Introduction to Bayesian Methods - 6

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

A short overview of model selection methods – ctd.

Akaike Information Criterion (AIC).

This was derived by Hirotugu Akaike in 1974, and takes the form

$$\mathrm{AIC} = -2 \ln \mathcal{L}_{\max} + 2k$$

where k is the number of parameters in the model. The subscript "max" indicates that one should find the parameter values yielding the highest possible likelihood within the model. ***This second term acts as a kind of "Occam factor"; initially, as parameters are added, the fit to data improves rapidly until a reasonable fit is achieved, but further parameters then add little and the penalty term 2k takes over.*** *The generic shape of the AIC as a function of number of parameters is a rapid fall, a minimum, and then a rise. The preferred model sits at the minimum.*

The AIC was derived from information-theoretic considerations, specifically an approximate minimization of the Kullback–Leibler information entropy which measures the distance between two probability distributions.

(Liddle & al., 2006)

Outline of Akaike's derivation

1. max log-likelihood ratio between conjectured model (*k*-dimensional parameter vector) and true model (*l*-dimensional parameter vector)

$$\ln \frac{f(x|\hat{\theta}^{(k)})}{f(x|\theta)}$$

2. this depends on the dataset, which is distributed according to the true model; to get rid of the fluctuations, we average the max log-likelihood ratio over the true distribution

$$\mathbb{E}\left[\ln \frac{f(x|\hat{\theta}^{(k)})}{f(x|\theta)}\right] = \int_{\Theta} f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)})}{f(x|\theta)} dx = -D_{\mathrm{KL}}\left(f(x|\theta)||f(x|\hat{\theta}^{(k)})\right)$$

3. here we remark that:
   - this is purely theoretical, since we do not know the true pdf
   - the r.h.s. expression is the negative of the Kullback-Leibler divergence between the conjectured and the true pdf
   - the l.h.s. is maximum when the KL divergence is at a minimum
   - the r.h.s. expression can be written as

$$\int_{\Theta} f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)})}{f(x|\theta)} dx = \int_{\Theta} f(x|\theta) \ln f(x|\hat{\theta}^{(k)}) - \int_{\Theta} f(x|\theta) \ln f(x|\theta)$$

Outline of Akaike's derivation

4. the second term in the expansion is unknown, but it is a constant and we can get rid of it, and change sign as well (with an additional factor 2, see later), so that by minimizing the first term we actually minimize the KL divergence

$$\int_\Theta f(x|\theta) \ln \frac{f(x|\hat{\theta}^{(k)})}{f(x|\theta)} dx = \int_\Theta f(x|\theta) \ln f(x|\hat{\theta}^{(k)}) - \int_\Theta f(x|\theta) \ln f(x|\theta) \quad \sim \quad -2 \int_\Theta f(x|\theta) \ln f(x|\hat{\theta}^{(k)})$$

5. going back to Wilks' theorem, we know that the remaining |*l*-*k*| degrees of freedom in the likelihood ratio are (asymptotically) normally distributed, therefore the -2log has a chi-square distribution with |*l*-*k*| degrees of freedom, with mean value |*l*-*k*|, and therefore the required mean value has an asymptotic bias 2|*l*-*k*|;
6. using the max likelihood as an estimator of the mean, we find that the discrepancy expressed by the equation above can be written as

$$-2 \ln f(x|\hat{\theta}^{(k)}) + 2k$$

after dropping the constant l

Bayesian Information Criterion (BIC).

This was derived by Gideon Schwarz in 1978 and strongly resembles the AIC. It is given by

$$\mathrm{BIC} = -2 \ln \mathcal{L}_{\max} + k \ln N$$

where N is the number of datapoints. Since a typical dataset will have $\ln N > 2$*, the BIC imposes a stricter penalty against extra parameters than the AIC.*

It was derived as an approximation to the Bayesian evidence**, to be discussed next, but **the assumptions required are very restrictive and unlikely to hold in practice, rendering the approximation quite crude.

(Liddle & al., 2006)

Bayesian evidence

Model selection aims to determine which theoretical models are most plausible given some data, without necessarily considering preferred values of model parameters.

Ideally, we would like to estimate posterior probabilities on the set of all competing models using Bayes' theorem:

$$P(M_i|D,I) = \frac{P(D|M_i,I)P(M_i|I)}{\sum_k P(D|M_k,I)P(M_k|I)}$$

and select the best model using the odds ratio

$$\mathcal{O}_{i,j} = \frac{P(M_i|D,I)}{P(M_j|D,I)} = \frac{P(D|M_i,I)P(M_i|I)}{P(D|M_j,I)P(M_j|I)}$$

or the Bayes factor, if we assume equal prior probabilities for the different models:

$$B_{i,j} = \frac{P(D|M_i,I)}{P(D|M_j,I)}$$

Thus, we see that the Bayes factor is a ratio of evidences

$$B_{i,j} = \frac{P(D|M_i, I)}{P(D|M_j, I)}$$

As usual, each evidence is obtained by marginalizing the likelihood with respect to the (potentially different) parameters:

$$P(D|M_i, I) = \int_{\Theta_i} P(D|\boldsymbol{\theta}_i, M_i, I) p(\boldsymbol{\theta}_i|M_i, I) d\boldsymbol{\theta}_i$$

The evidence of a model is … the average likelihood of the model in the prior.

Unlike the AIC and BIC, it does not focus on the best-fitting parameters of the model but asks "of all the parameter values you thought were viable before the data came along, how well on average did they fit the data?". Literally, it is the likelihood of the model given the data.

The evidence rewards predictability of models, provided they give a good fit to the data, and hence gives an axiomatic realization of Occam's razor.

A model with little parameter freedom is likely to fit data over much of its parameter space, whereas a model that could match pretty much any data that might have cropped up will give a better fit to the actual data but only in a small region of its larger parameter space, pulling the average likelihood down.

(Liddle & al., 2006)

Which statistics?

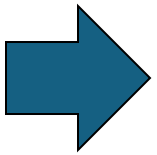
Of these statistics, we would advocate using – wherever possible – the Bayesian evidence, which is a full implementation of Bayesian inference and can be directly interpreted in terms of model probabilities. It is computationally challenging to compute, being a highly peaked multidimensional integral, but recent algorithm development has made it feasible in cosmological contexts.

If the Bayesian evidence cannot be computed, the BIC can be deployed as a substitute. *It is much simpler to compute as one need only find the point of maximum likelihood for each model.* ***However, interpreting it can be difficult. Its main usefulness is as an approximation to the evidence, but this holds only for gaussian likelihoods and provided the datapoints are independent and identically distributed.*** *The latter condition holds poorly for the current global cosmological dataset, though it can potentially be improved by binning of the data, hence decreasing the N in the penalty term.*

The AIC has been widely used outside astrophysics but is of debatable utility. ***It has been shown to be "dimensionally inconsistent", meaning that it is not guaranteed to give the right result even in the limit of infinite unbiased data.*** *It may be useful for checking the robustness of conclusions drawn using the BIC.* ***The evidence and BIC are dimensionally consistent.***

(Liddle & al., 2006)

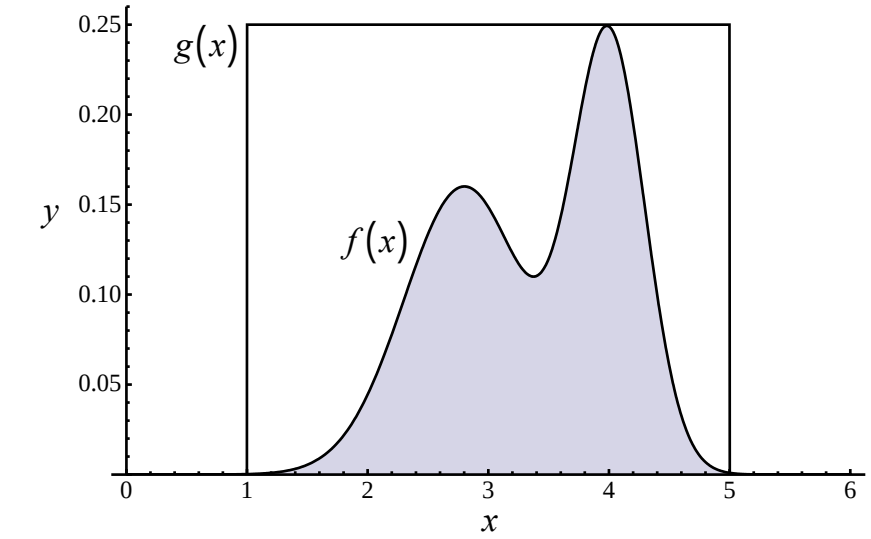
Our next important topic: Bayesian estimates often require complex numerical integrals. How do we confront this problem?

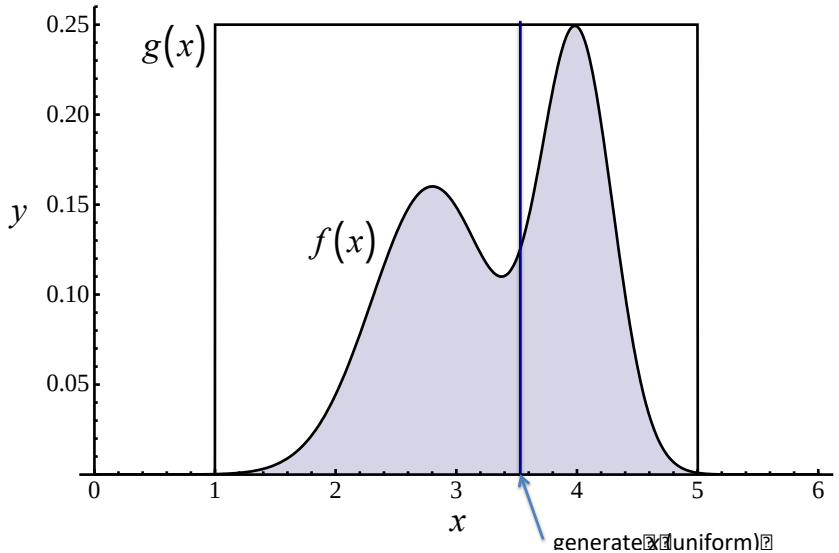


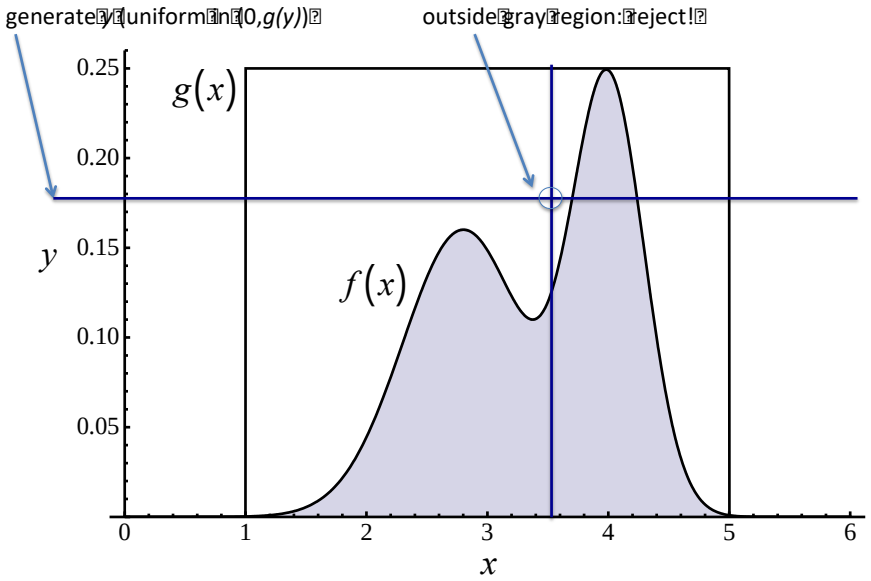
enter the Monte Carlo methods!

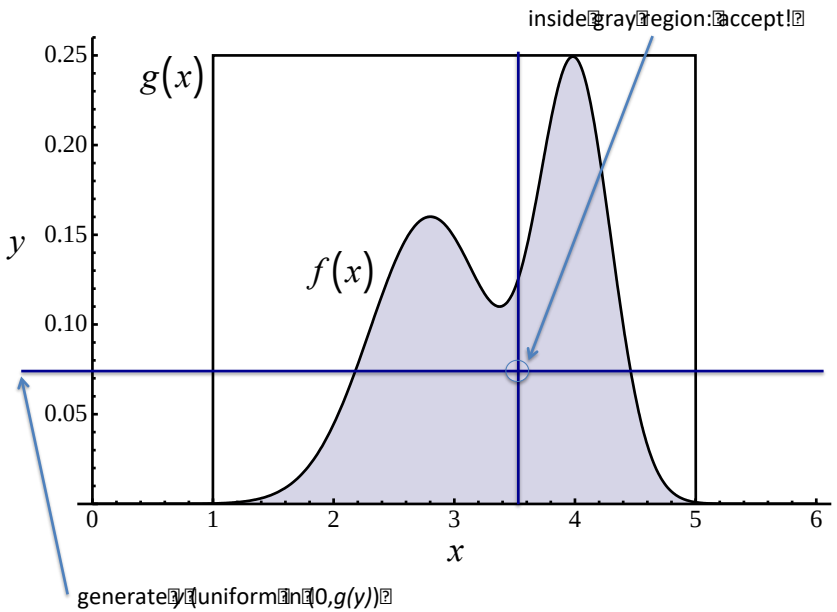
1. acceptance-rejection sampling
2. importance sampling
3. statistical bootstrap
4. Bayesian methods in a sampling-resampling perspective
5. Introduction to Markov chains and to Random Walks (RW)
6. Detailed balance and Boltzmann's H-theorem
7. The Gibbs sampler
8. More on Gibbs sampling
9. Simulated annealing and the Traveling Salesman Problem (TSP)
10. The Metropolis algorithm
11. Image restoration and Markov Random Fields (MRF)
12. The Metropolis-Hastings algorithm and Markov Chain Monte Carlo (MCMC)
13. The efficiency of MCMC methods
14. Affine-invariant MCMC algorithms (emcee)

1. The acceptance rejection method

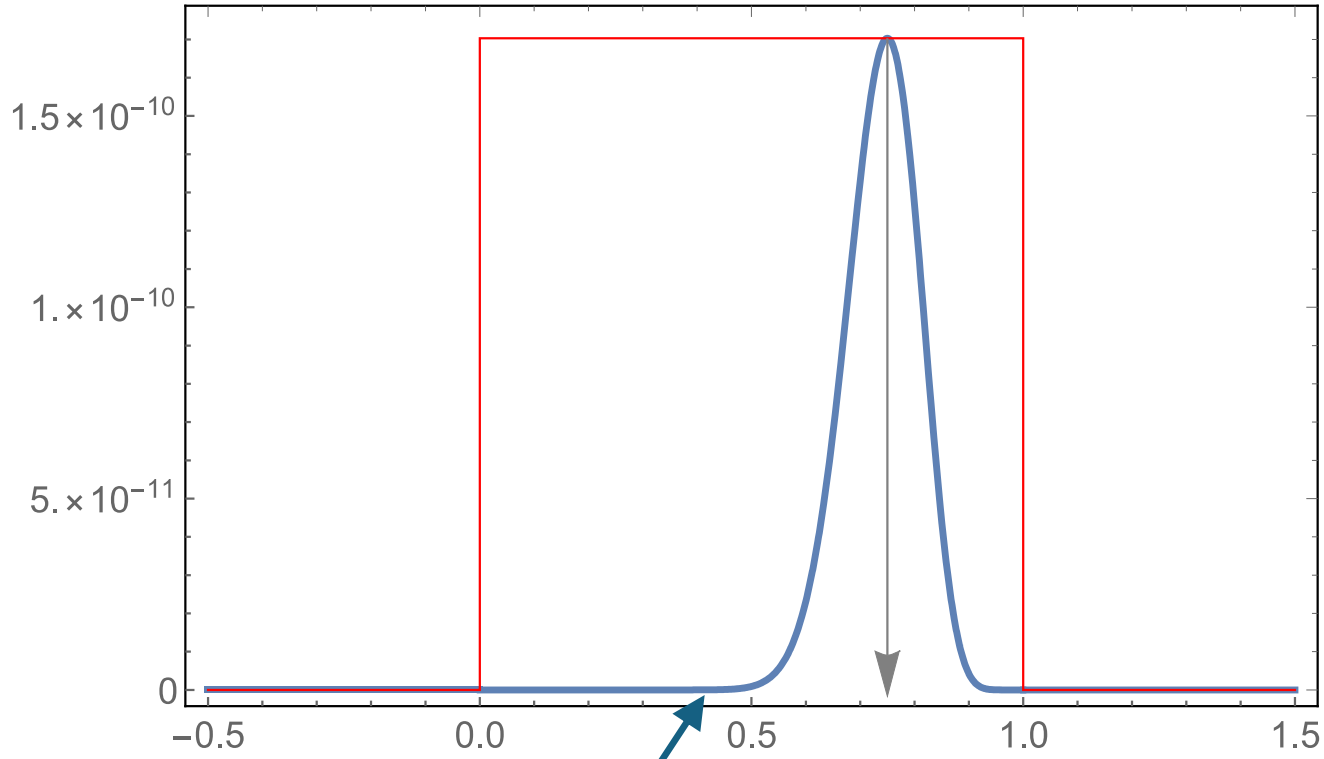








Example: generation of beta-distributed random numbers



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

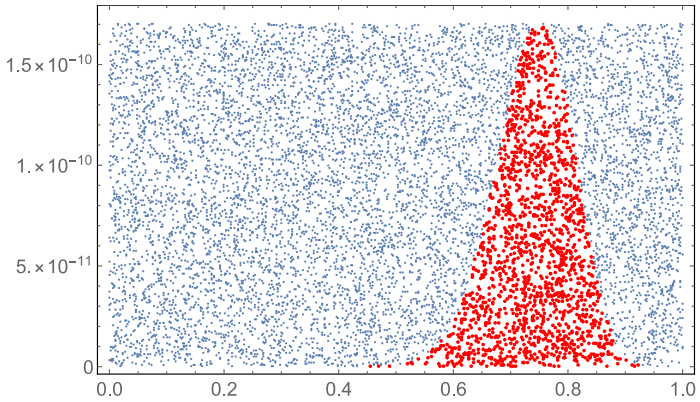
normalized distribution

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

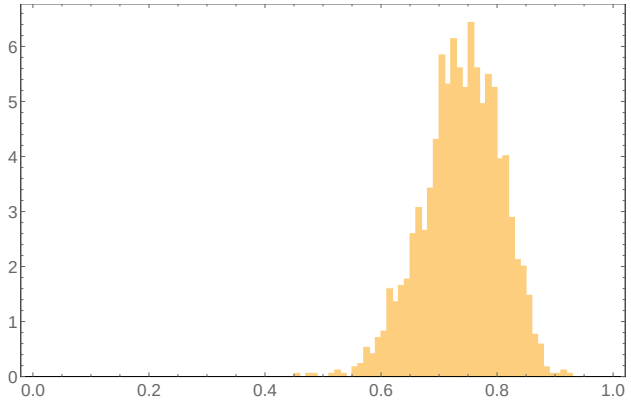
unnormalized distribution

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

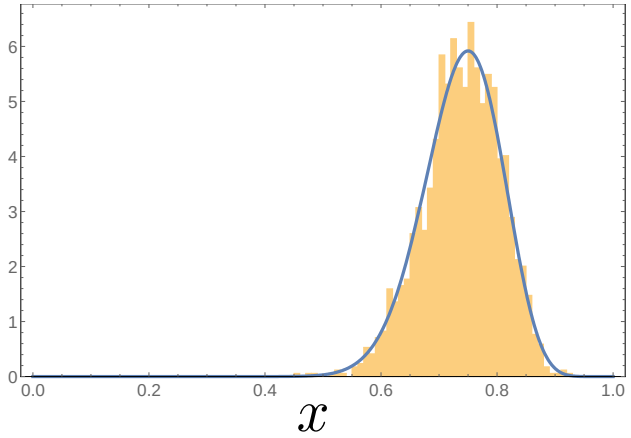
modal value



generated pairs
(red = accepted pairs)



normalized histogram of the accepted *x*'s



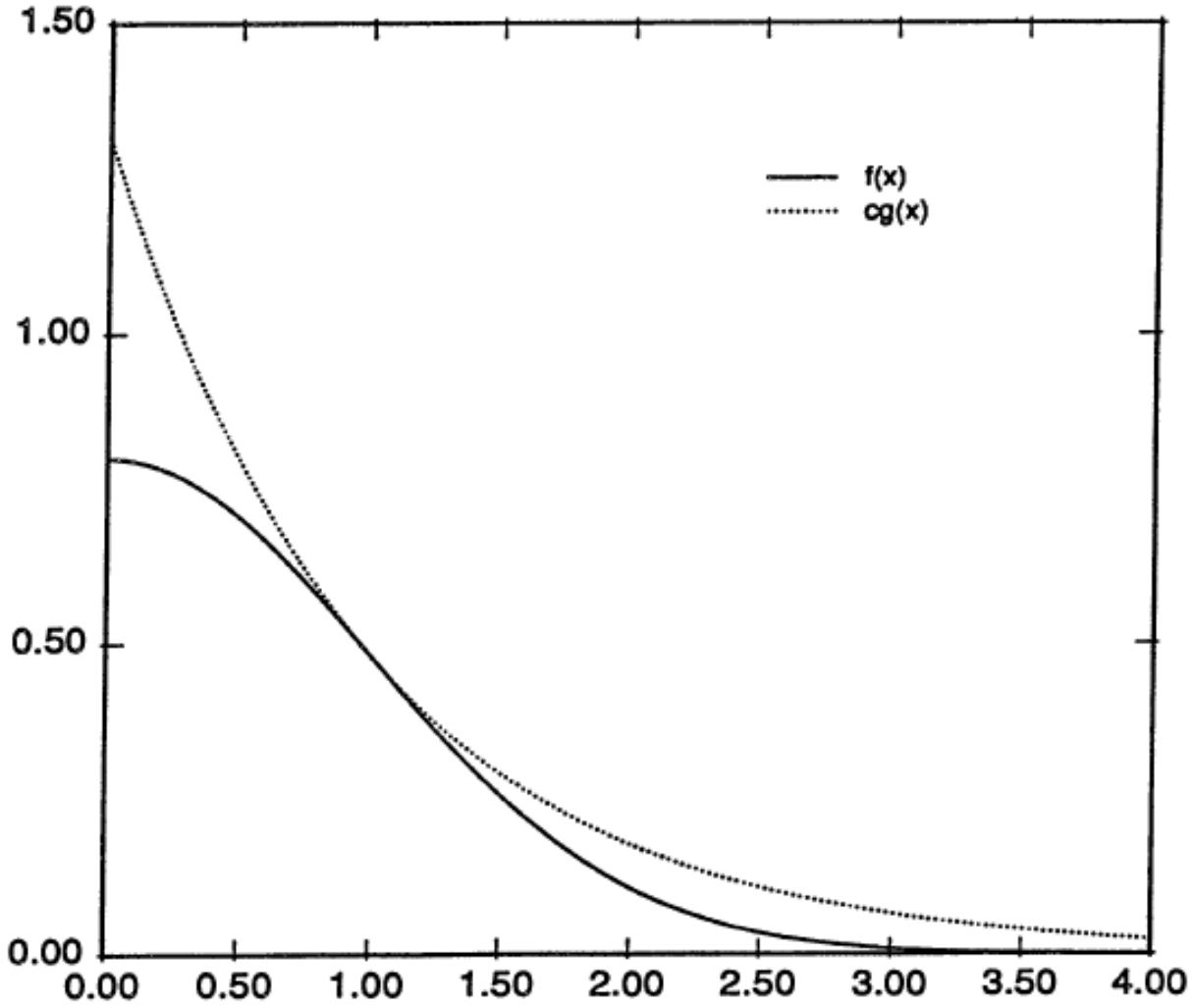
comparison with the plot of the normalized beta distribution

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

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

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

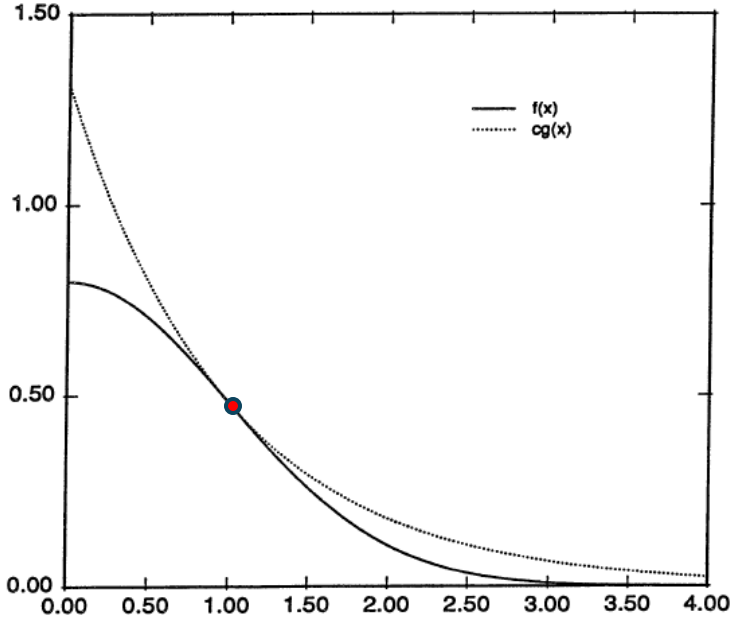
... determine *c* accordingly



Definition of contact point (to maximize efficiency)

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

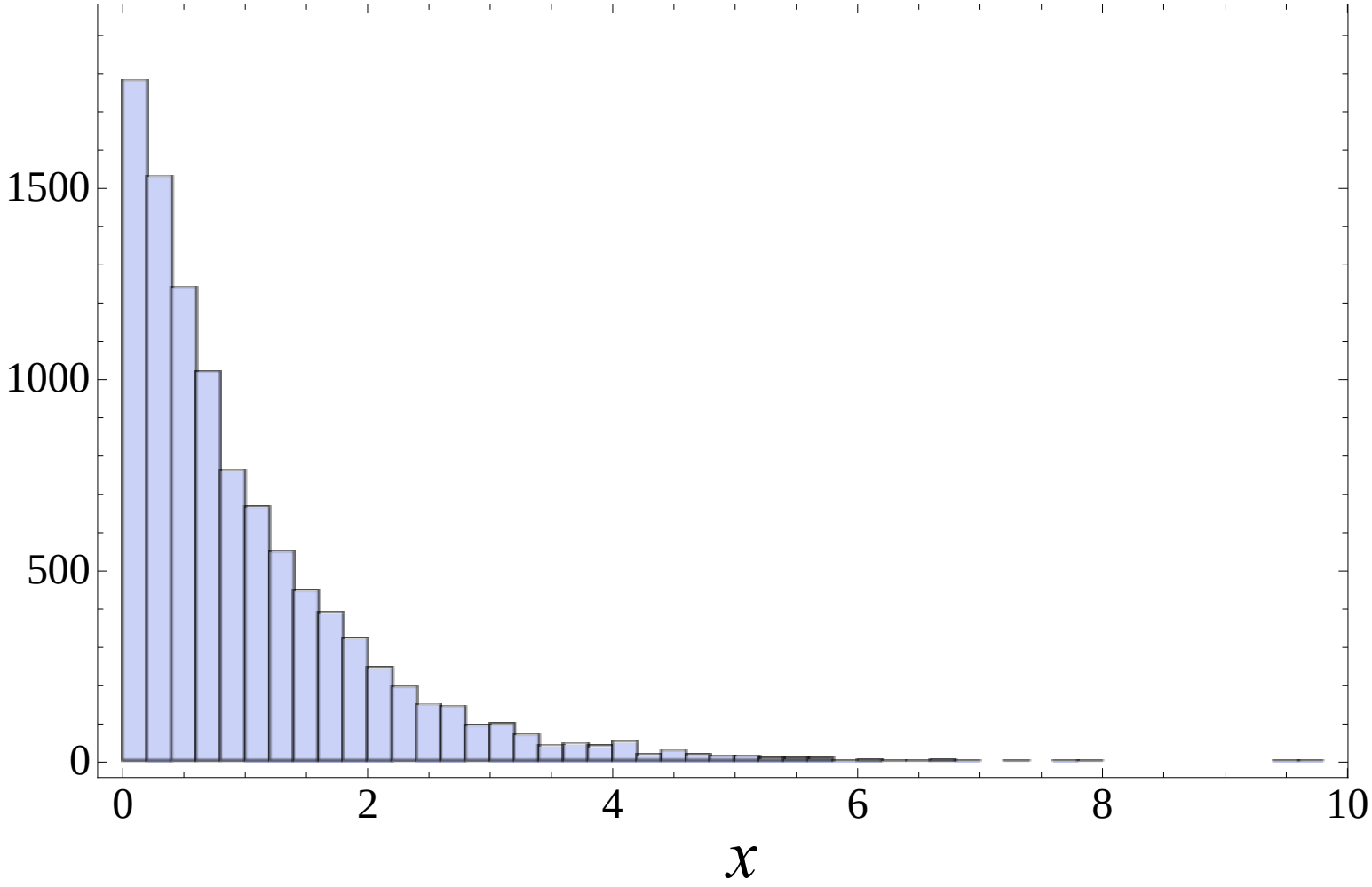
$$g(x) = \exp(-x)$$



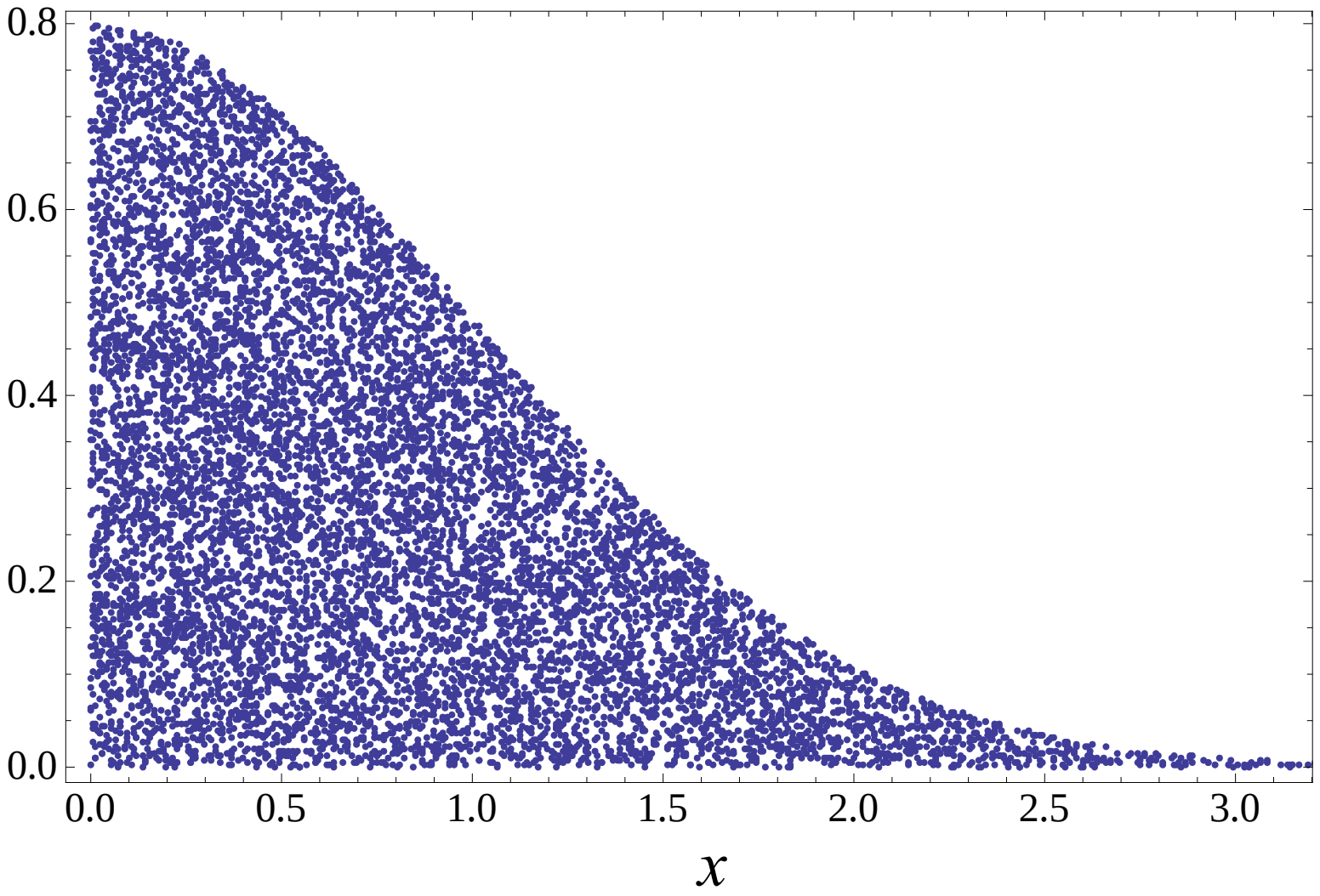
$$\Rightarrow \quad \begin{cases} f(x) = cg(x) \\ f'(x) = cg'(x) \end{cases} \quad \Rightarrow \quad \begin{cases} \sqrt{\frac{2}{\rho}} \exp\left(-\frac{x^2}{2} \right) = c \exp(-x) \\ x\sqrt{\frac{2}{\rho}} \exp\left(-\frac{x^2}{2} \right) = c \exp(-x) \end{cases}$$

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

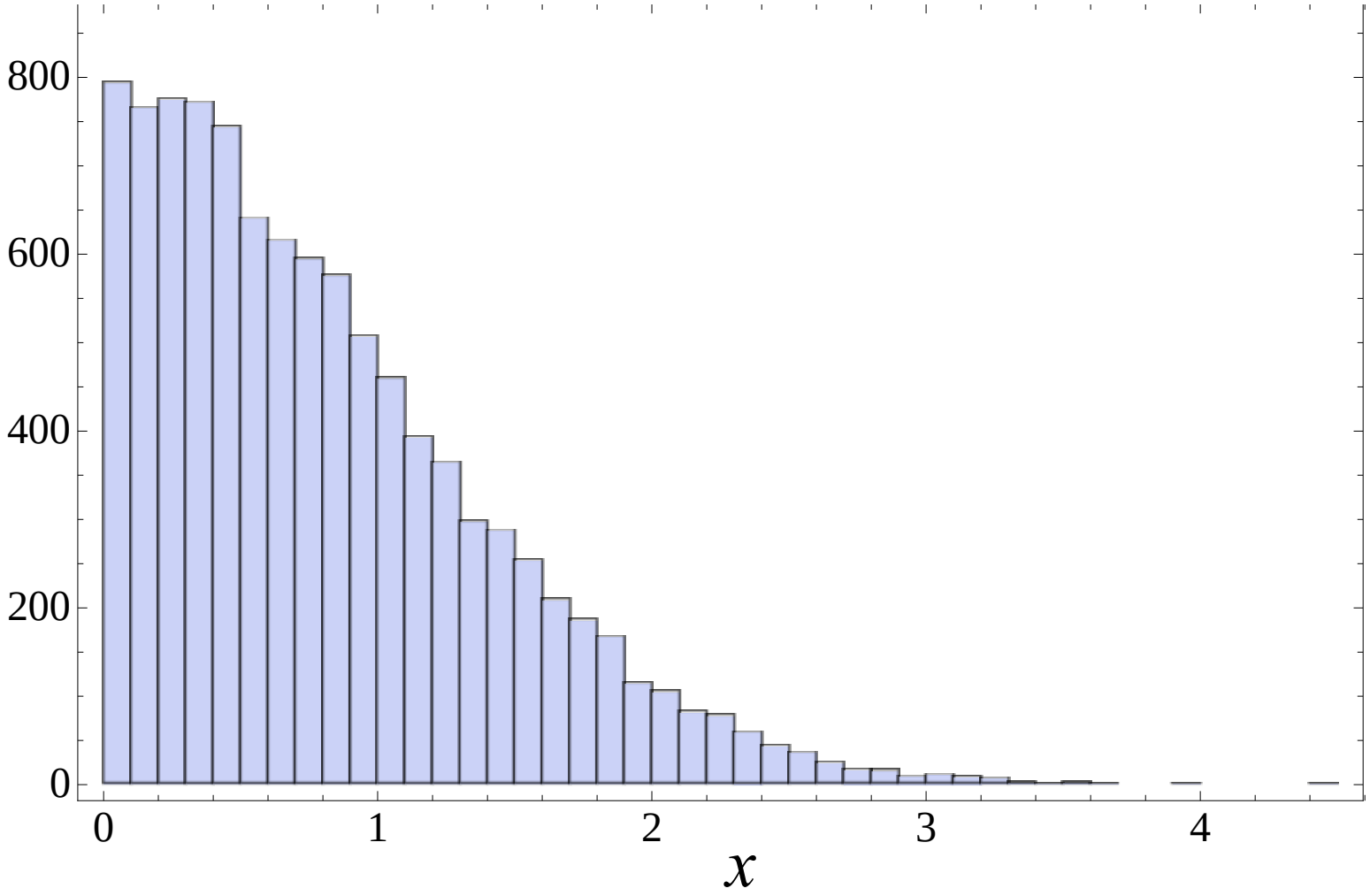
Exponentially distributed values



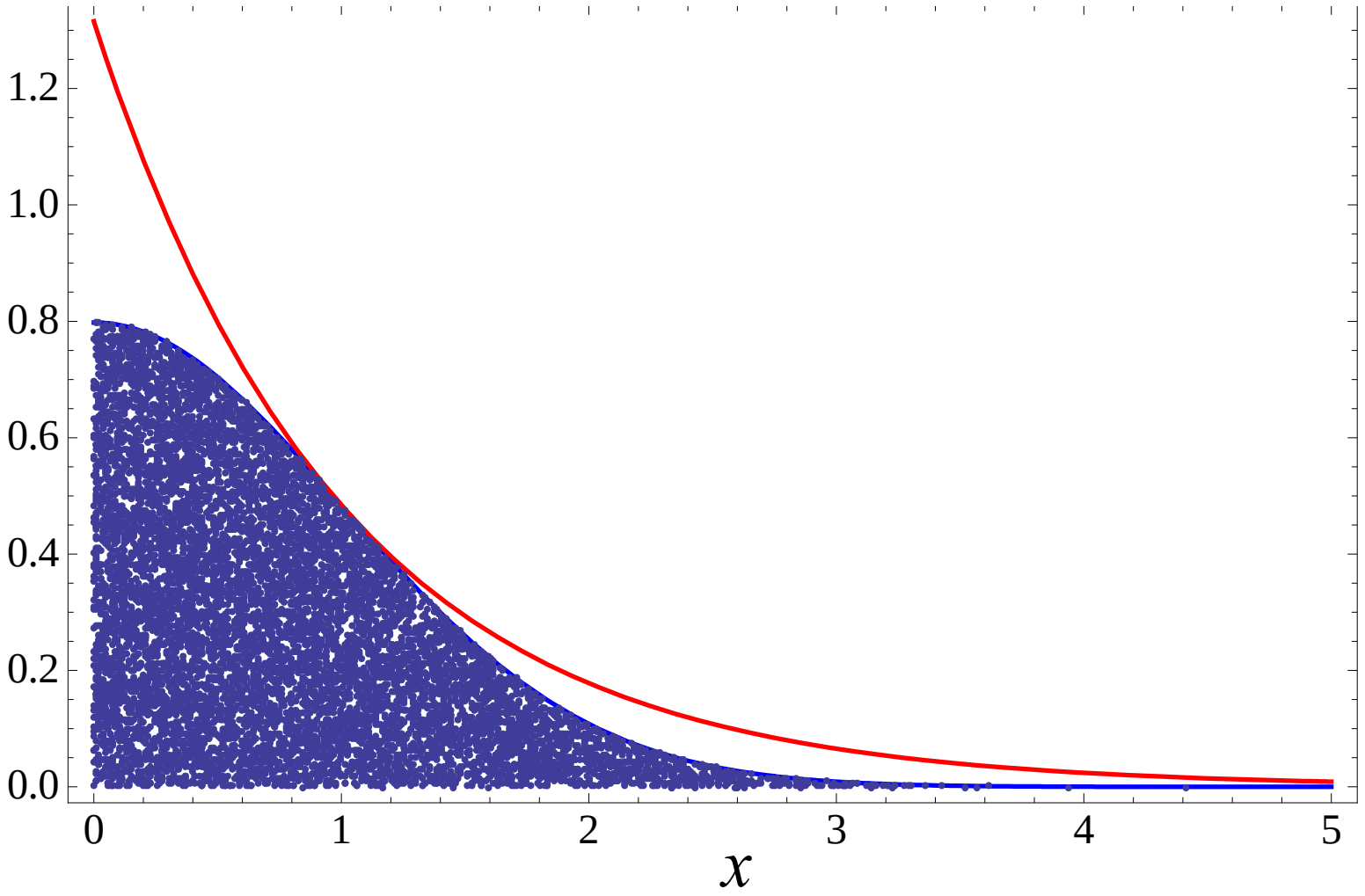
A/R accepted values (10000 accepted sample pairs)

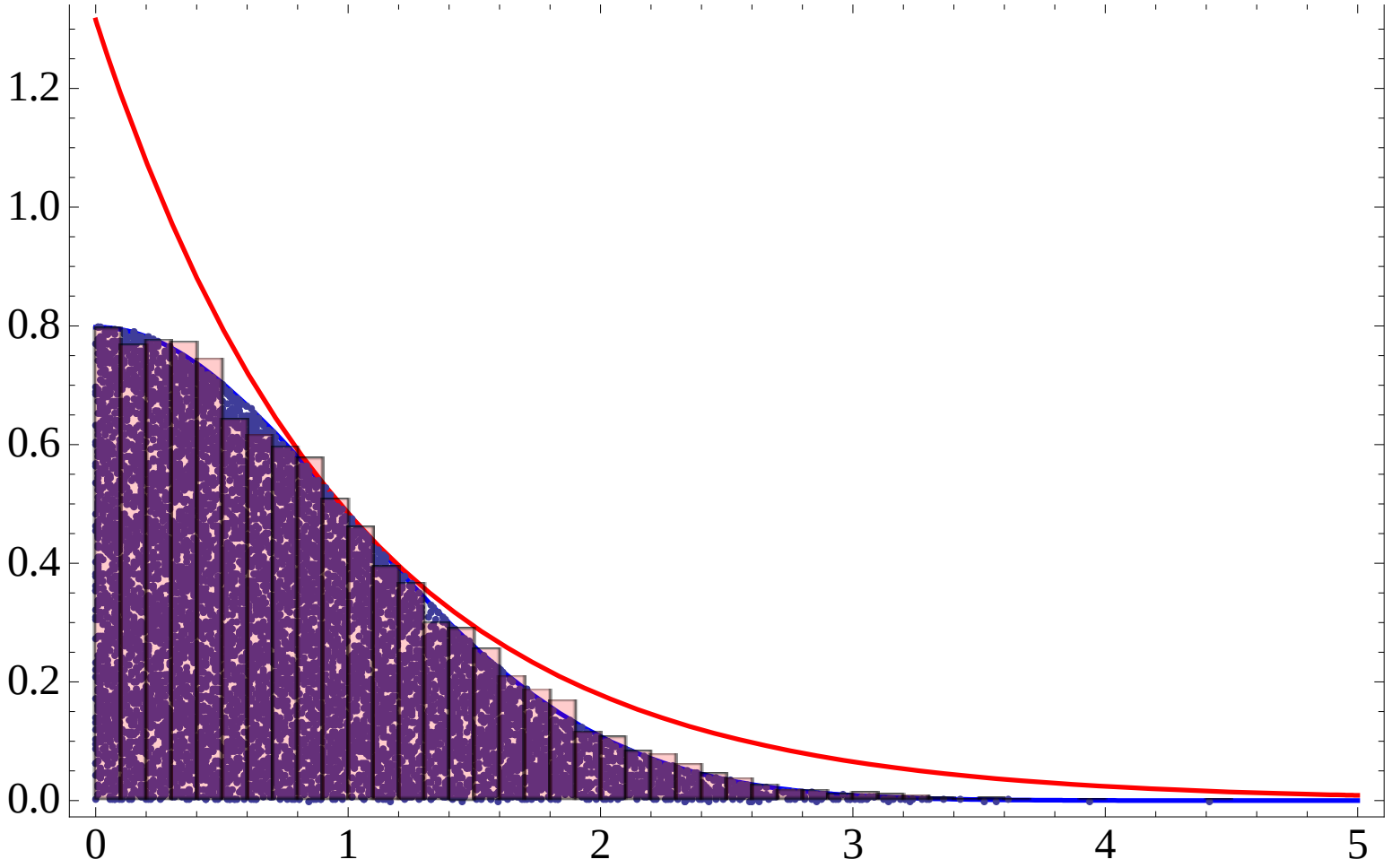


Histogram of accepted *x* values



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

Short summary:

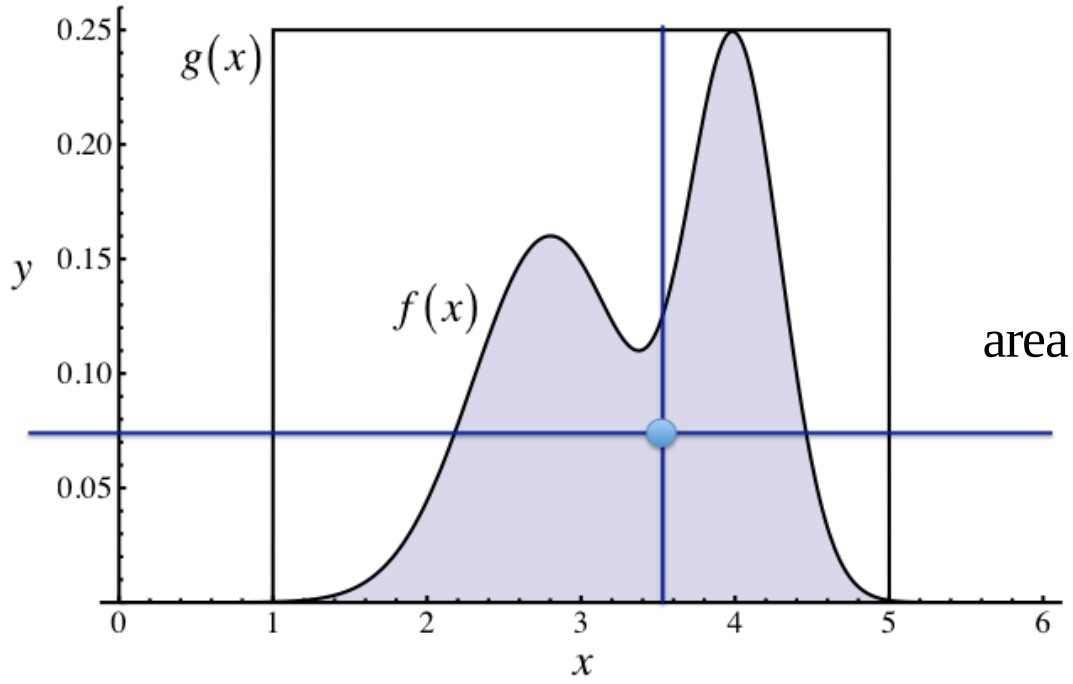
1. we create a data set by randomly sampling from the exponential distribution

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 ...)

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

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

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 …

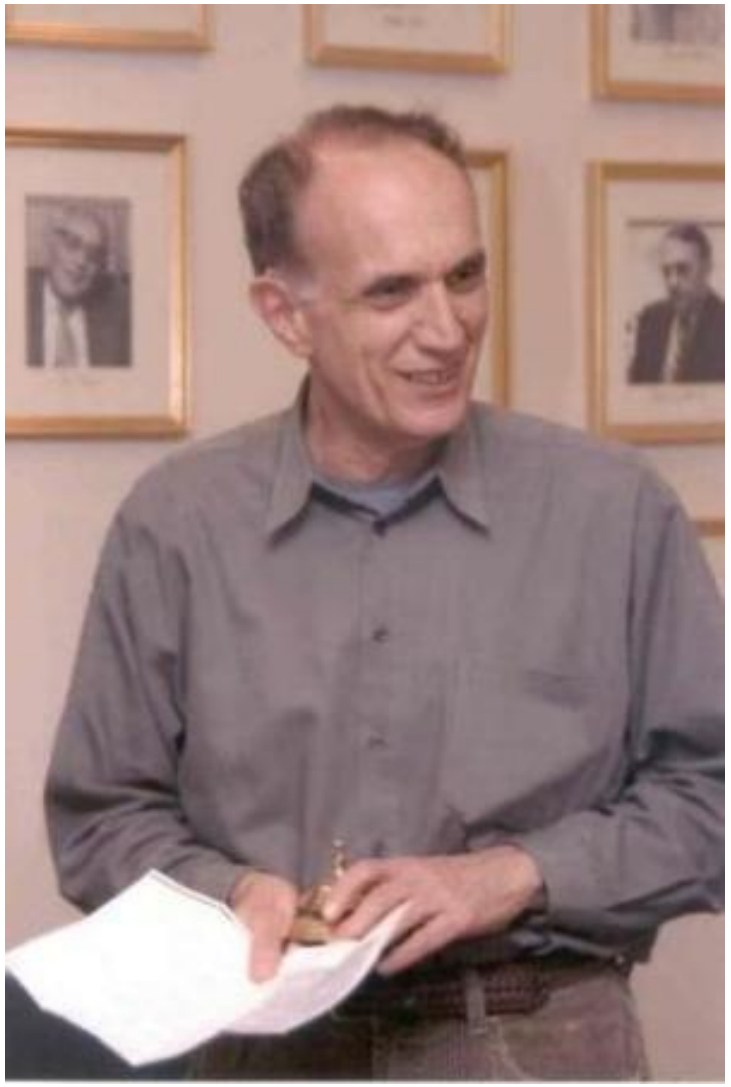
$$\int_a^b h(x)\,dx = \int_a^b f(x)\,p(x)\,dx = \int_a^b \left[f(x)\frac{p(x)}{q(x)} \right] q(x)\,dx$$

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

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

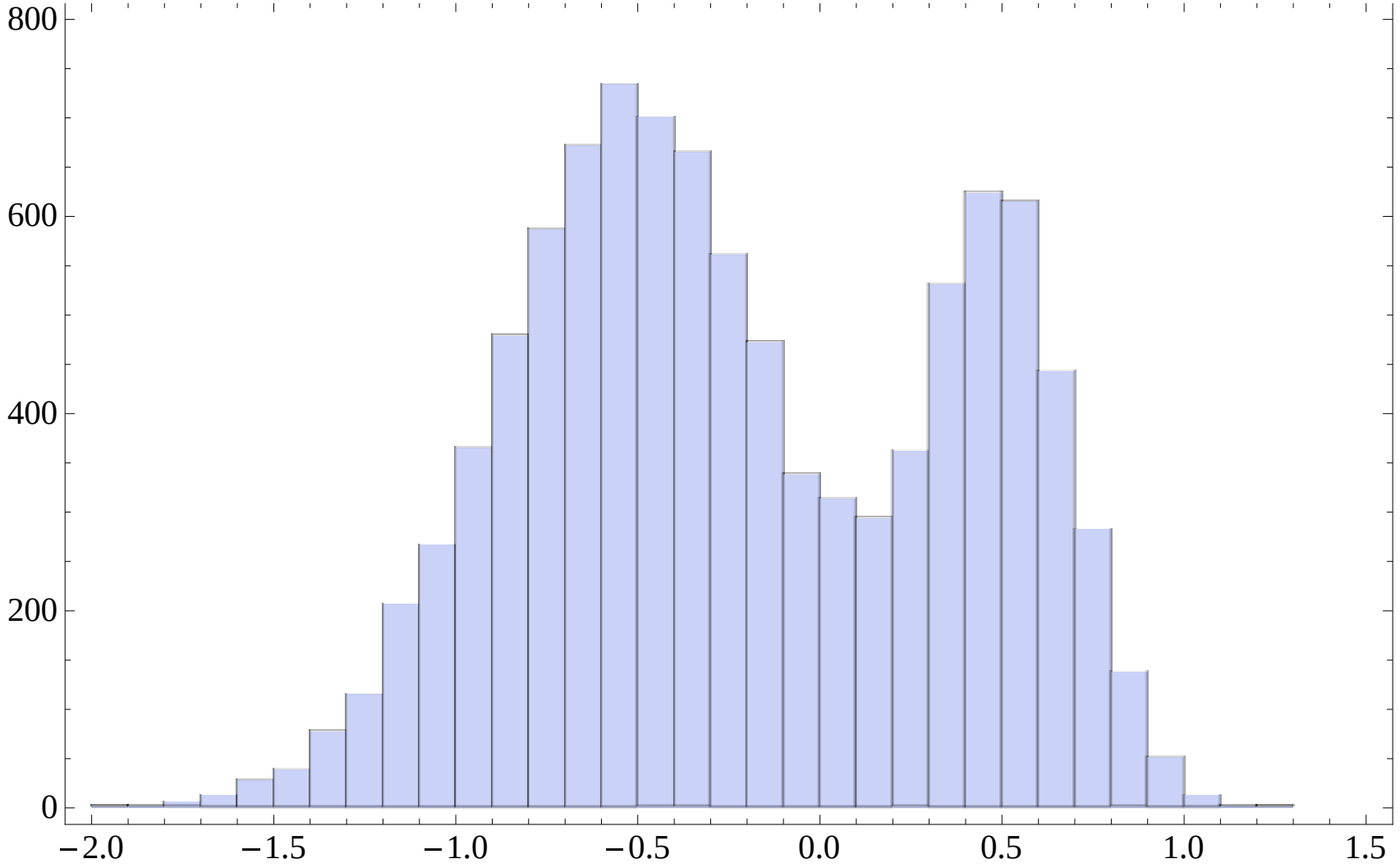
These methods are still not very efficient and there is a better alternative, the Markov Chain Monte Carlo method (see later)

3. An important resampling technique: the Bootstrap method (B. Efron, 1977)

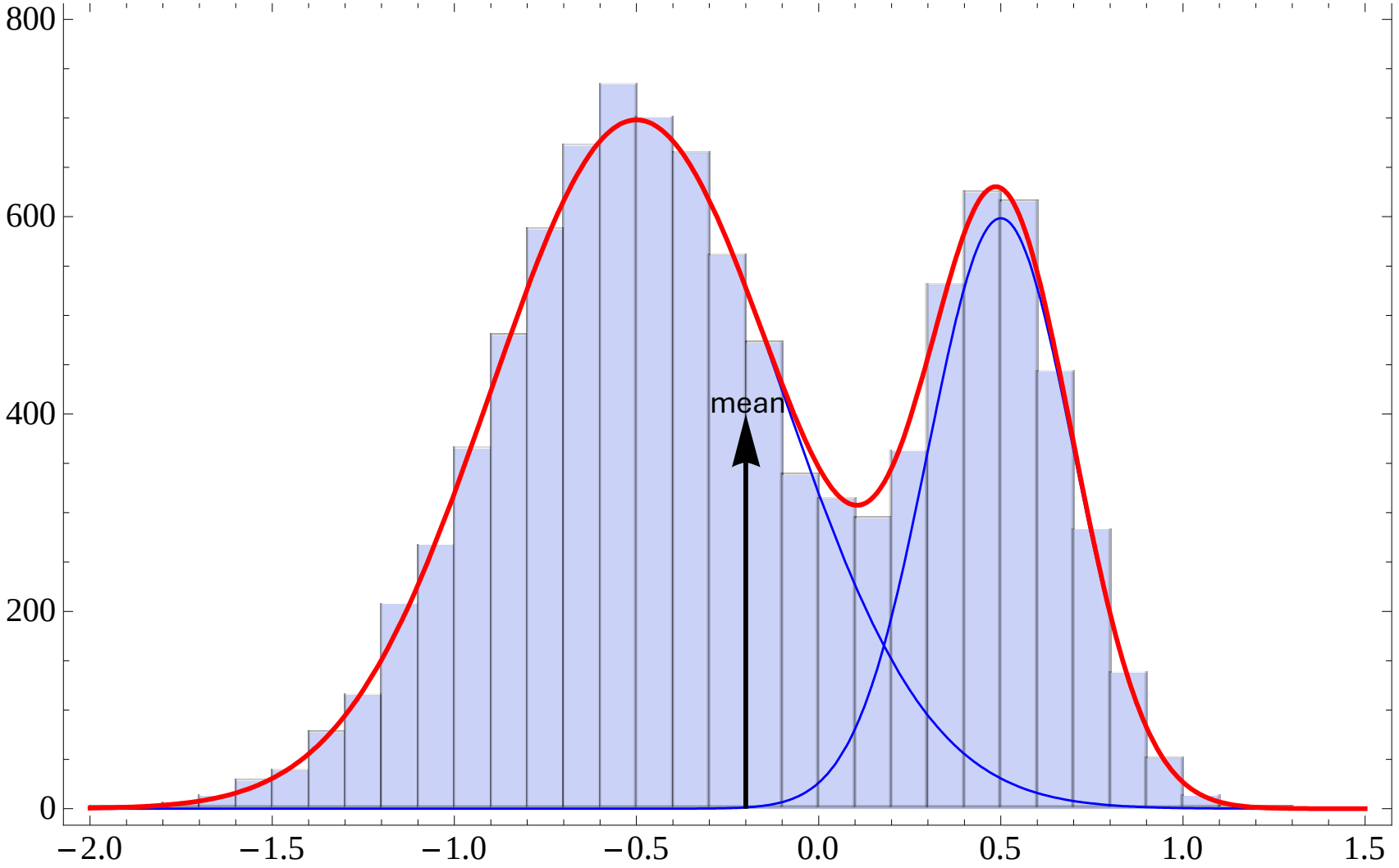


The bootstrap method is a resampling technique that helps calculating many statistical estimators

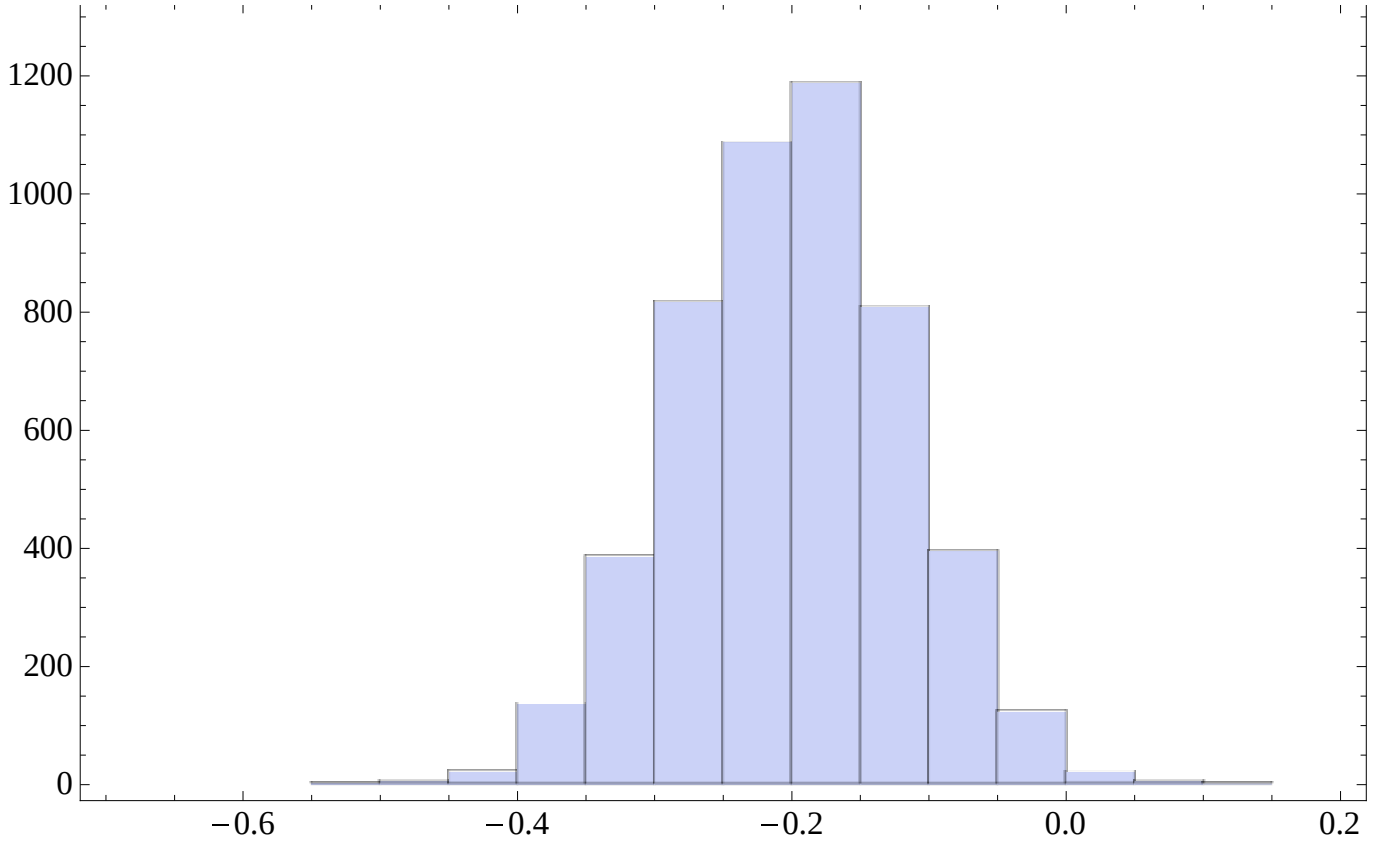
consider the distribution of a set of measurements



the distribution of data approximates the "true" underlying distribution (in this case a mixture model)

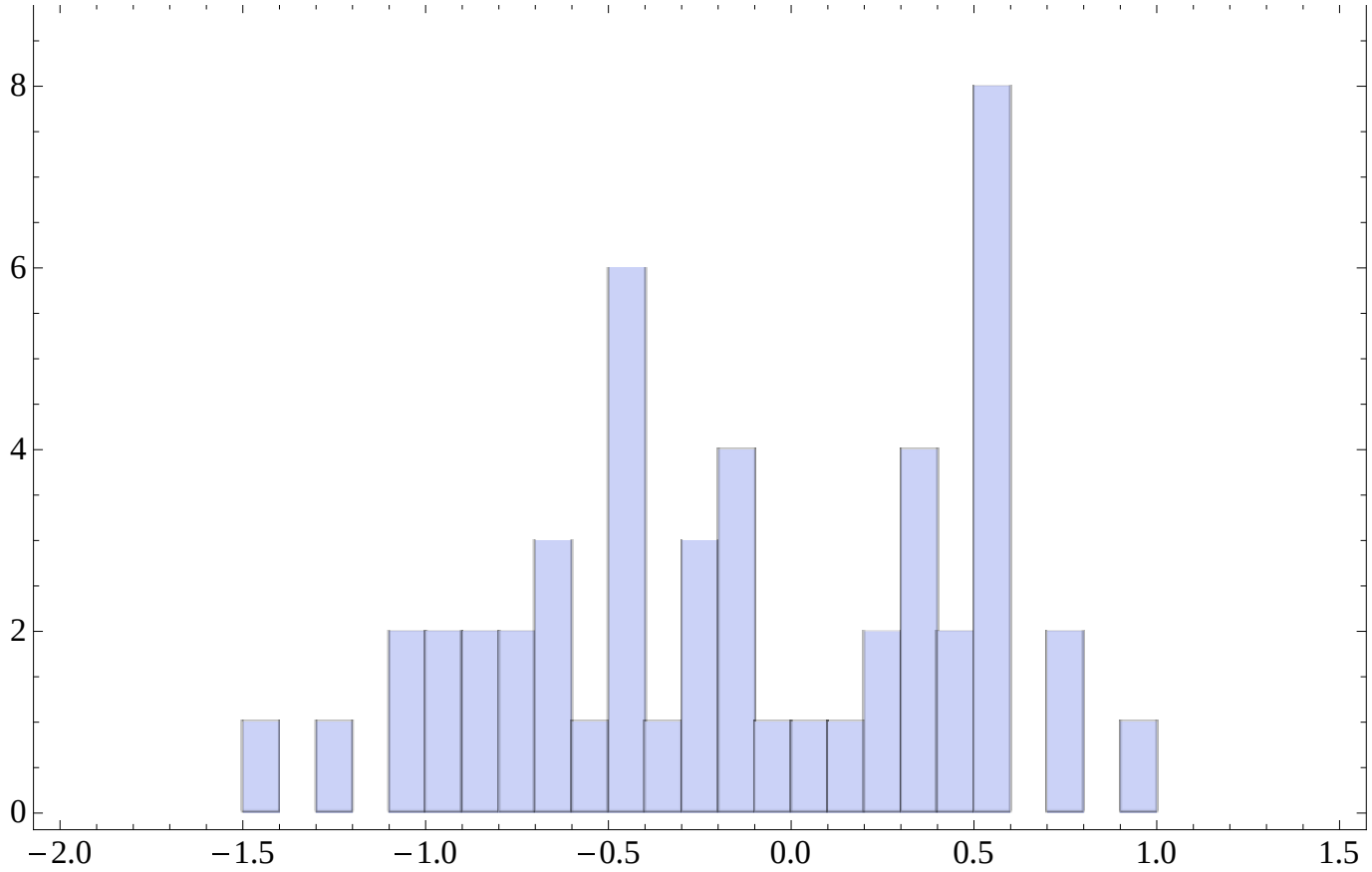


distribution of mean value obtained from 5000 sets of data (sample size = 50)



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

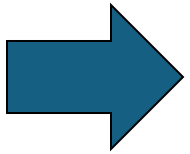
example: single dataset (same size as before, 50 measurements)



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

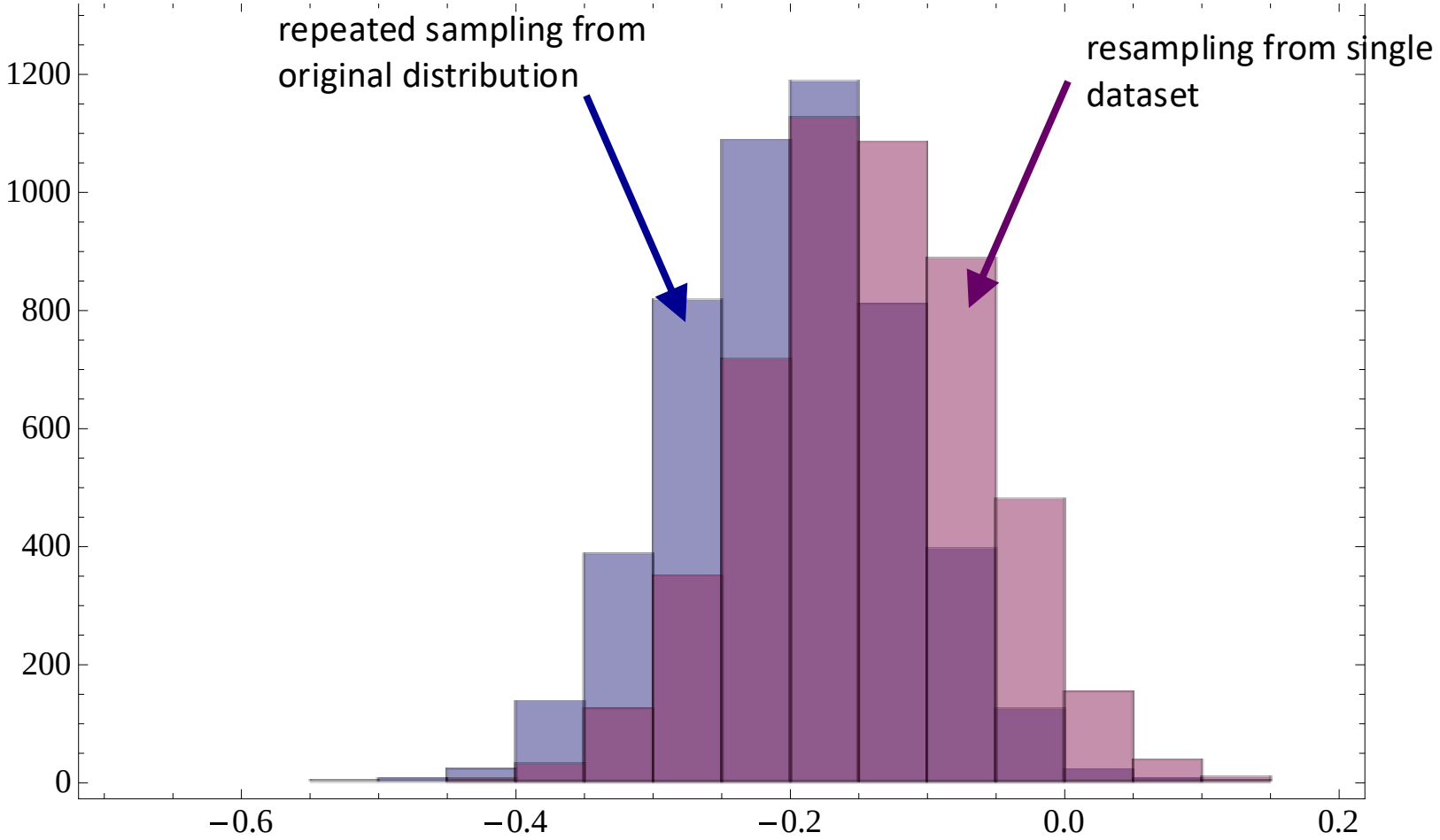
Bootstrap recipe:

if you need 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

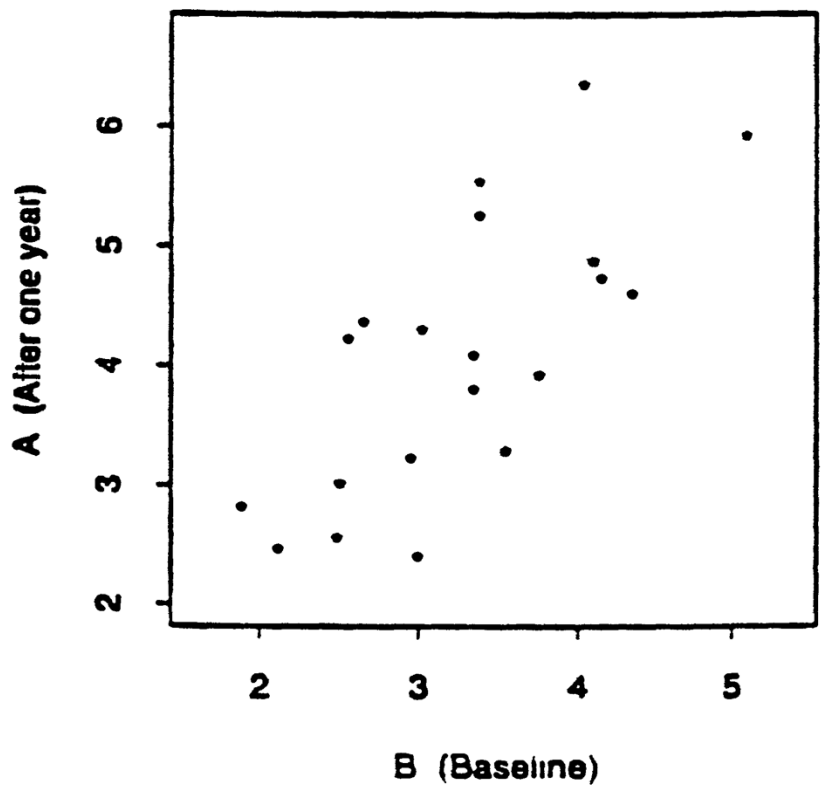


FIG. 1. *The cd4 data; cd4 counts in hundreds for 20 subjects, at baseline and after one year of treatment with an experimental anti-viral drug; numerical values appear in Table 1.*

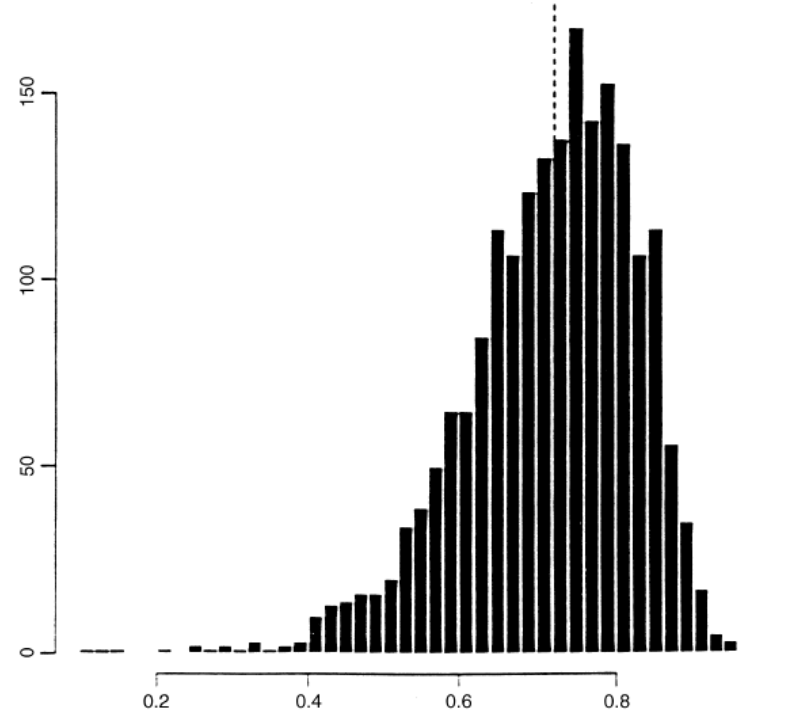


FIG. 3. *Histogram of 2,000 bootstrap correlation coefficients; bivariate normal sampling model.*

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

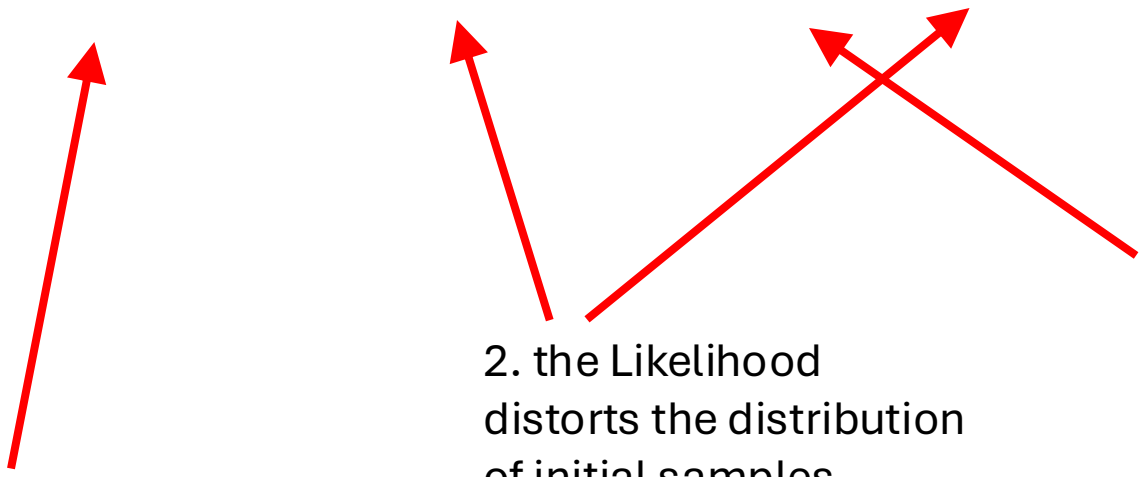
$$E[m(\theta)|x] = \int m(\theta)p(\theta|x)\, d\theta$$

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 (importance sampling-like scheme)

$$p(\theta_k|D) \propto p(D|\theta_k)p(\theta_k) = \ell(D|\theta_k)p(\theta_k)$$



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

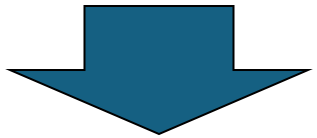
(sampling)

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

$$X_{i1} = \text{Binomial}\left(n_{i1}, q_1\right)$$
$$X_{i2} = \text{Binomial}\left(n_{i2}, q_2\right)$$

The observed random variables are the sums

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



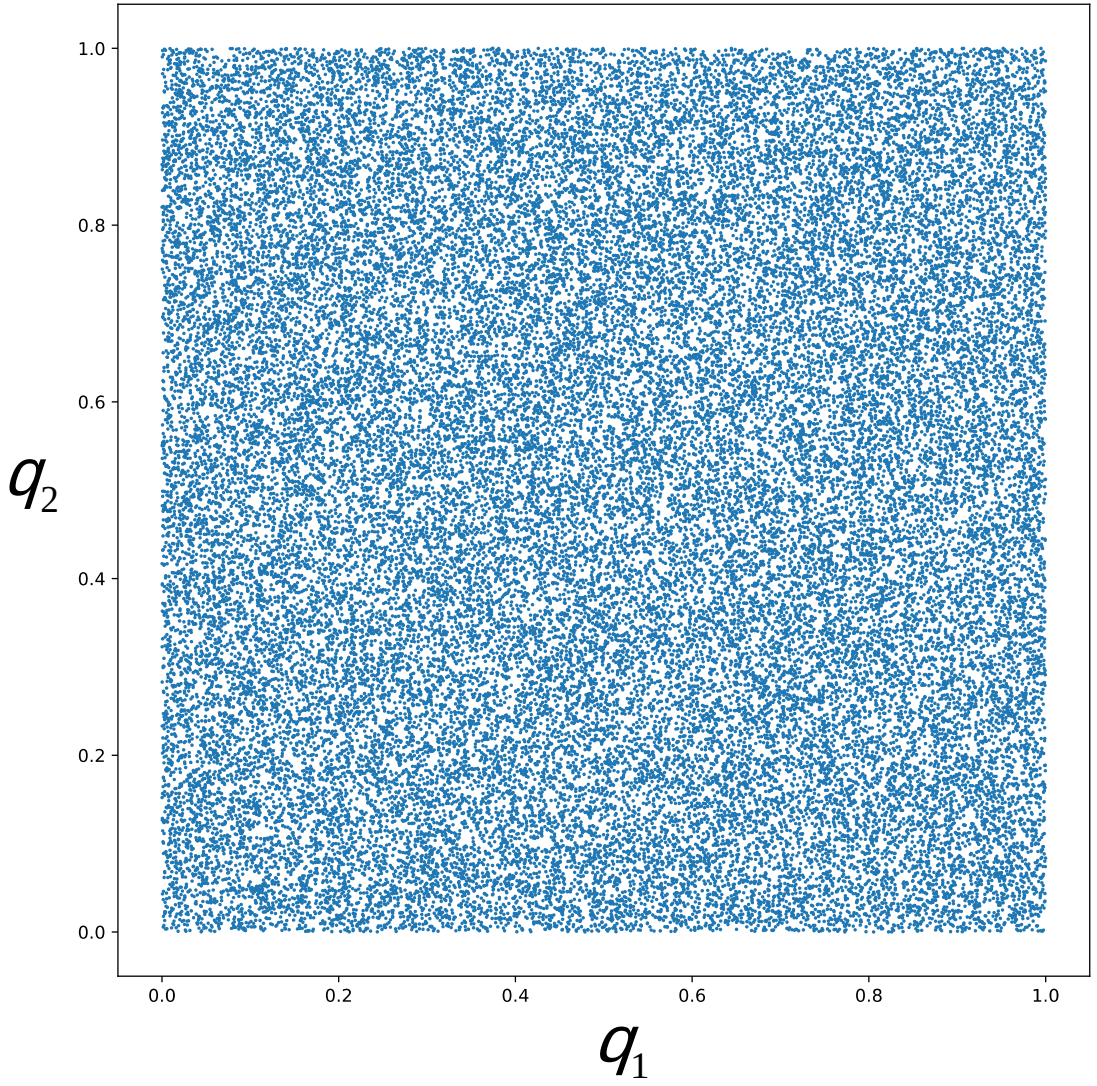
$$\text{likelihood} = \prod_{i=1}^{3} \sum_{j_i} \binom{n_{i1}}{j_i} \binom{n_{i2}}{y_i - j_i} \theta_1^{j_1} (1-\theta_1)^{n_{i1}-j_i} \theta_2^{y_i-j_i} (1-\theta_2)^{n_{i2}-y_i+j_i}$$

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

Sample data

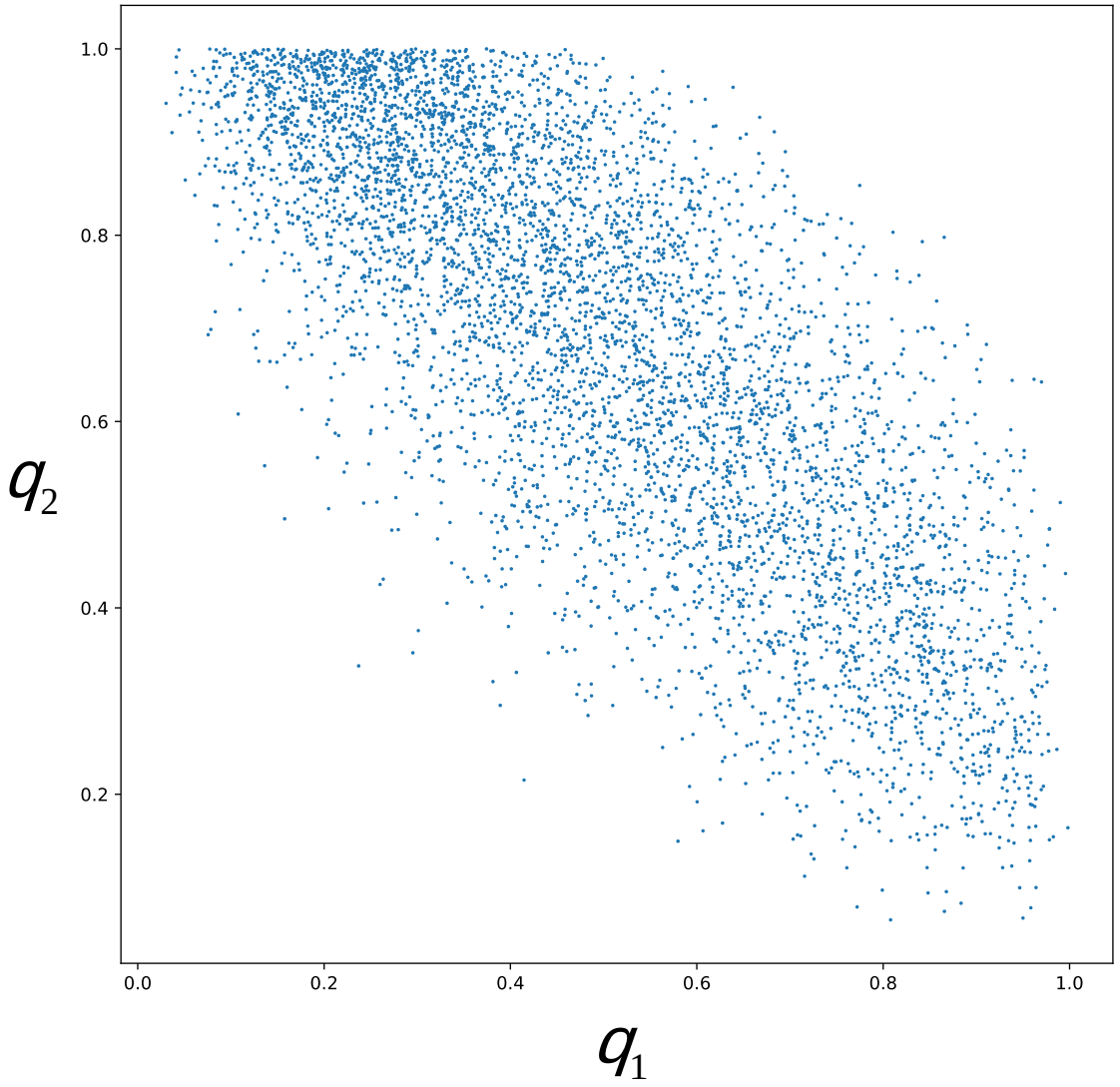
	1	**2**	**3**
n_{i1}	5	6	4
n_{i2}	5	4	6
y_i	7	5	6

Example of implementation in Python (see Jupyter notebook)

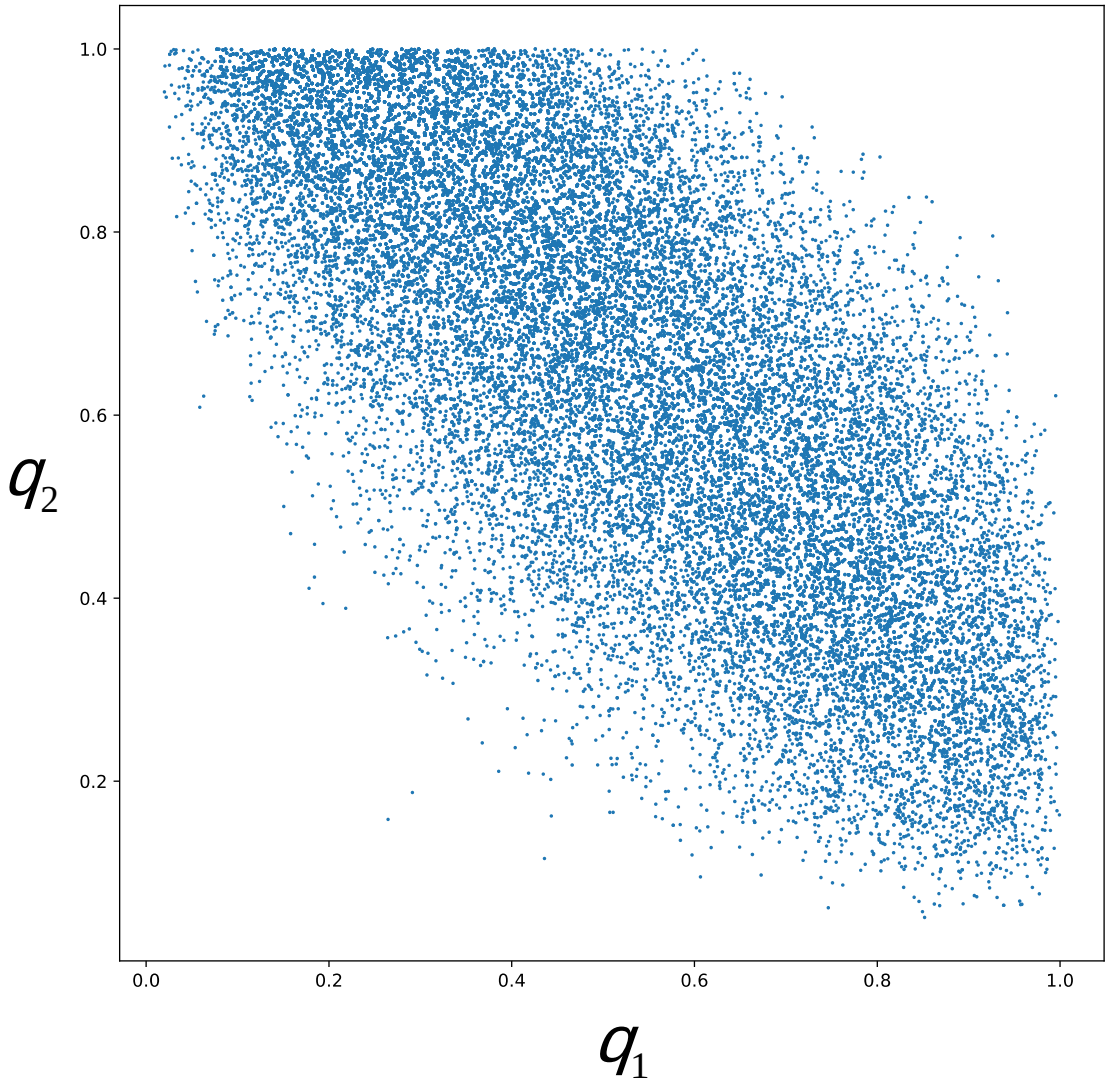


prior distribution
(50000 samples, uniform in 2D parameter space)

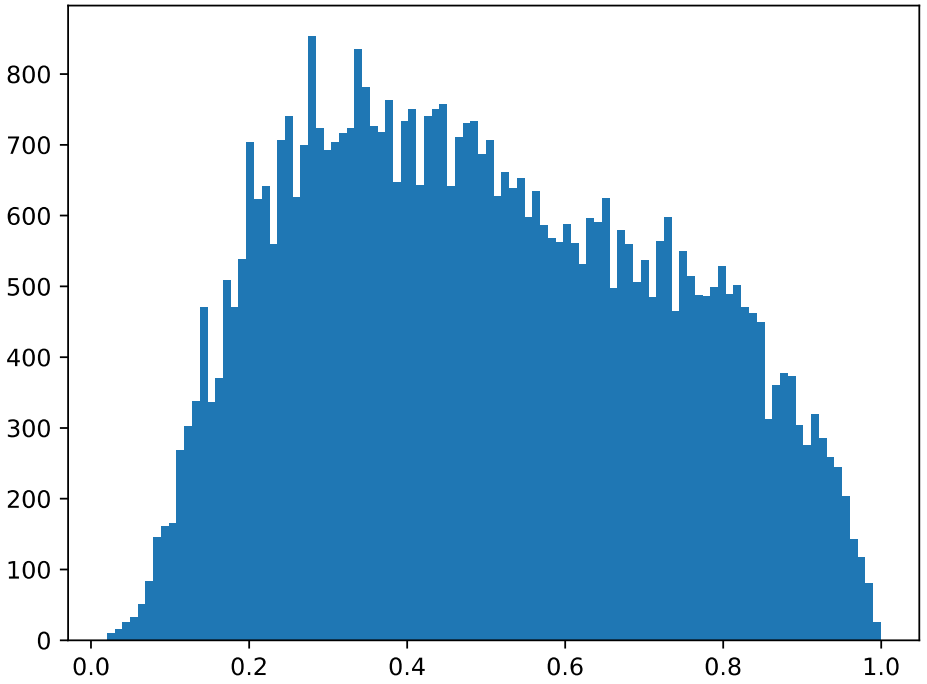
Posterior as a resampled prior using acceptance-rejection



Posterior as a resampled prior using weighted bootstrap

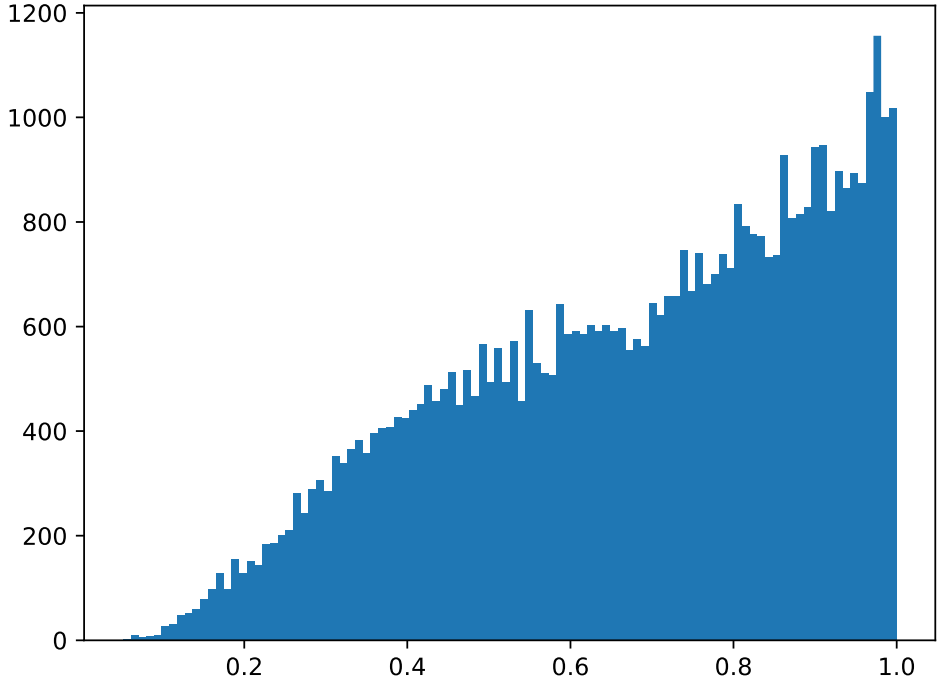


The resampled points are representative of the posterior distribution and can be used to evaluate any sample estimate



Marginalized histogram of q_1

Sample mean: 0.500 ± 0.001



Marginalized histogram of q_2

Sample mean: 0.677 ± 0.001

... 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