Probability in a world teeming with data: a neo-Baroque perspective on contemporary belief-events

Abstract

Since its Baroque invention [@hacking_emergence_1975], probability has been a double-sided coin. On one side, it concerns degrees of belief (the so-called 'subjective' view), and on the other side, frequencies, or how often things happen in the world (the so-called 'objective' view). In the last few centuries, one side of this coin has come up more often – the frequencies version of probability. Yet, as many historians of statistics, and statisticians themselve recognise, probability as degree of belief has never disappeared. It has only occurred less often, and been less often the object of belief. Indeed, this ineluctable entwining of belief and events, of subjective-objective, seems quintessentially Baroque in its interweaving and folding together of inside and outside. Drawing on both histories of statistics, and Gilles Deleuze's understanding of monads as 'simple, inverse, distribution numbers' [@deleuze_fold_1993], this paper examines the resurgence of the probability as degree of belief in the face of a world seemingly teeming with data. It argues that in the last few decades of statistical practice associated especially with 'Bayesian inference' and the techniques of Markov Chain Monte Carlo (MCMC) simulation, we see a re-configured and super-imposed concept of probability taking shape. As these practices pervade diverse scientific fields, commerce, government and industry, we might be seeing a different epistemic materialisation taking shape in which beliefs and events are less separate. On the contrary, through computation, subjective belief is exteriorised in simulated events, and a certain staging of events are reshaped as updateable beliefs.

Introduction

In the US Presidential elections of November 2012, the data analysis team supporting the re-election of Barack Obama were said to be running a statistical model of the election 66,000 times every night [@scherer_how_2012]. Their model, relying on polling data, records of past voting behaviour, and many other demographic features, was guiding tactical decisions about everything from where the presidential candidate would speak to the telephone calls that targeted specific groups of swing voters. In the media reports, the statistical model was favourably compared to the almost equally well-funded data analysis supplied to the Republican candidate, Mitch Romney. Widely reported in television news and internationally in print media (*Time*, *New York Times*, *The Observer*), the outstanding feature of Obama's re-election seems to me to be the figure of 66,000 nightly model runs.

Why so many thousand runs? This question was not addressed in the media

reports, nor surprisingly, addressed in the online discussion on blogs and other online forums that followed. This paper seeks to provide an answer to question of the power of repetition associated with data and models today. I only present this example as one amongst many recent illustrations of the power attributed to data. The answer is to be found, I suggest, in probability. Hardly ever discussed in media accounts of the growth of big data, certain shifts in the role played by probability changing the meaning and value of data as such, and hence, everything that depends on data.

In exploring recent mutations in probability, a Baroque perspective is not only useful but perhaps essential. Summarising his own account of the emergence of probability, the philosopher and historian Ian Hacking writes:

> I claimed in *The Emergence of Probability* that our idea of probability is a Janus-faced mid-seventheenth-century mutation in the Renaissance idea of signs. It came into being with a frequency aspect and a degree-of-belief aspect [@hacking_taming_1990, 96].

Indeed, in the work from 1975, Hacking, writing largely prior to the shifts in probability practice I discuss, claims that there was no probability prior to 1660 [@hacking_emergence_1975]. Not only is probability a Baroque invention, the fundamental instability that permits ongoing mutations in the concept has a distinctively Baroque flavour in the way that it combines something happening in the world with something that pertains to subjects. There is nothing controversial in this claim. Historian of statistics and statisticians themselves regularly speak about probability in the same way. Although the history of statistics shows various distributions and permutations of emphasis on the subjective and objective versions of probability, the parlance is now relatively normalised around a divided view of probability. For instance, a well-regarded textbook of statistics written by Larry Wasserman describes the situation as follows:

> We will assign a real number $Pr(A)$ to every event A, called the **probability** of A [@wasserman_all_2003,3]

Note that this number is 'real', meaning that it can take infinitely many values between 0 and 1; secondly that the number concerns events, where events are understood as subsets of all the possible outcomes in a given 'sample space' ('the **sample space** Ω is the set of possible outcomes of an experiment. ... Subsets of Ω are called **Events**' [@wasserman_all_2003,3]). Wasserman goes on to say:

> There are many interpretations of $Pr(A)$. The common interpretations are frequencies and degrees of belief. ... The difference in interpretation will not matter much until we deal with statistical inference. There the differing interpretations lead to two schools of inference: the frequentists and Bayesian schools [@wasserman_all_2003, 6].

The difference will only matter, suggests Wasserman, in relation to statistical inference. Or it may be that even before the different interpretations of probability even come into play, the grounds of probability are shifting.

Markov Chain Monte Carlo: an algorithm for subjectifying probability objectively?

While I depart from a Baroque event — the invention of probability — my primary focus here is on how contemporary probability has become entwined with a particular mode of computation. This paper will not in any way trace the complicated emergence of probability and its development in various statistical approaches to knowing, deciding, classifying, normalising, governing, breeding, predicting and modelling. Historians of statistics have documented this in great detail, and tracked how statistics is implicated in power-knowledge in various settings [@mackenzie_statistical_1978;@stigler_history_1986; @hacking_taming_1990; @daston_how_1994; @porter_trust_1996]. In examining a salient contemporary treatment of probability, my concern is the problem of invention of forms of thought able to critically affirm probability today. This problem arises, I suggest, in many, perhaps all, contemporary settings where populations, events, numbers and calculation are to be found. In seeking to unfold ways of thinking probability for social theory from computational practice, risks of scientism or scientocentrism abound. On this score, a Baroque sense of what happens offers at least tentative pointers to a different way of thinking about what is likely to happen in which the aleatory and the epistemic senses of probability find themselves recombined.

A single statistical simulation technique called MCMC – Markov Chain Monte Carlo simulation – has greatly transformed much statistical practices since the early 1990s (see [@mcgrayne_theory_2011] for a popular account). MCMC is sometimes called an algorithm: a series of precise operations that transform or reshape data. Moreover, MCMC has been called one of 'the ten most influential algorithms' in twentieth century science and engineering [@andrieu_introduction_2003, 5]. But MCMC is not really an algorithm, or at least, if it is, it is an algorithm subject to various algorithmic implementations (for instance, Metropolis-Hastings and Gibbs Sampler are two popular implementations). Invented during the 1950s, the MCMC technique is important in contemporary statistics, and especially in Bayesian statistics. It plays significant roles in areas such as images, speech and audio processing, computer vision, computer graphics, molecular biology and genomics, robotics, decision theory and information retrieval [@andrieu_introduction_2003, 37-38].

In all of these settings, MCMC is a way of simulating a sample of points distributed on a complicated curve or surface (see Figure 1). The MCMC technique addresses the problem of what to do with very uneven or folded distributions of numbers. It is a way of calculating areas or volumes whose curves, convolutions and hidden recesses elude perception. Accounts of MCMC emphasise

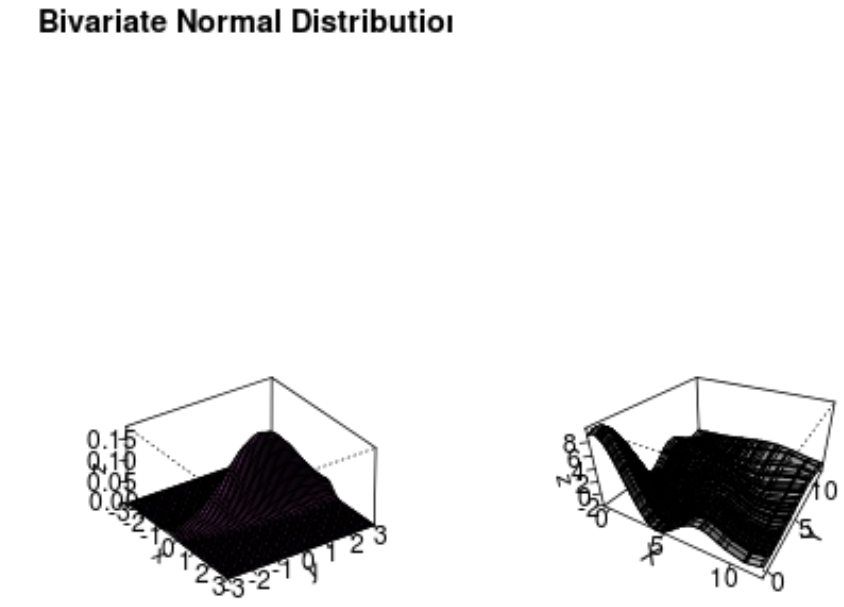

Figure 1: Folded surfaces

the 'high-dimensional' spaces in which the algorithm works: 'there are several high-dimensional problems, such as computing the volume of a convex body in d dimensions, for which MCMC simulation is the only known general approach for providing a solution within a reasonable time' [@andrieu_introduction_2003,5]. Indeed, we could say that MCMC increasingly facilitates the fabrication of high-dimensional, convoluted data spaces. Simulating a sample of points on folded surfaces, it becomes possible to calculate the area or volume enclosed by the surface. This area or volume typically equates to a probability. MCMC, put in terms of the minimal formal definition of probability is a way of assigning real numbers to events, but events occurring within complicated sample spaces.

What MCMC has added to the world is subtle yet indicative. In a history of the technique, Christian Robert and George Casella, two leading statisticians in the field, write that 'Markov chain Monte Carlo changed our emphasis from "closed form" solutions to algorithms, expanded our impact to solving "real" applied problems and to improving numerical algorithms using statistical ideas, and led us into a world where "exact" now means "simulated" ' [@robert_history_2008,18]. This shift from 'closed form' solution to algorithms and a world where 'exact means simulated' might be all too easily framed by a post-modern sensibility as another example of the primacy of the simulacra over the original. But here, a Baroque sensibility, awake to the at once objective and subjective senses of probability, might allow us to approach MCMC less precipitously and less in terms of a crisis of referentiality.

In order to make sense of the change in emphasis described by Robert and Casella, participant histories of the technique are useful. These histories do not, however, highlight the shifts and transitions in the senses of probability associated with the technique. Again, the Baroque sense of probability, especially as articulated by G.W. Leibniz, the 'first philosopher of probability' [@hacking_emergence_1975, 57], is helpful in keeping the concept more open. The brief version of the history of MCMC is as follows: physicists working on nuclear weapons at Los Alamas in the 1940s [@metropolis_monte_1949]} first devised ways of working with high-dimensional spaces in statistical mechanical approaches to physical processes such as crystallisation and nuclear fission. Their approach to statistical mechanics was then generalised by statisticians [@hastings_monte_1970]}. It was taken up by ecologists working on spatial interactions in plant communities during the 1970s [@besag_spatial_1974], revamped by computer scientists working on blurred image reconstruction [@geman_stochastic_1984], and then subsequently seized on again by statisticians in the early 1990s [@gelfand_sampling-based_1990]. In the 1990s, it became clear that the algorithm could make Bayesian inference — a general style of statistical reasoning that differs substantially from mainstream statistics in its treatment of probability [@mcgrayne_theory_2011] — practically useable in many situations. A vast, still continuing, expansion of Bayesian statistics ensued, nearly all of which relied on MCMC in some form or other. (Thompson Reuters Web of Knowledge shows 6 publications on MCMC in 1990, but over 1000 *each year* for the last five years in areas ranging from agricultural economics

to zoology, from wind-power capacity prediction to modelling the decline of lesser sand eels in the North Sea; similarly NCBI Pubmed lists close to 4000 MCMC-related publications since 1990 in biomedical and life sciences, ranging from classification of new-born babies EEGs to within-farm transmission of foot and mouth disease; searches on 'Bayesian' yield many more results). In the social sciences too, political scientists regularly use MCMC in their work [@gill_introduction_2011] because their research terrain — elections, opinions, voting patterns — little resembles the image of events projected by mainstream statistics: independent, identically distributed ('iid') events staged in experiments. When brought together with Bayesian inference, MCMC allows, as the political scientist Jeff Gill observes, all unknown quantities to be 'treated probabilistically' [@gill_introduction_2011,1]. We can begin to see why the Obama re-election team might have been running their model 66,000 times each night. In short, MCMC allows, at least in principle, *every* number to be treated as a probability.

The proliferation of probabilities is not unprecedented, at least philosophically. As Hacking reports, C.S Peirce, the American pragmatist philosopher who spent much of his life measuring things for the US Coastal Survey, was already arguing against *any* constant numbers, social or natural, in the late nineteenth century [@hacking_taming_1990, 200]. Only statistical stabilities mattered. A century later, the popularisation of MCMC perhaps surpasses what Peirce (and Hacking?) had in mind in saying there are only statistical stabilities. Peirce envisioned a universe filled with chance events ('chance pours in at every sense'), amidst which islands of pragmatic sense emerged standing on habit and consensus. By contrast, treating every number as a probability, as facilitated by MCMC, does not simply generalise probability by saying that the universe is indeed aleatory. On the contrary, as I will seek to show, it allows a hyper-subjectified sense of probability to take shape precisely through its exteriorisation in folded flows of random numbers. A technique of computational simulation distributes numbers in the world, it assigns numbers of events, but largely in the service of modifying, limited, quantifying belief and uncertainties associated with beliefs. This folding together of subjective and objective, of epistemic and aleatory senses of probability can be thought as a neo-Baroque monadological mode of probability.

Distributions: living in the margins

The flat operational definition of probability as mapping events to real numbers seethes with convolutions. The problem here is the continuum. While some events are discrete, many are happenings are not. In Lancaster, the probability of rain on a given day would be say 70%, however days have very different amounts of rain. Some days, it rains once briefly and lightly. Other days it rains frequently and heavily. A gamut of rain events can occur, and each would distribute different amounts of water on Lancaster. Given the amount of vari-

ation, much better then to say that rain on Lancaster is a *random variable*: 'a random variable is a mapping that assigns a real number to each outcome' [@wasserman_all_2003,19]. Again, the deceptive simplicity of 'mapping' hides many variations. Mapping is a form of one-to-one correspondence, usually expressed as a mathematical function. A random variable links events to numbers through functions. Again, all this remains rather formal. The practical reality is better expressed by curves, and ways of talking about what the curves express.

```
## Using as id variables
```

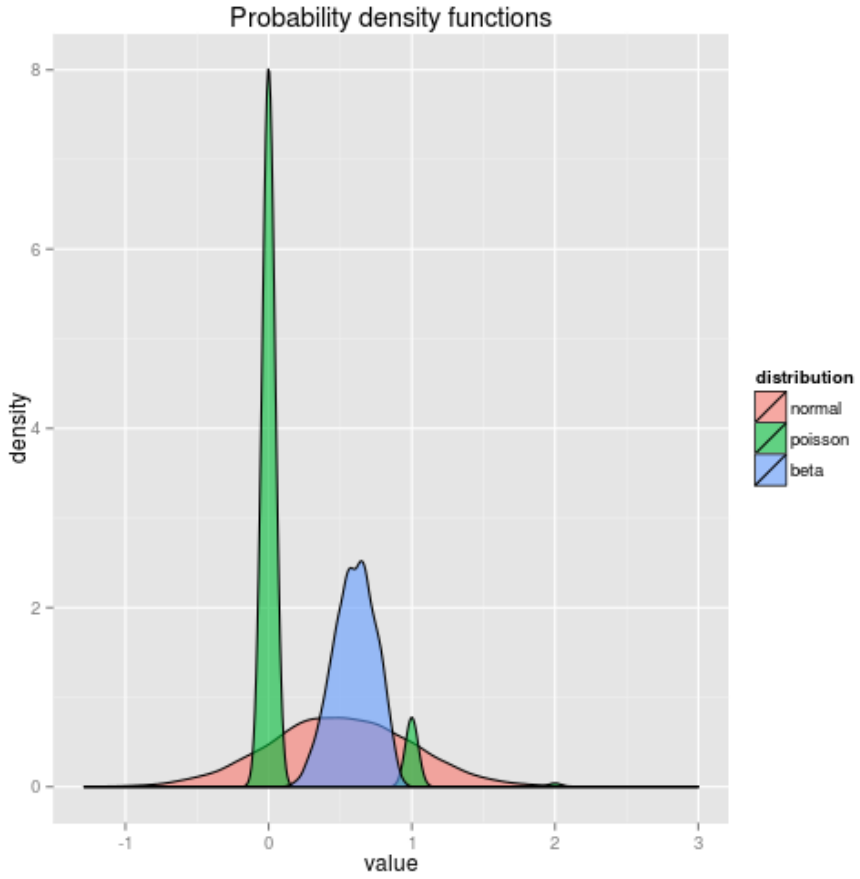

Figure 2: Distributions

Probability distributions are a common way of showing and talking about random variables. As shown in Figure 2, these distributions appear in countless shapes and forms in scientific, government and popular literature of many different kinds. Statistical graphics have a rich history and semiology that I do

not discuss here (see [@bertin_semiology_1983]). Perhaps the most famous function or mapping is the normal or Gaussian distribution. This distribution has a powerladen biopolitical history, since it is so closely tied with knowledges and governing of national and other populations. The normal distribution is so semiotically powerful that it is possible, under certain circumstances, to effectively subjectify someone on the spot. For instance, if an educational psychologist says to someone that their intelligence lies towards the left-hand size of the peak, they quickly render them somehow subject to the normal curve. But statistics uses dozens of different probability distributions to map continuous and discrete variations to real numbers. Probability distributions abound — normal (Gaussian), uniform, Cauchy exponential, gamma, beta, hypergeometric, binomial, Poisson, chi-squared, Boltzmann-Gibbs distributions, etc (see [@nist_2012] for a gallery of distributions) — because outcomes occur in widely differing patterns. The queuing times at airport check-ins do not, for instance, easily fit a normal distribution. Queues are usually modelled using a Poisson distribution, which unfortunately for travellers, distributes waiting times very differently. Similarly, it might be better to think of the probability of rain today in Lancaster in terms of a Poisson distribution that models that queue of clouds in the Atlantic just waiting to land on the northwest coast of England. Rather than addressing the question of if it will rain, the Poisson distribution addresses the question of how soon.

The diverse range of probability distributions — and we will see below some reasons why we can expect them to proliferate in certain settings — attests to the variable ways in which events might be mapped to real numbers. Crucially for our purposes, the term *distribution* emphasises a quite material or tangible way of thinking about probabilities, despite the sometime forbidding mathematical equations. The curves in both Figure and Figure 2 are examples of the most common mathematical descriptions in any data analysis setting: they are *probability density* functions (for continuously varying quantities). There are also * probability mass* functions for variables that have discrete values; for instance: 1,2,3,4,5). The probability density function (pdf) is a function, usually graphed as a curve, that describes how likely a random variable is to take on a particular value. In many cases, statistical practice seeks to estimate distribution functions such as pdfs (or their close relatives, cdfs — *cumulative distribution functions*) for the given data. Statisticians speak of 'fitting a density' to data, emphasising their assumption that events can be incorporated in the forms of probalitity distributions. The underlying probability distribution is 'unobservable' as such, but it is assumed to give rise to all the data gathered through experiments and observations. The task is to estimate the shape of that curve, and its defining parameters (means, variance, etc.). Given that curve, areas under the pdf equate to the likely range of value of a variable. While the total 'probability mass' under the probability density function curve always must be equal to one (since the probability of each individual outcome ranges between 0 and 1), finding the area under particular parts of the curve is a key issue. Finding the area under probability density (or mass) function curves be-

comes the way in which many epistemic processes link mathematical functions to lived states of affairs, such as populations.

Multiplying curves

In the Bayesian statistics popular since the 1990s, any number, including the defining parameters of other distributions such as the mean, can treated as a probability. Indeed, the 'Bayesian revolution' in statistics exemplifies a multiplying of probabilities. Bayes Theorem, known since the 18th century, is usually presented in the first few pages of any probability textbook, as a way of relating probabilities to each other:

$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)}$

where $Pr(A)$ and $Pr(B)$ are two probabilities, and $Pr(A|B)$ is the probability of A given B, and $Pr(B|A)$, conversely, is the probability of B given A. Again, bearing in mind the different notions of probability discussed above (subjective - objective; degree of belief vs. frequency of outcome), this formula can be read in different ways. But regardless of the interpretation, probabilities are being multiplied here. And if the A and B are random variables, then probability densities are being multiplied to produce higher-dimensional surfaces, as we saw in the very simple illustration of Figure 1. The *joint probability distribution* that results still has the same total mass (1), but now distributed in a different volume. As the number of random variables grows, the surfaces open onto higher dimensions, and cannot be graphed easily.

Knowledge of how to mathematically manipulate the functions associated with particular probability distributions has accreted over several centuries. They all share a common purpose: to express the distribution of outcomes associated with certain events. Bringing together random variables in models, even in the basic form of the Bayes Rule where B is conditioned by A, means multiplying probability density functions. The total mass of the the probability always remains the same (i.e. 1), but the question is where it is distributed. As the simulated joint probability density of Figure 1 shows, certain zones of a joint probability are much more mountainous that others, and these peaks suggest more likely events in the range of possible outcomes.

The mathematical difficulties posed as distributions multiply relates to that other great Baroque mathematical invention, calculus. If calculus made possible so many different calculations of rates of change, calculations that profoundly affected senses of space, time, and increasingly growth, variation and change more generally (hence, Deleuze's work both on Leibniz and in his philosophical conceptualisation of difference more generally is deeply imbricated with differential calculus [@deleuze_fold_1993]), it also ran into many obstacles in relations to calculations of probability. Calculating the area under a curve in order to estimate variables is a problem of integration. That is, the area under a curve is given by the integral of the probability density function. If the probability

distribution cannot be normalized, then the area under the curve is much harder to estimate. The ornate and at times bewildering apparatus of statistical tests and procedures, as well as the debates between different schools of statistics (Bayesian vs frequentist), largely obscures the continuous trajectory in which what happens is transformed into a problem of measuring areas under curves, or volumes under surfaces. Sometimes the estimates are understood as a measure of our belief about what happens (as in Bayesian analysis) and sometimes it is understood as a measure of the frequency with which events occur in the world (as in frequentist statistics). While there is now a very extensive technical and philosophical literature on the differences between Bayesian and frequentist statistics, more or less the same computations can be in the service of either standpoint. So this is not the main point I want to pursue here.

From the perspective of a Baroque sensibility, mathematical functions for working with different shapes, areas, densities and masses of probability distributions have been combined to support estimations, inferences and predictions of change and growth in many processes. Through the generating role played by probability distributions in almost any field of science, government, industry, technology and increasingly media and commerce we could name, probability mixes through almost all forms of relationality. In calculations of insurance risk, in algorithms for error correction, in psychological testing, in climate models or biodiversity surveys, just to name a few, probability distributions function ground all inference. Although certain distributions, such as the normal, Poisson or binomial, etc., have dominated in these developments, this was largely because it has been easier to calculate estimates of their main parameters (mean, variance, etc) than those pertaining to less familiar distributions. In terms of shape, area and hence probability density, the normal distribution is one of the most tractable curves to work with. Even with the various data transformations and normalizations developed over several centuries, other probability distributions have been harder to work with. (They lack the 'closed form' solutions that Robert and Casella refer to.) This occasions many disputes in the history of statistics over 'curve-fitting' to normal or other mathematically tractable distributions as arbitrary and unjustified [@hacking_taming_1990, 164]. In whatever way these disputes have been resolved (see [@mackenzie_statistical_1978] for an early 20th century example), the practical problem of calculating the area under all or some part of the curve has skewed what we believe about many different things (about sub-atomic particles, climate change, likelihood of glaucoma, the chances of rain today, Obama's chance of re-election, etc.) towards some forms of probabilitity distribution more than others. The normal probability density function tends to be the norm.

Good approximations to probabilities

The proliferation of normal curves and surfaces brings us back to MCMC, the technique that inaugurates 'a world where "exact" now means "simulated" '

[@robert_history_2008,18]. MCMC is, as mentioned above, a technique for simulating samples from high-dimensional or complicated concave volumes. In other words, it is a way of exploring the contoured and folded surfaces generated when flows of data or random variables come together in one joint probability distribution. These surfaces, generated by the combinations of mathematical functions or probability distributions are not easy to see or explore, except in the exceptional cases where calculus can deliver a deductive analytical 'closed form' solution to the problems of integration (finding the area) and differentiation (finding the distribution function for one variable). By contrast, MCMC effectively simulates some important parts of the surface, and in simulating convoluted volumes, loosens the analytical ties that bind probability to certain well-characterised analytical regular forms such as the normal curve.

In this simulation of folded and multiplied probability distributions, the lines between objective and subjective, or aleatory and epistemic probability, begin to shift. There is perhaps something increasingly monadological about MCMC, as we can see if we revisit the history of the technique with less an eye on the events leading up to the [Bayesian] revolution, and more with an eye on what is being folded in, and what is unfolding as the technique develops. The starting point here, and it is found in almost every textbook on MCMC-related methods is the computer as random number generator. Rather than Peirce's 'chance pouring in at every sense,' it might be better to speak of chance pouring out of MCMC on every event.

```
## evaluated this
```

Figure 3 shows two plots. The one on the left plots 10,000 computer generated random numbers between 0 and 1, and as expected, or hoped, they are more less uniformly distributed between 0 and 1. This is simulation of the simplest probability distribution of all, the *uniform* probability distribution in which all events are equally likely. The plot on the right derives from the same random numbers, but shows a different probability distribution in which events mapped to numbers close to 0 are much more likely than events close to 1. What has happened here? The reshaping of the flow of numbers depends a very simple multiplication of the simulated uniform distribution by itself.

> A real function of a random variable is another random variable. Random variables with a wide variety of distributions can be obtained by transforming a standard uniform random variable $U \approx UNIF(0,1)$. Let $U \approx UNIF(0,1)$. ... We seek the distribution of $X = U^2$ [@suess_introduction_2010, 32].

It happens that multiplying a uniform distribution by itself (U^2) produces an instance of important distribution, the *Beta* distribution, shown on the right of Figure 3. Now it would be possible to produce that curve of a beta distribution analytically, by plotting points generated by the *Beta* probability density function:

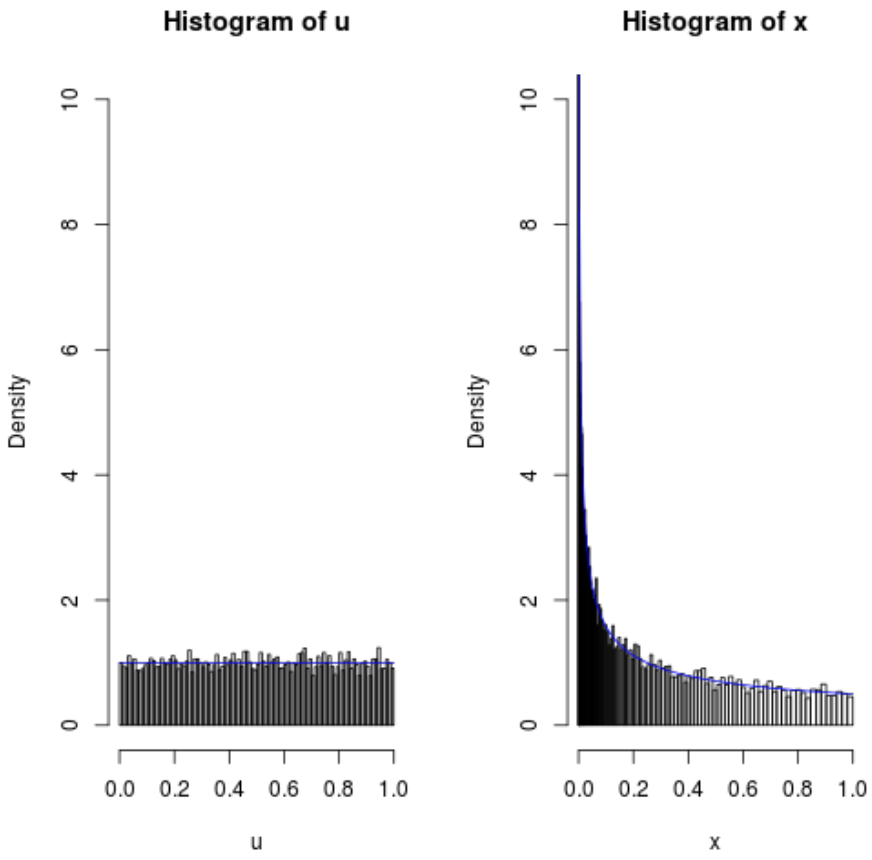

Figure 3: Simulated distributions

$$f(x;\alpha,\beta) = constant \bullet x^{\alpha-1}(1-x)^{\beta} - 1$$

where $\alpha = 0.5$ and $\beta = 1$. But in the case of the plots shown on the right of Figure 3, the shape has been generated from a flow of random variables, So, from a flow of random numbers, generated by the computer (using an *pseudo-random* number generator algorithm), more random variables result, but with different shapes or probability densities. As Robert and Casella write, 'the point is that a supply of random variables can be used to generate different distributions' [@robert_introducing_2010,p.44]. Indeed, this is the principle of all Monte Carlo simulations, methods that 'rely on the possibility of producing (with a computer) a supposedly endless flow of random variables for well-known or new distributions' [@robert_introducing_2010, 42]. The example shown here is really elementary in terms of the distribution and dimensionality of the random variables involve, but it illustrates a general practice underpinning the MCMC technique: the reshaping of the 'supposedly endless flow of random variables' to produce known or new distributions.

This is already monadological in the sense that it seems to bring probability inside the computer. Monte Carlo simulations regard computers as substitutes for chance in the world, and they render that world more manipulable by knowing subjects. It is hardly surprising that scientists working at the epicentre of the 'closed world' [@edwards_closed_1996] of post-WWII nuclear weapons research should develop a technique that allows the world to move in this way. In 1953, Metropolis, the Rosenbluths and and the Tellers were calculating 'the properties of any substance which may be considered as composing of interacting individual molecules' [@metropolis_equation_1953, 1087] (for instance, the flux of neutrons in a hydrogen bomb detonation). In their short, but still widely cited paper (over 20,000 citations according to Google Scholar; over 14,000 according to Thomson Reuters Web of Knowledge), they describe how they used computer simulation to deal with the number of possible interactions in a substance, and to thereby come up with a statistical description of the properties of the substance. Their model system consists of a square containing only a few hundred particles. These particles are at various distances from each other and exert forces (electric, magnetic, etc.) on each other dependent on the distance. In order to estimate the probability that the substance will be in any particular state (fissioning, vibrating, crystallising, cooling down, etc.), they needed to integrate over the many dimensional space comprising all the distance and forces between the particles. (This space is a typical multivariate joint distribution.) As they write, 'it is evidently impossible to carry out a several hundred dimensional integral by the usual numerical methods, so we resort to the Monte Carlo method' (1088), a method that Nicholas Metropolis and Stanislaw Ulam had already descibed in an earlier paper [@metropolis_monte_1949]. Here the problem is that the turbulent randomness of events in a square containing a few hundred particles thwarts calculations of the physical properties of the substance. They substitute for that non-integrable turbulent randomness a controlled flow of random variables generated by a computer. While still somewhat random

(i.e. pseudo-random), these Monte Carlo variables taken together approximate to the integral of the many dimensional space. In monadological terms, Monte Carlo simulation attenuates the distance between the aleatory and epistemic poles of probability. While the computer is regarded as aleatory in its capacity to generate seemingly random numbers, it is strongly epistemic in its power to marshall these numbers into shapes that cannot be analysed using the 'usual numerical methods' (for instance, of integral calculus).

Probability in Markov Chains

Metropolis and co-authors immediately go on to say, however, that they cannot just sample a random set of points. The range of events is not equally accessible to simulation. The joint probability distribution of the system is quite uneven (that is, highly folded). Randomly sampled points are likely to lie in low probability regions (valleys and plains), whereas they are interested in the high probability peaks. Instead they propose a move which becomes the modus operandi of subsequent MCMC work (and hence justifies the high citation count): 'we place the N particles in any configuration ... then we move each of particles in succession' (1088). Here is the beginning of the 'random walk' or 'Markov chain' technique that distinguishes MCMC from Monte Carlo simulation more generally. As well as generating a sample of random variables, they submit each variable to a test. Physically, the image here is that they displace each particle by a small random amount. Having moved the particle/variable, they calculate the resulting slight change in the overall system state, and then decide whether that particular move puts the system in a more or less probable state. If that state is more likely, the move is allowed; otherwise the particle goes back to where it was. Having carried out this process of small moves for all the particles, they can calculate the overall system state or property. The process of randomly displacing the particles by a small amount, and always moving to the more probable states, effectively explores the bumpy topography of the joint probability density. In many minute moves, the simulation begins to migrate all the randomly generated values points onto the peaks that represent interestingly high probabilities.

```
## [1] -0.0348 -0.0354  0.9966  0.9992  0.7994
```

In Figure 4, a toy example, the MCMC technique has been used to generate a simulate a bivariate normal probability distribution. (The example comes from [@suess_introduction_2010, 177-178].) We have already seen a bivariate normal distribution (see Figure 1), but now the distribution has been produced by drawing on the uniformly distributed flow of random variables produced as an MCMC algorithm (actually in this case, the Metropolis-Hastings implementation of MCMC) runs, rather than by generating values using the mathematical function for the probability density. On the left hand side, we see the path

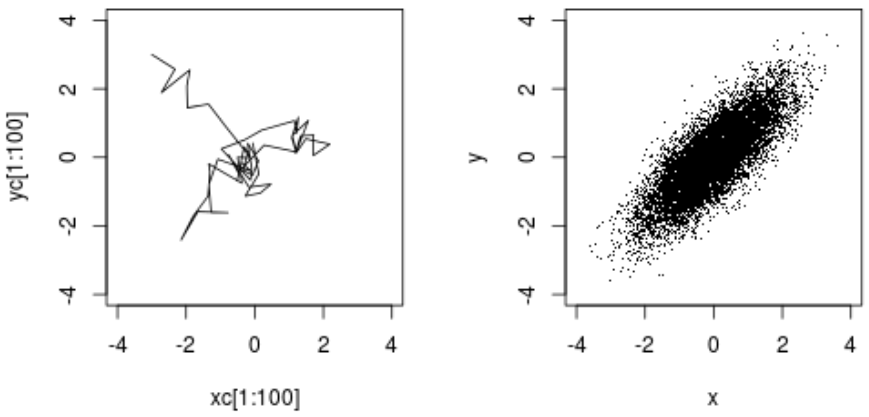

Figure 4: Bivariate normal distribution generated using MCMC

by taken a single random variable as it moves closer to the central peak of the distribution, the zone of highest probability. On the right hand side, we can see the cloud of variables clustered around the central peak of the bivariate normal distribution after the MCMC technique has run 40,000 times. The Metropolis-Hastings implementation and the perhaps more popular Gibbs sampler implementation of the MCMC technique share this idea of waiting to see where flows of random variables end up and hoping that distribution of these values will approximate a 'a desired long-run distribution' [@suess_introduction_2010, 150].

'Consider the Markov chain defined by $X^{t+1} = \sigma X^t + \epsilon(t)$ where $\epsilon(t)$ $\mathcal{U}(0, 1)$', write Robert & Casella [@robert_introducing_2010, p.169]. This Markov chain knows nothing of the normal distribution, yet simulates it by piling up large numbers of random numbers.

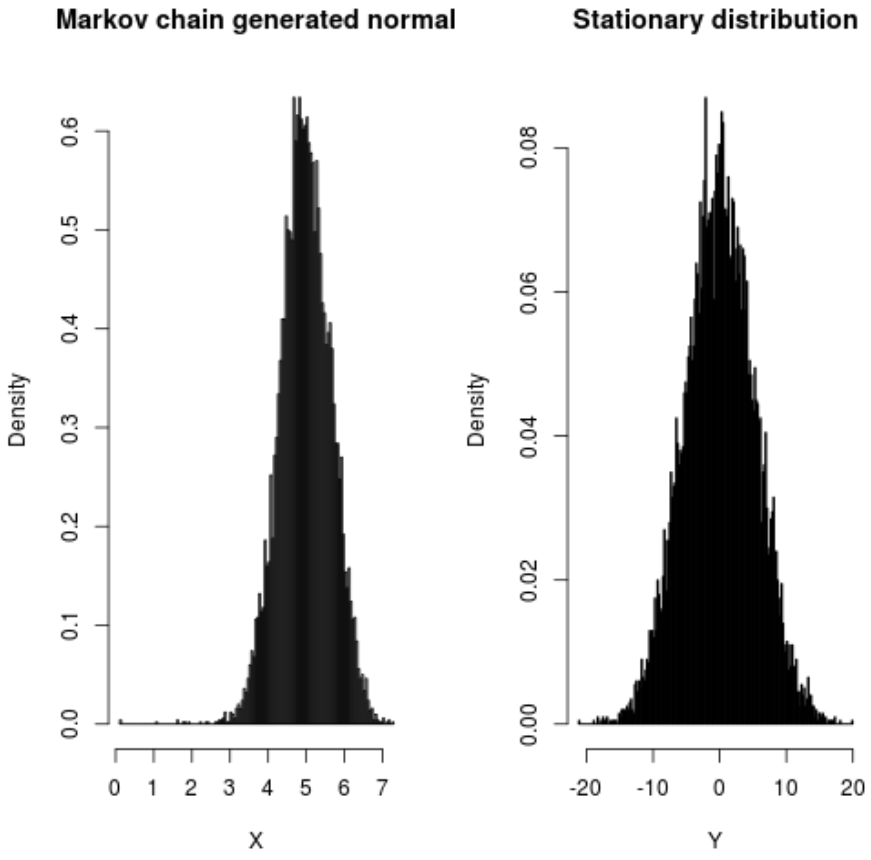

Figure 5: Markov-chain generated normal distribution

Again, it is possible this is something like what the Obama data analytics team were doing each night. For present purposes, however, the point is that these supplementary moves – the slight perturbations and adjustments that permit a kind of exploration of the features of the probability density surface – remain monadological in the sense that they express a world within a monad. But they also suggest that that world is not immediately accessible or transparent. It has to be explored through the invention of techniques that render its variations and differences somehow sensible.

How events and beliefs are combined in MCMC

With apologies for the slightly forbidding typography, the probability:

$$f(\mathbf{x}|\theta) = L(\theta) = \prod_i f(x_i|\theta)$$

is a likelihood function, a function that describes how likely some data (from measurements, observations, experiments, transactions, etc) is, given a particular value of θ: 'it provides the chances of each value of θ having led to that observed value of x' [@gamerman_markov_2006, 43]. θ is a parameter of some kind that describes the shape of a probability distribution (such as mean and variance for normal distributions, α, β for beta distributions, etc.). The typography of this expression is important. The value $\mathbf{x}$ is bold font because it stands for a random variable, a variable that it observed to take a range of values (continuous or discrete). The large symbol $\prod_i$ stands for the product or multiplication of the probabilities of all the different observed values of x (not bold), *given* a particular value of θ. All of this highlights a reversal that stands at the centre of contemporary usages of MCMC: the data is evaluated in the light of prior *belief* about the values of the key parameters of the probability distributions. Folding together data and belief is not exactly the same as a coalescence between the aleatory and epistemic poles of probability, but it certainly brings them much closer together. Yet, as we have seen, this increased proximity between the two faces of probability in MCMC comes about only via the many detours of the Markov-chained random variables creeping across the multi-dimensional topography of the joint probability distributions.

When in 1990, 'Sampling-Based Approaches to Calculating Marginal Densities' the article that set off the 'Bayesian revolution,' appeared in *Journal of the American Statistical Association* [@gelfand_sampling-based_1990], the statisticians Alan Gelfand and Adrian Smith refer to this altered topology. They state that the problem they are addressing is how 'to obtain numerical estimates of nonanalytically available marginal densities of some or all [the collection of random variables] simply by means of simulated samples from available conditional distributions, and without recourse to sophisticated numerical analytic methods' [@gelfand_sampling-based_1990, 398]. It would take some paraphrasing to develop all of the [TBA]

[REACHED HERE]

They take up the Gibbs sampler algorithm as developed by [@geman_stochastic_1984] for image-processing, investigate some of its formal properties (convergence), and then set out a number of mainstream statistical problems that could be done differently using MCMC and the Gibbs sampler in particular. This paper is sometimes said to have announced the 'Bayesian revolution' [@robert_introducing_2010, 9] because it made clear the links between MCMC and statistical inference more generally through six illustrative mainstream examples: multinomial models, hierarchical models, multivariate normal sampling, variance components, and the k-group normal means model. The details of these examples need not detain, but each of the illustrations in the paper shows how previously difficult problems of Bayesian inference can be carried out by sampling simulations. As they state in another paper from the same year, 'the potential of the methodology is enormous, rendering straightforward the analysis of a number of problems hitherto regarded as intractable' [@gelfand_illustration_1990, 984]. A rapid convergence on MCMC follows from the 1990s onwards. Gibbs samplers appear in desktop computer software such as the widely used WinBUGS ('Windows Bayes Using Gibbs Sampler') written by statisticians at Cambridge University in the early 1990s [@lunn_winbugs-bayesian_2000], and MCMC quickly moves into the different disciplines and applications found today.

References