# THE ISBA BULLETIN

*Vol. 9 No. 1* *March 2002*

The official bulletin of the International Society for Bayesian Analysis

## A MESSAGE FROM THE PRESIDENT

by David Draper
***ISBA President***
draper@ams.ucsc.edu

Broadly speaking (as you well know), the academic life is devoted to three kinds of tasks: research, teaching, and administration/service. In this issue of the *Bulletin* I have an item or two to share with you about each of these things, as they relate to us as Bayesians (in general) and as ISBA members (in particular). I'll address the three topics in reverse order.

First comes three announcements.

• After long, extremely effective, and much appreciated service as the Editor of the *ISBA Bulletin*, Fabrizio Ruggeri will be stepping down from this post after the issue you are now reading. (See the report from the Past President by Alicia Carriquiry, which follows this column, for a lengthier tribute to Fabrizio for his wonderful efforts.) I'm pleased to announce that the next Editor of the *Bulletin* will be **Hedibert Lopes** (hedibert@im.ufrj.br, also hedibert@stat.duke.edu). Hedibert earned Bachelor's and M.Sc. degrees in statistics from the Federal University of Rio de Janeiro (UFRJ) in Brazil, in 1991 and 1994, respectively, and went on to Ph.D. study at Duke University, obtaining his doctorate (with a dissertation on Bayesian analysis of latent factor and longitudinal models) in 2000. He is now an Associate Professor in the Department of Statistical Methods within the Institute of Mathematics at UFRJ. For the past several years he's been a member of the Macroeconometric Modeling group at the Brazilian Institute of Applied Economics Research, and he says that this has had a strong effect on his applied interests.

I ask you to join me in welcoming Hedibert to this important position. Please contact him if you're interested in becoming involved with the *Bulletin* editorial process.

• As a way to save money that will be invested in other aspects of producing the *Bulletin*, as a method for achieving timely delivery to the ISBA global membership, and as a small part of the process of firmly staking out ISBA's digital territory in the 21st century :-), after consultation with the membership the ISBA Executive has decided that, starting with the next issue, the default method of distribution of the *Bulletin* will be electronic. You will be notified sometime this summer about how to obtain the next (and subsequent) issues, and you will always have the option of requesting a hardcopy version instead.

• Planning is underway for **ISBA2004**, our next big quadrennial World Meeting, but the location of the meeting has not been finalized. Luis Raul Pericchi (luispericchi@yahoo.com or pericchi@goliath.cnnet.clu.edu), current Chair of the Program Council, would like to receive bids from people interested in hosting the meeting. As was the case in the planning for *ISBA2000* in Crete, priority will be given to locations that (a) are relatively easy to reach and accessible (not too expensive) for members from all over the

*Contents*

world, (b) have a good Local Arrangements Committee, and (c) are committed to either providing or helping to raise some funding. To get full consideration bids will need to be received by Luis no later than Monday 1 July 2002. More details on how to put together a successful bid will be available soon on the ISBA website and in an email to `bayes-news`.

Second, on the subject of teaching: as one of my "campaign promises" (an item in the statement I was asked to write for the September 2000 issue of the *Bulletin* on what I would try to do if elected President) I said that I thought it would be a good moment to take stock of how (Bayesian) statistics is taught (if at all) in educational systems around the world prior to Bachelor's study in universities. It's clear, at least in the United States, that *some* statistics is being taught in high school, because (as Jessica Utts told me on a recent visit to the University of California, Davis) 50,000 students will take the statistics Advanced Placement (AP) exam (which, in principle, gives students credit for one or more introductory courses in probability and statistics at university) in May 2002. As an example of what's going on out there, `www.cde.ca.gov/standards/math/ap.html` contains a summary of the curriculum taught in California high schools to prepare students for the AP exam. There are 19 categories of learning and achievement, covering the following topics: basic probability calculations involving *and*, *or*, *not*, and conditional probability on finite sample spaces; discrete and continuous random variables, including calculation of means and variances of discrete distributions; basic calculations with the normal, binomial, and exponential distributions; the Central Limit Theorem and its use in forming the normal approximation to the binomial distribution; simple descriptive statistics as applied to real data sets (including measures of center and spread, histograms, scatterplots, stem-and-leaf plots, and boxplots); calculating the equation of the least-squares line in simple linear regression; calculation and interpretation of the correlation coefficient; the use of repeated-sampling reasoning to construct confidence intervals for population means in simple Gaussian problems and to make simple sample size calculations; calculation of $P$-values based upon simple random sampling from a normal distribution; and the use and interpretation of the $\chi^2$ test.

I don't know about you, but I personally had no idea until quite recently that high school students were learning this much probability and statistics. In particular, did you know that 17-year-olds were already having their scientific lives ruined `:-)` by exposure to frequentist $P$-values as the only approach to non-interval-estimate-based inference? If you have strong views on the teaching of statistics to people before they arrive at university, I would like to hear them; please email me. In particular, I would be especially interested to hear from those of you who can relate facts and details about how the subjects of probability and statistics are being taught to pre-university students in countries other than the U.S. During my Presidential year I will try to draw together information on this topic and get a discussion going among ISBA members on what, if anything, we can do to improve the teaching of our subject, especially at the delicate moment when people are hearing about probabilistic and statistical ideas for the first time.

And finally, on the topic of research, since moving to the Engineering School at the University of California, Santa Cruz (UCSC), about a year ago (where my Department shares intellectual and physical space with a number of extremely sharp people in computer science, computer engineering, electrical engineering, and bioinformatics), I've begun to speak with some regularity, for the first time in my life, with people who are machine learning experts. You probably already know (I'm coming rather late to this party, I'm sure) that these people (a) are working on extremely interesting statistical problems and (b) have a lot of ideas that we can benefit from (we have some good ideas for them, too, but that's another story). I recently posed the following problem to some of the machine learning people at UCSC; they told me that as far as they knew it was still an unsolved problem, but they encouraged me to continue asking people, and I'd like to relate this problem to you in the hope that somebody knows if any progress has been made on it.

The history of how Markov chain Monte Carlo (MCMC) ideas entered Bayesian statistics (see the article on "MCMC and Bayesian Data Analysis: The Beginning" by Alan Gelfand in the March 2000 *Bulletin* for more on this topic) is both interesting and embarrassing: the former for the cautionary tale it provides about dissemination of scientific knowledge, the latter because it took us so long to stumble onto a set of techniques that (i) merely revolutionized our subject but which (ii) had been sitting around for more than 30 years waiting to be "discovered." In oversimplified sketch form, (A) Metropolis et al., writing in the chemical physics literature, published the key MCMC ideas in 1953, where they were unknown to (almost all) statisticians for at least 15 years; (B) Hastings generalized these ideas and tried to promote them to a statistical audience by publishing in *Biometrika* in 1970, but his work was also largely ignored for a very long time; (C) Geman and Geman independently reinvented a particular version of the Metropolis-Hastings idea (Gibbs sampling) in the early 1980s, but they published in the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, which was not (shall we say) widely read by Bayesians in those days; (D) Tanner and Wong independently developed something quite a lot like Gibbs sampling and published it in *JASA* in 1987; and (E) finally the penny dropped and the floodgates opened (to mix metaphors) with the appearance of the Gelfand and Smith paper in *JASA* in 1990. I am glossing over the key fact that computers weren't fast enough to do MCMC calculations on a widespread basis until the late 1980s, which certainly played a part, but I think you would agree with me that this is not a good track record for interdisciplinary communication of ideas.

And the problem is that it's only going to get worse, as the number of people doing interesting science in a growing number of fields steadily increases. It's hard enough to keep up with the literature in a single subject, and yet whatever hard problem you've recently been thinking about, it's fairly likely that somebody in some field has already had a good idea that would help you solve your problem, if you only knew about it. Search engines can help to some extent, but are often defeated by the fact that when good ideas develop in two or more different disciplines they will typically be referred to using totally different technical terms. Even if you had access to a good search engine in the period from 1953 to 1990, a fast computer to run it on, and a complete data base of titles, abstracts, and keywords of all papers in all fields, you would have been hard pressed to run a search to solve the fundamental problem of Bayesian computation (accurate approximation of high-dimensional integrals) in such a way that the Metropolis et al. paper would have appeared in your search, because the language with which Metropolis et al. talked about approximating *their* high-dimensional integrals was so different from the way Bayesians talk about the problem (even though it's the same mathematical challenge).

What we all need is an *analogy engine*: if I want to solve problem $A$, I'd like to be able to ask a computer program to find all of the literature, in any field whatever, that has made progress in solving problems *similar to* problem $A$. The machine learning people at UCSC tell me that the analogy engine doesn't exist yet, and not much progress has been made toward its creation, but there is no theoretical reason why it could not be built (possibly with a fair amount of initial human intervention in defining the operator *similar to* on problem space). My questions to the ISBA membership are this: Do you know of any significant progress toward the creation of something like the analogy engine? If so, please let me know, and I will pass on what I learn in future columns. If not, can you think of any other way to defeat the Tower-of-Babel problem (too much scientific work being published, and not enough tools for nonspecialists to gain ready access to it)?

## One last word from the Past President

by Alicia Carriquiry
***Past ISBA President***
alicia@iastate.edu

To those of you who are beginning to think that I will never go away, I would like to promise that this truly, inequivocably, is my very last farewell article.

I simply could not leave without thanking Fabrizio Ruggeri on behalf of the entire ISBA membership, as he embarks in the process of handing the ISBA Bulletin over to Hedibert Lopes. Fabrizio took over the Bulletin in 1998, and never looked back. During his tenure as Editor, the Bulletin underwent a deep transformation, both in format and in content. In fact, it even changed name, from Newsletter to ISBA Bulletin. As an Editor, Fabrizio has been imaginative, hard-working, forward-looking, and has provided a most significant service to the society, helping us all in spreading the Bayesian word. The Bulletin is interesting, always informative, and I for one look forward to each issue. [It is at this point that I should also point out that Fabrizio, in spite of relentness but invariably gentle nudging, has often waited for weeks and weeks to get promised articles, so delays in the publication of the Bulletin are strictly the responsibility of people like me! Hedibert, heads up!]

So in closing, I will simply say THANK YOU Fabrizio! It has been a pleasure working with you on Bulletin matters. You have been generous with your time and efforts in spite of some very difficult challenges in your personal life. As an ISBA member, I look forward to benefiting from your leadership in your new role as Director of ISBA.

## An update on Valencia 7

by Hal Stern
***SCP Selection Committee Chair***
hstern@iastate.edu

Planning for the Valencia 7 meeting, to be held June 2 through June 6, 2002 in Tenerife, Spain continues. The Program Committee (Bayarri, Berger, Bernardo, Dawid, Heckerman, Smith, West) has announced the invited program consisting of 25 invited speakers and 25 invited discussants. The invited program and local arrangements information can be found at the conference web site www.uv.es/valencia7 or the US mirror site www.stat.duke.edu/valencia7.

A new feature for this Valencia meeting is the incorporation of Selected Contributed Papers for oral presentation. Fifty contributed papers, chosen from more than 120 submitted by the October 15th deadline, were selected for presentation. These will be organized into sessions that will run concurrently with the invited program. The Selected Contributed Papers review committee (Hal Stern, Chair) found the contributions to be of extremely high quality and the selection of the final 50 was difficult. The selected papers consider a wide range of topics from foundational issues to applications in fields as diverse as software testing and neurophysiology. In addition, there will be the usual high quality poster presentations for which Valencia is famous! All of the contributed paper authors whose work was not selected for oral presentation have been encouraged to participate in the poster program.

### Latest news from Valencia 7

Brad Carlin is organizing the V7 cabaret, on Thursday June 6th night. He asked Jose Bernardo to distribute this message to all participants: "Anyone who is interested in performing at the cabaret (and this includes musicians, actors, jugglers, magicians, joke tellers, men that like dressing as women, etc.) should feel free to contact either me (brad@biostat.umn.edu), Mark Glickman (mg@math.bu.edu) or Tony O'Hagan (a.ohagan@sheffield.ac.uk) either now or at the conference. All contributions gratefully accepted, no matter how silly!"

## GOOD-BYE FROM THE EDITOR

by Fabrizio Ruggeri
***ISBA Bulletin Editor***
fabrizio@iami.mi.cnr.it

Here I am one more time but this looks like it is the last one. The search for a new Editor took a while but, eventually, it ended with an excellent choice. I know Hedibert Lopes since when he was a Ph.D. student at Duke University and, as soon as he went back to Brazil, I asked him to become the Corresponding Editor for the Bulletin from his country. He contributed a paper in the June 2001 issue on "Sailing the Bayesian Boat in a Hostile Sea", where he expressed his views on Bayesian teaching. Hedibert has been one of the most active Corresponding Editor the Bulletin had in the past. I am very confident Hedibert will do a great job as the Editor of the Bulletin: he has qualities and energy. An editorial change is a good opportunity to improve the Bulletin, bringing new ideas and people. I am sure Hedibert has plenty of ideas but he needs your comments and suggestions.

*In bocca al lupo, Hedibert!*

(Have you ever considered learning Italian? It is almost useless worldwide, but not at Bayesian meetings!)

I am really flattered by the words written by David and Alicia in the first pages of the current issue. As I already wrote in past issues, I believe all the achievements of the 13 issues I have been in charge of are the results of a team work: the people who wrote papers, the Associated Editors who "haunted" for papers, the Corresponding Editors (I apologise for not mentioning Polasek in the previous issue) who contributed news, the people who gave me advices at the very beginning and later, and the ISBA officers (first of all, the Presidents I have been working with: Bayarri, Geweke, Dawid, Carriquiry and Draper) with whom the cooperation has been very fruitful. All the names of the people (more than 100!) who contributed can be found at www.iami.mi.cnr.it/isba, the "old" Bulletin web page.

Special thanks go to the people who helped me in preparing the current issue and were able to get their papers very quickly once we were asked to serve for one more issue (nonetheless, the Bulletin is delivered late but, in a different part of it, you can find a honest, deeply appreciated, "confession" ...). You can see their names in the papers we published (and Brunero is an excellent "new entry"!), except for two who did not sign the papers they got for the Bulletin: Gabriel Huerta who went back to the job of Associate Editor for the Software section he held in 1999-2000 and Eugenio Regazzini, who kindly agreed to let me translate parts of an interesting work he wrote in Italian. I am aware that a short review on the laws of large numbers before 1950 is not "Bayesian history" but I believe it is a very interesting topic Bayesians should know about.

I have been very lucky in serving as the Bulletin Editor in the past years, in which ISBA has turned into a well established society. When I began (late 1998), big plans were under discussion: rules governing the Society, a new world conference (which turned out to be the impressive ISBA 2000 in Crete) and a renewed Bulletin, just to mention few of them. After achieving these goals, new, more ambitious, ones are ahead of us. David's opening paper illustrates some of them and ISBA members can discuss them at the General Assembly at Valencia 7 and on the Discussion Forum on the ISBA web site. I already had the opportunity to take part in a very lively and interesting discussion when those issues were raised in the ISBA Board of Directors. It was my first act in my new position, where I can serve the Society I am part of since its foundation (it looks like ages passed since the 1993 San Francisco's meeting!).

From these columns, I told you what was going on with the Bulletin. Now let me close with a thought at the world outside: here is the English translation (not mine!) of a poem from *If This is a Man* by the Italian writer Primo Levi, an Auschwitz survivor.

*You who live safe*
*In your warm houses,*
*You who find, returning in the evening,*
*Hot food and friendly faces:*
*Consider if this is a man*
*Who works in the mud*
*Who does not know peace*
*Who fights for a scrap of bread*
*Who dies because of a yes or a no.*
*Consider if this is a woman,*
*Without hair and without name*
*With no more strength to remember,*
*Her eyes empty and her womb cold*
*Like a frog in winter.*
*Meditate that this came about:*
*I commend these words to you.*
*Carve them in your hearts*
*At home, in the street,*
*Going to bed, rising;*
*Repeat them to your children,*
*Or may your house fall apart,*
*May illness impede you,*
*May your children turn their faces from you.*

## Rob Kass

by Brunero Liseo
brunero.liseo@uniroma1.it

Professor Kass does not need any introduction. He is Professor and Head of the Department of Statistics at Carnegie Mellon University. Among the many valuable contributions of Professor Kass to our discipline I would like to mention his effort to bridge theoretical and applied statistics, as the numerous Case Studies Workshops held at CMU testify.

You can find out more information from the Departmental homepage at:

www.stat.cmu.edu/~kass

We e-mailed Professor Kass a number of questions about his professional story and his personal view of Statistics. Here are his responses.

`1. Rob, you have recently said: ''Everything I've really needed to know (about Statistics), I learned in graduate school''. Who were, at that time, the most influential teachers for you?`

I had many wonderful teachers who were important to my development, including Michael Perlman and Steve Stigler. I was also fortunate to sit through a series of lectures by Dennis Lindley and another by Arnold Zellner, both of whom were very helpful to me through many subsequent discussions. But the most influential person was David Wallace, co-author of the greatest Bayesian case study, *Inference and Disputed Authorship: The Federalist*. David was (and still is) an insightful thinker with an unrelenting intellect. Philosophically, David seemed to be a combination of Tukey, Fisher, and Savage: he instilled in everyone around him a sense that data analysis is a deep subject, that Fisher's methods were nearly always sensible, but that when willing to think carefully, one ought to be Bayesian.

`2. Your research is characterized by strong mathematical knowledge and ingenuity. Still, you have paid a lot of attention to applications, as the Case Studies Workshop at CMU demonstrates. Is this the right mixing of abilities that a statistician should have nowadays?`

Throughout college and graduate school I had both theoretical and applied interests, but when I finished my Ph.D. and was looking for a job one of my professors, Paul Meier, told me that I had to decide whether I was a theoretical or applied statistician. What a strange statement! Yet, he was right: back then, in order to be successful, most of us had to choose. So I chose theory, but I missed scientific applications terribly. It was not until I became a full professor that I felt the security to pursue applied work seriously. These days our field is clearly righting the imbalance created by the excessive influence of mathematical statistics in the U.S., and many other countries, during roughly the third quarter of the 20th century. I trust that nearly everyone now beginning an academic career will be doing at least some cross-disciplinary work in conjunction with their methodological research.

`3. Tell us something about your interest in cognitive neuroscience; what is the role of statistics there?`

Neuroscience is a vast enterprise in need of statistical help with things ranging from clinical trials, to brain imaging, to understanding neuronal firing patterns. On the latter subject we are holding a workshop here at Carnegie Mellon, May 3-4. Details about this and my own interests, and also related links, may be found at my website www.stat.cmu.edu/~kass. It is a fascinating area of science.

`4. What is your personal view of a good graduate program in Statistics? Do you see this type of program in Statistics Departments around the world, or do you believe there are some important aspects that are missing?`

Together with my colleagues I have put much effort into constructing the graduate program at Carnegie Mellon. I think highly of the training we offer, but our program is only one among many fine alternatives throughout the world. The important thing is to help students become well-rounded statisticians who know basic theory and methods, can collaborate in cross-disciplinary projects, and have good communication skills.

`5. A major merit that is to be ascribed to you is the popularization of Jeffreys' ideas. In your opinion, what is the main message of his book` *Theory of Probability*`? Why should a young researcher read it?`

In terms of his influence on statistical practice, Jeffreys claimed some credit for getting physicists to include appropriate standard errors when reporting results. This may sound trivial, but obtaining simple and meaningful assessments of uncertainty remains the most important accomplishment of most Bayesian analyses.

Although he failed to appreciate fully the importance of the subjectivist viewpoint, Jeffreys correctly perceived the practical necessity of default methods. Thus, we have "Jeffreys's Prior" (really several priors). His other major contribution was development of the Bayesian theory of testing. This, however, remains problematic in many applications: there are important situations where Bayes factors are useful, but Jeffreys did not recognize their limitations, particularly in multiparameter settings. I would like to think my review papers, with Larry Wasserman and with Adrian Raftery, have helped to clarify the issues and thereby encouraged good practice. Among my own papers, these two are actually my favorites. I wish more people would write reviews. They are hard work, but very rewarding.

Interested students of statistical thinking who persevere in reading fundamental original sources, such as Jeffreys's book (which can be difficult because of its style), may be pleasantly surprised to find unique perspectives on old and seemingly standard ideas, which lead to valuable insights.

`6. What is your personal forecast about the role of Bayesian statistics within statistics and in connection with other disciplines? Should Statistics become just one?`

When I completed my Ph.D. thesis Bayesian statisticians were a small sect, convinced that they knew the true path to reasoned scientific inference but having no one else to talk to. Some of the people who advised me in 1981 wondered why I was interested in Carnegie Mellon. At that time it was known mainly as the place Morrie DeGroot happened to be. Jay Kadane, although Department Head, was a relatively young faculty member and Steve Fienberg had just been recruited. True, my advisors conceded, there was a core that could build a strong Bayesianly-oriented group but it was not clear that they would become outstanding among statistics departments as a whole. I remember explaining that I felt someday those impossible-looking integrals would be evaluated numerically and that the data-analytical world would then become much more Bayesian. I wanted to be in a place where Bayesian statistics was respectable. Fortunately for me, the revolution started shortly after I began my career.

Should Statistics become just one? In a certain sense, yes, and in my view it already has: Bayesian inference now stands alongside frequentist methodology as part of mainstream Statistics, each being used—by competent analysts—to do what it does well. (Frequentist methods remain good for very simple and standard analyses, for assessing goodness-of-fit, and for many nonparametric problems.) My forecast is that Bayesian statistics will flourish and grow rapidly along with the rest of the field, at least in the near future. I am continually hearing of new areas where Bayesian methods are being used, and it is rare to come across someone who refuses to apply Bayesian methods on principle. The days of stigma and serious dispute are over.

`7. Do you have any particular suggestion to give to a young graduate student in statistics?`

I would advise any young person to listen to their inner judgment and follow what inspires them most, while also considering what they are particularly good at doing. When we are young we don't know how high we can climb. It is best to aim for the top, however that might be defined, expecting eventually to find a comfortable niche that will continue to offer new challenges.

Thanks to Rob for his stimulating and informative answers.

# BAYESIAN ECOLOGICAL APPLICATION IN *Biometrics*

by Kate Cowles
kcowles@stat.uiowa.edu

In a recent issue of the *ISBA Bulletin*, we surveyed Bayesian applications papers that appeared in *JASA* over the past 5 years. In this column, we look at an unusual applied problem addressed by Bayesian methods in the December, 2001, issue of *Biometrics*.

In the context of decline in the population of many bird species, the research question is the reproductive potential of a particular species of birds.

One measure of interest is the probability of "nest success". The authors mention two definitions of "nest success", which have been used in past studies by biologists. The first is that at least one egg hatches; the second is that at least one young bird fledges. The typical design for a nest-survival study is for the researchers to locate active nests in the study area and then to recheck each at least once. At each visit, the nest status (failed, still active, or has succeeded) is recorded.

He, Sun and Tra cite previous frequentist approaches to analyzing nest survival data, which assume that the hazard of nest failure is constant over the active life of a nest. This assumption probably is not realistic, particularly when the definition of success involves fledging, since the risks to unhatched eggs are likely to be very different from those to hatched chicks.

He, Sun and Tra analyze data from a nest success study of mourning doves conducted at the Patuxent Wildlife Research Station in Laurel, MD (Nichols, *et al.*, 1984). In this study, successful fledging was defined as at least one living nestling in the nest 10 days after hatching. A nest was required to survive for 26 days after egg-laying to be considered successful. In analyzing the data collected, Pollock and Cornelius (1988) and Heisey and Nordheim (1995) grouped the data into consecutive intervals of 8, 8, and 10 days. Because the original daily data were not available to them, He, Sun, and Tra used the grouped data. Thus a nest had to survive through the third interval ($J = 3$) to be successful.

He, Sun, and Tra use the following notation:

- $y = 1$ if the nest is successful and 0 otherwise
- $u$ is the time interval after egg-laying during which the nest is discovered
- $z$ is the number of time intervals during which the nest is observed before a failure or a success is determined
- $t$ is the age of the nest when the failure or success is determined
- $\delta_k$ is the conditional probability that a nest is discovered during time interval $k$ given that it is discoverable: $\delta_k \; = \; P(u = k|u \leq J)$
- $q_k$ is the probability of nest failure during the time interval $q_k \; = \; P(t = k, \; Y = 0), k = 1, 2, 3$
- $q_4 = 1 - q_1 + q_2 + q_3$ is the probability of nest success

In the study, 59 nests were observed of which 25 were successful and 34 failed. Of the successful nests, 11 were discovered no more than 8 days after egg-laying (u = 1), 8 were discovered when the eggs were 9-16 days old (u = 2), and for the remaining 6 nests, u = 3. Among unsuccessful nests, 16 failed within 8 days of discovery (z = 1), 9 failed between 9 and 16 days after discovery (z = 2), and the remaining 9 failed after the 16th day (z = 3).

The likelihood of $\delta$ and $\mathbf{q}$ given the data $(y_1, z_1; y_2, z_2; \cdots; y_n, z_n)$ from $n$ observed nests is

$$L(\delta, \mathbf{q}) \; = \; A^{-n}$$

$$\times \prod_{k=1}^{n} (\delta_{J-z_k+1} q_{J+1})^{y_k} \left( \sum_{l=z_k}^{J} \delta_{l-z_k+1} q_l \right)^{1-y_k}$$

where

$$A = \left( \sum_{j=1}^{J} \delta_j \right) q_{J+1} + \sum_{i=1}^{J} \left( q_i \sum_{j=1}^{i} \delta_j \right)$$

He, Sun, and Tra placed independent noninformative Dirichlet priors on the probability vectors $\delta$ and $\mathbf{q}$, i.e. $\delta \sim D(1, 1, 1)$ and $\mathbf{q} \sim D(1, 1, 1, 1)$.

The authors used Gibbs sampling with data augmentation to fit their model. The posterior mean (standard deviation) of $q_4$ (probability of nest success) was 0.3481 (0.0625).

The proposed Bayesian modeling strategy does not require grouping of data, but it does require that nests be visited frequently and regularly, which may not be the case in real field studies.

**➤ References**

HE, C.A., SUN, D. and TRA, Y. (2001). Bayesian Modeling of Age-Specific Survival in Nesting Studies Under Dirichlet Priors. *Biometrics*, **57**, 1059-1068.

HEISEY, D.M. and NORDHEIM, E.V. (1995). Modeling age-specific survival in nesting studies, using a general approach for doubly-censored and truncated data. *Biometrics*, **46**, 51-60.

NICHOLS, J.D., PERCIVAL, H.F., COON, R.A., CONROY, M.J., HENSLER, G.L. and HINES, J.E. (1984). Observer visitation frequency and success of mourning dove nests: A field experiment. *The Auk*, **101**, 398-402.

# A LIBRARY FOR REGENERATIVE MCMC SIMULATION

by Anthony E. Brockwell
abrock@stat.cmu.edu

Regenerative Markov chain Monte Carlo simulation is a technique which relies on the ability to break up a Markov chain into a sequence of independent and identically distributed segments. It is proposed by a number of authors, including Crane and Lemoine (1977), Ripley (1987), Mykland *et al.* (1995), Robert and Casella (1999), Brockwell and Kadane (2001), Jones and Hobert (2001), as a means of (1) obtaining honest variance estimates for parameters and (2) avoiding the burn-in issue. Further benefits include the ability to construct a single chain by "patching" together segments constructed on parallel processors, as well as the ability [see Gilks *et al.*, 1998] to introduce adaptivity into an MCMC kernel.

Mykland *et al.* (1995) develop a nice method for identifying regeneration points in a chain, based on establishing that the chain satisfies a so-called "minorization" condition. Unfortunately, their method has not yet been widely adopted, possibly because people cannot be bothered showing that the minorization condition is satisfied for their particular problems.

The Regenerative MCMC Simulation (RMCMCS) Library aims to make regenerative simulation transparent to a user, and is based on Algorithm 3.1 of Brockwell and Kadane (2001). The idea [although presented in a different way to that of Mykland *et al.* (1995)] is essentially to construct a hybrid transition kernel, which by its construction, automatically satisfies the minorization condition. The hybrid kernel consists of two components. One of these looks like the usual transition kernel for an MCMC problem, and the other (implemented within the RMCMCS library) ensures that the chain will be regenerative, and provides a simple way of identifying the regeneration times.

The user has only to supply C/C++ code which implements a standard (non-regenerative) MCMC simulation. To link with the RMCMCS Library, the user's MCMC code must conform to a set of specifications described in the RMCMCS Library documentation. In addition, a function must be supplied evaluating the log-joint likelihood (plus an arbitrary constant) of the parameters. After this, there are a couple of additional tuning-parameters which must be chosen.

The end result is that the user can make a call to a function which constructs a regenerative chain and returns estimates of parameters, as well as estimates of the variances of the estimates themselves. In addition, for systems which have the standard "Message-Passing Interface" (MPI) library [see, e.g. Gropp *et al.* (1999)], the RMCMCS Library can automatically "parallelize" generation of the chain, yielding a speed increase roughly proportional to the number of processors available. (Note that the MPI library is a freely-distributed package available for most operating systems, which should be relatively straightforward for most system administrators to install. No special hardware is required - it can function on, for instance, a department network of Unix machines, simply using existing ethernet connections.)

Of course, Algorithm 3.1 of Brockwell and Kadane (2001) is not the only way (or indeed, the best way) to obtain a regenerative chain. However, it does provide a simple way, as implemented in the RMCMCS library, of introducing regeneration into an existing non-regenerative chain, without doing too much thinking.

The RMCMCS Library is freely available, for non-commercial purposes, at the StatLib archive at `lib.stat.cmu.edu/general`. In its current form, it is distributed as source code, along with documentation and some simple examples. In order to use the library, the user needs to have some C programming skills, and ideally will have already written C code to implement a standard MCMC simulation. To compile with the RMCMCS library, the user's computer is required to have the (standard) LAPACK library installed, and, if the user wants to use parallel processing, the MPI library as well. Current limitations of the algorithm mean that it is only effective for problems in which the state-space dimension is no larger than around 30 or so, although the author hopes to get around this problem in the future.

**➤ References**

BROCKWELL, A.E and KADANE, J.B. (2001). Practical regeneration for Markov chain Monte Carlo simulation. *Technical Report 757*, Carnegie Mellon University Dept. of Statistics.

CRANE, M.A. and LEMOINE, A.J (1977). *An Introduction to the Regenerative Method for Simulation Analysis*, Lecture

Notes in Control and Information Sciences, **4**, Springer.

GILKS, R., ROBERTS, G.O. and SAHU, S.K. (1998). Adaptive Markov chain Monte Carlo through regeneration. *Journal of the American Statistical Association*, **93**, 1045-1054.

GROPP, W., LUSK, E. and SKJELLUM, A. (1999). *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. The MIT Press.

JONES, G.L. and HOBERT, J.P. (2001). Honest exploration of intractable probability distributions via Markov chain Monte Carlo. *Statistical Science*, **16**, 312-334.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association*, **90**, 233-241.

RIPLEY, B.D. (1987). *Stochastic Simulation*. Wiley.

ROBERT, C.P. and CASELLA, G. (1999). *Monte Carlo Statistical Methods*. Springer.

## STOCKHOLM WATER PRIZE AWARDED TO IGNACIO RODRÍGUEZ-ITURBE

by Bruno Sansó
bruno@ams.ucsc.edu

We all know that there is no Nobel Prize in Mathematics but probably few of us know that there is something very close to a Nobel "Water" Prize. In fact the $150,000 Stockholm Water Prize has been presented for the last eleven years by the Stockholm Water Foundation. The Prize recognizes outstanding research and activities that increase knowledge of water as a resource and protects its usability for all life. King Carl XVI Gustaf of Sweden is the patron of the Stockholm Water Prize and will present the award to Professor Ignacio Rodríguez-Iturbe this year.

Ignacio is a Venezuelan hydrologist who, developing statistical and applied probability methods, has made a substantial contribution in making Hydrology a respected and mature science area.

This is relevant to the Bayesian community since in the mid 1970s, Ignacio introduced Bayesian approaches to improve different models for river flows and to predict the likelihood of extreme hydrological events.

Ignacio has made profuse use of statistical methods to formulate a mathematical representation of rainfall as random, active point processes. Because of this it is now possible to simulate rainfall patterns in time and space over many years, creating sequences that mimic how nature may behave in the future and use the results in engineering design or analysis.

Ignacio was born in Venezuela, he is a retired Professor of the Universidad Simón Bolívar, Caracas, and now works in Princeton University (his web page is www.princeton.edu/~irodrigu/. In Venezuela and the USA he supervised several students, both in engineering and statistics, working on dissertations that applied Bayesian thinking for a wide variety of hydrological problems. Like, for example, using regional information to assess priors to enlarge scarce data sets, selecting hydrological models using Bayes Factors and risk assessment of floods. As I pointed out in the article about Bayesians in Venezuela in the March 1999 issue of the Bulletin, when Ignacio went back to Venezuela after finishing his PhD, he was the spark that fired the interest for Bayesian statistics in the country. Something for which Venezuelan statisticians are very grateful. Please join me in congratulating Ignacio for his new achievement.

(In producing this note I made liberal use of the information in www.siwi.org/swp/swp.html)

### NEW ISBA BULLETIN EDITOR

HEDIBERT FREITAS LOPES IS THE NEW ISBA BULLETIN EDITOR

PLEASE CONTACT HIM AT

hedibert@im.ufrj.br

FOR COMMENTS AND SUGGESTIONS

# VARIABLE SELECTION IN REGRESSION

by Duncan K. H. Fong
i2v@psu.edu

We present recent Bayesian work on variable selection in linear regression models, with special emphasis on the case where the number of independent variables may be larger than the number of observations. More references on variable selection methods can be found in George (2000) [*Journal of the American Statistical Association*, **95**, 1304–1308.]

• E. I. GEORGE AND R. E. MCCULLOCH (1993) **Variable Selection via Gibbs Sampling**, *Journal of the American Statistical Association*, **88**, 881–889.

A crucial problem in building a multiple regression model is the selection of predictors to include. The main thrust of this article is to propose and develop a procedure that uses probabilistic considerations for selecting promising subsets. This procedure entails embedding the regression setup in a hierarchical normal mixture model where latent variables are used to identify subset choices. In this framework the promising subsets of predictors can be identified as those with higher posterior probability. The computational burden is then alleviated by using the Gibbs sampler to indirectly sample from this multinomial posterior distribution on the set of possible subset choices. Those subsets with higher probability–the promising ones–can then be identified by their more frequent appearance in the Gibbs sample. Using proper priors, the method can be applied even when the number of independent variables is greater than the sample size.

• A. O'HAGAN (1995) **Fractional Bayes Factors for Model Comparison**, *Journal of the Royal Statistical Society, Series B*, **57**, 99–138.

Bayesian comparison of models is achieved simply by calculation of posterior probabilities of the models themselves. However, there are difficulties with this approach when prior information about the parameters of the various models is weak. Partial Bayes factors offer a resolution of the problem by setting aside part of the data as a training sample. The training sample is used to obtain an initial informative posterior distribution of the parameters in each model. Model comparison is then based on a Bayes factor calculated from the remaining data. In this paper, properties of partial Bayes factors are discussed, particularly in the context of weak prior information, and they are found to have advantages over other proposed methods of model comparison. A new variant of the partial Bayes factor, the fractional Bayes factor, is advocated on grounds of consistency, simplicity, robustness and coherence.

• J. O. BERGER AND L. R. PERICCHI (1996) **The Intrinsic Bayes Factor for Linear Models**, in *Bayesian Statistics–Proceedings of the 5th Valencia International Meetings Held in Alicante, June 5-9, 1994*, (Eds.) J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford Science Publication, **5**, 25-44.

In Berger and Pericchi (1996) [*Journal of the American Statistical Association*, **91**, 109-122.] a general automatic Bayesian method for comparing models, the Intrinsic Bayes Factor (IBF) was proposed. One version, the Arithmetic IBF, was shown to essentially correspond to an actual Bayes factor for a reasonable Intrinsic Prior. A second version, the Geometric IBF, is justified in Pericchi and Smith (1994), using a prequential type of loss function, without assuming that one of the entertained models is the true sampling model. In this paper the authors analyze the general normal linear model, determining the intrinsic Bayes factors for any model comparisons, nested or separate, as well as for multiple model comparisons. In these situations they also calculate the Expected Arithmetic IBF. The authors also generalize model elaboration ideas to linear models with fixed mean structure but arbitrary error distributions. The method is illustrated on examples and compared with other model selection methods.

• J. HOETING, A. E. RAFTERY AND D. MADIGAN (1996) **A Method for Simultaneous Variable Selection and Outlier Identification in Linear Regression**, *Computational Statistics and Data Analysis*, **22**, 251-270.

In this paper the authors suggest a method for simultaneous variable selection and outlier identification based on the computation of posterior model probabilities. This avoids the problem that the model you select depends upon the order in which variable selection and outlier identification are carried out. Their method can find multiple outliers and appears to be successful in identifying masked outliers.

The authors also address the problem of model uncertainty via Bayesian model averaging.

For problems where the number of models is large, they suggest a Markov chain Monte Carlo approach to approximate the Bayesian model average over the space of all possible variables and outliers under consideration. Software for implementing this approach is described. In an example, they show that model averaging via simultaneous variable selection and outlier identification improves predictive performance and provides more accurate prediction intervals as compared to any single model that might reasonably be selected.

• J. GEWEKE (1996) **Variable Selection and Model Comparison in Regression**, in *Bayesian Statistics–Proceedings of the 5th Valencia International Meetings Held in Alicante, June 5-9, 1994*, (Eds.) J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Oxford: Oxford Science Publication, **5**, 609-620.

In the specification of linear regression models it is common to indicate a list of candidate variables from which a subset enters the model with nonzero coefficients. This paper interprets this specification as a mixed continuous-discrete prior distribution for coefficient values. It then utilizes a Gibbs sampler to construct posterior moments. It is shown how this method can incorporate sign constraints and provide posterior probabilities for all possible subsets of regressors. The methods are illustrated using some standard data sets.

• E. I. GEORGE AND R. E. MCCULLOCH (1997) **Approaches for Bayesian Variable Selection**, *Statistica Sinica*, **7**, 339–373.

This paper describes and compares various hierarchical mixture prior formulations of variable selection uncertainty in normal linear regression models. These include the nonconjugate SSVS formulation of George and McCulloch (1993), as well as conjugate formulations which allow for analytical simplification. Hyperparameter settings which base selection on practical significance, and the implications of using mixtures with point priors are discussed. Computational methods for posterior evaluation and exploration are considered. Rapid updating methods are seen to provide feasible methods for exhaustive evaluation using Gray Code sequencing in moderately sized problems, and fast Markov Chain Monte Carlo exploration in large problems. Estimation of normalization constants is seen to provide improved posterior estimates of individual model probabilities and the total visited probability. Various procedures are illustrated on simulated sample problems and on a real problem concerning the construction of financial in dex tracking portfolios.

• H. CHIPMAN, M. HAMADA AND C. F. J. WU (1997) **A Bayesian Variable Selection Approach for Analyzing Designed Experiments with Complex Aliasing**, *Technometrics*, **39**, 372–381.

Experiments using designs with complex aliasing patterns are often performed-for example, two-level nongeometric Plackett-Burman designs, multilevel and mixed-level fractional factorial designs, two-level fractional factorial designs with hard-to-control factors, and supersaturated designs. Hamada and Wu (1992) [*Journal of Quality Technology*, **24**, 130–137.] proposed an iterative guided stepwise regression strategy for analyzing the data from such designs that allows entertainment of interactions. Their strategy provides a restricted search in a rather large model space, however. This article provides an efficient methodology based on a Bayesian variable selection algorithm for searching the model space more thoroughly. The authors show how the use of hierarchical priors provides a flexible and powerful way to focus the search on a reasonable class of models. The proposed methodology is demonstrated with four examples, three of which come from actual industrial experiments.

• S. D. BEATTIE, D. K. H. FONG AND D. K. J. LIN (2002) **A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs**, *Technometrics*, **44**, 55–63.

In early stages of experimentation, one often has many candidate factors of which only few have significant influence on the response. Supersaturated designs can offer important advantages. However, standard regression techniques of fitting a prediction line using all candidate variables fail to analyze data from such designs. Stepwise regression may be used but has drawbacks as reported in the literature. A two-stage Bayesian model selection strategy, able to keep all possible models under consideration while providing a level of robustness akin to Bayesian analyses incorporating noninformative priors, is proposed. The strategy is demonstrated on a well-known dataset and compared to competing methods via simulation.

# Bayesians in South Africa

by Paul J. Mostert
pjmos@akad.sun.ac.za

Escaping winterly climates all around the world many Bayesians visited the ISBA 1996 World Meeting in Cape Town. Many from South Africa made presentations and/or chaired sessions at this ISBA meeting and impressed participants with their deep knowledge of modern Bayesian principles and techniques. At this World Meeting it was decided to have an international ISBA conference every two years, starting in 2000. The Southern Africa Chapter (ISBASA) was founded in 1998 at the annual South African Statistical Association (SASA) conference. The members of the ISBASA Executive Committee (EC) are:

*Chairperson:* P.J. Mostert (University of Stellenbosch)
*Secretary:* A. de Waal (CSIR)
*Members:* AL Pretorius (University of the Free State), J.W. Kruger (University of the Witwatersrand), M.V. Muddapur (University of Botswana), C.T. Tharakkan (University of Botswana), E. Keogh (University of Zimbabwe) and D.J. de Waal (University of the Free State)

The EC meets annually during the SASA Conference in November. In recent years, we have established two ISBA Southern Africa sessions, where we are exclusively promoting Bayesian activities and where the speakers are presenting Bayesian related papers. These sessions are organised by the EC of the Chapter and are also recognised as official Bayesian streams at the Conference. In April 2002 we hosted ISBASA 2002, the first Workshop of the Chapter and we foresee that this will be a bi-annual event. These workshops are intended to promote Bayesian activities in Southern Africa, and we invite specialists in Bayesian Statistical Analysis to participate in these workshops. We then invite *frequentists* all over Southern Africa to attend these workshops. ISBASA 2002 has attracted a number of frequentists and a significant number of the post-graduate students at the different Southern African Universities. The EC had invited Jim Berger, Peter Müller, Daan de Waal and Sanjib Basu to present sessions in the Workshop. Anders Madsen (Hugin, Denmark) agreed to host a session in Bayesian Networks. Their participation made the Workshop truly a world event. Their practical experience of data analysis is evident in the tens of papers and books they wrote. The major sponsors for this Workshop were the South African National Research Foundation (NRF), the Council for Science and Industrial Research (CSIR) and the University of Stellenbosch. For more details about the workshop, visit `www.sun.ac.za/isbasa/` or contact Paul Mostert at pjmos@sun.ac.za.

The number of Bayesian Statisticians in South Africa and neighbouring countries are still limited. A number of the Universities in South Africa present Bayesian Statistical Analysis at the post-graduate level. Unfortunately, we are still far away from the ideal, where people will think Bayes in solving practical and industrial problems. Our biggest group of Bayesians is at the University of the Free State (UFS). It all began there with Morris de Groot's book *Optimal Statistical Decisions*, which was used in the first post-graduate course in Bayesian Statistical Analysis at the University of the Free State. A number of other Universities present post-graduate courses in Bayesian Statistical Analysis, like the University of Stellenbosch (US), UNISA, University of Cape Town (UCT), Potchefstroom University (PU) and the Witwatersrand University (Wits). The first Bayesian symposium was held in Bloemfontein in 1986, where Jim Berger was the guest speaker.

A number of people in South Africa are actively involved in Bayesian statistical research. Daan de Waal (UFS) specialises in "Extreme value theory", Piet Groenewald (UFS) in "Changepoint analysis", Abri van der Merwe and Bertus Pretorius (UFS) in "Mixed Linear model", Isabelle Garish (UFS) in "Decision Theory" and Piet Steyn (UFS) in "Discrete distributions". At UCT Theo Stewart is busy with "Decision systems" and Paul Fatti from Wits has done some work using the Bayesian approach. Kotie Roux, Andriette Bekker (UNISA) and Paul Mostert (US) are involved in "Bayesian Reliability analysis" and "Bayesian survival models". Jan du Plessis (PU) works on "Bayesian extreme type models".

In recent years some Bayesians have established ties for Business in South Africa. One such relationship is between the Department of Mathematical Statistics at the University of the Free State and ESKOM (National Electricity supplier). Other Statistical Departments in South Africa have also strong ties with Business and Industry, but few of them involve Bayesian analysis in solving problems.

# Bayesian Instruction in Warsaw

by Tomasz Szapiro
and Marek Męczarski
[tszapiro-mecz]@sgh.waw.pl

Our experiences concern Warsaw School of Economics (WSE) but we know enough about University of Warsaw and Warsaw University of Technology.

Our WSE is a university level school of economics and business with 1250 students at each year. Students follow five years program for the Master of Economics degree. There are six profiles offered at WSE. For the purpose of this text we split them into two groups labeled "regular" and "advanced". It is understood that students interested in the regular profiles (ca 70%) follow obligatory curriculum in mathematics, econometrics and general statistics, while advanced profiles expands significantly this basic formal content offering numerous courses on mathematically based methods applicable to economics and management. Students are allowed to take any curriculum they wish — they are also allowed to switch curricula under some rules.

Students are obliged to pass exams after one-year course on *Mathematics* (a lecture provided with classes). During the course students are introduced to the concept of conditional probability and basic formulae including the Bayes Theorem. Only very simple short proofs are presented. It is assumed that a student may need to read an argument employing Bayesian inference and hence the task of an instructor is to help him to understand calculations and interpretations presented in typical economic texts.

What concerns advanced profiles, courses including Bayesian inference (and taught by the authors) are: *Calculus*, *Probability Theory*, *Mathematical Statistics* and *Statistical Decision Analysis*. The Bayes Theorem is presented repeatedly in all of these courses but it is interpreted from different points of view. In the course of *Calculus* the conditional probability is introduced as an example of families of probability measures parametrized by a set $B$ in the definition $P(A|B)$. In the course of *Probability Theory* more detailed classical mathematical presentation is given. For both lectures Polish textbooks are required, including M. Fisz's well known book as the auxiliary one. The course of *Mathematical Statistics* gives a standard presentation of the topic, including two or three lectures on Bayesian inference. S. D. Silvey's *Statistical Inference* is one of two main textbooks (another one is a Polish book by R. Zieliński). In the course of *Statistical Decision Analysis* a different view is presented which focuses on the interpretation of the set B in the definition $P(A|B)$ as a constraint. The Bayes Theorem states the rule on constraints fulfillment. A lecture notes manuscript is in use and general textbooks on calculus, probability and statistics are recommended.

Concerning teaching methods, at general (regular) level only numerical examples are presented and trained. This solution is forced by short time given for this topic. Moreover, lack of economic background impedes introduction of case studies or illustrations. The advanced level courses on *Calculus*, *Probability Theory* and *Mathematical Statistics* are supported with classes where mini-cases are presented to students. Students are obliged to do homework, which requires creation of mini-cases analogous to those presented in class. During the course on *Statistical Decision Analysis* students are obliged to present research articles and chapters from monographs. Students work in teams and they are encouraged to participate in class discussions.

A strictly "applied" area with typical basic role of Bayesian approach is insurance risk theory, which constitutes a content of another lecture.

We often experience organizational problems with computer supported problem analysis and real life cases to present in class.

At the Institute of Applied Mathematics of University of Warsaw and the Department of Mathematics and Information Sciences of Warsaw University of Technology theoretical foundations are not a concern. Bayesian statistics is a standard part of the courses on mathematical statistics. Sometimes courses on particular problems may happen, e. g. robust methods (including Bayesian robustness) or Markov Chain Monte Carlo. They are usually based on their authors' concepts and M. Sc. theses on these subjects are often prepared by their students, who are supervised to study more advanced and detailed issues.

# TEACHING BAYESIAN STATISTICS AT UCSC

by Raquel Prado
and Bruno Sansó
[raquel-bruno]@ams.ucsc.edu

The new Department of Applied Mathematics and Statistics (AMS) belongs to the Baskin School of Engineering at the University of California, Santa Cruz. After being here for a few months, it is easy to understand why. In fact, a modern school of engineering is very naturally Bayesian. These days when engineers deal with uncertainty they very willingly consider Bayesian methods. Here is an example. The vector $U$ represents information to be transmitted reliably over an unreliable channel. $U$ is encoded, that is, mapped into a codeword $X$ of the form $X = (U, X_1)$. This codeword is transmitted over a noisy channel and received as $Y = (Y_s, Y_1)$, with transmission probabilities $p(y|x) = Pr(Y = y|X = x)$, where $Y_s$ corresponds to $U$ and $Y_1$ to $X_1$. Usually $p(y|x)$ factors as

$$p(y|x) = \left( \prod_{i=1}^{k} p(y_{s_i}|u_i) \right) p(y_1|x_1) \, .$$

The *Decoding Problem* for electrical engineers is that of inferring the values of the hidden variables $U$ given the observed values $Y$. This is done by maximising the *Belief*,

$$BEL_i(a) = Pr(U_i = a|Y_s = y_s, Y_1 = y_1) \, ,$$

which for us is the posterior distribution of $U_i$ given the data. This is a very fundamental problem and it is very Bayesian in nature.

Many more examples can be found: machine learning, image tracking, applications of hidden Markov models to genomics, just to mention a few areas. So it is by no means unexpected that you attend one of the seminars of the Computer Engineering Seminar Series at UCSC and the talk abounds in details of, for example, the last generation of algorithms for sequential updating of posterior distributions in dynamic models.

Given that incorporating randomness into models is done more and more in engineering and the sciences, it is not surprising that Bayesian statistics has become so popular. In such an environment, it was all too natural to place a new department of statistics within the School of Engineering, and build it with a very strong Bayesian flavour. One of the challenges we face now is how to teach courses in Bayesian statistics in a way that is appealing and useful to engineering students. We think that providing the students with methodological courses where most of the ideas are inspired and developed through real-world case studies, i.e. finding a good compromise between theory and applications, is the way to go.

Our teaching duties include teaching statistics to undergraduate and graduate students with different backgrounds. A big proportion of the students we teach are majoring in computer sciences and computer engineering, and many of them have strong interests in machine learning or bioinformatics. However, we also teach many students majoring or pursuing graduate degrees in areas such as biology, environmental sciences and astronomy. They are eager to discover the use of Bayesian statistics in topics that range from oceanography, and marine sciences, to developing good quality local wine.

In our short teaching experience in Santa Cruz, we have been faced again with a problem that has been discussed several times: how to teach Bayesian statistics at an introductory level. Unfortunately, no matter what type of field they are meant for, there are still very few introductory Bayesian textbooks out there. In that respect we were pleased to see the new edition of an old favorite: DeGroot's book on probability and statistics, which Mark Schervish has taken care of (Addison Wesley).

The availability of textbooks does not get much better for more advanced statistical topics that can be of interest to engineers and scientists. A look at the list of Bayesian books compiled on the SBSS web page (`www.amstat.org/sections/SBSS/books/index.html`) reveals that Bayesians have put a lot of emphasis in writing books in biostatistics and econometrics, but have overlooked other areas, even if closely related. An example: even though areas like survival analysis or time series are very close to reliability and signal processing, respectively, they very seldom get treated in the way that engineers are familiar with. Hoping to be contradicted, we found only a couple of counterexamples: the book on Numerical Bayesian Methods Applied to Signal Processing by J. O Ruanaidh and W. Fitzgerald, and the book on Digital Audio Restoration by Simon J. Godsill and Peter J.W. Rayner, both by Springer.

Our department has offered a couple of Bayesian courses at the graduate level this year.

The first course, ENG-206, provides an introduction to Bayesian methods for inference and prediction to students who do not necessarily have a strong statistical background. Topics such as exchangeability, coherence and calibration, conjugate analysis and basic MCMC methods for simulation-based computations were discussed in the course. Although this first course was conceived as a methodological course, it was guided by real-world case studies. The course, offered in the Winter quarter, was really popular among graduate students from Computer Sciences, Computer Engineering and Bioinformatics. After this first exposure, many students thought Bayesian statistics was so useful for their research that they decided to take a second course, ENG-207, which is being offered this Spring. It was a pleasant surprise to see that about 20 graduate students from disciplines other than statistics are attending this new course and are really excited about the topic. ENG-207 is a statistical modelling course. The students will be exposed to hierarchical models, regression models and GLMs from a Bayesian perspective. The last portion of the course will be devoted to models that are particularly interesting for many of the graduate students here, such as hidden Markov models.

Our degree programs are not yet fully in place, but we hope to admit our first cohorts of Master and Ph.D. students in the fall of 2002, when a joint Graduate Program in applied mathematics and statistics at UCSC is expected to be functional. This presents a new challenge, for, taking advantage of the strengths of the department, we expect to give the students a modern scientific approach that combines mathematical modelling and statistical tools. This idea is at the heart of new developments in statistics where the Bayesian community is having a leading role.

---

## A LAST COMMENT

by M. Eugenia Castellanos and Javier Morales
me.castella@umh.es
j.morales@umh.es

Along the last year we have published articles about recent Ph.D's, some advices about young statisticians that were working and several possible opportunities in the job market. Our goal was to show different topics that could be interesting for the students and let them know about those. We hope, at least partially, to have achieved this goal. Of course there is a lot of work to do, and we think the *Student's Corner* is appropriate for this purpose. We hope the new associate editors will work along this direction and we are sure that they will enjoy this work because it is a fruitful experience that allows to contact many people. We thank Fabrizio Ruggeri for his help and his useful suggestions. We also thank all people who have helped us with their articles and abstracts. In this last issue we include the abstract of the thesis of Dr. Antonella Bodini.

**Antonella Bodini**
CNR - IMATI
anto@iami.mi.cnr.it
Advisor: Eugenio Regazzini

The extraordinary development of the modern statistical methods has gone with the critical examination of the deterministic conception of science and, in particular, of the meaning of the term "scientific law". In the paper *Why Isn't Everyone a Bayesian* (Am. Stat., 40, pp. 1-11, 1986) Efron reminds us that "the current era is the first century in which statistics has been widely used for scientific reporting…". As far as the role of statistics is concerned, it can be condensed into the proposal of logically consistent methods, both for the design of informative experiments and for the elaboration of observed results in order to making inference about future observations. There are various ways that have been devised to accommodate statistical predictions that are not Bayesian: classical confidence approach, various methods that depend on the likelihood, etc. In any case, only the Bayesian formulation teaches how to take account of "each new bit of knowledge: the initial opinion (or, technically, distribution) is replaced, step by step with each new datum, to give, when the process is arrested, the final opinion (or distribution)." (de Finetti, 1959, *Probability, Induction and Statistics. The art of guessing*, Wiley, London). On the other hand the Bayes theorem is applied, very often, through initial distributions for unknowin parameters of probabilistic models, without caution or consideration on the fact that parameters, such as their "generating" models, are mere abstract constructions.

In view of the above remarks, in this thesis we have described and studied the essential characteristics of an approach to statistical inference that: (a) preserves the peculiarities of the Bayesian formulation with respect to the dynamics of the process of learning from experience; (b) does not require the assessment of any distribution for unobservables parameters; (c) can incorporate any assigned model. The domain of application of the approach that we get briefly to present is the field of all observables phenomena when: (i) there is at least one probabilistic model, based on hypotheses regarding the development of the phenomenon being studied, which allows to make predictions on some aspect of the phenomenon itself; (ii) this model is not completely known, in general, and its specification is based on the results of observations on the phenomenon under study; (iii) models are only viewed as instruments for prediction, so that unknown parameters are thought of as devoid of any objective meaning, in general.

We suppose that any observation takes its values in $(X,\mathcal{X})$ where $X$ is a Polish space and $\mathcal{X}$ is its Borel $\sigma$-algebra. The probabilistic model above-mentioned takes the form of a probability measure $\alpha = \alpha(\bullet;\theta)$ defined on $(X,\mathcal{X})$, $\theta$ being an unknown parameter which belongs to a well-defined parameter space $\Theta$.

As far as the real process of observation is concerned, "set" observations are generally admitted but their probability law is defined as a probability distribution (abb. *p.d.*) $\rho$ on $(X^N,\mathcal{X}^N)$, where $N \leq \infty$. The assessment of $\rho$ is of fundamental importance for the implementation of the inferential process that we are discussing. In fact, it is based on the assumption that inferential processes stem from the combination of the theoretical model (fixed *a priori*) with the way in which - in accordance to personal choice - observed data are considered with respect to the prevision of future facts. In order to reflect the second element in that combination we introduce an initial distribution $\rho_1$ on $(X,\mathcal{X})$ and a sequence $\{\rho_{n+1}\}_{n\geq 1}$ of probability kernels from $X^{n-1}$ to $\mathcal{X}$. If $\tilde{x}_k$ is defined to be the $k-$th projection of $X^N$ into $X$, according to the Ionescu-Tulcea theorem, there exists a unique probability measure $\rho$ on $(X^N,\mathcal{X}^N)$ such that $\rho_1$ is the p.d. of $\tilde{x}_1$ and $\rho_{n+1}$ is the conditional distribution of $\tilde{x}_{n+1}$ given $(\tilde{x}_1,\ldots,\tilde{x}_n)$. The first element, i.e. the theoretical model, is taken into account by restricting the choice of the $\rho_n$ to those predictive p.d.'s for which $\alpha(\bullet;\theta)$ is the p.d. of each $\tilde{x}_k$. The value of a model, as a tool for prevision, springs from the logical connection between its mathematical assumptions and the peculiarities of the phenomenon which is object of prevision. Moreover, models are, generally speaking, mathematically tractable entities. In view of these facts, they are generally preferred to predictive distributions, for scientific report, even in presence of statistical data. As a matter of fact, any predictive distribution incorporates the influence of observations on prevision, fixed subjectively by the one which defines $\{\rho_{n+1}\}_{n\geq 1}$, and this may bear some trouble for the intelligibility and the tractability of each $\rho_{n+1}$. In view of all these remarks one can try to employ data in order to decide if it is reasonable using the model as substitute for a predictive distribution. In other words, it is natural to consider some interesting consequence, expressed by a loss function, of the effect of such substitution and, then, to find a value of $\theta$ that optimizes such consequence. The resulting value of $\theta$ can be used to complete the assessment of each $\rho_n$, $n = 1, 2, \ldots$, until additional data are gathered or some new loss function draws the statistician's attention.

In this way, one avoids introducing prior distributions for unobservables parameters, which are thought as control variables. As by-product of this procedure, one obtains a direct way to compare statistical models in relation to some specific data set and with respect to some fixed loss function. In order to illustrate the previous remarks, in this work we have pointed out the general lines of solutions of some specific problems connected with a couple of topics that occupy a central position: point estimation and regression analysis. We have considered, in particular, the case in which the predictive distribution $\rho_{n+1}$ is defined as

$$\rho_{n+1}(\tilde{x}_1,\ldots,\tilde{x}_n;\bullet) = \frac{a\alpha_1(\bullet,\theta) + \sum_{k=1}^{n}\delta_{x_k}(\bullet)}{a+n} \qquad (n \geq 1)$$

where $a$ is a positive constant, so the observation sequence $\{\tilde{x}_k\}_{1\leq k\leq N}$ is exchangeable.

# NEWS FROM THE WORLD

by Antonio Lijoi
lijoi@unipv.it

✻ *denotes an ISBA activity*

## ➤ Events

**Gnedenko Anniversary Meeting** *June 3-7, 2002, Kyiv, Ukraine*
The conference will focus on topics close to the wide range of scientific interests of B. V. Gnedenko. The aim is twofold: first, to highlight the contributions of B. V. Gnedenko in probability theory and its applications, history of mathematics, problems of educations, and, second, to present the developments of his ideas as well as the current trends in the theory of probability and related fields. Web page: `ln.com.ua/~tbimc/gnedenko/`

**EURANDOM Workshop on Discrete Probability** *June 17-20, 2002, Eindhoven, The Netherlands*
The workshop will focus on discrete probability and related areas, including, e.g., combinatorics, randomized algorithms, and scaling limits of discrete random structures. Web page: `www.eurandom.tue.nl/workshops/discrete_probability_2002.htm`

**The New Frontiers of Statistical Data Mining, Knowledge Discovery and E-Business** *June 22-25, 2002, Knoxville, Tennessee, U.S.A.*
The primary focus of this conference will be to bring national and international experts and practitioners together to share and disseminate new research and developments covering the wide spectrum of areas such as: data storage, warehousing, access and processing to visual and algorithmic analysis and reporting, to the role of information theoretic and Bayesian statistical modeling techniques in Data Mining, Knowledge Discovery, and E-Business with emphasis to real-world applications. There will also be a Special Workshop on Data Mining to be delivered by a prominent and experienced Data Mining Group. Deadline for registration is May 15, 2002. Conference web page at `stat.bus.utk.edu/conference/`

**ASA Conference on Radiation and Health** *June 23-26, 2002, Deerfield Beach, Florida, U.S.A.*
The Conference offers a unique forum for discussing the qualitative aspects of radiation health research in a multi-disciplinary setting. The Conference also furnishes investigators in health related disciplines the opportunity to learn about new quantitative approaches to their problems and furnishes statisticians the opportunity to learn about new applications for their discipline. The focus of the 2002 Conference on Radiation and Health is "Current Issues in Radiation Health". Conference web page at `www.amstat.org/meetings/radiation/general.html`.

**28th Conference on Stochastic Processes and their Applications** *July 1-5, 2002, Melbourne, Australia*
The 28th Conference on Stochastic Processes and their Applications is organized under the auspices of the Bernoulli Society for Mathematical Statistics and Probability. The program consists of fifty-minute lectures delivered by invited speakers and twenty-minute contributed talks by participants on various topics related to stochastic processes and their applications. The topics include, but are not limited to: stochastic analysis, discrete random processes and randomised algorithms, topics in limit theorems, Markov chain Monte Carlo, Markov processes, random processes in random environments, point processes, as well as application areas such as: stochastic processes in finance and insurance, stochastic processes in physics, applications to telecommunications, time series, modelling in biology and medicine. Abstract for contributed talks need to be sent by May, 3, 2002. URL of the conference `www.spa28.ms.unimelb.edu.au/`

**Summer School on Inference for Stochastic Processes** *July 1-20, 2002, Torgnon, Italy*
University "L. Bocconi", Milan, organizes a summer school on Inference for Stochastic Processes. Lectures will be held by Profs. B.L.S. Prakasa Rao and Anand Vidyashankar. The topics of the course will range from Markov and Branching Processes to Tree Structured Processes and Diffusions. Main applications will involve clinical trials, cell biology, internet traffic and finance. Application form for registration can be downloaded from the web site of the school and should be sent before May, 21, 2002. The program of the course and the application form can be found at `www.uni-bocconi.it/imqcorsi/`

**XXXII International Probability School** *July 7-24, 2002, Saint-Flour (Cantal), France*

Founded in 1971, this school is organized every year by the Laboratory of Applied Mathematics (Clermont-Ferrand). It is supported by the Blaise Pascal University and the C.N.R.S. It is intended for PhD students, teachers and researchers who are interested in probability theory, statistics, and in applications of these techniques. The school has three main goals: (1) to provide, in three high level 15-hour courses, a comprehensive study of a field in probability theory or statistics; (2) to allow the participants to present their work in lectures; (3) to promote exchanges between the participants. Lectures will be held by Profs. Jim Pitman, Boris Tsirelson and Wendelin Werner. Deadline for registration is March 31, 2002. More detailed information at the web site `wwwlma.univ-bpclermont.fr/stflour/stflour-en.html`

**From Lévy Processes to Semimartingales - Recent Theoretical Developments and Applications to Finance** *August 20-27, 2002, Aarhus, Denmark*
It is a summer school jointly organized by MaPhySto, DynStoch and CAF and dealing with recent theoretical achievements in the theory of Lévy processes and of semimartingales, having applications to finance in view. There will be three courses: two of them, namely "Introduction to Semimartingales and their Basic Algebra" and "Change of Time and Measure" will be held by Prof. A.N. Shiryaev, whereas talks by Profs. N. Shephard, O.E. Barndorff-Nielsen and F. Hublaek will focus on "Stochastic Volatility Models and some of their Applications". Deadline for registration is July 1, 2002. Additional information is available from the web site `www.maphysto.dk/events/LPS2002/`

**Fourth International Conference on Statistical Data Analysis Based on the L1-Norm and Related Methods**
*August 4-9, 2002, Neuchatel, Switzerland*
The purpose of the conference is to bring new developments, in computational and inferential aspects of statistical data analysis based on the L1-norm, together in order both to simplify and enrich the theory, and discuss possibilities for new applications. Besides a few invited lectures, some contributed talks are planned as well. Those wishing to present papers at the conference are asked to submit to the conference organizer an abstract no later than May 31, 2002. Web page of the conference `www.unine.ch/statistics/`

**ISC6 - Sixth International Statistics Conference** *August 26-28, 2002, Tehran, Iran*
The conference, sponsored by the Iranian Statistical Society, will be held at Tarbiat Modarres University. Web page of the conference `www.modares.ac.ir/isc6/`.

**Reason Park - Second International Summer School on "Reasoning under Partial Knowledge"** *August 26 - September 14, 2002, Foligno, Italy*
The scope of the School (sponsored by Regione Umbria : lectures, lodging, breakfast, lunch, and a stipend - about 100 EUR - are granted to the participants) is to provide Ph.D. students and, in general, young researchers with a basic training in some different topics which play an important role in "Reasoning under Partial Knowledge" and their application in various fields, including Computer Science, Economics, Engineering, Medicine, Biology.
Regarding the level of the courses, the first two or three lectures of each course will provide a tutorial and simple introduction to the field, while the remaining part should provide a complete and updated information. In this way, all the courses should be easily accessible also to an audience that has not been previously acquainted with the subject. Web page of the school `www.dipmat.unipg.it/reasonpark`.

**11th International Workshop on Matrices and Statistics**
*August 29-31, 2002, Lyngby, Denmark*
The workshop will provide a forum through which statisticians working in the field of linear algebra and matrix theory may be better informed of the latest developments and newest techniques, and may exchange ideas with researchers from a wide variety of countries. Presentation of both Invited and Contributed papers on matrices and statistics are planned. It is expected that many of these papers will be published, after refereeing, in a Special Issue of Linear Algebra and its Applications - with the theme Linear Algebra and Statistics. Deadline for submission of abstracts of Contributed papers is June, 3, 2002. Interested people may visit the web site `www.imm.dtu.dku/matrix02/`

**Physics - Signal - Physics: On the Links between Nonlinear Physics and Information Sciences** *September 8-13, 2002, Les Houches, France*
Nonlinear Physics is a rapidly

growing field. Recent theoretical and phenomenological approaches allowed great progress in many domains as different as turbulence, disordered systems, critical phenomena, ... Approaches developed by physicists have led theoreticians in signal and image processing to define new models, and to develop new processing techniques applicable in many fields of information sciences. In parallel and independently, signal and image processing have experienced numerous theoretical and algorithmic developments in the field of nonstationary and/or non Gaussian signal, scaling processes, nonlinear systems, ... These new techniques have in return allowed significant progress in physics.
The aim of the session "Physics - Signal - Physics" is to review some of these two way interactions between physics and signal, and from a more general point of view, between physics and information sciences. More information can be found at www.lis.inpg.fr/houches.htm.

**Course on Statistical Methods for Evaluating the Effectiveness of Advertising** *Rimini 25-27, 2002, Rimini, Italy*
ENBIS (European Network for Business and Industrial Statistics) is organising the course, right after its Second Annual Conference. The course is hosted by the University of Bologna at Rimini. More information on the ENBIS web page www.ibisuva.nl/ENBIS/.

**Workshop on Genomic Signal Processing and Statistics (GENSIPS)** *October 12-13, 2002, Raleigh, North Carolina, U.S.A.*
The Workshop on Genomic Signal Processing and Statistics is a cooperating conference of the IEEE Signal Processing Society and will be sponsored by DARPA, NSF and the Kenan Institute. The aim of this two-day workshop is to identify potential areas of collaboration between the biological, statistical, and signal processing communities and to open new avenues of research to address new challenges in genetics by exploiting potential synergies between signal processing, statistics and Genomics and by building on their respective strengths. This workshop will consist of both invited sessions and contributed sessions. Those interested in contributing should submit a four-page summary by May, 1, 2002. More information are available at www.gensips.gatech.edu

**Euroworkshop on Statistical Modelling – Model Building and Evaluation** *October 31 - November 3, 2002, Bernried, Germany*
The Euroworkshop on Statistical Modelling is a project which is funded by the European Commission (CORDIS) in the programme High Level Scientific Conferences. The present workshop is designed to have lectures from Keynote Speakers and further contributed presentations. The number of participants is limited to 30, and young researchers are particularly encouraged to participate actively by presenting their work. Among the main areas of interest there are: "Graphical Model Diagnostics", "Model validation using smoothing techniques" and "Bayesian Models and their validation". Prospective participants should send an e-mail to Göran Kauermann: goeran@stats.gla.ac.uk not later than July 17, 2002. For recent updates, refer to the web page www.stat.uni-muenchen.de/euroworkshop/2002.html

**KONBiN 2003 - Third Safety and Reliability International Conference** *May 26-30, 2003, Gdynia, Poland*
The conference is focused on the problems of the creation and assurance of safety and reliability in the human-technology-environment systems. The conference is addressed to university and research institutes scientists, industry and transport employees, government and municipal bodies, reliability and safety consultants and other persons interested in the conference topics. Venue of the Conference will be "Gdynia Orbis Hotel", situated in the central district of the city. Web page of the conference www.wsm.gdynia.pl/konbin/.

**International Statistical Institute, 54th Biennial Session** *August 13-20, 2003, Berlin, Germany*
Information on the conference can be found at the web page www.isi-2003.de/index.htm.

## ➤ Internet Resources

**HYDRA**
HYDRA is an open-source, platform-neutral library for performing Markov Chain Monte Carlo. It implements the logic of standard MCMC

samplers within a framework designed to be easy to use and to extend while allowing integration with other software tools. In addition, it provides classes implementing several unique adaptive and multiple chain/parallel MCMC methods. Web site for full description of the software and download is `software.biostat.washington.edu/statsoft/MCMC/Hydra/`

**Bayesian Belief Network Software**
The package consists of three systems: Belief Network (BN) PowerConstructor, BN PowerPredictor and Data PreProcessor. The first one is a system that learns Bayesian belief network structures and parameters from data. The second is a data mining system *for data modeling/classification/*prediction. It extends BN PowerConstructor to BN based classifier learning. The last system is a tool used with BN PowerConstructor and BN PowerPredictor for pre-processing the training data. Web page for download: `www.cs.ualberta.ca/~jcheng/bnsoft.htm`

**BASSIST v0.8**
Bassist is a tool that automates the use of hierarchical Bayesian models in complex analysis tasks. Bassist Version 0.8 supports most of the "standard" features of hierarchical Bayesian models. Instruction for installation and files for download are available at `www.rni.helsinki.fi/cs/bassist/v0.8/index.html`

**Statistical Pattern Recognition Toolbox**
The toolbox contains pattern recognition methods and was written by Vojtech Franc. It focuses on linear discriminant function including its generalization to quadratic discriminant function by non-linear data mapping, unsupervised learning algorithms and minimax learning algorithms. The toolbox is build on top of the Matlab, version 5.2. Web page for download `cmp.felk.cvut.cz/~xfrancv/stprtool/html/mainpage.html`

**Computer-Intensive Statistical Methods**
C-ISM2 is a freeware package for testing one- and two-sample hypotheses by means of permutation (randomization) and bootstrap techniques. All computer-intensive modules are based on ideas gathered by in a book by B.F.J. Manly: "Randomization, Bootstrap and Monte Carlo Methods in Biology". The software can be downloaded from the web site `pjadw.tripod.com/download.htm`

## ➤ Awards and Prizes.

**Gertrude Cox Scholarship**
The Gertrude Cox Scholarship is sponsored by the ASA Committee on Women in Statistics and the Caucus for Women in Statistics and is awarded each year to encourage women to enter statistically oriented professions.
Application is limited to women who are citizens or permanent residents of the United States or Canada and its deadline is April 30, 2002. The award will be presented at the Joint Statistical Meetings in August.
Application form and further information at `www.amstat.org/awards/cox-scholarship.html`

**Edward C. Bryant Scholarship**
Each year an outstanding graduate student in Survey Statistics is awarded the Edward C. Bryant Scholarship to help support the student's graduate education. Selection of the scholarship recipient is made by the ASA Bryant Scholarship Award Committee. Applications and letters of recommendation must be received by April 30, 2002, for consideration. The award will be presented at the Joint Statistical Meetings in August. Application form and further information at `www.amstat.org/awards/bryant.html`

## ➤ Miscellanea

**Day for Seymour Geisser**
On July 1, 2001, Seymour Geisser resigned as Director of the School of Statistics at University of Minnesota after 30 years. He will be honored for these years of service on Saturday, May 11, 2002.
In the opening session Jim Press will give an overview of Geisser's contributions to the profession. The title of his talk is "Seymour Geisser As Statistician, Part I–The First 50 Years"
There will be three more sessions during the day where six of his former PhD students, Jim Hodges, Wes Johnson, Michael Lavine, Jack Lee, Rob McCulloch and Ming-Dauh Wang will give technical talks.
Since a book of recollections about Geisser is under preparation, students of the School, colleagues and friends are encouraged to send Jane Sell (`jane@stat.umn.edu`) anecdotes and pictures that can be included in the book.
More details can be found at `www.stat.umn.edu`.

## Some historical notes on the Laws of Large Numbers before 1950

The "classical" law of large numbers states the convergence of the sequence of the frequencies of success to the probability of success in a Bernoulli scheme. In general, the "large numbers problem" can be stated in the following way:

*Given a sequence of random numbers $X_1, X_2, \ldots, X_n, \ldots$, with finite expected value, which further conditions are needed in order that*

$$T_n := \sum_{k=1}^{n} \{X_k - E(X_k)\}/n$$

*may converge to 0, as $n \to \infty$, in a way to be properly specified?*

Convergence in probability leads to the *weak law of large numbers*, whereas the *strong law of large numbers* refers to almost sure convergence. It is well known that almost sure convergence implies the one in probability, whereas the vice versa does not generally hold. The intuitive phenomenon of the frequencies concentrating, on the long run, around the value of the probability finds a mathematical description in the almost sure convergence and not in the one in probability.

From an historical point of view, the first law of large numbers refers to the convergence in probability of the sequence of the frequencies of success - in a sequence of independent and identically distributed trials - towards the probability of success in each trial; it is the classical Bernoulli's theorem. The first extension of the previous theorem to random numbers with uniformly bounded variance goes back to Pafnuty L. Chebysheff (1821-1894) in 1867. The proof of the resulting weak law relies on an inequality, nowadays named after both Bienaymé and Chebysheff [it seems that Irenee Jules Bienaymé (1796-1878) was the first to state it around 1850]. In 1899 Andrei A. Markov (1856-1922) slightly improved Chebysheff's law. In any case, these results do not solve the pressing problem of the sequence $(T_n)_{n\geq 1}$ eventually concentrating around $0$. At the beginning of the XX-th century, some people, mainly among pragmatic ones, thought the previous laws very sufficient to justify the phenomenon, felt very intuitive by many. The shortcoming of that belief was realised by Emile Borel (1871-1956) during his research generated by a result in Number Theory which states that, with respect to Lebesgue measure, the number of zeroes in the binary representation of almost every number $x$ in $[0, 1)$ is $1/2$. Although his arguments were not completely right, Borel proved, in 1909, the first strong law of large numbers about the frequencies in a Bernoulli scheme with probability of success $1/2$. It took some time before the distinction between convergence in probability and almost sure one was clearly made. The distinction goes back to Francesco Paolo Cantelli (1875-1966) as a result of his deep analysis, started in 1916-1917, about the meaning of the law of large numbers. Remarkably, he introduced the term *uniform law* and not *strong law*, actually due to Alexander I. Khinchin (1894-1959). Cantelli is the author of an important extension of Borel's theorem to independent [but not necessarily identically distributed] random numbers with a sequence of uniformly bounded fourth moments. The proof is based, besides some Chebysheff-like inequalities, on the famous result, proved in that occasion, which is known nowadays as the Borel-Cantelli's lemma (see Cantelli, 1917a, 1917b).

Andrei Nicolaevic Kolmogorov (1903-1987) significantly improved Cantelli's law. In 1930 he proved (Kolmogorov, 1930) that the convergence of the series $\sum_{n=1}^{\infty}(v_n/n^2)$, where $v_n$ is the variance of $X_n$ (and $m_n$, used later, its expected value), is a sufficient condition to ensure the validity of the strong law for sequences of independent random numbers, even if not identically distributed. In 1933, assuming the identity in distribution, Kolmogorov proved (see Kolmogorov, 1933) that the finite expected value of any element in the sequence is a necessary and sufficient condition for the strong law to hold. The analysis of the large numbers problem leads, naturally, to the study of the series of random numbers; the contributions by Khinchin and, mainly, Kolmogorov (between 1925 and 1930) were fundamental in this field of research, too.

Connected to the limit law of the empirical distribution function $F_n$, the *fundamental theorem of Mathematical Statistics* derives, in some sense, from the strong law for frequencies in a Bernoulli scheme; it was proved by Valerii I. Glivenko (1897-1940) in a particular case and, in the general case, by Cantelli (1933). It states that $\sup_{x\in\mathbf{R}} |F_n(x) - F(x)|$ converges to zero almost surely, when $n$ goes to infinity. It is worth mentioning that most of Cantelli's strong law of large numbers in Bernoulli trials was proved in de Finetti (1933), which precedes Cantelli's paper.

Up to now we have mentioned only researches on the laws of large numbers for sequences of independent random variables. The independence assumption is crucial but, nonetheless, there are interesting results under particular assumptions on dependence. Historically, the first (weak) law of large numbers for dependent sequences was proved by Markov in 1907, in the case of homogeneous chain, with a finite number of states and strictly positive transition probabilities. Under these assumptions, the elements in the sequence are "asymptotically independent" because of the Markov ergodic theorem. Actually, the asymptotic independence ensures the validity of the same result obtained under the more restrictive assumption of independence. These findings suggested Serghei N. Bernstein (1880-1968) to investigate if the same conclusion could be achieved under more general conditions. Bernstein (1946) states, essentially, that the weak law of large numbers holds if $\lim_{n\to\infty}\sum_{k=1}^{n} m_k/n$ exists, and if there exists $K$ such that $v_n \leq K$ for all $n$ and the correlation coefficient between $X_i$ and $X_j$, $\rho_{ij}$, is such that $|\rho_{ij}| \leq R(|i-j|)$ where $R(n)$ is a nonnegative function of $n$, such that $R(0) = 1$ and $\sum_{k=1}^{n} R(k)/n \to 0$ for $n \to \infty$. The last assumption replaces, in some sense, the condition on the asymptotic independence of Markov chains with a weaker form of "asymptotic orthogonality". In any case, it seems that the validity of the strong law of large numbers, in its classical formulation, strongly depends on the vanishing of the dependence between two elements in the sequence as their positions get farther and farther.

However, there are remarkable examples of sequences of random numbers in which the above phenomenon does not hold, but, nonetheless, the convergence of the "average" series $(\sum_{k=1}^{n} X_k/n)_{n\geq 1}$ holds, possibly to a new random quantity [not necessarily degenerate as in the classical case]. As an example, if $(X_n)_{n\geq 1}$ is exchangeable, then it can be shown, under the assumption $E|X_1| < \infty$, that $|\sum_{i=1}^{n} X_i/n - \sum_{i=1}^{m} X_i/m|$ converges to 0 as $m$ and $n$ diverge. The assumption is equivalent to the existence of a random number $X$ to which $(\sum_{i=1}^{n} X_i/n)_{n\geq 1}$ converges almost surely. Because of the importance of the exchangeable sequences in statistical inference, the above strong law of large numbers has a considerable interest; furthermore, it includes, as a particular case, the Kolmogorov's law for sequences of i.i.d. random numbers. It has been proved, under restrictive conditions on the moments of $X_1$, by Bruno de Finetti (1906-1985) in 1933 (see de Finetti, 1993b). Now we have a simple proof stemming from the martingale convergence theorem. Furthermore, de Finetti's theorem can be presented as a corollary of the Birkhoff's ergodic theorem, relative to stationary sequences.

The notion of stationarity has origin in Physics: Liouville's theorem, which provides an invariant (over time) measure for the evolution of conservative dynamic systems in their phase space. Khinchin, in 1932, transferred the physical notion into a probabilistic one: a sequence of random numbers $X_1, X_2, \ldots$ is stationary when the probability law of $n$ elements is invariant for translation with respect to the indices, and it holds for any $n = 1, 2, \ldots$ (see Khinchin, 1932). In a probabilistic scenario, Birkoff's ergodic theorem has a counterpart in the following law of large numbers: if $(X_n)_{n\geq 1}$ is stationary and $E|X_1| < \infty$, then $(\sum_{i=1}^{n} X_i/n)_{n\geq 1}$ converges, in a strong sense, to a random number coincident with $E(X_1|\mathcal{I})$, a distinguished conditional expectation of $X_1$, being $\mathcal{I}$ a $\sigma$-algebra generated by $(X_n)_{n\geq 1}$.

In the 30's and 40's Paul Lévy (1886-1971) and, specially, Joseph Leo Doob (1910- ) proved some fundamental theorems on the convergence of submartingales which allowed for new proofs of many results mentioned before (see, e.g., Hall and Heyde, 1980).

The interested reader can find detailed information on the laws of large numbers in the book by Revesz (1968), besides the classical ones by Gnedenko e Kolmogorov (1954), Ibragimov and Linnik (1971), Petrov (1975) and Hald (1998).

### ➤ References

BERNSTEIN, S.N. (1946). Sur le theoreme limite de la theorie des probabilites *Bulletin Math. Mech. Inst. Univ. Tomsk*. **3**, 174-190.

CANTELLI, F.P. (1917a). Sulla probabilità come limite della frequenza. *Atti dell'Accademia Nazionale dei Lincei*, **Serie** $5^a$, **Rendiconti 26**, 39-45.

CANTELLI, F.P. (1917b). Su due applicazioni di un teorema di G. Boole alla statistica matematica. *Atti dell'Accademia Nazionale dei Lincei*, **Serie** $5^a$, **Rendiconti 26**, 295-302.

CANTELLI, F.P. (1933). Considerazioni sulla legge uniforme dei grandi numeri e sulla generalizzazione di un fondamentale teorema del signor Lévy. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 327-350.

DE FINETTI, B. (1933a). Sull'approssimazione empirica di una legge di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 415-420.

DE FINETTI, B. (1933b). La legge dei grandi numeri nel caso dei numeri aleatori equivalenti. *Atti dell'Accademia Nazionale dei Lincei*, **Serie** $6^a$, **Rendiconti 18**, 203-207.

GNEDENKO, B.V. and KOLMOGOROV, A.N. (1954). *Limit distributions for sums of independent random variables*. Addison-Wesley. (Published in Russian in 1949).

HALD, A. (1998). *A History of Mathematical Statistics from 1750 and 1930*. Wiley.

HALL, P. and HEYDE, C.C. (1980). *Martingale limit theory and its applications*. Academic Press.

IBRAGIMOV, I.A. and LINNIK, Y.V. (1971). *Independent and stationary sequences of random variables*. Wolters-Noordhoff. (Published in Russian in 1965).

KHINCHIN, A.I. (1932). Zur Birkhoffs Losüng des Ergodenproblems. *Mathematische Annalen*, **107**, 485-488.

KOLMOGOROV, A.N. (1930). Sur la lois forte des grandes nombres. *Comptes Rendus de l'Academie des Sciences, Paris*, **191**, 910-912.

KOLMOGOROV, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 92-99.

PETROV, V. (1975). *Sums of independent random variables*. Springer.

REVESZ, P. (1968). *The laws of large numbers*. Academic Press.

*(This paper is based on some historical notes, mainly written in Italian, by Eugenio Regazzini, whom we heartily thank. F.R.)*