Introduction to Bayesian Statistics - 5

Edoardo Milotti

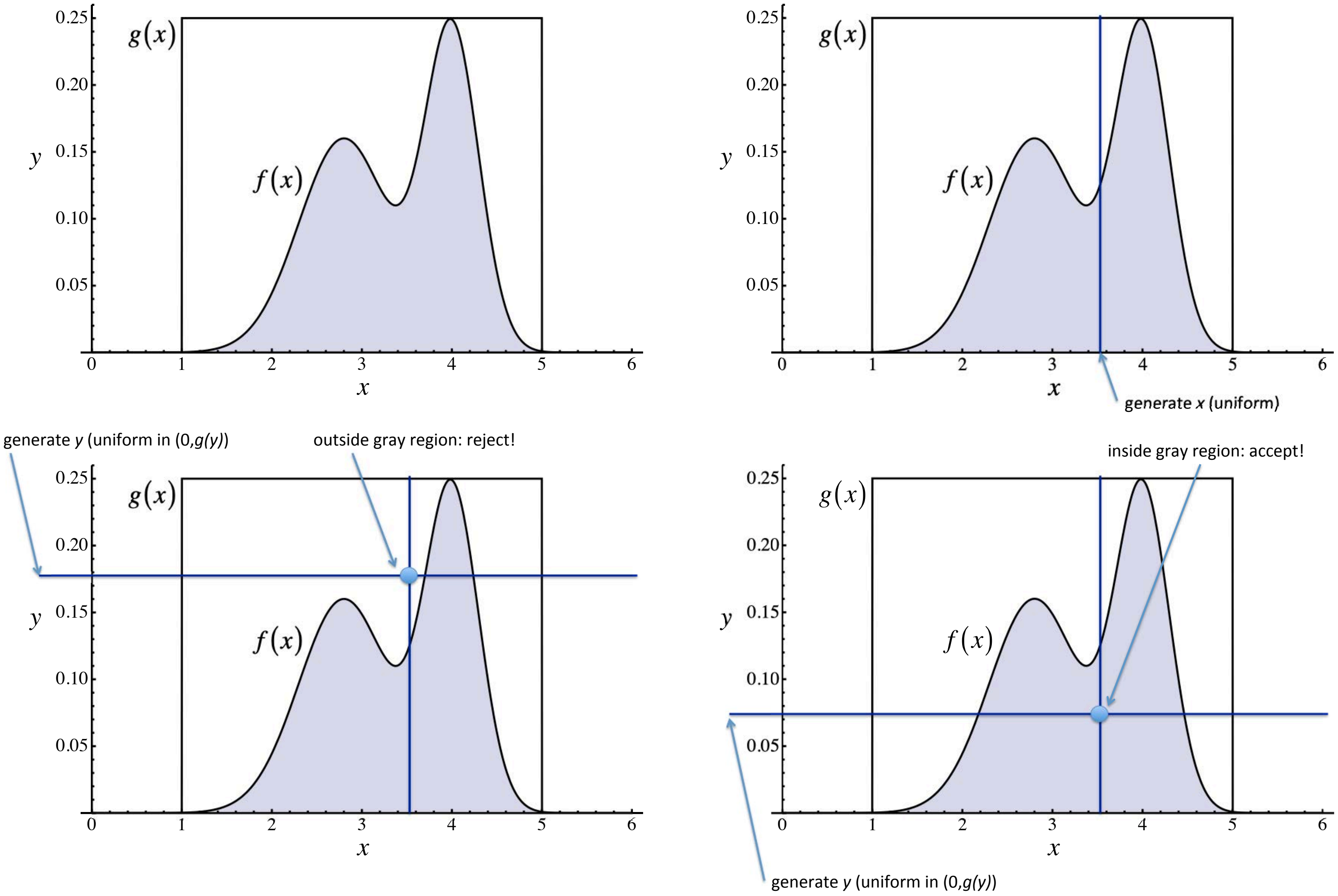
Università di Trieste and INFN-Sezione di Trieste

Bayesian estimates often require the evaluation of complex integrals. Usually these integrals can only be evaluated with numerical methods.

➡ enter the Monte Carlo methods!

1. acceptance-rejection sampling
2. importance sampling
3. statistical bootstrap
4. Bayesian methods in a sampling-resampling perspective
5. introduction to Markov chains and to the Metropolis algorithm
6. Markov Chain Monte Carlo (MCMC)

1. The acceptance rejection method



generate *x* (uniform)

generate *y* (uniform in (0,*g(y)*)

outside gray region: reject!

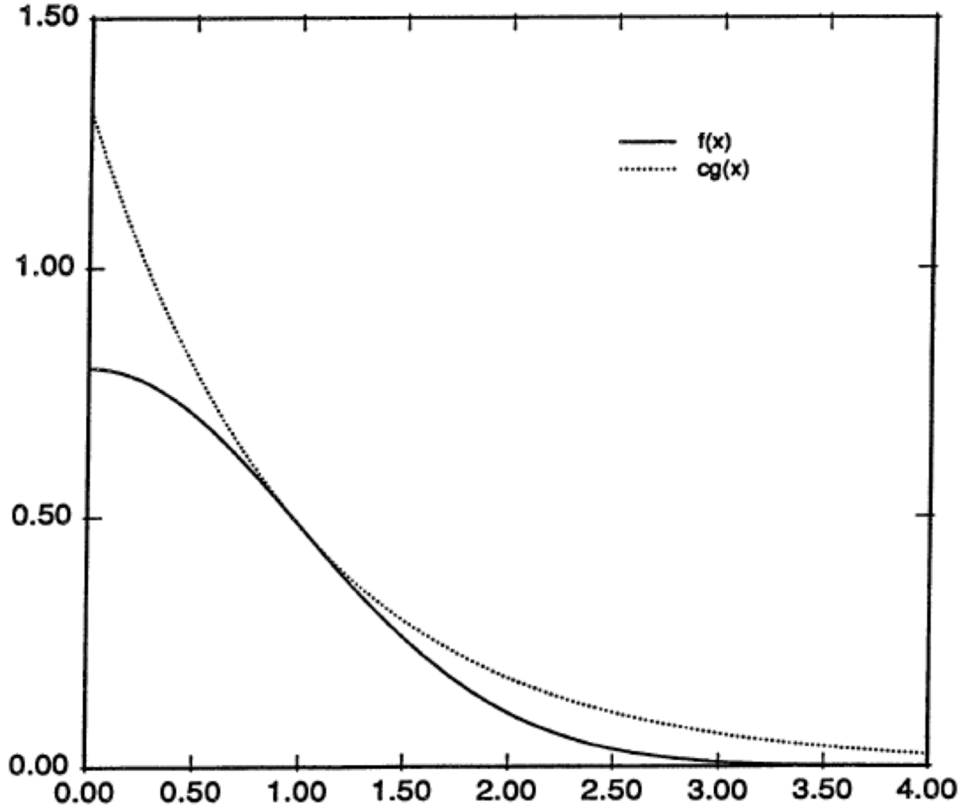
inside gray region: accept!

generate *y* (uniform in (0,*g(y)*)

Example: random numbers with semi-Gaussian distribution from exponentially distributed random numbers.

$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad x \geq 0$$

$$g(x) = \exp(-x)$$



Definition of contact point (to maximize efficiency)

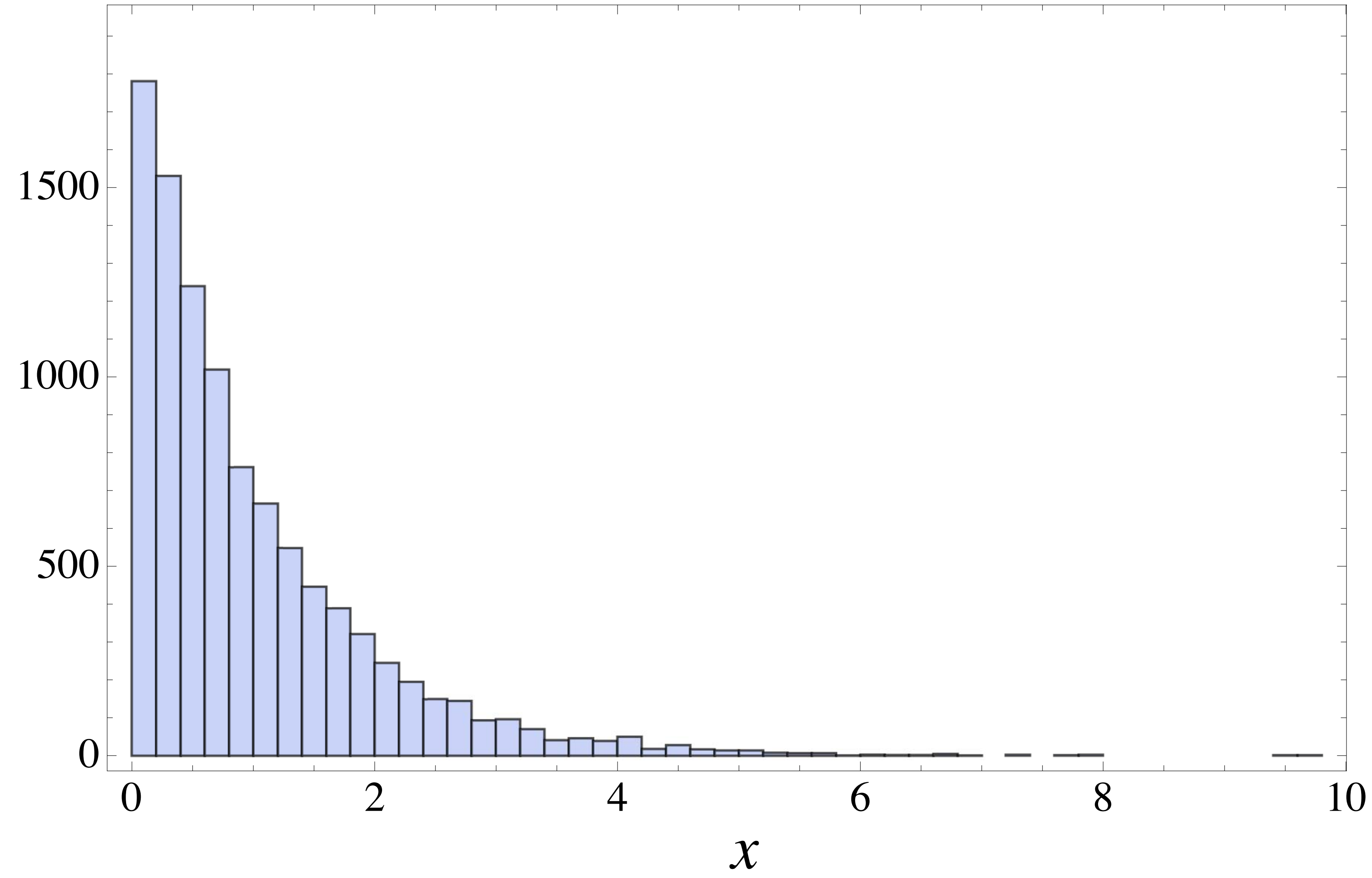
$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad x \geq 0$$

$$g(x) = \exp(-x)$$

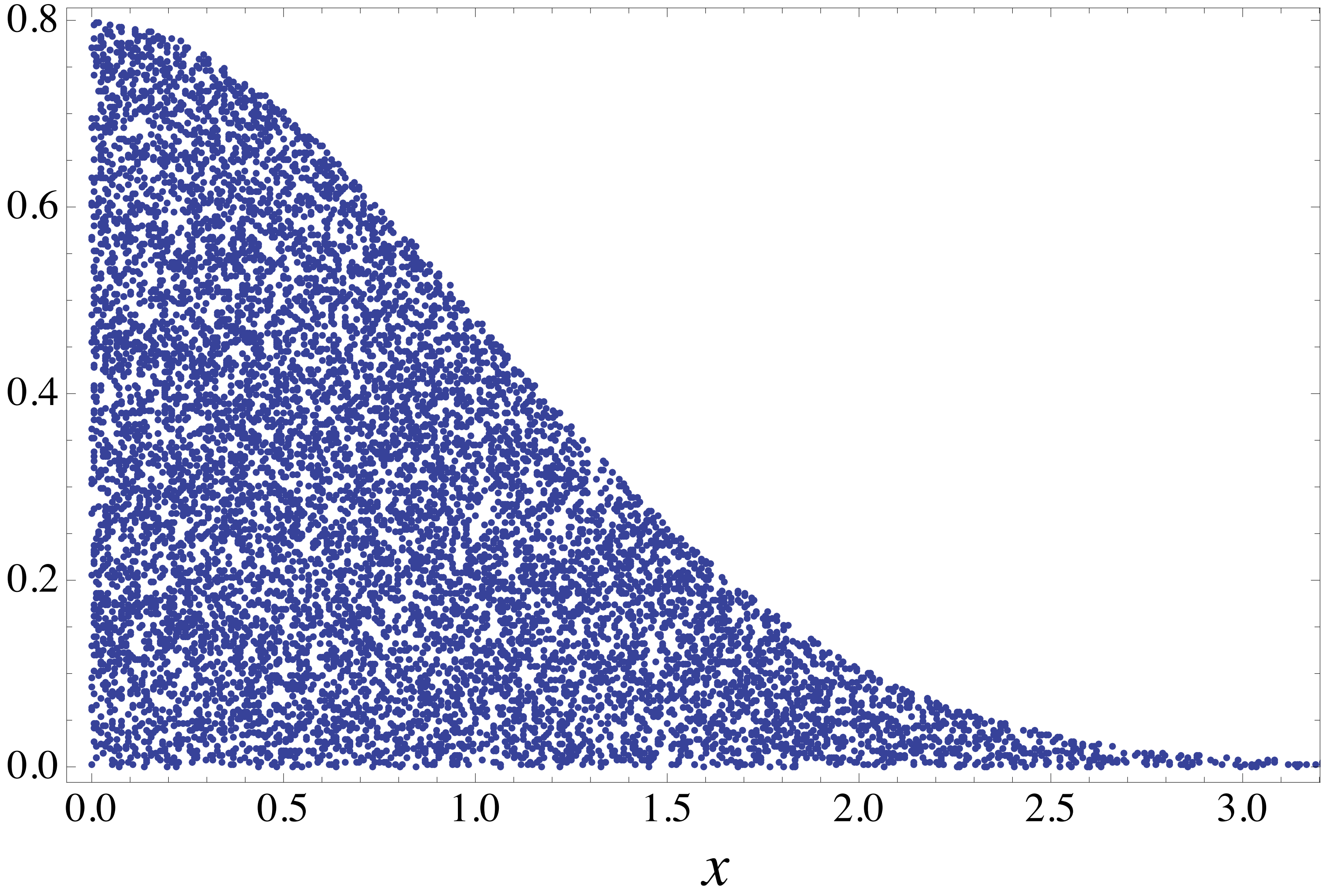
$$\Rightarrow \quad \begin{cases} f(x) = cg(x) \\ f'(x) = cg'(x) \end{cases} \quad \Rightarrow \quad \begin{cases} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) = c\exp(-x) \\ x\sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) = c\exp(-x) \end{cases}$$

$$\Rightarrow \quad x = 1; \quad c = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2} + x\right) \approx 1.31549$$

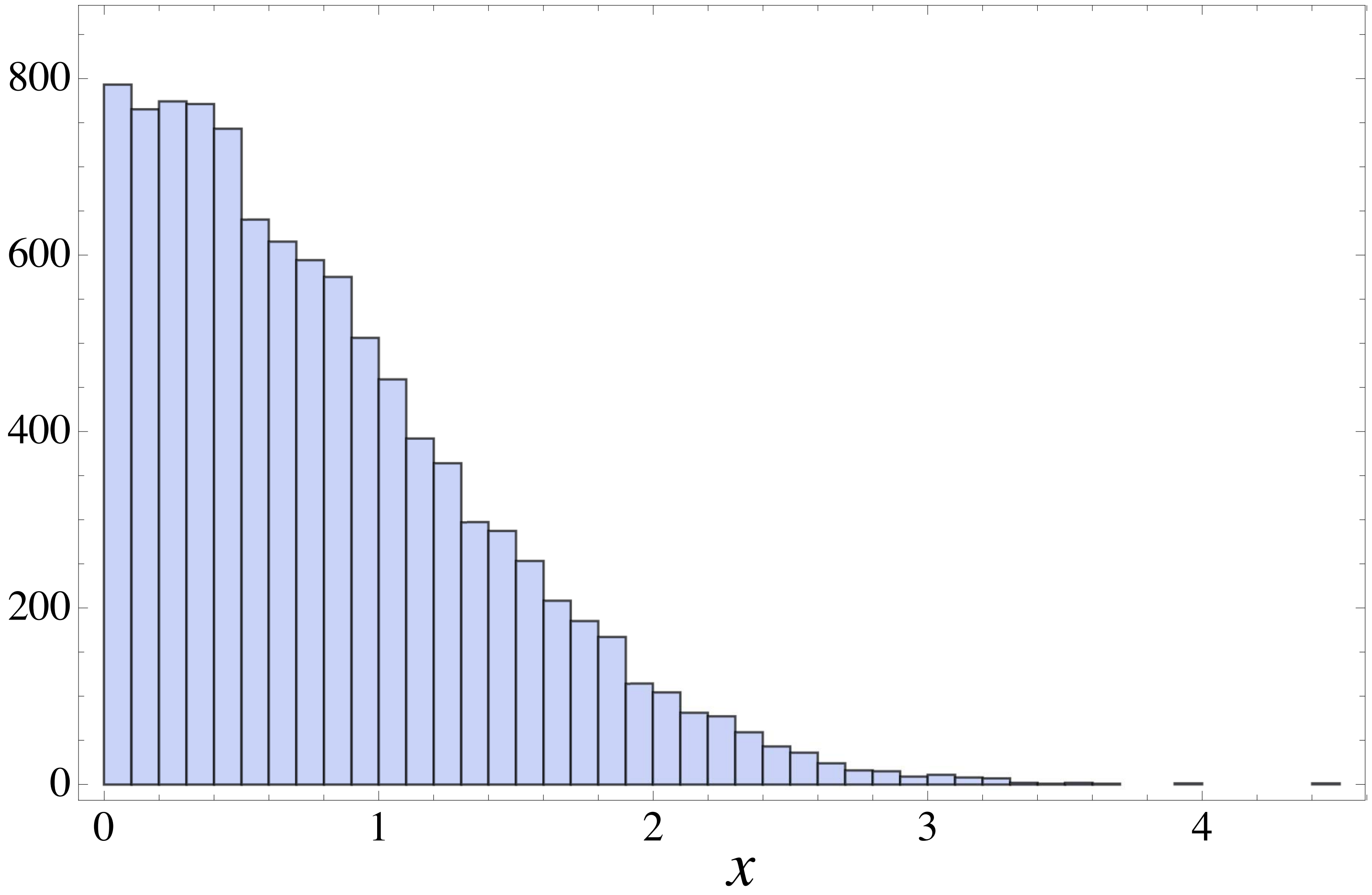
Exponentially distributed values



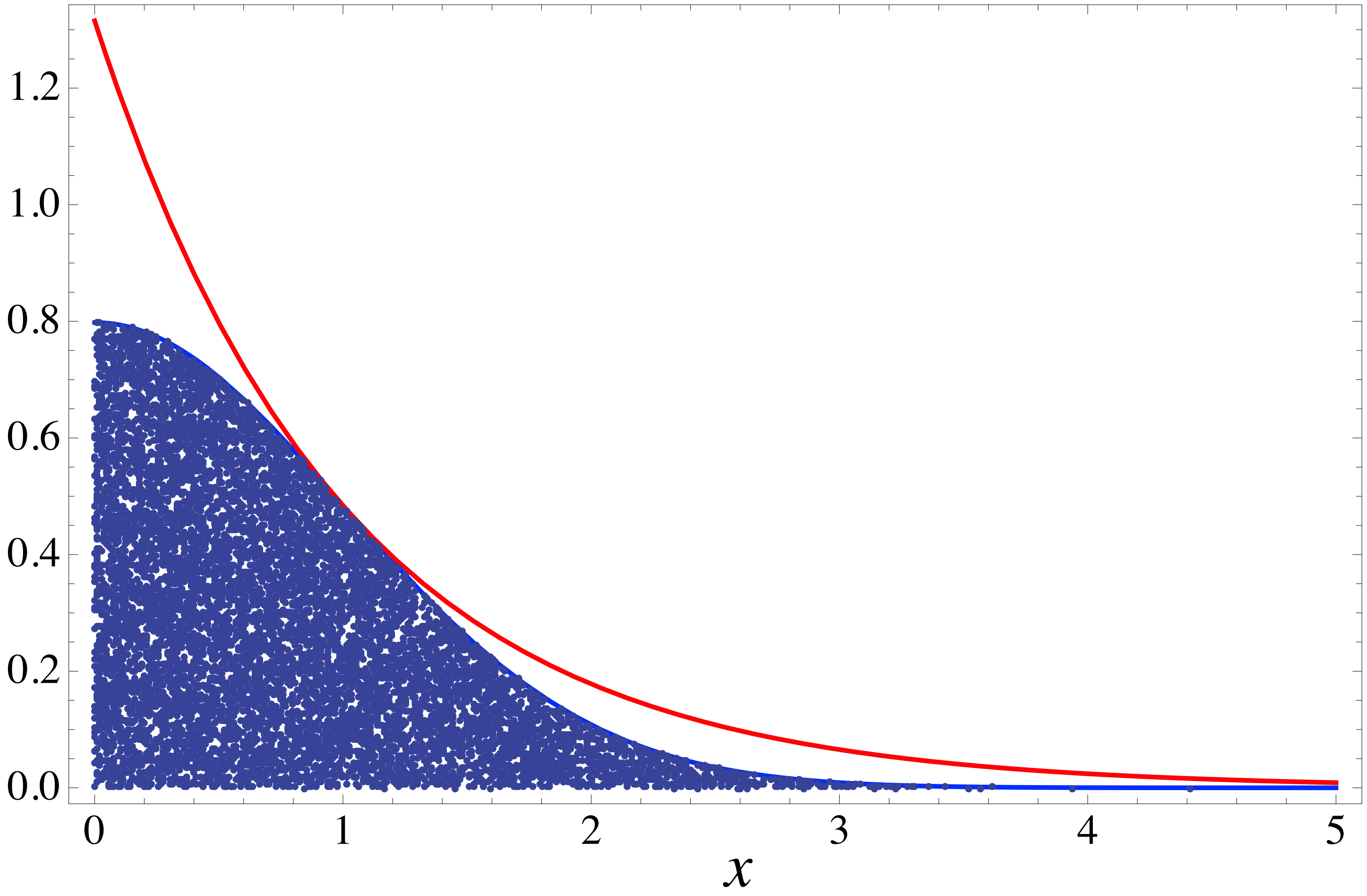
A/R accepted values (10000 accepted sample pairs)

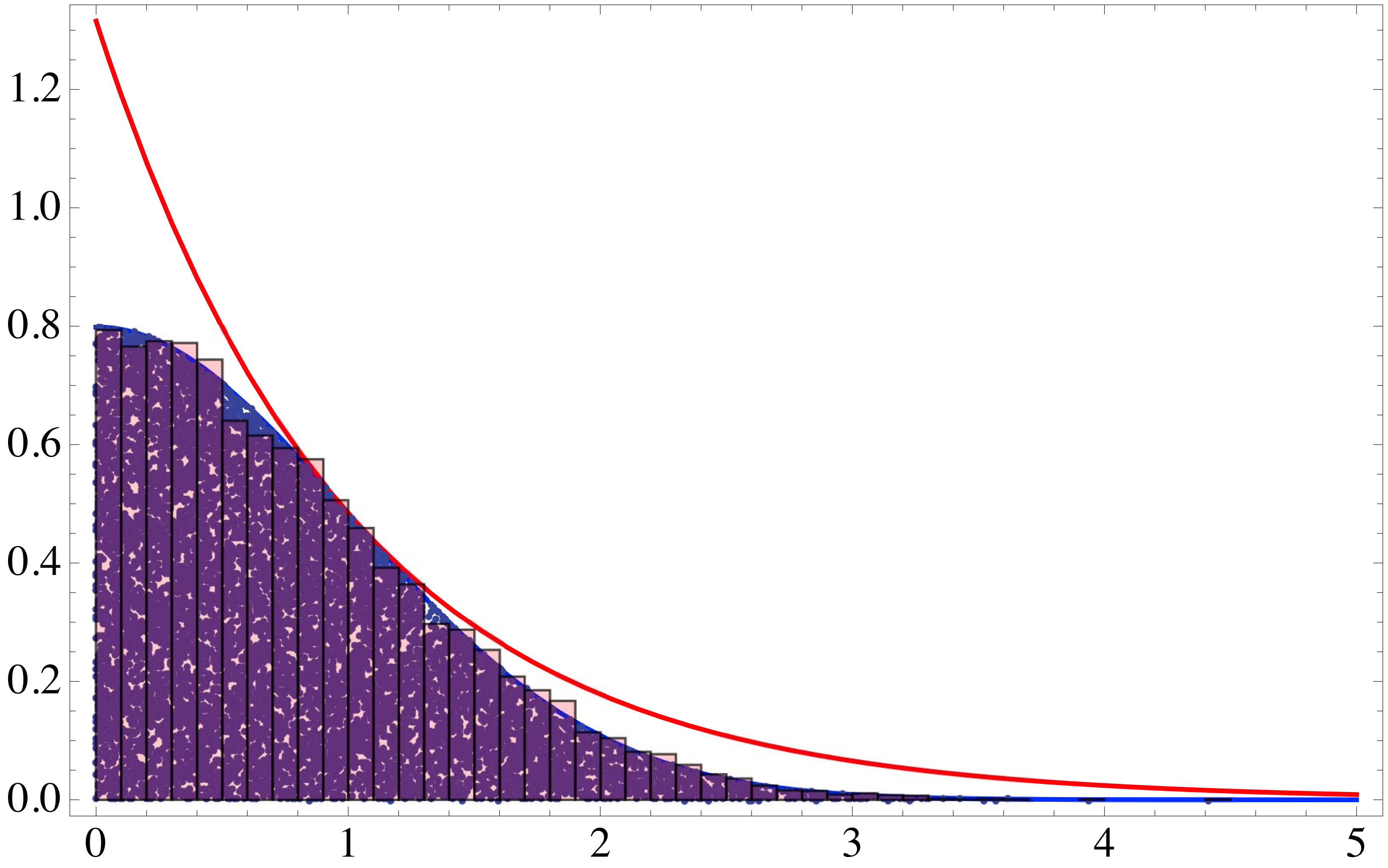


Histogram of accepted *x* values



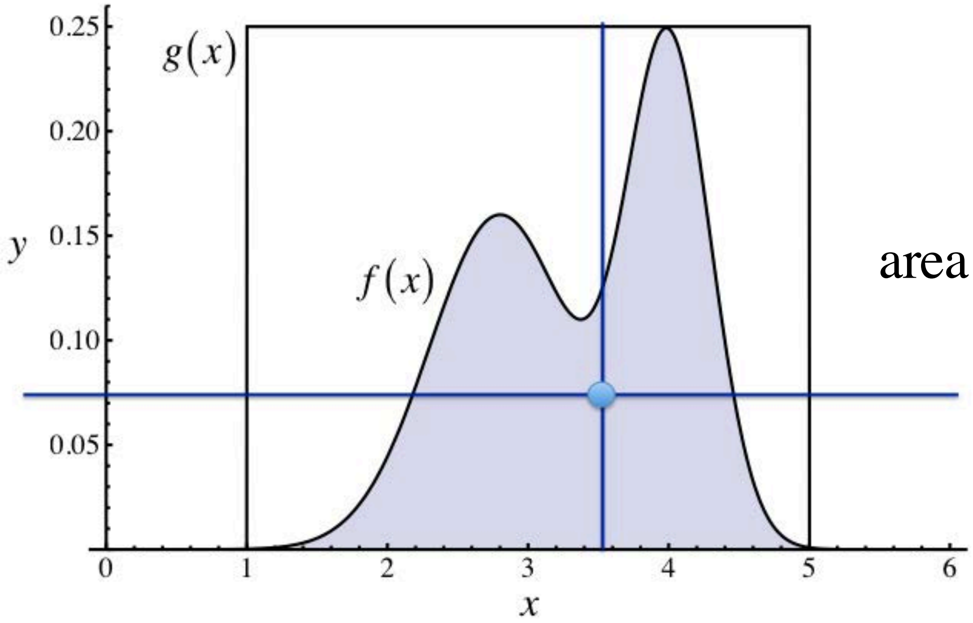
Comparison with the original distributions




1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5

Now notice that in this method we generate pairs of real numbers (u,θ) that are uniformly distributed between $f(\theta)$ and the x-axis, therefore we can use these pairs to estimate the total area under the curve

(here the reference area is the area of the enclosing rectangle which corresponds to a uniform distribution)



$$\text{area} = \frac{\#\text{ of accepted pairs}}{\#\text{ of pairs}}\text{reference area}$$

In general, if $h(x) = f(x)p(x)$, where *p* is a pdf

$$\int_a^b h(x)dx = \int_a^b f(x)p(x)dx = E_p\left[f(x)\right] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

here the *x* are i.i.d with pdf *p(x)*

and we find that the variance of this estimate of the integral is

$$\frac{1}{N}\left\{\frac{1}{N-1}\sum_{n=1}^{N}\left[f(x_n) - E_p\left[f(x)\right]\right]^2\right\}$$

We encounter a problem with this method when we must sample functions that have many narrow peaks.

2. *Importance sampling*

this pdf is troublesome ...

therefore we use this ...

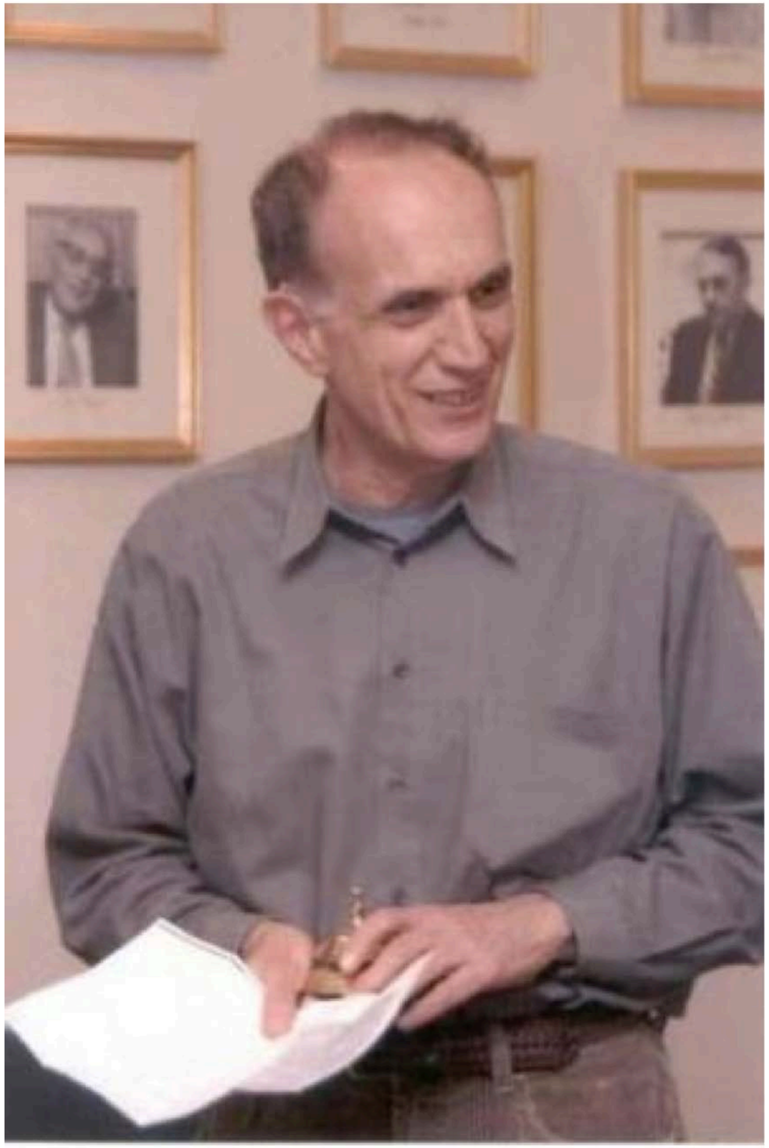
$$\int_a^b h(x)\,dx = \int_a^b f(x)\,p(x)\,dx = \int_a^b \left[f(x)\frac{p(x)}{q(x)} \right] q(x)\,dx$$

$$= E_q\left[f(x)\frac{p(x)}{q(x)} \right] \approx \frac{1}{N}\sum_{n=1}^{N} f(x_n)\frac{p(x_n)}{q(x_n)}$$

here the *x* are i.i.d with pdf *q(x)*

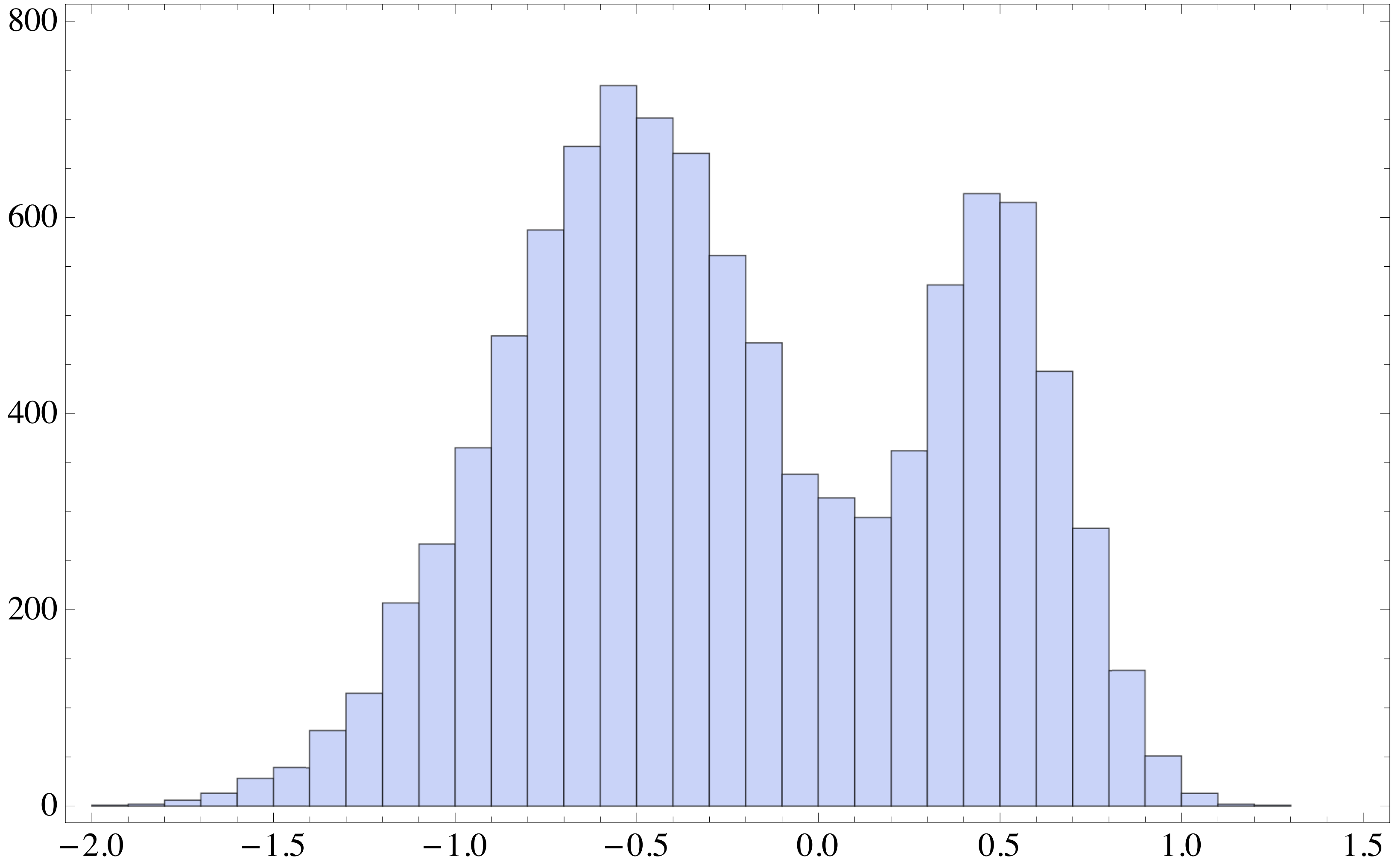
These methods are still not very efficient and there is a better alternative, the Markov Chain Monte Carlo method

3. Bootstrap (B. Efron, 1977) and the importance of edf's

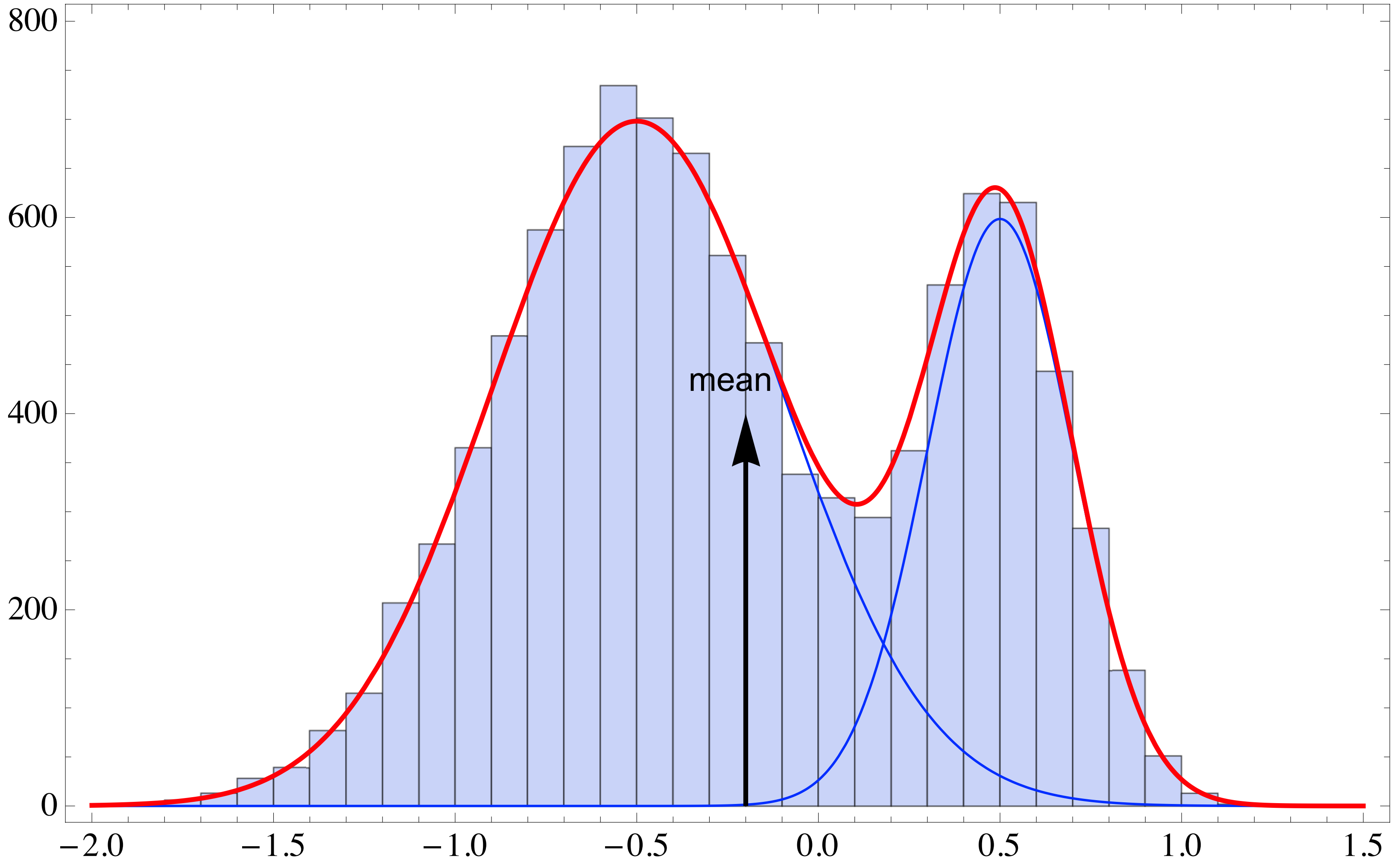


The bootstrap method is a resampling technique that helps calculate many statistical estimators

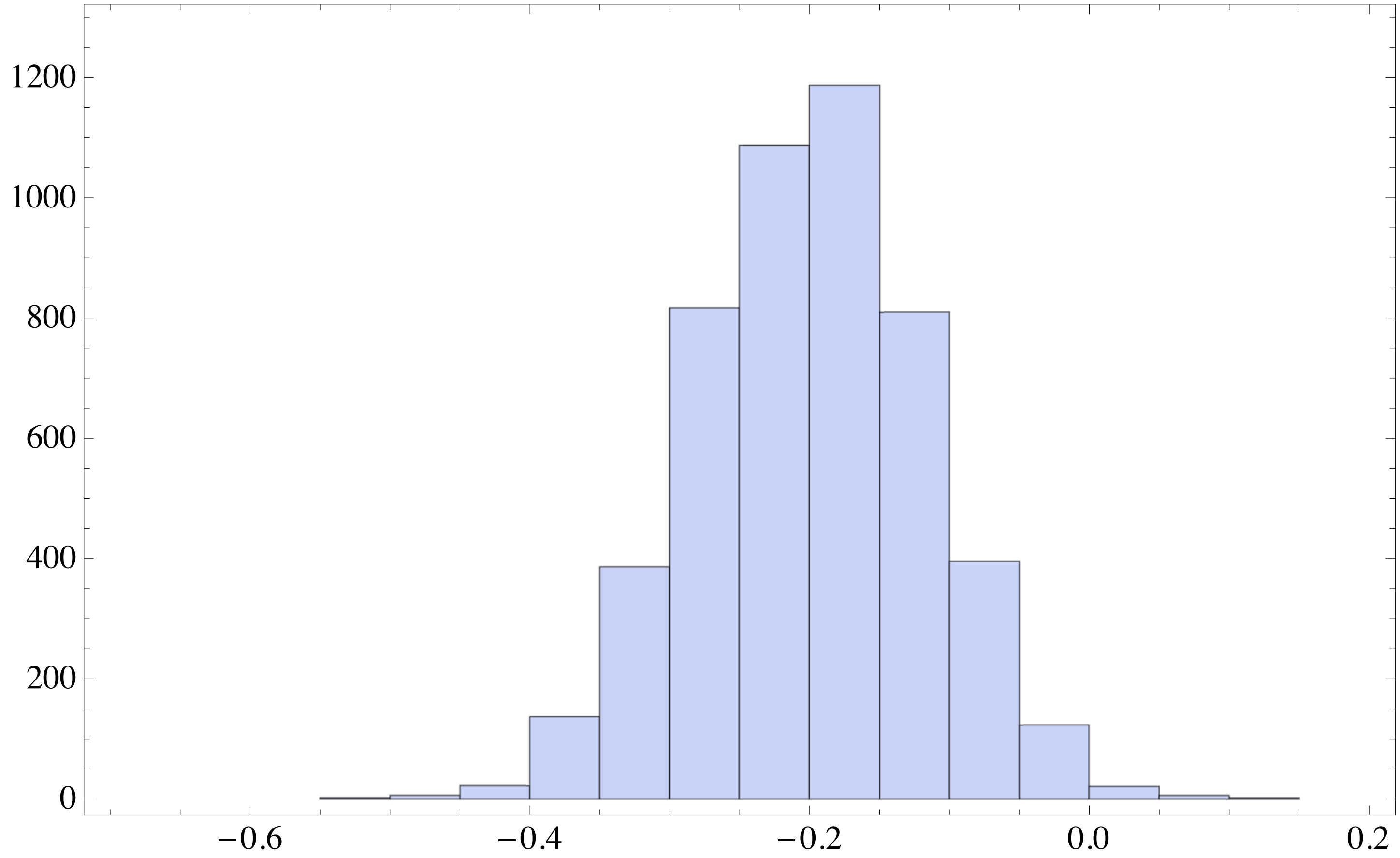
consider the distribution of a set of measurements



the distribution of data is an approximation of the “true” underlying distribution (in this case a mixture model)

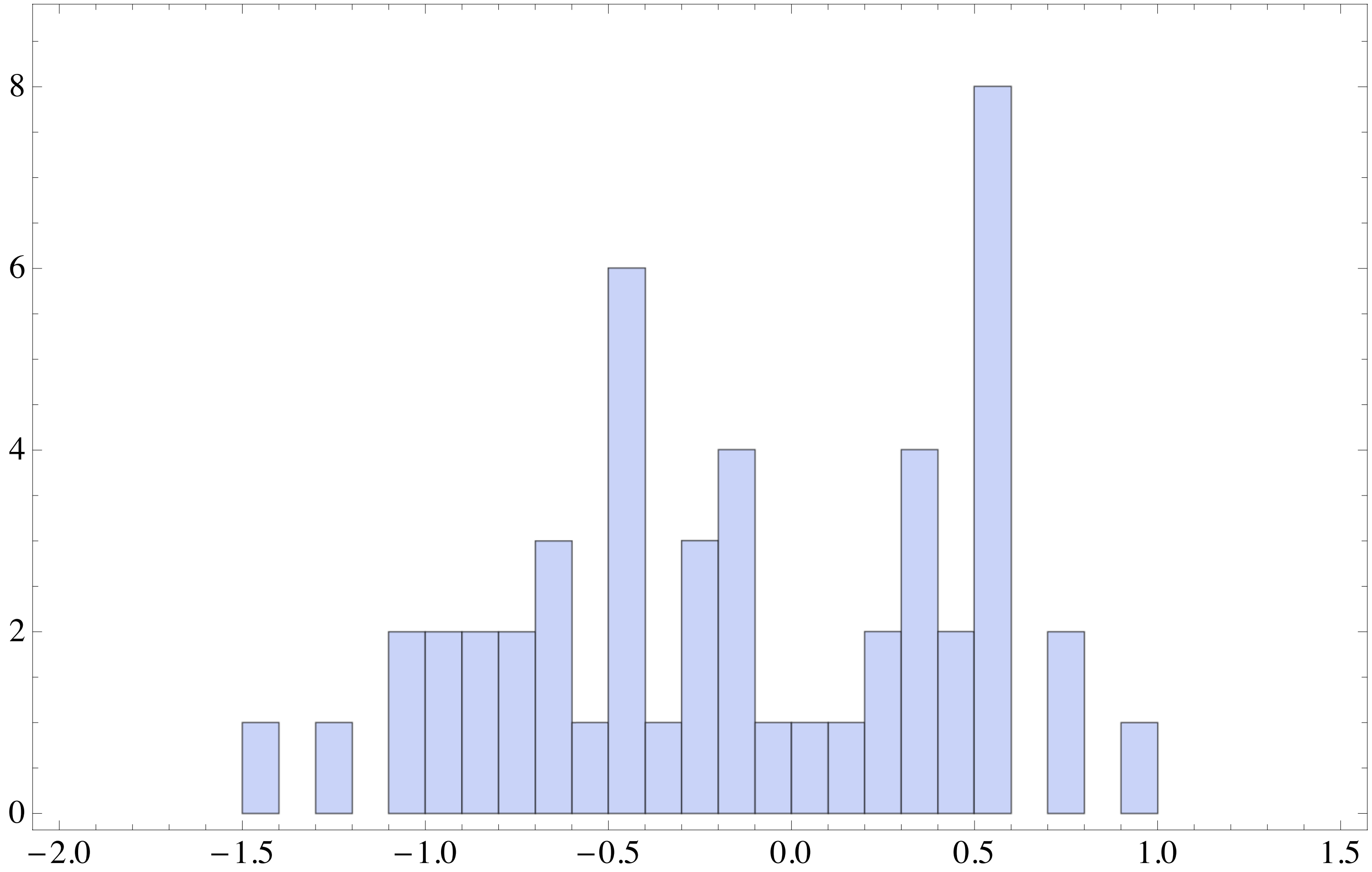


distribution of mean value obtained from 5000 sets of data (sample size = 50)



You can do this if you have large datasets ... but what if you have only a handful of measurements?

example: single dataset (same size as before, 50 measurements)



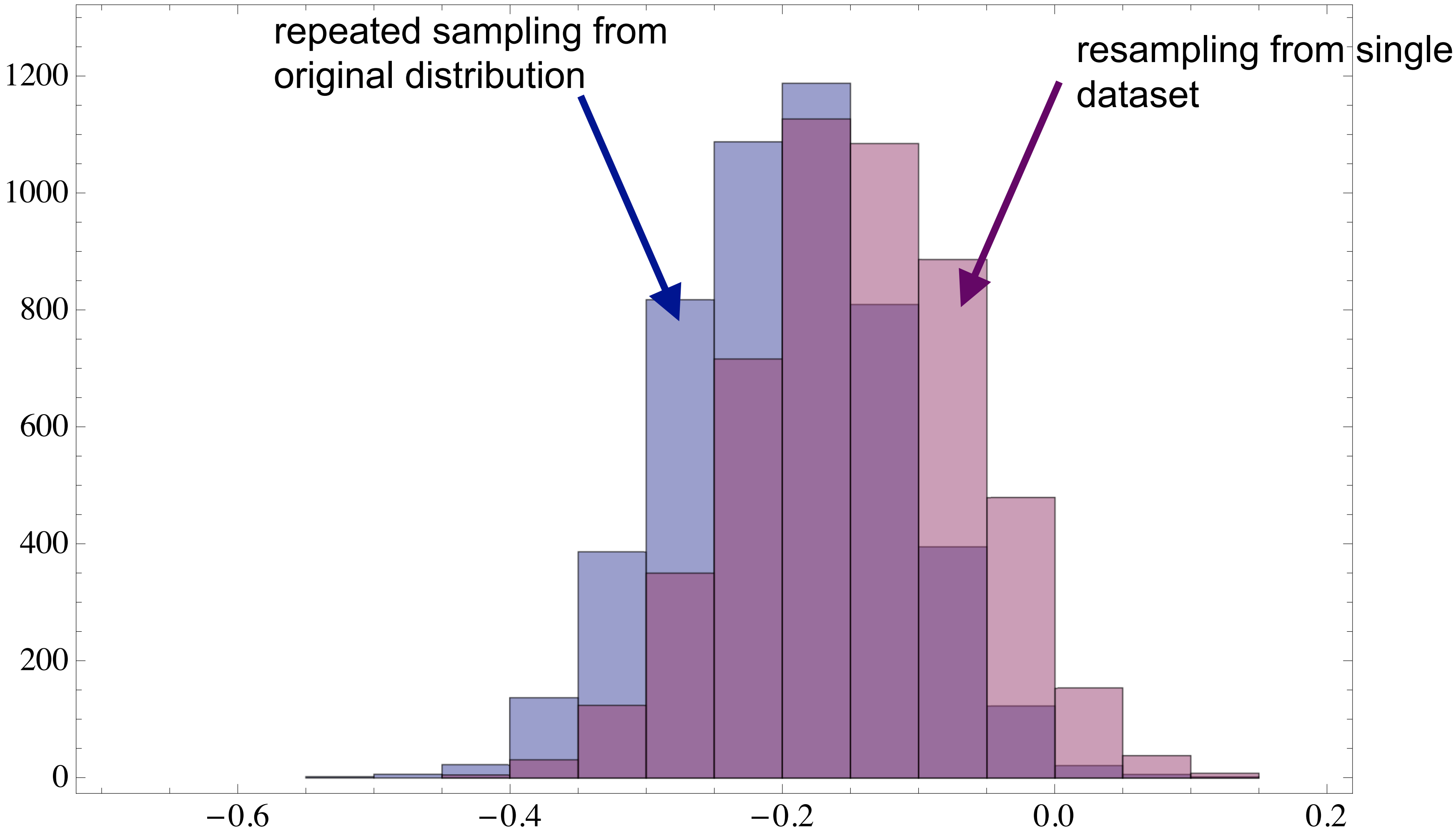
the distribution is a rough representation of the underlying distribution ... and yet it can be used just as before ...

Bootstrap recipe:

if you want to find the distribution of the mean (or any other statistical estimator) use the dataset itself to generate new datasets

➡ resample from dataset (with replacement)

distribution of mean value



true mean: -0.2
mean from repeated sampling (size = 250000): -0.200222 ± 0.0813632
mean from resampling dataset (size = 50): -0.142699 ± 0.0838678

counts of CD4 limphocytes

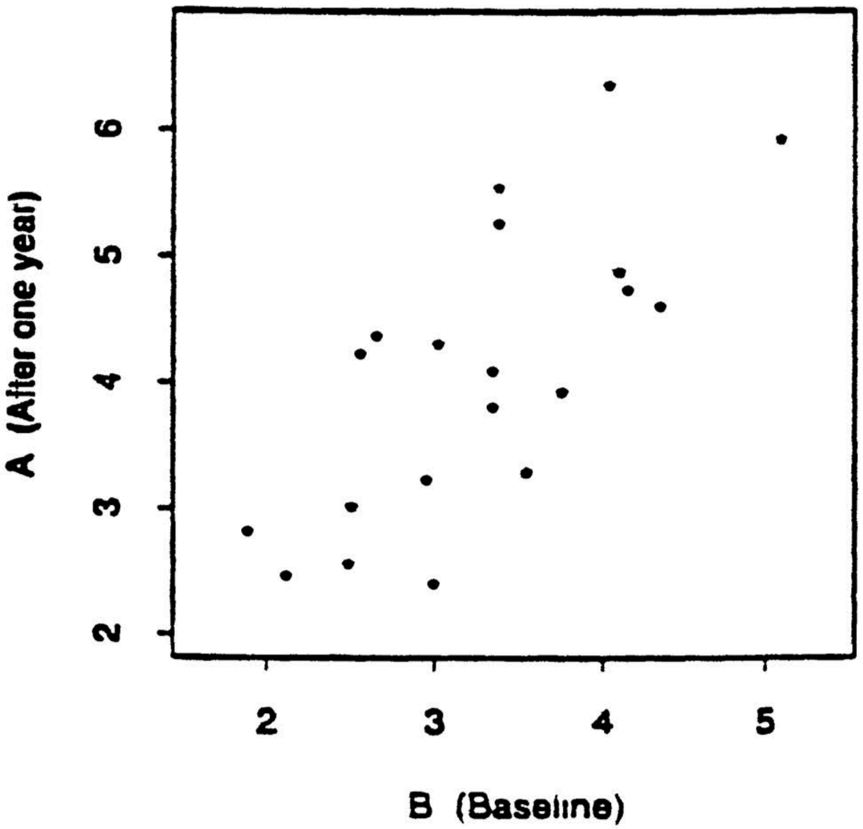


FIG. 1. *The cd4 data; cd4 counts in hundreds for 20 subjects, at baseline and after one year of treatment with an experimental anti-viral drug; numerical values appear in Table 1.*

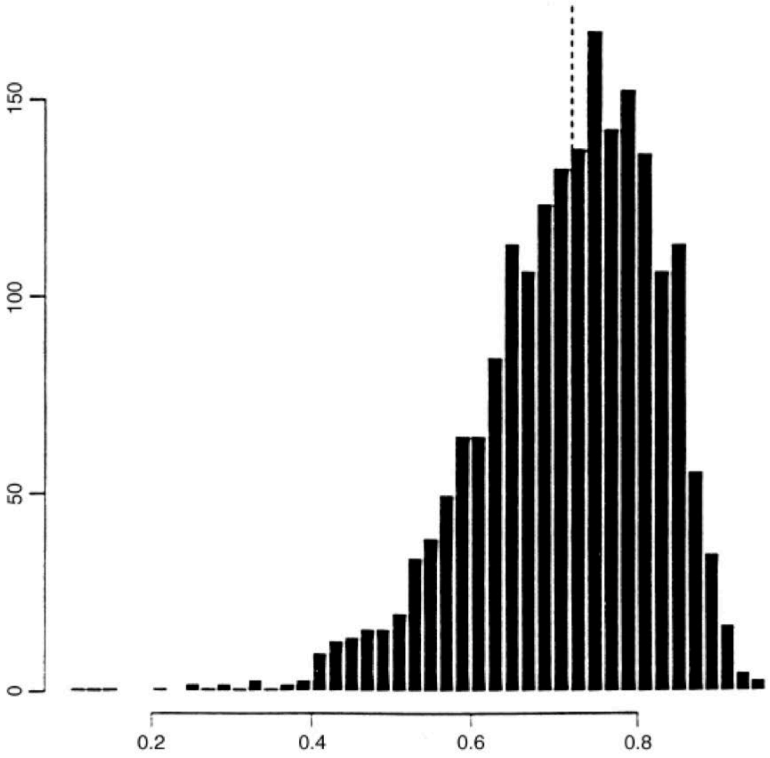
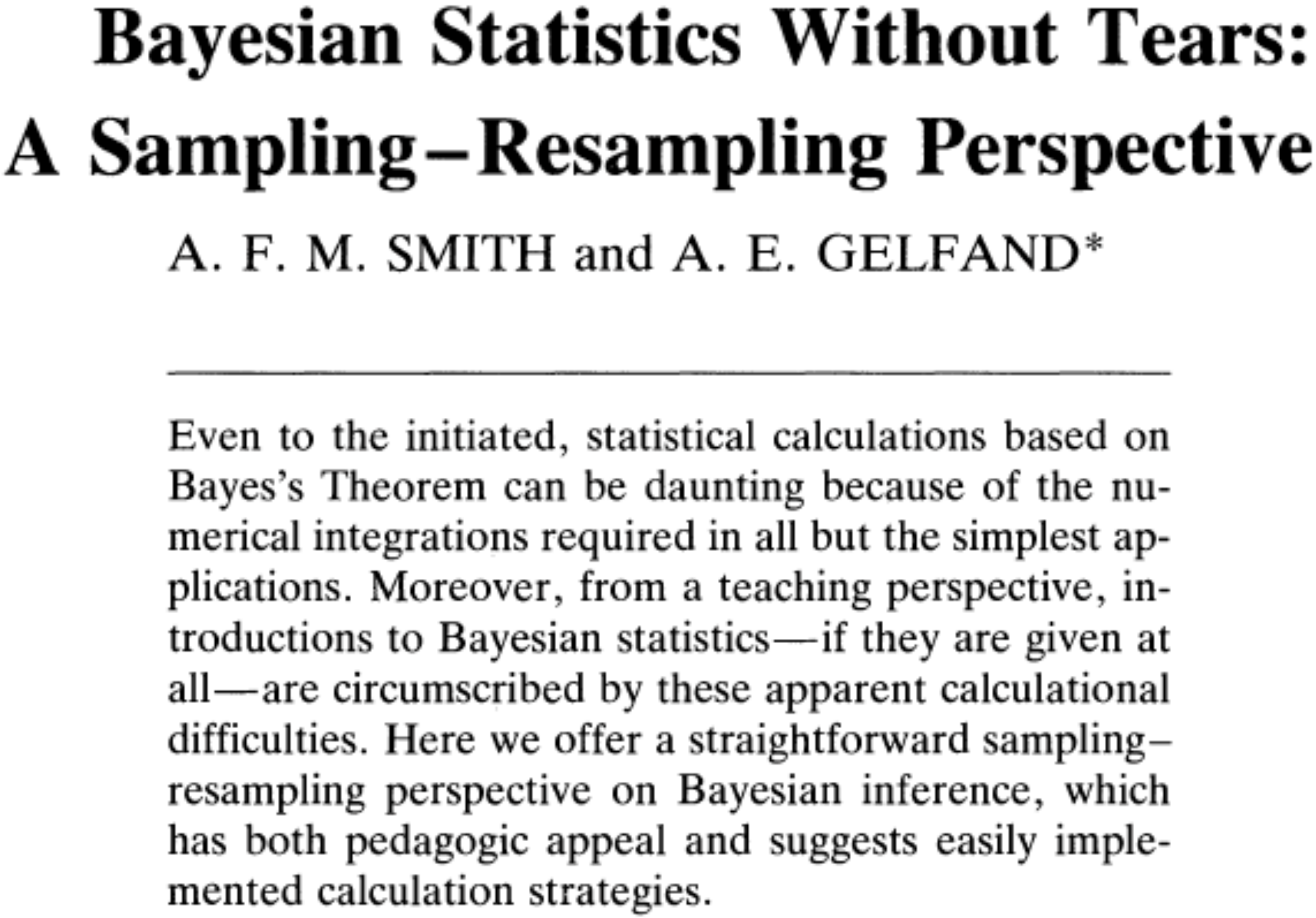


FIG. 3. *Histogram of 2,000 bootstrap correlation coefficients; bivariate normal sampling model.*

bootstrap estimate of correlation coefficient distribution

Example from Di Ciccio & Efron, Statistics of Science **11** (1996) 189 and Efron, Statistics of Science **13** (1998) 95

4. Bayesian methods in a sampling-resampling perspective (Smith & Gelfand, 1992)



Bayesian Statistics Without Tears: A Sampling–Resampling Perspective

A. F. M. SMITH and A. E. GELFAND*

Even to the initiated, statistical calculations based on Bayes's Theorem can be daunting because of the numerical integrations required in all but the simplest applications. Moreover, from a teaching perspective, introductions to Bayesian statistics—if they are given at all—are circumscribed by these apparent calculational difficulties. Here we offer a straightforward sampling–resampling perspective on Bayesian inference, which has both pedagogic appeal and suggests easily implemented calculation strategies.

In Bayesian methods we have to evaluate many integrals, like, e.g.,

$$p(\theta|x) = \frac{l(\theta;\, x)p(\theta)}{\int l(\theta;\, x)p(\theta)\; d\theta}$$

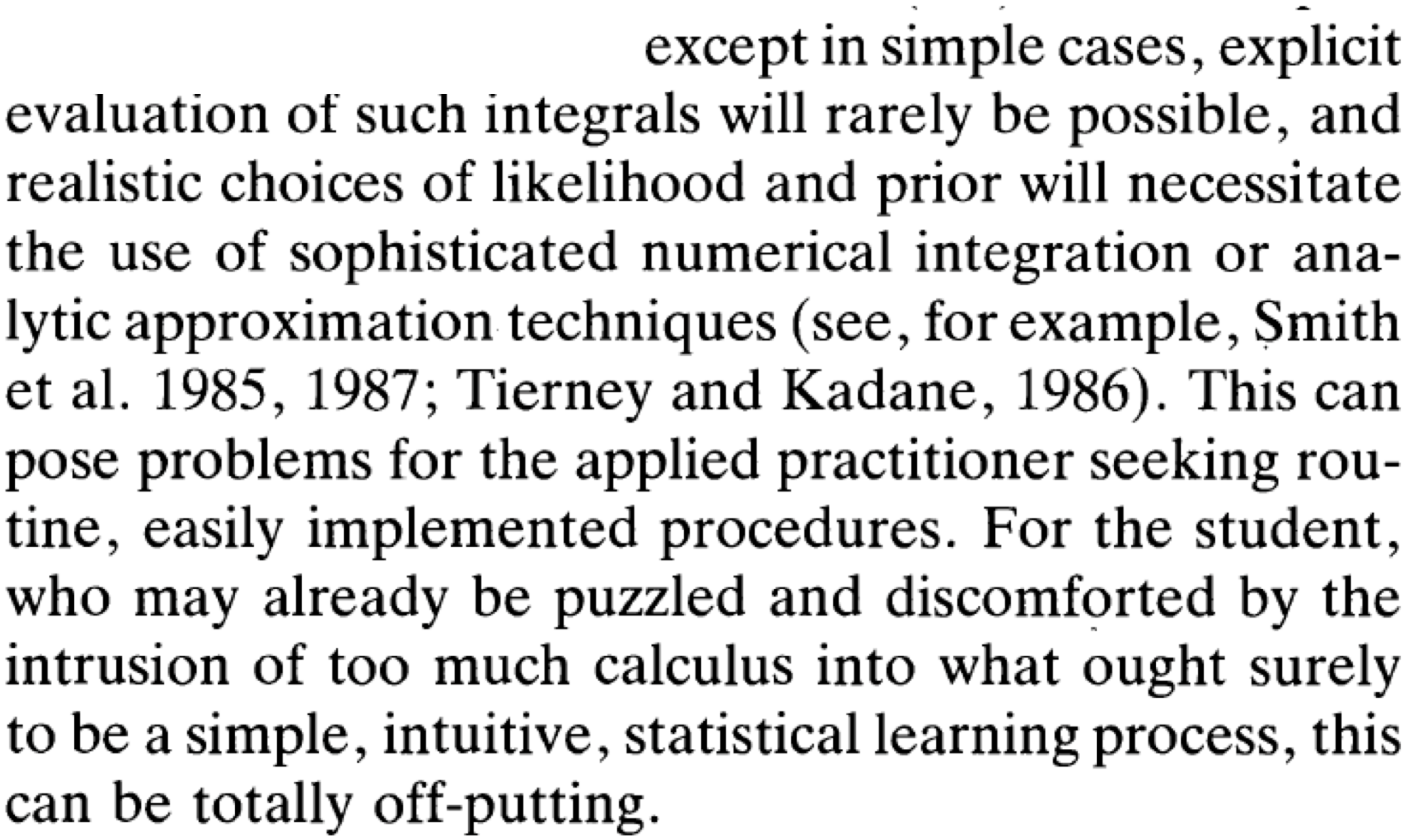
normalization (evidence)

$$p(\phi|x) = \int p(\phi,\, \psi|x)\; d\psi.$$

marginalization

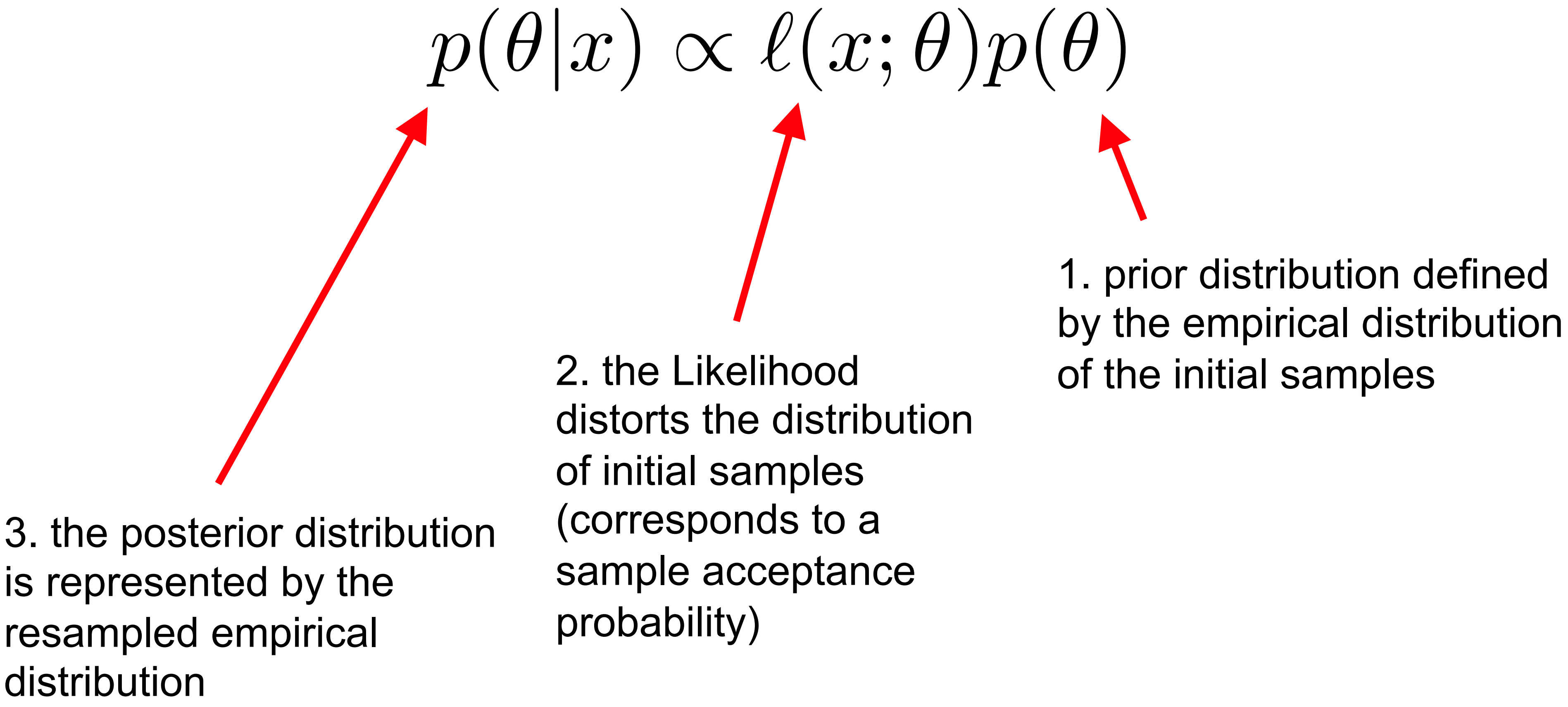
$$E[m(\theta)|x] = \int m(\theta)p(\theta|x)\; d\theta$$

averages (statistical estimators)



except in simple cases, explicit evaluation of such integrals will rarely be possible, and realistic choices of likelihood and prior will necessitate the use of sophisticated numerical integration or analytic approximation techniques (see, for example, Smith et al. 1985, 1987; Tierney and Kadane, 1986). This can pose problems for the applied practitioner seeking routine, easily implemented procedures. For the student, who may already be puzzled and discomforted by the intrusion of too much calculus into what ought surely to be a simple, intuitive, statistical learning process, this can be totally off-putting.

Bayesian learning as a resampling procedure



$$p(\theta|x) \propto \ell(x;\theta)p(\theta)$$

1. prior distribution defined by the empirical distribution of the initial samples

2. the Likelihood distorts the distribution of initial samples (corresponds to a sample acceptance probability)

3. the posterior distribution is represented by the resampled empirical distribution

Example (McCullagh & Nelder): take two sets of binomially distributed independent random variables X_{i1} and X_{i2} (i=1,2,3)

$$X_{i1} = \mathrm{Binomial}\left(n_{i1}, \theta_1\right)$$

$$X_{i2} = \mathrm{Binomial}\left(n_{i2}, \theta_2\right)$$

The observed random variables are the sums

$$Y_i = X_{i1} + X_{i2}$$

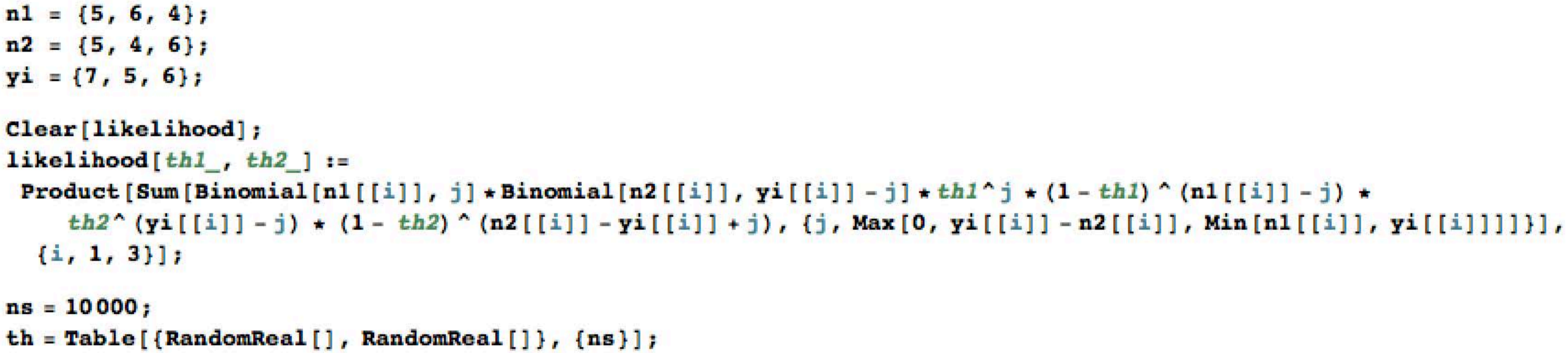
$$\text{likelihood} = \prod_{i=1}^{3} \sum_{j_i} \binom{n_{i1}}{j_i} \binom{n_{i2}}{y_i - j_i} \theta_1^{j_1} (1-\theta_1)^{n_{i1}-j_i} \theta_2^{y_i - j_i} (1-\theta_2)^{n_{i2}-y_i+j_i}$$

$$\max(0, y_i - n_{i2}) \le j_i \le \min(n_{i1}, y_i)$$

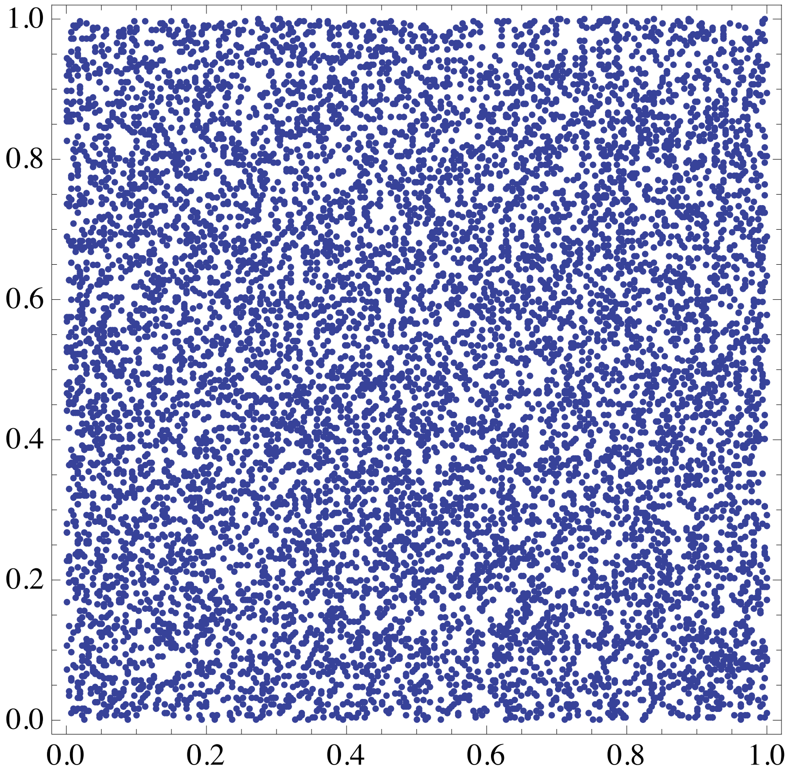
Sample data

	1	**2**	**3**
n_{i1}	5	6	4
n_{i2}	5	4	6
y_i	7	5	6

Example of implementation in *Mathematica*

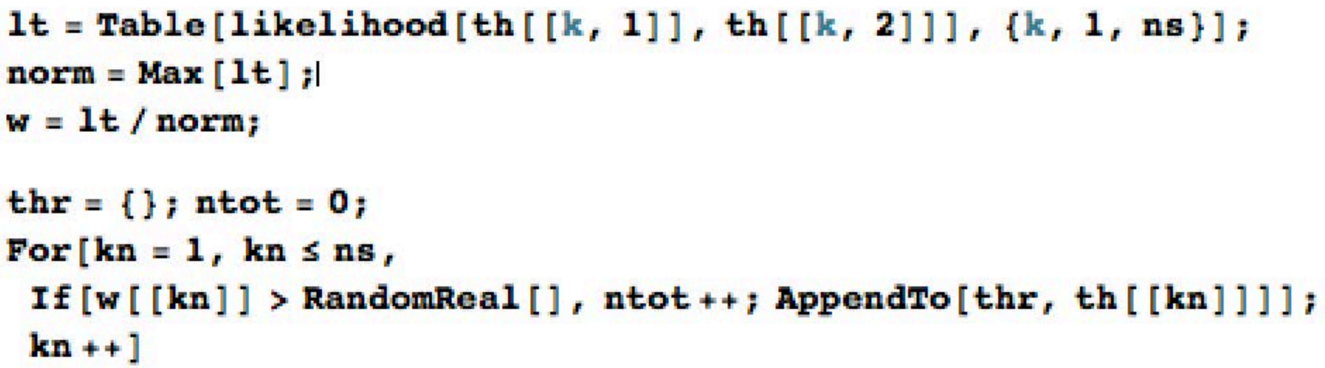


```
n1 = {5, 6, 4};
n2 = {5, 4, 6};
yi = {7, 5, 6};

Clear[likelihood];
likelihood[th1_, th2_] :=
 Product[Sum[Binomial[n1[[i]], j] * Binomial[n2[[i]], yi[[i]] - j] * th1^j * (1 - th1)^(n1[[i]] - j) *
    th2^(yi[[i]] - j) * (1 - th2)^(n2[[i]] - yi[[i]] + j), {j, Max[0, yi[[i]] - n2[[i]]], Min[n1[[i]], yi[[i]]]}],
  {i, 1, 3}];

ns = 10000;
th = Table[{RandomReal[], RandomReal[]}, {ns}];
```

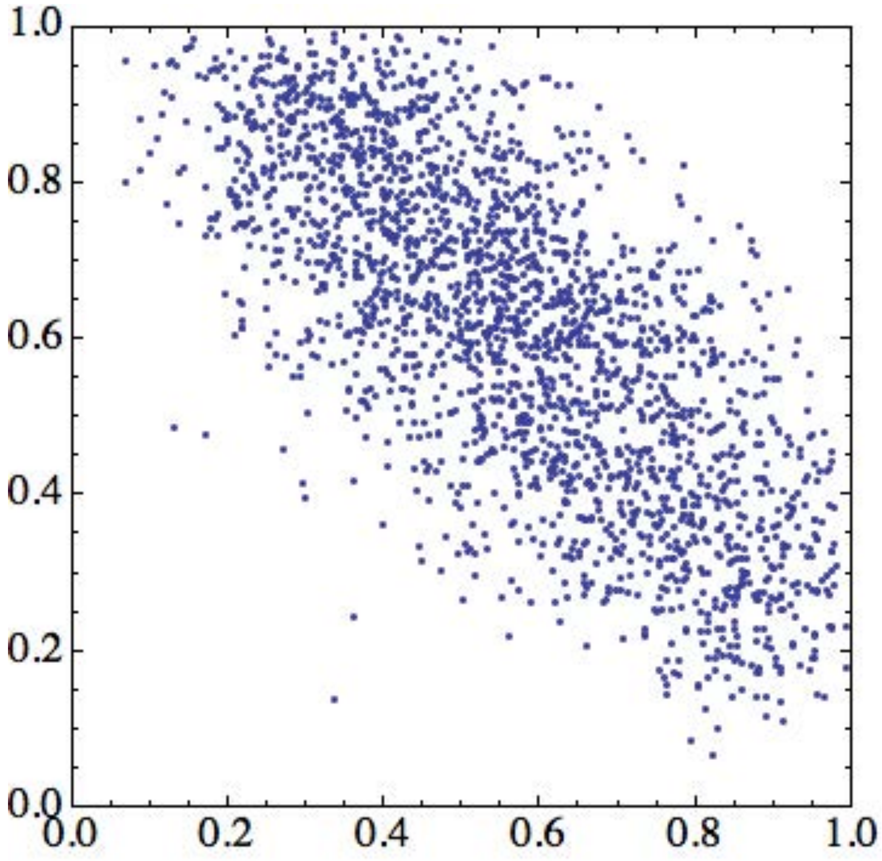


prior distribution (uniform in 2D parameter space)

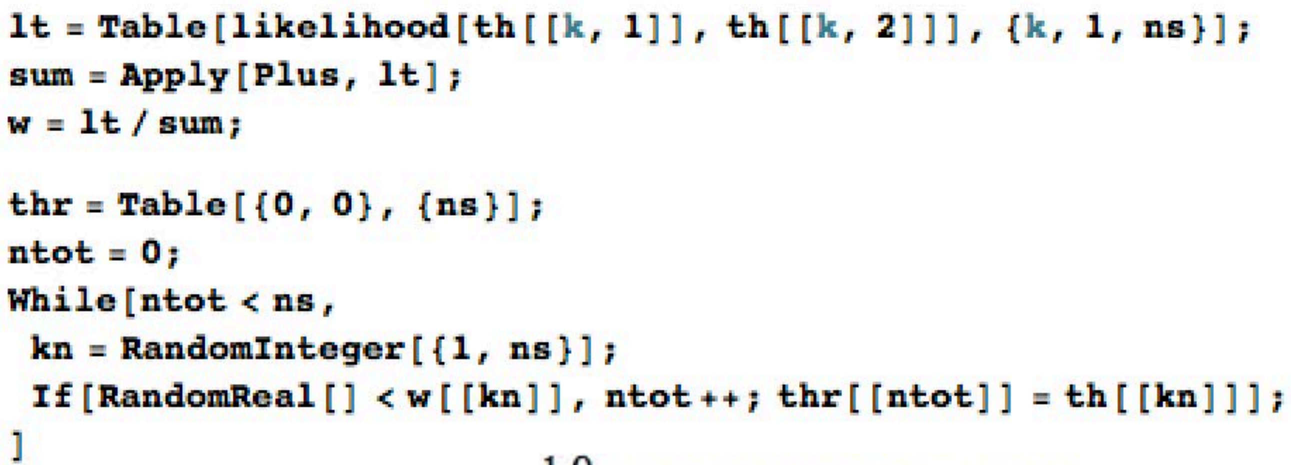
Posterior as a resampled prior using acceptance-rejection



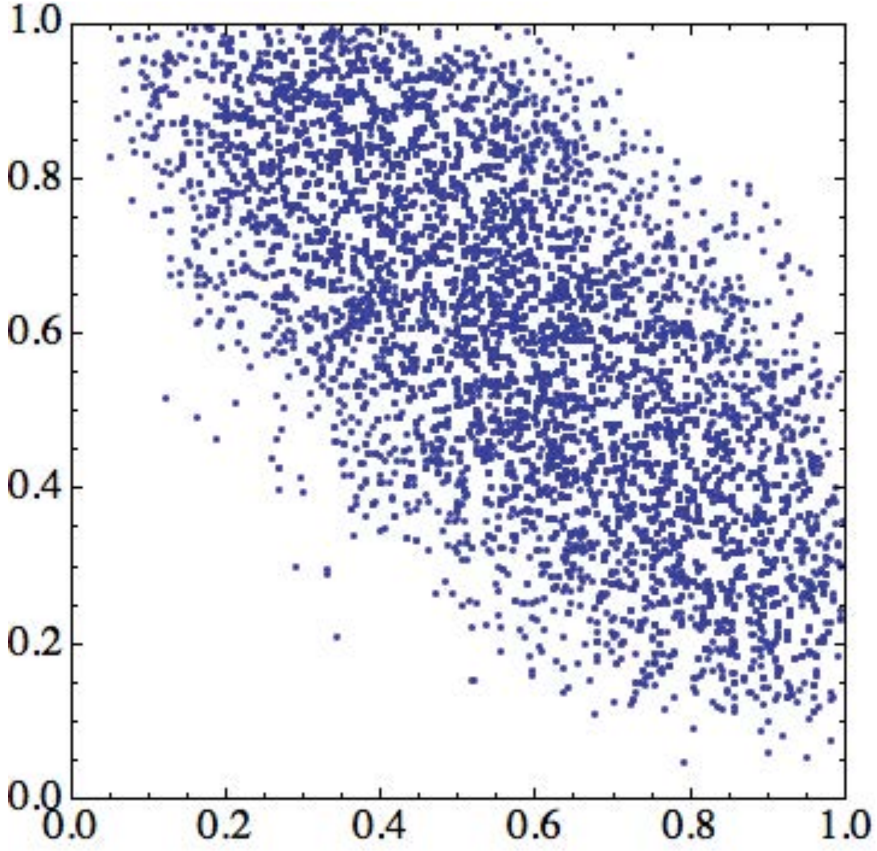
```
lt = Table[likelihood[th[[k, 1]], th[[k, 2]]], {k, 1, ns}];
norm = Max[lt];
w = lt / norm;

thr = {}; ntot = 0;
For[kn = 1, kn ≤ ns,
 If[w[[kn]] > RandomReal[], ntot++; AppendTo[thr, th[[kn]]]];
 kn++]
```



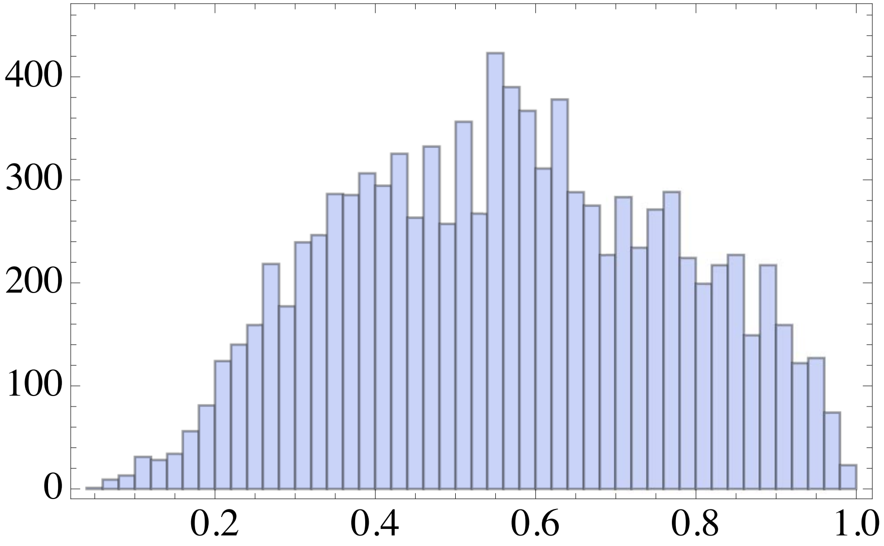
Posterior as a resampled prior using weighted bootstrap



```
lt = Table[likelihood[th[[k, 1]], th[[k, 2]]], {k, 1, ns}];
sum = Apply[Plus, lt];
w = lt / sum;

thr = Table[{0, 0}, {ns}];
ntot = 0;
While[ntot < ns,
 kn = RandomInteger[{1, ns}];
 If[RandomReal[] < w[[kn]], ntot++; thr[[ntot]] = th[[kn]]];
]
```

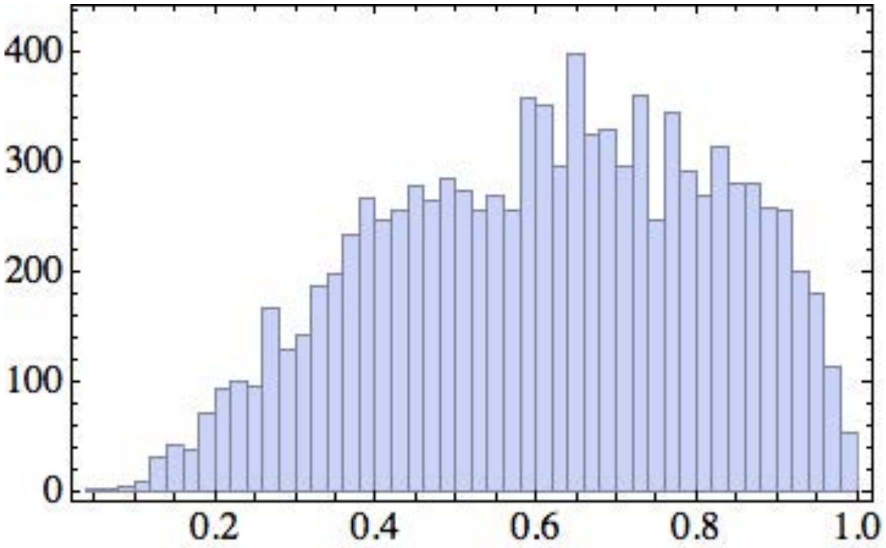


The resampled points are representative of the posterior distribution and can be used to evaluate any sample estimate



Marginalized distribution of θ_1

Sample mean: 0.564±0.002



Marginalized distribution of θ_2

Sample mean: 0.613±0.002

5. ***Very*** *short introduction to Markov chains*

Consider a system such that

- the system can occupy a finite or countably infinite set of states S_n;
- the system changes state randomly at discrete times $t = 1, 2, \ldots$;
- if the system is in state S_i, then the probability that the system goes into state S_j is

  $$p_{ij} = P[S(n+1) = S_j | S(n) = S_i] \qquad i, j = 1, 2, \ldots$$

  i.e., this probability depends only on the previous state, and is independent o all previous states (this is the *Markov property*);
- the *transition probabilities* p_{ij} do not depend on time n.

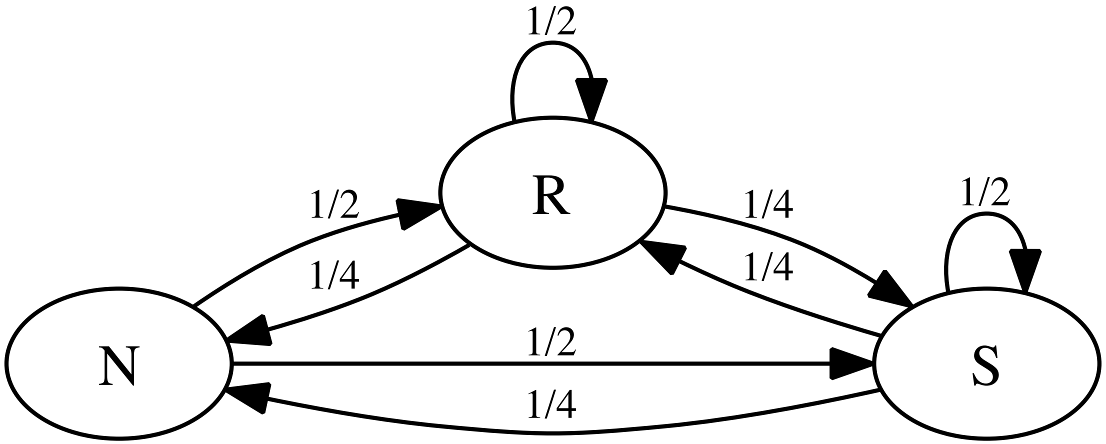
Such a system is a special type of discrete time stochastic process, which is called *Markov chain*.

Example:

in the Land of Oz they never have two nice days in a row, rather, after a sunny day it either rains or snows.

If they have a nice day, they are just as likely to have snow as rain the next day. If they have snow or rain, they have an even chance of having the same the next day. If there is change from snow or rain, only half of the time is this a change to a nice day. When we denote the three states with the symbols N (Nice), R (Rain), or S (Snow), the transition probabilities are:

$$\begin{array}{lll} p_{NN} = 0; & p_{NR} = 1/2; & p_{NS} = 1/2 \\ p_{RN} = 1/4; & p_{RR} = 1/2; & p_{RS} = 1/4 \\ p_{SN} = 1/4; & p_{SR} = 1/4; & p_{SS} = 1/2 \end{array}$$



(representation as a directed graph)

Matrix of transition probabilities (also called *transition kernel*)

$$\mathbf{P} = \begin{pmatrix} p_{NN} & p_{NR} & p_{NS} \\ p_{RN} & p_{RR} & p_{RS} \\ p_{SN} & p_{SR} & p_{SS} \end{pmatrix} = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}$$

This is a *row stochastic matrix*, where all rows are such that

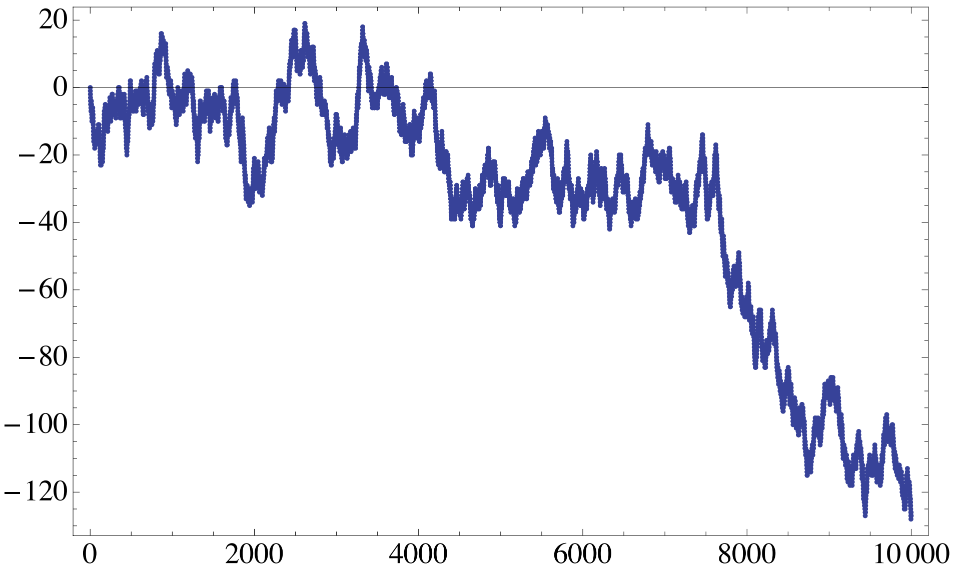
$$\textstyle\sum_j p_{ij} = 1$$

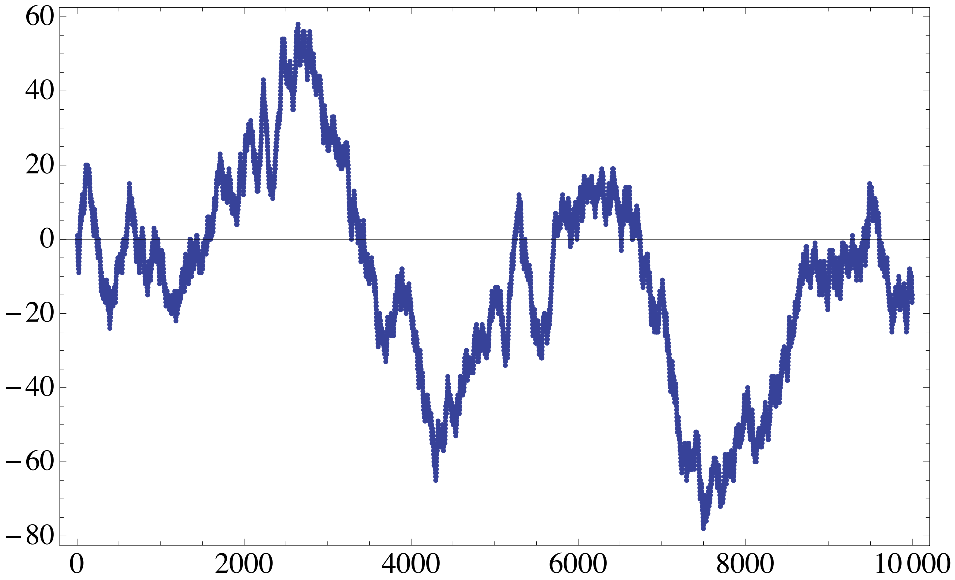
There are also column stochastic matrices, and doubly stochastic matrices that are necessarily square:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = \sum_{i=1}^{n} 1 = n$$

$$\Rightarrow \qquad m = n$$

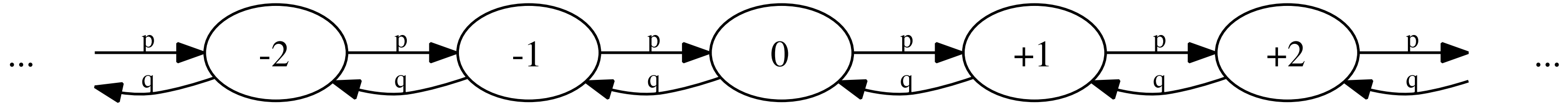
$$\sum_{j=1}^{m} \sum_{i=1}^{n} p_{ij} = \sum_{j=1}^{m} 1 = m$$





Discrete-time discrete-space random walks are an example of Markov chains with infinite states.

$$p_{i,i+1} = p; \quad p_{i,i-1} = q$$



Now let

$$\pi_i^{(n)} = P[S(n) = S_i]$$

be the probability that at time *n* the system is in state S_i , then:

$$\pi_j^{(n+1)} = \sum_i P[S(n+1) = S_j | S(n) = S_i] P[S(n) = S_i] = \sum_i p_{ij} \pi_i^{(n)}$$

When we define the vector $\boldsymbol{\pi}^{(n)} = \{\pi_j^{(n)}\}$and the matrix $\mathbf{P} = \{p_{ij}\}$ we see that the equation becomes

$$\boldsymbol{\pi}^{(n+1)} = \boldsymbol{\pi}^{(n)} \mathbf{P}$$

$$\Rightarrow \quad \boldsymbol{\pi}^{(n)} = \boldsymbol{\pi}^{(0)} \mathbf{P}^n$$

n-step transition kernel

For example, the transition kernels for the weather in the Land of Oz are

$$\mathbf{P} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

$$\mathbf{P}^2 = \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.375 & 0.4375 \end{pmatrix}$$

$$\mathbf{P}^5 = \begin{pmatrix} 0.199219 & 0.400391 & 0.400391 \\ 0.200195 & 0.400391 & 0.399414 \\ 0.200195 & 0.399414 & 0.400391 \end{pmatrix}$$

$$\mathbf{P}^{10} = \begin{pmatrix} 0.200001 & 0.4 & 0.4 \\ 0.2 & 0.400001 & 0.4 \\ 0.2 & 0.4 & 0.400001 \end{pmatrix}$$

the transition kernels seem to converge to a fixed matrix ...

$$\mathbf{P}^{20} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

$$\mathbf{P}^{100} = \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}$$

Notice that if the transition kernel converges to a fixed matrix where all rows are equal, then the distribution of states also converges to a fixed distribution which does not depend on the initial distribution:

$$\mathbf{P}^n \xrightarrow[n\to\infty]{} \mathbf{P}_\infty$$

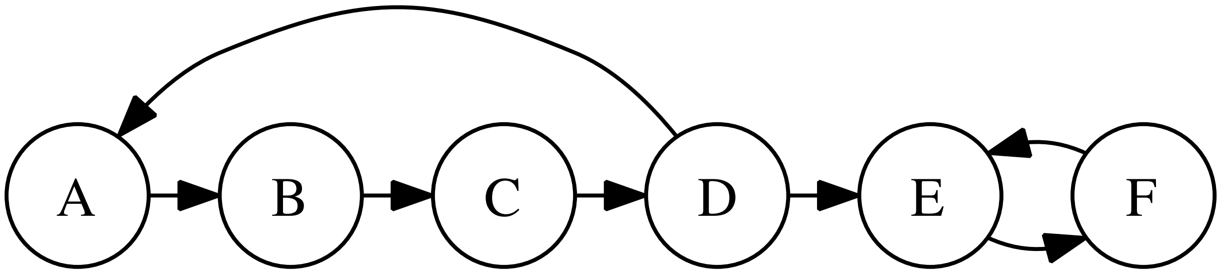
$$(\mathbf{P}_\infty)_{i,j} = f_j$$

all rows equal

$$\pi_j^{(\infty)} = \sum_i \pi_i^{(0)} (\mathbf{P}_\infty)_{i,j} = \sum_i \pi_i^{(0)} f_j = f_j$$

Persistent and transient states ...

Type of state	Definition of state (assuming, where applicable, that the state is initially occupied)
Periodic	Return to state possible only at times t, $2t$, $3t$, $\ldots$, where $t > 1$
Aperiodic	Not periodic
Recurrent/Persistent	Eventual return to state certain
Transient	Eventual return to state uncertain
Ephemeral	Is a state j such that $p_{ij} = 0$ for every i
Positive-recurrent	Recurrent/persistent, finite mean recurrence time
Null-recurrent	Recurrent, infinite mean recurrence time
Ergodic	Aperiodic, positive-recurrent



This graph represents the states and the transition probabilities of a finite Markov chain with 6 states.

The arrows correspond to nonzero transition probabilities. If the chain starts with any one of states A, B, C or D, it can loop around these four states until a transition D to E occurs, then the system is locked in the E-F loop.

States A, B, C, and D are transient, while states E and F are persistent (and periodic, with period 2). A Markov chain with just one class, such that all states communicate, is said to be irreducible. This Markov chain is not irreducible.

VERY INTERESTING MATH ON PERSISTENT STATES, HOWEVER WE DO NOT PURSUE IT FURTHER, WE DO NOT NEED IT NOW.

Limiting probabilities and stationary distributions

Here we prove that the convergence that we saw in the Land of Oz example is a general feature of Markov chains, under the assumption that the chain is irreducible, and that for some N we have

$$\min_{i,j} p_{ij}^{(N)} = \delta > 0$$

Now let

$$r_j^{(n)} = \min_i p_{ij}^{(n)}; \quad R_j^{(n)} = \max_i p_{ij}^{(n)}$$

be the min and max of the j-the column vector in the n-step transition matrix.

Recall the example:

$$
\begin{aligned}
\mathbf{P}^2 &= \begin{pmatrix} 0.25 & 0.375 & 0.375 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.375 & 0.4375 \end{pmatrix} \\
\mathbf{P}^5 &= \begin{pmatrix} 0.199219 & 0.400391 & 0.400391 \\ 0.200195 & 0.400391 & 0.399414 \\ 0.200195 & 0.399414 & 0.400391 \end{pmatrix} \\
\mathbf{P}^{10} &= \begin{pmatrix} 0.200001 & 0.4 & 0.4 \\ 0.2 & 0.400001 & 0.4 \\ 0.2 & 0.4 & 0.400001 \end{pmatrix} \\
\mathbf{P}^{20} &= \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix} \\
\mathbf{P}^{100} &= \begin{pmatrix} 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \\ 0.2 & 0.4 & 0.4 \end{pmatrix}
\end{aligned}
$$

we shall show that, in each column, the min and the max become closer and closer as n grows and bracket a value that is the asymptotic matrix element (the same for all rows in a given column)

Then we find

$$r_j^{(n+1)} = \min_i p_{ij}^{(n+1)} = \min_i \mathbf{P}_{ij}^{n+1} = \min_i (\mathbf{P}\mathbf{P}^n)_{ij} = \min_i \sum_k p_{ik} p_{kj}^{(n)}$$
$$\geq \min_i \sum_k p_{ik} r_j^{(n)} = r_j^{(n)}$$

and

$$R_j^{(n+1)} = \max_i p_{ij}^{(n+1)} = \max_i \mathbf{P}_{ij}^{n+1} = \max_i (\mathbf{P}\mathbf{P}^n)_{ij} = \max_i \sum_k p_{ik} p_{kj}^{(n)}$$
$$\leq \max_i \sum_k p_{ik} R_j^{(n)} = R_j^{(n)}$$

This means that, as *n* grows, the minimum and the maximum values in a column vector get closer and closer (the components of the column vector get closer and closer). *But do they converge to the same value ???*

We must consider the difference

$$R_j^{(n)} - r_j^{(n)} = \max_i p_{ij}^{(n)} - \min_k p_{kj}^{(n)} = \max_{i,k} \left[p_{ij}^{(n)} - p_{kj}^{(n)} \right]$$

Then, shifting the difference by N, we find

$$R_j^{(n+N)} - r_j^{(n+N)} = \max_{i,k} \left[p_{ij}^{(n+N)} - p_{kj}^{(n+N)} \right] = \max_{i,k} \left\{ \sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)} \right] p_{lj}^{(n)} \right\}$$

Next we split the difference enclosed in braces into sums of negative and positive contributions

$$\begin{aligned} \sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)} \right] p_{lj}^{(n)} &= \sum_l^{+} [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} + \sum_l^{-} [p_{il}^{(N)} - p_{kl}^{(N)}] p_{lj}^{(n)} \\ &\leq \sum_l^{+} [p_{il}^{(N)} - p_{kl}^{(N)}] R_j^{(n)} + \sum_l^{-} [p_{il}^{(N)} - p_{kl}^{(N)}] r_j^{(n)} \end{aligned}$$

Now consider the structure of the positive sum, it must contain at least one term where one subtracts the smallest element in the column, so that

$$\sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] = \sum_l^+ p_{il}^{(N)} - \sum_l^+ p_{kl}^{(N)} \leq \sum_l p_{il}^{(N)} - \delta = 1 - \delta$$

Similarly, for the negative sum we find

$$\sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] = \sum_l^- p_{il}^{(N)} - \sum_l^- p_{kl}^{(N)} \geq \delta - \sum_l p_{kl}^{(N)} = -(1 - \delta)$$

and therefore

$$\begin{aligned}\sum_l \left[p_{il}^{(N)} - p_{kl}^{(N)}\right] p_{lj}^{(n)} &\leq \sum_l^+ [p_{il}^{(N)} - p_{kl}^{(N)}] R_j^{(n)} + \sum_l^- [p_{il}^{(N)} - p_{kl}^{(N)}] r_j^{(n)} \\ &\leq (1-\delta) R_j^{(n)} - (1-\delta) r_j^{(n)} = (1-\delta)(R_j^{(n)} - r_j^{(n)})\end{aligned}$$

so that taking strides of N steps at a time, and recalling that $0 < 1 - \delta < 1$

$$\boxed{R_j^{(kN)} - r_j^{(kN)} < (1-\delta)^k \left[R_j^{(N)} - r_j^{(N)}\right] \xrightarrow[k \to \infty]{} 0}$$

Since

$$R_j^{(kN)} - r_j^{(kN)} < (1-\delta)^k \left[R_j^{(N)} - r_j^{(N)} \right] \xrightarrow[k\to\infty]{} 0$$

the matrix elements in the column converge to a single value p_j^* , i.e.,

$$p_{ij}^* = \lim_{n\to\infty} [\mathbf{P}^n]_{ij} = p_j^*$$

and

$$\pi_j^* = \sum_k \pi_k^{(0)} p_{kj}^* = \sum_k \pi_k^{(0)} p_j^* = p_j^*$$

This asymptotic distribution is stable, indeed from

$$\pi_j^{(n)} = \sum_k \pi_k^{(n-1)} p_{kj}$$

we find

$$[\pi^* \mathbf{P}]_j = \sum_k \pi_k^* p_{kj} = \sum_k p_k^* p_{kj} = \sum_k p_{ik}^* p_{kj} = p_{ij}^* = p_j^* = \pi_j^*$$

or, in matrix form

$$\pi^* = \pi^* \mathbf{P}$$

i.e., the asymptotic probability vector is the left eigenvector with eigenvalue 1 of the transition probability matrix. The distribution expressed by the probability vector π^* is called *invariant distribution* or *stationary distribution*.

Detailed balance

From the definition of conditional probabilities we find

$$\begin{aligned} P[S(n) = S_i \text{ and } S(n+1) = S_j] &= P[S(n) = S_i | S(n+1) = S_j] P[S(n+1) = S_j] \\ &= P[S(n+1) = S_j | S(n) = S_i] P[S(n) = S_i] \end{aligned}$$

therefore, when a Markov chain is time reversed we find

$$\begin{aligned} & P[S(n) = S_i | S(n+1) = S_j] \\ & \qquad = P[S(n+1) = S_j | S(n) = S_i] \frac{P[S(n) = S_i]}{P[S(n+1) = S_j]} \end{aligned}$$

i.e.,

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^{(n)}}{\pi_j^{(n+1)}}$$

which shows that the reversed chain is time-dependent.

However if states are distributed according to the invariant distribution, we have

$$P[S(n) = S_i | S(n+1) = S_j] = p_{ij} \frac{\pi_i^*}{\pi_j^*}$$

which means that the backward transition probabilities are again time-independent, and in particular they must coincide with the forward transition probabilities, i.e.,

$$p_{ji} \pi_j^* = p_{ij} \pi_i^*$$

a condition which is called *detailed balance*.

So *if* stationary distribution *then* detailed balance ... however the reverse also holds

$$\pi_j^{(n+1)} = \sum_i \pi_i^{(n)} p_{ij} = \sum_i \pi_j^{(n)} p_{ji} = \pi_j^{(n)} \sum_i p_{ji} = \pi_j^{(n)}$$

Physical aside: continuous-time Markov processes

The time-dependence of the reversed chain is a manifestation of the dissipative character of the chain. Another important related result is the validity of the H-theorem for Markov processes.

In the case of continuous-time processes we can write

$$\begin{aligned} & P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = \\ & \qquad = P\left(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) P\left(S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) \end{aligned}$$

Memoryless processes

$$P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = P\left(S_{i_k}, t_k\right)$$

Markov processes

$$P\left(S_{i_k}, t_k; S_{i_{k-1}}, t_{k-1}; \ldots; S_{i_0}, t_0\right) = P\left(S_{i_k}, t_k | S_{i_{k-1}}, t_{k-1}\right) P\left(S_{i_{k-1}}, t_{k-1}\right)$$

For Markov processes the following equation also holds

$$P(S_n, t + \Delta t) = P(S_n, t) + \\ + \sum_j \left[P\left(S_n, t + \Delta t | S_j, t\right) P(S_j, t) - P\left(S_j, t + \Delta t | S_n, t\right) P(S_n, t) \right]$$

(*master equation*).

When we assume that the transition probabilities are time-invariant and we define the transition rates T

$$P\left(S_n, t + \Delta t | S_j, t\right) = T_{n,j} \Delta t$$

we find the differential form of the master equation

$$\frac{d}{dt} P(S_n, t) = \sum_j \left[T_{n,j} P(S_j, t) - T_{j,n} P(S_n, t) \right]$$

Using the previous notation for the probability distribution on states, we can rewrite the master equation as follows

$$\frac{d\pi_n}{dt} = \sum_j \left[T_{n,j}\pi_j(t) - T_{j,n}\pi_n(t)\right]$$

Next, we assume that transition probabilities are "reversible"

$$T_{n,j} = T_{j,n}$$

so that

$$\frac{d\pi_n}{dt} = \sum_j T_{n,j}\left[\pi_j(t) - \pi_n(t)\right]$$

and therefore, at equilibrium

$$\sum_j T_{n,j}\left(\pi_j^* - \pi_n^*\right) = 0 \quad \Rightarrow \quad \pi_j^* = \pi_n^*$$

all states are equally likely at equilibrium

Now consider the following sum

$$H = \sum_n \pi_n \ln \pi_n$$

Using the master equation we find a differential equation for H

$$\frac{dH}{dt} = \sum_n \frac{d}{dt}(\pi_n \ln \pi_n) = \sum_n \frac{d\pi_n}{dt}(\ln \pi_n + 1)$$

$$= \sum_{n,j} T_{n,j}\,(\pi_j - \pi_n)\,(\ln \pi_n + 1)$$

Exchanging indexes ...

$$\frac{dH}{dt} = \sum_{n,j} T_{n,j}\,(\pi_n - \pi_j)\,(\ln \pi_j + 1)$$

Adding the two differential equations we find

$$\frac{dH}{dt} = \frac{1}{2}\sum_{n,j} T_{n,j}\left(\pi_n - \pi_j\right)\left(\ln \pi_j - \ln \pi_n\right)$$

Since

$$\left(\pi_n - \pi_j\right)\left(\ln \pi_j - \ln \pi_n\right) \leq 0$$

we find

$$\boxed{\frac{dH}{dt} \leq 0}$$

Boltzmann's H-theorem

The derivative vanishes at equilibrium, and we find that it is a stable point for *H*. Since *H* is essentially the negative of Gibbs' entropy, the theorem states that the entropy of a Markov chain increases up to a maximum which is reached at equilibrium.