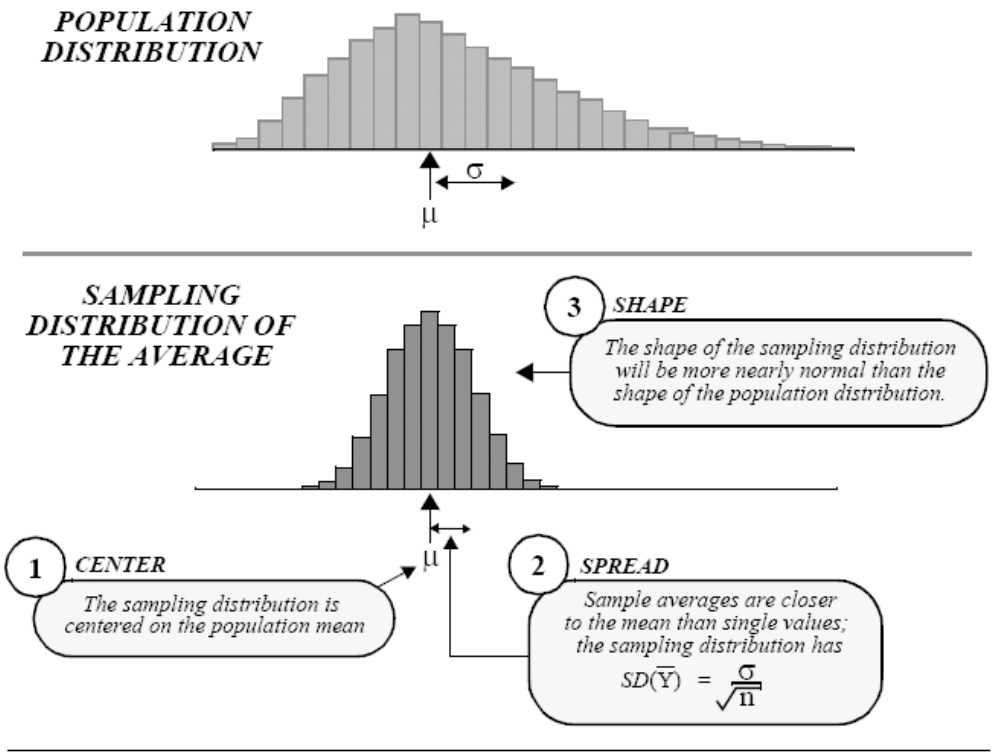
The Greatest Statistical Magic

Estimate the errors in our estimate without knowing the truth



Graph is taken from *Statistical Sleuth*

The relationship between the population distribution and the sampling distribution of the average in random sampling



Graph is taken from *Statistical Sleuth*

Statistical Inference

- Point Estimator: $\bar{g}_n = \frac{1}{n}\sum_{t=1}^{n} g(x^{(t)})$
- Variance Estimator: $V(\bar{g}_n) \approx \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$,

$\sigma^2 = Var(g(x))$ estimated by $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{t=1}^{n}(g(x^{(t)}) - \bar{g}_n)^2$,

$\rho = corr(g(x^{(t)}), g(x^{(t-1)}))$ estimated by

$$\hat{\rho} = \frac{1}{n-1}\frac{\sum_{t=2}^{n}(g(x^{(t)}) - \bar{g}_n)(g(x^{(t-1)}) - \bar{g}_n)}{\sqrt{\sum_{t=1}^{n-1}(g(x^{(t)}) - \bar{g}_n)^2 \sum_{t=2}^{n}(g(x^{(t)}) - \bar{g}_n)^2}}.$$

- Interval Estimator:

$$(\bar{g}_n - t_d\sqrt{\widehat{V}(\bar{g}_n)}, \quad \bar{g}_n + t_d\sqrt{\widehat{V}(\bar{g}_n)}),$$

where $d = n\frac{1-\rho}{1+\rho} - 1$, and $t_d \to 1.96$ as $n \to \infty$.

Data Augmentation(数据扩大法)

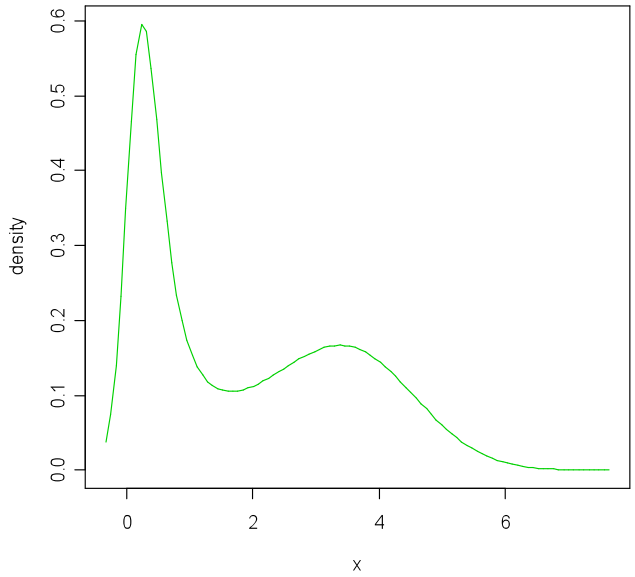
- We want to simulate from

$$f(x) \propto \frac{1}{\sqrt{1+x^2}} \exp\{-\frac{1}{2}(x^2 - 8x - \frac{16}{1+x^2})\}.$$

But this is just the marginal distribution (边缘分布) of

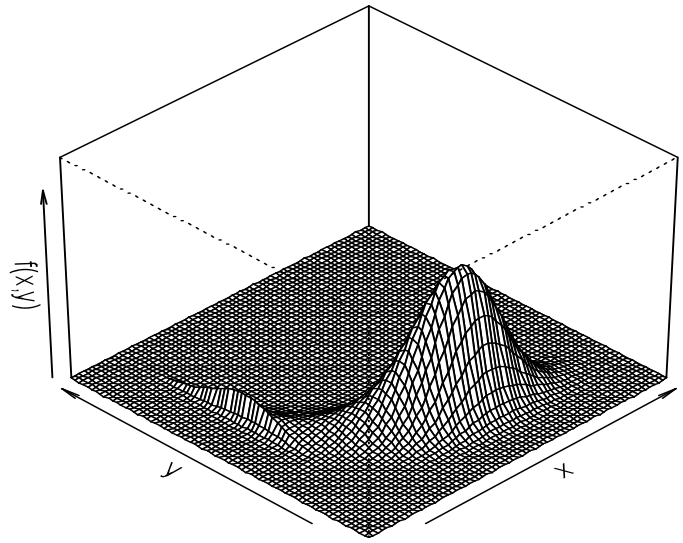
$$f(x,y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y)).$$

So once we have simulations:
{($x^{(t)}$, $y^{(t)}$: t= 1,2,…,N)},
we also obtain draws:
{$x^{(t)}$: t= 1,2,…,N)}



A More Complicated Example

$$f(x,y) \propto \exp(-\frac{1}{2}(|x|y^2 + x^2 + y^2 - 8x - 8y))$$



$$f(x,y) = \exp\{-\frac{1}{2}(x-4)^2\}\exp\{-\frac{1}{2}(y-4)^2\}\exp\{-\frac{1}{2}|x|y^2\}$$

Metropolis-Hastings algorithm

- Simulate from an approximate (近似) distribution $q(z_1|z_2)$, then
  - Step 0: Select $z^{(0)}$;

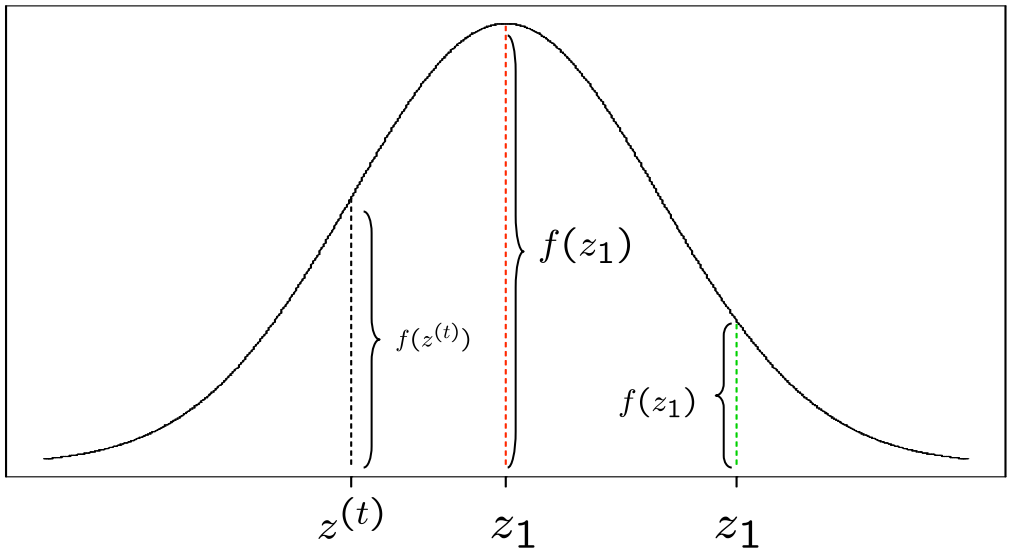
    Now for t = 1,2,…,N, repeat
  - Step 1: draw z_1 from $q(z_1|z_2=z^{(t)})$
  - Step 2: Calculate

$$\alpha(z_1, z^{(t)}) = \frac{f(z_1)q(z^{(t)}|z_1)}{f(z^{(t)})q(z_1|z^{(t)})}$$

  - Step 3: set $z^{(t+1)} = \begin{cases} z_1, & with\ p = \min\{\alpha, 1\} \\ z^{(t)}, & with\ 1-p \end{cases}$ (接受) (拒绝)
- Discard the first N_0 draws

M-H Algorithm: An Intuitive Explanation

Assume $q(z_1|z_2) = q(z_2|z_1)$, then $\alpha(z_1, z^{(t)}) = \frac{f(z_1)}{f(z^{(t)})}$



M-H: An Ugly Implementation

$$f(x,y) = \Phi(x-4)\Phi(y-4)\exp\{-\tfrac{1}{2}|x|y^2\}$$

[$\Phi(x)$ is the density function of $N(0,1)$]

We choose $q(z|z_2)=q(z)=\Phi(x-4)\Phi(y-4)$(独立正态)

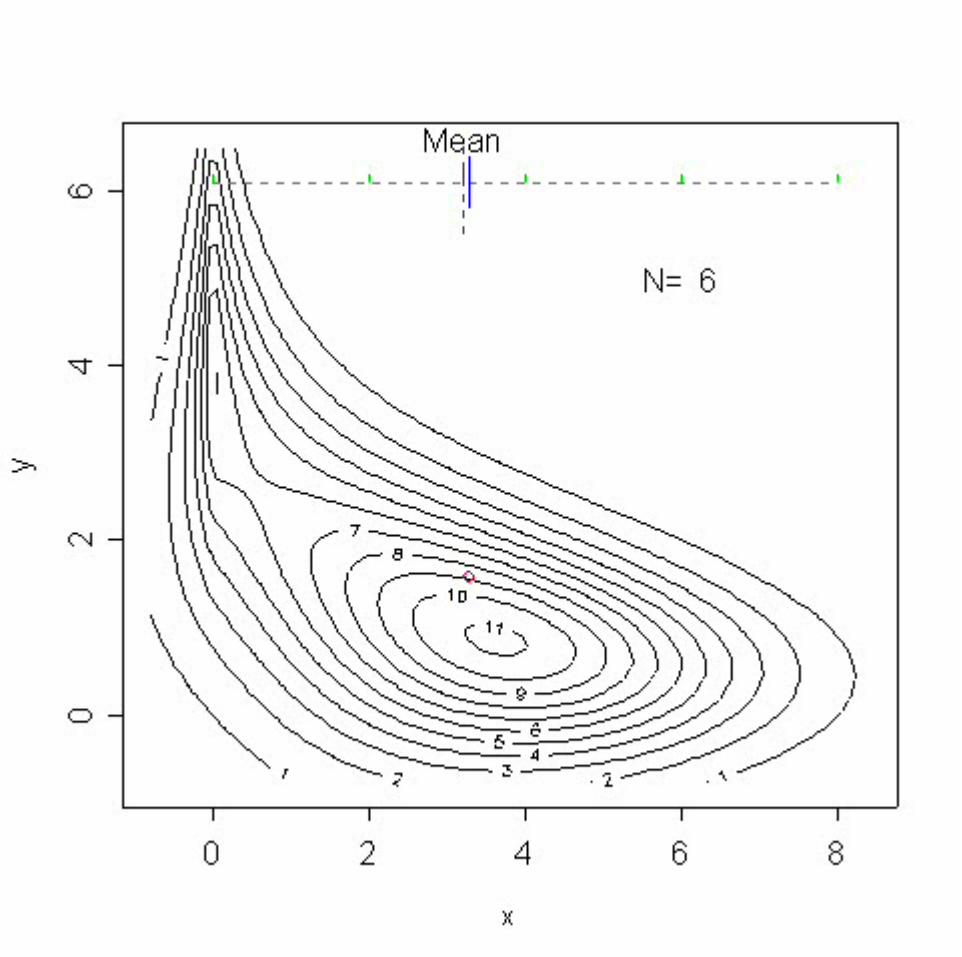
- Step 1: draw x ~ N(4,1), y ~ N(4,1);
  Dnote $z_1=(x,y)$
- Step 2: Calculate

$$\alpha(z_1, z^{(t)}) = \frac{\exp\{-\frac{1}{2}|x|y^2\}}{\exp\{-\frac{1}{2}|x^{(t)}|[y^{(t)}]^2\}}$$

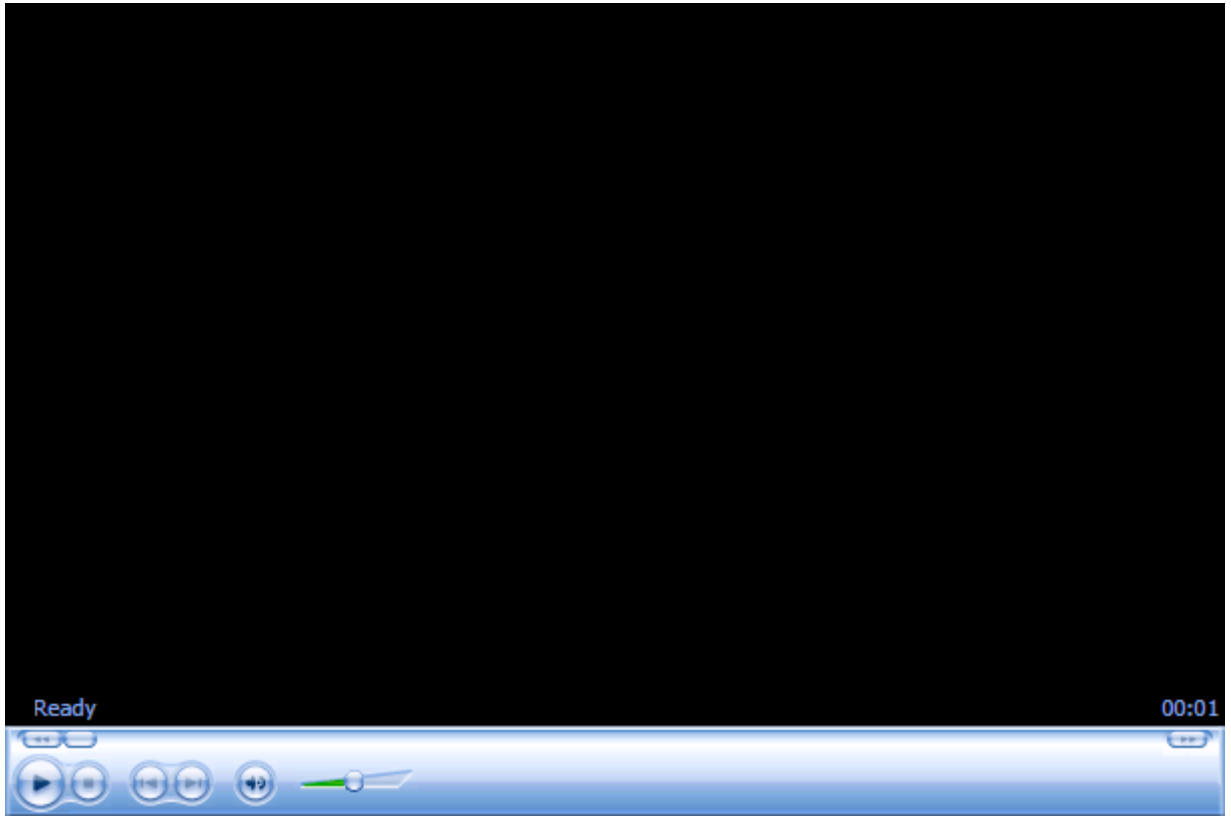
- Step 3: draw u ~U[0,1]
  Let

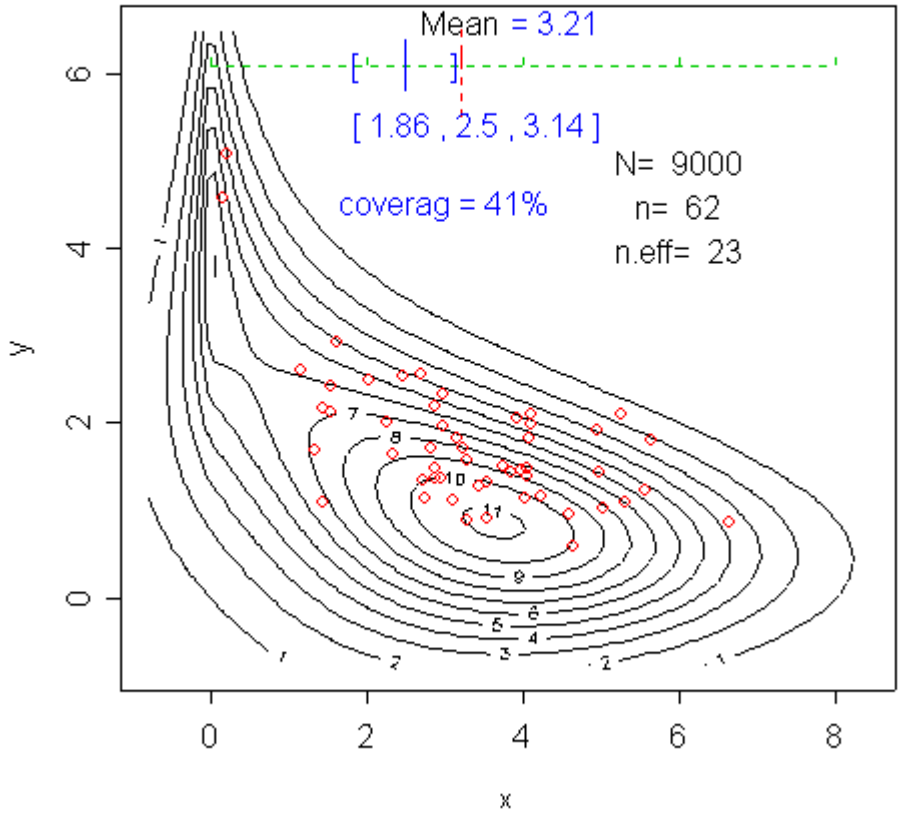
$$z^{(t+1)} = \begin{cases} z_1, & if\ u \le \min\{1, \alpha\} \\ z^{(t)}, & otherwise \end{cases}$$

Why is it so ugly?



Why is it so ugly?




Mean = 3.21
[1.86 , 2.5 , 3.14]
coverag = 41%
N= 9000
n= 62
n.eff= 23
y
x
0
2
4
6
8

M-H: A Bad Implementation

Starting from some arbitrary $(x^{(0)},y^{(0)})$

- Step 1: draw x ~ $N(x^{(t)},1)$, y ~ $N(y^{(t)},1)$
  "random walk" $x = x^{(t)} + U_x,\ y = y^{(t)} + U_y$

$$U_x, U_y \overset{iid}{\sim} N(0,1)$$

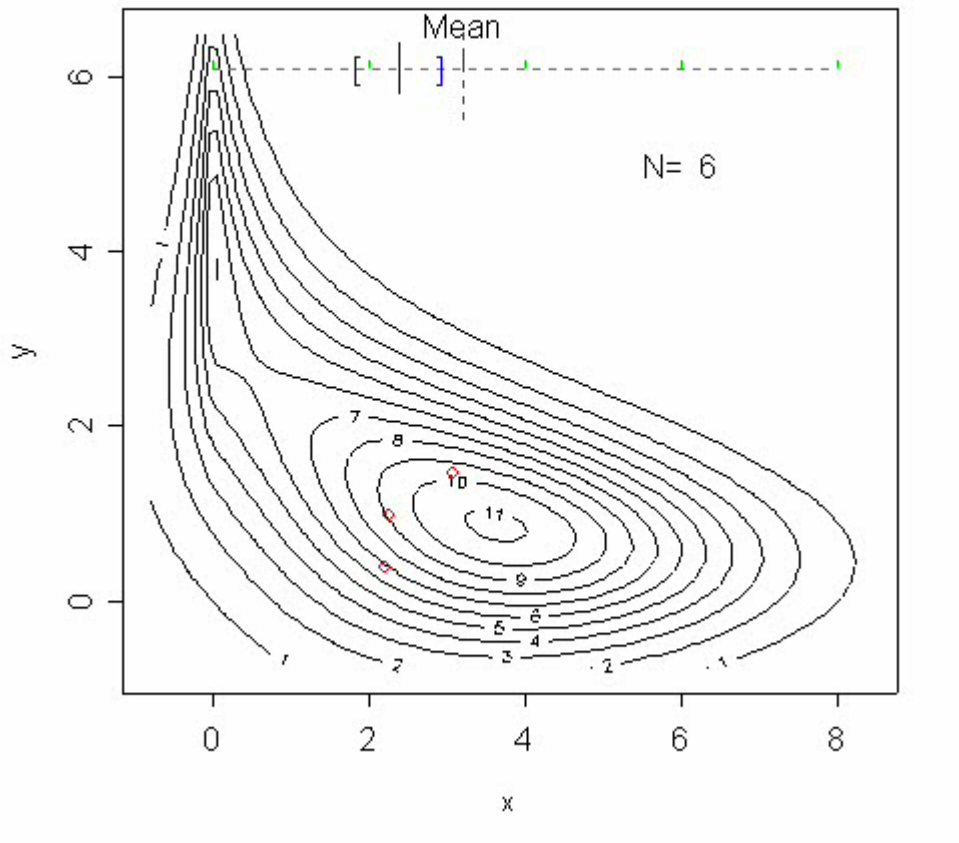
- Step 2: dnote z_1=(x,y), calculate

$$\alpha(z_1, z^{(n)}) = \frac{f(z_1)}{f(z^{(t)})}$$

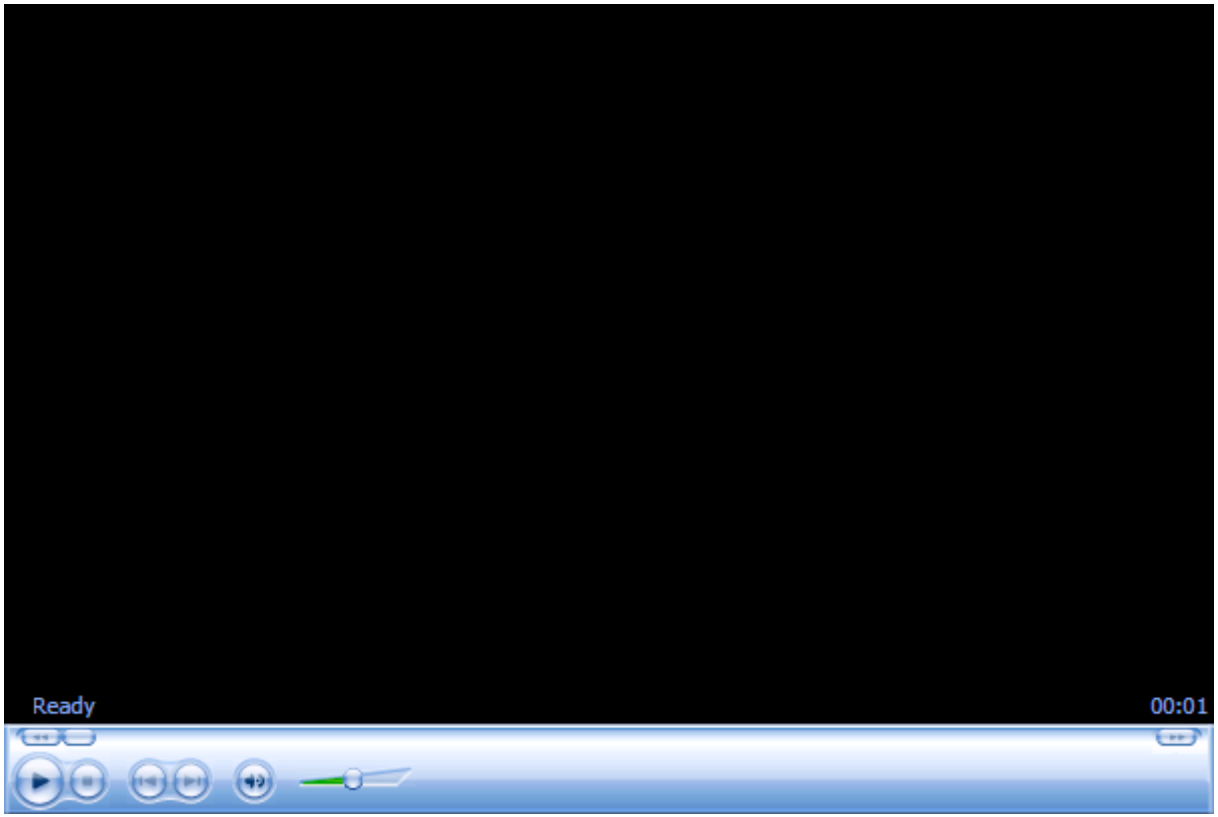
- Step 3: draw u ~ U[0,1]
  Let

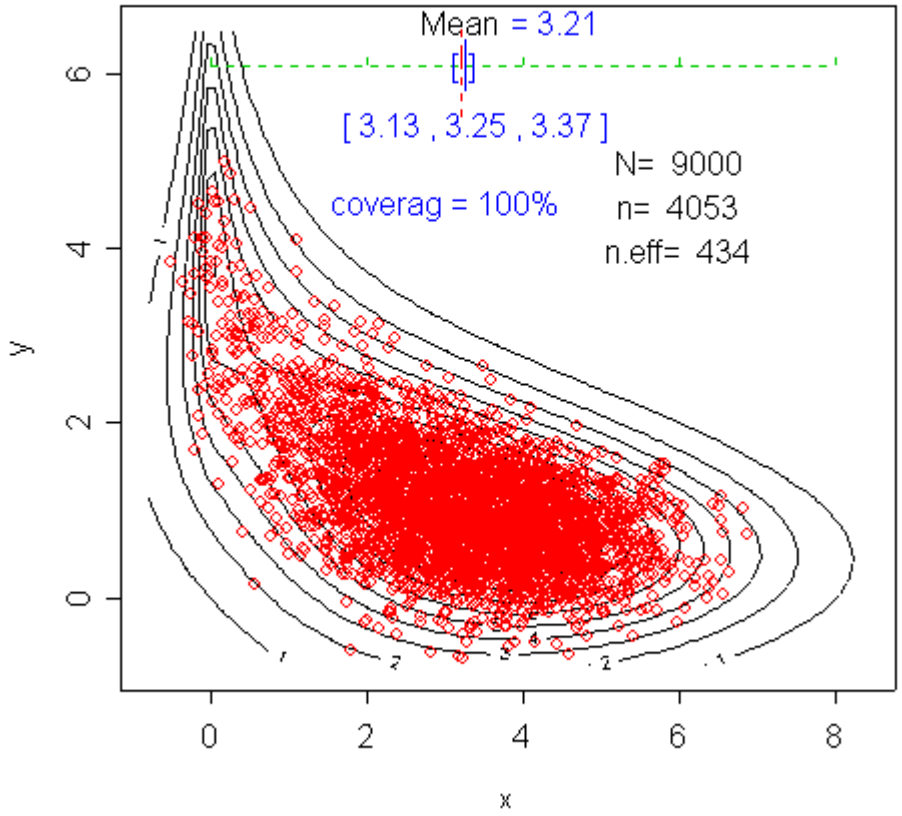
$$z^{(n+1)} = \begin{cases} z_1, & if\ u \le \min\{1, \alpha\} \\ z^{(n)}, & otherwise \end{cases}$$

Much Improved, but still bad!



Much Improved, but still bad!




Mean = 3.21
[3.13 , 3.25 , 3.37]
coverag = 100%
N= 9000
n= 4053
n.eff= 434
y
x
0
2
4
6
8

MH: good implementation

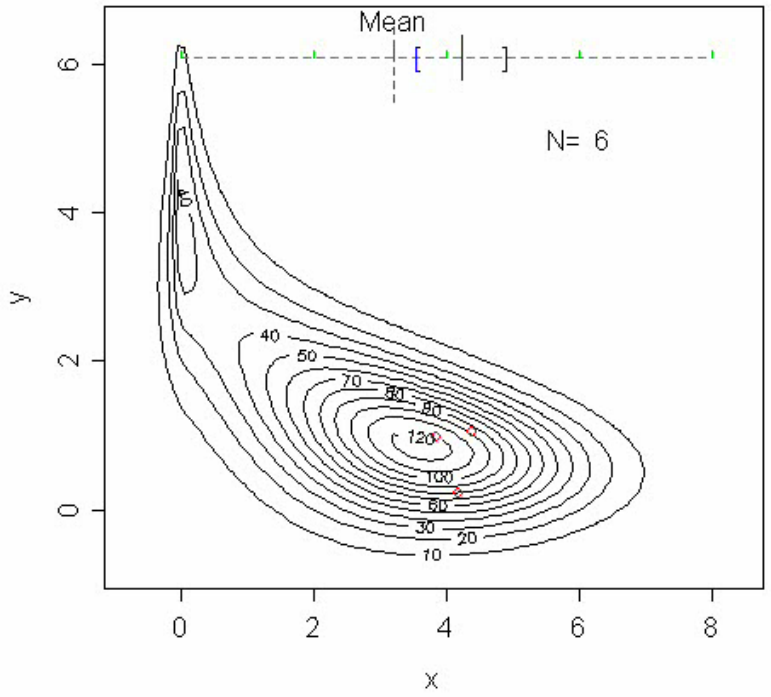
Change proposal to

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x^{(t)} \\ y^{(t)} \end{pmatrix} + \Sigma^{\frac{1}{2}} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

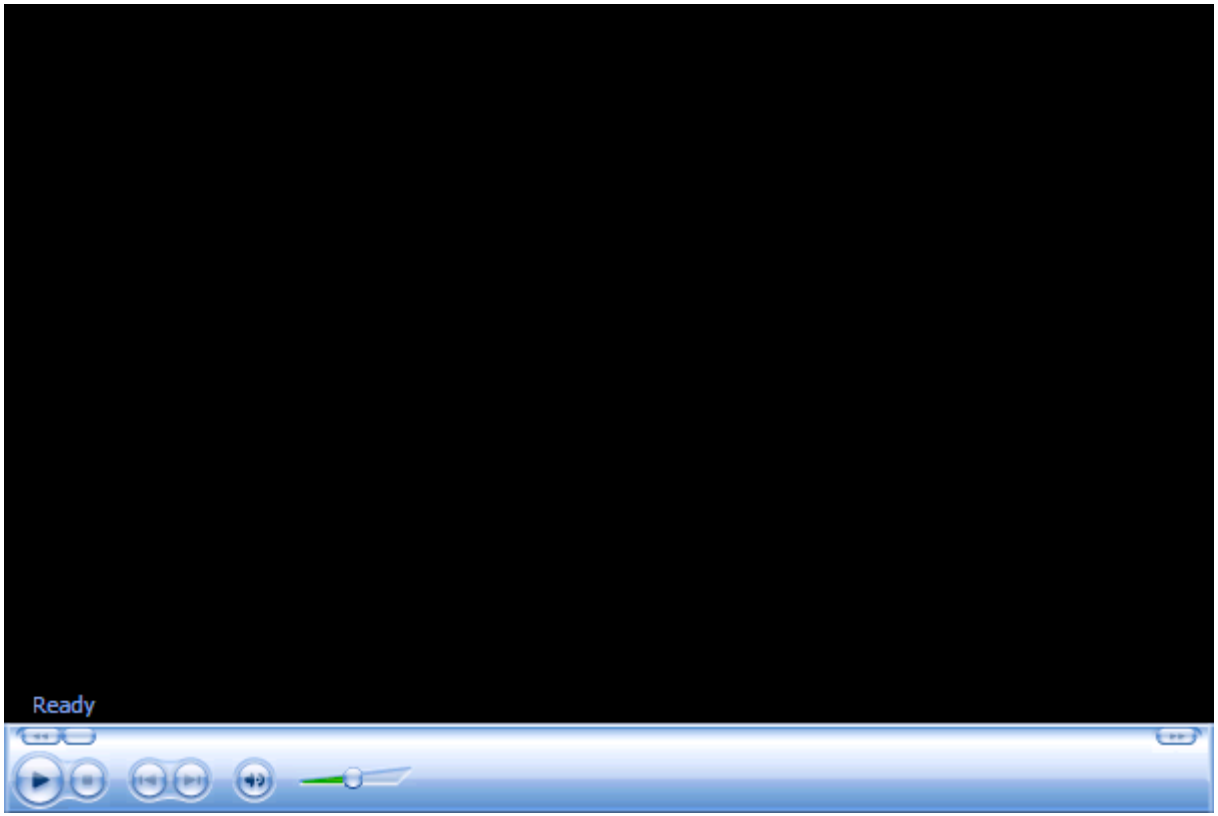
where,

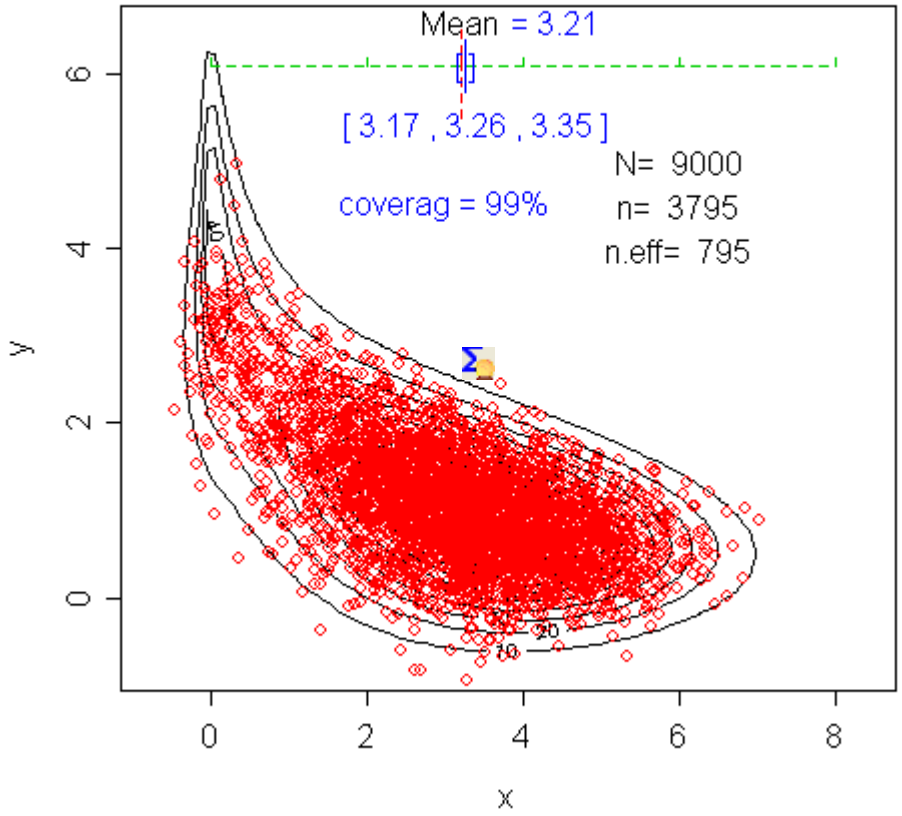
$$\Sigma = 1.2 \cdot \begin{pmatrix} 1.58 & -0.55 \\ -0.55 & 0.53 \end{pmatrix}$$

MH: good implementation



MH: good implementation




Mean = 3.21
[3.17 , 3.26 , 3.35]
coverag = 99%
N= 9000
n= 3795
n.eff= 795
Σ
y
x
0
2
4
6
8
0
2
4
6
20
10

Further Discussion

- How large should N_0 and N be?

  Not an easy problem!

- Key difficulty:

  multiple modes(多峰) in unknown area

- We would like to know all (major) modes, as well as their surrounding mass.

  Not just the global mode (最高峰)

  We need "automatic, Hill-climbing" algorithms.

- The Expectation/Maximization (EM) Algorithm, which can be viewed as a deterministic version of Gibbs Sampler.

Drive/Drink Safely,

Don’t become a Statistic;

Go to Graduate School,

Become a Statistician!

“Unfortunately, there’s no law against driving after doing triple shifts.”