

Introduction to Bayesian Methods - 5

Edoardo Milotti

Università di Trieste and INFN-Sezione di Trieste

Example: expert elicitation

(Morgan, PNAS 111 (2014) 7176)

Consider a problem where no experimental data exist, and where you wish to make an informed guess.

You can rely on expert opinion and ask the expert to provide his/her own estimate of a probability distribution.

You can also rely on a population of experts and construct averaged probability distributions from their guesses.

The Interpretation of Probability

A subjectivist or Bayesian interpretation of probability (5, 26–28) is used when one makes subjective probabilistic assessments of the present or future value of uncertain quantities, the state of the world, or the nature of the processes that govern the world. In such situations, probability is viewed as a statement of an individual's belief, informed by all formal and informal evidence that he or she has available. Although subjective, such judgments cannot be arbitrary. They must conform to the laws of probability. Further, when large quantities of evidence are available on identical repeated events, one's subjective probability should converge to the classical frequentist interpretation of probability.

Expert elicitation is a complex procedure, and one should resort to it only when absolutely necessary.

Morgan concludes that

“... it may be tempting to view expert elicitation as a low-cost, low-effort alternative to conducting serious research and analysis, it is neither. Rather, expert elicitation should build on and use the best available research and analysis and be undertaken only when, given those, the state of knowledge will remain insufficient to support timely informed assessment and decision making.”

Partial list of problems in expert elicitation

1. Words convey different meanings

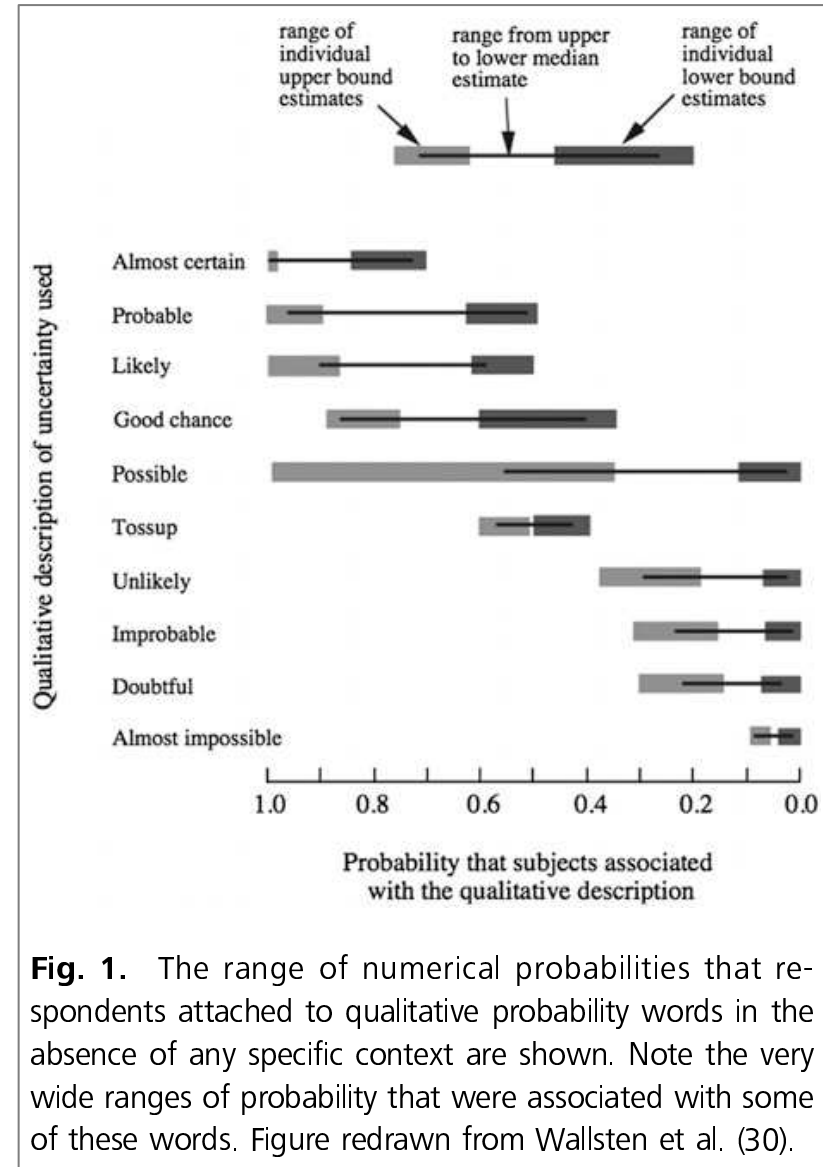




Fig. 2. Results obtained by Morgan (32) when members of the Executive Committee of the EPA Science Advisory Board were asked to assign numerical probabilities to uncertainty words that had been proposed for use with EPA cancer guidelines (33). Note that even in this relatively small and expert group, the minimum probability associated with the word “likely” spans 4 orders of magnitude, the maximum probability associated with the word “not likely” spans more than 5 orders of magnitude, and there is an overlap of the probabilities the different experts associated with the two words.

2. Cognitive heuristics and bias

... When presented with an estimation task, if people start with a first value (i.e., an anchor) and then adjust up and down from that value, they typically do not adjust sufficiently. ...

To minimize the influence of this heuristic when eliciting probability distributions, it is standard procedure not to begin with questions that ask about “best” or most probable values but rather to first ask about extremes: “What is the highest (lowest) value you can imagine for coefficient X?” or “Please give me a value for coefficient X for which you think there is only one chance in 100 that actual value could be larger (smaller).”

Having obtained an estimate of an upper (lower) bound, it is then standard practice to ask the expert to imagine that the uncertainty about the coefficient’s value has been resolved and the actual value has turned out to be 10% or 15% larger (smaller) than the bound they offered. We then ask the expert, “Can you offer an explanation of how that might be possible?”

Sometimes experts can offer a perfectly plausible physical explanation, at which point we ask them to revise their bound. After obtaining estimates of upper and lower bounds on the value of a coefficient of interest, we then go on to elicit intermediate values across the probability distribution [“What is the probability that the value of X is greater (less) than Y?”].

3. Ubiquitous overconfidence

... One reason for adopting this rather elaborate procedure is that there is strong evidence that most such judgments are overconfident.

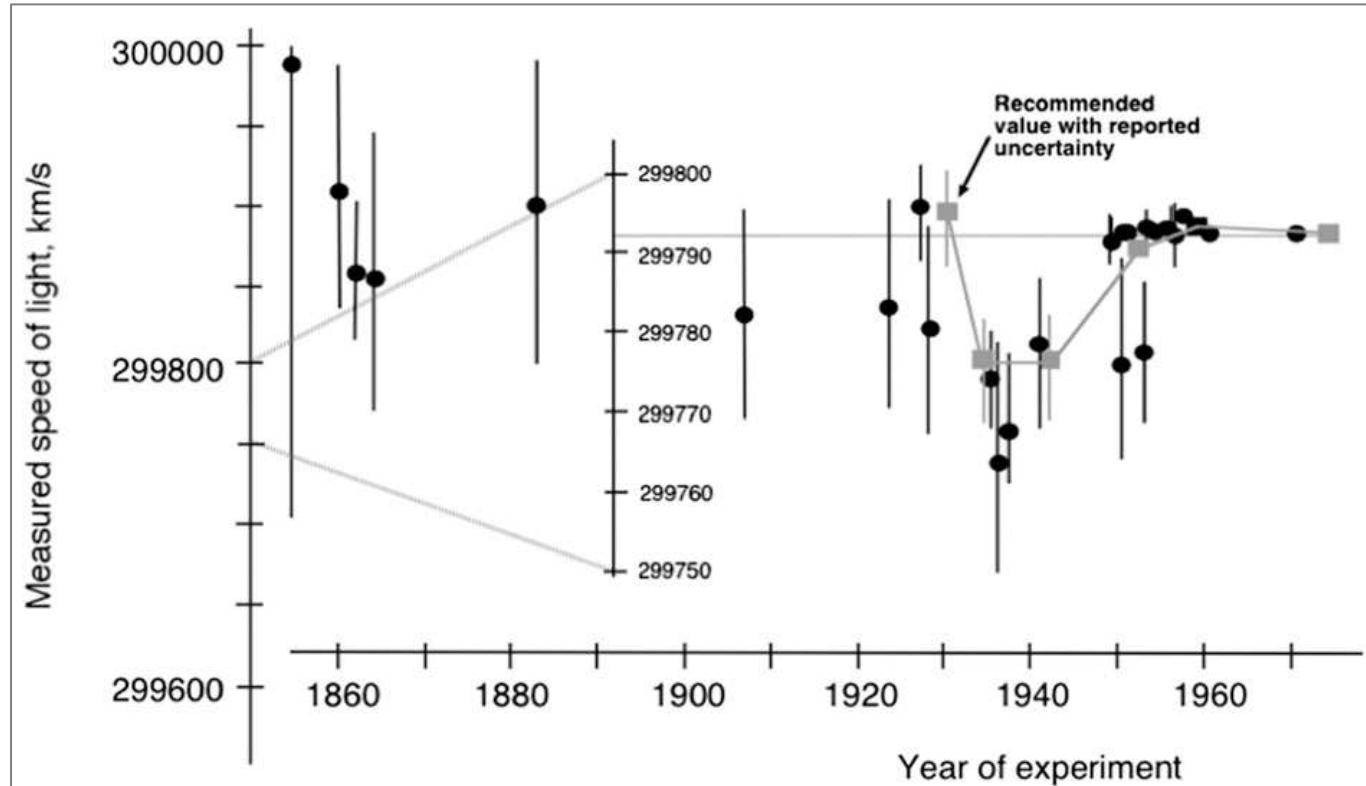


Fig. 4. Published estimates of the speed of light. The light gray boxes that start in 1930 are the recommended values from the particle physics group that presumably include an effort to consider uncertainty arising from systematic error (40). Note that for over two decades the reported confidence intervals on these recommended values did not include the present best-measured value. Henrion and Fischhoff (40), from which this figure is combined and redrawn, report that the same overconfidence is observed in the recommended values of a number of other physical constants.

4. Expert calibration

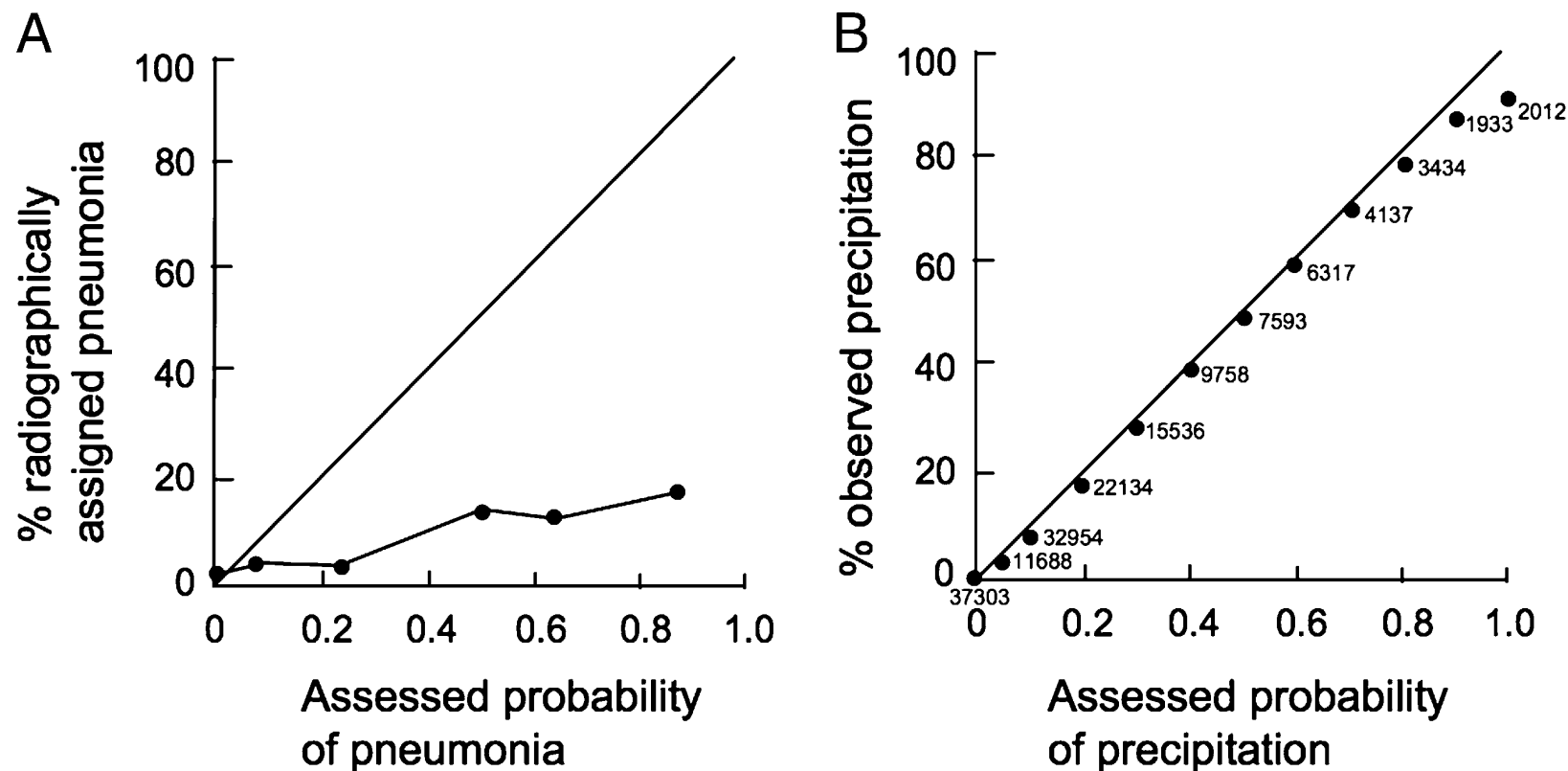


Fig. 5. Illustration of two extremes in expert calibration. (A) Assessment of probability of pneumonia (based on observed symptoms) in 1,531 first-time patients by nine physicians compared with radiographically assigned cases of pneumonia as reported by Christensen-Szalanski and Bushyhead (44). (B) Once-daily US Weather Service precipitation forecasts for 87 stations are compared with actual occurrence of precipitation (April 1977 to March 1979) as reported by Charba and Klein (43). The small numbers adjacent to each point report the number of forecasts.

5. Range of opinions in the scientific community

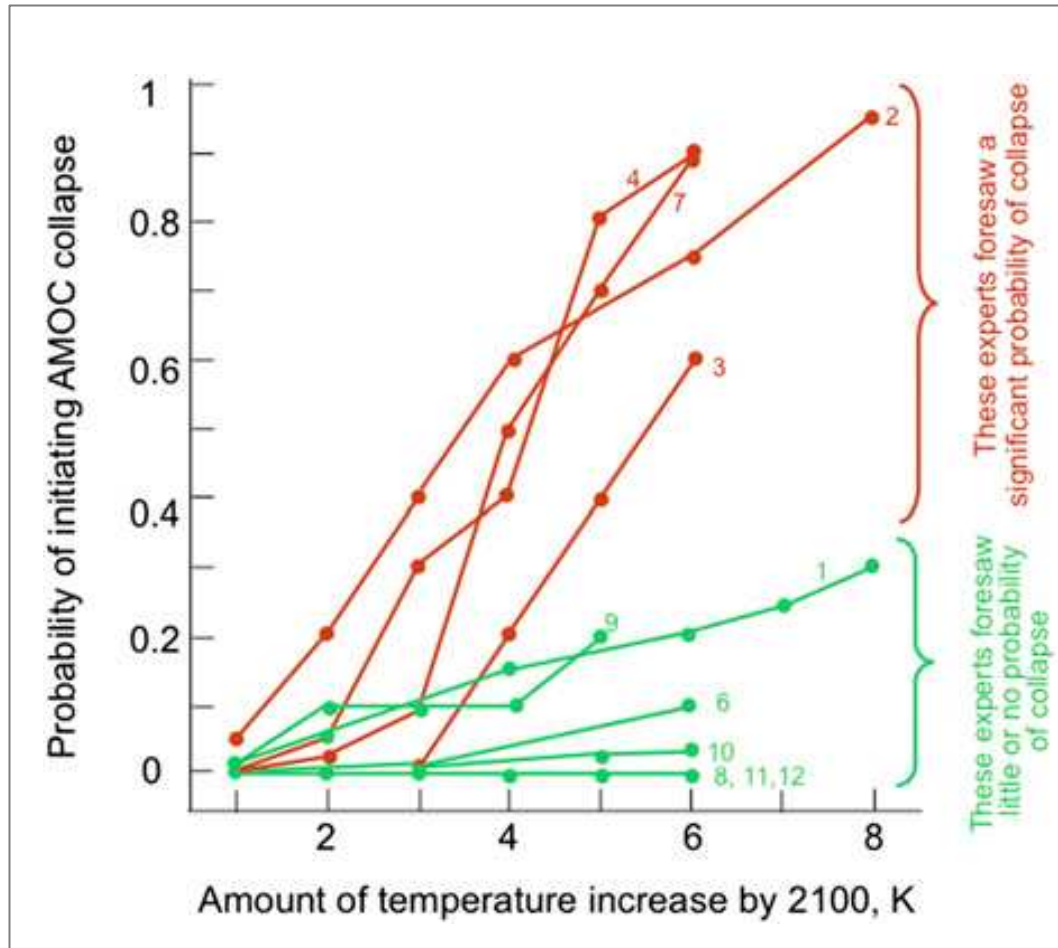


Fig. 8. Expert elicitation can be effective in displaying the range of opinions that exist within a scientific community. This plot displays clearly the two very different schools of thought that existed roughly a decade ago within the community of oceanographers about the probability “that a collapse of the AMOC will occur or will be irreversibly triggered as a function of the global mean temperature increase realized in the year 2100.” Each curve shows the subjective judgments of one of 12 experts. Four experts (2, 3, 4, and 7 in red) foresaw a high probability of collapse, while seven experts (1, 5, 6, 8, 9, 10, 11, 12 in green) foresaw little, if any, likelihood of collapse. Collapse was defined as a reduction in AMOC strength by more than 90% relative to present day. Figure redrawn from Zickfeld et al. (18).

(AMOC = Atlantic Meridional Overturning Circulation)

Expert elicitation and the Yucca Mountain nuclear waste depository
(Alley and Alley, “Too Hot to Touch – The Problem of High-Level Nuclear Waste, CUP, 2012)



Figure 19.2 Lathrop Wells cinder cone. Photograph by Greg Valentine.



FEDERAL LANDS IN SOUTHERN NEVADA

... In 1980, the first estimates of volcanic activity at the repository site put the annual probability at about 1 in 100 million. This was right at the Nuclear Regulatory Commission's cutoff point for inclusion in the TSPA, but not below it. By some accounts, 1 in 100 million is also roughly the same possibility as the ultimate low-probability, high-consequence event – global mass extinction from the impact of an asteroid or comet.

In the mid 1990s, DOE convened a panel of ten experts, mostly volcanologists, to conduct a formalized “ask-the-experts” approach to estimating the probability of volcanism and its uncertainty. The method, called expert elicitation, brings together a panel of experts and mathematically combines their individual estimates.

The goal is to obtain a probability distribution and range of uncertainty representative of the larger scientific community. Of course, the end result is affected by who serves on the panel, and the pool of qualified participants is not very large.

Using a formal nomination process, ten panel members were selected from a group of 70 scientists. Expertise mattered, but equally important were strong communication and interpersonal skills, as well as flexibility and impartiality.

The experts were asked to act as objective evaluators of the various theories. Their job was to listen to proponents of different positions and then weigh each of these theories in making their estimates.

After workshops and field trips to bring everyone up to speed, professional interviewers spent two days with each panel member extracting key information. Each of the ten experts independently arrived at an annual probability distribution for a volcanic event intersecting the repository. The average of these estimates was about 1 in 70 million, later revised to about 1 in 60 million. In the scheme of things, this was not far from the original 1 in 100 million estimate.

The expert elicitation did not end the debate. The results were challenged not only by the State of Nevada but also by scientists working for the NRC. Arguing that conservatism was needed, the NRC used an estimate of 1 event in 10 million in their assessments.

Still portending a rare event, the higher probability by NRC scientists came about, in part, from their assumption that faults and deep tectonic structures may provide pathways for the ascent of magma directly into Yucca Mountain. There was also disagreement about whether the time between eruptions was increasing or decreasing. Volcanism is known to be episodic. While most geoscientists consider the volcanism in the Yucca Mountain region to be waning, a few argued that we could be in the middle or end of a quiescent period. ...

Distr.: General
07 November 2019

English only

Economic Commission for Europe

Conference of European Statisticians

Work Session on Demographic Projections

Belgrade, 25–27 November 2019

Item No. 4 of the provisional agenda

Methodology

Using expert elicitation to build long-term projection assumptions

Note by Statistics Canada*

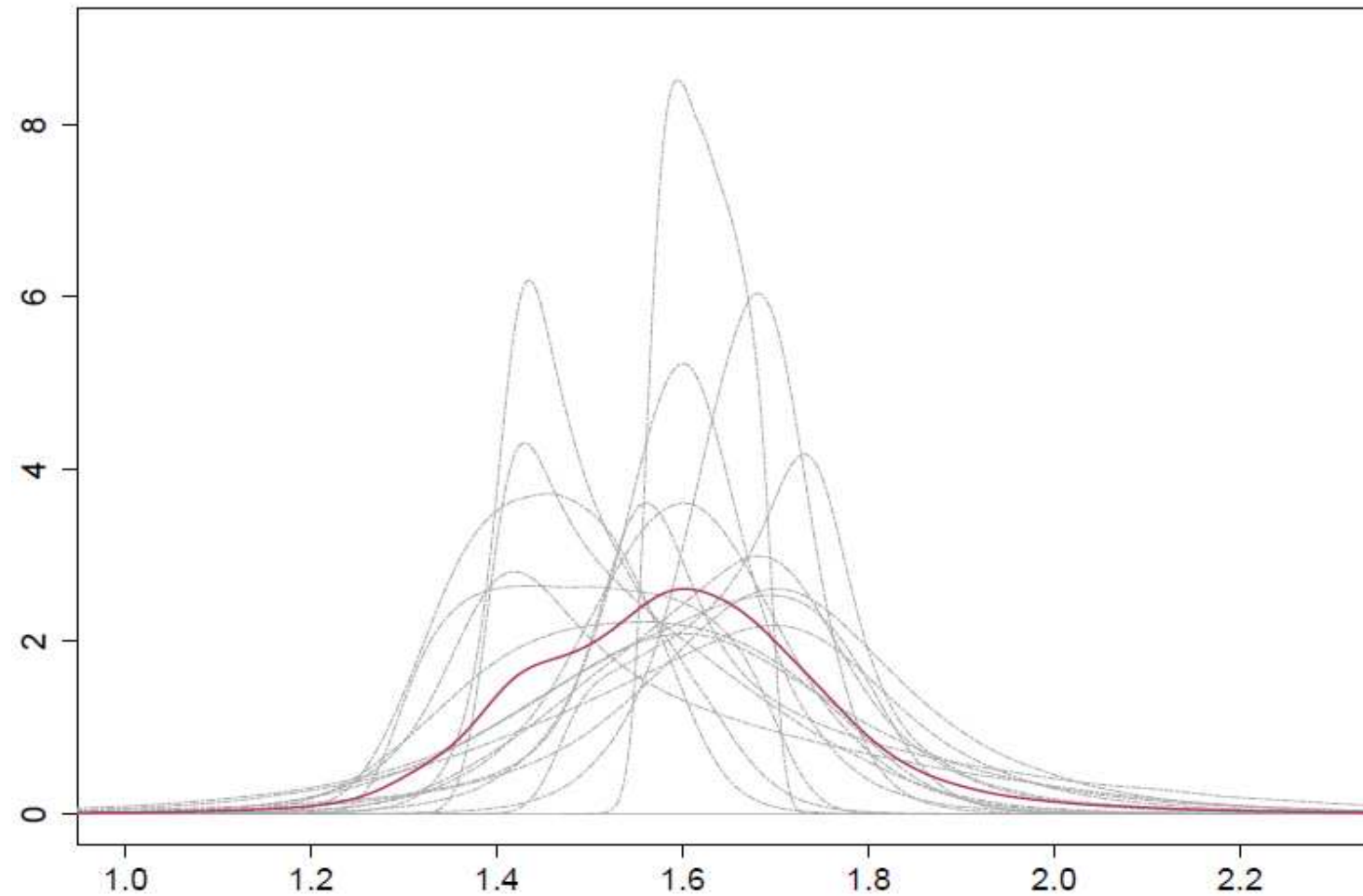
Summary

Expert judgment is valuable when there is either a lack of good data, insufficient knowledge about underlying causal mechanisms, or apparent randomness in trends. Most statistical agencies consult with experts in some manner prior to formulating their assumptions about the future. The development of probabilistic projections in recent years has triggered an increasing interest in formal methods of expert elicitation.

In 2013, Statistics Canada adopted a rather formal approach to expert elicitation by asking experts to provide estimates of ‘most likely’ values for a series of demographic indicators, along with corresponding 80% prediction intervals. Motivated in large part by a will to improve the characterization of uncertainty, Statistics Canada has recently refined its consultation process, delving further into the science of Expert knowledge elicitation.

In this paper, we describe the expert elicitation protocol used by Statistics Canada in 2018 to inform the development of assumptions. This information may be useful for demographers looking to adopt a formal approach to eliciting expert judgements, and can be pertinent in the context of producing probabilistic projections, where it is necessary – but often difficult – to obtain plausible estimates of uncertainty for components of population growth.

Figure 1 Period total fertility rate, Canada, 2043: Individual expert probability distributions (grey dashed curves) and aggregate mixture distribution (red curve) of the 17 fertility respondents of *the 2018 Survey of Experts on Future Demographic Trends*



Source: Statistics Canada, Demography Division.

Possible pitfalls when dealing with experts are described here:

<https://www.theatlantic.com/magazine/archive/2019/06/how-to-predict-the-future/588040/>

Hunches based on previous knowledge are not (unfortunately) a substitute for real data, see, e.g., here:

<https://www.theatlantic.com/science/archive/2021/03/americas-coronavirus-catastrophe-began-with-data/618287/>

WILLIAM M. ALLEY and ROSEMARIE ALLEY

TOO HOT TO TOUCH

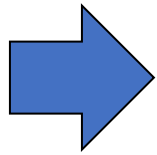
THE PROBLEM OF HIGH-LEVEL NUCLEAR WASTE



This book contains an excellent discussion on expert elicitation when handling the problem of nuclear waste.

Bayesian estimates often require the evaluation of complex integrals.

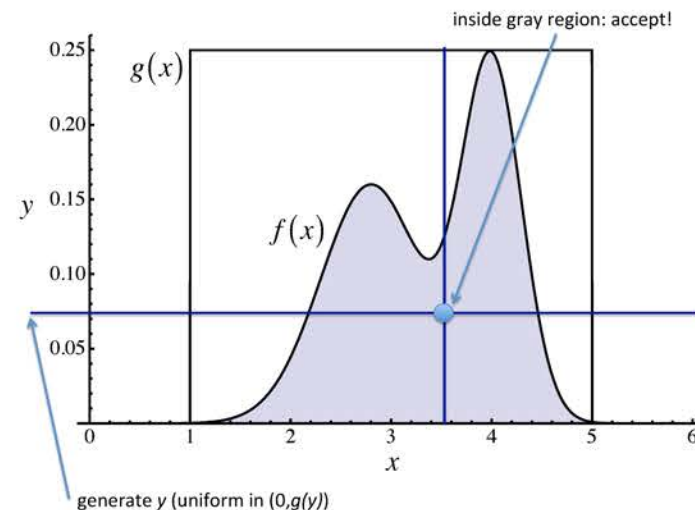
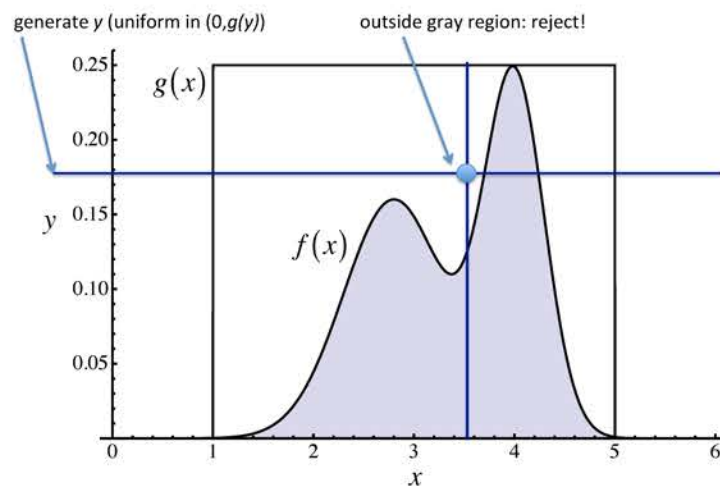
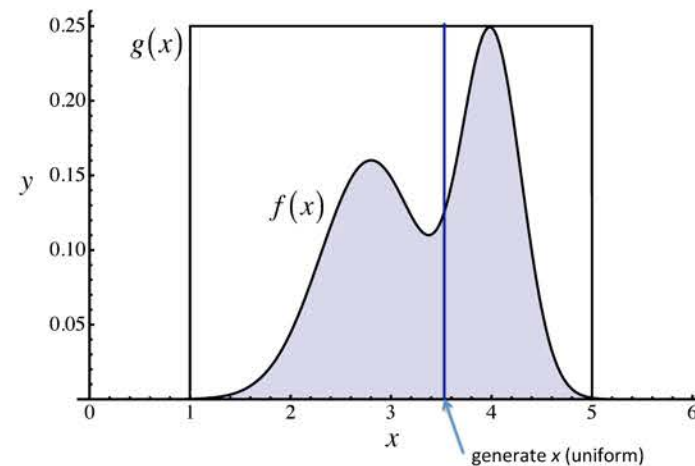
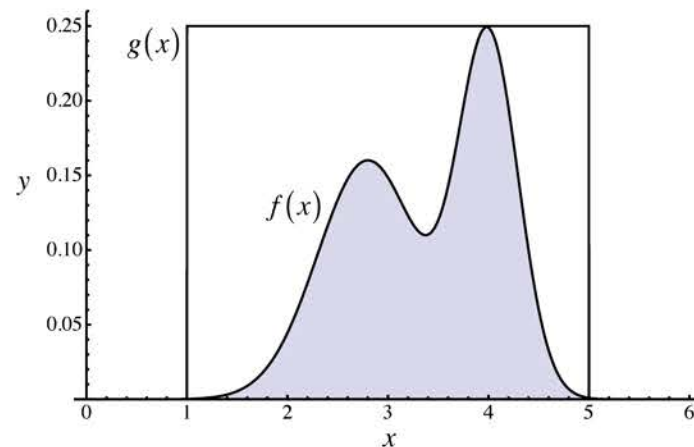
Usually these integrals can only be evaluated with numerical methods.



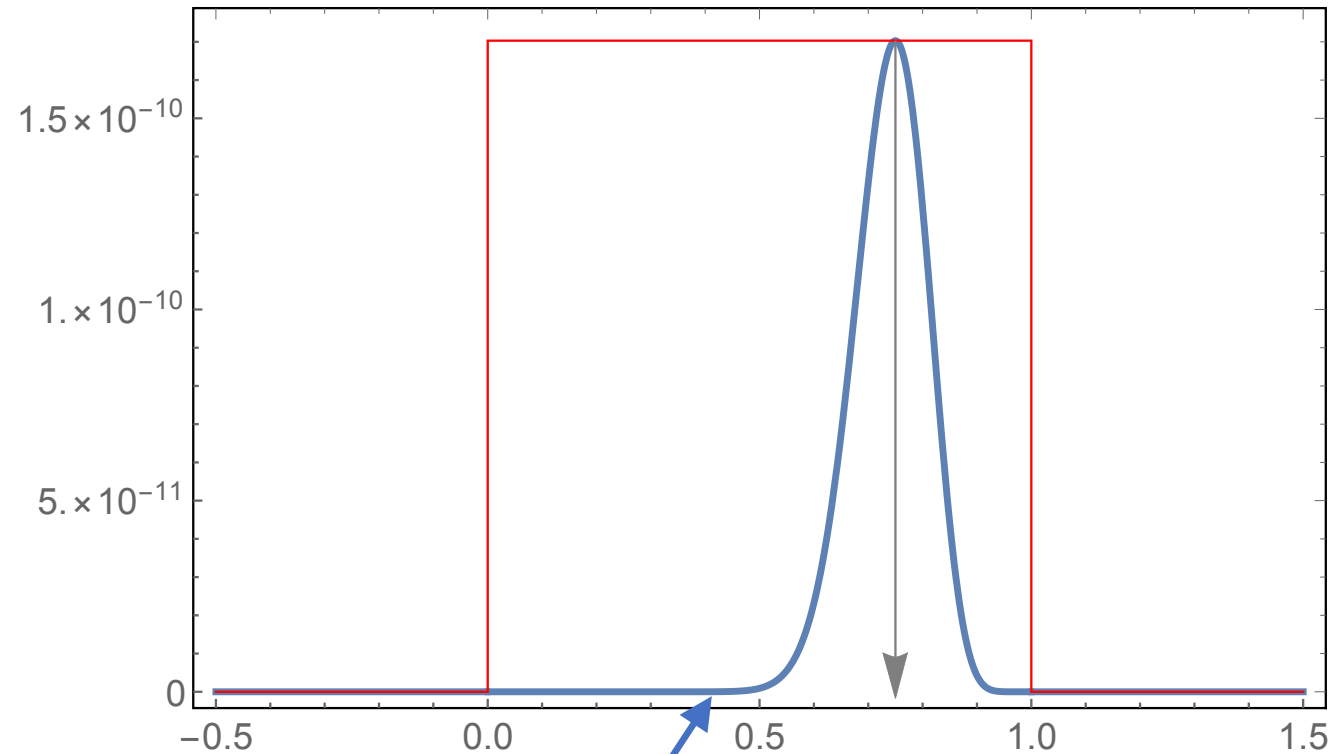
enter the Monte Carlo methods!

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



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

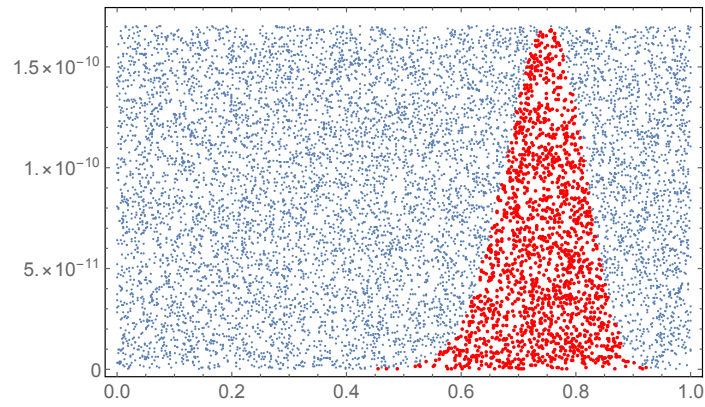
normalized distribution

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

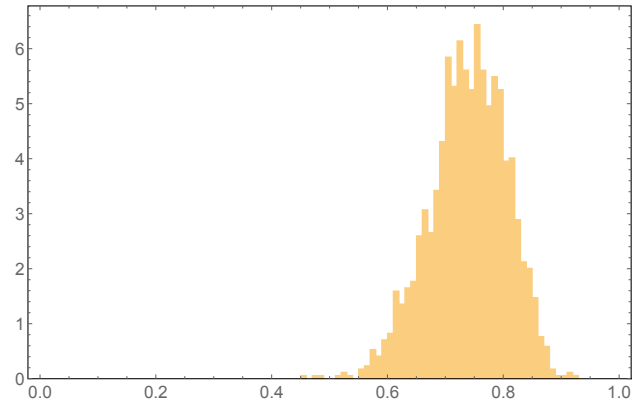
unnormalized distribution

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

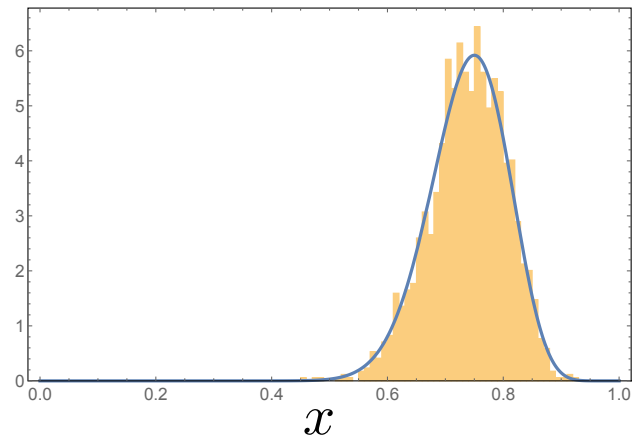
modal value



generated pairs
(red = accepted pairs)



normalized histogram of the
accepted x's

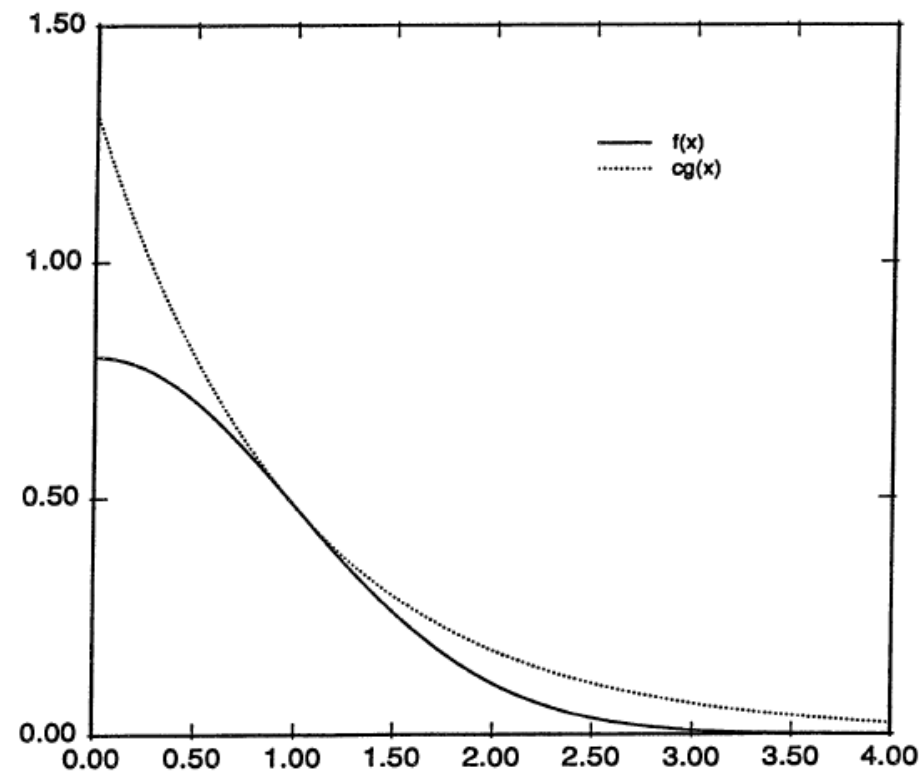


comparison with the plot of the
normalized beta distribution

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)

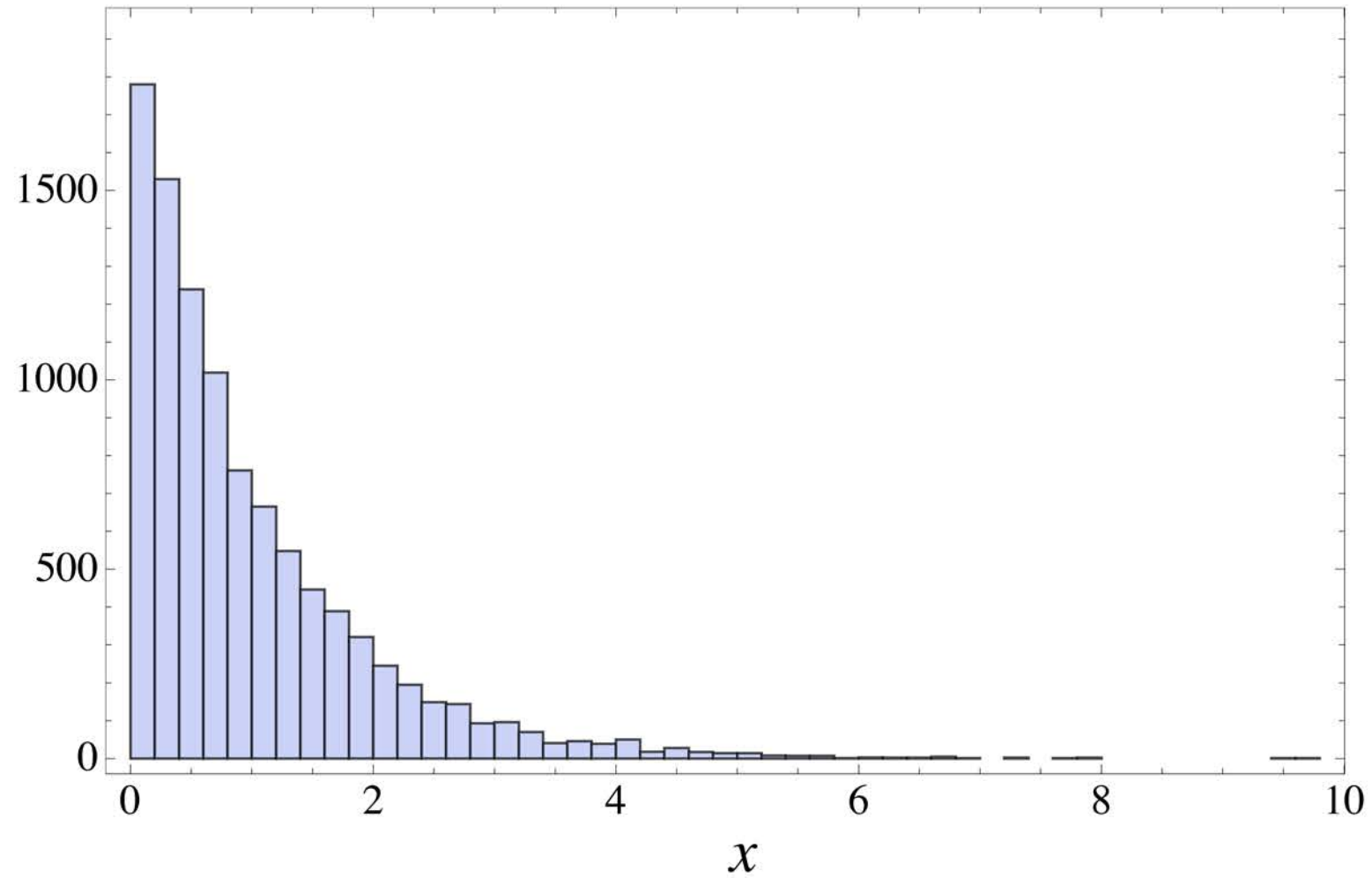
$$f(x) = \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) \quad x \geq 0$$

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

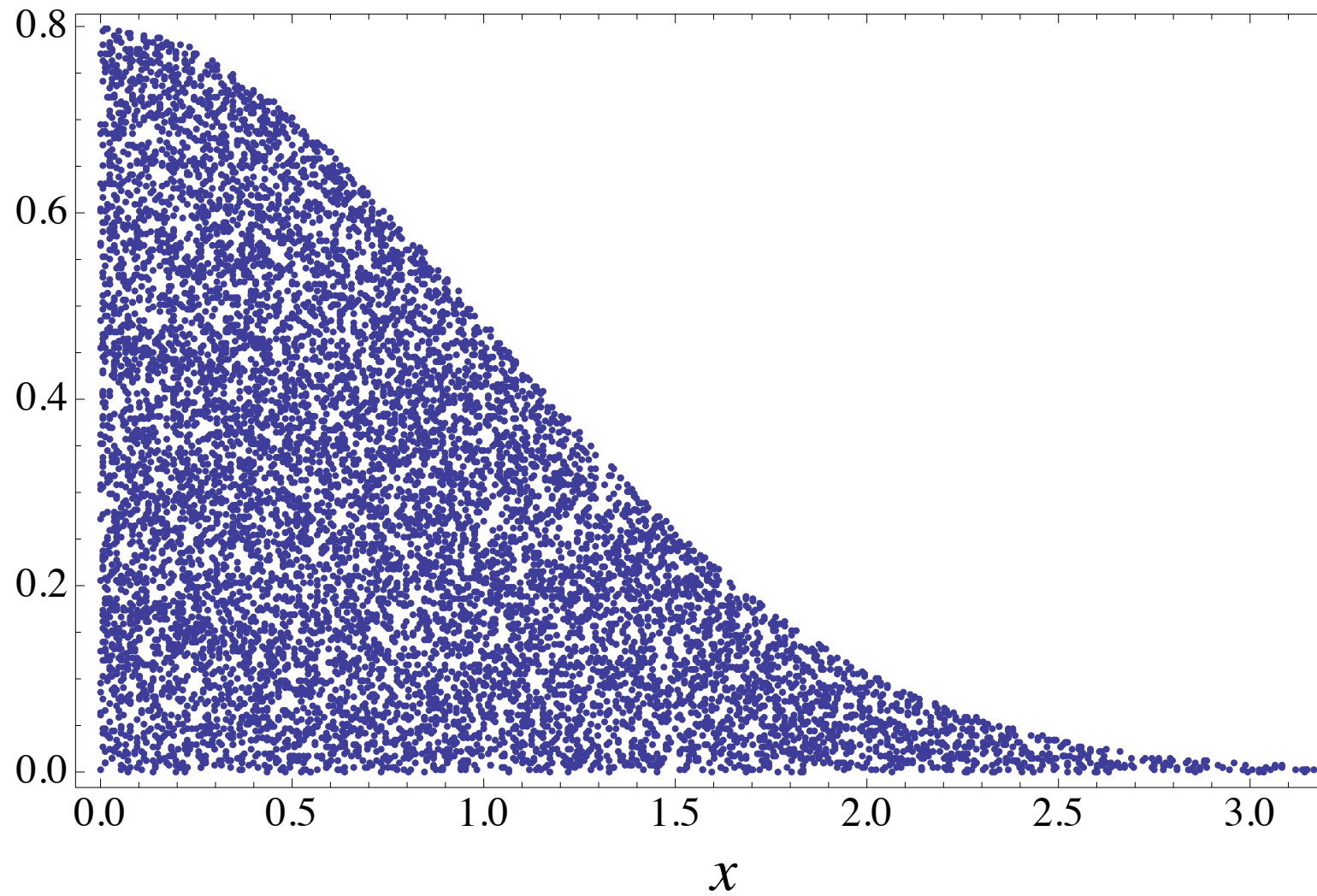
$$\Rightarrow \begin{cases} f(x) = c g(x) \\ f'(x) = c g'(x) \end{cases} \Rightarrow \begin{cases} \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) = c \exp(-x) \\ x \sqrt{\frac{2}{\pi}} \exp\left(-\frac{x^2}{2}\right) = c \exp(-x) \end{cases}$$

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

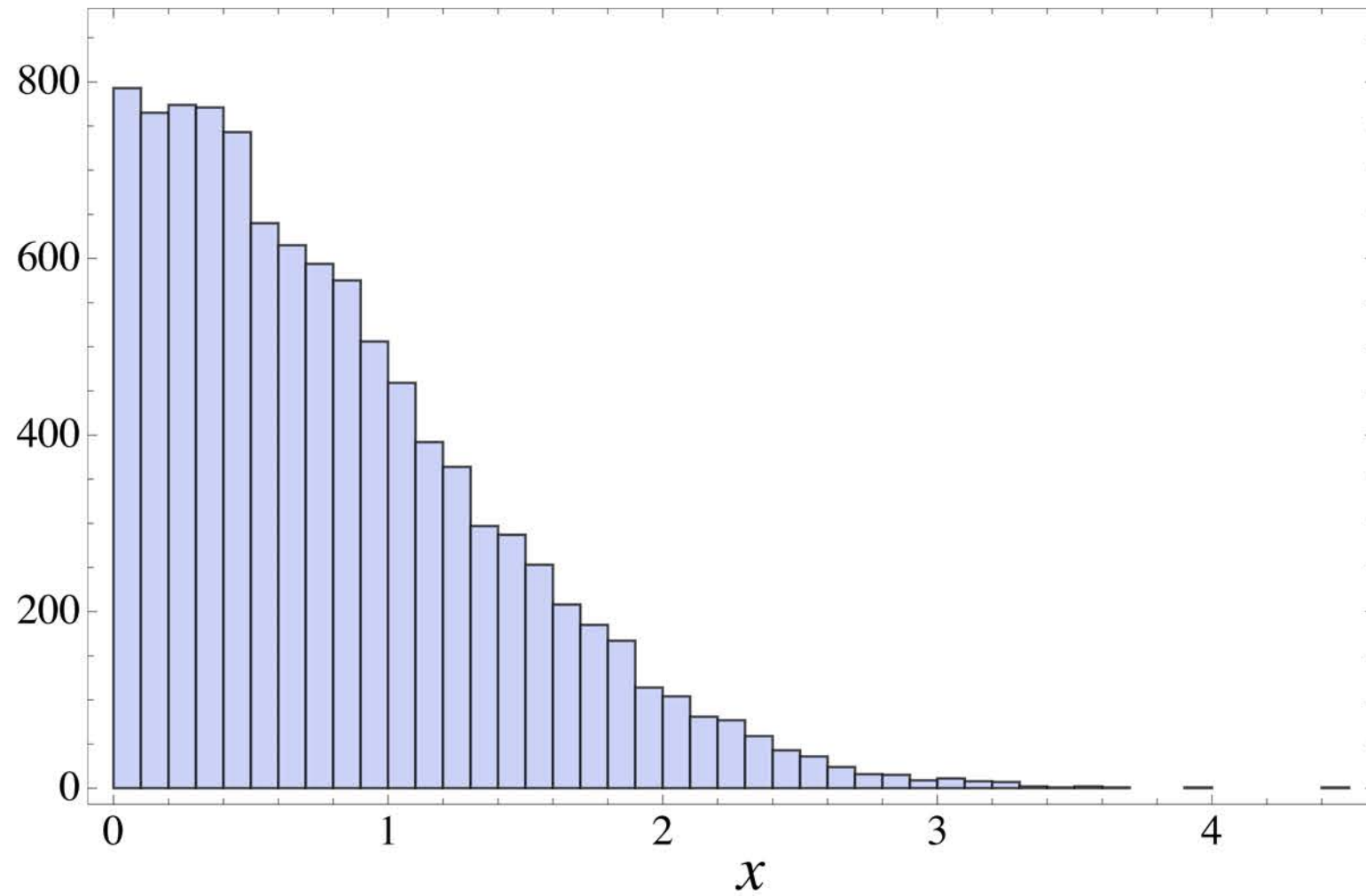
Exponentially distributed values



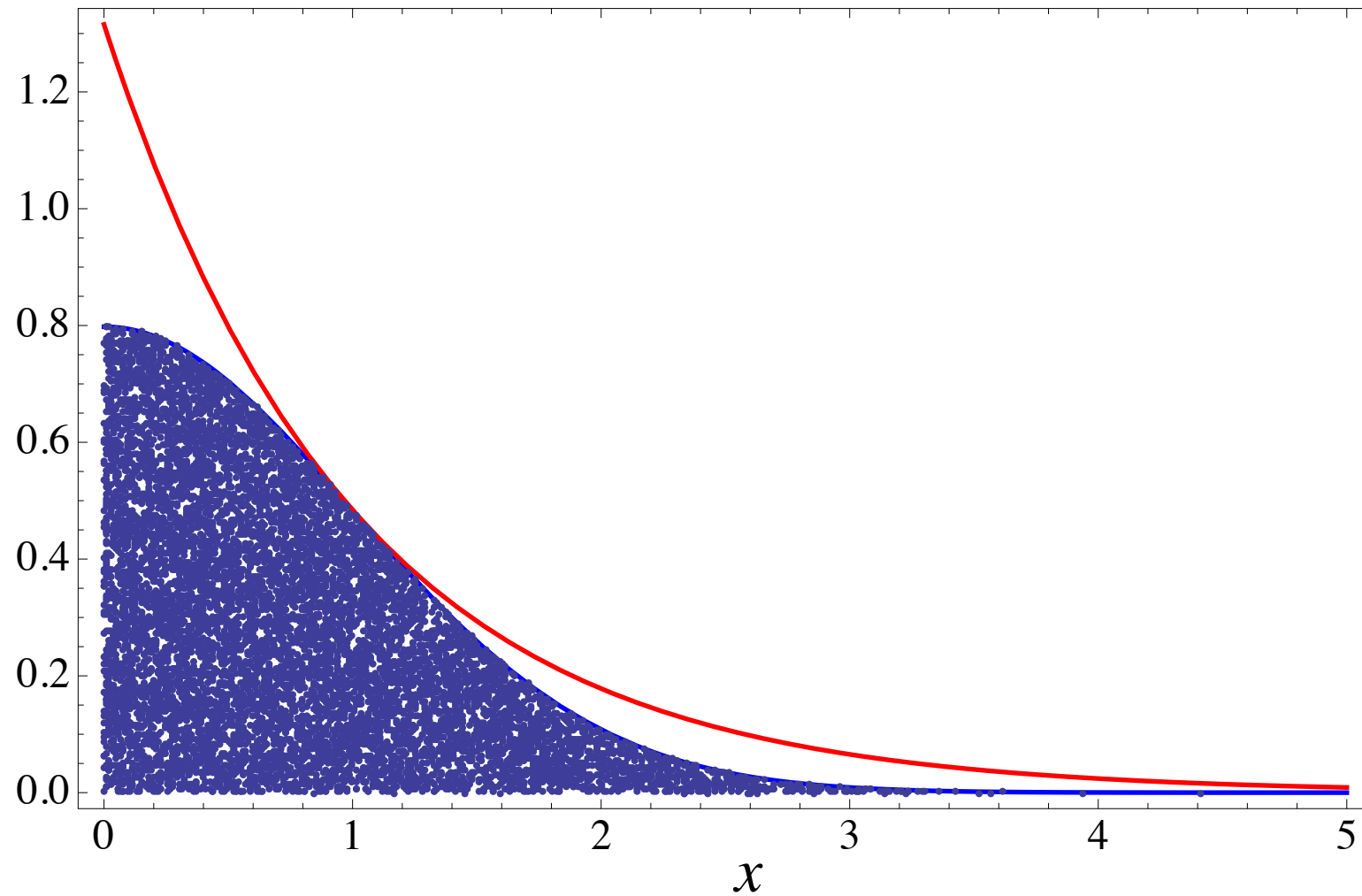
A/R accepted values (10000 accepted sample pairs)

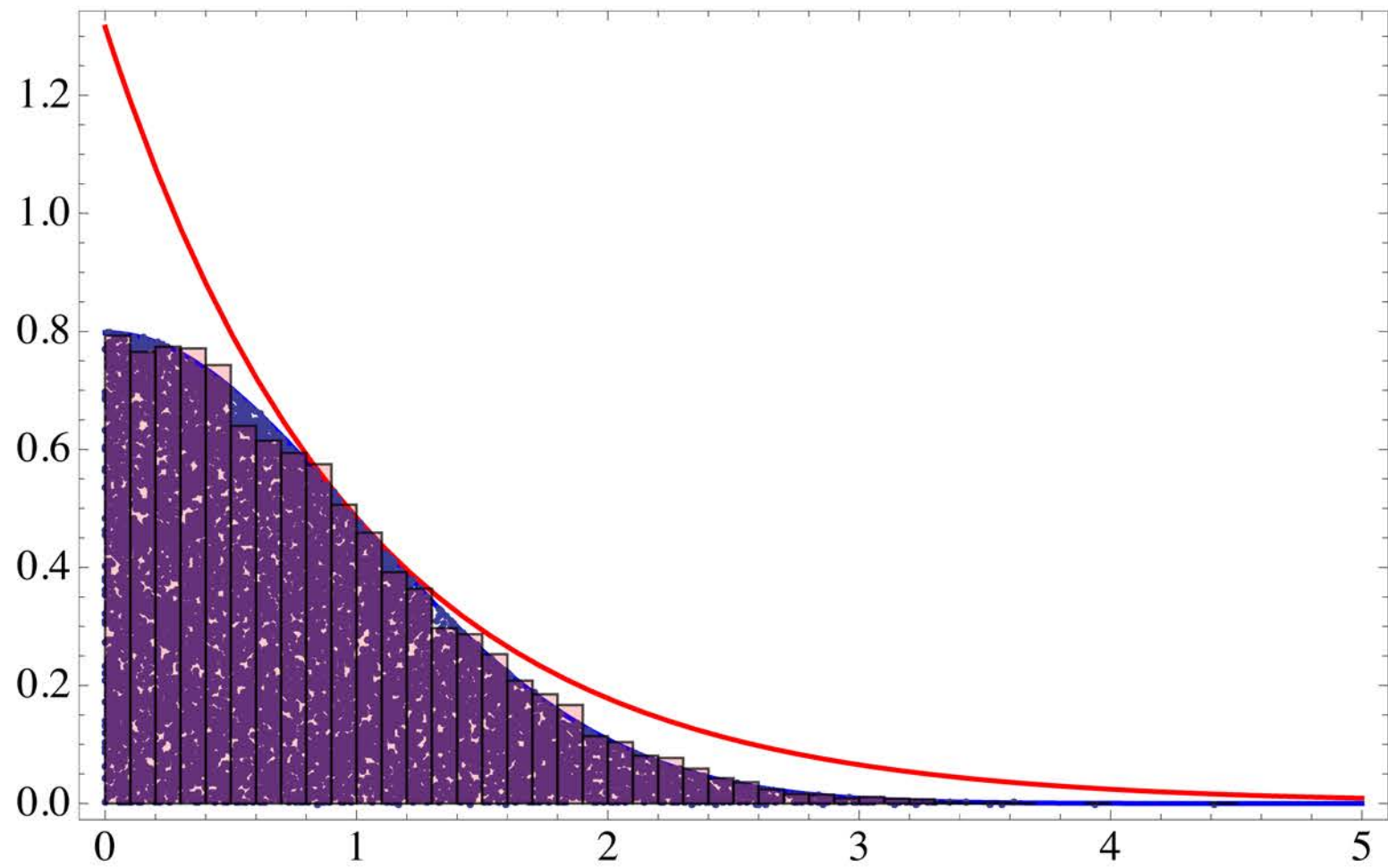


Histogram of accepted x values



Comparison with the original distributions





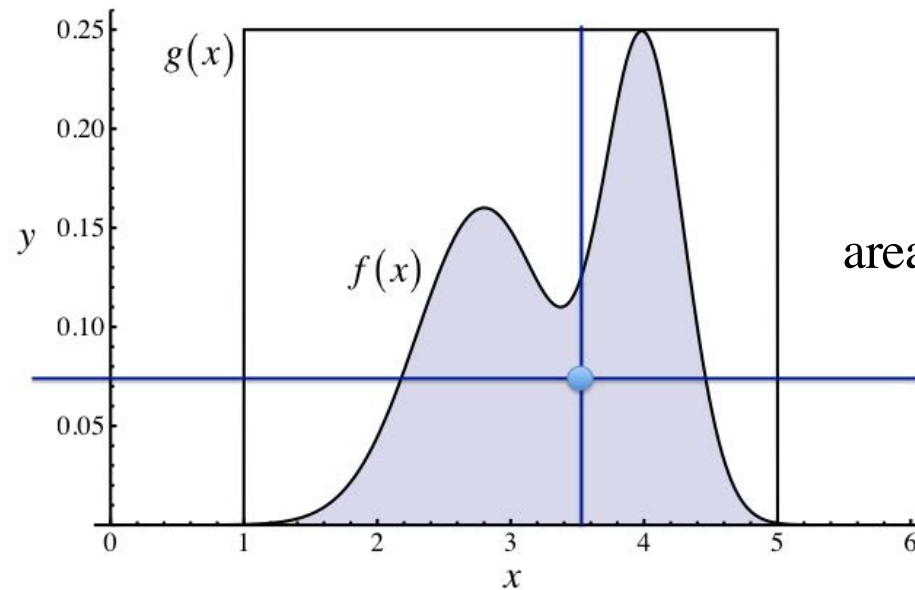
Short summary:

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 ...)

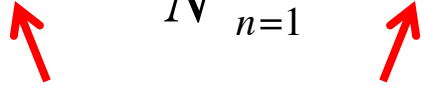
Now notice that in this method we generate pairs of real numbers that are uniformly distributed between $f(x)$ and the x-axis, therefore we can use these pairs to estimate the total area under the curve

(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 ...

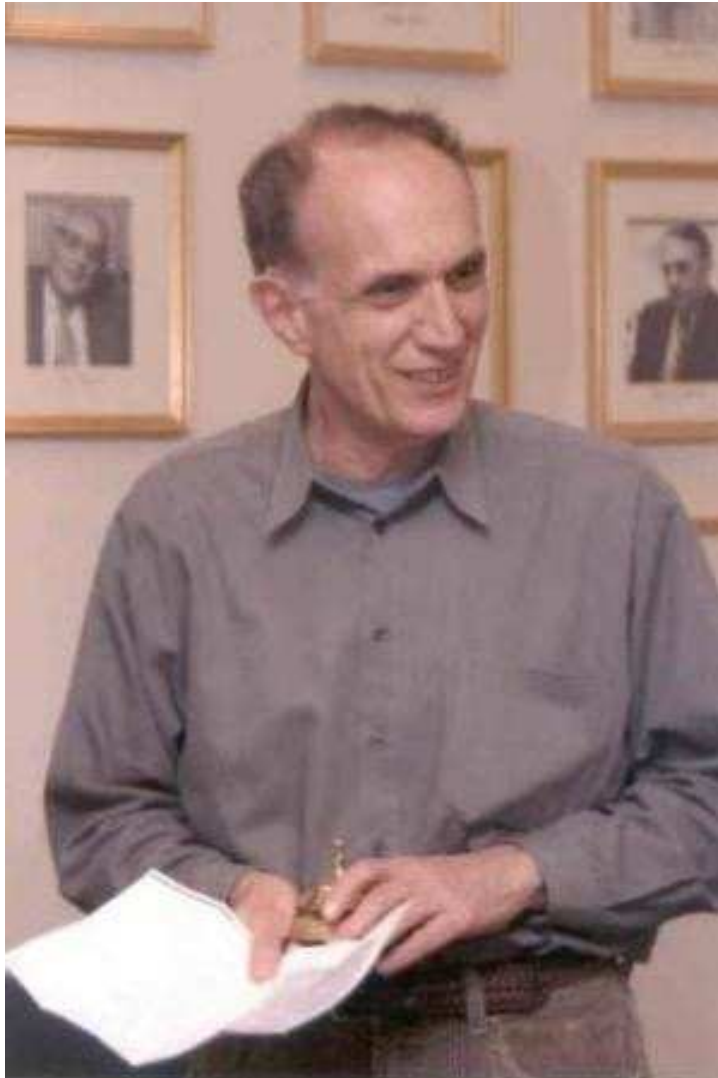
therefore we use this ...

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

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

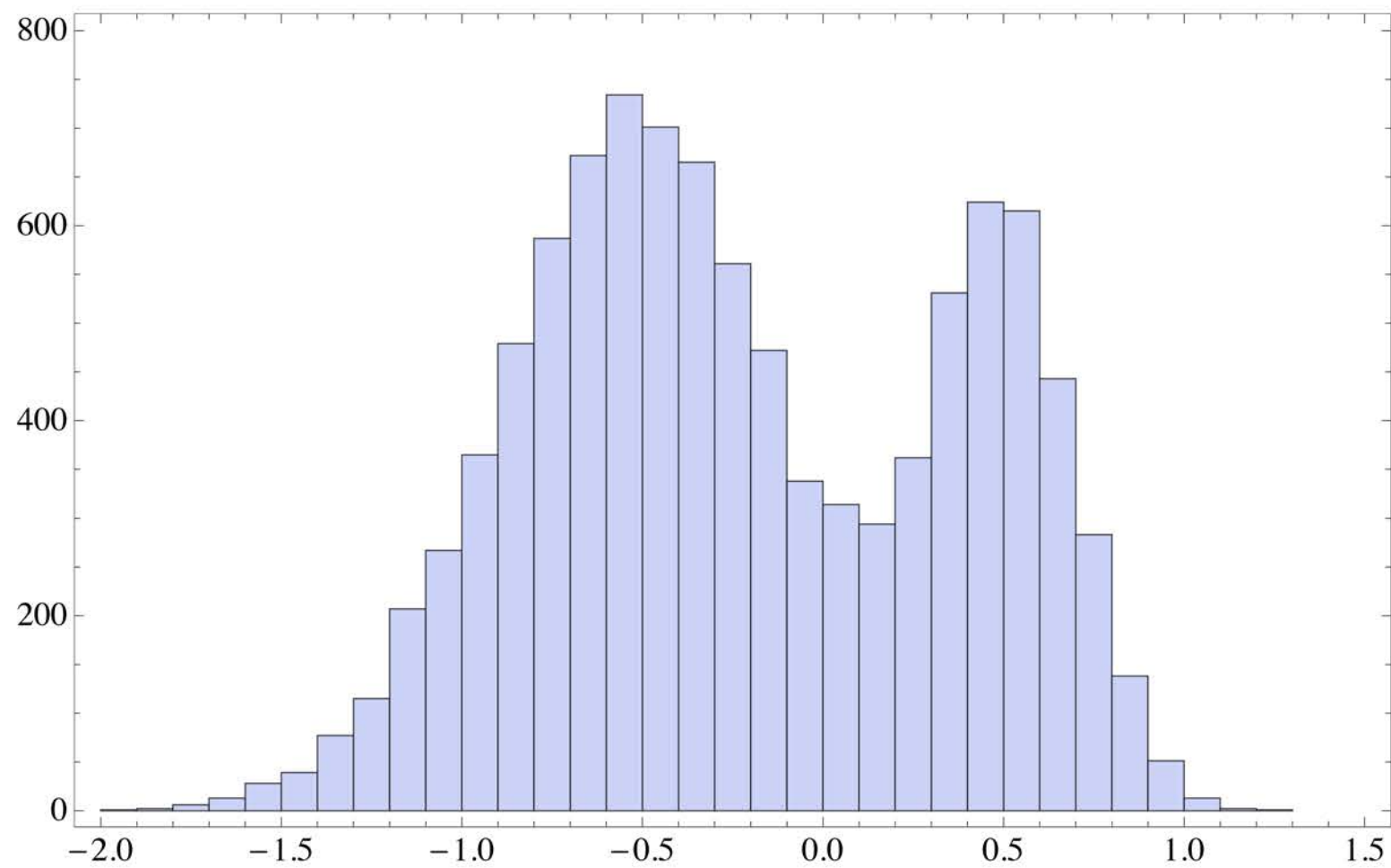
These methods are still not very efficient and there is a better alternative, the Markov Chain Monte Carlo method

3. Bootstrap (B. Efron, 1977)

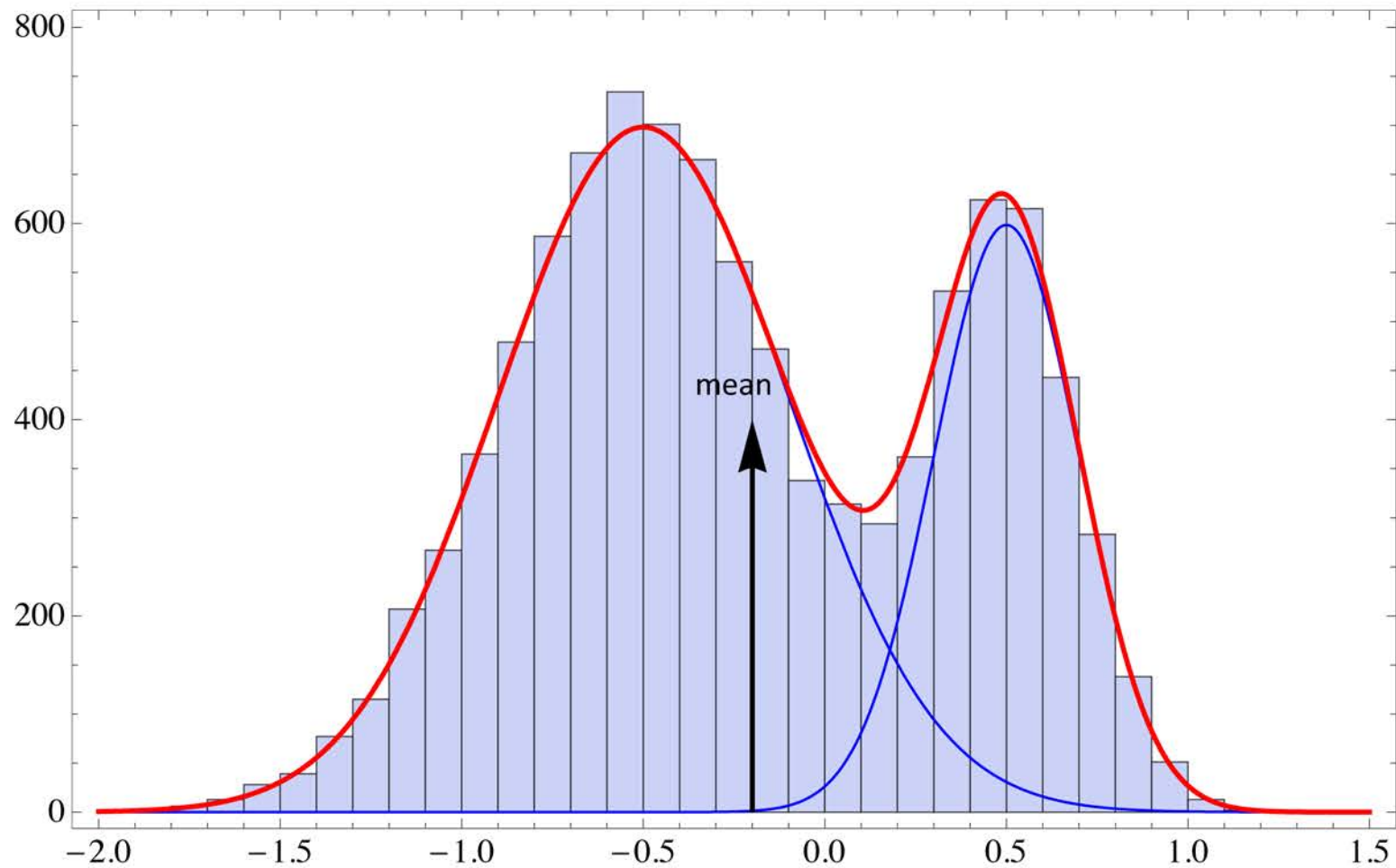


The bootstrap method is a resampling technique that helps calculate many statistical estimators

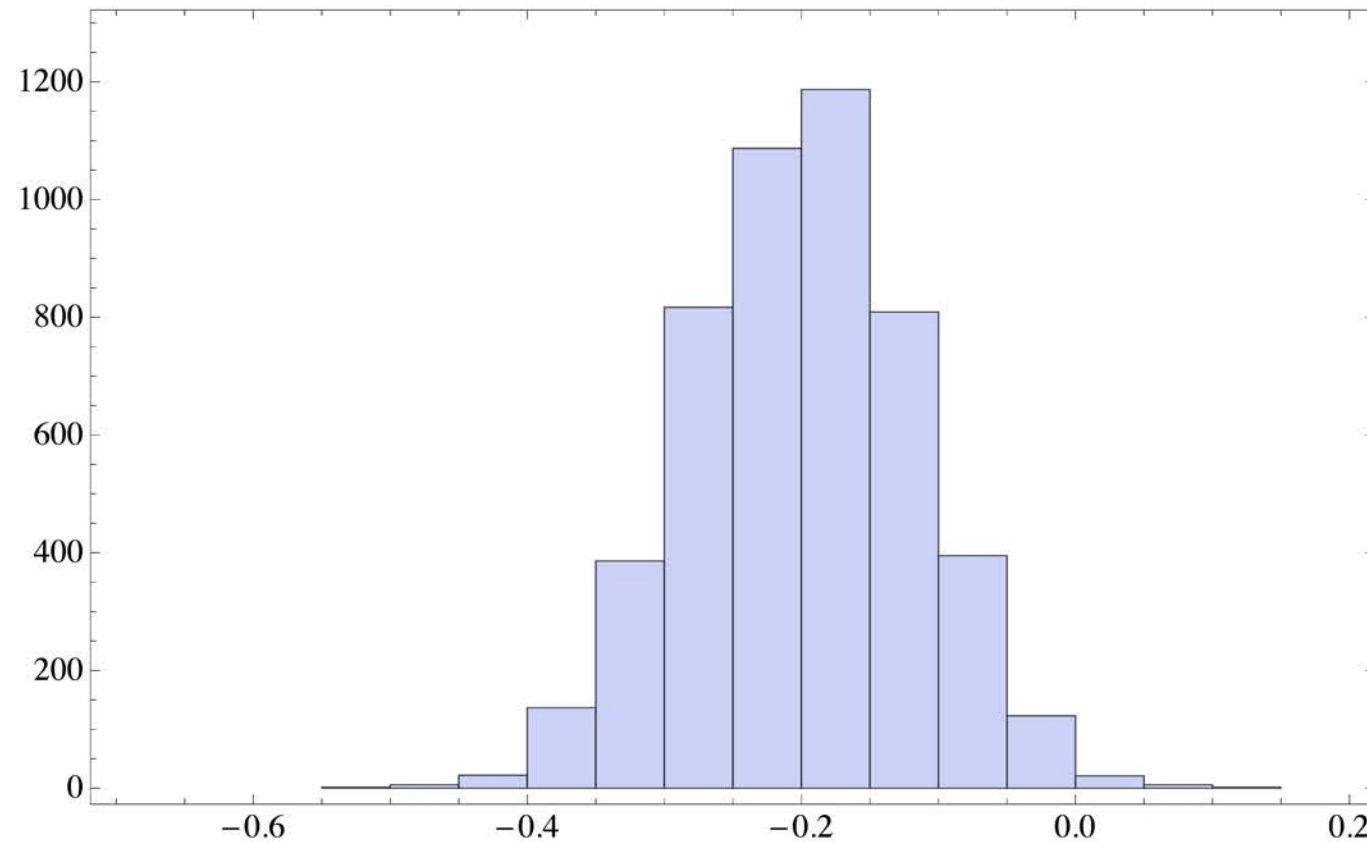
consider the distribution of a set of measurements



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

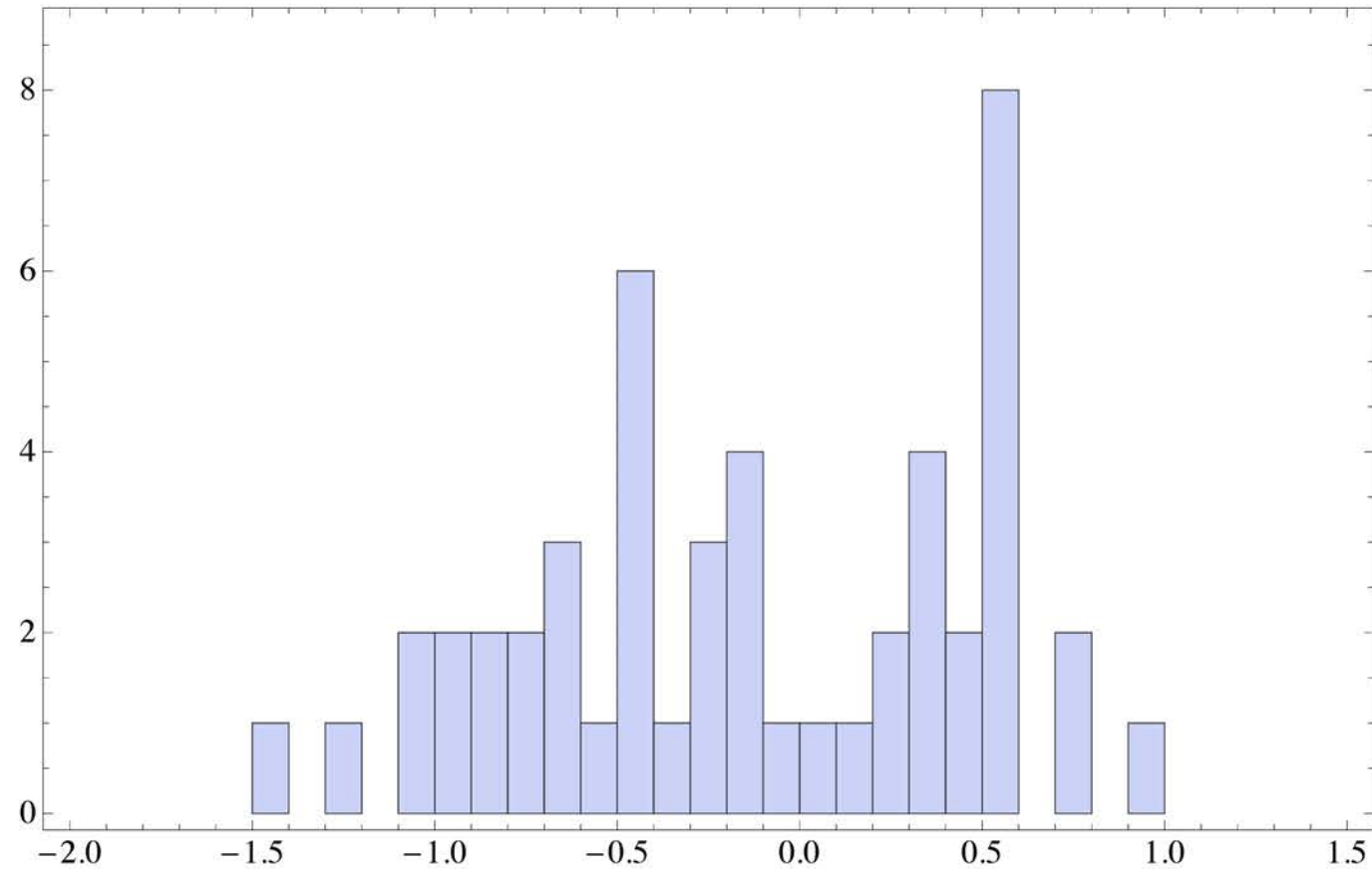


distribution of mean value obtained from 5000 sets of data
(sample size = 50)



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

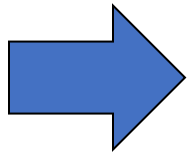
example: single dataset (same size as before, 50 measurements)



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

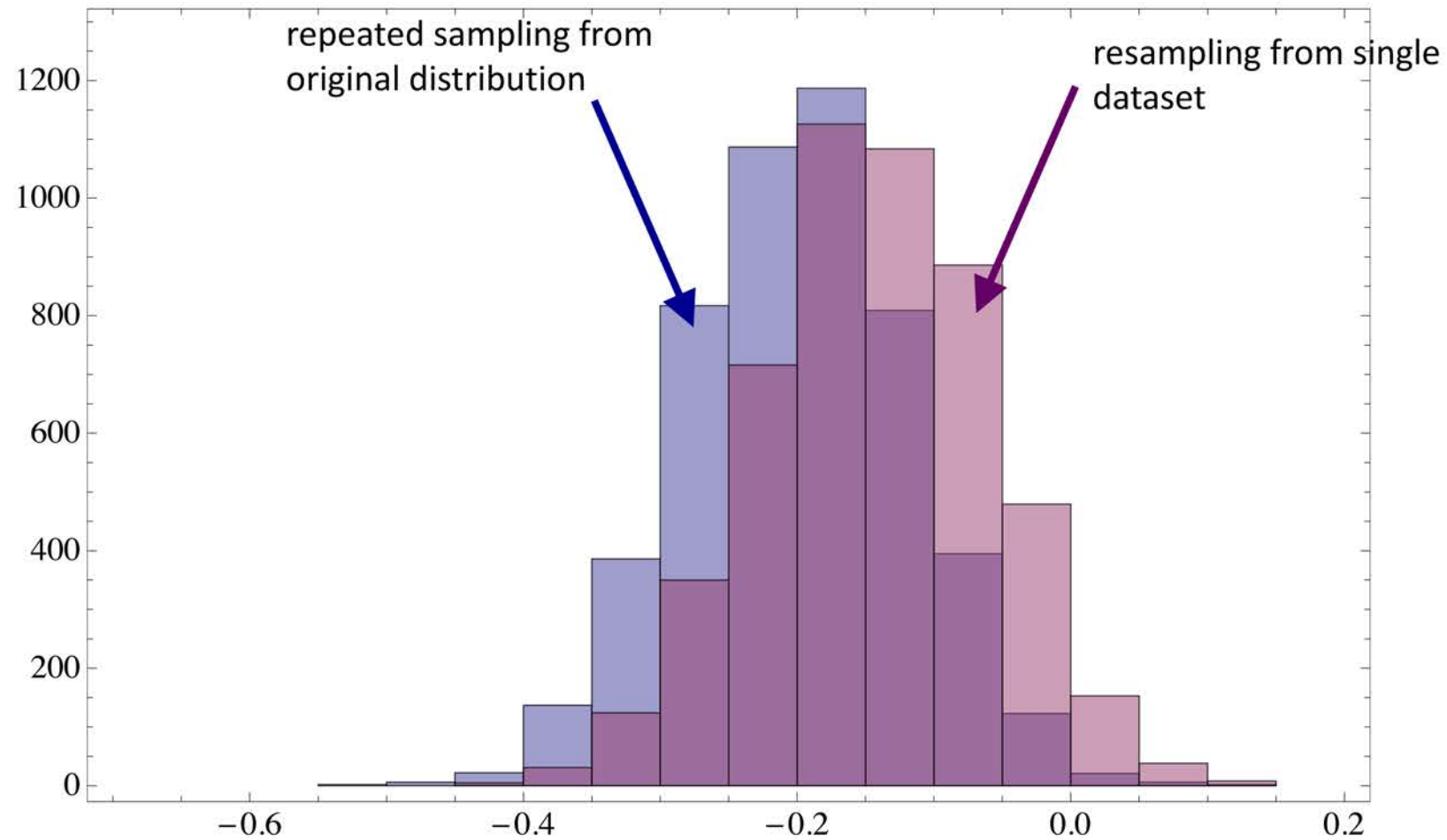
Bootstrap recipe:

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 ± 0.0813632

mean from resampling dataset (size = 50): -0.142699 ± 0.0838678

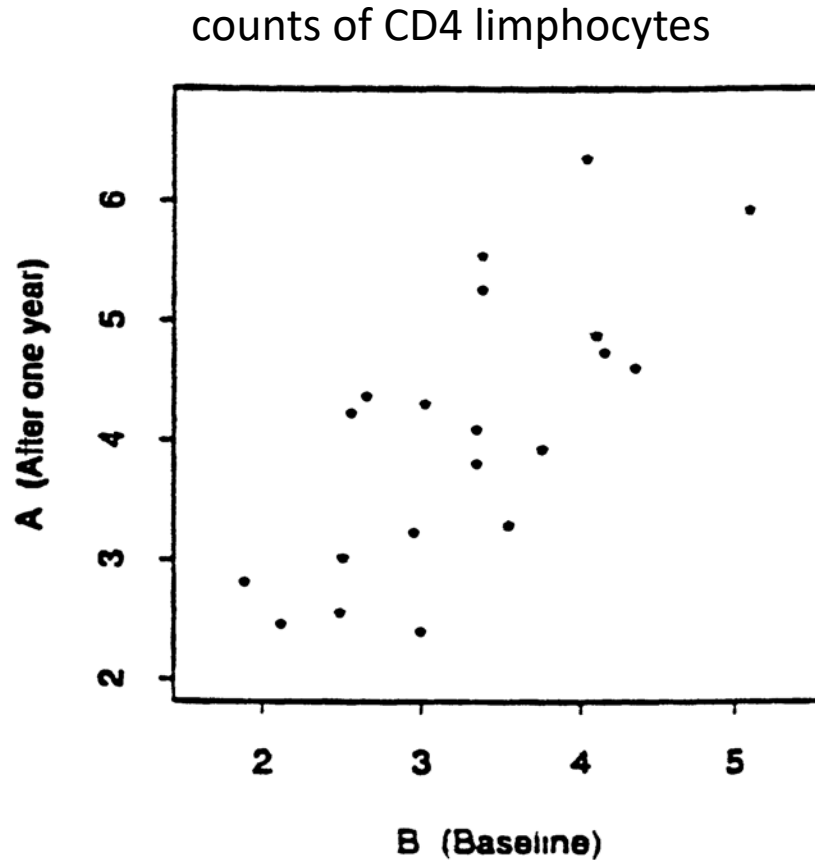


FIG. 1. *The cd4 data; cd4 counts in hundreds for 20 subjects, at baseline and after one year of treatment with an experimental anti-viral drug; numerical values appear in Table 1.*

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

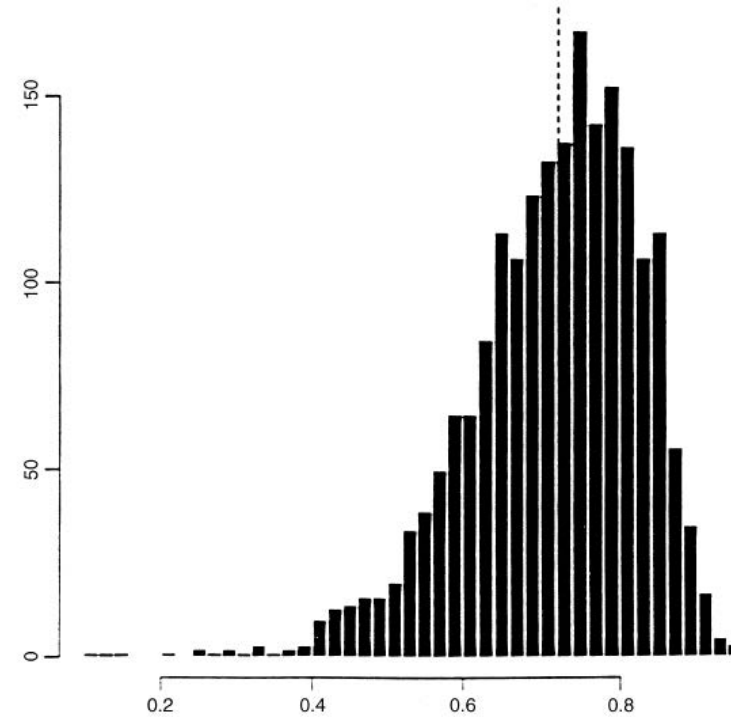


FIG. 3. *Histogram of 2,000 bootstrap correlation coefficients; bivariate normal sampling model.*

bootstrap estimate of correlation
coefficient distribution

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)

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

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)

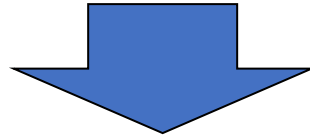
Example (McCullagh & Nelder): take two sets of binomially distributed independent random variables X_{i1} and X_{i2} ($i=1,2,3$)

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

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

The observed random variables are the sums

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



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

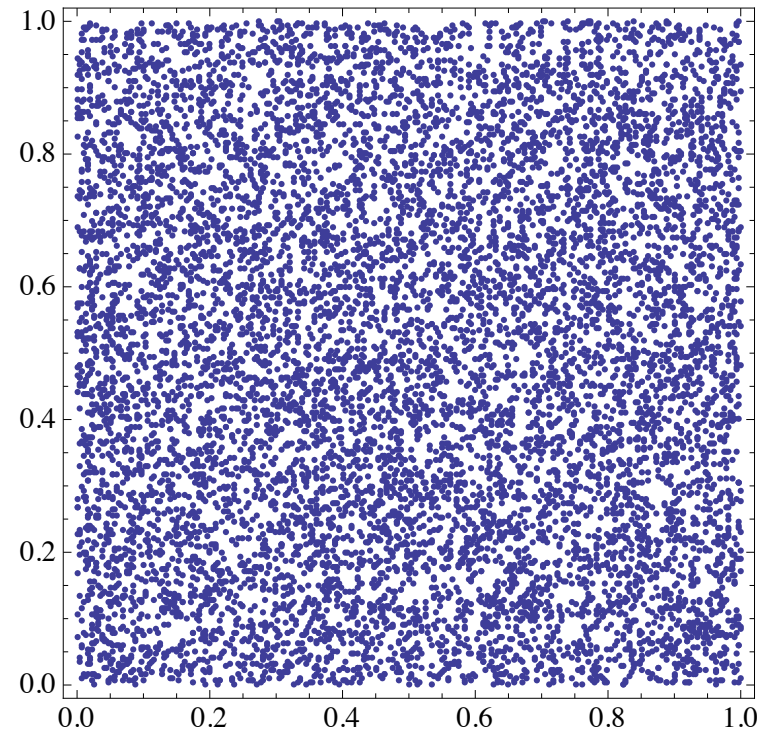
$$\max(0, y_i - n_{i2}) \leq j_i \leq \min(n_{i1}, y_i)$$

Sample data

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

Example of implementation in *Mathematica*

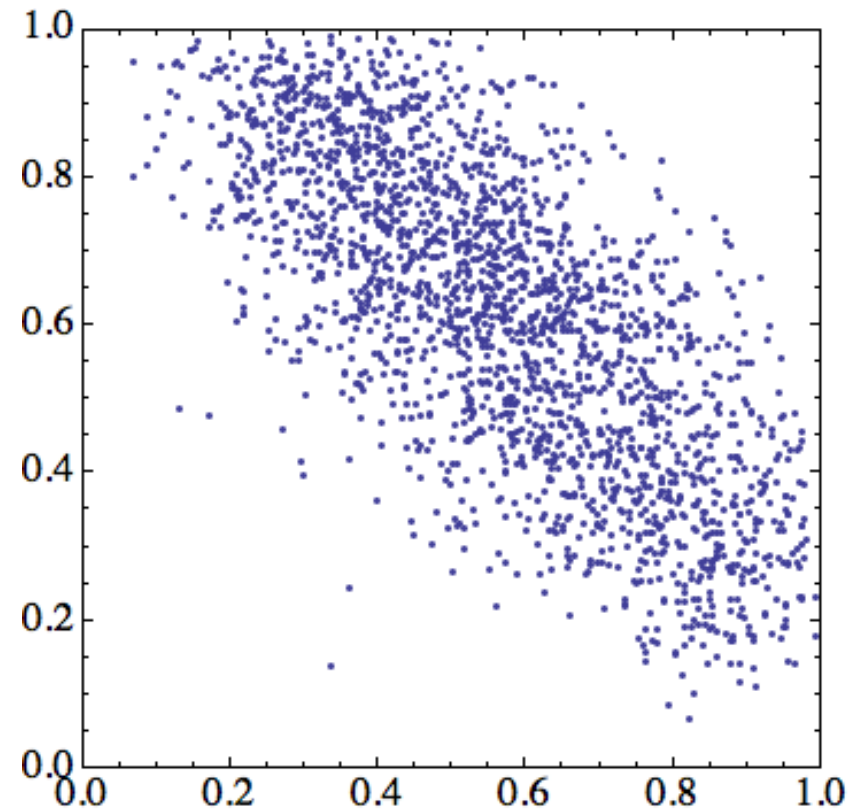
```
n1 = {5, 6, 4};  
n2 = {5, 4, 6};  
yi = {7, 5, 6};  
  
Clear[likelihood];  
likelihood[th1_, th2_] :=  
  Product[Sum[Binomial[n1[[i]], j] * Binomial[n2[[i]], yi[[i]] - j] * th1^j * (1 - th1)^(n1[[i]] - j) *  
    th2^(yi[[i]] - j) * (1 - th2)^(n2[[i]] - yi[[i]] + j), {j, Max[0, yi[[i]] - n2[[i]], Min[n1[[i]], yi[[i]]]}],  
    {i, 1, 3}];  
  
ns = 10000;  
th = Table[{RandomReal[], RandomReal[]}, {ns}];
```



prior distribution (uniform in 2D
parameter space)

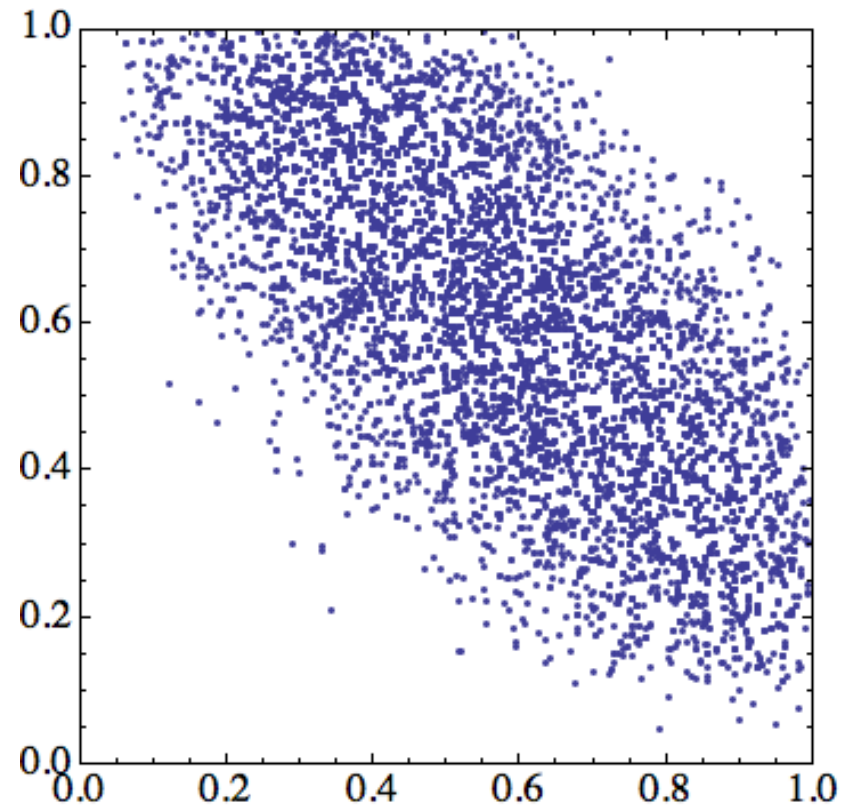
Posterior as a resampled prior using acceptance-rejection

```
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++]
```

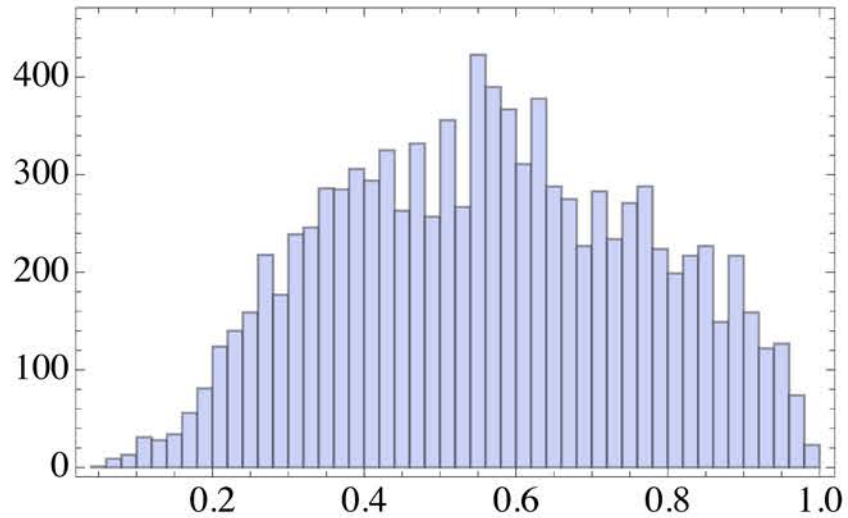


Posterior as a resampled prior using weighted bootstrap

```
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]]];  
]
```

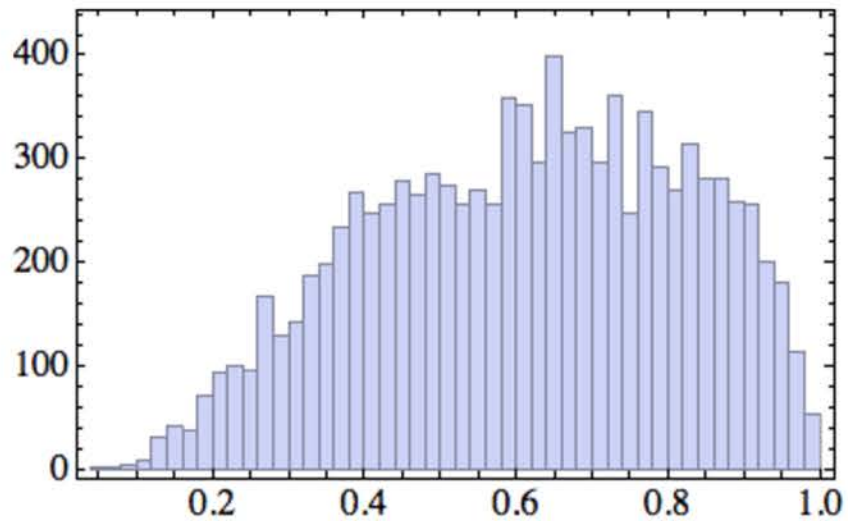


The resampled points are representative of the posterior distribution and can be used to evaluate any sample estimate



Marginalized distribution of θ_1

Sample mean: 0.564 ± 0.002



Marginalized distribution of θ_2

Sample mean: 0.613 ± 0.002

... 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