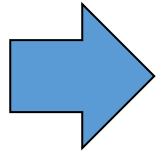


# Introduction to Bayesian Statistics - 5

*Edoardo Milotti*

Università di Trieste and INFN-Sezione di Trieste

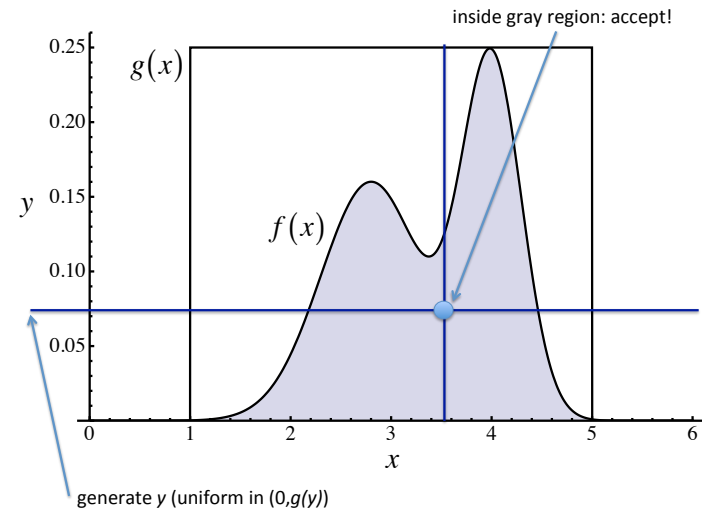
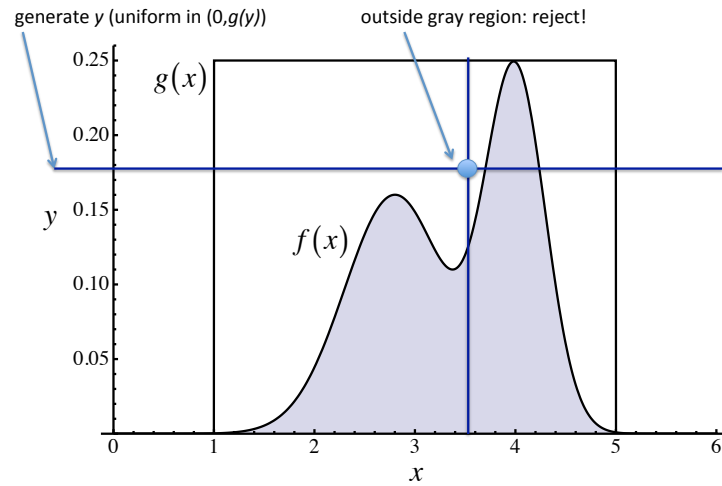
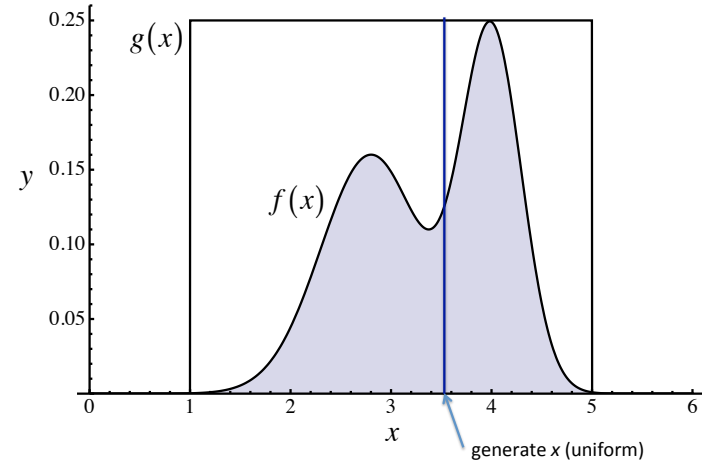
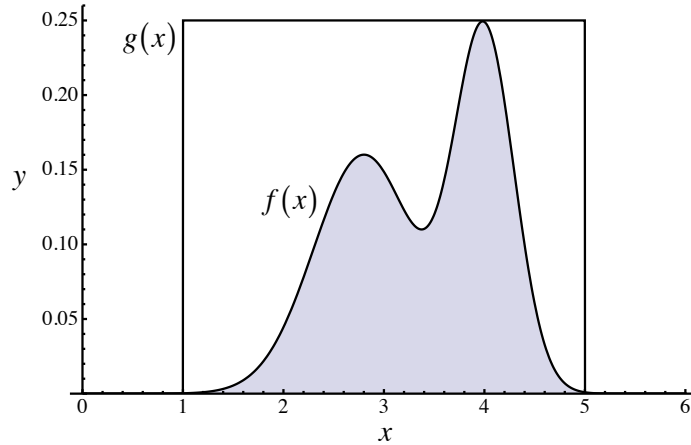
Bayesian estimates often require the evaluation of complex integrals. Usually these integrals can only be evaluated with numerical methods.



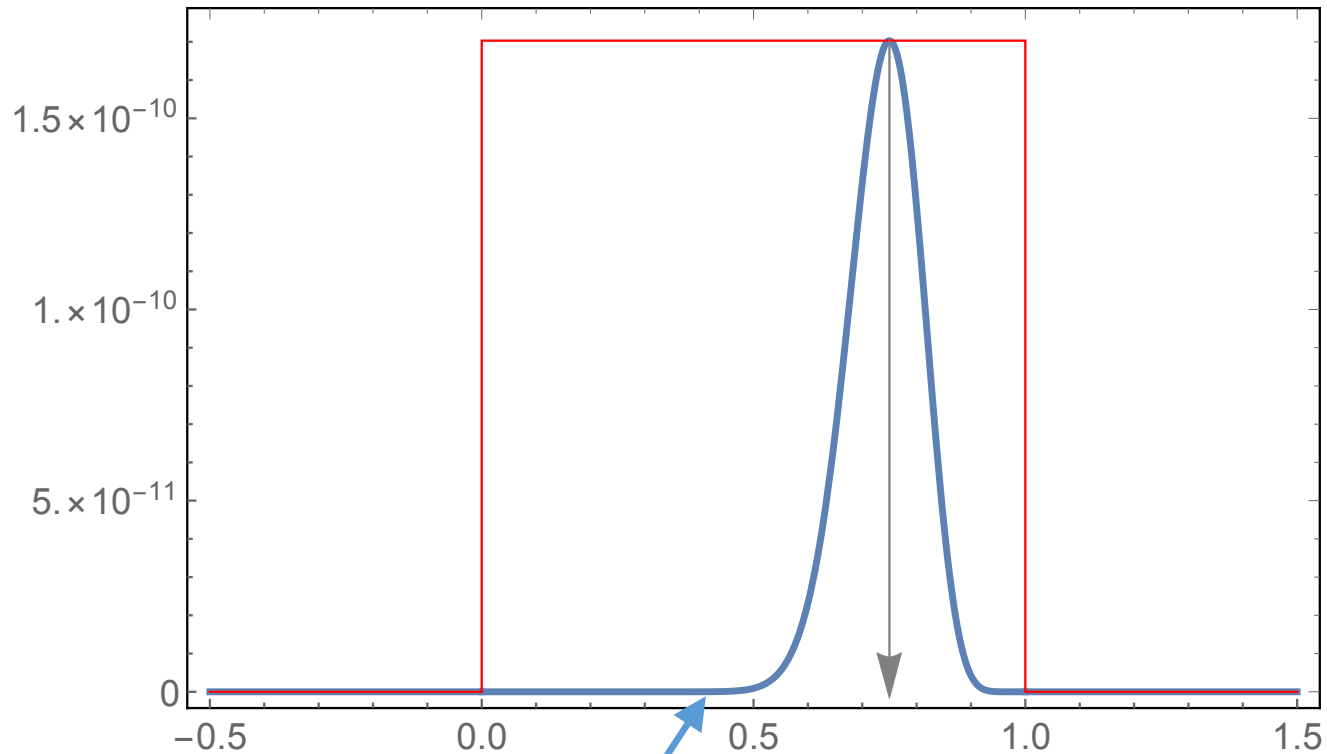
enter the Monte Carlo methods!

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



# Example: generation of beta-distributed random numbers



$$p(x) = \frac{x^a(1-x)^b}{B(a+1, b+1)}$$

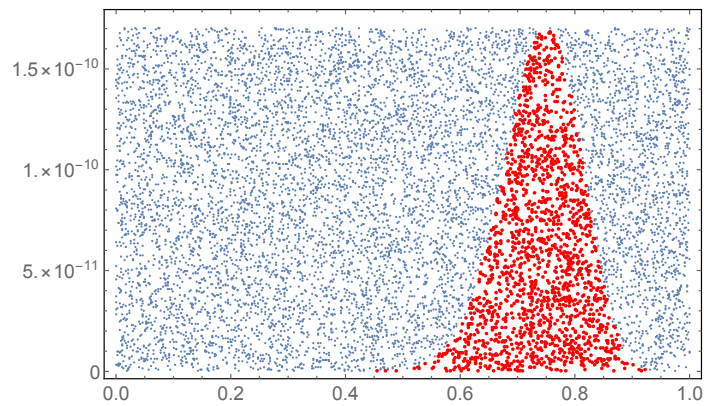
normalized distribution

$$p_0(x) = x^a(1-x)^b$$

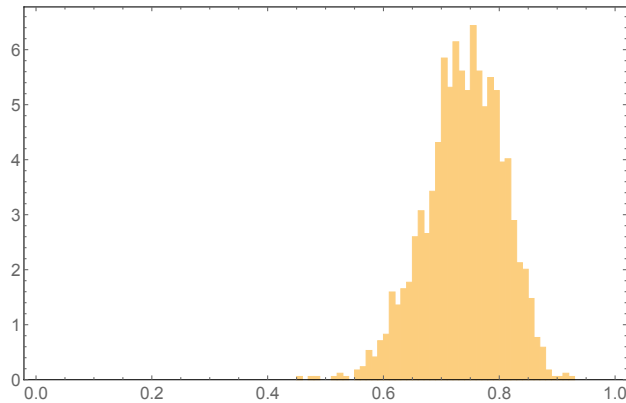
unnormalized distribution

$$x_{\max} = \frac{a}{a+b}$$

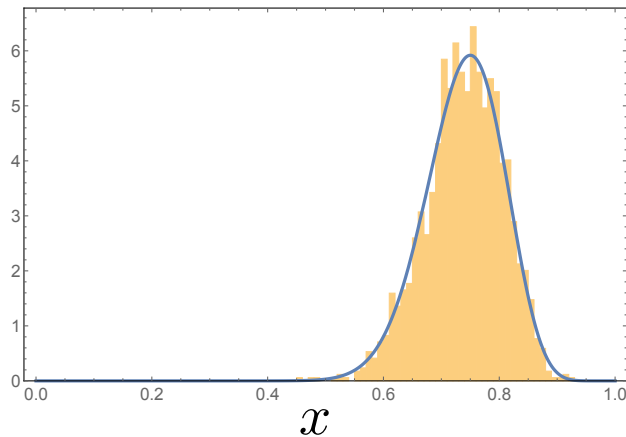
modal value



generated pairs  
(red = accepted pairs)



normalized histogram of the  
accepted  $x$ 's

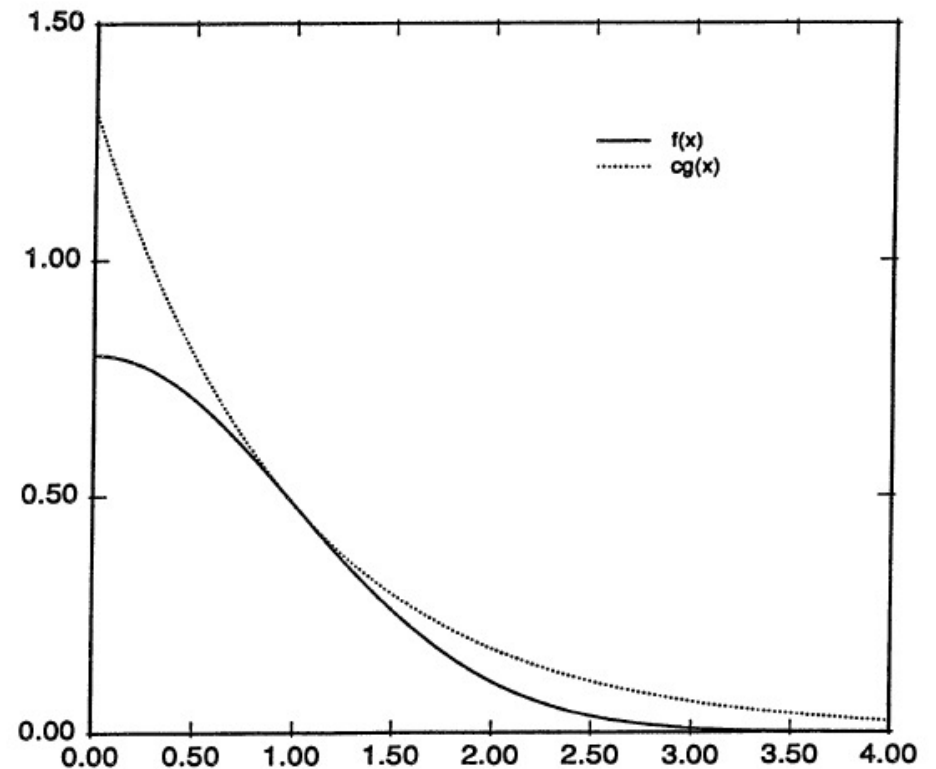


comparison with the plot of the  
normalized beta distribution

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

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

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



## Definition of contact point (to maximize efficiency)

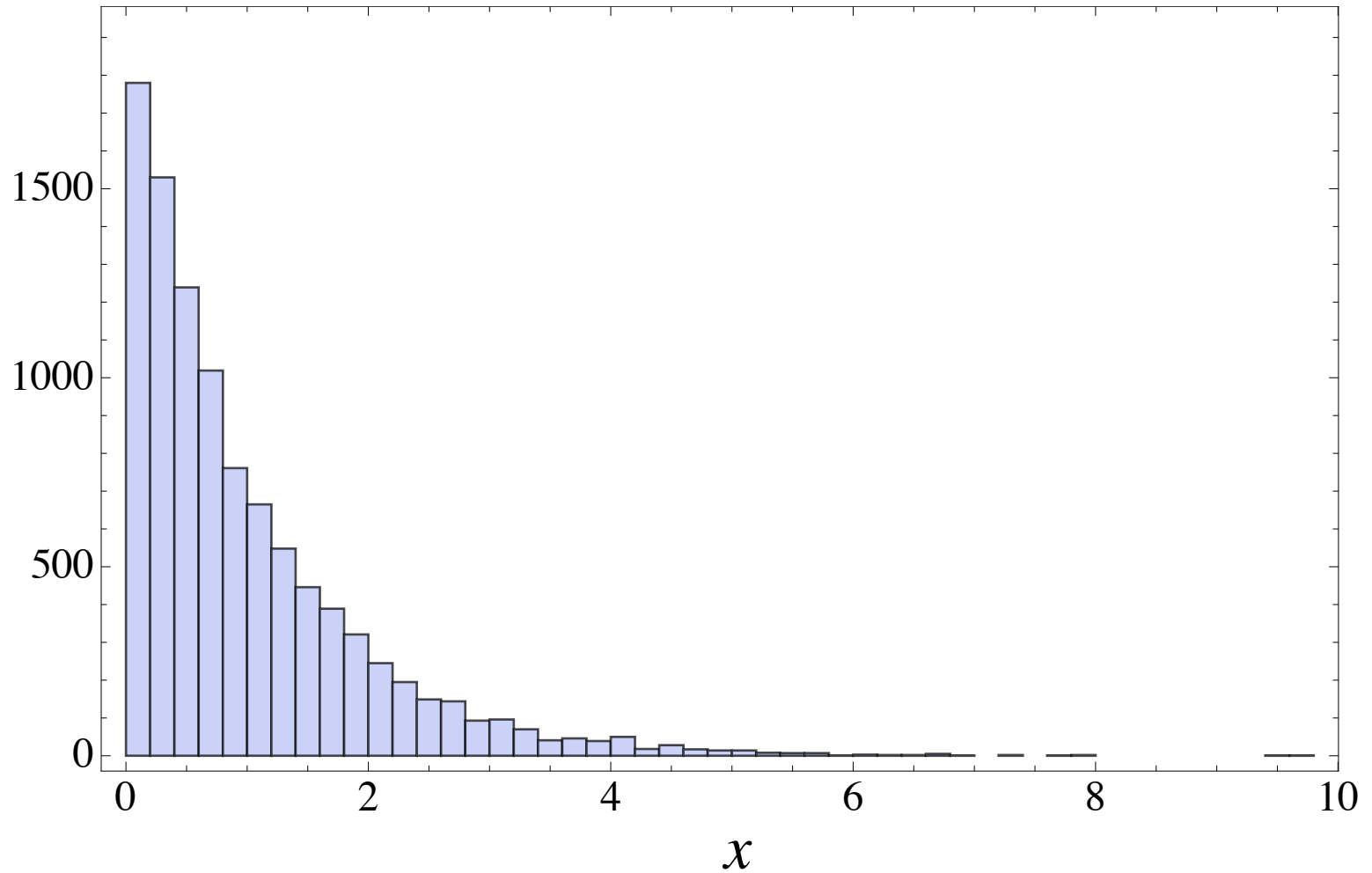
$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \quad x \geq 0$$

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

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

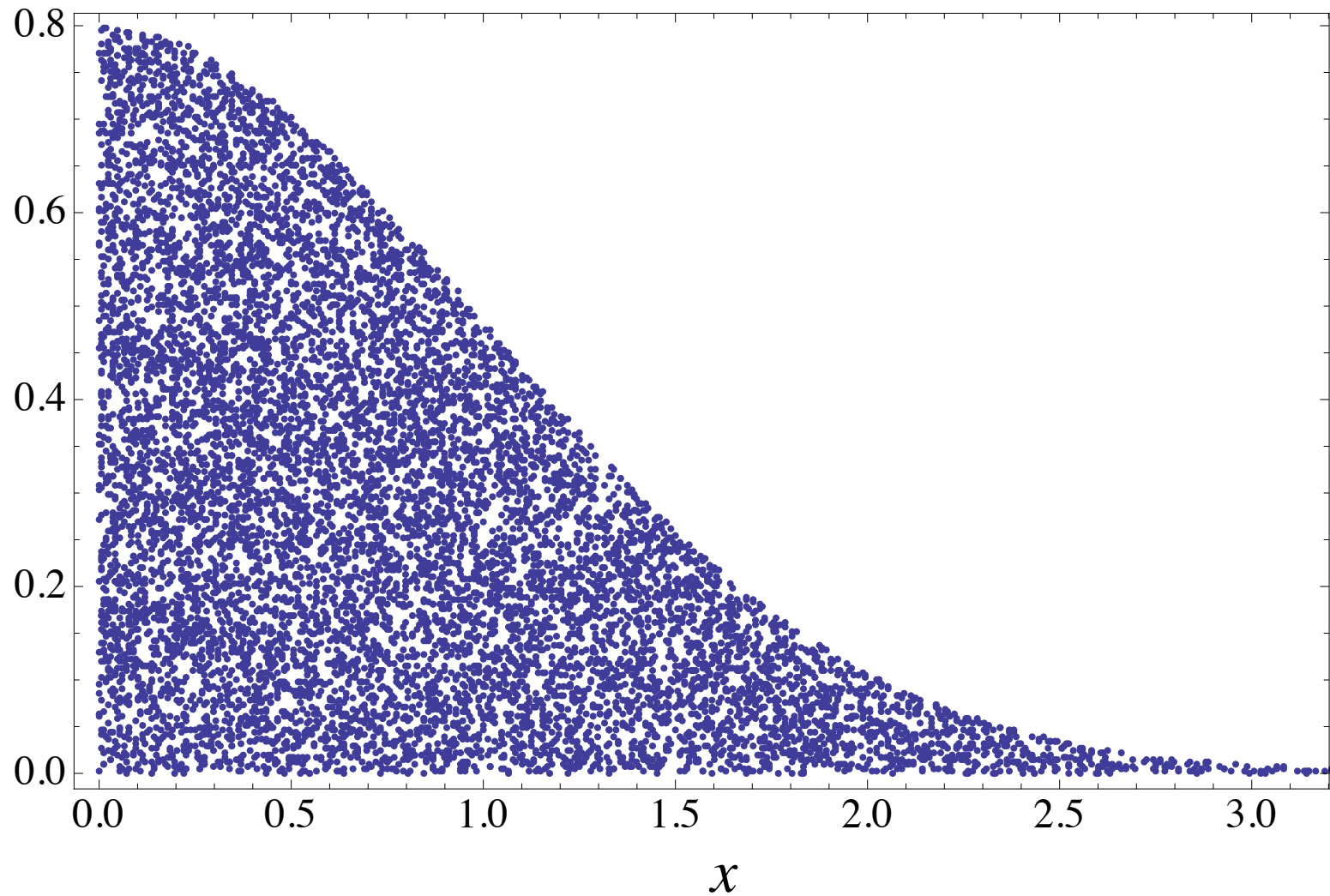
$$\Rightarrow x = 1; \quad c = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2} + x\right) \approx 1.31549$$

# Exponentially distributed values

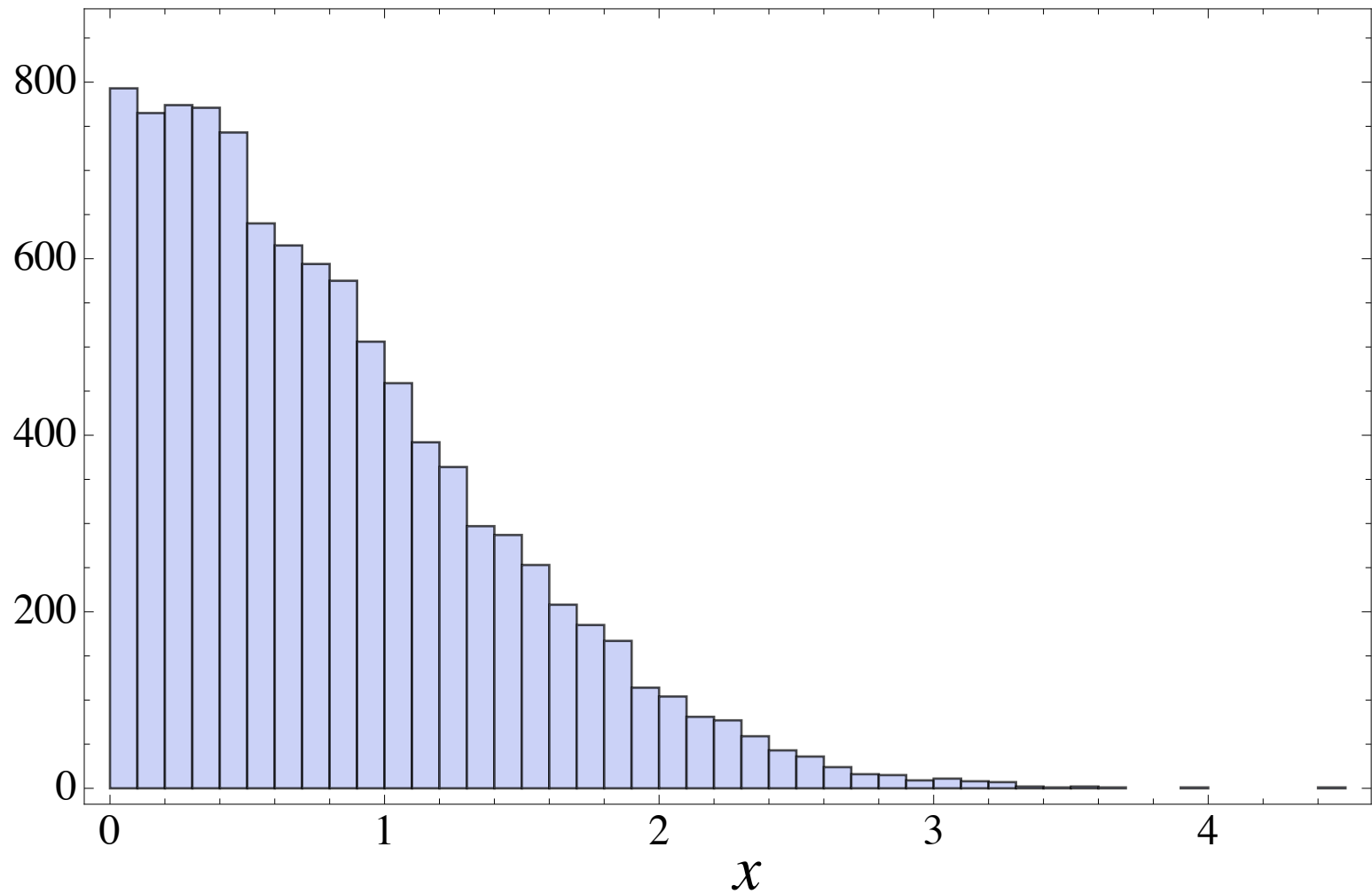




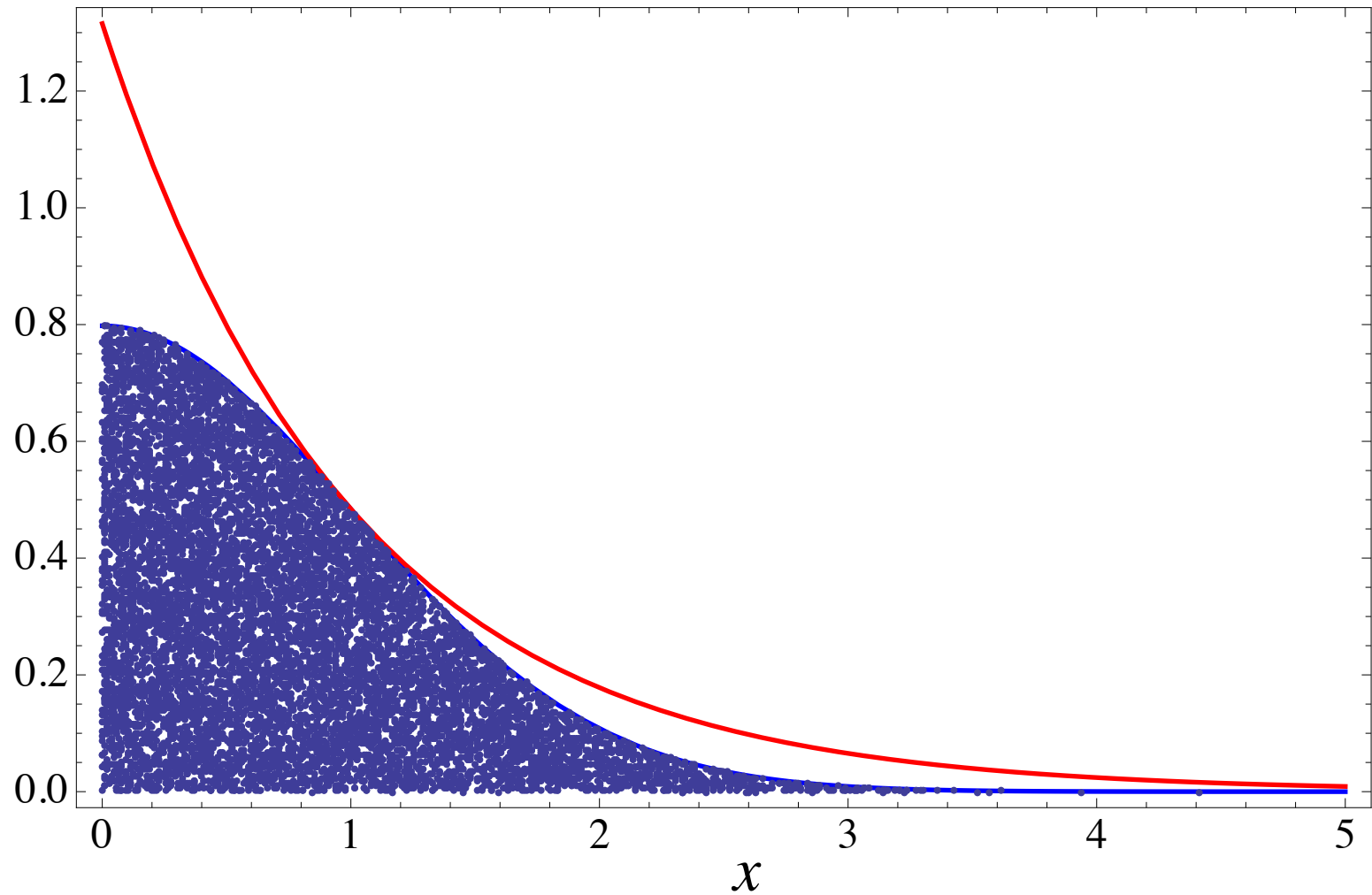
# A/R accepted values (10000 accepted sample pairs)

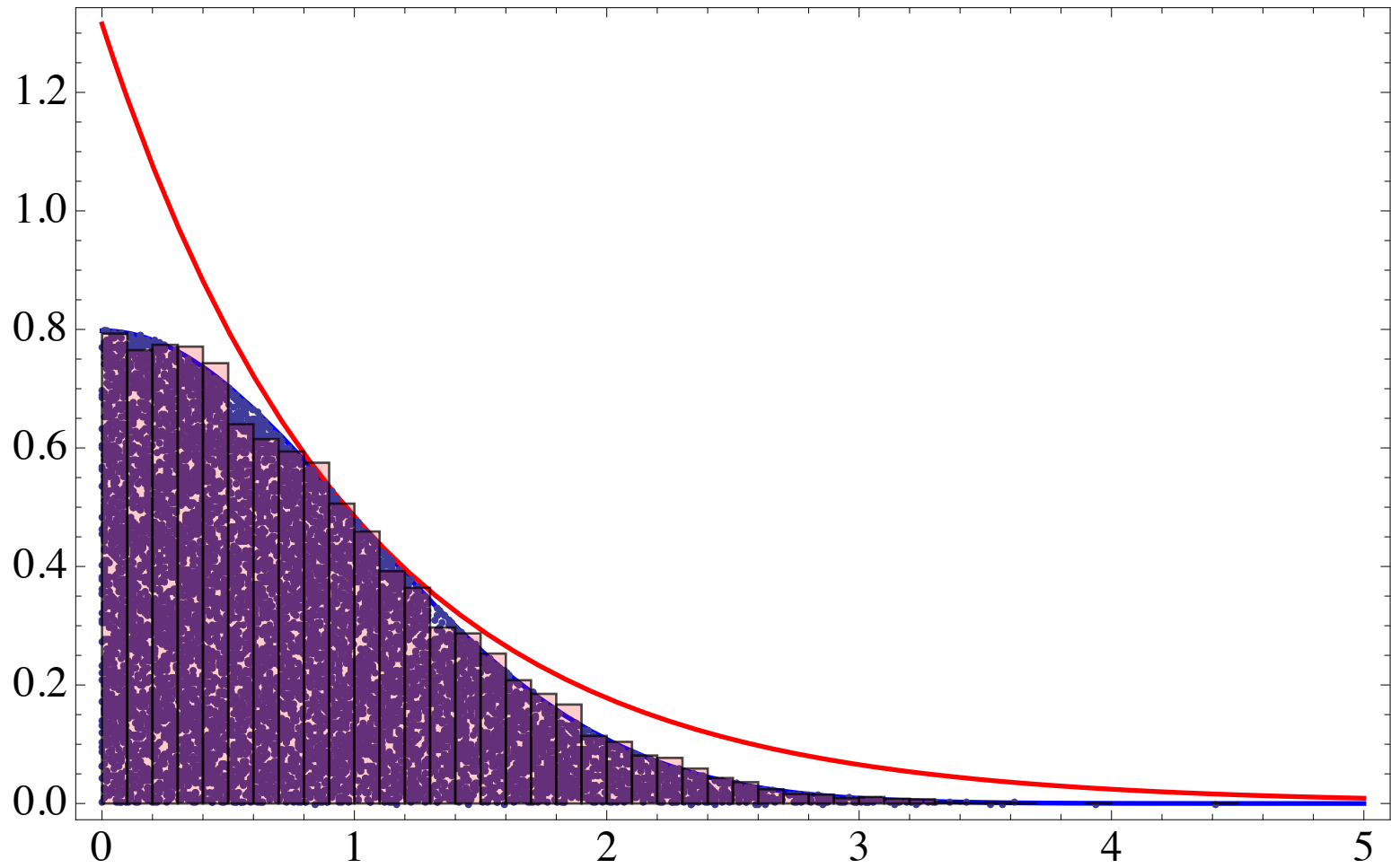


# Histogram of accepted $x$ values



# Comparison with the original distributions





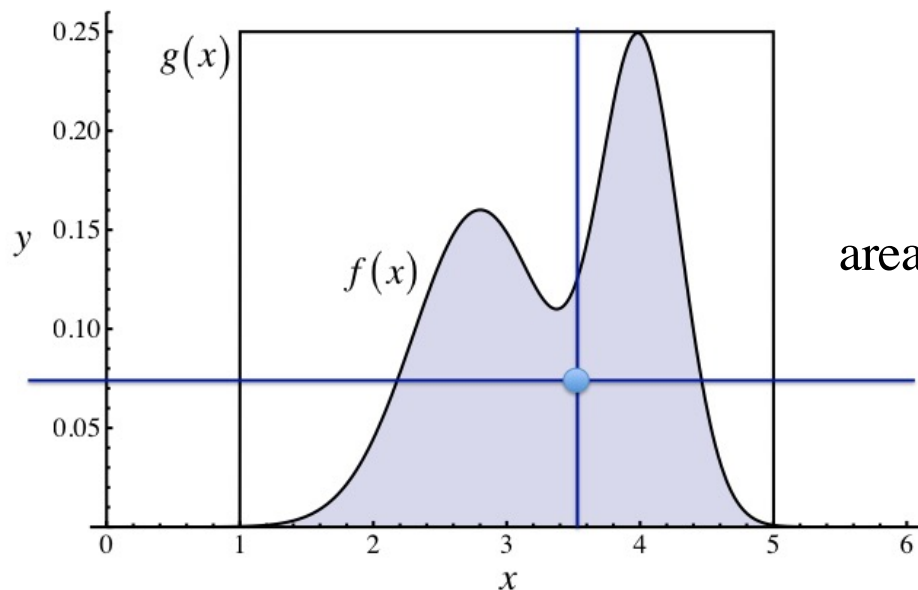
## Short summary:

1. we create a data set by randomly sampling from the exponential distribution
2. we use the acceptance-rejection algorithm to resample the data set with the target distribution (the half-Gaussian)

*This is a sampling – resampling technique (see later ... )*

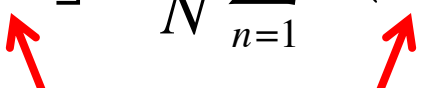
Now notice that in this method we generate pairs of real numbers that are uniformly distributed between  $f(x)$  and the  $x$ -axis, therefore we can use these pairs to estimate the total area under the curve

(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

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


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

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

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

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 ...

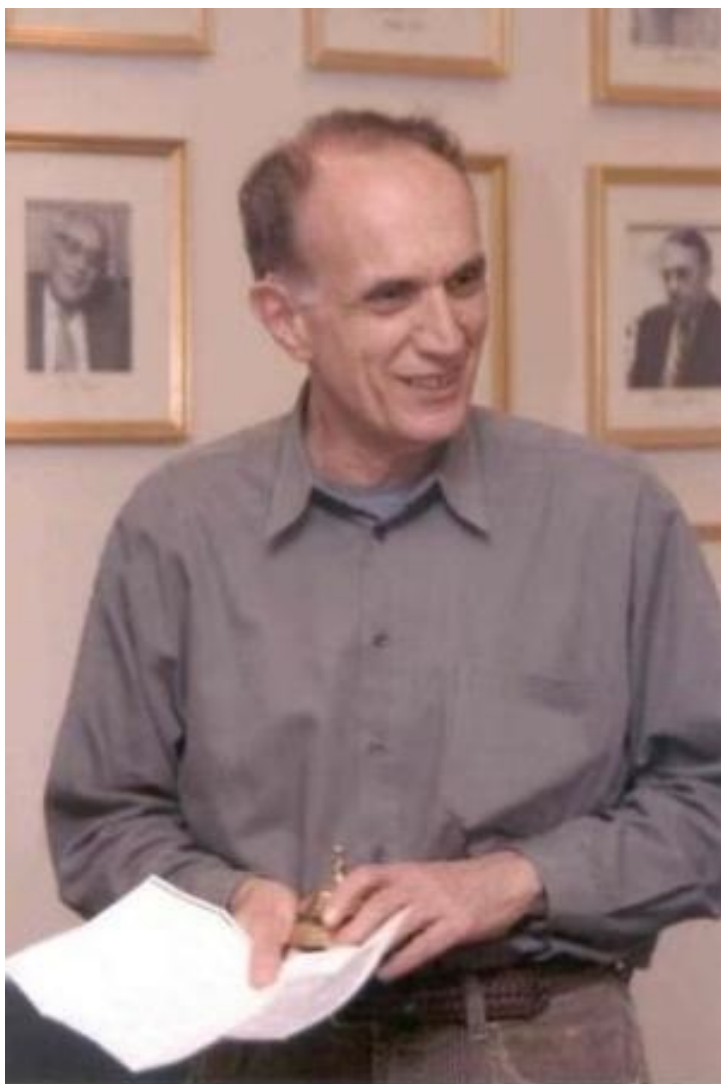
$$\int_a^b h(x) dx = \int_a^b f(x) p(x) dx = \int_a^b \left[ f(x) \frac{p(x)}{q(x)} \right] q(x) dx$$
$$= E_q \left[ f(x) \frac{p(x)}{q(x)} \right] \approx \frac{1}{N} \sum_{n=1}^N f(x_n) \frac{p(x_n)}{q(x_n)}$$

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

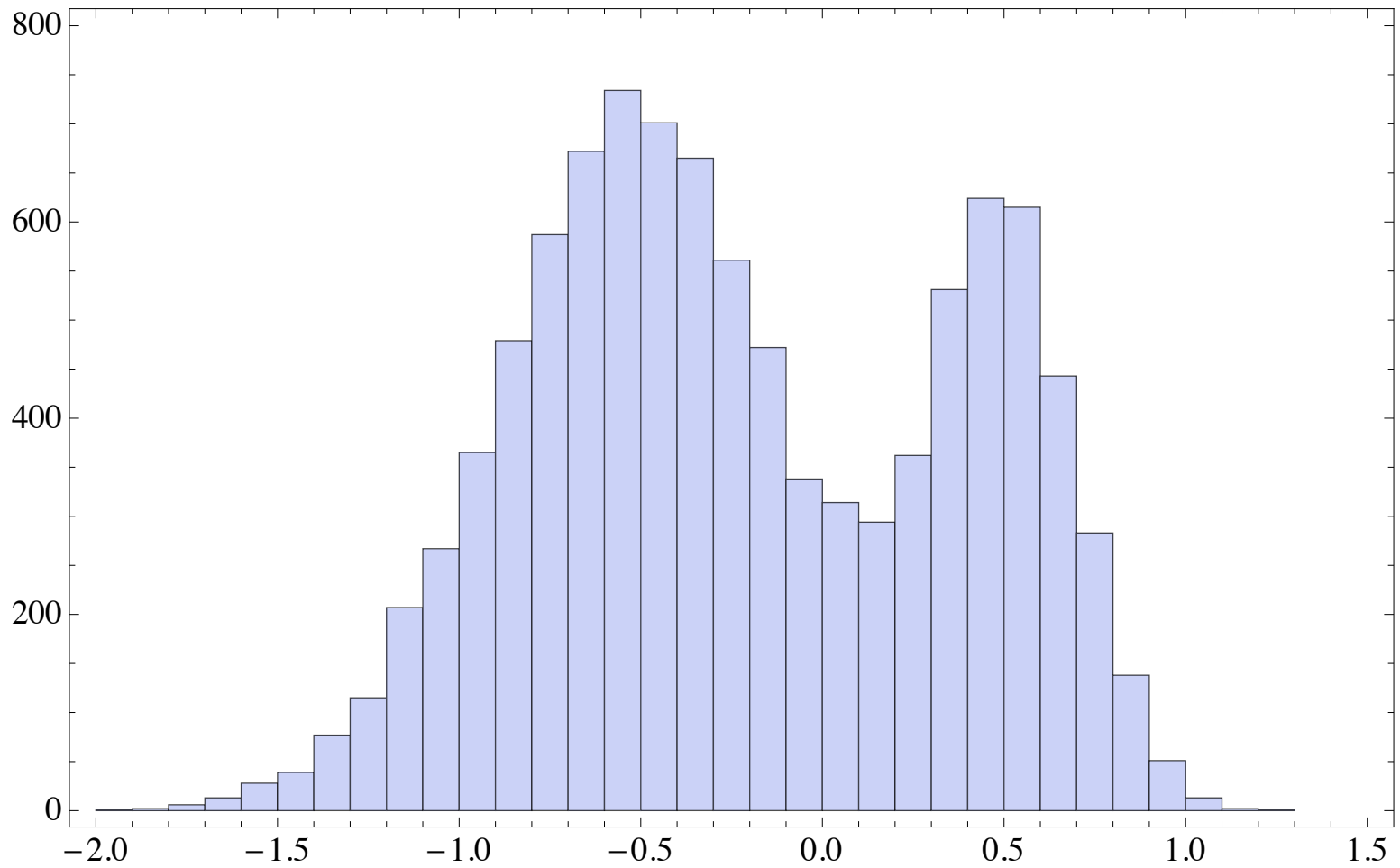


### 3. Bootstrap (B. Efron, 1977)

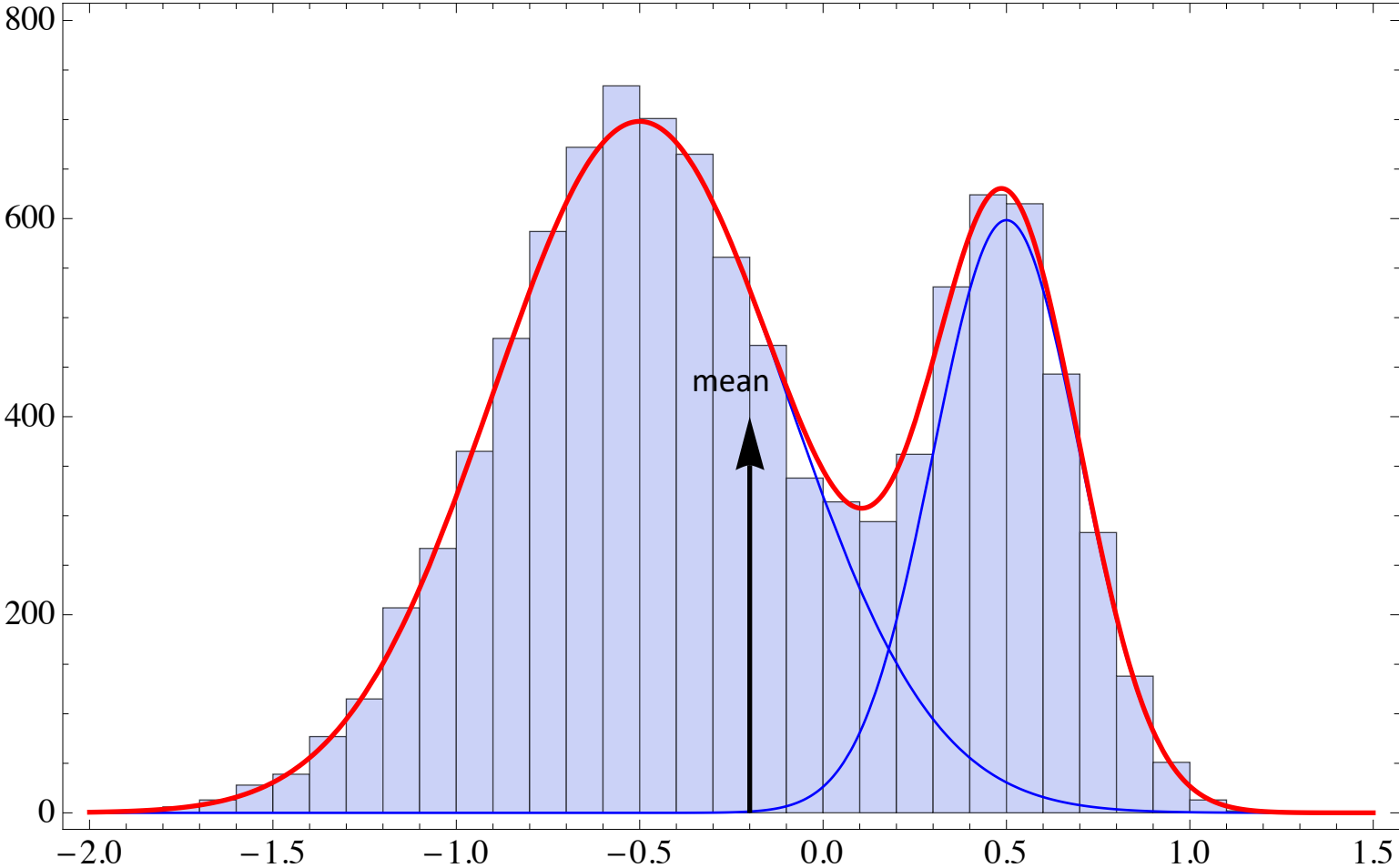


The bootstrap method is a resampling technique that helps calculate many statistical estimators

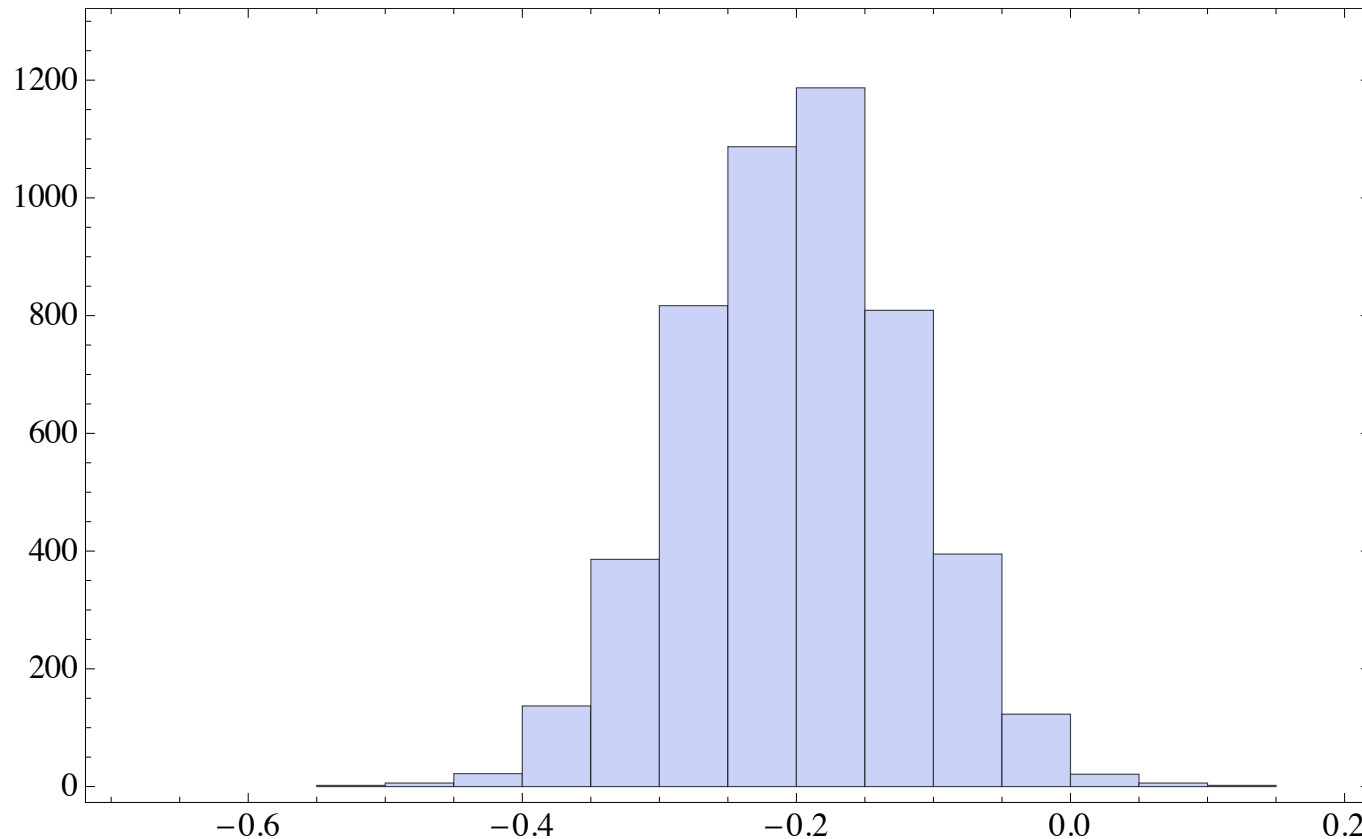
consider the distribution of a set of measurements



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

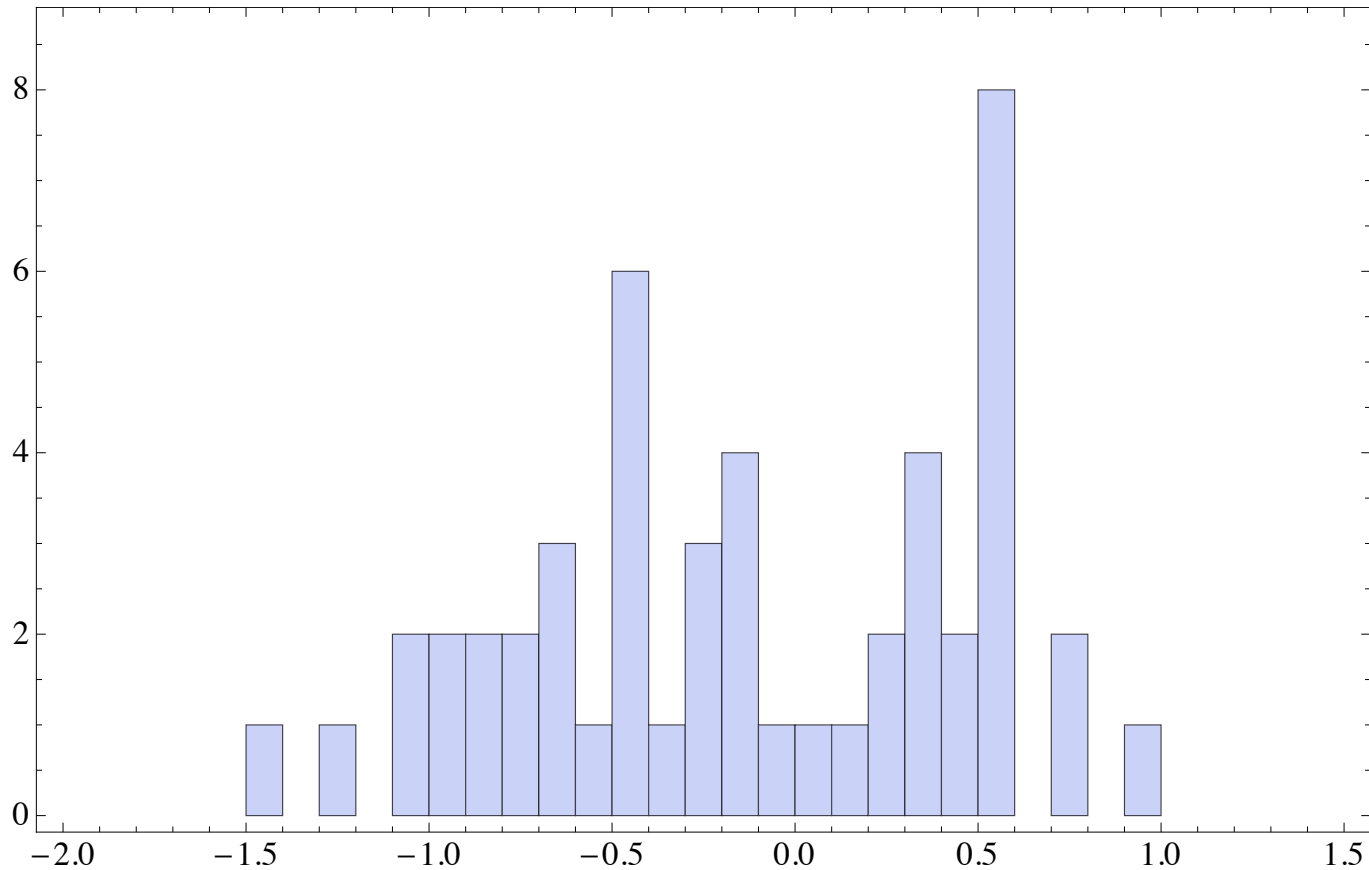


distribution of mean value obtained from 5000 sets of data  
(sample size = 50)



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

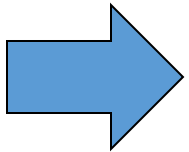
example: single dataset (same size as before, 50 measurements)



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

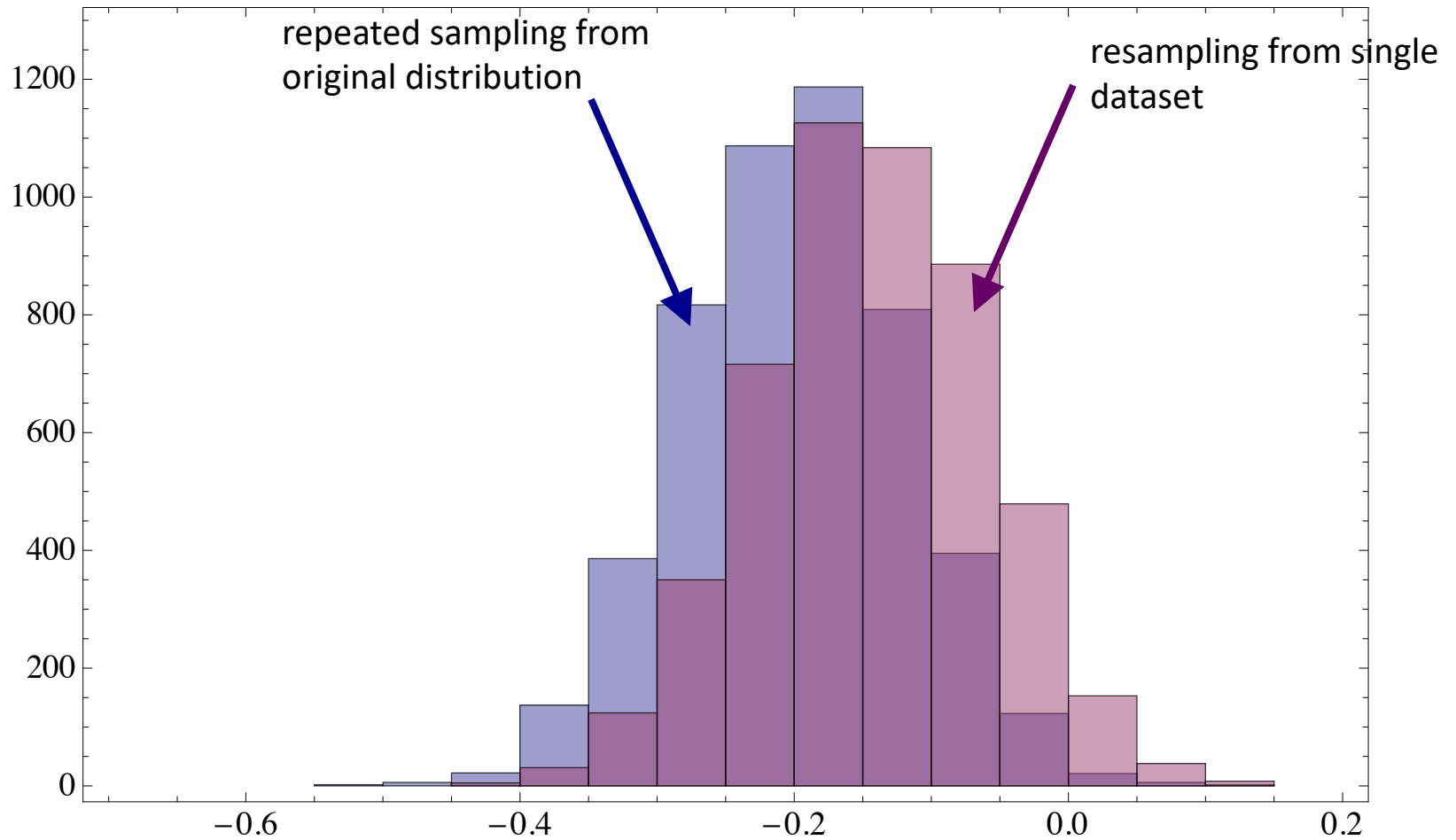
## Bootstrap recipe:

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):  $-0.200222 \pm 0.0813632$

mean from resampling dataset (size = 50):  $-0.142699 \pm 0.0838678$

counts of CD4 lymphocytes

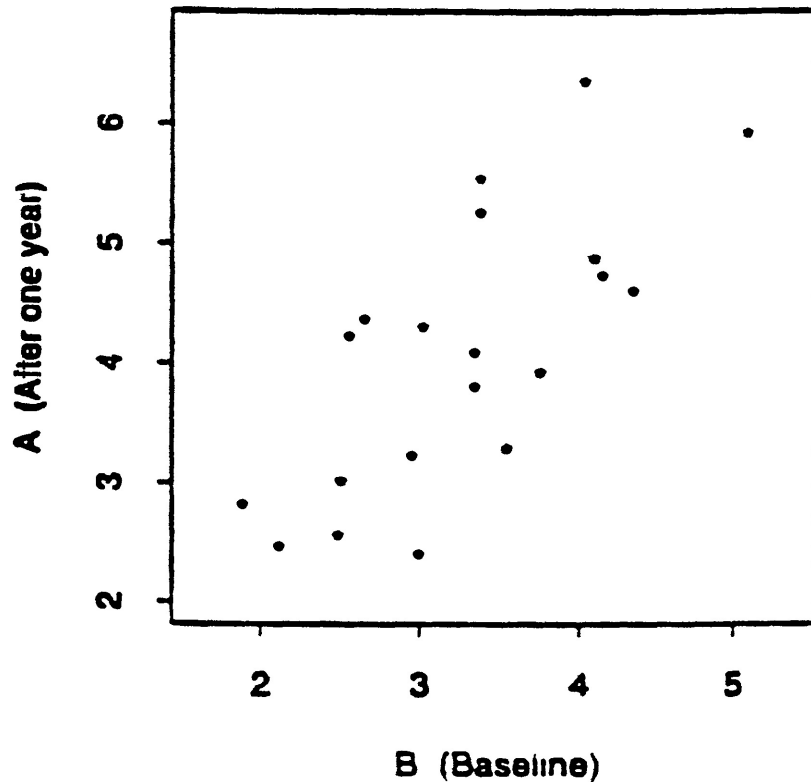


FIG. 1. *The cd4 data; cd4 counts in hundreds for 20 subjects, at baseline and after one year of treatment with an experimental anti-viral drug; numerical values appear in Table 1.*

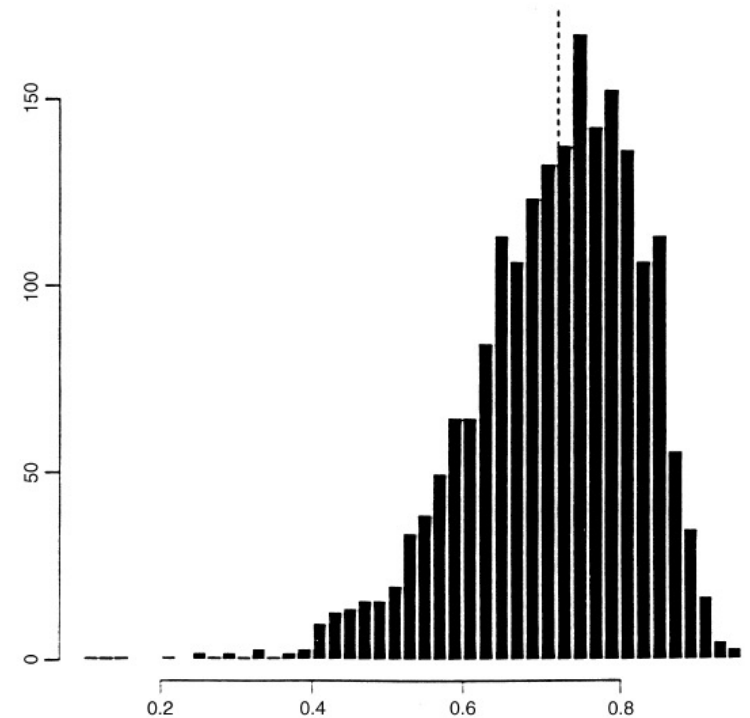


FIG. 3. *Histogram of 2,000 bootstrap correlation coefficients; bivariate normal sampling model.*

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.,*

$$p(\theta|x) = \frac{l(\theta; x)p(\theta)}{\int l(\theta; x)p(\theta) d\theta} \leftarrow \text{normalization (evidence)}$$

$$p(\phi|x) = \int p(\phi, \psi|x) d\psi. \leftarrow \text{marginalization}$$

$$E[m(\theta)|x] = \int m(\theta)p(\theta|x) d\theta \leftarrow \text{averages (statistical estimators)}$$

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 (importance sampling-like scheme)

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

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

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

(resampling))

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

(sampling)

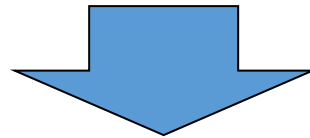
Example (McCullagh & Nelder): take two sets of binomially distributed independent random variables  $X_{i1}$  and  $X_{i2}$  ( $i=1,2,3$ )

$$X_{i1} = \text{Binomial}(n_{i1}, \theta_1)$$

$$X_{i2} = \text{Binomial}(n_{i2}, \theta_2)$$

The observed random variables are the sums

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



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

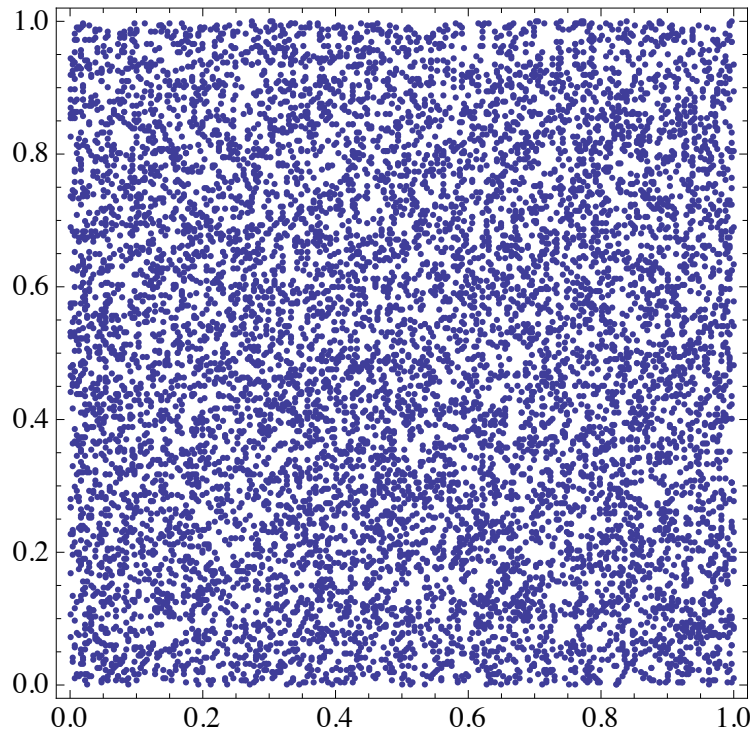
$$\max(0, y_i - n_{i2}) \leq j_i \leq \min(n_{i1}, y_i)$$

# Sample data

	<b>1</b>	<b>2</b>	<b>3</b>
$n_{i1}$	5	6	4
$n_{i2}$	5	4	6
$y_i$	7	5	6

## Example of implementation in *Mathematica*

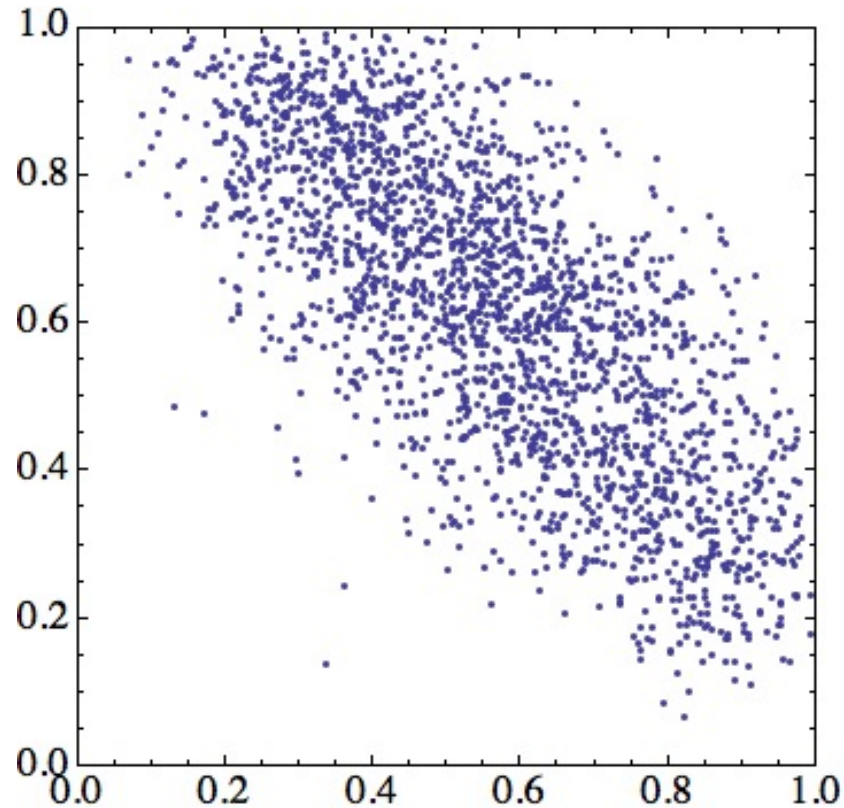
```
n1 = {5, 6, 4};  
n2 = {5, 4, 6};  
yi = {7, 5, 6};  
  
Clear[likelihood];  
likelihood[th1_, th2_] :=  
  Product[Sum[Binomial[n1[[i]], j] * Binomial[n2[[i]], yi[[i]] - j] * th1^j * (1 - th1)^(n1[[i]] - j) *  
    th2^(yi[[i]] - j) * (1 - th2)^(n2[[i]] - yi[[i]] + j), {j, Max[0, yi[[i]] - n2[[i]], 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

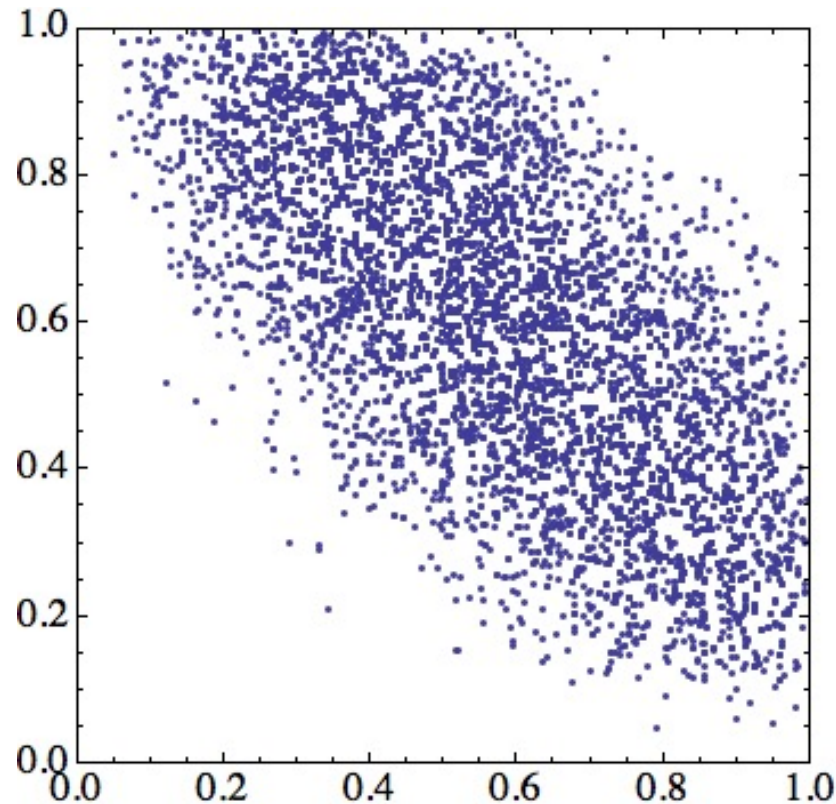
```
lt = Table[likelihood[th[[k, 1]], th[[k, 2]]], {k, 1, ns}];  
norm = Max[lt];  
w = lt / norm;  
  
thr = {}; ntot = 0;  
For[kn = 1, kn ≤ ns,  
  If[w[[kn]] > RandomReal[], ntot++; AppendTo[thr, th[[kn]]];  
  kn++]
```



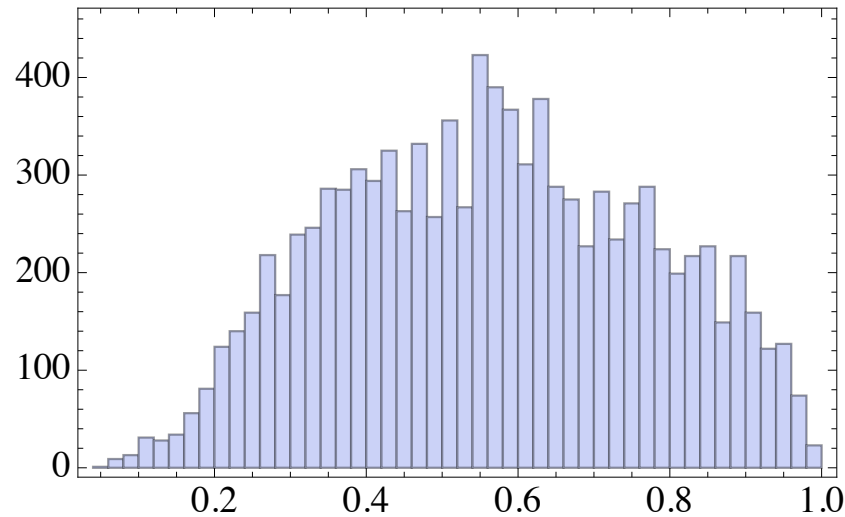


## Posterior as a resampled prior using weighted bootstrap

```
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]]];  
]
```

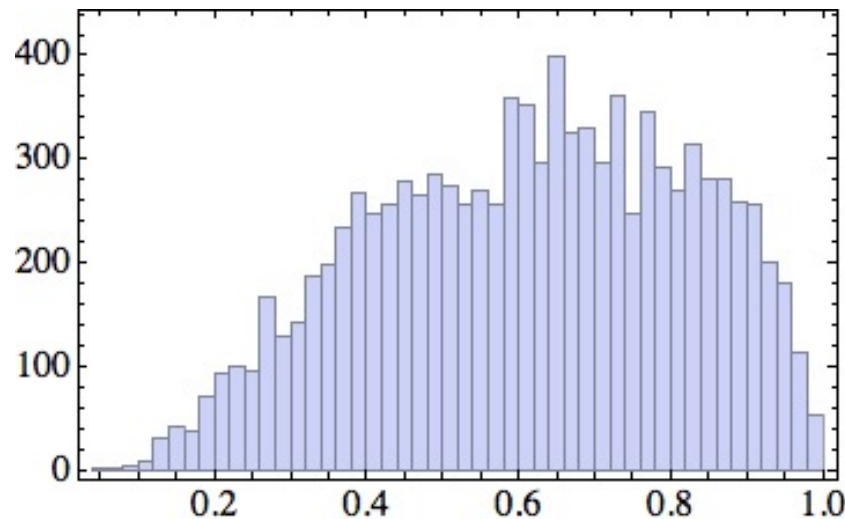


The resampled points are representative of the posterior distribution and can be used to evaluate any sample estimate



Marginalized distribution of  $\theta_1$

Sample mean:  $0.564 \pm 0.002$



Marginalized distribution of  $\theta_2$

Sample mean:  $0.613 \pm 0.002$

... these calculational methodologies have also had an impact on theory. By freeing statisticians from dealing with complicated calculations, the statistical aspects of a problem can become the main focus.

Casella & George, in their description of the Gibbs sampler.  
Am. Stat. **46** (1992) 167