Plying R: a statistical programming language and the credibility of data

Adrian Mackenzie, Cesagen, Lancaster University, a.mackenzie@lancaster.ac.uk

The position paper that follows did not work out as expected. The general topic is how things become data. I thought it would track how a statistical programming language called R moves between biology, mathematics, finance, field sciences, business analytics, epidemiology, and social sciences, etc. The interest for me was in finding commonalities, lateral movements, borrowings or contagions of belief and desire (Tarde) associated with data.

What happened instead was that I found myself sidetracked by a 17th century event: the emergence of probability. If the contemporary case study of R is already crowded with sidetracks, the history of probability is full of even more chance connections. As a result, the position paper moves somewhat uncertainly between 1660 and the 2000s. The 1660s is relevant because it gives rise to the intrinsically bifurcated concept of probability, oscillating between belief and events. The 1990-2000s episode is also bifurcated: between very mundane operations seeking to reduce data friction, and computational intensive statistical models seeking to enhance the power of statistics to assign cases to source populations via controlled injections of randomness. The contrast between the practical *plying* – as in using, supplying, moving between, and folding – of data and calculation of degrees of *credibility* interests me greatly. My position, if I have one, is that this confluence of plying and calculating data occurring today might be the sign of a shift in credibility. This shift is best understood in the light of the ongoing life of probability.

The 1660s: belief in chance

In his widely discussed account of the emergence of probability, Ian Hacking (Hacking 1975) locates a shift around 1660: 'there was no probability until about 1660' (17). For the 17th century figures that Hacking discusses – Leibniz, Pascal, Arnauld, Huyghens and Petty – the question of whether probability concerns what happens or something about our beliefs remains somewhat undecided. The mixing of belief and events can be seen in various debates about chance, about miracles, etc. in Leibniz and others. This is an ongoing historical instability in the notion of probability: probability is both something to do with beliefs (pre-17th century, probability refers to the authority of religious texts), and probability is something to do with how often things happen in world (especially when chance is involved, as in the rolling of dice or at what age people die). The epistemic-ontological duality continues unabated through the 18-19th centuries (see (Daston 1994) for a survey) well into the late 20th century (Fienberg 2006). Rather than this uncertainty being something that the development of mathematics of probability cleared away, Hacking frames it as it allowing probability to emerge.

The emergence of probability can be seen in terms of a difference between external and internal evidence. External evidence comes from outside the thing. Testimony, for instance, is external evidence, and is well understood already in the Renaissance (Hacking 1975, p.34). Internal evidence consists in a 'thing pointing beyond itself' (34), and is specifically new to the 17th century. The model for internal evidence comes from the 'low sciences' of alchemy, geology, medicine and astrology, knowledges that cannot argue through demonstration, but rely on diagnoses, exploration or 'adventure' to form opinions (36). This new empiricism works with signs, especially 'natural signs.' Signs-as-evidence, especially as they inform the physicians' diagnosis, pass over into a more widespread conception of the sign as the basis of the empirical knowledge that will 'come to dominate European thought' (46). Natural signs are not conventional or conjectural; instead they are frequently observed in experience. In the ways they point beyond themselves, frequently observed signs lie somewhere between demonstration (the ideal of high sciences) and testimony. Because the naturalness of such signs depend on repetition, counting their occurrence becomes central. During the second half of the 17th century, in the wake of this shift in what counts as evidence, calculations of probability based on counting what happens can gain ground. Games of chance (lotteries, dices games, etc.) had long been subject to attempts at calculation, but now calculations of annuities based on expected lifespans as well as gambling become amenable to the new empiricism of countable sign-as-evidence. Interestingly, the emergence of a concept of probability was not dependent on games of chance. While Pascal and Huyghens, for instance, departed from games of chance, Leibniz, as Hacking reports, working in Germany, not knowing much about games of chance, developed an account of probability based on numerical degrees of belief derived from jurisprudence (89; Lorraine Daston develops a similar argument at greater length (Daston 1988)).

'Internal evidence' of this kind becomes for twentieth century epistemology 'fundamental evidence' (83). Throughout much of the subsequent history of statistics in science, government, economics and technology, debates over the character of probability continued. Only in the early twentieth century, in the mark of R.A. Fisher, did the 'frequentist' (Fisher's term) interpretation prevail . From then, statistics mostly reckons how often some event occur, and then tries to use those measures of frequency to infer something about the likely distribution of events in a population (of things, of lives, etc.). The apparatus of statistical reasoning and techniques for inference henceforth rests largely on this conception of probability as something to with what happens, not with our opinions or belief about what happens. The notions of populations, of distributions, of parameters (means, variances, etc.), of tests of significance and association, estimations of error, confidence intervals and expected values, of hypothesis testing – largely remains frequentist. The meta-statistical concerns with norms, normality and populations (human, species, particles, etc.) – hence biopolitical governmentality too – are highly imbued with frequencies, with counting how often things occur, and with probability as the way of making sense of frequencies.

Yet like a tectonic faultline, treatments of probability as belief run through the history of statistical thought, resurfacing at various points, for instance in the eighteenth century in Bayes' Theorem, and again from the mid-20th century, arguably, especially in much of the statistical work called 'Bayesian' (Fienberg 2006), work that has surged in popularity since the early 1990s. A Baroque sense of probability, where the mixing of belgenomicevent is less bifurcated, and more richly material-semiotic, could be useful in following that faultline as it ripples across contemporary data.

The 1990s: belief in R

I have been studying – well more than studying, actually – a well-known and widely used statistical programming language and environment called R (R Development Core Team n.d.). According to surveys of business and scientific users, R is replacing popular software packages such as SPSS, SAS and Stata as the statistical and data analysis tool of choice for many people in business, government, and sciences ranging from political science to genomics, from quantitative finance to climatology (Rexer Analytics 2010). Developed in New Zealand in the mid-1990s, and like many open source software projects, emulating a commercialised predecessor (the language S), R is now extremely widely used across life and physical sciences, as well as quantitative social sciences. Many undergraduate and graduate students learn R as a basic tool for statistics. Skills in R are often seen as essential pre-requisite for scientific researchers, especially in the life sciences. Estimates of its number of users range between 250000 and 2 million. Increasingly, R is integrated into commercial services and products (for instance, SAS, a very important statistics package now has an R interface; Norman Nie, one of the original developers of the SPSS package heavily used in social sciences, now leads a business, Revolution, devoted to commercialising R; R is heavily used at Google, at FaceBook, and by quantitative traders in hedge funds, etc.).

R is an interestingly diffuse entity. Some ways of working with data are way more clearly focused on equations and calculation. For instance, in order to pursue numbers in physical sciences and engineering, maybe MATLAB or Mathematica would be better. Other ways of working with data are more focused on ordering and searching. For instance, in order to the look at the organisation of large aggregates of data, relational databases, query languages, data-centre architectures, and perhaps the techniques of aggregating and disaggregating data en-masse would be worth studying. Why then study R, a statistical programming language, a language initiated by statisticians rather than computer scientists, software engineers or hackers? Because it embodies something of the mainstream practice of working across measurements, numbers, texts, images, models and equations, with techniques for sampling and sorting, with probability distributions and random numbers, R is an evocative object. It engages immediately, practically and widely with words, numbers, images, symbols, signals, sensors, forms, instruments and above all virtual forms such as distributions to make data. By virtue of thousands of packages that flow across boundaries between nature and culture, between aesthetic, epistemic and pragmatic values, R embodies a wide-with data economies, cultures, sciences, politics and technologies. Somewhere between calculation and searching, R channels counting and sorting, in the estimation of likelihoods.

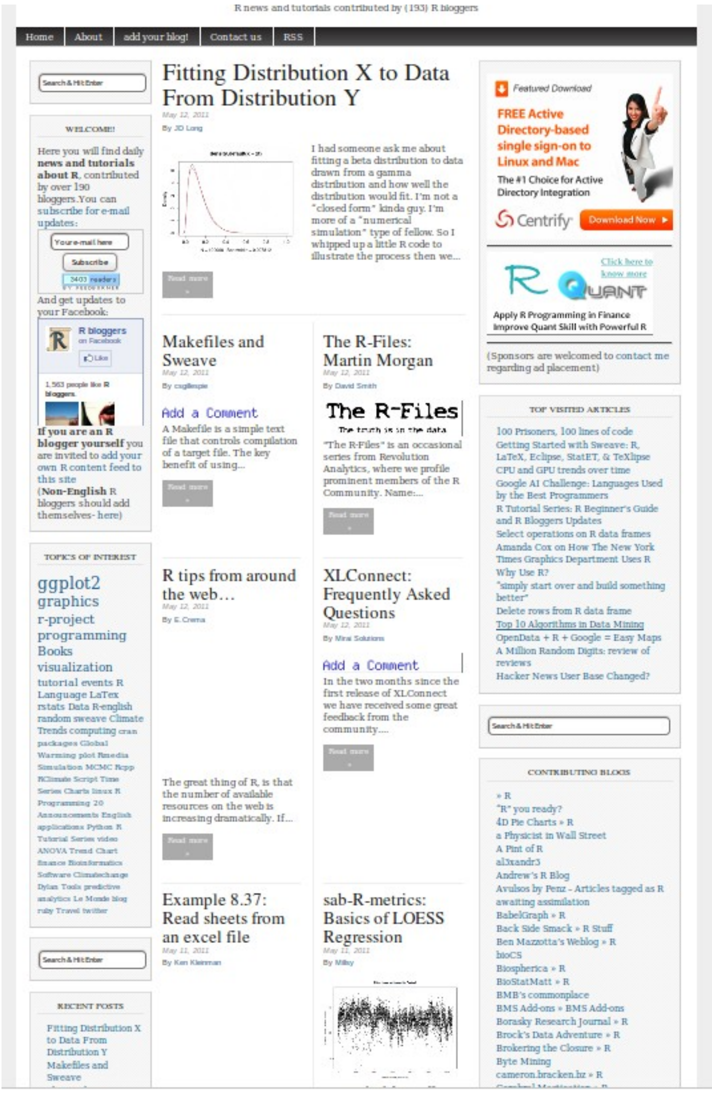
It is easy today to find economised belief in data:

> “The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data. So the complimentary [sic] scarce factor

> is the ability to understand that data and extract value from it." Hal Varian, chief economist at Google: (McKinsey & Company 2009)

What Varian presents as an 'important skill,' the 'scarce factor' is the 'ability to take data' he refers to is something quite complicated, that will be temporally emergent, not just existing. Some people – admittedly a small group – explicitly promote R in this context. They associate R with the wider growth of data democracy or open data. Here is Norman Nie interviewed in *Forbes*, a leading business news magazine, evangelising for R:

> Everyone can, with open-source R, afford to know exactly the value of their house, their automobile, their spouse and their children--negative and positive … It's a great power equalizer, a Magna Carta for the devolution of analytic rights (Hardy 2010).



R-bloggers.com: aggregation of several hundred R-related blogs

It would be possible to cite many other instances of this belief and desire in R since they are rampant in the blogs, publications, and events associated with R. A good source is the aggregate blog, R-bloggers.com, that brings together several hundred R-related blogs in one place. The range of topics discussed there on any one day – Merrill-Lynch Bond Return indices, how to run R on an iPhone, using US Department of Agriculture commodity prices, analysing human genetic variation using PLINK/SEQ via R, using the 'Rhipe' (*sic!)* package to program big data applications on Map-Reduce cloud architectures, doing meta-analysis of phylogenetic trees, etc., etc (R-bloggers 2011) – indicates how widely desires/beliefs in R travel, almost oblivious to the changing scenery.

Data frictions

Actually, this is a key problem for me – having just said that desires/beliefs in R travel widely, I'm not sure what I mean by 'widely' or 'travel.' Maybe it would be better to say they that they do not travel very far, but undergo some kind of modulation that reduces data and computational friction (Edwards, 2010), and allows them to slip around the corners that separate things. As a preliminary take on this question, I would say that frictions are handled R in several ways:

1. **Habit formation**. Because R is a programming language, it allow scripts to be written and run very on everything from laptops to cloud computing. In this it differs greatly from statistics software applications such as SPSS, SAS, or STATA that, arguably, change less often and under the control of business or scientific needs. Learning R is learning to read, write and run scripts, rather than how to carry out operations using the standard interfaces of menus, buttons and windows found in more conventional software applications. There is also much about its connection to programming and computing cultures (Unix operating system, open source scripting languages such as Perl, scientific programming languages such as FORTRAN) that mobilises R in a different way to statistics software applications. Scripts are much more fragmented, mobile elements than software applications. Their mutability and malleability means they can make and move data across many different interstices. At the same time, learning to use and write R opens pathways to habit-forming investments.
2. **Believing is seeing.** Given that seeing something in data, finding patterns, showing connections or regularities, or locating important or relevant parameters animates all work with data, the endless variations in visualization found in R play an important role. The many forms of visualization it allows are very closely connected to debates, trends, fashions concerning how data should be made visible – as table, scatterplot, network diagram, tag-cloud, histogram, heatmap, map, etc. The associated minutiae of visual grammars also connect R directly into the literary technologies of many scientific, government, and business knowledges. It would be hard to over-estimate the role played by 'low science' graphing and diagramming in some of these domains.
3. **The latest and greatest.** The latest statistical and algorithmic techniques and models developed by statisticians, domain scientists and computer scientists are often published as R packages (or libraries), especially on CRAN, the Comprehensive R Archive Network (CRAN 2010). The deposit of an R package - a collection of functions, data structures, data sets or interfaces to data sources – into CRAN often accompanies journal or book publication in certain fields. The quick rollout of techniques in the form of packages means that people resort to R to keep up. To understand what precipitates the data deluge, how people handle it, and what kinds of resistances, we might think more about the different things that become a package. Maybe it would be worth thinking about packages as monads, as world-making inscriptive-perspectives. This question is approached below in the discussion of the 'Bayesian revolution.'

All of this might indicate how R becomes believable, or how people are prepared to believe in R. But it doesn't say how belief actually is actually worked on in R. Here I think my position could move in two directions – either heading towards an analysis of some localised practices of working with data in R or towards an analysis of how the grounds on which data are made and analysed shift. This is returning to my broader position or hunch that a fairly subtle but important change in what counts as empirically credible is occurring in many domains. This change troubles the negotiated settlement between beliefs and events achieved in the last 200-300 years of statistical reasoning practice. Any such change is inevitably hard to infer, especially as it is taking place. How could we apprehend something whose forms, practices and position are fluxing?

What counts as data is subject to constant reshaping. Taking data and changing its form through various rearrangements has become a matter of central economic, commercial, political and scientific importance. While data has long been gathered and published in tables, in files and in databases, the injunction to re-use and re-analyse generates re-shaping or 'data munging' imperatives. We could approach these imperatives from the perspective of various discourses (transparency, sharing, accountability, democracy, Mertonian science, etc), or from the perspective of the world of analytics, data-mining and machine learning where data is seen as a valuable material whose inherent patterns promise new forms of economic, aesthetic or epistemic value. All of this deserves further analysis and comment. However, between the strategies of openness or pattern-finding and the data sources lies a terrain where data is re-shaped, transformed and plied into forms and patterns. The practices of plying, multiplying and applying data seem to me crucial in understanding belief in data.

As a programming language, R is striking for its '-ply' constructs. There are several in the core language and many to be found in packages (especially the popular 'plyr' package). These include *apply, sapply, tapply, lapply, mapply.* All of the -ply constructs have a common feature: they take some collection of things (it may be ordered in many different ways – as a list, as a table, as an array, etc), do something to it, and return something else. This hardly sounds startling. But while most programming languages in common use offer constructs to help deal with collections of things sequentially (for instance, by accessing each element of data set in turn and doing something), R offers ways of dealing with with them all at once. Deriving its -ply constructs ultimately from the functional logic developed by Alonzo Church in the 1930s, R presents difficulties for programmers used to so-called procedural programming languages. The functional programming style of applying functions to functions seems strangely abstract. Once you get a feel for them, these -ply constructs they are very convenient to write, and the code runs much faster. They implicitly parallelise operations on data, and this adapts well to the increasingly parallel contemporary chip architectures. The -ply operations reduce both data and computational frictions. The real stake in -plying data, however, is not speed but transformation. -Plying make data less like inscription (number, text, table, list, etc), and more like a pliable material, something that can be reshaped and folded. Such shifts in feeling for data are very mundane yet crucial to the flow of data. Plying, in the sense of plying trade, seems much closer to the kinds of work done on data than the figures of data flow or data deluge.

An endless supply of random variables

If -plying affects the flow of data, shifts in treatment of practices of probability might re-code belief in data. Here I mean 'in' in both senses – 'in' as internal to; in as concerning. Growing data flow, and desire for data flow is interlaced with growing belief in data. In order to frame this question more narrowly: what does it mean if we have, as we do in the form of R, 'the possibility of producing a supposedly endless flow of random variables for well-known or new distributions' (C. P. Robert & Casella 2010, p.42)? While random numbers are not new, an increase in supply of randomness has some interesting influences on belief in data. I don't have a lot of evidence for this, but one statistical approach that has become very widely used across various settings ranging from clinical trials to spam filters seems to me to point towards it: Bayesian statistics, a way of making statistical inferences that takes into account prior hypotheses or beliefs into account. Contemporary undergraduate statistics textbooks barely mention the groundswell of Bayesian change that has washed over statistical practice in the last twenty years. The average statistics textbook will mention Reverend Thomas Bayes' (1702-1762) rule or law for calculating conditional probability, but it will remain staunchly 'frequentist' in its treatment of inference.

Bayes' Rule is a rule concerning probability rather than statistics. In Bayes' own words:

> Given the number of times in which an unknown event has happened and failed [... Find] the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named (Bayes & Price 1763)

In probability talk: given the probability of A independently occurring, given the probability of B independently, and given the probability of B occurring if A occurs, what is the probability of A occurring given that B occurs? Like many 17-18th century approaches to probability (Pascal, Leibniz, Huyghens), Bayes' exposition is framed as a game of chance. Like Pascal's wager, and Leibniz's ars combinatoria, the probability talk, the examination of bets, wagers, chance (hazard) uses mathematics, but probability is itself still entwined with matters of belief.[1] The Bayesian revolution, as it is sometimes called, relies on degrees of 'subjective belief' and then modifies them in the light of new evidence. The most mundane example is probably the so-called 'naïve Bayes classifiers' used by many email spam filters. They constantly update their classification rules for spam depending on how previous email has been classified. For me, the Markov Chain Monte Carlo (MCMC) simulation-based processes are more evocative because they rely on massive injections of randomness to increase the credibility or believability of data. Since the 1990s, they have become increasingly prominent as a way of rendering Bayesian inference actually doable (C. Robert & Casella 2008, p.1). The so-called 'Bayesian revolution' in statistics has unfolded in the last two decades almost solely on the basis of this technique. If the -ply operations practically multiply the folds of data, methods like MCMC re-work belief in data in many different domains, but especially in statistically saturated fields of genomics, epidemiology, clinical trials, econometrics (etc., etc.,) and machine learning (as used in data-mining) more generally. MCMC is now said to be one of the 'top ten science and engineering algorithms of the 20 century' (Andrieu et al. 2003).

The origins of Monte Carlo methods in WWII atomic weapons research are well-known (Metropolis & Ulam 1949). Monte Carlo methods are simulation techniques used when it is impossible to calculate the results exactly. They repeatedly generate random sample variables, and then aggregate those results to produce an average result.

A standard Monte Carlo simulation can be seen as kind of random walk by a group of walkers, jumping all over the place randomly sampling values. Each sample is statistically independent of the others. The underlying assumptions for Monte Carlo techniques are thoroughly probabilistic, relying particularly on key 18-19th century formulations of the Law of Large Numbers and the Central Limit Theorem to warrant inference based on large numbers of random samples. Monte Carlo methods find broad application. Techniques that in 1950 were used to calculate radiation shielding are today used in everything from weather forecasting to to calculate pension yields personal finance planning in volatile market conditions (Savage 2005).

The confluence of Markov Chains with Monte Carlo methods to make MCMC is less familiar. Following the ins and outs of that confluence would perhaps be a little too baroque. I'm less interested in describing the technique itself than seeing how it brings changes in what data does. A Markov chain is a way to constrain the choice of random samples to follow particular paths. In an MCMC, rather than being independent, each sample is correlated. So an MCMC might be seen as a walk where a large ensemble of walkers has some relation – or more strictly speaking, correlation – to each other. The ensemble has a particular distribution. The implication is that the values resulting from their movements can be treated as random samples from a population distribution, and that together, these values provide a way of sampling the parameters of a population.

What does such a simulation arrangement deliver in to statistical inference, and to belief in data more generally? By staging many trials in silico, it opens the possibility of solidying inferences about the parameters of populations in the light of available evidence. Like the -plying operations, it reshapes forms, but this time not the given form of the data, but the inferred form – the distribution – of the population. What does all of this mean practically? It sounds so profound that one might expect startingly new forms of evidence or visualisations to result. The R code to run an MCMC-based inference is rather unexciting, and it produces rather conventional plots of population parameters (such as probability density function plots). So the visual culture of MCMC itself is familiar. However, we could track how this techniques feeds into the construction of more sophisticated modelling of populations of things on the move. Illustrative examples of how this works in contemporary scientific practice would include a study of the association of human *Campylobacter jejuni* infections in Lancashire with meat eating (Wilson et al. 2008); the use of Bayesian graphical models in the Manchester Asthma Study to 'take full advantage of the data-intensive environment provided by a birth cohort study (Simpson et al. 2010); or the statistical modelling of the dynamics of friendship networks in US high school students (GOODREAU et al. 2009).

Final remarks

Let me try to back out of this labyrinth. Between the 1660s and 1990s, what has happened to the internal evidence that connected belief and events?

1. Draping. These two examples – the -ply operations on data and MCMC sampled from R – could be taken as different edges or tendencies in contemporary data cultures more generally. The way in which they make data count differ; the levels of mathematical and computational sophistication they rely on vary. On the one hand, minute or specific techniques and practices address data frictions and seek to make data into something that changes shape like a lightly spun cloth, perhaps more like the fabric of talking. On the other hand, techniques that modulate concepts of evidence, inference and exactness through intensive injections of randomness seek to chisel out underlying forms – population parameters – that datasets drape.
2. Twisting. Statisticians, who are aware of the philosophical struggles over epistemic and frequentist notions of probability, readily make use of Bayesian approaches in a frequentist way. Regardless of what theory you hold about probability, the concept itself seems to operate autonomously. So the philosophical splits over probability are perhaps less interesting than the twisting or torsions in living, breathing data. Perhaps we can attribute this to the probability as a form of thought. Hacking's thesis about the status of internal evidence in allowing probability to emerge as a middle ground between demonstration and testimony is very resonant and provocative for thinking about data today. Data as 'internal evidence' – signs pointing beyond themselves that things display– is both a matter of belief and something in the world. The folding or plying operations, and the concern with making data flow differently – these practices can be understood differently if probability as sign-evidence is borne in mind. Data faces in one direction and move in another. It twists.

References

Andrieu, C. et al., 2003. An introduction to MCMC for machine learning. *Machine learning*, 50(1), p.5–43.

Bayes, T. & Price, R., 1763. An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S. *Philosophical Transactions of the Royal Society of London*, 53, p.370–418.

CRAN, 2010. The Comprehensive R Archive Network. Available at: http://www.stats.bris.ac.uk/R/ [Accessed May 5, 2010].

Daston, L., 1988. *Classical Probability in the Enlightenment*, Princeton Univ Press.

Daston, L., 1994. How probabilities came to be objective and subjective. *Historia mathematica*, 21(3), p.330–344.

Fienberg, Stephen E., 2006. When Did Bayesian Inference Become “Bayesian”? *Bayesian Analysis*, 1(1), pp.1–40.

GOODREAU, S.M., KITTS, J.A. & MORRIS, M., 2009. Birds of a Feather, or Friend of a Friend? Using Exponential Random Graph Models to Investigate Adolescent Social Networks. , 46(1), pp.103-125.

Hacking, I., 1975. The emergence of probability.

Hardy, Q., 2010. Power in the Numbers Page 2 of 3 - Forbes.com. *Forbes.com*. Available at: http://www.forbes.com/forbes/2010/0524/opinions-software-norman-nie-spss-ideas-opinions_2.html [Accessed June 10, 2010].

McKinsey & Company, 2009. Hal Varian on how the Web challenges managers -. *McKinsey Quarterly - Strategy - Innovation*. Available at: http://www.mckinseyquarterly.com/Hal_Varian_on_how_the_Web_challenges_managers_2286 [Accessed June 10, 2010].

Metropolis, N. & Ulam, S., 1949. The Monte Carlo method. *Journal of the American Statistical Association*, (44), p.335−341.

R-bloggers, 2011. R-bloggers | R news & tutorials from the web. Available at: http://r-bloggers.com/ [Accessed May 5, 2011].

R Development Core Team, The R Project for Statistical Computing. Available at: http://www.r-project.org/ [Accessed June 11, 2010].

Rexer Analytics, 2010. Rexer Analytics 4th Annual Data Miner Survey - 2010. Available at: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html [Accessed May 9, 2011].

Robert, C. & Casella, G., 2008. A History of Markov Chain Monte Carlo-Subjective Recollections from Incomplete Data. *0808.2902*. Available at: http://arxiv.org/abs/0808.2902 [Accessed January 1, 2011].

Robert, C.P. & Casella, G., 2010. *Introducing Monte Carlo Methods with R*, New York: Springer.

Savage, T., 2005. *The Savage Number: How Much Money Do You Need to Retire?*, John Wiley & Sons.

Simpson, A. et al., 2010. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *American Journal of Respiratory and Critical Care Medicine*, 181(11), pp.1200-1206.

Wilson, D.J. et al., 2008. Tracing the Source of Campylobacteriosis. *PLoS Genet*, 4(9), p.e1000203.

1 In the introduction he wrote to the essay discovered after Bayes' death, Richard Price interpreted Bayes' work theologically:

> The purpose I mean is, to shew what reason we have for believing that there are in the constitution of things fixt laws according to which things happen, and that, therefore, the frame of the world must be the effect of the wisdom and power of an intelligent cause; and thus to confirm the argument taken from final causes for the existence of the Deity (Bayes & Price 1763).

I don't want to make too much the historical connection between the law of conditional probability and religious belief, but Bayesian thought has continued to provoke philosophical discussion to this day. While most working statisticians acknowledge this, and then quickly move on, I want to suggest that the practices of doing Bayesian inferences cannot fully eradicate the ontological instabilities of probability.