Lecture Notes in Statistics 205
Proceedings

Probability Approximations and Beyond

Springer

Lecture Notes in Statistics

Proceedings

Volume 205

For further volumes:
http://www.springer.com/series/8440

Andrew D. Barbour · Hock Peng Chan
David Siegmund
Editors

Probability Approximations and Beyond

Springer

Andrew D. Barbour
Institut für Mathematik
Universität Zürich
Winterthurerstr. 190
8057 Zürich, Switzerland
e-mail: A.D.Barbour@math.uzh.ch

Hock Peng Chan
Department of Statistics
and Applied Probability
National University of Singapore
Singapore 119260
Republic of Singapore
e-mail: stachp@nus.edu.sg

David Siegmund
Department of Statistics
Stanford University
Serra Mall 390, Sequoia Hall
94305 Stanford
CA, USA
e-mail: dos@stat.stanford.edu

ISSN 0930-0325
ISBN 978-1-4614-1965-5 e-ISBN 978-1-4614-1966-2
DOI 10.1007/978-1-4614-1966-2
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011941623

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Louis Chen: A Celebration

On 25 and 26 June 2010, a conference, Probability Approximations and Beyond, was held at the National University of Singapore (NUS) to honor Louis Chen on his 70th birthday. Professor Chen is the Tan Chin Tuan Centennial Professor and Professor in both the Department of Mathematics and the Department of Statistics and Applied Probability. He is also the founding Director of the Institute for Mathematical Sciences at the NUS.

Growing up as one of five brothers and a sister during WorldWar II and the immediate postwar period, Louis developed his life-long interests in mathematics and music. He graduated from the University of Singapore[1] in 1964; and after teaching briefly in Singapore, he began graduate studies in the United States. He earned a Master's and a Ph.D. in Statistics at Stanford University, where he wrote his Ph.D. thesis under the supervision of Professor Charles Stein. During his time at Stanford, Louis met his future wife, Annabelle, who was then a summer school student at Stanford.

During his Ph.D. studies, Louis made the first of several seminal contributions to the theory and application of Stein's method. This appeared in his famous 1975 paper on Poisson approximation for dependent trials, and laid the foundation for what is now known simply as the Stein–Chen method. The Poisson approximation, sometimes called the "law of small numbers," has been known for nearly two centuries, and is taught in introductory probability courses as the limiting approximation for the distribution of the number of occurrences of independent, rare events. Louis showed that independence is not a necessary prerequisite for the law to hold, and proved, by a simple and elegant argument, that the error in the approximation can be explicitly bounded (and shown to be small) in an amazingly large number of problems involving dependent events. This approximation has

[1] NUS was formed through the merger of the University of Singapore and Nanyang University in 1980.

found widespread application, in particular in the field of molecular sequence comparison.

For much of his research career, Louis has been fascinated by a circle of ideas centered on probability inequalities and the central limit theorem. Apart from his work on Poisson and compound Poisson approximation, he has written a number of papers exploring the relationships between Stein's method and Poincaré inequalities; he has established martingale inequalities that, in particular, sharpen Burkholder's inequalities; and he has returned again and again to the central limit theorem. One of his most important contributions here has been to turn Stein's concentration inequality idea into an effective tool for providing error bounds for the normal approximation in many settings, and in particular for sums of random variables exhibiting only local dependence. He has recently co-authored a book, 'Normal approximation by Stein's method', that promises to be the definitive text on the subject for years to come.

After his graduate studies, Louis spent almost a year as Visiting Assistant Professor at Simon Fraser University in Canada, before returning to Singapore in 1972. Since then, he has been engaged in teaching and research at NUS, apart from short visiting appointments in France, Sweden and the United States. Annabelle worked for many years for IBM, and together they raised two daughters, Carmela and Jacinta. In addition to research and teaching, Louis has played a leading role in the transition of NUS from a largely teaching institution to a leading research university. Louis has served as Chair of Mathematics, helped to found the Department of Statistics and Applied Probability, where he was also Chair, and since 2000 has been the director of the Institute for Mathematical Sciences (IMS). Under Louis's leadership, the IMS has developed short programs to bring international groups of mathematicians and related scientists to Singapore, to discuss recent research and to work with the local mathematical community on problems of common interest, both theoretical and applied. It has also pursued outreach programs and organized public lectures to stimulate interest in mathematics and science among Singapore students at the high school/junior college level.

Louis's professional service has not been confined to NUS. He has also served as President of the Bernoulli Society (1997–1999), of the Institute of Mathematical Statistics (2004–2005), and as Vice President of the International Statistical Institute (2009–2011). He has also served on numerous committees of these and other international organizations.

Along with this extraordinary level of administrative activity, Louis has continued a very active program of research, infecting students and colleagues alike with his enthusiasm for probability and its applications. As well as exploring new directions in probability theory, he has developed a recent interest in applications of his work on Poisson approximation to problems of signal detection in computational biology. Several of the papers in this volume provide ample evidence that these subjects continue to provide exciting theoretical developments and scientific applications.

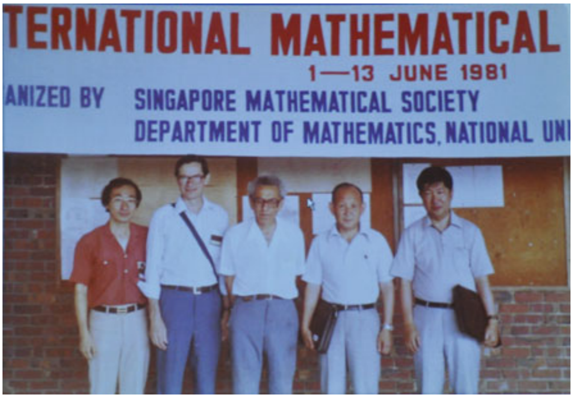
In summary, Louis Chen's professional life has combined outstanding scholarship with exemplary service, to strengthen scientific institutions in Singapore and internationally, and to provide more and better opportunities for all mathematical scientists. This volume is only a small expression of the many contributions he has made to students and colleagues. We hope to see him continuing to participate in mathematical research and enjoying music for many years to come.

Andrew D. Barbour
Hock Peng Chan
David Siegmund

Conference participants at the University Hall

A candid shot of Louis captured during the conference

Chatting with friends during the conference dinner



A younger Louis

晓露天初曙，
云高万里晴！

Poem composed by Lou Jiann-Hua and presented to Louis during the conference dinner, on behalf of the Department of Mathematics. The poem meant that the first dew appearing early in the morning, clouds are high and it is sunny for ten thousand miles. Key in this poem is that the first word in each line forms Louis' Chinese given name

人生七十载，
学海五十年。
探赜获骊珠，
致远登丹墀。
晓云荡环宇，
鲲鹏翱九天。
一臻如如境，
乾坤亦等闲。

Poem composed by Chen Zehua and presented to Louis during the conference dinner, on behalf of the Department of Statistics and Applied Probability

Contents

Contributors

Andrew D. Barbour Universität Zürich, Zurich, Switzerland, e-mail: a.d.barbour @math.uzh.ch

Mu-Fa Chen Beijing Normal University, Beijing, China, e-mail: mfchen@bnu. edu.cn

Yanchu Chen Hong Kong University of Science and Technology, Hong Kong, China, e-mail: cyxab@ust.hk

Goedele Dierckx Hogeschool-Universiteit Brussel and Katholieke Universiteit Leuven, Brussel, Leuven, Belgium, e-mail: Goedele.Dierckx@hubrussel.be

Larry Goldstein University of Southern California, California, CA, USA, e-mail: larry@math.usc.edu

Takeyuki Hida Nagoya University and Meijo University, Nagoya, Japan, e-mail: takeyuki@math.nagoya-u.ac.jp

Guo-Lie Lan Guangzhou University, Guangzhou, China, e-mail: langl@gzhu.edu.cn

Zheng-yan Lin Zhejiang University, Hangzhou, China, e-mail: zlin@zju.edu.cn

Weidong Liu Shanghai Jiao Tong University, Shanghai, China, e-mail: liuweidong99@gmail.com

Wei-Yin Loh University of Wisconsin, Madison, WI, USA, e-mail: loh@stat.wisc. edu

Zhi-Ming Ma Academy of Math and Systems Science, Beijing, China, e-mail: mazm@amt.ac.cn

Anna Pósfai Tufts University and University of Szeged, Medford, MA, USA, e-mail: anna.posfai@tufts.edu

Adrian Röllin National University of Singapore, Singapore, Singapore, e-mail: adrian.roellin@nus.edu.sg

Qi-Man Shao Hong Kong University of Science and Technology, Hong Kong, China, e-mail: maqmshao@ust.hk

Su-Yong Sun Academy of Math and Systems Science, Beijing, China, e-mail: sunsuy@amss.ac.cn

Jef Teugels Katholieke Universiteit Leuven, Leuven, Belgium, e-mail: Jef.Teugels@wis.kuleuven.be

Aihua Xia The University of Melbourne, Melbourne, VIC, Australia, e-mail: xia@ms.unimelb.edu.au

Part I
Stein's Method

Chapter 1
Couplings for Irregular Combinatorial Assemblies

Andrew D. Barbour and Anna Pósfai

Abstract When approximating the joint distribution of the component counts of a decomposable combinatorial structure that is 'almost' in the logarithmic class, but nonetheless has irregular structure, it is useful to be able first to establish that the distribution of a certain sum of non-negative integer valued random variables is smooth. This distribution is not like the normal, and individual summands can contribute a non-trivial amount to the whole, so its smoothness is somewhat surprising. In this paper, we consider two coupling approaches to establishing the smoothness, and contrast the results that are obtained.

1.1 Introduction

Many of the classical random decomposable combinatorial structures have component structure satisfying a *conditioning relation*: if $C_i^{(n)}$ denotes the number of components of size i in a randomly chosen element of size n, then the distribution of the vector of component counts $(C_1^{(n)}, \ldots, C_n^{(n)})$ can be expressed as

A. D. Barbour (✉)
Angewandte Mathematik, Universität Zürich,
Winterthurertrasse 190, 8057 Zürich, Switzerland
e-mail: a.d.barbour@math.uzh.ch

A. Pósfai
Department of Mathematics, Tufts University,
503 Boston Avenue, Medford, MA 02155, USA

and

Analysis and Stochastics Research Group of the Hungarian Academy of Sciences, Bolyai Institute, University of Szeged,
Aradi vértanúk tere 1, Szeged 6720, Hungary
e-mail: anna.posfai@tufts.edu

A. D. Barbour et al. (eds.), *Probability Approximations and Beyond*, Lecture Notes in Statistics 205, DOI: 10.1007/978-1-4614-1966-2_1,

$$\mathscr{L}(C_1^{(n)}, \ldots, C_n^{(n)}) = \mathscr{L}(Z_1, \ldots, Z_n | T_{0,n} = n), \tag{1.1}$$

where $(Z_i, i \geq 1)$ is a fixed sequence of independent non-negative integer valued random variables, and $T_{a,n} := \sum_{i=a+1}^{n} i Z_i,\ \ 0 \leq a < n$. Of course, $T_{0,n}$ is just the total size of the chosen element, and by definition has to be equal to n; the interest in (1.1) is that, given this necessary restriction, the joint distribution of the component counts is 'as independent as it possibly could be'. The most venerable of these structures is that of a randomly chosen permutation of n elements, with its cycles as components, where one has $Z_i \sim \mathrm{Po}(1/i)$. Random monic polynomials over a finite field of characteristic $q \geq 2$ represent another example, with size measured by degree, and with irreducible factors as components; here, $Z_i \sim \mathrm{NB}(m_i, q^{-i})$, and $q^{-i} m_i \sim 1/i$. Many other examples are given in [1].

In both of the examples above (with $\theta = 1$), and in many others, the Z_i also satisfy the asymptotic relations

$$i\mathbb{P}[Z_i = 1] \to \theta \qquad \text{and} \qquad \theta_i := i\mathbb{E}Z_i \to \theta, \tag{1.2}$$

for some $0 < \theta < \infty$, in which case the combinatorial structure is called *logarithmic*. Arratia, Barbour and Tavaré [1] showed that combinatorial structures satisfying the conditioning relation and slight strengthenings of the logarithmic condition share many common properties, which were traditionally established case by case, by a variety of authors, using special arguments. For instance, if $L^{(n)}$ is the size of the largest component, then

$$n^{-1} L^{(n)} \to_d L, \tag{1.3}$$

where L has probability density function $f_\theta(x) := e^{\gamma\theta}\Gamma(\theta+1)x^{\theta-2}p_\theta((1-x)/x)$, $x \in (0, 1]$, and p_θ is the density of the Dickman distribution P_θ with parameter θ, given in [11, p. 90]. Furthermore, for any sequence $(a_n, n \geq 1)$ with $a_n = o(n)$,

$$\lim_{n\to\infty} \mathrm{d_{TV}}\left(\mathscr{L}(C_1^{(n)}, \ldots, C_{a_n}^{(n)}), \mathscr{L}(Z_1, \ldots, Z_{a_n})\right) = 0. \tag{1.4}$$

Both of these convergence results can be complemented by estimates of the approximation error, under appropriate conditions.

If the logarithmic condition is not satisfied, as in certain of the additive arithmetic semigroups introduced in [5], the results in [1] are not directly applicable. However Manstavičius [7] and Barbour and Nietlispach [4] showed that the logarithmic condition can be relaxed to a certain extent, without disturbing the validity of (1.4), and that (1.3) can also be recovered, if the convergence in (1.2) is replaced by a weaker form of convergence. A key step in the proofs of these results is to be able to show that, under suitable conditions, the distribution of $T_{a_n,n}$ is smooth, in the sense that

$$\lim_{n\to\infty} \mathrm{d_{TV}}\big(\mathscr{L}(T_{a_n,n}), \mathscr{L}(T_{a_n,n}+1)\big) = 0, \qquad \text{for all } a_n = \mathrm{o}(n), \tag{1.5}$$

and that the convergence rate in (1.5) can be bounded by a power of $\{(a_n+1)/n\}$.

Intuitively, the limiting relation (1.5) should hold if (1.4) does, because the approximate independence of $C_1^{(n)},\dots,C_{a_n}^{(n)}$ suggests that the event $\{T_{0,n}=n\}$ has much the same conditional probability, whatever the values taken by $C_1^{(n)},\dots,C_{a_n}^{(n)}$; in other words, the distribution of $T_{a_n,n}+r$ should be much the same, whenever the value r taken by T_{0,a_n} is not too large. Somewhat more formally, using the conditioning relation, and writing $t_{0,a}(c) := \sum_{j=1}^{a} jc_j$, we have

$$\frac{\mathbb{P}[C_1^{(n)}=c_1,\dots,C_a^{(n)}=c_a]}{\mathbb{P}[Z_1=c_1,\dots,Z_a=c_a]}=\frac{\mathbb{P}[T_{a,n}=n-t_{0,a}(c)]}{\mathbb{P}[T_{0,n}=n]},$$

and

$$\frac{\mathbb{P}[T_{0,n}=n]}{\mathbb{P}[T_{a,n}=n-t_{0,a}(c)]}=\sum_{r\ge 0}\mathbb{P}[T_{0,a}=r]\frac{\mathbb{P}[T_{a,n}=n-r]}{\mathbb{P}[T_{a,n}=n-t_{0,a}(c)]},$$

with the right hand side close to *1* if $\mathbb{P}[T_{a,n}=n-r]$ is close to being constant for r in the range of values typically taken by $T_{0,a}$. This latter argument indicates that it is actually advantageous to show that the probability mass function of $T_{a_n,n}$ is flat over intervals on a length scale of a_n, for sequences $a_n=\mathrm{o}(n)$. This is proved in [1, 4] by showing that the normalized sum $n^{-1}T_{a_n,n}$ converges not only in distribution but also locally to the Dickman distribution P_θ, and that the error rates in these approximations can be suitably controlled.

Now, in the case of Poisson distributed Z_i, the distribution of $T_{a,n}$ is a particular compound Poisson distribution, with parameters determined by n and by the θ_i. In [1], the θ_i are all close to a single value θ, and the distribution of $T_{a_n,n}$ is first compared with that of the simpler, standard distribution of $T^*_{0,n} := \sum_{j=1}^{n} jZ^*_j$, where the $Z^*_j \sim \mathrm{Po}(\theta/j)$ are independent. The comparison is made using Stein's method for compound Poisson approximation (cf. [3]), and the argument can be carried through, under rather weak assumptions, even when the Z_i are not Poisson distributed. In a second step, Stein's method is used once more to compare the distribution of $n^{-1}T^*_{0,n}$ with the Dickman distribution P_θ. Both approximations are made in a way that allows the necessary local smoothness of the probability mass function of $T_{a_n,n}$ to be deduced. In [4], the same strategy is used, but the fact that the θ_i may be very different from one another causes an extra term to appear in the bound on the error in the first approximation. In order to control this error, some *a priori* smoothness of the distribution of $T_{a_n,n}$ needs to be established, and a suitable bound on the error in (1.5) turns out to be exactly what is required.

In this note, we explore ways of using coupling to prove bounds on the rate of convergence in (1.5), in the case in which the Z_i have Poisson distributions. This is now just a problem concerning a sum of independent random variables with well-known distributions, and it is tempting to suppose that its solution would be rather simple. For instance, one could take the classical coupling approach to such bounds, known as the Mineka coupling, and described in the next section. The Mineka coupling is very effective for sums T_n of independent and identically distributed