Statistics for Industry and Technology

In Honor of
Sri Gopal Mohanty

SRI GOPAL MOHANTY

Advances in Combinatorial Methods and Applications to Probability and Statistics

N. Balakrishnan
Editor

1997
Birkhäuser
Boston • Basel • Berlin

N. Balakrishnan
Department of Mathematics
and Statistics
McMaster University
Hamilton, Ontario L8S 4K1
Canada

Library of Congress Cataloging-in-Publication Data

Advances in combinatorial methods and applications to probability and statistics / N. Balakrishnan, editor.
p. cm. -- (Statistics for industry and technology)
Includes bibliographical references and index.

ISBN-13: 978-1-4612-8671-4 e-ISBN-13: 978-1-4612-4140-9
DOI: 10.107/978-1-4612-4140-9

paper)
1. Combinatorial probabilities. 2. Mathematical statistics.
I. Balakrishnan, N., 1956- . II. Series.
QA273.45.A38 1997
519.2--DC21 97-6185
CIP

Printed on acid-free paper

Birkhäuser
Softcover reprint of the hardcover 1st edition 1997

ISBN-13: 978-1-4612-8671-4

Typeset by the Editor in LaTeX.
Cover design by Vernon Press, Boston, MA.

9 8 7 6 5 4 3 2 1

Contents

Preface

Sri Gopal Mohanty has made pioneering contributions to lattice path counting and its applications to probability and statistics. This is clearly evident from his lifetime publications list and the numerous citations his publications have received over the past three decades.

My association with him began in 1982 when I came to McMaster University. Since then, I have been associated with him on many different issues at professional as well as cultural levels; I have benefited greatly from him on both these grounds. I have enjoyed very much being his colleague in the statistics group here at McMaster University and also as his friend. While I admire him for his honesty, sincerity and dedication, I appreciate very much his kindness, modesty and broad-mindedness.

Aside from our common interest in mathematics and statistics, we both have great love for Indian classical music and dance. We have spent numerous hours discussing many different subjects associated with the Indian music and dance. I still remember fondly the long drive (to Amherst, Massachusetts) I had a few years ago with him and his wife, Shantimayee, and all the hearty discussions we had during that journey.

Combinatorics and applications of combinatorial methods in probability and statistics has become a very active and fertile area of research in the recent past. This volume has been put together in order to (i) review some of the recent developments in this area, (ii) highlight some of the new noteworthy results and illustrate their applications, and (iii) point out possible new directions in this fruitful area of research.

With these goals in mind, a number of authors actively involved in theory and/or applications of combinatorial methods were invited to write an article for this volume. The articles so collected have been carefully organized into this volume in the form of 32 chapters. For the convenience of the readers, the volume has been divided into following seven parts:

- Lattice Paths and Combinatorial Methods
- Applications to Probability Problems
- Applications to Urn Models
- Applications to Queueing Theory
- Applications to Waiting Time Problems
- Applications to Distribution Theory
- Applications to Nonparametric Statistics

From the above list, it should be clear to the readers that advances in both theory and applications of combinatorial methods have received due attention in this volume. Furthermore, it should also be stressed here that this volume is **not** a proceedings, but rather a volume comprised of carefully collected articles with specific editorial goals (mentioned earlier) in mind.

It has been a very pleasant experience corresponding with all the authors involved, and it is with great pleasure that I dedicate this volume to **Sri Gopal Mohanty**. I sincerely hope that this work will be of interest to mathematicians, theoretical and applied statisticians, and graduate students working on combinatorial methods and their applications to probability and statistics.

Acknowledgments: My sincere thanks go to all the authors who have contributed to this volume, and provided great support and encouragement throughout the course of this project. Special thanks go to Mrs. Debbie Iscoe for the excellent typesetting of the entire volume. Thanks are also due to Shantimayee, Pritidhara, Niharika and Suvankar Mohanty and Dr. Ihor Chorneyko for providing help whenever needed. My final thanks go to Mr. Wayne Yuhasz (Editor, Birkhäuser) for the invitation to undertake this project, and to Ms. Lauren Lavery for her assistance in the production of the volume.

N. BALAKRISHNAN
Hamilton, Ontario, Canada

December 1996

Sri Gopal Mohanty—Life and Works

Sri Gopal Mohanty was born on February 11, 1933 in the village of Soro in Orissa, an eastern state of India on the Bay of Bengal. Orissa is also known as the "land of temples." Sri Gopal is the eldest of four children. His father was a school teacher. His mother was his loving and stern teacher of family and community values. He lived in a large household with his immediate family, two paternal uncles, aunts, and cousins. The village life left an indelible impression on Sri Gopal. Many experiences there would later have bearing: growing up in a joint family, observations of the family's involvement in village drama, his artistic endeavors, and writing a published novel in his native tongue, Oriya.

Sri Gopal attended Satyananda High school in Soro. He earned his BA from Fakhir Mohan College in a town nearby. While in New Delhi from 1951 to 1959, Sri Gopal was working at the Ministry of Food and Agriculture, Directorate of Economics and Statistics and received a diploma from the Indian Council of Agricultural Research in 1957. He continued to work at the Ministry of Food and Agriculture when he pursued and obtained his MA in Mathematics from Punjab University in 1959. Subsequently, he went abroad and was conferred with a PhD in Statistics (based on the Thesis entitled *On some properties of compositions of an integer and their application to probability theory and statistics*, written under the supervision of the late Prof. T. V. Narayana) by the University of Alberta, Edmonton, Canada. He became an associate professor in the University of Buffalo, USA in 1962. In 1963, he returned to Orissa to marry Shantimayee Das. Shantimayee had been a lecturer in Botany after having obtained her MSc. After the birth of their first child in 1964, they moved across the border to Hamilton, Canada where Sri Gopal joined McMaster University as an associate professor in Statistics. During the years 1966 to 1968, he took a leave of absence and travelled to India with his small family to work at the Indian Institute of Technology, Delhi. He resumed his position at McMaster University in 1968 to become a full professor in the early 70's. Until this day, he still holds the position of professor in Statistics at McMaster.

Forging a path in a new country that had adapted multiculturalism as a framework, Sri Gopal brought forth his talents and community values. Fostering the Indian community in Hamilton was one of his passions. He led a local Indian community organization. He was also involved in numerous stage productions for the Indian community and multi-ethnic events; he wore many hats for these cultural productions: conceptualizer, director, and stage manager.

Drama ran in the blood after all. Sri Gopal's eldest uncle, the patriarch of his childhood home, was the village sponsor of local dramatic productions. Sri Gopal's cousin went on to become a movie and TV director in India. Sri Gopal, himself, wrote, adapted, directed, and produced plays about the Indian immigrant experience, women's oppression, and Indian village life conflict.

Sri Gopal has spent considerable time in promoting his own culture and heritage in North America: staging productions involving local community talents, encouraging and inspiring artists, creating awareness of Indian culture in his community and the community at large. These efforts have also won him recognition amongst the people of his state, Orissa, who are in North America: he was awarded the Kalashree Award in 1995 by the Orissa Society of Americas. In 1996, he also won the Community Award of Excellence given by the Hamilton Mayor's Committee Against Racism and Discrimination.

Involvement in the local community at large was important too. He participated as an executive member in the Home and School Association, a Canadian network of parents–teacher organization that advised and assisted schools. Concerned about the age segregation observed in Canada, he encouraged a local school to organize student visits to a nearby home for the elderly people. Sri Gopal also arranged for cultural performances by children from the Canadian Indian community at the elderly home.

Sri Gopal has an avid interest in travel and learning about all international matters. He always encouraged his children to learn about the different peoples and places of the world.

Sri Gopal and Shantimayee Mohanty have three children: Pritidhara, Niharika, and Suvankar. Their eldest daughter, Pritidhara, currently works for the US Environmental Protection Agency, in Washington, DC; Niharika is an Odissi (style of Indian classical dance of Orissa) dancer completing her MA in Dance at York University in Toronto; Suvankar is an undergraduate student in Criminology at Carleton University in Ottawa.

Publications

Book

1. *Lattice Path Counting and Applications*, New York: Academic Press (1979).
2. *A Course on Queueing Models* (with J. L. Jain and W. Böhm), under preparation.
3. *Combinatorics in Lattice Paths* (with C. Krattenthaler), under preparation.

Articles

1. A note on difference equations and combinatorial identities arising out of coin-tossing problems (with T. V. Narayana), *Journal of the Indian Society of Agricultural Statistics*, **17**, 83–86 (1955).
2. Distribution of χ^2-analogue for normal population with class intervals defined in terms of the sample median (with A. R. Roy), *Journal of the Indian Society of Agricultural Statistics*, **10**, 90–98 (1958).
3. A combinatorial problem and its application to probability theory II (with T. V. Narayana and J. C. Ladouceur), *Journal of the Indian Society of Agricultural Statistics*, **12**, 182–189 (1960).
4. Some properties of compositions and their application to probability and statistics I (with T. V. Narayana), *Biometrische Zeitschrift*, **3**, 252–258 (1961).
5. Some properties of compositions and their application to probability and statistics II (with T. V. Narayana), *Biometrische Zeitschrift*, **5**, 8–18 (1963).
6. Some properties of compositions and their application to the ballot problem, *Canadian Mathematical Bulletin*, **8**, 359–372 (1965).
7. An urn problem related to the ballot problem, *The American Mathematical Monthly*, **73**, 526–528 (1966).
8. On a generalized two-coin tossing problem, *Biometrische Zeitschrift*, **8**, 266–272 (1966).
9. Some convolutions with multinomial coefficients and related probability distribution, *SIAM Review*, **8**, 501–502 (1966).

10. An r-coin tossing game and the associated partition of generalized Fibonacci numbers, *Sankhyā, Series A*, **29**, 207–214 (1967).

11. Restricted compositions, *Fibonacci Quarterly*, **5**, 223–234 (1967).

12. On a partition of generalized Fibonacci numbers, *Fibonacci Quarterly*, **6**, 22–33 (1968).

13. On some generalization of a restricted random walk, *Studia Scientiarum Mathematicarum Hungarica*, **3**, 225–241 (1968).

14. On lattice paths with several diagonal steps (with B. R. Handa), *Canadian Mathematical Bulletin*, **12**, 537–545 (1969).

15. Extensions of Vandermonde-type convolutions with several summations and their applications (with B. R. Handa), *Canadian Mathematical Bulletin*, **12**, 45–62 (1969).

16. On some distributions concerning a restricted random walk (with B. R. Handa), *Studia Scientiarum Mathematicarum Hungarica*, **4**, 99–108 (1969).

17. On two types of queueing process involving batches (with J. L. Jain), *Canadian Operational Research Society Journal*, **8**, 38–43 (1970).

18. A generalized Vandermonde-type convolution and associated inverse series relations (with B. R. Handa), *Proceedings of the Cambridge Philosophical Society*, **68**, 457–474 (1970).

19. Rank order statistics related to a generalized random walk (with B. R. Handa), *Studia Scientiarum Mathematicarum Hungarica*, **5**, 267–276 (1970).

20. A short proof of Steck's result on two-sample Smirnov Statistics, *Annals of Mathematical Statistics*, **42**, 413–414 (1971).

21. A note on combinatorial identities for partial sums, *Canadian Mathematical Bulletin*, **14**, 65–67 (1971).

22. The distribution of the maximum queue length, the number of customers served and the duration of the busy period for the queueing systems M/M/1 involving batches, *INFOR*, **8**, 161–166 (1971).

23. On queues involving batches, *Journal of Applied Probability*, **9**, 430–435 (1972).

24. On some distributions of a generalized restricted random walk (with S. Vellore), In *Colloquia Mathematica Societatis Janos Bolyai* (European Meeting of Statisticians, Budapest), **9**, 547–555 (1972).

25. On the enumeration of pseudo-search codes (with I. Z. Chorneyko), *Studia Scientiarum Mathematicarum Hungarica*, **7**, 47–54 (1972).

26. On more combinatorial methods in the theory of queues, In *Mathematical Methods in Queueing Theory*, Lecture Notes in Economics and Mathematical Systems, **98**, pp. 365–371, New York: Springer-Verlag (1974).

27. On the enumeration of certain sets of planted plane trees (with I. Z. Chorneyko), *Journal of Combinatorial Theory*, **18**, 209–221 (1975).

28. Higher dimensional lattice paths with diagonal steps (with B. R. Handa), *Discrete Mathematics*, **15**, 137–140 (1975).

29. Appendix of the book *Lattice Path Combinatorics with Statistical Applications* by T. V. Narayana, University of Toronto Press, Toronto—based on an invited paper with B. R. Handa, presented at the American Mathematical Society, Toronto (1976).

30. Lattice paths and trees, In *Proceedings of the Seventh Southeastern Conference on Combinatorics, Graph Theory and Computing, Congressus Numerantium*, **17**, 447–452 (1976).

31. Enumeration of higher-dimensional paths under restriction (with B. R. Handa), *Discrete Mathematics*, **26**, 119–128 (1979).

32. On Dwass' method for deriving the distribution of rank order statistics (with B. R. Handa), *Aplikace Mathematiky*, **24**, 458–468 (1979).

33. On some computational aspects of rectangular probabilities, In *Colloquia Mathematica Societatis Janos Bolyai* **32**–*Nonparametric Statistical Inference*, 597–617 (1980).

34. Critical values of the Kolmogorov-Smirnov one-sample tests (with H. C. Kaner and J. C. Lyons), *Psychological Bulletin*, **88**, 498–501 (1980).

35. On q-binomial coefficients and some statistical applications (with B. R. Handa), *SIAM Journal of Mathematical Analysis*, **11**, 1027–1035 (1980).

36. The exact distributions of some Kolmogorov-Smirnov type k-sample statistics (with B. R. Handa), In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics and New Directions*, 431–441 (1981).

37. On some distributions related to lattice path combinatorics, In *Proceedings of Statistics '81 Canada Conference*, Montreal: Concordia University, 789–802 (1981).

38. Busy period distributions for two heterogeneous queueing models involving batches (with J. L. Jain), *INFOR*, **19**, 133–139 (1981).

39. Excursion and meander in random walk (with E. Csáki), *Canadian Journal of Statistics*, **9**, 57–70 (1981).

40. A formula for $P(b_i \leq R_i \leq a_i,\, 1 \leq i \leq m)$ under a general class of alternatives (with B. R. Handa), *Journal of Statistical Planning and Inference*, **6**, 335–343 (1982).

41. On the distribution in a graph process (with B. R. Handa), *Discrete Mathematics*, **50**, 319–322 (1984).

42. On a multiserver Markovian queueing system with balking and reneging (with A. J. Montazer-Haghighi and J. Medhi), *Computers and Operations Research*, **13**, 421–425 (1986).

43. Some joint distributions for conditional random walks (with E. Csáki), *Canadian Journal of Statistics*, **14**, 19–28 (1986).

44. On non-null distribution of ranks (with D. Callfas), *Communications in Statistics-Simulation and Computation*, **15**, 591–608 (1986).

45. On a property of lattice paths (with B. R. Handa), *Journal of Statistical Planning and Inference*, **14**, 59–62 (1986).

46. On an inclusion-exclusion formula based on the reflection principle (with T. Watanabe), *Discrete Mathematics*, **64**, 281–288 (1987).

47. A discrete-time analogue of the M/M/1 queue and the transient solution: an analytic approach (with W. Panny), In *Colloquia Mathematica Societatis Janos Bolyai* **57**–*Limit Theorems in Probability and Statistics*, 417–424 (1989).

48. On random walks in a plane (with E. Csáki and J. Saran), *Ars Combinatoria*, **29**, 309–318 (1990).

49. A discrete-time analogue of the M/M/1 queue and the transient solution: a geometric approach (with W. Panny), *Sankhyā, Series A*, **52**, 364–370 (1990).

50. On random walks in a plane (with E. Csáki and J. Saran), *Ars Combinatoria*, **29**, 309–318 (1990).

51. The transient solution of M/M/1 queues under (M,N)-policy. A combinatorial approach (with W. Böhm), *Journal of Statistical Planning and Inference*, **34**, 23–33 (1993).

52. Enumeration of restricted three-dimensional lattice paths with fixed number of turns and an application (with I. Z. Chorneyko and Kanwar Sen), *Journal of Statistical Planning and Inference*, **34**, 57–62 (1993).

53. On the transient behavior of a finite birth-death process with an application (with A. Montazer-Haghighi and R. Trueblood), *COR*, **20**, 230–248 (1993).

54. On lattice path counting by major index and descents (with C. Krattenthaler), *European Journal of Combinatorics*, **14**, 43–51 (1993).

55. On zero-avoiding transition probabilities on an r-node tandem queue. A combinatorial approach (with W. Böhm and J. L. Jain), *Journal of Applied Probability*, **30**, 737–741 (1993).

56. On discrete time Markovian N-policy queues involving batches (with W. Böhm), *Sankhyā, Series A*, **56**, 144–163 (1994).

57. On random walks with barriers and their application to queues (with W. Böhm), *Studia Scientiarum Mathematicarum Hungarica*, **29**, 397–413 (1994).

58. Transient analysis of M/M/1 queues in discrete time with general server vacations (with W. Böhm), In *Studies in Probability:* A special volume of the *Journal of Applied Probability*, 115–129 (1994).

59. q-Generalization of a ballot problem (with C. Krattenthaler), *Discrete Mathematics*, **126**, 195–208 (1994).

60. Transient analysis of queues with heterogeneous arrivals (with W. Böhm), *QUESTA*, **18**, 27–45 (1994).

61. Success runs of length k in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **46**, 777–796 (1994).

62. Counting tableaux with row and column bounds (with C. Krattenthaler), *Discrete Mathematics*, **139**, 273–285 (1995).

63. Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials (with S. Aki and N. Balakrishnan), *Annals of the Institute of Statistical Mathematics*, **48**, 773–787 (1996).

64. Start-up demonstration tests under Markov dependence model with corrective actions (with N. Balakrishnan and S. Aki), To appear in *Annals of the Institute of Statistical Mathematics*, **49** (1997).

65. Extensions of Bonferroni type inequalities (with B. R. Handa), To appear in *Studia Scientiarum Mathematicarum Hungarica* (1997).

66. On the Karlin-McGregor theorem and applications (with W. Böhm), To appear in *Annals of Probability* (1997).

Contributors

Agarwal, Manju Department of Operational Research, Faculty of Mathematical Sciences, University of Delhi, Delhi-110007, India

Aki, Sigeo Department of Mathematical Science, Facultyof Engineering Science, Osaka University, Toyonaka, Osaka 560, Japan
e-mail: *aki@sigmath.es.osaka-u.ac.jp*

Andersen, Erik Sparre University of Copenhagen, Høeghsmindeparken 15, DK 2900 Hellerup, Denmark
e-mail: *sparre@math.ku.dk*

Antzoulakos, Demetrios L. The School of Social Sciences, Department of Economics, University of Crete, Perivolia - Rethymno 74100, Greece

Balakrishnan, N. Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1
e-mail: *bala@mcmail.cis.mcmaster.ca*

Böhm, Walter Department of Statistics, University of Economics, A-1090 Vienna, Augasse 2-6, Austria
e-mail: *boehm@wu-wien.ac.at*

Chakraborti, S. Department of Management Science andStatistics, University of Alabama, P.O. Box 870226, Tuscaloosa, AL 35487-0226
e-mail: *schakrab@ua1vm.ua.edu*

Charalambides, Ch. A. University of Athens, Panepistemiopolis, 15784 Athens, Greece
e-mail: *ccharal@atlas.uoa.ariadne-t.gr*

Csáki, Endre Mathematical Institute of the Hungarian Academy of Sciences, 1053 Budapest V., Reáltanoda u. 13-15, Hungary

Galambos, Janos Department of Mathematics, Temple University, Philadelphia, PA 19122
e-mail: *janos@euclid.math.temple.edu*

Gessel, Ira M. Department of Mathematics, BrandeisUniversity, Waltham, MA 02254-9110
e-mail: *gessel@math.brandeis.edu*

Govindarajulu, Z. Department of Statistics, University of Kentucky, 839 Patterson Office Tower, Lexington, KT 40506-0027
e-mail: *raju@ms.uky.edu*

Günalay, Yavuz Michael G. DeGroote School of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada

Gupta, Diwakar Michael G. DeGroote School of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada
e-mail: *guptad@mcmail.cis.mcmaster.ca*

Hirano, Katuomi The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo 106, Japan

Hušková, M. Department of Statistics, Charles University, Sokolovksá 83, 186 00 Praha, Czech Republic
e-mail: *huskova@karlin.mff.cuni.cz*

Janardan, K. G. Department of Mathematics, Eastern Michigan University, Ypsilanti, MI 48197

Katzenbeisser, Walter Institut für Statistik, Abteilung für Mathematische, Methoden der Statistik, University of Economics, A-1090 Wien, Augasse 2-6, Austria

Kemp, Adrienne W. Mathematical Institute, University of St Andrews, North Haugh, St Andrews, KY16 9SS, Scotland
e-mail: *awk@st-and.ac.uk*

Kemp, C. David Mathematical Institute, University of St Andrews, North Haugh, St Andrews, KY16 9SS, Scotland
e-mail: *cdk@st-and.ac.uk*

Kotz, Samuel Department of Management Sciences and Statistics, University of Maryland, College Park, MD 20742

Koutras, Markos Department of Mathematics, University of Athens, Panepistemiopolis, 15784 Athens, Greece
e-mail: *mkoutras@atlas.uoa.gr*

Krattenthaler, C. Institut für Mathematik der Universität Wien, Strudlhofgasse 4, A-1090 Wien, Austria
e-mail: *KRATT@pap.Univie.Ac.At*

Larsen, Mogens Esrom Department of Mathematics, University of Copenhagen, Universitetsparken 5, Copenhagen, Denmark
e-mail: *mel@math.ku.dk*

Medhi, J. Institute of Advanced Study in Science and Technology, Guwahati, U.N. Bezbarua Road, Silpukhuri, Guwahati-781 003, India

Nevzorov, Valery B. Department of Mathematics and Mechanics, St.-Petersburg State University, Bibliotechnaya Square-2, St.-Petersburg, Russia 198904
e-mail: *uuvaler@valery.niimm.spb.su*

Niederhausen, Heinrich Department of Mathematics, Florida Atlantic University, Boca Raton, FL 33431
e-mail: *niederha@acc.fau.edu*

Nikulin, M. S. UFR MIS2, Université Bordeaux 2, Bordeaux, France; Laboratory of Statistical Methods, Steklov Mathematical Institute, St. Petersburg, Russia
e-mail: *M.S.Nikouline@u-bordeaux2.fr*

Panny, W. Institut für Statistik, Abteilung für Mathematische, Methoden der Statistik, University of Economics, A-1090 Wien, Augasse 2-6, Austria
e-mail: *panny@exai2.wu-wien.ac.at*

Philippou, Andreas N. Department of Mathematics, University of Patras, 261 10 Patras, Greece

Rahman, M. Mahibbur University of Swaziland, Swaziland

Rani, Sarita Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi-110007, India

Ree, Sangwook Department of Mathematics, SuWon University, Kyung-Ki-Do, Korea, 445 743
e-mail: *swree@math.snu.ac.kr*

Saran, Jagdish Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi-110007, India

Schaafsma, W. Department of Mathematics, University of Groningen, P.O. Box 800, 9700 AV Groningen, The Netherlands

Sen, Kanwar Department of Statistics, Faculty of Mathematical Sciences, University of Delhi, Delhi-11007, India

Sobel, Milton Department of Statistics and Applied Probability, University of California, Santa Barbara, CA 93106
e-mail: *sobel@pstat.ucsb.edu*

Takács, Lajos Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106

Tőrös, R. Research Institute of Technical Chemistry of Hungarian Academy of Sciences, Budapest, Hungary

Tripsiannis, Gregory A. Department of Medical Statistics, Faculty of Medicine, Democritus University of Thrace, Alexandroupolis, Greece

Uchida, Masayuki The Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-Ku, Tokyo 106, Japan

Vincze, Istvan Mathematical Institute of the Hungarian Academy of Sciences, Budapest, P.O.B. 127, H-1364, Hungary

Voinov, V. G. Institute of Pure and Applied Mathematics, Ministry of Science - Academy of Sciences of the Republic of Kazakhstan, Almaty, Kazakhstan
e-mail: *voinov@itvms.itpm.alma-ata.su*

Watanabe, Toshihiro Department of Applied Mathematics, Faculty of Engineering, GIFU University, Gifu, 501-11, Japan
e-mail: *wata@gifu-u.ac.jp*

Zeilberger, Doron Department of Mathematics, Temple University, Philadelphia, PA 19122
e-mail: *zeilberg@euclid.math.temple.edu*

Tables

Figures

PART I

Lattice Paths and Combinatorial Methods

1

Lattice Paths and Faber Polynomials

Ira M. Gessel and Sangwook Ree

Brandeis University, Waltham, MA
SuWon University, Kyung-Ki-Do, Korea

Abstract: The r-th Faber polynomial of the Laurent series $f(t) = t + f_0 + f_1/t + f_2/t^2 + \cdots$ is the unique polynomial $F_r(u)$ of degree r in u such that $F_r(f) = t^r +$ negative powers of t. We apply Faber polynomials, which were originally used to study univalent functions, to lattice path enumeration.

Keywords and phrases: Lattice path enumeration, ballot problem, Faber polynomials

1.1 Introduction

The classical ballot problem [see, for example, Mohanty (1979)] asks for the number $B(m, n)$ of paths from $(1, 0)$ to (m, n) (where $m > n$), with unit steps up and to the right, that never touch the line $x = y$. The number $B(m, n)$ can easily be computed by the recurrence

$$B(m,n) = B(m-1,n) + B(m,n-1) \quad \text{for } m > n \geq 0,\ (m,n) \neq (1,0),$$

with the initial condition $B(1, 0) = 1$ and the boundary conditions $B(m, -1) = 0$ and $B(m, m) = 0$ for all $m \geq 0$. Displaying these values on the corresponding lattice points, we have the following array, showing $B(m, n)$ for $m \geq n \geq 0$:

n/m	0	1	2	3	4	5
5						0
4					0	14
3				0	5	14
2			0	2	5	9
1		0	1	2	3	4
0	0	1	1	1	1	1

Let us now extend the values of $B(m,n)$ to the region in which $n > m \geq 0$ so that the same recurrence is satisfied; this can be done in only one way, since we may write the recurrence as $B(m-1,n) = B(m,n) - B(m,n-1)$. We obtain the following array:

$$\begin{array}{rrrrrr}
-1 & -4 & -9 & -14 & -14 & 0 \\
-1 & -3 & -5 & -5 & 0 & 14 \\
-1 & -2 & -2 & 0 & 5 & 14 \\
-1 & -1 & 0 & 2 & 5 & 9 \\
-1 & 0 & 1 & 2 & 3 & 4 \\
0 & 1 & 1 & 1 & 1 & 1
\end{array}$$

We observe that the recurrence $B(m,n) - B(m-1,n) - B(m,n-1) = 0$ is now satisfied for all $m,n \geq 0$ except $(m,n) = (1,0)$ and $(m,n) = (0,1)$, as long as we take $B(m,n)$ to be 0 for $m < 0$ or $n < 0$. In terms of generating functions, the recurrence and initial conditions are equivalent to the formula

$$(1-x-y) \sum_{m,n=0}^{\infty} B(m,n)x^m y^n = x - y,$$

which gives

$$\sum_{m,n=0}^{\infty} B(m,n)x^m y^n = \frac{x-y}{1-x-y}. \tag{1.1}$$

Following MacMahon, we may call (1.1) a "redundant generating function," since it contains some terms which are not part of the solution of the original problem.

From (1.1) we may derive the well-known formula for the ballot numbers,

$$B(m,n) = \binom{m+n-1}{m-1} - \binom{m+n-1}{m} = \frac{m-n}{m+n}\binom{m+n}{m}. \tag{1.2}$$

There is a gap in our derivation of (1.1). It is clear that the numbers $B(m,n)$ defined by (1.1) do indeed have the property that for $m > n \geq 0$,

$$B(m,n) - B(m-1,n) - B(m,n-1) = \begin{cases} 1 & \text{if } (m,n) = (1,0) \\ 0 & \text{otherwise.} \end{cases}$$

However, we have not yet proved that the boundary condition $B(m,m) = 0$ is satisfied. This follows easily from the explicit formula (1.2), or from the fact that the generating function (1.1) is anti-symmetric. The proof that the coefficients of (1.1) are indeed the solution to our problem is now complete.

By exactly the same reasoning, we find that for any positive integer r and any nonnegative integers $m > n \geq 0$, the number of paths from $(r,0)$ to (m,n) that never touch the line $x = y$ is the coefficient of $x^m y^n$ in $(x^r - y^r)/(1-x-y)$.

We can try a similar approach to paths that begin at $(1,0)$ and stay below the line $x = 2y$. Here the recurrence is again $C(m,n) = C(m-1,n)+C(m,n-1)$, but the boundary condition is $C(2n,n) = 0$. Extending the recurrence to the region $m < 2n$, we obtain the following array:

$$\begin{array}{rrrrrrr}
-2 & -5 & -8 & -10 & -10 & -7 & 0 \\
-2 & -3 & -3 & -2 & 0 & 3 & 7 \\
-2 & -1 & 0 & 1 & 2 & 3 & 4 \\
0 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}$$

As before, we find that the extended function $C(m,n)$, with $C(m,n) = 0$ for $m < 0$ or $n < 0$, satisfies the recurrence $C(m,n) = C(m-1,n) + C(m,n-1)$ everywhere except when (m,n) is $(1,0)$ or $(0,1)$, and thus the generating function for the extended function is apparently

$$\frac{x-2y}{1-x-y}, \tag{1.3}$$

from which we may derive the formula

$$C(m,n) = \binom{m+n-1}{m-1} - 2\binom{m+n-1}{m} = \frac{m-2n}{m+n}\binom{m+n}{n}.$$

To complete the proof, we must show that the coefficient of $x^{2n}y^n$ in (1.3) is indeed zero. Although this may be seen from the explicit formula for the coefficients, we use a different method that we will need later on. Let us substitute xt for x and y/t^2 for y in (1.3). Then it suffices to show that the constant term in t in

$$\frac{xt-2y/t^2}{1-xt-y/t^2},$$

when expanded as a power series in x and y, is zero. But

$$\frac{xt-2y/t^2}{1-xt-y/t^2} = t\,\frac{d}{dt}\left\{\log\frac{1}{1-xt-y/t^2}\right\},$$

and since the coefficient of $1/t$ in the derivative with respect to t of a Laurent series in t is 0, the desired conclusion follows.

Note that this approach cannot easily be applied to paths that are required to stay below the line $y = 2x$: here we would require the boundary conditions $C(m,2m) = 0$ and $C(m,2m+1) = 0$, and this is not so easily achieved. However, there is no problem with paths starting at $(1,0)$ that stay below the line $x = py$, where p is a positive integer, and we find in this case the generating function $(x-py)/(1-x-y)$.

We now consider one final example before embarking on the general case. Suppose we want to count paths from $(3,0)$ to (m,n) that stay below the line

$x = 2y$, where $m > 2n$. The same recurrence is satisfied, and as before, we may extend its solution into the region where $m < 2n$, obtaining the following array:

$$\begin{array}{rrrrrrr}
-2 & -7 & -15 & -25 & -35 & -42 & -42 \\
-2 & -5 & -8 & -10 & -10 & -7 & 0 \\
0 & -3 & -3 & -2 & 0 & 3 & 7 \\
0 & 0 & 0 & 1 & 2 & 3 & 4 \\
0 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}$$

The recurrence is satisfied except at the points $(3,0)$, $(1,2)$, and $(0,3)$, so the generating function is apparently

$$\frac{x^3 - 3xy^2 - 2y^3}{1 - x - y}. \tag{1.4}$$

To prove this we must show that the coefficient of $x^{2n}y^n$ in (1.4) is zero, which we can do just as in the previous example: we replace x with xt and y with y/t^2. Then we have

$$\frac{x^3t^3 - 3xy^2/t^3 - 2y^3/t^6}{1 - xt - y/t^2} = t\,\frac{d}{dt}\left\{\log\frac{1}{1-P(t)} - P(t) - P(t)^2/2\right\}, \tag{1.5}$$

where $P(t) = xt + y/t^2$, so the constant term in t in (1.5) is zero.

In the remainder of this paper, we shall develop the general theory of which (1.5) is a special case. It will turn out that the numerator in (1.5) and its generalizations are closely related to certain polynomials called *Faber polynomials* which have been studied in connection with univalent functions [see Schiffer (1948); and also Brini (1984), Jabotinsky (1953) and Schur (1945)]. Faber polynomials were first applied to lattice path enumeration, in the special case we consider in Section 1.5, by Ree (1994).

1.2 Faber Polynomials

Let

$$f(t) = t + f_0 + \frac{f_1}{t} + \frac{f_2}{t^2} + \cdots.$$

In the original applications of Faber polynomials, the f_i are complex numbers, and the series converges in some neighborhood of infinity. However, for our applications we take t and the f_i to be indeterminates; i.e., we work in the ring of formal Laurent series $\boldsymbol{C}[[t, f_0, f_1/t, f_2/t^2, \ldots]]$.

Let $F(u)$ be a polynomial in u of degree r such that

$$F(f) = t^r + \text{negative powers of } t.$$

We say that $F(u)$ is a *Faber polynomial* of f. It is easy to prove by induction that there is exactly one Faber polynomial $F_r(u)$ of degree r, which we call the *rth Faber polynomial of f*. For example, we have $F_1(u) = u - f_0$ and $F_2(f) = u^2 - 2f_0u + (f_0^2 - 2f_1)$.

Schiffer (1948) gave the generating function

$$\log \frac{f(v) - u}{v} = -\sum_{r=1}^{\infty} F_r(u)\frac{v^{-r}}{r}. \tag{1.6}$$

If we set $f(v) = vh(1/v)$, so that $h(w) = 1 + \sum_{i=0}^{\infty} f_i w^{i+1}$ is a power series in w, then (1.6) may be rewritten in terms of formal power series as

$$\log\{h(w) - uw\} = -\sum_{r=1}^{\infty} F_r(u)\frac{w^r}{r}\ . \tag{1.7}$$

Expanding (1.6) or (1.7) gives the explicit formula

$$F_r(u) = \sum_{i=0}^{r} u^i \sum_{i+j_0+2j_1+3j_2+\cdots=r} (-1)^{j_0+j_1+\cdots} r\ \frac{(i-1+j_0+j_1+\cdots)!}{i!\,j_0!\,j_1!\cdots}\ f_0^{j_0} f_1^{j_1}\cdots .$$

1.3 Counting Paths

Let r be a positive integer and let k and n be nonnegative integers. Let S be a subset of the set $\{1, 0, -1, -2, \cdots\}$. We call the elements of S *steps*. We want to count sequences $(s_1, s_2, \ldots, s_n)$ of elements of S such that every partial sum $r + s_1 + s_2 + \cdots + s_i$ is positive and $r + s_1 + s_2 + \cdots + s_n = k$. We call such a sequence of steps a *good path of length n from r to k*. The ballot problem is equivalent to the case $S = \{1, -1\}$, with $r = 1$, and the other problems discussed in Section 1.1 are all equivalent to specializations of the case $S = \{1, -p\}$ for various values of p, r, and k.

It is convenient to consider a somewhat more general problem: We take as our set of steps the entire set $\{1, 0, -1, -2, \cdots\}$, but we assign to each path $(s_1, s_2, \ldots, s_n)$ the weight $f_{-s_1} f_{-s_2} \cdots f_{-s_n}$, where f_0, f_1, f_2,... are indeterminates and $f_{-1} = 1$. The condition that $f_{-1} = 1$ simplifies all our formulas, but does not lose any information.

Lemma 1.3.1 *A path from r to k with weight* $f_0^{j_0} f_1^{j_1} \cdots$ *has* $k - r + j_1 + 2j_2 + \cdots$ *steps equal to 1, and length* $k - r + j_0 + 2j_1 + \cdots$.

PROOF. Let j_{-1} be the number of steps equal to 1. Since the path is from r to k, we have $r + j_{-1} - 0j_0 - 1j_1 - 2j_2 - \cdots = k$, and the first assertion follows. Then the length of the path is $j_{-1} + j_1 + j_2 + \cdots = k - r + j_0 + 2j_1 + \cdots$. ∎

We now fix r throughout the rest of this section. Let $G(n,k)$ be the sum of the weights of all good paths of length n from r to k. Thus, the coefficient of $f_0^{j_0} f_1^{j_1} \cdots$ in $G(n,k)$ is the number of good paths of length n from r to k with j_0 steps equal to 0, j_1 steps equal to -1, and so on.

The following is clear.

Lemma 1.3.2

(i) $G(n,0) = 0$ *for all* n;

(ii) $G(0,r) = 1$ *and* $G(0,k) = 0$ *for* $k \neq r$;

(iii) For $n > 0$, $G(n,k) = \sum_{i=-1}^{\infty} f_i G(n-1, k+i)$.

Moreover, $G(n,k)$ *is uniquely determined by conditions (i)–(iii).*

Now, let us define

$$G_k = \sum_{n=0}^{\infty} G(n,k).$$

By Lemma 1.3.1, we can recover $G(n,k)$ from G_k as the sum of all terms in G_k involving $f_0^{j_0} f_1^{j_1} \cdots$, where $k - r + j_0 + 2j_1 + \cdots = n$.

Now let $f(t)$, as in Section 1.2, be the formal Laurent series

$$f(t) = t + f_0 + \frac{f_1}{t} + \frac{f_2}{t^2} + \cdots.$$

We use the notation $[t^i]A(t)$ to denote the coefficient of t^i in $A(t)$.

Lemma 1.3.3 *Let* $N(t)$ *be a Laurent series in* t *such that*

(a) $N(t) = t^r +$ *negative powers of* t

(b) $[t^0]N(t)/\{1 - f(t)\} = 0$.

Then for $k > 0$, $G_k = [t^k]N(t)/\{1 - f(t)\}$.

Proof. Suppose that the hypotheses of the lemma are satisfied. For $k \geq 0$, let

$$g_k = [t^k]\frac{N(t)}{1 - f(t)}$$

and for each integer n, let $g(n,k)$ be the sum of all terms in g_k involving $f_0^{j_0} f_1^{j_1} \cdots$, where

$$k - r + j_0 + 2j_1 + \cdots = n. \tag{1.8}$$

By Lemma 1.3.2, it suffices to show

(i) $g(n,0) = 0$ for all n;

(ii) $g(0, r) = 1$ and $g(0, k) = 0$ for $k \neq r$;

(iii) For $n > 0$, $g(n, k) = \sum_{i=-1}^{\infty} f_i g(n-1, k+i)$.

First note that (i) follows immediately from (b). By the definition of g_k, we have

$$\frac{N(t)}{1 - f(t)} = \sum_{k=1}^{\infty} g_k t^k + t^{-1} R(t),$$

where $R(t)$ is a power series in t^{-1}. Multiplying both sides by $1 - f(t)$, we get

$$N(t) = \{1 - f(t)\} \sum_{k=1}^{\infty} g_k t^k + S(t),$$

where $S(t) = \{1 - f(t)\} t^{-1} R(t)$ is a power series in t^{-1}. Equating coefficients of t^k for $k > 0$ on both sides and using (a), we obtain

$$g_k - \sum_{i=-1}^{\infty} f_i g_{k+i} = \begin{cases} 1 & \text{if } k = r \\ 0 & \text{if } k \neq r. \end{cases} \tag{1.9}$$

Extracting the terms in $f_0^{j_0} f_1^{j_1} \cdots$, where $k - r + j_0 + 2j_1 + \cdots = n$, we obtain

$$g(n, k) - \sum_{i=-1}^{\infty} f_i g(n-1, k+i) = \begin{cases} 1 & \text{if } k = r \text{ and } n = 0 \\ 0 & \text{otherwise,} \end{cases} \tag{1.10}$$

since the nonzero case of (1.9) contributes to (1.10) only when $k = r$ and $j_0 = j_1 = \cdots = 0$. This proves (iii). Finally, (ii) will follow from the $n = 0$ case of (1.10) once we show that $g(-1, k) = 0$ for all k. We show in fact that $g(n, k) = 0$ for all $n < 0$: It is clear from (1.8) that $g(n, k) = 0$ for $n < -r$. It then follows from (1.10) by induction on n that $g(n, k) = 0$ for all negative n. Thus, (ii) holds. ∎

Theorem 1.3.1 *G_k is the coefficient of t^k in*

$$\frac{t}{r} \frac{d}{dt} F_r(f)/(1 - f),$$

where $F_r(u)$ is the r-th Faber polynomial of f.

PROOF. It follows from the definition of Faber polynomials that

$$\frac{t}{r} \frac{d}{dt} F_r(f) = t^r + \text{negative powers of t.}$$

In view of Lemma 1.3.3, it is sufficient to show that

$$\frac{d}{dt} F_r(f)/(1 - f)$$

is the derivative of some Laurent series in t, since this will imply that it has no term in t^{-1}.

Let $F_r(u) = \sum_{i=0}^{r} c_i u^i$. Then

$$\frac{d}{dt}\, F_r(f)/(1-f) = \sum_{i=1}^{r} i c_i f^{i-1} f' \sum_{j=0}^{\infty} f^j = \sum_{i=1}^{r} \sum_{j=0}^{\infty} i c_i f^{i+j-1} f'.$$

But $f^{i+j-1} f' = \frac{d}{dt}\, f^{i+j}/(i+j)$, so

$$\frac{d}{dt}\, F_r(f)/(1-f) = \frac{d}{dt} \sum_{i=1}^{r} \sum_{j=0}^{\infty} \frac{i c_i}{i+j}\, f^{i+j}. \qquad \blacksquare$$

1.4 A Positivity Result

Let $N_r = \frac{t}{r} \frac{d}{dt}\, F_r(f)$ be the numerator in Theorem 1.3.1. We know that $N_r = t^r - M_r$, where M_r contains only negative powers of t.

Theorem 1.4.1 *The coefficients of M_r, as a power series in $t^{-1}, f_0, f_1, \ldots$ are nonnegative integers.*

PROOF. By setting $u = f(t)$ in Schiffer's formula (1.6), and then differentiating with respect to t, we obtain

$$\frac{t f'(t)}{f(v) - f(t)} = \sum_{r=1}^{\infty} N_r v^{-r}. \tag{1.11}$$

Thus,

$$\begin{aligned} \sum_{r=1}^{\infty} M_r v^{-r} &= \sum_{r=1}^{\infty} \left[\left(\frac{t}{v} \right)^r - N_r v^{-r} \right] \\ &= \frac{t}{v-t} - \frac{t f'(t)}{f(v) - f(t)} \\ &= t \frac{v-t}{f(v) - f(t)} \left[\frac{f(v) - f(t)}{(v-t)^2} - \frac{f'(t)}{v-t} \right]. \end{aligned} \tag{1.12}$$

We shall show that the last two factors in (1.12) have positive coefficients when expanded as power series in v^{-1} and t^{-1}. First, we have

$$\frac{f(v) - f(t)}{v-t} = \sum_{i=-1}^{\infty} f_i \left(\frac{v^{-i} - t^{-i}}{v-t} \right) = 1 - \sum_{i=1}^{\infty} f_i \left(\frac{1}{v t^i} + \frac{1}{v^2 t^{i-1}} + \cdots + \frac{1}{v^i t} \right).$$

Thus, $(v-t)/\{f(v) - f(t)\}$ has nonnegative coefficients.

Next, we have

$$\frac{f(v)-f(t)}{(v-t)^2}-\frac{f'(t)}{v-t}=\sum_{i=-1}^{\infty} f_i\left(\frac{v^{-i}-t^{-i}}{(v-t)^2}+\frac{it^{-i-1}}{v-t}\right). \tag{1.13}$$

The coefficient of f_i in (1.13) is zero for $i=-1$ and $i=0$. It is easily verified (by multiplying both sides by $(v-t)^2$, for example) that for $i \geq 1$,

$$\frac{v^{-i}-t^{-i}}{(v-t)^2}+\frac{it^{-i-1}}{v-t}=\sum_{j=1}^{i}\frac{j}{v^{i-j+1}t^{j+1}},$$

and thus it follows that the coefficients of (1.13) are nonnegative. ∎

1.5 Examples

Let us now return to the problem discussed in the first section: given positive integers r and p, count paths in the plane with steps $(1,0)$ and $(0,1)$, from $(r,0)$ to (m,n), where $m > pn$, that never touch the line $x = py$. (Note that any starting point below the line $x = py$ would give an equivalent problem.) We can convert this problem to an instance of the problem introduced in Section 1.3 by representing a horizontal step by a step equal to 1 and a vertical step by a step equal to $-p$. The transformed path will then go from r to k, where $k = m - pn$. The solution to the transformed problem is then obtained by setting to zero all the f_i except f_p in the general solution given in Theorem 1.3.1, where the weight of the transformed path is f_p^n. Explicitly, the required number is the coefficient of $t^{m-pn}f_p^n$ in

$$\frac{t}{r}\frac{d}{dt}F_r(t+f_p/t^p)/(1-t-f_p/t^p), \tag{1.14}$$

where the Faber polynomials $F_r(u)$ are given from (1.7) by

$$\begin{aligned}
\sum_{r=1}^{\infty} F_r(u)\frac{w^r}{r} &= -\log(1+f_p w^{p+1}-uw)\\
&= \sum_{j=1}^{\infty}(uw-f_p w^{p+1})^j/j\\
&= \sum_{j=1}^{\infty}\sum_{i=0}^{j}\frac{(-1)^i}{j}\binom{j}{i}(f_p w^{p+1})^i(uw)^{j-i}\\
&= \sum_{j=1}^{\infty}\sum_{i=0}^{j}\frac{(-1)^i}{j}\binom{j}{i}f_p^i w^{pi+j}u^{j-i}.
\end{aligned}$$

Setting $j = r - pi$ and equating coefficients of w^r, we obtain

$$\frac{F_r(u)}{r} = \sum_{i \le r/(p+1)} \frac{(-1)^i}{r - pi} \binom{r - pi}{i} f_p^i u^{r-(p+1)i},$$

and thus the numerator in (1.14) is

$$\begin{aligned}
&\frac{t}{r}\, F_r'(t + f_p/t^p)(1 - pf_p/t^{p+1}) \\
&= (t - pf_p/t^p) \sum_{i \le r/(p+1)} (-1)^i\, \frac{r-(p+1)i}{r - pi} \binom{r - pi}{i} f_p^i (t + f_p/t^p)^{r-(p+1)i-1} \\
&= (t - pf_p/t^p) \sum_{i < r/(p+1)} (-1)^i \binom{r - pi - 1}{i} f_p^i (t + f_p/t^p)^{r-(p+1)i-1}. \qquad (1.15)
\end{aligned}$$

To recover a generating function in x and y, as in Section 1.1, we substitute x for t and $x^p y$ for f_p. Then, (1.14) and (1.15) give as the redundant generating function for our problem

$$\frac{x - py}{1 - x - y} \sum_{i < r/(p+1)} (-1)^i \binom{r - pi - 1}{i} (x^p y)^i (x + y)^{r-(p+1)i-1}. \qquad (1.16)$$

For example, if we take $p = 2$ and $r = 3$, (1.16) gives

$$\frac{x - 2y}{1 - x - y}\,(x + y)^2 = \frac{x^3 - 3xy^2 - 2y^3}{1 - x - y},$$

as in (1.4).

Now, let (1.16) equal $\widetilde{N}_r/(1 - x - y)$ and let $\widetilde{N}_r = x^r - \widetilde{M}_r$. Then, $\widetilde{N}_r$ and $\widetilde{M}_r$ can be obtained from N_r and M_r as defined in Section 1.4 by setting $t = x$, $f_p = x^p y$, and $f_i = 0$ for $i \ne p$. Since it is clear from (1.16) that $\widetilde{N}_r$ and $\widetilde{M}_r$ are homogeneous of degree r in x and y, they are determined by the sums $\sum_r \widetilde{N}_r$ and $\sum_r \widetilde{M}_r$. The formulas in the proof of Theorem 1.4.1 give

$$\sum_{r=1}^{\infty} \widetilde{N}_r = \frac{x - py}{1 - x - y + x^p y} \qquad (1.17)$$

and

$$\sum_{r=1}^{\infty} \widetilde{M}_r = y\, \frac{p + (p-1)x + (p-2)x^2 + \cdots + x^{p-1}}{1 - y(1 + x + x^2 + \cdots + x^{p-1})}. \qquad (1.18)$$

For $p = 1$, (1.18) gives $\widetilde{M}_r = y^r$, so that $\widetilde{N}_r = x^r - y^r$, as we observed in Section 1.1. We can also obtain a simple explicit formula when $p = 2$. In this case, (1.18) gives

$$\sum_{r=1}^{\infty} \widetilde{M}_r = y \frac{(2 + x)}{1 - y(1 + x)} = \sum_{i,j=0}^{\infty} x^i y^{j+1} \left[\binom{i}{j} + \binom{i+1}{j} \right].$$

Extracting the terms that are homogeneous of degree r and simplifying, we obtain

$$\widetilde{M}_r = \sum_{i=0}^{[r/2]} \frac{2r-3i}{r-i}\binom{r-i}{i} x^i y^{r-i}.$$

The method we have described can be used for counting paths in the plane that stay below the line $x = py$, with arbitrary starting and ending points, and an arbitrary set of allowed steps subject only to the condition that every step (i,j) satisfies $i - pj \le 1$. Another method, also using Laurent series, which is not subject to the restriction on steps, but does not allow an arbitrary starting point, is described in Gessel (1980).

References

1. Brini, A. (1984). Higher dimensional recursive matrices and diagonal delta sets of series, *Journal of Combinatorial Theory, Series A*, **36**, 315–331.

2. Gessel, I. M. (1980). A factorization for formal Laurent series and lattice path enumeration, *Journal of Combinatorial Theory, Series A*, **28**, 321–337.

3. Jabotinsky, E. (1953). Representation of functions by matrices. Application to Faber polynomials, *Proceedings of the American Mathematical Society*, **4**, 546–553.

4. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

5. Ree, S. (1994). Enumeration of lattice paths and P-partitions, *Ph.D. Thesis*, Brandeis University.

6. Schiffer, M. (1948). Faber polynomials in the theory of univalent functions, *Bulletin of the American Mathematical Society*, **54**, 503–517.

7. Schur, I. (1945). On Faber polynomials, *American Journal of Mathematics*, **67**, 33–41.

2

Lattice Path Enumeration and Umbral Calculus

Heinrich Niederhausen

Florida Atlantic University, Boca Raton, FL

Abstract: The Umbral Calculus is an excellent tool for solving systems of difference equations with given initial values. Many lattice path enumeration problems can be formulated as such systems. Examples are paths underneath a boundary of straight lines, path inside a diagonal band, weighted paths, paths with several step directions, and paths crossing some line a given number of times.

Keywords and phrases: Umbral Calculus, lattice path enumeration, Kolmogorov-Smirnov tests

2.1 Introduction

Twenty years ago, when I saw the "Finite Operator Calculus" [Rota, Kahaner and Odlyzko (1973)] for the first time, I was captivated by its beauty and inspired by all the roads it opened up for further exploration. Sheffer polynomials became the magic tools for my thesis work on Ballot problems and Kolmogorov-Smirnov distributions, and I started to work on some generalizations, like piecewise polynomial Sheffer functions ("Sheffer splines") and multi-indexed Sheffer sequences. However, none of the *generalizations* I have studied were as satisfying to me as the *specializations* that lead to real applications. There is of course a considerable amount of details necessary before we can actually calculate a significance level, say, when we start with the Umbral Calculus.

All the results in this paper have been published earlier, except Theorem 2.5.1 on geometric Sheffer sequences, and perhaps formula (2.8) on counting lattice path with weighted left turns staying *above* a parallel to the diagonal. However, this paper is not intended to be a survey on lattice path problems, but to show how the Umbral Calculus can serve as a tool in certain situations.

2.1.1 Notation

A *polynomial sequence* $\{p_n\}_{n\geq 0}$ is a sequence of polynomials such that $\deg p_n = n$ for all $n = 0, 1, \ldots$.

$$p(x,t) = \sum_{n\geq 0} p_n(x) t^n$$

is the generating function of this polynomial sequence. For convenience, we will henceforth assume that $p_n \equiv 0$ for negative n.

A *delta operator* B is a formal power series of order 1 in the derivative operator D_x,

$$B(D_x) = D_x + b_2 D_x^2 + \ldots$$

A *Sheffer sequence* $\{s_n\}$ (for B) is a polynomial sequence such that

$$Bs_n = s_{n-1}$$

for all $n = 0, 1, \ldots$. The basic sequence $\{b_n\}$ (for B) is the Sheffer sequence with initial values $b_n(0) = \delta_{0,n}$. Basic sequences and Sheffer sequences have generating functions of the form

$$b(x,t) = e^{x\beta(t)}, \quad s(x,t) = s(t) e^{x\beta(t)}$$

where β is the compositional inverse of B, and $s(t) = \sum_{n\geq 0} s_n(0) t^n$ is a formal power series of order 0. A straightforward consequence of those generating functions is the *binomial theorem for Sheffer sequences*,

$$s_n(x+y) = \sum_{i=0}^{n} s_i(x) b_{n-i}(y) \tag{2.1}$$

2.2 Initial Value Problems

In lattice path enumeration, we frequently have to solve the system of difference equations

$$Br_n(x) = r_{n-1}(x)$$

for all $n = 0, 1, \ldots$, where B is a given difference operator and r_0 is a non-zero constant. This implies that $\{r_n\}$ is a Sheffer sequence for B. Finding a solution to this system usually requires expanding $r_n(x)$ in terms of the corresponding basic sequence $\{b_n\}$ such that certain initial values are met (which are set by path boundaries),

$$r_n(x_n) = y_n$$

say, for $n = 0, 1, \ldots$. Such initial values uniquely determine $\{r_n\}$. The binomial theorem for Sheffer sequences in (2.1) can be utilized for such an expansion if we know an initial value *at the same* input for all n, like $r_n(0)$ for example:

$$r_n(x) = \sum_{i=0}^{n} r_j(0) b_{n-j}(x). \tag{2.2}$$

The same theorem can help us now to recursively determine $r_n(0)$ from the given initial values, because

$$y_n = r_n(x_n) = \sum_{j=0}^{n} r_j(0) b_{n-j}(x_n).$$

In other words, we must solve the matrix equation $Y = AR$ for R, where $Y = (y_i)_{i=0\ldots n}$, $R = (r_i(0))_{i=0\ldots n}$, and $A = (b_{i-j}(x_i))_{i=0\ldots n, j=0\ldots i}$ is lower-triangular. Cramer's rule will easily express $r_n(0)$ as a determinant (note that $|A| = 1$), if necessary. However, here we do not consider the well-known determinant as an *explicit* solution, because of its inherent recursive nature. To get $r_n(x)$ we need another triangular matrix, $C = (b_{i-j}(x))_{i=0\ldots n, j=0\ldots i}$, and find

$$(r_i(x))_{i=0\ldots n} = CR = CA^{-1}Y.$$

We will see below that Umbral Calculus can find an *explicit* solution to the initial value problem if the inputs x_n are piecewise affine in n. The size of the initial values y_n is of minor importance; suppose we know a family of Sheffer sequences $\left\{t_n^{(i)}\right\}_{n\geq 0}$ for B with initial values $t_n^{(i)}(x_{n+i}) = \delta_{0,n}$ for all $i = 0, 1, \ldots$. Then it is straightforward to verify that

$$r_n(x) = \sum_{i=0}^{n} y_i t_{n-i}^{(i)}(x) \tag{2.3}$$

solves the original initial value problem.

It can be helpful to have a mental image of the solutions. In the context of initial value problems, I visualize a Sheffer sequence as rows of values:

...	*	*	*	*	$y_4 = 5$	*	*	*	*	...	$s_4(x)$	quartic
...	*	*	*	*	*	*	$y_3 = 3$	*	*	...	$s_3(x)$	cubic
...	$y_2 = 3$	*	*	*	*	*	*	*	*	...	$s_2(x)$	quadratic
...	*	*		*	*	*	*	$y_1 = 2$	*	...	$s_1(x)$	linear
...	1	1	1	1	$y_0 = 1$	1	1	1	1	...	$s_0(x)$	constant
...	tan 2	-2	-1	tan 4	tan 0	1	tan 3	tan 1	2	...	x	

Example: $s_n(\tan n) = n + 1$

An important aspect of this example is that the recurrence need not take place in an integer-lattice. The difference operator and the derivative are both delta operators. In other words, we can simultaneously study lattice paths and empirical distribution functions as in the Kolmogorov two-sample and one-sample tests.

2.2.1 The role of e^x

The Finite Operator Calculus [Rota, Kahaner and Odlyzko (1973)] is based on the reference sequence $\{x^n/n!\}$ and its generating function e^{xt}. The special analytical properties of this generating function stand behind many interesting results in the Finite Operator Calculus.

Suppose $\{b_n\}$ is the basic sequence for the delta operator B, with generating function

$$\sum_{n\geq 0} b_n(x)t^n = e^{x\beta(t)}.$$

It is of great importance for our initial value problem that $D_t e^{x\beta(t)} = x\beta'(t)e^{x\beta(t)}$, because this implies that

$$p_n(x) := \frac{n+1}{x}\, b_{n+1}(x) \tag{2.4}$$

has the generating function $\beta'(t)e^{x\beta(t)}$, and therefore must be a Sheffer polynomial. The linear combination

$$s_n(x) := b_n(x-c) - ap_{n-1}(x-c) = \frac{x-an-c}{x-c}\, b_n(x-c)$$

of Sheffer polynomials is again a Sheffer polynomial for the same operator, and solves the initial value

$$Bs_n = s_{n-1}, \text{ and } s_n(an+c) = \delta_{0,n}$$

for all $n = 0, 1, \ldots$, where a and c are given constants. This solution has already been given in Rota, Kahaner and Odlyzko (1973). In order to solve the problem

$$Br_n = r_{n-1}, \text{ and } r_n(an+c) = y_n$$

for all $n = 0, 1, \ldots$ using the expansion (2.3), we must define $t_n^{(i)}(x) := s_n(x-ai)$ and get

$$r_n(x) = \sum_{i=0}^{n} y_i s_{n-i}(x-ai) = \sum_{i=0}^{n} y_i \frac{x-an-c}{x-ai-c}\, b_{n-i}(x-ai-c). \tag{2.5}$$

2.2.2 Piecewise affine boundaries

Suppose we want to solve the system with initial values first along the line $x_n = an + c$ given by

$$y_n = r_n(an+c)$$

for all $n = 0, \ldots, L-1$, and thereafter on the line $x_n = \tilde{a}n + \tilde{c}$ given by

$$y_n = r_n(\tilde{a}n + \tilde{c})$$

for all $n = L, \ldots,$ where $a, \tilde{a}, c, \tilde{c}$ and L are all given constants. By (2.5), the beginning of the sequence can be calculated along the *second* line as

$$r_i(\tilde{a}i + \tilde{c}) = \sum_{j=0}^{i} y_j \, \frac{(\tilde{a} - a)i + \tilde{c} - c}{\tilde{a}i - aj + \tilde{c} - c} \, b_{i-j}(\tilde{a}i - aj + \tilde{c} - c)$$

for all $i = 0, \ldots, L-1$, and applying (2.5) again gives for $n \geq L$ the expansion

$$r_n(x) = \sum_{i=0}^{L-1} r_i(\tilde{a}i+\tilde{c}) \, \frac{x - \tilde{a}n - \tilde{c}}{x - \tilde{a}i - \tilde{c}} \, b_{n-i}(x-\tilde{a}i-\tilde{c}) + \sum_{i=L}^{n} y_i \, \frac{x - \tilde{a}n - \tilde{c}}{x - \tilde{a}i - \tilde{c}} \, b_{n-i}(x-\tilde{a}i-\tilde{c}). \tag{2.6}$$

Substituting for $r_i(\tilde{a}i + \tilde{c})$ from the beginning of the sequence into the first sum in (2.6) finishes the expansion, but makes it into a double sum.

This procedure can be repeated for initial values on more affine pieces. Obviously, the multiplicity of the summation will grow with the number of pieces. In the example presented in Section 2.2, $D_x s_n(x) = s_{n-1}(x)$ and $s_n(\tan(n)) = n + 1$, and it suffices to use two pieces if we only want to find $s_4(x) = \frac{1}{4!}x^4 + .07\,3765x^3 + .78994x^2 + 4.\,2212x - 1.\,1357$:

$$\begin{aligned} x_n &= (\tan 2 - \tan 1)\, n + 2\tan 1 - \tan 2 && \text{for } n = 0, 1, 2 \\ x_n &= (\tan 4 - \tan 3)\, n + 4\tan 3 - 3\tan 4 && \text{for } n \geq 3. \end{aligned}$$

2.2.3 Applications: Bounded paths

Some of the best known applications occur in the enumeration of lattice paths, sequences of horizontal $\rightarrow$ and vertical $\uparrow$ steps starting at the origin. Let $r_n(m)$ be the number of paths that reach the point (m, n) under some kind of restriction following the recurrence

$$r_n(m) = r_n(m-1) + r_{n-1}(m).$$

The (generalized) ballot problem requires the paths to remain below some boundary line; this translates into initial conditions of the form $r_n(-1) = \delta_{0,n}$ for all $n = 0, \ldots, L-1$, and $r_n(an + c) = 0$ for all $n = L, \ldots$.

...	*	*	*	*	*	*	*	*	$\boxed{0}$	409	1034	...	$r_5(x)$
...	*	*	*	*	*	$\boxed{0}$	52	132	248	409	625	...	$r_4(x)$
...	*	*	$\boxed{0}$	6	$\underrightarrow{16}$	$\underrightarrow{31}$	$\underrightarrow{52}$	$\underrightarrow{80}$	$\underrightarrow{116}$	$\underrightarrow{161}$	$\underrightarrow{216}$	...	$r_L(x)$
...	$\boxed{0}$	$\underrightarrow{1}$	$\underrightarrow{3}$	$\underrightarrow{6}$	$\uparrow 10$	15	21	28	36	45	55	...	$r_2(x)$
...	$\boxed{0}$	$\uparrow 1$	2	3	4	5	6	7	8	9	10	...	$r_1(x)$
...	1	$\uparrow 1$	1	1	1	1	1	1	1	1	1	...	$r_0(x)$
...	−1	0	1	2	3	4	5	6	7	8	9	...	x

Path boundary $3n - 8$, with sample path

Such initial value problems must be solved for calculating the exact distribution of the (one-sided) two-sample Kolmogorov-Smirnov test. The exact distribution of the *one*-sample Kolmogorov-Smirnov test derives from the Lebesque measure of certain empirical distribution functions instead of the counting measure of lattice paths; the same techniques apply, however the derivative operator D_x takes the place of the backwards difference operator. Sheffer sequences are also employed for the exact distribution of some multivariate generalizations. The two-sided distribution of the two-sample test has a "closed form"; why that is not the case for the *one*-sample case is explained in Section 2.4.1. Details about these applications can be found in Niederhausen (1979). A general reference in this direction is Mohanty (1979); see also Mohanty (1968).

2.3 Systems of Operator Equations

In the last section, we discussed the rather simple system $Br_n = r_{n-1}$, for $n = 0, 1, \ldots$ with relatively general initial conditions. In this section we concentrate on finding basic solutions b_n, where $b_n(0) = \delta_{0,n}$, of more complicated systems of the form

$$Qb_n = Rb_{n-1} + Sb_{n-c} \tag{2.7}$$

for all $n = 0, 1, \ldots$, where c is a positive integer, R and S are translation invariant invertible operators (i.e., power series of order 0 in D_x), and Q is a delta operator. An example for R could be $Rb_n(x) = \sum_{i=1}^{r} \rho_i b_n(x - r_i)$ for some given constants $\rho_1, \ldots, \rho_k$ and $r_1, \ldots, r_k$. By a *solution* of (2.7), we mean an expansion of $b_n(x)$ in terms of the basic sequence $\{q_n\}$ of Q.

Suppose the unknown solution $\{b_n\}$ is the basic sequence for some delta operator B. If we can construct a solution under this hypothesis, then the assumption will be justified. Because B and Q are both delta operators, there exists a translation invariant and invertible operator T such that $B = TQ$ [see Corollary 4 of Rota, Kahaner and Odlyzko (1973)]. Any such T which solves the equation

$$I = RT + ST^cQ^{c-1}$$

also solves the equation

$$Q = RTQ + ST^cQ^c$$

which is equivalent to the system (2.7) (any two translation invariant operators commute). Equivalently, $I - RT = (RT)^c R^{-c} SQ^{c-1}$ and $(RT)^{-c} - (RT)^{1-c} = R^{-c}SQ^{c-1}$. Lagrange inversion then gives

$$T^{-n} = \sum_{k \geq 0} \binom{n - k(c-1)}{k} \frac{n}{n - (c-1)k} R^{n-kc} B^k Q^{k(c-1)}$$

and the Transfer Formula [Section 4 of Rota, Kahaner and Odlyzko (1973)] gives

$$b_n(x) = xT^{-n}x^{-1}q_n(x) = \sum_{k\geq 0} \binom{n-k(c-1)}{k} R^{n-kc}B^k q_{n-k(c-1)}(x)$$

[we need to use that $\left\{\frac{n+1}{x}q_{n+1}(x)\right\}$ is a Sheffer sequence; see (2.4)]. It is now easy to verify that this basic sequence really solves (2.7).

2.3.1 Applications: Lattice paths with several step directions

Counting (weighted) lattice paths with several step directions leads to recurrence relations of the form

$$d_n(x) = d_n(x-1) + \sum_{i=1}^{r} \rho_i b_{n-1}(x-r_i) + \sum_{i=1}^{s} \sigma_i b_{n-c}(x-s_i)$$

if the step vectors are $(1,0), (r_1,1), \ldots, (r_r,1), (s_1,c), \ldots, (s_s,c)$. In this case, $Q = \nabla$, the backwards difference operator. More details about the simple case $r = 1 = s$ are given in Niederhausen (1979). An application to a gamblers ruin problem and expected game duration can be found in Niederhausen (1986).

2.4 Symmetric Sheffer Sequences

In Section 2.2.3, we mentioned the general ballot problem as an application of formula (2.6). In the classical ballot problem, the paths stay below the diagonal, or some line parallel to the diagonal. The initial values are, therefore, $r_n(-1) = 0$ for all $n = 1, \ldots, L-1$, and $r_n(n-L) = 0$ for all $n \geq L$.

...	*	*	*	0	28	90	207	...	$r_5(m)$
...	*	*	0	9	28	62	117	...	$r_4(m)$
...	*	0	3	9	$\underrightarrow{19}$	$\underrightarrow{34}$	$\underrightarrow{55}$	...	$r_L(m)$
...	0	$\underrightarrow{1}$	$\underrightarrow{3}$	$\underrightarrow{6}$	$\uparrow 10$	15	21	...	$r_2(m)$
...	0	$\uparrow 1$	2	3	4	5	6	...	$r_1(m)$
...	1	$\uparrow 1$	1	1	1	1	1	...	$r_0(m)$
...	-1	0	1	2	3	4	5	...	m

The ballot problem ($L = 3$), with sample path

However, the solution to this initial value problem is much simpler than the sum in formula (2.6) indicates. It is given by

$$r_n(m) = \binom{n+m}{n} - \binom{n+m}{n-L}.$$

This well-known solution is easily identified as a difference of two Sheffer polynomials for the backwards difference operator ∇, and it is obviously zero at $m = -1$ for all $n = 1, \ldots, L-1$. However, for $n \geq L$ the initial values are attained because of a very special property:

$$r_n(n-L) = \binom{n+n-L}{n} - \binom{n+n-L}{n-L} = \binom{n+n-L}{n} - \binom{n+n-L}{n}.$$

In other words, for nonnegative integers n and m, we can interchange the degree n with the argument m in the polynomial $s_n(m) := \binom{n+m}{n}$, and get again a polynomial $s_m(n) = \binom{m+n}{m} = s_n(m)$. We call such a Sheffer sequence *symmetric*—obviously, all symmetric Sheffer sequences can be used to construct the very simple solution $s_n(m) - s_{n-L}(L+m)$ to the above boundary problem. But are there any other symmetric Sheffer sequences besides $\binom{n+x}{n}$? Niederhausen (1996) has shown that, except for a scaling factor, there is only one parameter that describes the whole class of symmetric Sheffer sequences.

Theorem 2.4.1 *All symmetric Sheffer sequences are of the form* $\left\{\alpha s_n^{(\mu)}(x)\right\}_{n \geq 0}$, *where* α *is a nonzero scaling factor, and*

$$s_n^{(\mu)}(x) = \sum_{l=0}^{n} \binom{n}{l} \binom{x}{l} \mu^l$$

($\mu \neq 0$). The corresponding delta operator $\Omega^{(\mu)}$ has the expansion

$$\Omega^{(\mu)} = \frac{\Delta}{\mu + \Delta}$$

in terms of the forward difference operator Δ.

2.4.1 Applications: Weighted left turns

If $\mu = 1$, we obtain $s_n^{(1)}(x) = \binom{n+x}{n}$ and $\Omega^{(1)} = \nabla$, the backwards difference operator. In general, $s_n^{(\mu)}(m)$ equals the weighted sum of lattice paths from $(0,0)$ to (m,n), where every left turn $\rightarrow \overset{\uparrow}{\circ}$ gets the weight μ. Because of symmetry, the classical ballot problem has a simple solution for this kind of weighted paths:

$$s_n^{(\mu)}(m) - s_{n-L}^{(\mu)}(m+L)$$

which is the weighted sum of lattice paths from $(0,0)$ to (m,n) that stay below the line $m = n - L$. Switching to paths above the line $m = n + K$ is no longer an equivalent problem (except if we also switch from weighted left turns to weighted right turns.)

1	$1+3\mu$	$1+5\mu+3\mu^2$	$1+6\mu+6\mu^2+\mu^3$	$1+6\mu+6\mu^2+\mu^3$	$0\ldots$	$w_3(m)$
1	$1+2\mu$	$1+3\mu+\mu^2$	$1+3\mu+\mu^2$	0		$w_2(m)$
1	$1+\mu$	$1+\mu$	0			$w_1(m)$
1	1	0				$w_0(m)$
0	1	K	3	4	$5\ldots$	m

Counts of paths with weighted left turns above $m = n + 2$

However, it is easy to verify that the Sheffer polynomial

$$w_n(m) := s_n^{(\mu)}(m) - \sum_{i\geq 1} \mu^i \binom{m-K+2}{i+1}\binom{n+K-2}{i-1} \tag{2.8}$$

solves the problem for $m = 0, \ldots, n+K-1$, because it satisfies the necessary and sufficient condition $w_n(n+K-1) = w_n(n+K-2)$. Note that $w_n(n+K) \neq 0$.

More details and further references are present in Niederhausen (1996) and Sulanke (1993). Related topics are correlated random walks, and nonintersecting pairs of weighted lattice paths. The q-binomial coefficients are obviously also symmetric. How to use transforms of operators [Freeman (1985)] to count lattice paths with q-weighted left turns is explained in Niederhausen (1994).

2.4.2 Paths inside a band

The exact distribution of the two-sided two-sample Kolmogorov-Smirnov test requires counting the number of lattice paths inside a band parallel to the diagonal. This number can be described by piecewise polynomial functions. The initial conditions are $t_n(-1) = 0$ for $n = 0, \ldots, L-1$, $t_n(n-L) = 0$ for all $n \geq L$, and $t_n(n+K) = 0$ for all n.

...	$*$	$*$	$*$	0	21	55	89	...	$t_5(m)$
...	$*$	$*$	0	8	21	$\uparrow 34$	34	...	$t_4(m)$
...	$*$	0	3	8	$\underrightarrow{13}$	$\uparrow 13$	0	...	$t_L(m)$
...	0	$\underrightarrow{1}$	$\underrightarrow{3}$	$\underrightarrow{5}$	$\uparrow 5$	0			$t_2(m)$
...	0	$\uparrow 1$	2	2	0				$t_1(m)$
...	1	$\uparrow 1$	1	0					$t_0(m)$
...	-1	0	1	K	3	4	5	...	m

Paths inside a band ($L = 3$, $K = 2$), with sample path

Symmetry of the polynomials $s_n^{(1)}(x) = \binom{n+x}{n}$ is the reason why a (relatively simple) expansion of this function exists (in our view). We want to recall this expansion, because it is so often omitted in the literature. We saw that $r_n(m) = s_n^{(1)}(m) - s_{n-L}^{(1)}(m+L)$ is the number of lattice paths below the line

$m = n - L$, and reaching (m, n); a ballot number. A sum gives the number of such paths that also stay above $m = n + K$:

$$t_n(m) = \sum_{i \geq 0} (r_{n-i(K+L)}(m + i(K+L)) - r_{m-i(K+L)-K}(n + i(K+L) + K))$$

It is amazing that $t_n(m)$ satisfies the recurrence, and both types of boundary values. The telescopic nature of this sum becomes essential if we verify the condition $t_n(n-L) = 0$ for $n \geq L$. For $K = L$, the formula was derived by Koroljuk (1955). See Fray and Roselle (1971) for another proof of the general case.

In the two-sided one-sample case, the distribution must be expanded in terms of $x^n/n!$. Because of its lack of symmetry, no closed form is known.

2.5 Geometric Sheffer Sequences

A Sheffer sequence $\{s_n\}_{n \geq 0}$ is *geometric* if $s_0 \equiv 1$ and if there exists a pair of constants a and $\hat{a}$ such that

$$s_n(an) = \hat{a} s_{n-1}(an)$$

for all $n = 1, 2, \ldots$.

...	5	15	35	2×35	126	...	$s_4(x)$
...	4	10	2×10	35	56	...	$s_3(x)$
...	3	2×3	10	15	21	...	$s_2(x)$
...	2×1	3	4	5	6	...	$s_1(x)$
...	1	1	1	1	1	...	$s_0(x)$
...	1	2	3	4	5	...	x

Example for a geometric Sheffer sequence $(a = 1, \hat{a} = 2)$

There exists a geometric Sheffer sequence for any delta operator B and for any pair $a, \hat{a}$ (with $\hat{a} \neq 0$) , because this initial value problem always has a solution; it can be recursively calculated from the expansion (2.5). The following theorem explains why they are called "geometric".

Theorem 2.5.1 *The Sheffer sequence* $\{s_n\}_{n \geq 0}$ *with generating function* $\sum_{n \geq 0} s_n(x) t^n = s(t) e^{x\beta(t)}$ *is geometric iff* $\hat{a}^n$ *is the coefficient of* t^n *in the expansion of the formal power series* $\sum_{n \geq 0} s_n(an) t^n e^{-n\beta(t)}$.

PROOF. $E^{-a} = e^{-aD_x}$ denotes the translation operator by $-a$,

$$E^{-a} p(x) = p(x - a)$$

for any polynomial $p(x)$. $\{s_n(an+x)\}_{n\geq 0}$ is a Sheffer sequence for the delta operator BE^{-a}. Denote the compositional inverse of $B(t)e^{-at}$ by $\alpha(t)$. $\{s_n\}$ is geometric iff

$$\begin{aligned}\sum_{n\geq 0} s_n(an)t^n &= 1+\sum_{n\geq 1}\hat{a}s_{n-1}(an)t^n = 1+t\hat{a}\sum_{n\geq 0}s_n(an+a)t^n \\ &= 1+t\hat{a}\left(\sum_{n\geq 0}s_n(an)t^n\right)e^{a\alpha(t)}.\end{aligned}$$

Solving for $\sum_{n\geq 0}s_n(an)t^n$ gives $\sum_{n\geq 0}s_n(an)t^n = 1/\left(1-\hat{a}te^{a\alpha(t)}\right)$. Substitute $B(t)e^{-at}$ for t in order to get

$$\sum_{n\geq 0}s_n(an)\left(B(t)e^{-at}\right)^n = \frac{1}{1-\hat{a}B(t)}\,. \tag{2.9}$$

Finally, substituting $\beta(t)$ for t yields

$$\sum_{n\geq 0}s_n(an)t^n e^{-na\beta(t)} = \frac{1}{1-\hat{a}t}\,. \tag{2.10}$$

■

The above identities are special cases of Lagrange inversion [see Pólya and Szegö (1972)]. We check some examples:

1. $\{x^n/n!\}$ is a geometric Sheffer sequence for D_x with $a=\hat{a}$, and $\beta(t)=t=B(t)$. Both identities give $\sum_{n\geq 0}(nte^{-t})^n = 1/(1-t)$ [Pólya and Szegö (1972), Part III, Problem 214].
2. $\binom{n+x}{n}$ is a geometric Sheffer sequence for $\nabla = 1-E^{-1}$ with $a+1=\hat{a}$, and $\beta(t)=-\ln(1-t)$. Identity (2.10): $\sum_{n\geq 0}\binom{n+an}{n}t^n(1-t)^{an} = 1/\left(1-(a+1)t\right)$ [Pólya and Szegö (1972), Part III, special case of Problem 216].
3. $\left\{\frac{x-an+1}{x+1}\binom{n+x}{n}\right\}$ is a geometric Sheffer sequence for ∇ with $\hat{a}=1$. Identity (2.9): $\sum_{n\geq 0}\frac{1}{an+1}\binom{n+an}{n}(e^{-at}-e^{-(a+1)t})^n = e^t$ [Pólya and Szegö (1972), Part III, Problem 211].

2.5.1 Applications: Crossings

Denote by $D(n,m;l)$ the number of (restricted) lattice paths from $(0,0)$ to (m,n) with steps $\rightarrow$ and $\uparrow$ that go through at least l of some given nodes in the plane. It is usually not difficult to calculate $D(n,m;l)$ recursively. Closed form expressions are known if the nodes lie on a line, $(n,an+c)$ for $n>\bar{n}$, where a,c and $\bar{n}$ are given constants. Additional restrictions may be imposed on the path; for example, the path may be required

- to cross through the node coming from below,
  - and leave in a vertical step,
- to stay above a line (or some line segments),
- to walk in a higher dimensional lattice.

The surprisingly "simple" closed forms for $D(n,m;l)$ occur when $D(n,m;0)$ can be expressed by a *geometric* Sheffer polynomial $s_n(m)$. Only in this case we get for paths terminating on a node the recurrence relation

$$\begin{aligned} D(n,an+c;l) &= \sum_{i=\tilde{n}+l-1}^{n-1} \left(D(ai+c,i;l-1) - D(ai+c,i;l)\right) s_{n-i}(a(n-i)) \\ &= \hat{a} D(n-1,an+c;l-1) \end{aligned}$$

which is essential for further simplifications in calculating $D(n,m;l)$, $m \geq an+c$. In statistical inference, tests based on the number of crossings are called Takács tests [Takács (1971b)]. The above method applies to the one-sample Takács distribution [Takács (1971a)] as well (empirical distribution functions instead of lattice paths), because $\{x^n/n!\}$ is geometric too. More details are given in Mohanty (1967, 1968, 1979) and Niederhausen (1982).

References

1. Fray, R. D. and Roselle, D. P. (1971). Weighted lattice paths, *Pacific Journal of Mathematics*, **37**, 85–96.

2. Freeman, J. M. (1985). Transforms of operators on $K[x][[t]]$, *Congressus Numerantium*, **48**, 115–132.

3. Koroljuk, V. S. (1955). On the discrepancy of empiric distributions for the case of two independent samples, *Izv. Akad. Nauk SSSR Ser. Mat.*, **19**, 81–96 [*IMS & AMS Selected Translations in Mathematical Statistics and Probability*, **4** (1963), 105–122].

4. Mohanty, S. G. (1966). On a generalized two-coin tossing problem, *Biometrische Zeitschrift*, **8**, 266–272.

5. Mohanty, S. G. (1967). Restricted compositions, *Fibonacci Quarterly*, **5**, 223–234.

6. Mohanty, S. G. (1968). On some generalization of a restricted random walk, *Studia Scientiarum Mathematicarum Hungarica*, **3**, 225–241.

7. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

8. Niederhausen, H. (1979). Lattice paths with three step directions, *Congressus Numerantium*, **14**, 753–774.

9. Niederhausen, H. (1981). Sheffer polynomials for computing exact Kolmogorov–Smirnov and Rényi type distributions, *Annals of Statistics*, **9**, 923–944.

10. Niederhausen, H. (1982). How many paths cross at least l given lattice points? *Congressus Numerantium*, **36**, 161–173.

11. Niederhausen, H. (1986). The enumeration of restricted random walks by Sheffer polynomials with applications to statistics, *Journal of Statistical Planning and Inference*, **14**, 95–114.

12. Niederhausen, H. (1992). Fast Lagrange inversion, with an application to factorial numbers, *Discrete Mathematics*, **104**, 99–110.

13. Niederhausen, H. (1994). Counting intersecting weighted pairs of lattice paths using transforms of operators, *Congressus Numerantium*, **102**, 161–173.

14. Niederhausen, H. (1996). Symmetric Sheffer sequences and their applications to lattice path counting, *Journal of Statistical Planning and Inference*, **54**, 87–100.

15. Pólya, G. and Szegö, G. (1972). *Problems and Theorems in Analysis–I*, New York: Springer-Verlag.

16. Rota, G.-C., Kahaner, D. and Odlyzko, A. (1973). On the foundations of combinatorial theory, VIII. Finite operator calculus, *Journal Mathematical Analysis and Applications*, **42**, 684–760.

17. Sulanke, R. A. (1993). Refinements of the Narayana numbers, *Bulletin of the ICA*, **7**, 60–66.

18. Takács, L. (1971a). On the comparison of a theoretical and an empirical distribution function, *Journal of Applied Probability*, **8**, 321–330.

19. Takács, L. (1971b). On the comparison of two empirical distribution functions, *Annals of Mathematical Statistics*, **42**, 1157–1166.

3

The Enumeration of Lattice Paths With Respect to Their Number of Turns

C. Krattenthaler

Institut für Mathematik der Universität Wien, Vienna, Austria

Abstract: We survey old and new results on the enumeration of lattice paths in the plane with a given number of turns, including the recent developments on the enumeration of nonintersecting lattice paths with a given number of turns. Motivations to consider such enumeration problems come from various fields, e.g. probability, statistics, combinatorics, and commutative algebra. We show that the appropriate tool for treating turn enumeration of lattice paths is the encoding of lattice paths in terms of two-rowed arrays.

Keywords and phrases: Turns, lattice paths, nonintersecting lattice paths, coin tossing, run statistics, non-crossing two-rowed arrays, determinantal rings, pfaffian rings, Hilbert series, tableaux, plane partitions

3.1 Introduction

In this article we consider lattice paths in the plane consisting of unit horizontal and vertical steps in the positive direction. We will be concerned with enumerating such lattice paths which have a given number of *turns*. By a *turn*, we mean a vertex of a path where the direction of the path changes. For example, the turns of the path P_0 in Figure 3.1 are $(1,1)$, $(2,1)$, $(2,3)$, $(5,3)$, $(5,4)$, and $(6,4)$. Distinguishing between the two possible types of turns, we call a vertex of a path a *North-East turn* (*NE-turn*, for short) if it is the end point of a vertical step and at the same time the starting point of a horizontal step, and we call a vertex of a path an *East-North turn* (*EN-turn*, for short) if it is a point in a path P which is the end point of a horizontal step and at the same time the starting point of a vertical step. The NE-turns of the path in Figure 3.1 are $(1,1)$, $(2,3)$, and $(5,4)$, and the EN-turns of the path in Figure 3.1 are

$(2, 1)$, $(5, 3)$, and $(6, 4)$.

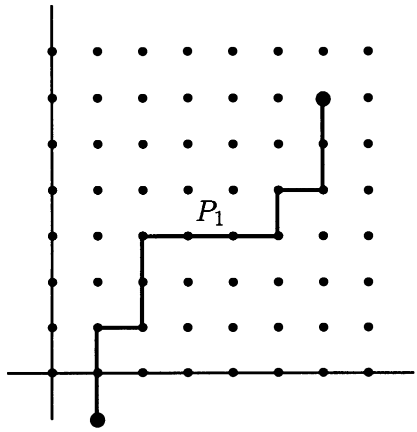


Figure 3.1

There are various motivations to be interested in the turn enumeration of lattice paths. We describe three such motivations, from probability, statistics, and commutative algebra, respectively, in more detail, in Section 3.3. The examples from probability and statistics (correlated random walk, run and Kolmogorov-Smirnov statistics) in Section 3.3 lead to the enumeration of paths, with given starting and end points, with a given number of turns, which are bounded by lines. This is classical today. The example from commutative algebra (Hilbert series of determinantal and pfaffian rings) however leads to the enumeration of *families of nonintersecting lattice paths*, with given starting and end points, with a given number of turns, and subject to certain restrictions. Interest in this subject arose only recently, mainly due to the path-breaking work of Abhyankar (1987, 1988). A number of remarkable formulas were discovered to solve most of these problems. But there are still some important open questions.

The problem of turn enumeration of lattice paths was attacked in many different ways. However, there is a uniform approach which is able to handle all these problems, which is by encoding paths in terms of *two-rowed arrays*. Actually, this is the way in which Narayana (1959, 1979, Section II.2), who probably was the first to count paths with respect to their turns, used to see turn enumeration problems. However, he did not use the combinatorics of two-rowed arrays. His proofs are manipulatory and usually work by induction. The purpose of this survey article is to show that two-rowed arrays allow to handle turn enumeration in a purely combinatorial way. The combinatorics of two-rowed arrays is able to explain all the existing formulas in a conceptual way. What is very appealing is that all the standard techniques from ordinary path counting, such as reflection principle, iterated reflection principle, interchanging procedure for nonintersecting lattice paths, have their analogues in the "world of two-rowed arrays."

Another purpose of this survey is to show the wide diversity of connections and applications in other fields like combinatorics, representation theory, and q-series. Moreover, it is not unreasonable to expect that the recent subject of turn enumeration of nonintersecting lattice paths will also have its applications in probability, statistics, or physics. Evidence for this feeling comes from the fact that turn enumeration of (single) lattice paths is of importance in these fields, and (plain) enumeration of nonintersecting lattice paths is too [see, for example, Essam and Guttmann (1995), Fisher (1984), Karlin (1988) and Karlin and McGregor (1959a,b)].

This exposition brings together ideas from several papers of this author and Mohanty [see, for example, Krattenthaler (1989, 1993, 1995a, 1995b, 1996a) and Krattenthaler and Mohanty (1993)] . The proof of Theorem 3.4.2 is new.

The paper is organized in the following way. In the next section, we introduce some basic notations which we use throughout the paper. Section 3.3 contains the announced motivating examples. In Section 3.4, we address the turn enumeration of (single) lattice paths. The results of Section 3.4 are then applied in Section 3.5 to solve some of the problems in the mentioned examples. Finally, Section 3.6 is devoted to turn enumeration of nonintersecting lattice paths. The results of this section answer most of the problems of the third example in Section 3.3. Open problems are listed at the end of Section 3.6.

3.2 Notation

Given two lattice points A and E, we denote the set of all lattice paths from A to E by $L(A \to E)$. If P is a path from A to E, we will symbolize this sometimes by $P : A \to E$. If R is some property of paths, we use the "probability-like" notation $L(A \to E \mid R)$ for the set of all paths from A to E satisfying property R.

3.3 Motivating Examples

Example 3.3.1 A TWO COIN TOSSING GAME; CORRELATED RANDOM WALK. Mohanty (1966) considered the following game. Take two coins 1 and 2 with probabilities p_1 and p_2 of obtaining heads, respectively. The rules for the game are:

1. start with coin i, $i = 1, 2$;

2. if the last trial was a tail, then make the next trial with coin 1, otherwise with coin 2;

3. stop making further trials when for the first time the total number of heads exceeds μ times the total number of tails by exactly a, with a fixed $a > 0$.

The question is: Provided the game was started by tossing coin i, $i = 1$ or 2, what is the distribution of the duration of the game?

This game has also an equivalent formulation in terms of a "correlated" random walk; see, for example, Mohanty (1979, Section 5.2). In sampling plan terminology [DeGroot (1959)], these games describe sequential sampling plans for binomial populations with $y = \mu x + a$ as the boundary line.

It is an easy observation that any game can be represented in terms of a lattice path, by starting in $(0, 0)$ and proceeding by a horizontal step if tail (T) was tossed and by a vertical step if head (H) was tossed. Thus, the game $THHHTHTHHHH$ (which is a game for $\mu = 2$ and $a = 2$) would be represented by the lattice path P_2 in Figure 3.2. The condition (3) is reflected by the fact that any such lattice path, except for the final vertical step, stays below the line $y = \mu x + a - 1$ (being allowed to touch it).

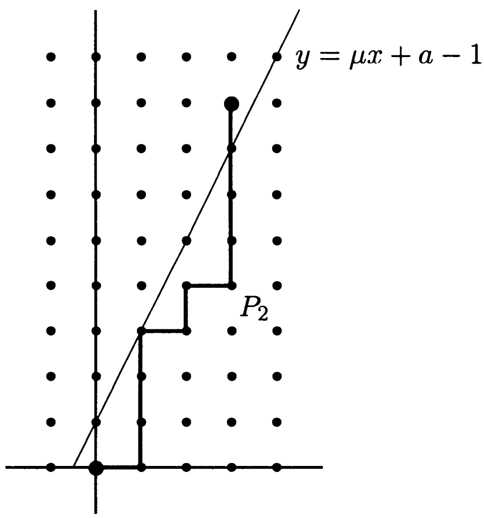


Figure 3.2

The probability of a game of length $(\mu + 1)n + a$ (n tails and $\mu n + a$ heads) is given as follows. If the first toss was with coin 1, then the probability of a game, corresponding to a path P as described above, is

$$p_1^{\mathrm{NE}(P)+1}(1 - p_1)^{n-\mathrm{NE}(P)} p_2^{\mu n+a-\mathrm{NE}(P)-1}(1 - p_2)^{\mathrm{NE}(P)}, \tag{3.1}$$

where $\mathrm{NE}(P)$ denotes the number of NE-turns of P. On the other hand, if the first toss was with coin 2, then the probability of a game, corresponding to path P, is

$$p_1^{\mathrm{NE}(P)+1}(1 - p_1)^{n-\mathrm{NE}(P)-1} p_2^{\mu n+a-\mathrm{NE}(P)-1}(1 - p_2)^{\mathrm{NE}(P)+1}, \tag{3.2}$$

if the first toss resulted in tail, and

$$p_1^{\mathrm{NE}(P)}(1-p_1)^{n-\mathrm{NE}(P)}p_2^{\mu n+a-\mathrm{NE}(P)}(1-p_2)^{\mathrm{NE}(P)}, \tag{3.3}$$

if the first toss resulted in head, respectively.

Therefore, to determine the probability of games of length $(\mu+1)n+a$, we need to enumerate lattice paths from $(0,0)$ to $(n, \mu n+a-1)$ staying below the line $y = \mu x + a - 1$, being allowed to touch it, which have a given number of NE-turns.

Example 3.3.2 RUNS AND KOLMOGOROV-SMIRNOV STATISTICS. Two common rank order statistics for nonparametric testing problems in the two-sample case are the *run statistics* and the (one- and two-sided) *Kolmogorov-Smirnov statistics*. We consider just the case of equal sample size. Recall [see, for example, Mohanty (1979, Section 4.3)] that there are two sets of independent and identically distributed random variables $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$ and $\mathcal{Y} = \{Y_1, Y_2, \dots, Y_n\}$ of size n. These are then put together and ordered into $\mathcal{Z} = (Z_1, Z_2, \dots, Z_{2n})$ according to size. The *run statistics* counts the number of maximal consecutive subsequences in $\mathcal{Z}$ the members of which belong to just one of the sets $\mathcal{X}$ or $\mathcal{Y}$. Thus, if $n = 5$, and if $\mathcal{Z} = (X_1, Y_1, Y_2, Y_3, X_2, X_3, Y_4, X_4, X_5, Y_5)$, then the number of runs in $\mathcal{Z}$ is 6. The *one-sided Kolmogorov-Smirnov statistic* $D^+_{n,n}$ is defined by

$$D^+_{n,n} = \frac{1}{n}\max_i\{a_i - b_i\},$$

where a_i is the number of occurrences of X_j's in the initial segment $Z_1, Z_2, \dots, Z_i$ of $\mathcal{Z}$, while b_i is the number of occurrences of Y_j's in this initial segment. The *two-sided Kolmogorov-Smirnov statistic* $D_{n,n}$ is defined by

$$D_{n,n} = \frac{1}{n}\max_i\{|a_i - b_i|\}.$$

Thus, we have for our combined sample $\mathcal{Z}$ that $D^+_{5,5} = 1/5$ and $D_{5,5} = 2/5$.

Each such sequence $\mathcal{Z}$ can be represented by a lattice path in the obvious way. Namely, start at $(0,0)$, then read through the sequence from left to right and proceed by a vertical step if some X_j is encountered and by a horizontal step if some Y_j is encountered. Thus, the above set $\mathcal{Z}$ corresponds to the lattice path P_3 in Figure 3.3. The run statistics obviously translates into the number of maximal horizontal and vertical pieces in the corresponding path. The one-sided Kolmogorov-Smirnov statistic is basically the maximal deviation from the main diagonal in direction $(1,-1)$. The two-sided Kolmogorov-Smirnov statistic is basically the maximal deviation from the main diagonal, in either direction. So in Figure 3.4, paths which stay in the region between the indicated lines $y = x+2$ and $y = x-2$ correspond to sequences $\mathcal{Z}$ with two-sided Kolmogorov-Smirnov statistic $D_{n,n} \le 2/5$.

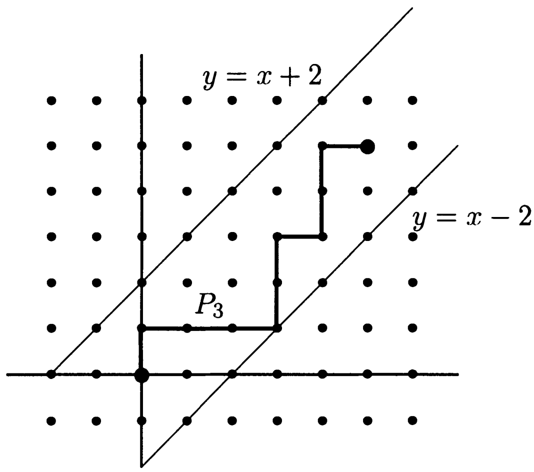


Figure 3.3

Since the number of runs of a lattice path equals 1 plus the number of turns of the path, we see that to determine the distribution of the run statistics we need to count lattice paths from $(0,0)$ to (n,n) with a given number of turns (both, NE- and EN-turns). If, in addition, we want to know the joint distribution of runs and the Kolmogorov-Smirnov statistic, then we have to count paths from $(0,0)$ to (n,n) with a given number of turns which in addition stay below a line $y = x + t$ for the one-sided Kolmogorov-Smirnov statistic and between lines $y = x + t$ and $y = x - t$ for the two-sided Kolmogorov-Smirnov statistic.

Example 3.3.3 DETERMINANTAL RINGS. Determinantal rings are frequently studied objects in commutative algebra and algebraic geometry. We start with the classical case. Let $X = (X_{i,j})_{0 \le i \le b,\ 0 \le j \le a}$ be a $(b+1) \times (a+1)$ matrix of indeterminates. Let $K[X]$ denote the ring of all polynomials over some field K in the $X_{i,j}$'s, $0 \le i \le b$, $0 \le j \le a$, and let $I_{n+1}(X)$ be the ideal in $K[X]$ that is generated by all $(n+1) \times (n+1)$ minors of X. The ideal $I_{n+1}(X)$ is called a *determinantal ideal.* The associated *determinantal ring* is $R_{n+1}(X) := K[X]/I_{n+1}(X)$. This is a graded ring. The obvious question to ask is what the dimensions of the homogeneous components $R_{n+1}(X)_\ell$ of dimension ℓ, $\ell = 0, 1, \dots$, of $R_{n+1}(X)$ are. This information is recorded in terms of the Hilbert series of $R_{n+1}(X)$, which is simply the generating function $\sum_{\ell=0}^{\infty} \dim_K (R_{n+1}(X)_\ell)\, z^\ell$. It was shown in several ways [Abhyankar (1988), Abhyankar and Kulkarni (1989), Conca and Herzog (1994), Kulkarni (1996), Modak (1992) and also Ghorpade (1996)] that this problem relates to counting lattice paths with respect to turns, more precisely, to counting *families of nonintersecting lattice paths* with respect to turns. A family $(P_1, P_2, \dots, P_n)$ of paths P_i, $i = 1, 2, \dots, n$, is called *nonintersecting* if no two paths in the family have a point in common, otherwise it is called *intersecting.*

Theorem 3.3.1 *Let $A_i = (0, n-i)$ and $E_i = (a-n+i, b)$, $i = 1, 2, \dots, n$. Then, the Hilbert series of the determinantal ring $R_{n+1}(X) = K[X]/I_{n+1}(X)$ equals*

$$\sum_{\ell=0}^{\infty} \dim_K (R_{n+1}(X)_\ell)\, z^\ell = \frac{\sum_{\mathbf{P}} z^{\mathrm{NE}(\mathbf{P})}}{(1-z)^{(a+b+1)n-2\binom{n}{2}}}, \tag{3.4}$$

where the sum on the right-hand side is over all families $\mathbf{P} = (P_1, P_2, \dots, P_n)$ of nonintersecting lattice paths, with P_i running from A_i to E_i, $i = 1, 2, \dots, n$. Here, the number $\mathrm{NE}(\mathbf{P})$ is defined to be the total number $\sum_{i=1}^{n} \mathrm{NE}(P_i)$ of NE-turns of the family $\mathbf{P}$.

Figure 3.4 contains an example of such a family of nonintersecting lattice paths for $a = 13$, $b = 15$, and $n = 4$. The NE-turns are marked by bold dots.

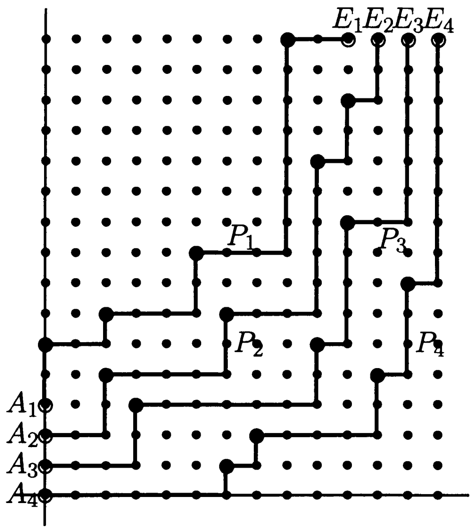


Figure 3.4

Several generalizations of this concept have also been considered. These pose even more difficult turn enumeration problems. We describe just one such generalization in detail. Let $\mathbf{a} = (a_1, a_2, \dots, a_n)$ and $\mathbf{b} = (b_1, b_2, \dots, b_2)$ be two vectors of nonnegative integers which are in strictly increasing order. Let $I_{n+1}^{\mathbf{a},\mathbf{b}}(X)$ denote the ideal in $K[X]$ that is generated by all $t \times t$ minors of the restriction of X to rows $0, 1, \dots, a_t - 1$ and columns $0, 1, \dots, b_t - 1$, $t = 1, 2, \dots, n$, and by all $(n+1) \times (n+1)$ minors of X. What we considered before is the special case $\mathbf{a} = (0, 1, \dots, n-1)$ and $\mathbf{b} = (0, 1, \dots, n-1)$. Again, the associated *determinantal ring* is $R_{n+1}^{\mathbf{a},\mathbf{b}}(X) := K[X]/I_{n+1}^{\mathbf{a},\mathbf{b}}(X)$. For more information on these rings, see Herzog and Trung (1992) and the references therein. In the papers by Abhyankar (1988), Abhyankar and Kulkarni (1989), Conca and Herzog (1994), and Kulkarni (1996), it is shown that this relates to counting lattice paths with respect to turns in much the same way. The difference is that the starting and end points of the lattice paths now depend on the vectors **a** and **b**, respectively.

Theorem 3.3.2 *Let $A_i = (0, a_{n-i+1})$ and $E_i = (a - b_{n-i+1}, b)$, $i = 1, 2, \dots, n$. Then, the Hilbert series of the determinantal ring $R^{\mathbf{a},\mathbf{b}}_{n+1}(X) = K[X]/I^{\mathbf{a},\mathbf{b}}_{n+1}(X)$ equals*

$$\sum_{\ell=0}^{\infty} \dim_K \left(R^{\mathbf{a},\mathbf{b}}_{n+1}(X)_\ell\right) z^\ell = \frac{\sum_{\mathbf{P}} z^{\mathrm{NE}(\mathbf{P})}}{(1-z)^{(a+b+1)n - \sum_{i=1}^n (a_i+b_i)}}, \tag{3.5}$$

where the sum on the right-hand side is over all families $\mathbf{P} = (P_1, P_2, \dots, P_n)$ of nonintersecting lattice paths, with P_i running from A_i to E_i, $i = 1, 2, \dots, n$.

Finally, we remark that similar constructions are studied with minors of "ladder-shaped" matrices, of symmetric matrices, and with minors of pfaffians. It was shown by Abhyankar (1988) and Abhyankar and Kulkarni (1989) for the ladder case, by Conca (1994) for minors of a symmetric matrix, and by Ghorpade and Krattenthaler (1996) for minors of pfaffians, that the computation of Hilbert series for the resulting rings again requires enumeration of families of nonintersecting lattice paths, restricted to certain regions, with respect to their number of turns. In particular, the pfaffian case leads to the enumeration of families of nonintersecting lattice paths with given starting and end points which stay below a diagonal line.

3.4 Turn Enumeration of (Single) Lattice Paths

Examples 3.3.1 and 3.3.2 of the previous section, and the $n = 1$ case of Example 3.3.3, lead to the problem of turn enumeration of lattice paths, in some way, as explained above. In the next section, we show that if one knows the answer for the enumeration of lattice paths with a given number of *NE-turns*, then this implies solutions for all the aforementioned enumeration problems. Therefore, it is sufficient to concentrate on the enumeration of lattice paths with given starting and end points, satisfying certain restrictions, and with a given number of NE-turns. This is exactly what we do in this section.

The first question, namely '*what is the number of paths from $A = (a_1, a_2)$ to $E = (e_1, e_2)$ with exactly ℓ NE-turns*', is immediately answered by

$$\left| L((a_1, a_2) \to (e_1, e_2) \mid \mathrm{NE}(.) = \ell) \right| = \binom{e_1 - a_1}{\ell} \binom{e_2 - a_2}{\ell}. \tag{3.6}$$

This comes from the observation that any path from (a_1, a_2) to (e_1, e_2) is uniquely determined by its NE-turns. There are $e_1 - a_1$ integers from which we can choose the x-coordinates of the NE-turns, and there are $e_2 - a_2$ integers from which we can choose the y-coordinates. And, we have to choose ℓ for each of those. Thus (3.6) is explained.

The fact that paths with given starting and end points are uniquely determined by their NE-turns suggests that we should actually encode paths by their NE-turns themselves, more precisely, by the coordinates of their NE-turns. Let (p_1, q_1), (p_2, q_2), ..., (p_ℓ, q_ℓ) be the NE-turns of a path P. Then the *NE-turn representation* of P is defined by the two-rowed array

$$\begin{matrix} p_1 & p_2 & \dots & p_\ell \\ q_1 & q_2 & \dots & q_\ell, \end{matrix} \tag{3.7}$$

which consists of two strictly increasing sequences. Sometimes, we will also use a one-line notation, $(p_1, \dots, p_\ell \mid q_1, \dots, q_\ell)$, or even shorter $(\mathbf{p} \mid \mathbf{q})$ where $\mathbf{p} = (p_1, \dots, p_\ell)$ and $\mathbf{q} = (q_1, \dots, q_\ell)$.

Clearly, if P runs from (a_1, a_2) to (e_1, e_2), then $a_1 \le p_1 < p_2 < \dots < p_\ell \le e_1 - 1$ and $a_2 + 1 \le q_1 < q_2 < \dots < q_\ell \le e_2$. If we wish to make this fact transparent, we write

$$\begin{matrix} a_1 \le & p_1 & p_2 & \dots & p_\ell & \le e_1 - 1 \\ a_2 + 1 \le & q_1 & q_2 & \dots & q_\ell & \le e_2. \end{matrix} \tag{3.8}$$

For a given starting point and a given end point, by definition the empty array is the representation for the only path that has no NE-turn. For example, the two-rowed array representation of the path in Figure 3.1 would be

$$\begin{matrix} 1 & 2 & 5 \\ 1 & 3 & 4, \end{matrix}$$

or with bounds included,

$$\begin{matrix} 1 \le & 1 & 2 & 5 & \le 5 \\ 0 \le & 1 & 3 & 4 & \le 6. \end{matrix}$$

Apparently, in order to find the distribution for the game of Example 3.3.1 with $\mu = 1$, and to find the joint distribution for runs and one-sided Kolmogorov-Smirnov statistic, we need to count lattice paths, with given starting and end point, and with a given number of NE-turns, which stay below a given diagonal line. This is addressed in the following theorem.

Theorem 3.4.1 *Let $a_1 \ge a_2$ and $e_1 \ge e_2$. The number of all lattice paths from (a_1, a_2) to (e_1, e_2) staying below the diagonal line $x = y$ (being allowed to touch it) with exactly ℓ NE-turns is given by*

$$\begin{aligned} &\left| L((a_1, a_2) \to (e_1, e_2) \mid x \ge y,\ \mathrm{NE}(.) = \ell) \right| \\ &\quad = \binom{e_1 - a_1}{\ell}\binom{e_2 - a_2}{\ell} - \binom{e_1 - a_2 - 1}{\ell - 1}\binom{e_2 - a_1 + 1}{\ell + 1}. \end{aligned} \tag{3.9}$$

Remark 3.4.1 Before we sketch a proof of this theorem, a remark is in order. Recall that plain enumeration of lattice paths from (a_1, a_2) to (e_1, e_2) staying below $x = y$ (without fixing the number of NE-turns) is usually done by means of the *reflection principle* [see, for example, Comtet (1974, p. 22)]. We promised to treat all the turn enumeration problems by using two-rowed arrays. In fact, the proof below can be considered as the *reflection principle for two-rowed arrays*.

PROOF. The paths from (a_1, a_2) to (e_1, e_2) staying below $x = y$ with exactly ℓ NE-turns by the NE-turn representation can be represented by

$$\begin{array}{rcccl} a_1 \le & p_1 & p_2 & \dots & p_\ell & \le e_1 - 1 \\ a_2 + 1 \le & q_1 & q_2 & \dots & q_\ell & \le e_2, \end{array} \tag{3.10}$$

where

$$p_i \ge q_i, \qquad i = 1, 2, \dots, \ell. \tag{3.11}$$

The number of these two-rowed arrays is the number of *all* two-rowed arrays of the type (3.10) minus those two-rowed arrays of the type (3.10) which violate (3.11), i.e. where $p_i < q_i$ for some i between 1 and ℓ. We know the first number from (3.6).

Concerning the second number, we claim that two-rowed arrays of the type (3.10) which violate (3.11) are in one-to-one correspondence with two-rowed arrays of the type

$$\begin{array}{rcccccl} a_2 + 1 \le & & & r_2 & \dots & r_\ell & \le e_1 - 1 \\ a_1 \le & s_0 & s_1 & s_2 & \dots & s_\ell & \le e_2. \end{array} \tag{3.12}$$

The number of all these two-rowed arrays is $\binom{e_1 - a_2 - 1}{\ell - 1}\binom{e_2 - a_1 + 1}{\ell + 1}$, as desired. So it only remains to construct the one-to-one correspondence.

Take a two-rowed array $(\mathbf{p} \mid \mathbf{q})$ of the type (3.10) such that $p_i < q_i$ for some i. Let I be the largest integer such that $p_I < q_I$. Then map $(\mathbf{p} \mid \mathbf{q})$ to

$$\begin{array}{ccccccccc} & & q_1 & & \cdots\cdots & q_{I-1} & p_{I+1} & \cdots & p_\ell \\ p_1 & p_2 & \cdots\cdots & & p_I & q_I & q_{I+1} & \cdots & q_\ell. \end{array} \tag{3.13}$$

Observe that both rows are strictly increasing because of $q_{I-1} < q_I < q_{I+1} \le p_{I+1}$ (since I is largest with $p_I < q_I$) and $p_I < q_I$. By a case by case analysis, it can be seen that (3.13) is of type (3.12).

The inverse of this map is defined in the same way. Let $(\mathbf{r} \mid \mathbf{s})$ be a two-rowed array of the type (3.12). Let J be the largest integer such that $r_J < s_J$. If there is no such J, take $J = 1$. Then map $(\mathbf{r} \mid \mathbf{s})$ to

$$\begin{array}{cccccccc} s_0 & \cdots\cdots & & s_{J-1} & r_{J+1} & \cdots & r_\ell \\ r_2 & \cdots & r_J & s_J & \cdots\cdots\cdots & & s_\ell. \end{array} \tag{3.14}$$

It is not difficult to check that the mappings (3.13) and (3.14) are inverses of each other. This completes the proof of (3.9). ■

In order to solve the generalized problem in Example 3.3.1 (where the game is stopped when the number of heads exceeds μ times the total number of tails by exactly a), we need to count lattice paths, with given starting and end points, and with a given number of NE-turns, which stay below a line of the form $y = \mu x$. As in the situation encountered for plain counting (i.e., disregarding the number of turns), there is no nice formula for arbitrary starting and end points. But, there is if the end point lies on the boundary line. Luckily, this is exactly our situation in Example 3.3.1.

We formulate the result in an equivalent form. Namely, we consider paths bounded by a line of the form $x = \mu y$ (instead of $y = \mu x$) where the *starting* point lies on the boundary. That this is indeed equivalent is obvious from reversal of paths. Of course, we use two-rowed arrays in the proof. In contrast to the proof of Theorem 3.4.1, this proof is not purely bijective, as is pointed out in more detail after the proof. However, from the proof it can be seen very clearly where the limitations are, and in particular, why it does not generalize to an arbitrary location of the starting point.

Theorem 3.4.2 *Let μ be a positive integer and let $e_1 \geq \mu e_2$. The number of all lattice paths from $(0,0)$ to (e_1, e_2) staying below the line $x = \mu y$ (being allowed to touch it) with exactly ℓ NE-turns is given by*

$$\begin{aligned} &\left|L((0,0) \to (e_1, e_2) \mid x \geq \mu y,\ \mathrm{NE}(.) = \ell)\right| \\ &\quad = \binom{e_1}{\ell}\binom{e_2}{\ell} - \mu \binom{e_1 - 1}{\ell - 1}\binom{e_2 + 1}{\ell + 1}. \end{aligned} \tag{3.15}$$

PROOF. Again we represent our paths from $(0,0)$ to (e_1, e_2) staying below $x = \mu y$ with exactly ℓ NE-turns, by their NE-turn representation. It is

$$\begin{array}{llllll} 0 \leq & p_1 & p_2 & \dots & p_\ell & \leq e_1 - 1 \\ 1 \leq & q_1 & q_2 & \dots & q_\ell & \leq e_2, \end{array} \tag{3.16}$$

where

$$p_i \geq \mu q_i, \qquad i = 1, 2, \dots, \ell. \tag{3.17}$$

Once again, the number of these two-rowed arrays is the number of *all* two-rowed arrays of the type (3.16) minus those two-rowed arrays of the type (3.16) which violate (3.17), i.e. where $p_i < \mu q_i$ for some i between 1 and ℓ. We know the first number from (3.6).

This time, we claim that there are as many two-rowed arrays of the type (3.16) which violate (3.17) as μ times the number of two-rowed arrays of the type

$$\begin{array}{llllll} 1\le & & r_2 & \dots & r_\ell & \le e_1-1 \\ 0\le & s_0\quad s_1 & s_2 & \dots & s_\ell & \le e_2. \end{array} \tag{3.18}$$

The number of all these two-rowed arrays is $\binom{e_1-1}{\ell-1}\binom{e_2+1}{\ell+1}$, as desired. What remains to be done is to find a $(\mu:1)$ correspondence between the two-rowed arrays of type (3.16), violating (3.17), and those of type (3.18).

Take a two-rowed array $(\mathbf{p}\mid\mathbf{q})$ of the type (3.16) such that $p_i<\mu q_i$ for some i. Let I be the largest integer such that $p_I<\mu q_I$. The two-rowed array $(\mathbf{p}\mid\mathbf{q})$ then looks like

$$\begin{array}{lllllll} 0\le & p_1 & \dots\dots\dots & p_I & \dots & p_\ell & \le e_1-1 \\ 1\le & q_1 & \dots\quad q_{I-1} & q_I & \dots & q_\ell & \le e_2 \end{array}. \tag{3.19}$$

Now we fix the right portion, i.e., the entries $p_{I+1},\dots,p_\ell$ and $q_I,\dots,q_\ell$. With this fixed right portion, there are

$$\binom{\mu q_I}{I}\binom{q_I-1}{I-1} \tag{3.20}$$

possible left portions.

On the other hand, let $(\mathbf{r}\mid\mathbf{s})$ be a two-rowed array of the type (3.18). Let J be maximal with $r_J<\mu s_J$ (if there is no such J, take $J=1$), so that $(\mathbf{r}\mid\mathbf{s})$ looks like

$$\begin{array}{lllllll} 1\le & & r_2 & \dots\dots\dots & r_J & \dots\quad r_\ell & \le e_1-1 \\ 0\le & s_0\quad s_1 & s_2 & \dots\quad s_{J-1} & s_J & \dots\quad s_\ell & \le e_2 \end{array}. \tag{3.21}$$

Again, fix the right portion, i.e., the entries $r_{J+1},\dots,r_\ell$ and $s_J,\dots,s_\ell$. Furthermore, assume that the right portion in (3.21) is equal to the right portion in (3.19), i.e., assume that $J=I$, $r_i=p_i$, $i=I+1,\dots,\ell$, and $s_i=q_i$, $i=I,\dots,\ell$. With this fixed right portion in (3.21) there are

$$\binom{\mu q_I-1}{I-1}\binom{q_I}{I}=\frac{1}{\mu}\binom{\mu q_I}{I}\binom{q_I-1}{I-1} \tag{3.22}$$

possible left portions. By comparing with (3.20), we see that, for a fixed right portion, there are μ times as many two-rowed arrays of the type (3.19), with $p_I<\mu q_I$, as there are two-rowed arrays of the type (3.21), with $r_I<\mu s_I=\mu q_I$. This proves our claim and hence completes the proof of the theorem. ∎

Remark 3.4.2 The above proof could be made purely bijective if one could find a bijection for the binomial identity (3.22), i.e., for

$$\mu\binom{\mu q_I-1}{I-1}\binom{q_I}{I}=\binom{\mu q_I}{I}\binom{q_I-1}{I-1}. \tag{3.23}$$

I have not been able to find any.

On the other hand, it is exactly identity (3.23) which constitutes the limitations towards a formula for an arbitrary starting point. One may check that there is no such binomial identity in this latter situation. The appearance of a factor μ on the left-hand side of (3.23) is rather special.

There is a companion of Theorem 3.4.2 for the enumeration with respect to EN-turns. By a rotation by 180°, it can easily be transformed into a result for counting paths which stay *above* the line $x = \mu y$ with respect to NE-turns. We state the result without proof. It can be established in much the same way as Theorem 3.4.2.

Theorem 3.4.3 *Let μ be a positive integer and let $e_1 \geq \mu e_2$. The number of all lattice paths from $(0,0)$ to (e_1, e_2) staying below the line $x = \mu y$ (being allowed to touch it) with exactly ℓ EN-turns is given by*

$$\begin{aligned}&\left|L((0,0) \to (e_1,e_2) \mid x \geq \mu y,\ \mathrm{EN}(.) = \ell)\right| \\ &\quad = \binom{e_1+1}{\ell}\binom{e_2-1}{\ell-1} - \mu\binom{e_1}{\ell-1}\binom{e_2}{\ell}. \end{aligned} \tag{3.24}$$

Now, in order to find the joint distribution of *two-sided* Kolmogorov-Smirnov and run statistics, we need to count lattice paths, with given starting and end points, and with a given number of NE-turns, which stay *between two* given diagonal lines. The result which solves this problem is as follows.

Theorem 3.4.4 *Let $a_1 + t \geq a_2 \geq a_1 + s$ and $e_1 + t \geq e_2 \geq e_1 + s$. The number of all paths from (a_1, a_2) to (e_1, e_2) staying below the line $y = x + t$ and above the line $y = x + s$ (being allowed to touch them) with exactly ℓ NE-turns is given by*

$$\begin{aligned}&\left|L((a_1,a_2) \to (e_1,e_2) \mid x+t \geq y \geq x+s,\ \mathrm{NE}(.) = \ell)\right| \\ &\quad = \sum_{k=-\infty}^{\infty}\left\{\binom{e_1-a_1-k(t-s)}{\ell+k}\binom{e_2-a_2+k(t-s)}{\ell-k}\right. \\ &\qquad \left. -\binom{e_1-a_2-k(t-s)+s-1}{\ell+k}\binom{e_2-a_1+k(t-s)-s+1}{\ell-k}\right\}. \end{aligned} \tag{3.25}$$

Remark 3.4.3 Again, a remark is in order before we begin the proof. Recall that plain enumeration of lattice paths from (a_1, a_2) to (e_1, e_2) staying between two diagonal lines is usually done by means of *iterated* reflection principle [see, for example, Mohanty (1979, proof of Theorem 2 on p. 6)]. The proof below can be considered as the analogue of iterated reflection principle for two-rowed arrays.

PROOF. By the NE-turn representation, the paths under consideration are in one-to-one correspondence with two-rowed arrays of the type

$$\begin{array}{rcccl} a_1 \le & p_1 & \dots & p_\ell & \le e_1 - 1 \\ a_2 + 1 \le & q_1 & \dots & q_\ell & \le e_2, \end{array} \tag{3.26}$$

where

$$p_i + t \ge q_i \ge p_{i+1} + s. \tag{3.27}$$

The proof of this theorem is by a "cancelling" bijection on certain two-rowed arrays, which we introduce now. In fact, there are two types of arrays. Let us call two-rowed arrays of the type

$$\begin{array}{rccccccl} a_1 + k(t-s) \le & p_{1-k} & \dots & p_{1+k} & \dots & p_\ell & \le e_1 - 1 \\ a_2 + 1 - k(t-s) \le & & & q_{1+k} & \dots & q_\ell & \le e_2 \end{array} \qquad \text{for } k \ge 0$$

and

$$\begin{array}{rccccccl} a_1 + k(t-s) \le & & & p_{1-k} & \dots & p_\ell & \le e_1 - 1 \\ a_2 + 1 - k(t-s) \le & q_{1+k} & \dots & q_{1-k} & \dots & q_\ell & \le e_2 \end{array} \qquad \text{for } k < 0$$

type I arrays. Similarly, we call two-rowed arrays of the type

$$\begin{array}{rccccccl} a_2 + 1 - s + k(t-s) \le & p_{1-k} & \dots & p_{1+k} & \dots & p_\ell & \le e_1 - 1 \\ a_1 + s - k(t-s) \le & & & q_{1+k} & \dots & q_\ell & \le e_2 \end{array} \qquad \text{for } k \ge 0$$

and

$$\begin{array}{rccccccl} a_2 + 1 - s + k(t-s) \le & & & p_{1-k} & \dots & p_\ell & \le e_1 - 1 \\ a_1 + s - k(t-s) \le & q_{1+k} & \dots & q_{1-k} & \dots & q_\ell & \le e_2 \end{array} \qquad \text{for } k < 0$$

type II arrays. We shall set up a bijection between type I arrays not being of the type (3.26) – (3.27) [which means that (3.27) must be violated if both rows have equal length] and type II arrays. Given such a bijection, we could deduce

$$|\{\text{type I arrays}\}| - |\{\text{type II arrays}\}| = |\{\text{arrays of type (3.26) – (3.27)}\}|. \tag{3.28}$$

The arrays of type (3.26) – (3.27) exactly correspond to the paths we are intending to enumerate. By definition of type I and type II arrays, the left-hand side in (3.28) equals the right-hand side in (3.25). Thus (3.25) would be established.

The definition of the bijection and its inverse can be given in a unified form. Let $(\mathbf{p} \mid \mathbf{q})$ be a type I array not of the type (3.26) – (3.27) or a type II array,

$$\begin{array}{cccc} p_{1-k} & \cdots\cdots\cdots\cdots & & p_\ell \\ & q_{1+k} & \cdots & q_\ell . \end{array}$$

(This representation has to be understood symbolically. k could be also negative, whence the upper row would be shorter.) Let I be the largest integer, $1 \le I \le \ell$, such that either

$$q_I > p_I + t \quad \text{or} \quad I = -k, \tag{3.29}$$

or

$$q_I < p_{I+1} + s \quad \text{or} \quad I = k. \tag{3.30}$$

If (3.29) is satisfied, then map $(\mathbf{p} \mid \mathbf{q})$ to

$$\begin{array}{ccccccc} & (q_{1+k} - t) & \dots\dots\dots & (q_{I-1} - t) & p_{I+1} & \dots & p_\ell \\ (p_{1-k} + t) & \dots\dots\dots\dots & (p_I + t) & q_I & & \dots\dots & q_\ell . \end{array}$$

Note that both rows are strictly increasing because of $q_{I-1} < q_{I+1} \le p_{I+1} + t$ and $p_I + t < q_I$. If (3.29) is not satisfied, and hence (3.30) is, map $(\mathbf{p} \mid \mathbf{q})$ to

$$\begin{array}{cccccc} & (q_{1+k} - s) & \dots & (q_I - s) & p_{I+1} & \dots & p_\ell \\ (p_{1-k} + s) & \dots\dots\dots\dots & & (p_I + s) & q_{I+1} & \dots & q_\ell . \end{array}$$

Again note that both rows are strictly increasing, this time because of $q_I - s < p_{I+1}$ and $p_I + s < p_{I+2} + s \le q_{I+1}$.

It is not difficult to verify that this mapping maps type I arrays not being of type (3.26) – (3.27) to type II arrays not being of type (3.26) – (3.27), and vice versa. Besides, by applying this map to some array twice, one would obtain that array back. Therefore, this mapping is the desired bijection. ∎

Theorem 3.4.4 and its proof are basically from Krattenthaler and Mohanty (1995). Actually, Theorem 1 of Krattenthaler and Mohanty (1995) provides a q-analogue. A closely related paper is by Burge (1993). There, "restricted partition pairs" are considered, which are nothing but two-rowed arrays with restrictions very similar to (3.27). Burge proves a generating function result for these restricted partitions. It turns out that the above proof generalizes to prove Burge's main theorem, also. (Burge gives a different, slightly involved proof.) Remarkably, (among other results) Burge derives a number of identities expressing a Gaussian binomial coefficient as difference of two terminating basic hypergeometric sums. These identities combine two well-known but previously unrelated identities into a single one. In particular, he finds an identity which contains Rogers' proof as well as Schur's proof of the Rogers–Ramanujan identities, which were previously considered to be unrelated. Eventually, the notion of partition pairs was generalized to r-tuples of partitions and were investigated by Gessel and Krattenthaler (1996) under the name of "cylindric partitions". Again, these objects could be used to derive identities in a simple way. The resulting identities are identities for multiple basic hypergeometric series, some of them known, but many of them new.

Counting paths subject to general boundaries with respect to NE-turns is what is needed to compute the Hilbert series of ladder determinantal rings generated by 2×2 minors. "Nice" formulas cannot be expected here in general. Solutions for "one-sided" ladders were proposed by Kulkarni (1993) and Krattenthaler and Prohaska (1996). A solution for two-sided ladders is proposed by Ghorpade (private communication). Niederhausen's (1996) approach using umbral calculus methods is also worth mentioning here, though it is formulated only for EN-turns.

3.5 Applications

In this section, we apply the results from the previous section to solve (some of) the problems mentioned in Section 3.3.

ad Example 3.3.1. We saw that any game of length $(\mu+1)n+a$ corresponds to a path from $(0,0)$ to $(n, \mu n+a-1)$ staying below the line $y=\mu x+a-1$. Equivalently, by reversal of paths, it corresponds to a path from $(0,0)$ to $(\mu n+a-1, n)$ staying below the line $x=\mu y$. Also, in (3.1)–(3.3), we expressed the probability of a game of length $(\mu+1)n+a$ in terms of the NE-turns of the corresponding path. In particular, the probability that a game with first toss by coin 1 has length $(\mu+1)n+a$, is immediately obtained from Theorem 3.4.2 with $e_1=\mu n+a-1$ and $e_2=n$:

A game starting with a toss of coin 1 has length $(\mu+1)n+a$ with probability

$$\sum_{\ell=0}^{n}\left\{\binom{\mu n+a-1}{\ell}\binom{n}{\ell}-\mu\binom{\mu n+a-2}{\ell-1}\binom{n+1}{\ell+1}\right\} \\ \times p_1^{\ell+1}(1-p_1)^{n-\ell}p_2^{\mu n+a-\ell-1}(1-p_2)^{\ell}. \tag{3.31}$$

Of course, also games starting with a toss of coin 2 can be represented by a path from $(0,0)$ to $(\mu n+a-1, n)$ staying below the line $x=\mu y$. However, we have a split expression, namely (3.2) and (3.3), for the corresponding probabilities of the length of the game. The situation can be made uniform if we attach a horizontal step at the end of each path, so that we now consider paths $\bar{P}$ from $(0,0)$ to $(\mu n+a, n)$ *ending with a horizontal step* and staying below the line $x=\mu y$. Then it is easy to see that (3.2) and (3.3), in terms of $\bar{P}$, become

$$p_1^{\mathrm{NE}(\bar{P})}(1-p_1)^{n-\mathrm{NE}(\bar{P})}p_2^{\mu n+a-\mathrm{NE}(\bar{P})}(1-p_2)^{\mathrm{NE}(\bar{P})}. \tag{3.32}$$

Since the number of paths in question which have ℓ NE-turns is just the difference of the number of paths from $(0,0)$ to $(\mu n+a, n)$ staying below $x=\mu y$ and having ℓ NE-turns, minus the number of paths from $(0,0)$ to $(\mu n+a, n-1)$ staying below $x=\mu y$ and having ℓ NE-turns, we obtain from Theorem 3.4.2 by simplifying the difference:

A game starting with a toss of coin 2 has length $(\mu+1)n+a$ with probability

$$\sum_{\ell=0}^{n}\left\{\binom{\mu n+a}{\ell}\binom{n-1}{\ell-1}-\mu\binom{\mu n+a-1}{\ell-1}\binom{n}{\ell}\right\} \times p_1^{\ell}(1-p_1)^{n-\ell}p_2^{\mu n+a-\ell}(1-p_2)^{\ell}. \tag{3.33}$$

ad Example 3.3.2. We have to convert our enumeration results for NE-turns into ones for runs. Recall that the number of runs of a path is exactly one more than the number of turns (both, NE-turns and EN-turns). To avoid case by case formulation, depending on whether the number of runs is even or odd, we prefer to consider generating functions. Suppose we know the number of all paths from A to E satisfying some property R and containing a given number of NE-turns. Then we also know the generating function $\sum_P x^{\mathrm{NE}(P)}$, where the sum is over all paths P from A to E satisfying R. Let us denote it by $F(A\to E\mid R;x)$. We define four refinements of $F(A\to E\mid R;x)$. Let $F_{hv}(A\to E\mid R;x)$ be the generating function $\sum_P x^{\mathrm{NE}(P)}$ where the sum is over all paths in $L(A\to E\mid R)$ that start with a horizontal step and end with a vertical step. Similarly define $F_{hh}(A\to E\mid R;x)$, $F_{vh}(A\to E\mid R;x)$, and $F_{vv}(A\to E\mid R;x)$. The relation between enumeration by runs and enumeration by NE-turns is given by

$$\begin{aligned}\sum_{P\in L(A\to E\mid R)} x^{\mathrm{runs}(P)} &= xF_{hh}(A\to E\mid R;x^2)+x^2F_{hv}(A\to E\mid R;x^2)\\ &\quad + F_{vh}(A\to E\mid R;x^2)+xF_{vv}(A\to E\mid R;x^2).\end{aligned} \tag{3.34}$$

All the four refinements of the NE-turn generating function can be expressed in terms of NE-turn generating functions. This is seen by setting up a few linear equations and solving them. Evidently,

$$\begin{aligned}F(A\to E\mid R;x) &= F_{hh}(A\to E\mid R;x)+F_{hv}(A\to E\mid R;x)\\ &\quad + F_{vh}(A\to E\mid R;x)+F_{vv}(A\to E\mid R;x).\end{aligned}$$

Besides, if $E_1=(1,0)$ and $E_2=(0,1)$ denote the standard unit vectors, we have

$$\begin{gathered}F_{hh}(A\to E\mid R;x)+F_{hv}(A\to E\mid R;x)=F(A+E_1\to E\mid R;x),\\ F_{hv}(A\to E\mid R;x)+F_{vv}(A\to E\mid R;x)=F(A\to E-E_2\mid R;x),\\ F_{hv}(A\to E\mid R;x)=F(A+E_1\to E-E_2\mid R;x).\end{gathered}$$

Solving for F_{hh}, F_{hv}, F_{vh} and F_{vv}, we get

$$F_{hh}(A \to E \mid R; x) = F(A+E_1 \to E \mid R; x) - F(A+E_1 \to E-E_2 \mid R; x), \tag{3.35}$$

$$F_{hv}(A \to E \mid R; x) = F(A+E_1 \to E-E_2 \mid R; x), \tag{3.36}$$

$$\begin{aligned} F_{vh}(A \to E \mid R; x) &= F(A \to E \mid R; x) + (A+E_1 \to E-E_2 \mid R; x) \\ &\quad - F(A+E_1 \to E \mid R; x) - F(A \to E-E_2 \mid R; x), \end{aligned} \tag{3.37}$$

$$F_{vv}(A \to E \mid R; x) = F(A \to E-E_2 \mid R; x) - F(A+E_1 \to E-E_2 \mid R; x). \tag{3.38}$$

Now, turning to the joint distribution of runs and two-sided Kolmogorov-Smirnov statistics, we noted earlier that we have to count paths from $(0,0)$ to (n,n) staying between the lines $y=x+t$ and $y=x-t$ and which contain r runs. We do this by using (3.34) with $A=(0,0)$, $E=(n,n)$, R meaning the property to 'stay between $y=x+t$ and $y=x-t$', then using Eqs. (3.35)–(3.38) for $F_{hh}, F_{hv}, F_{vh}, F_{vv}$, respectively, in (3.34), and finally applying Theorem 3.4.4 to obtain explicit expansions for various generating functions $F(\dots)$. A comparison of coefficients of powers of z then gives, after some manipulation of binomials:

For the joint distribution of runs, denoted by $R_{n,n}$, and the two-sided Kolmogorov–Smirnov statistics $D_{n,n}$, we have

$$\begin{aligned} &\binom{2n}{n} \Pr[D_{n,n} \le t/n,\ R_{n,n} = 2r+1] \\ &= \sum_{k=-\infty}^{\infty} \left\{ \binom{n-2kt-1}{r+k} \binom{n+2kt-1}{r-k-1} + \binom{n-2kt-1}{r+k-1} \binom{n+2kt-1}{r-k} \right. \\ &\qquad \left. -2 \binom{n-2kt+t-1}{r+k-1} \binom{n+2kt-t-1}{r-k} \right\}, \end{aligned}$$

and

$$\begin{aligned} &\binom{2n}{n} \Pr[D_{n,n} \le t/n,\ R_{n,n} = 2r] \\ &= \sum_{k=-\infty}^{\infty} \left\{ 2 \binom{n-2kt-1}{r+k-1} \binom{n+2kt-1}{r-k-1} - \binom{n-2kt+t-1}{r+k-2} \binom{n+2kt-1}{r-k} \right. \\ &\qquad \left. - \binom{n-2kt+t-1}{r+k-1} \binom{n+2kt-t-1}{r-k-1} \right\}. \end{aligned}$$

Thus, we recover the results of Vellore (1972, Theorems 8 and 9). She derived these results by very different means. (The expressions therefore look differently. But it is not difficult to show that they are really equivalent.) The path of derivation we have chosen here is from Krattenthaler and Mohanty (1993) where it was also used to obtain extensions and q-analogues of the above result.

ad Example 3.3.3. By Theorems 3.3.1 and 3.3.2, the case of $n = 1$ in Example 3.3.3, i.e., the case of rings generated by (at most) 2×2 minors in the way described above, leads to the problem of enumerating paths with given starting and end points which have a given number of NE-turns. Clearly, this is done by (3.6).

Besides, we indicated that the case of pfaffian rings generated by 4×4 pfaffians leads to the enumeration of paths with given starting and end points which have a given number of NE-turns and stay below a diagonal line. Clearly, this is done by Theorem 3.4.1.

3.6 Nonintersecting Lattice Paths and Turns

Here, we complete the solutions to our Examples of Section 3.3. More precisely, we address the problem of enumerating nonintersecting lattice paths with a given number of NE-turns, which is the problem to be solved in order to compute Hilbert series of determinantal and pfaffian rings, as we described earlier in Example 3.3.3. If one forgets about the number of turns, i.e., if one is interested in the plain enumeration of nonintersecting lattice paths with given starting and end points, then the solution is a certain determinant. This is a classical result now [*cf.* Gessel and Viennot (1985 and 1989, Corollary 2); Stembridge (1990, Theorem 1.2)]. In fact, it has been realized over the past ten years that nonintersecting lattice paths have innumerable applications in combinatorics, probability, statistics, physics, etc. [see the references in Krattenthaler (1996b) for combinatorial applications, and the references in the Introduction for applications in physics and probability; in fact, most of the determinantal formulas in probability and statistics, like "Steck's determinants" [Mohanty (1971), Pitman (1972) and Steck (1969, 1974)] follow easily from nonintersecting lattice paths; see also Sulanke (1990)]. However, the method that is used for the plain enumeration [the "Gessel–Viennot involution", which actually can be traced back to Lindström (1973) and Karlin and McGregor (1959a)], is not appropriate to keep track of turns. Still, the answers to "turn enumeration" are determinants. But, alternative methods are needed now. It is the combinatorics of two-rowed arrays which explains these determinants. In fact, it is the context of nonintersecting lattice paths in which the usefulness of working with two-rowed arrays becomes most striking. Interestingly, the techniques developed here arose in the study of plane partition and tableaux generating functions [Krattenthaler (1995a)] and of identities for Schur functions [Krattenthaler (1993)].

From Theorems 3.3.1 and 3.3.2, we know for the computation of the Hilbert series for the determinantal rings $R_{n+1}(X)$ and $R^{\mathbf{a},\mathbf{b}}_{n+1}(X)$ that we need to enumerate families $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ of nonintersecting lattice paths, where P_i runs from $(0, a_{n-i+1})$ to $(a - b_{n-i+1}, b)$, $i = 1, 2, \ldots, n$, where the total number of NE-turns in $\mathbf{P}$ is some fixed number. Here, the starting points are lined up vertically and the end points are lined up horizontally. In fact, we are able to answer the problem even if the starting and end points are (basically) in general position. Let $\mathcal{A} = (A_1, A_2, \ldots, A_n)$ and $\mathcal{E} = (E_1, E_2, \ldots, E_n)$ be points in the two-dimensional integer lattice $\mathbf{Z^2}$. The restriction on the location of the points which we have to impose is the one which is always necessary with nonintersecting lattice paths [see Gessel and Viennot (1989) and Stembridge (1990)]. Namely, we assume that the starting points are lined up north-west to south-east, strictly from north to south, and that the end points are also lined up north-west to south-east, but strictly from west to east. We have the following theorem.

Theorem 3.6.1 *Let* $A_i = (a_1^{(i)}, a_2^{(i)})$ *and* $E_i = (e_1^{(i)}, e_2^{(i)})$, $i = 1, 2, \ldots, n$, *be lattice points satisfying*

$$a_1^{(1)} \le a_1^{(2)} \le \cdots \le a_1^{(n)}, \quad a_2^{(1)} > a_2^{(2)} > \cdots > a_2^{(n)},$$

and

$$e_1^{(1)} < e_1^{(2)} < \cdots < e_1^{(n)}, \quad e_2^{(1)} \ge e_2^{(2)} \ge \cdots \ge e_2^{(n)}.$$

The generating function $\sum_{\mathbf{P}} z^{\mathrm{NE}(P)}$, *where the sum is over all families* $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ *of nonintersecting lattice paths* $P_i : A_i \to E_i$, *equals*

$$\det_{1 \le i,j \le n} \left\{ \sum_{k \ge 0} \binom{e_1^{(i)} - a_1^{(j)} + j - i}{k + j - i} \binom{e_2^{(i)} - a_2^{(j)} - j + i}{k} z^k \right\}. \tag{3.39}$$

Remark 3.6.1 This theorem was independently proved by Kulkarni (1993), who derived it from a theorem on determinantal rings due to Abhyankar, by Modak (1992), who found a manipulatory proof, and for the first time by combinatorial means by Krattenthaler (1995b, 1996a), using two-rowed arrays. See also Ghorpade (1996).

Sketch of proof. If we want to prove this theorem by means of two-rowed arrays, we have to first work out how the condition of two paths to be nonintersecting translates into the corresponding two-rowed arrays.

Let P_1, P_2 be two paths, $P_1 : A \to E$, $P_2 : B \to F$, where $A = (a_1, a_2)$, $B = (b_1, b_2)$, $E = (e_1, e_2)$, $F = (f_1, f_2)$, A located in the north-west of B (strictly in direction north and weakly in direction west), and E located in the north-west of F (weakly in direction north and strictly in direction west), i.e., with

$$a_1 \le b_1, \; a_2 > b_2, \; e_1 < f_1, \; e_2 \ge f_2.$$

Let the array representations of P_1 and P_2 be

$$P_1: \qquad \begin{array}{rcccl} a_1 \le & p_1 & \ldots & p_k & \le e_1 - 1 \\ a_2 + 1 \le & q_1 & \ldots & q_k & \le e_2 \end{array} \tag{3.40}$$

and

$$P_2: \qquad \begin{array}{rcccl} b_1 \le & r_1 & \ldots & r_l & \le f_1 - 1 \\ b_2 + 1 \le & s_1 & \ldots & s_l & \le f_2, \end{array} \tag{3.41}$$

respectively.

Suppose that P_1 and P_2 intersect, i.e. have a point in common. Let $\mathcal{M}$ be a meeting point of P_1 and P_2. For technical reasons, set $p_{k+1} := e_1$ and $q_0 := a_2$. (Note that the thereby augmented sequences a and b remain strictly increasing.)

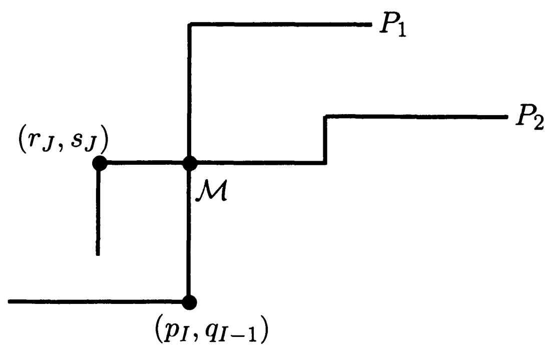


Figure 3.5

Considering the east-north turn (p_I, q_{I-1}) in P_1 immediately preceding $\mathcal{M}$ (and being allowed to be equal to $\mathcal{M}$) and the north-east turn (r_J, s_J) in P_2 immediately preceding $\mathcal{M}$ (and being allowed to be equal to $\mathcal{M}$), we get the inequalities (*cf.* Figure 3.5)

$$r_J \le p_I, \tag{3.42}$$
$$q_{I-1} \le s_J, \tag{3.43}$$

where

$$1 \le I \le k+1, \quad 1 \le J \le l. \tag{3.44}$$

Of course, k, l, p_I, q_I, r_J, s_J, etc., refer to the array representations of P_1 and P_2. It now becomes apparent that the above assignments for p_{k+1} and q_0 are needed for the inequalities (3.42) and (3.43) to make sense for $I = 1$ or $I = k+1$. Note that $\mathcal{M} = (p_I, s_J)$. Vice versa, if (3.42) – (3.44) are satisfied, then there must be a meeting point between P_1 and P_2 (because of the particular location of the starting and end points A, B, E, F).

Summarizing, the existence of I, J satisfying (3.42) – (3.44) characterize the array representations of *intersecting* pairs of paths. Therefore, we call two-rowed arrays P_1 and P_2 of the form (3.40) and (3.41), respectively, *intersecting* if (3.42) – (3.44) are satisfied, for some I and J, otherwise *nonintersecting.* The point $\mathcal{M} = (p_I, s_J)$ is called their *intersection point.*

We also need to consider *skew* two-rowed arrays. For convenience, we introduce some terminology. Let $j > 0$. We say that the two-rowed array P *is of the type* j if P has the form

$$\begin{array}{cccccccc} p_{-j+1} & p_{-j+2} & \cdots & p_{-1} & p_0 & p_1 & \cdots & p_k \\ & & & & & q_1 & \cdots & q_k \end{array}$$

for some $k \geq 0$. We say that P *is of the type* $-j$ if P has the form

$$\begin{array}{cccccccc} & & & & & p_1 & \cdots & p_k \\ q_{-j+1} & q_{-j+2} & \cdots & q_{-1} & q_0 & q_1 & \cdots & q_k \end{array}$$

for some $k \geq 0$. Note that the placement of indices is chosen such that non-positive indices can occur only in *one* row of P, while the positive indices occur in both rows of P. The meaning of non-skew two-rowed arrays being intersecting, and nonintersecting, and of intersection points, is extended to skew two-rowed arrays in the obvious way. In abuse of its actual literal meaning, we define the "number of NE-turns" of a two-rowed array P to be one half of the number of entries of P. (Recall that, under the correspondence between paths and two-rowed arrays, the number of NE-turns of the path equals one half of the number of entries of the corresponding two-rowed array.) We use the same short notation $\mathrm{NE}(P)$ for this number.

Now, we are in the position to actually begin with the proof of (3.39). First, we give the combinatorial interpretation of the determinant (3.39) in terms of two-rowed arrays. Expanding the determinant in (3.39), we obtain

$$\begin{aligned} \sum_{\sigma \in \mathcal{S}_n} \operatorname{sgn} \sigma \prod_{i=1}^{n} \binom{e_1^{(i)} - a_1^{(\sigma(i))} + \sigma(i) - i}{k_i + \sigma(i) - i} \binom{e_2^{(i)} - a_2^{(\sigma(i))} - \sigma(i) + i}{k_i} z^{k_i} \\ = \sum_{(\sigma, \mathbf{P})} \operatorname{sgn} \sigma \, z^{\mathrm{NE}(\mathbf{P})}, \end{aligned} \tag{3.45}$$

where $\mathcal{S}_n$ denotes the symmetric group of order n, and the sum on the right-hand side is over all pairs $(\mathbf{P}, \sigma)$ of permutations σ in $\mathcal{S}_n$, and families $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ of two-rowed arrays, P_i being of type $\sigma(i) - i$, and the bounds for the entries of P_i being as follows:

$$\begin{array}{rcl} a_1^{(\sigma(i))} + i - \sigma(i) \leq & \cdots & p_{\ell_i}^{(i)} \leq e_1^{(i)} - 1 \\ a_2^{(\sigma(i))} - i + \sigma(i) + 1 \leq & \cdots & q_{\ell_i}^{(i)} \leq e_2^{(i)}, \end{array} \tag{3.46}$$

$i = 1, 2, \ldots, n.$

The outline of the proof is as follows. We show that in the sum on the right-hand side of (3.45) all contributions corresponding to pairs $(\mathbf{P}, \sigma)$ where $\mathbf{P}$ is an intersecting family of two-rowed arrays cancel. (We call a pair $(\mathbf{P}, \sigma)$ *intersecting* if $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ contains two two-rowed arrays P_i and P_{i+1} with consecutive indices that have an intersection point. Otherwise it is called *nonintersecting*. In the sequel, two-rowed arrays with consecutive indices will be called *neighbouring* two-rowed arrays.) This is done by constructing a sign-reversing (with respect to $\operatorname{sgn}\sigma$) involution on these pairs, which keeps the total number of entries in the two-rowed arrays fixed. (Recall that, under the correspondence between paths and two-rowed arrays, the number of NE-turns of the path equals one half of the number of entries of the corresponding two-rowed array.) Finally, it is shown that, in a pair $(\mathbf{P}, \sigma)$ with $\sigma \neq \mathrm{id}$, the family $\mathbf{P}$ must be intersecting. This establishes that only pairs $(\mathbf{P}, \mathrm{id})$ where $\mathbf{P}$ is a nonintersecting family of two-rowed arrays contribute to the sum on the right-hand side of (3.45). But these pairs correspond exactly to the families of nonintersecting paths under consideration, and hence Theorem 3.6.1 would be proved.

Let $(\mathbf{P}, \sigma)$ be a pair under consideration for the sum on the right-hand side of (3.45). Besides, we assume that $\mathbf{P}$ contains two neighbouring two-rowed arrays P_i and P_{i+1} that have an intersection point. Consider all intersection points of neighbouring arrays. Among these points, choose those with maximal x-coordinate, and among all those choose the intersection point with maximal y-coordinate. Denote this intersection point by $\mathcal{M}$. Let i be minimal such that $\mathcal{M}$ is an intersection point of P_i and P_{i+1}. Let $P_i = (a \mid b) = (\ldots p_{\ell_i} \mid \ldots q_{\ell_i})$ and $P_{i+1} = (c \mid d) = (\ldots r_{\ell_{i+1}} \mid \ldots s_{\ell_{i+1}})$. Recall that P_i is of type $\sigma(i) - i$ and P_{i+1} is of type $\sigma(i+1) - i - 1$ and that the bounds of the entries in P_i and P_{i+1} are determined by (3.46). By (3.42) – (3.44), $\mathcal{M}$ being an intersection point of P_i and P_{i+1} means that there exist I and J such that P_i looks like

$$\begin{array}{rccccccl} a_1^{(\sigma(i))} + i - \sigma(i) \leq & \ldots & p_{I-1} & p_I & \ldots & p_{\ell_i} & \leq e_1^{(i)} - 1 \\ a_2^{(\sigma(i))} - i + \sigma(i) + 1 \leq & \ldots & q_{I-1} & q_I & \ldots & q_{\ell_i} & \leq e_2^{(i)}, \end{array} \tag{3.47}$$

P_{i+1} looks like

$$\begin{array}{rccccccl} a_1^{(\sigma(i+1))} + i + 1 - \sigma(i+1) \leq & \ldots\ldots\ldots & & r_J & r_{J+1} & \ldots & r_{\ell_{i+1}} & \leq e_1^{(i+1)} - 1 \\ a_2^{(\sigma(i+1))} - i + \sigma(i+1) \leq & \ldots & s_{J-1} & s_J & \ldots\ldots\ldots & & s_{\ell_{i+1}} & \leq e_2^{(i+1)}, \end{array} \tag{3.48}$$

$\mathcal{M} = (p_I, s_J)$,

$$r_J \leq p_I \tag{3.49}$$

$$q_{I-1} \leq s_J \tag{3.50}$$

and

$$1 \leq I \leq \ell_i + 1, \quad 0 \leq J \leq \ell_{i+1}. \tag{3.51}$$

Because of the construction of $\mathcal{M}$, the indices I *and* J are maximal with respect to (3.49) – (3.51).

We map $(\mathbf{P},\sigma)$ to the pair $(\bar{\mathbf{P}},\sigma\circ(i,i+1))$ [$(i,i+1)$ denotes the transposition interchanging i and $i+1$], where $\bar{\mathbf{P}}=(P_1,\dots,P_{i-1},\bar{P}_i,\bar{P}_{i+1},P_{i+2},\dots,P_n)$ with $\bar{P}_i$ being given by

$$\begin{matrix} \dots & r_J-1 & p_I & \dots & p_{\ell_i} \\ \dots & s_{J-1}+1 & q_I & \dots & q_{\ell_i}, \end{matrix} \tag{3.52}$$

$\bar{P}_{i+1}$ being given by

$$\begin{matrix} \dots & \dots & p_{I-1}+1 & r_{J+1} & \dots & r_{\ell_{i+1}} \\ \dots & q_{I-1}-1 & s_J & \dots & \dots & s_{\ell_{i+1}}. \end{matrix} \tag{3.53}$$

First of all, this operation is well-defined, i.e., all the rows in (3.52) and (3.53) are strictly increasing. To see this, we have to check $r_J-1<p_I$, $s_{J-1}+1<q_I$, $p_{I-1}+1<r_{J+1}$, and $q_{I-1}-1<s_J$. This is obvious for the first and last inequalities, because of (3.49) and (3.50). As for the second inequality, let us suppose $s_{J-1}+1\ge q_I$. Then, by (3.49), we have $r_J\le p_I<p_{I+1}$ and $q_I\le s_{J-1}+1\le s_J$. This means that (p_{I+1},s_J) is an intersection point of P_i and P_{i+1}, with an x-coordinate larger than that of $\mathcal{M}=(p_I,s_J)$, contradicting the "maximality" of $\mathcal{M}$. Similarly, if we assume $p_{I-1}+1\ge r_{J+1}$, we have $r_{J+1}\le p_{I-1}+1\le p_I$ and, by (3.50), $q_{I-1}\le s_J<s_{J+1}$. This means that (p_I,s_{J+1}) is an intersection point of P_i and P_{i+1}, with a y-coordinate larger than that of $\mathcal{M}=(p_I,s_J)$, again contradicting the "maximality" of $\mathcal{M}$.

We claim that $(\bar{\mathbf{P}},\sigma(i,i+1))$ is again a pair under consideration for the generating function (3.45). That is, we claim that $\bar{P}_i$ is of type $(\sigma\circ(i,i+1))(i)-i=\sigma(i+1)-i$, that $\bar{P}_{i+1}$ is of type $(\sigma\circ(i,i+1))(i+1)-i-1=\sigma(i)-i-1$, and that the bounds for the entries of $\bar{P}_i$ are given by

$$\begin{matrix} a_1^{(\sigma(i+1))}+i-\sigma(i+1)\le & \dots & r_J-1 & p_I & \dots & p_{\ell_i} & \le e_1^{(i)}-1 \\ a_2^{(\sigma(i+1))}-i+\sigma(i+1)+1\le & \dots & s_{J-1}+1 & q_I & \dots & q_{\ell_i} & \le e_2^{(i)}, \end{matrix} \tag{3.54}$$

and that those for $\bar{P}_{i+1}$ are given by

$$\begin{matrix} a_1^{(\sigma(i))}+i+1-\sigma(i)\le & \dots & \dots & p_{I-1}+1 & r_{J+1} & \dots & r_{\ell_{i+1}} & \le e_1^{(i+1)}-1 \\ a_2^{(\sigma(i))}-i+\sigma(i)\le & \dots & q_{I-1}-1 & s_J & \dots & \dots & s_{\ell_{i+1}} & \le e_2^{(i+1)}. \end{matrix} \tag{3.55}$$

The claims concerning the types of $\bar{P}_i$ and $\bar{P}_{i+1}$ are trivial. The claim concerning the bounds requires some case-by-case analysis, which we leave to the reader. One may also refer to Krattenthaler (1995b, 1996a). Obviously, the map (3.52) – (3.53) reverses the sign of the associated permutation. Besides, it can be checked that it is an involution. The proof that, given a pair $(\mathbf{P},\sigma)$, $\mathbf{P}=(P_1,P_2,\dots,P_n)$, $\sigma\ne\mathrm{id}$, there exist neighbouring two-rowed arrays P_i and P_{i+1} having an intersection point, is slightly technical. We refer the reader to Krattenthaler (1995b, 1996a) for the details. ∎

Remark 3.6.2 The map from (3.47) and (3.48) to (3.52) and (3.53) can be considered as the analogue in the "world of two-rowed arrays" for the interchanging of paths which is usually done with nonintersecting lattice paths [see, for example, Gessel and Viennot (1985), Stembridge (1990), and Krattenthaler (1995a, Section 2.2)].

Another problem that is posed by Example 3.3.3 is the enumeration of families of nonintersecting lattice paths which are bounded by a diagonal line with respect to their number of turns. Recall that this is necessary for the computation of the Hilbert series of pfaffian rings and of ladder determinantal rings where the ladder restriction is a diagonal boundary. Also here, we have a result where the location of the starting and end points is more general than needed.

Theorem 3.6.2 *Let $A_i = (a_1^{(i)}, a_2^{(i)})$ and $E_i = (e_1^{(i)}, e_2^{(i)})$, $i = 1, 2, \ldots, n$, be lattice points satisfying*

$$a_1^{(1)} \le a_1^{(2)} \le \cdots \le a_1^{(n)}, \quad a_2^{(1)} > a_2^{(2)} > \cdots > a_2^{(n)},$$

$$e_1^{(1)} < e_1^{(2)} < \cdots < e_1^{(n)}, \quad e_2^{(1)} \ge e_2^{(2)} \ge \cdots \ge e_2^{(n)},$$

and $a_1^{(i)} \ge a_2^{(i)}$, $e_1^{(i)} \ge e_2^{(i)}$, $i = 1, 2, \ldots, n$. The generating function $\sum_{\mathbf{P}} z^{\mathrm{NE}(P)}$, where the sum is over all families $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ of nonintersecting lattice paths $P_i : A_i \to E_i$, which stay below the line $x = y$ (being allowed to touch it), equals

$$\det_{1\le i,j\le n}\left(\left\{\sum_{k\ge 0}\binom{e_1^{(i)} - a_1^{(j)} + j - i}{k + j - i}\binom{e_2^{(i)} - a_2^{(j)} - j + i}{k}\right.\right.$$
$$\left.\left. - \binom{e_1^{(i)} - a_2^{(j)} - j - i + 1}{k - i}\binom{e_2^{(i)} - a_1^{(j)} + j + i - 1}{k + j}\right\} z^k\right). \tag{3.56}$$

Sketch of proof. Again, we work with families of two-rowed arrays. This time we consider triples $(\mathbf{P}, \sigma, \eta)$, where σ is a permutation in $\mathcal{S}_n$, $\eta \in \{-1, 1\}^r$, and $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ is a family of two-rowed arrays, with P_i being of type $\eta_i\sigma(i) - i$ and the bounds of P_i being given by

$$\begin{array}{rcl} a_1^{(\sigma(i))} + i - \sigma(i) \le & \ldots & \le e_1^{(i)} - 1 \\ a_2^{(\sigma(i))} - i + \sigma(i) + 1 \le & \ldots & \le e_2^{(i)} \end{array}, \qquad \text{for } \eta = 1, \tag{3.57}$$

and

$$\begin{array}{rcl} a_2^{(\sigma(i))} + i + \sigma(i) - 1 \le & \ldots & \le e_1^{(i)} - 1 \\ a_1^{(\sigma(i))} - i - \sigma(i) + 2 \le & \ldots & \le e_2^{(i)} \end{array}, \qquad \text{for } \eta = -1. \tag{3.58}$$

Define $\operatorname{sgn}\eta := \prod_{i=1}^{n}\eta_i$. It is easy to see that (3.56) is the generating function

$$\sum_{(\mathbf{P},\sigma,\eta)} \operatorname{sgn}\eta \operatorname{sgn}\sigma\, z^{\mathrm{NE}(\mathbf{P})}, \tag{3.59}$$

where the sum is over all triples which have been described above.

Now, the basic idea is as follows. We show that in the sum (3.59) all contributions cancel which correspond to triples $(\mathbf{P},\sigma,\eta)$, where $\mathbf{P}$ is an intersecting family of two-rowed arrays, or where the two-rowed array P_1 "crosses" $y = x$, by which we mean that there is an entry in the upper row of P_1 which is smaller than its neighbour in the bottom row of P_1. Again, this is done by constructing a sign-reversing involution (with respect to $\operatorname{sgn}\eta\operatorname{sgn}\sigma$) on those triples. Roughly described, this involution combines the "reflection principle for two-rowed arrays" with the "interchanging procedure for two-rowed arrays". Namely, this involution is defined to be the map (3.47) and (3.48) to (3.52) and (3.53) if $\mathbf{P}$ contains neighbouring two-rowed arrays which are intersecting, and if not, but the first two-rowed array P_1 "crosses" $y = x$, then it is defined to be basically the map (3.13), applied to P_1. It can be shown that in a triple $(\mathbf{P},\sigma,\eta)$ with $\sigma \neq \mathrm{id}$ or $\eta \neq (1,1,\ldots,1)$, the family $\mathbf{P}$ must be intersecting or P_1 "crosses $y = x$". This establishes that only triples $(\mathbf{P},\mathrm{id},(1,1,\ldots,1))$, where $\mathbf{P}$ is a nonintersecting family of two-rowed arrays which do not cross $y = x$, contribute to the sum (3.59). But these triples exactly correspond to the families of nonintersecting paths under consideration, and hence Theorem 3.6.2 would be proved. We refer the reader to Krattenthaler (1995b, 1996a) for the details. ■

As mentioned before, Theorem 3.6.2 can be applied to the computation of the Hilbert series of certain ladder determinantal rings (one sided, with a diagonal upper bound) and also of pfaffian rings. The computation of Hilbert series of rings generated by minors of a symmetric matrix as considered by Conca (1994) can also be solved by using the method of two-rowed arrays; see Krattenthaler (1996a). For arbitrary one-sided ladders, there is a solution when the starting points, and end points, are located "successively" (such as in Figure 3.4) by Krattenthaler and Prohaska (1996) proving a remarkable formula conjectured by Conca and Herzog (1994). For "generally" located starting and end points, there is a solution in terms of a determinant with entries counting certain two-rowed arrays by Krattenthaler (1996a). The case of two-sided ladder determinantal rings appears to be out of reach by the method of two-rowed arrays. Perhaps, the extension of the dummy path idea in Krattenthaler and Mohanty (1995) will be useful in this context. Finally, we want to point the reader to a refined turn counting for pairs of paths [Krattenthaler and Sulanke (1996)] which relates this subject also to polyomino counting.

References

1. Abhyankar, S. S. (1987). Determinantal loci and enumerative combinatorics of Young tableaux, In *Algebraic Geometry and Commutative Algebra* in honor of M. Nagata, pp. 1–26.

2. Abhyankar, S. S. (1988). *Enumerative Combinatorics of Young Tableaux*, New York: Marcel Dekker.

3. Abhyankar, S. S. and Kulkarni, D. M. (1989). On Hilbertian ideals, *Linear Algebra and its Applications*, **116**, 53–76.

4. Burge, W. H. (1993). Restricted partition pairs,*Journal of Combinatorial Theory, Series A*, **63**, 210–222.

5. Comtet, L. (1974). *Advanced Combinatorics*, Dordrecht: Reidel.

6. Conca, A. (1994). Symmetric ladders, *Nagoya Mathematical Journal*, **136**, 35–56.

7. Conca, A. and Herzog, J. (1994). On the Hilbert function of determinantal rings and their canonical module, *Proceedings of the American Mathematical Society*, **122**, 677–681.

8. DeGroot, M. H. (1959). Unbiased sequential estimation for binomial populations, *Annals of Mathematical Statistics*, **30**, 80–101.

9. Essam, J. W. and Guttmann, A. J. (1995). Vicious walkers and directed polymer networks in general dimensions, *Physical Review E*, **52**, 5849–5862.

10. Fisher, M. E. (1984). Walks, walls, wetting, and melting, *Journal of Statistical Physics*, **34**, 667–729.

11. Gessel, I. M. and Krattenthaler, C. (1996). Cylindric partitions, *Transactions of the American Mathematical Society* (to appear).

12. Gessel, I. M. and Viennot, X. (1985). Binomial determinants, paths, and hook length formulae, *Advances in Mathematics*, **58**, 300–321.

13. Gessel, I. M. and Viennot, X. (1989). Determinants, paths, and plane partitions, *Preprint.*

14. Ghorpade, S. R. (1996). Young bitableaux, lattice paths and Hilbert functions, *Journal of Statistical Planning and Inference* **54**, 55–66.

15. Ghorpade, S. R. and Krattenthaler, C. (1996). On pfaffian ideals, *Preprint.*

16. Herzog, J. and Trung, N. V. (1992). Gröbner bases and multiplicity of determinantal and Pfaffian ideals, *Advances in Mathematics*, **96**, 1–37.

17. Karlin, S. (1988). Coincident probabilities and applications in combinatorics, *Journal of Applied Probability*, **25**, 185–200.

18. Karlin, S. and McGregor, J. L. (1959a). Coincidence probabilities, *Pacific Journal of Mathematics*, **9**, 1141–1164.

19. Karlin, S. and McGregor, J. L. (1959b). Coincidence properties of birth-and-death processes, *Pacific Journal of Mathematics*, **9**, 1109–1140.

20. Krattenthaler, C. (1989). Counting lattice paths with a linear boundary, Part 2: q-ballot and q-Catalan numbers, *Sitz. ber. d. ÖAW, Mathnaturwiss. Klasse*, **198**, 171–199.

21. Krattenthaler, C. (1993). Non-crossing two-rowed arrays and summations for Schur functions, In *Proceedings of the 5th Conference on Formal Power Series and Algebraic Combinatorics, Florence, 1993* (Eds., A. Barlotti, M. Delest and R. Pinzani), pp. 301–314, Università di Firenze: D.S.I.

22. Krattenthaler, C. (1995a). *The Major Counting of Nonintersecting Lattice Paths and Generating Functions for Tableaux*, Providence, Rhode Island: Memoirs of the American Mathematical Society, **115**.

23. Krattenthaler, C. (1995b). Counting nonintersecting lattice paths with respect to weighted turns, *Seminaire Lotharingien Combin.*, **34**, paper B34i, 17 pp.

24. Krattenthaler, C. (1996a). Non-crossing two-rowed arrays, *Preprint.*

25. Krattenthaler, C. (1996b). Nonintersecting lattice paths and oscillating tableaux, *Journal of Statistical Planning and Inference* **54**, 75–85.

26. Krattenthaler, C. and Mohanty, S. G. (1993). On lattice path counting by major and descents, *European Journal of Combinatorics*, **14**, 43–51.

27. Krattenthaler, C. and Mohanty, S. G. (1995). Counting tableaux with row and column bounds, *Discrete Mathematics*, **139**, 273–286.

28. Krattenthaler, C. and Prohaska, M. (1996). A remarkable formula for counting nonintersecting lattice paths in a ladder with respect to turns, *Transactions of the American Mathematical Society* (to appear).

29. Krattenthaler, C. and Sulanke, R. A. (1996). Counting pairs of nonintersecting lattice paths with respect to weighted turns, *Discrete Mathematics* **153**, 189–198.

30. Kulkarni, D. M. (1993). Hilbert polynomial of a certain ladder-determinantal ideal, *Journal of Algebraic Combinatorics*, **2**, 57–72.

31. Kulkarni, D. M. (1996). Counting of paths and coefficients of Hilbert polynomial of a determinantal ideal, *Discrete Mathematics*, **154**, 141–151.

32. Lindström, B. (1973). On the vector representations of induced matroids, *Bulletin of the London Mathematical Society*, **5**, 85–90.

33. Modak, M. R. (1992). Combinatorial meaning of the coefficients of a Hilbert polynomial, *Proceedings of the Indian Academy of Science (Mathematical Sciences)*, **102**, 93–123.

34. Mohanty, S. G. (1966). On a generalised two-coin tossing problem, *Biometrische Zeitschrift*, **8**, 266–272.

35. Mohanty, S. G. (1971). A short proof of Steck's result on two-sample Smirnov statistics, *Annals of Mathematical Statistics*, **42**, 413–414.

36. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

37. Narayana, T. V. (1959). A partial order and its applications to probability theory, *Sankhyā*, **21**, 91–98.

38. Narayana, T. V. (1979). Lattice path combinatorics with statistical applications, *Mathematical Statistics Expositions, No. 23*, Toronto: University of Toronto Press.

39. Niederhausen, H. (1996). Symmetric Sheffer sequences and their applications to lattice path counting, *Journal of Statistical Planning and Inference* **54**, 87–100.

40. Pitman, E. J. G. (1972). Simple proofs of Steck's determinantal expressions for probabilities in the Kolmogorov and Smirnov tests, *Bulletin of the Astralian Mathematical Society*, **7**, 227–232.

41. Steck, G. P. (1969). The Smirnov tests as rank tests, *Annals of Mathematical Statistics*, **40**, 1449–1466.

42. Steck, G. P. (1974). A new formula for $P(R_i \leq b_i,\ 1 \leq i \leq m \mid m, n, F = G^k)$, *Annals of Probability*, **2**, 155–160.

43. Stembridge, J. R. (1990). Nonintersecting paths, pfaffians and plane partitions, *Advances in Mathematics*, **83**, 96–131.

44. Sulanke, R. A. (1990). A determinant for q-counting lattice paths, *Discrete Mathematics*, **81**, 91–96.

45. Vellore, S. (1972). Joint distributions of Kolmogorov–Smirnov statistics and runs, *Studia Scientiarum Mathematica Hungarica*, **7**, 155–165.

4

Lattice Path Counting, Simple Random Walk Statistics, and Randomization: An Analytic Approach

Wolfgang Panny and Walter Katzenbeisser

University of Economics, Augasse 2–6, Vienna

Abstract: In this paper an approach to lattice paths, simple random walks and randomized random walks is presented, which emphasizes the common features and permits to treat various aspects in a unified framework.

Keywords and phrases: Lattice paths, simple random walks, randomized random walks, rank order statistics, Dwass's method

4.1 Introduction

The purpose of this paper is to present an approach which has proved useful in dealing with various aspects of the simple random walk. The approach involves generating functions and is mainly of an analytical nature. A striking feature is that the simple random walk results happen to essentially comprise their lattice path counterparts as special cases, wich is due to the use of generalized trinomial coefficients. Moreover, the simple random walk results can be taken as a starting point to derive their continuous time counterparts by a limiting process, which takes us to randomized random walks. Our randomization procedure, in fact, may be seen as an alternative to Feller's randomization technique.

Our plan is as follows. In Section 4.2 we confine ourselves to the lattice path context, while in Section 4.3 we present the general method for simple random walks. Section 4.4 is devoted to a presentation of our randomization procedure. Each section also includes an example in order to illustrate the application.

4.2 Lattice Paths

Consider the lattice points $(0, y_0)$, $(1, y_1)$, $(2, y_2)$, $\ldots$ in the (x, y)-grid. We are dealing with lattice paths $(y_0, y_1, y_2, \ldots, y_n)$ with $y_\kappa = y_{\kappa-1} + \epsilon_\kappa$ where $\epsilon_\kappa \in \{-1, +1\}$, $\kappa = 1, 2, \ldots, n$ and $y_0 = 0$, $y_n = \ell$. Such a lattice path starts at the origin and leads to ℓ after n steps. Confining to $y_0 = 0$ actually constitutes no restriction at all. So, this assumption will be made in the sequel if not explicitly stated otherwise. Of course, $|y_\kappa| \leq \kappa$ and $y_\kappa \equiv \kappa(2)$. In particular, $y_n = \ell$ may only be reached if $|\ell| \leq n$ and $\ell \equiv n(2)$.

All results contained in this section are based on the generating function

$$\Psi_{h,m,\ell}(z) = \sum_{n \geq 0} N(h, m, \ell, n) z^n, \quad h, m > 0.$$

$N(h, m, \ell, n)$ counts the number of lattice paths starting at the origin and leading to the point (n, ℓ), where these paths are subject to the following restriction:

$$-m < y_0 = 0 < h, \ -m < y_1 < h, \ \ldots, \ -m < y_{n-1} < h, \ -m < y_n = \ell < h .$$

The above definition shows that the paths must lie entirely within the stripe defined by the lines $y = -m$ and $y = h$, where the paths are not even allowed to touch these lines.

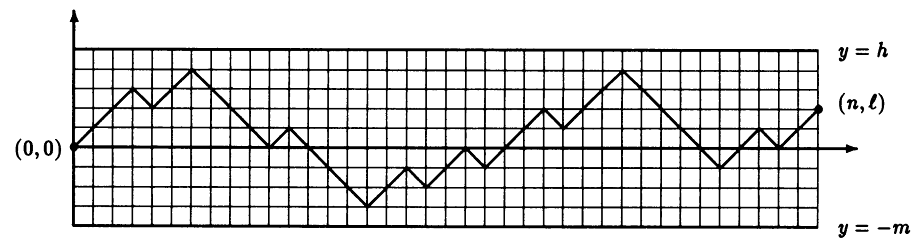


Figure 4.1: Example of a lattice path

Considering all those points from which the point (n, ℓ) can be reached by the next step, we are led to the following system of recurrence relations for Ψ:

$$\begin{pmatrix} 1 & -z & & & & & & & \\ -z & 1 & -z & & & & & & \\ & \ddots & \ddots & \ddots & & & & & \\ & & -z & 1 & -z & & & & \\ & & & -z & 1 & -z & & & \\ & & & & -z & 1 & -z & & \\ & & & & & \ddots & \ddots & \ddots & \\ & & & & & & -z & 1 & -z \\ & & & & & & & -z & 1 \end{pmatrix} \begin{pmatrix} \Psi_{h,m,h-1} \\ \Psi_{h,m,h-2} \\ \vdots \\ \Psi_{h,m,-1} \\ \Psi_{h,m,0} \\ \Psi_{h,m,1} \\ \vdots \\ \Psi_{h,m,-m+2} \\ \Psi_{h,m,-m+1} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} .$$

The matrix of this system has order $(h+m-1)\times(h+m-1)$ and will be denoted by $\mathbf{A}_{h+m}(z)$. It is not hard to find out that

$$|\mathbf{A}_k(z)| = \frac{1}{\sqrt{1-4z^2}}\left[\left(\frac{1+\sqrt{1-4z^2}}{2}\right)^k - \left(\frac{1-\sqrt{1-4z^2}}{2}\right)^k\right].$$

$|\mathbf{A}_k(z)|$ is a polynomial of degree $2\lceil\frac{k}{2}\rceil - 2$. Incidentally, $|\mathbf{A}_k(z)| = F_k(z^2)$, where $F_j(z)$ denotes the j-th Fibonacci polynomial [see Panny (1984)]. The Fibonacci polynomials are connected to the Fibonacci numbers F_j by $F_j(-1) = F_j$. Applying Cramer's rule, we get

$$\Psi_{h,m,\ell}(z) = \frac{|\mathbf{A}_{m-(|\ell|-\ell)/2}(z)|\,|\mathbf{A}_{h-(|\ell|+\ell)/2}(z)|}{|\mathbf{A}_{h+m}(z)|}\,z^{|\ell|}.$$

The substitution

$$z = g(v) = \frac{v}{1+v^2}, \tag{4.1}$$

is crucial for our approach because it considerably simplifies the original generating function, which now becomes

$$\Psi_{h,m,\ell}(z) = v^{|\ell|}\,\frac{1+v^2}{1-v^2}\,\frac{(1-v^{2m-(|\ell|-\ell)})(1-v^{2h-(|\ell|+\ell)})}{1-v^{2(h+m)}}. \tag{4.2}$$

The generating function also comprises the one-sided cases, viz.

$$\Psi_{\infty,m,\ell}(z) = v^{|\ell|}\left(1+v^2\right)\frac{1-v^{2m-(|\ell|-\ell)}}{1-v^2},$$

$$\Psi_{h,\infty,\ell}(z) = v^{|\ell|}\left(1+v^2\right)\frac{1-v^{2h-(|\ell|+\ell)}}{1-v^2},$$

and the unrestricted case

$$\Psi_{\infty,\infty,\ell}(z) = v^{|\ell|}\,\frac{1+v^2}{1-v^2}.$$

These generating functions may be taken as building blocks to derive appropriate generating functions for various path counting problems, as will be shown by the following example. As direct results, they furnish explicit expressions for the numbers $N(h,m,\ell,n)$ by applying Cauchy's integral formula

$$N(h,m,\ell,n) = \frac{1}{2\pi i}\oint\frac{\Psi_{h,m,\ell}(z)}{z^{n+1}}\,dz\,. \tag{4.3}$$

Since $v \approx z$ when $|z| \ll 1$, we may change variables in (4.3). From (4.1), we have

$$\frac{dz}{z^{n+1}} = \frac{g'(v)}{g^{n+1}(v)}\,dv = \frac{1-v^2}{v^{n+1}}\,(1+v^2)^{n-1}dv$$

and, hence,

$$N(h,m,\ell,n) = \frac{1}{2\pi i}\oint \frac{\Psi_{h,m,\ell}(g(v))}{g^{n+1}(v)}\, g'(v)\; dv.$$

Technically, this means that

$$N(h,m,\ell,n) = [v^n]\left\{v^{|\ell|}\left(1+v^2\right)^n \frac{\left(1-v^{2m-(|\ell|-\ell)}\right)\left(1-v^{2h-(|\ell|+\ell)}\right)}{1-v^{2(h+m)}}\right\},$$

where $[v^n]\,\{P(v)\}$ denotes the coefficient of v^n in $P(v)$. Consequently,

$$N(\,h\,,m,\ell,n) \;=\; \sum_{j=0,\pm1,\ldots}\left[\binom{n}{\frac{n+\ell}{2}+jd} - \binom{n}{\frac{n+\ell}{2}-h+jd}\right], \tag{4.4}$$

$$N(\infty,m,\ell,n) \;=\; \binom{n}{\frac{n+\ell}{2}} - \binom{n}{\frac{n+\ell}{2}+m}, \tag{4.5}$$

$$N(\,h\,,\infty,\ell,n) \;=\; \binom{n}{\frac{n+\ell}{2}} - \binom{n}{\frac{n+\ell}{2}-h}, \tag{4.6}$$

$$N(\infty,\infty,\ell,n) \;=\; \binom{n}{\frac{n+\ell}{2}}, \tag{4.7}$$

where $d = h+m$.

Of course, the formulas (4.4)–(4.7) are usually derived by path combinatorial arguments [see Mohanty (1979)]. In particular, (4.5) and (4.6) can be obtained using the method of reflections due to André. (4.4) may be found by applying this method repeatedly.

As a further remark, we would like to mention that (4.2) gives us the right clues on the location of the poles of $\Psi_{h,m,\ell}(z)$. Hence, (4.2) may be taken as a starting point for deriving the partial fraction expansion of $\Psi_{h,m,\ell}(z)$, which furnishes the following expression for $N(h,m,\ell,n)$ and $\ell \equiv n(2)$:

$$\frac{4}{h+m}\sum_{j=1}^{\lceil\frac{h+m}{2}\rceil-1} \sin h\theta_j\, \sin(h-\ell)\theta_j\,(2\cos\theta_j)^n,$$

where $\theta_j = j\pi/(h+m)$. Consequently, the asymptotic behavior as $n\to\infty$ can be described by

$$N(h,m,\ell,n) \sim \frac{4}{h+m}\sin\frac{h}{h+m}\pi\,\sin\frac{h-\ell}{h+m}\pi\left(2\cos\frac{\pi}{h+m}\right)^n.$$

Example 4.2.1 Let D_n^+ denote the maximum of the lattice path, i.e. $D_n^+ = \max\{y_0, y_1, \ldots, y_n\}$ and let Q_n denote the number of times that the maximum is reached. We are interested in the number $N(D_n^+ = k,\, Q_n = r)$ of paths with $D_n^+ = k$ and $Q_n = r$, where $y_0 = 0$ and $y_n = \ell$. Of course, one always has $k \geq 0$, $k \geq \ell$ and $r > 0$. Now, we will illustrate the usefulness of the generating functions $\Psi_{h,m,\ell}(z)$ in such counting problems.

Let us first consider the case $k > 0, k > \ell$. The following figure shows an appropriately decomposed path (with $Q_n = 3$).

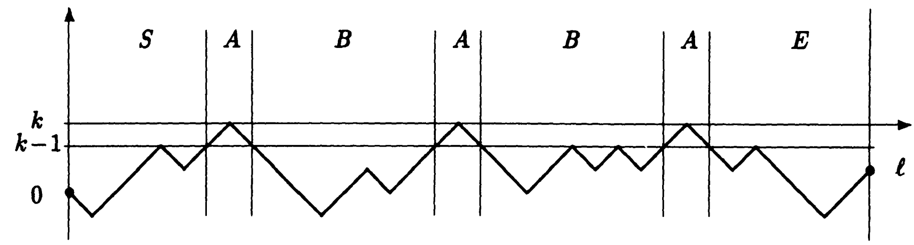


Figure 4.2: Number of times where the maximum is achieved

A path with $D_n^+ = k$, $Q_n = r$ can symbolically be written as $SA(BA)^{r-1}E$. The generating functions for the individual segments are

$$\begin{aligned}
\Upsilon_S(z) &= \Psi_{k,\infty,k-1}(z) = (1+v^2)v^{k-1}, \\
\Upsilon_A(z) &= z^2 = \frac{v^2}{(1+v^2)^2}, \\
\Upsilon_B(z) &= \Psi_{1,\infty,0}(z) = 1+v^2, \\
\Upsilon_E(z) &= \Psi_{1,\infty,\ell-k+1}(z) = (1+v^2)v^{k-1-\ell}.
\end{aligned}$$

Hence, the generating function $\phi_{k,\ell,r}(z)$ for $N(D_n^+ = k,\ Q_n = r)$, $k > 0$, $\ell < k$, is

$$\Upsilon_S(z)\Upsilon_A^r(z)\Upsilon_B^{r-1}(z)\Upsilon_E(z),$$

which yields

$$\phi_{k,\ell,r}(z) = \left(1+v^2\right)^{1-r} v^{2(r+k-1)-\ell}.$$

The remaining three cases are $\ell < k = 0$, $\ell = k > 0$, and $\ell = k = 0$. These cases can be investigated in the same way as the first case. It turns out that the generating function $\phi_{k,\ell,r}(z)$ in fact applies for all cases, i.e. for $k \geq 0$, $\ell \leq k$. The coefficient of z^n can be most conveniently extracted by means of Cauchy's integral formula. We only have to take into account that

$$\frac{dz}{z^{n+1}} = \frac{g'(z)}{g^{n+1}(v)}\, dv = \frac{(1-v^2)}{v^{n+1}}\left(1+v^2\right)^{n-1} dv.$$

Consequently, $N(D_n^+ = k,\ Q_n = r)$ can be expressed as

$$\frac{1}{2\pi i}\oint \frac{v^{2(r+k-1)-\ell}}{v^{n+1}}\left(1-v^2\right)\left(1+v^2\right)^{n-r} dv\,.$$

Determining the coefficient of v^n in $v^{2(r+k-1)-\ell}(1-v^2)(1+v^2)^{n-r}$ yields

$$N(D_n^+ = k,\ Q_n = r) = \binom{n-r}{\frac{n+\ell}{2}-k-r+1} - \binom{n-r}{\frac{n+\ell}{2}-k-r} \qquad (4.8)$$

or, equivalently,

$$N(D_n^+ > k,\, Q_n = r) = \binom{n-r}{\frac{n+\ell}{2} - k - r}\,. \tag{4.9}$$

Summation over all possible values of k (i.e., over all $k \geq \max\{0, \ell\}$) leads to [Mohanty (1979, p. 93)]:

$$N(Q_n = r) = \binom{n-r}{\frac{n+|\ell|}{2} - 1}\,. \tag{4.10}$$

The above results are comparatively easily obtained by our approach, since the generating functions become rather simple owing to the substitution $z = g(v)$ and because concatenation and summation is implicitly done by power series algebra. Since the substitution $z = g(v)$ is compatible with Cauchy's integral formula, we still may apply it to determine the coefficients of interest.

Katzenbeisser and Panny (1996) have shown that all results on rank order statistics given by Dwass (1967) can be obtained by this approach as well. Moreover, all of Dwass's results have been extended to arbitrary endpoints (n, ℓ), enabling one to deal with rank order statistics for unequal sample sizes also.

4.3 Simple Random Walks

Let X_k, $k = 1, 2, \ldots$, be independent and identically distributed random variables with

$$\Pr[X_k = 1] = \alpha, \quad \Pr[X_k = 0] = \beta, \quad \Pr[X_k = -1] = \gamma,$$

where $\alpha + \beta + \gamma = 1$. Consider the random walk

$$S_k = S_0 + \sum_{j=1}^{k} \mathbf{X}_j, \quad k = 1, 2, \ldots, n, \quad \text{with} \quad S_n = \ell\,,$$

i.e., a simple random walk in the sense of Cox and Miller (1965) starting at S_0 and leading to ℓ after n steps. Confining to $S_0 = 0$ actually constitutes no restriction at all. So, this assumption will be made in the sequel if not explicitly stated otherwise. In this section, $\Psi_{h,m,\ell}(z)$ denotes the probability generating function

$$\Psi_{h,m,\ell}(z) = \sum_{n \geq 0} p(h, m, \ell, n) z^n, \quad h, m > 0,$$

where $p(h, m, \ell, n)$ gives the probability that a particle obeying a random walk with absorbing barriers at $-m$ and h reaches the state ℓ when it started from state 0, i.e.,

$$p(h, m, \ell, n) = \Pr[-m < S_1 < h, \dots, -m < S_{n-1} < h, -m < S_n = \ell < h \mid S_0 = 0].$$

The definition of $p(h, m, \ell, n)$ shows that it is not even admissible to touch the barriers. Following Barton and Mallows (1965), this type of absorption could be termed as *strong sense absorption.*

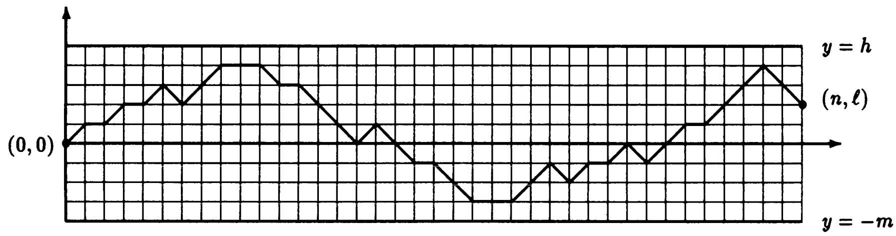


Figure 4.3: Sample path of a simple random walk

In the following, we show how the approach adopted earleir for path counting can be generalized to determine the probabilities $p(h, m, \ell, n)$. It is not hard to see [Panny (1984) and Katzenbeisser and Panny (1984)] that the matrix of the system of recurrence relations for Ψ becomes

$$\mathbf{A}_{h+m}(z) = \begin{pmatrix} (1-\beta z) & -\alpha z & & & & \\ -\gamma z & (1-\beta z) & -\alpha z & & & \\ & -\gamma z & (1-\beta z) & -\alpha z & & \\ & & \ddots & \ddots & \ddots & \\ & & & -\gamma z & (1-\beta z) & -\alpha z \\ & & & & -\gamma z & (1-\beta z) \end{pmatrix}$$

in the present setting. As before, $\mathbf{A}_k(z)$ is of order $(k-1) \times (k-1)$. The determinant is

$$|\mathbf{A}_k(z)| = \frac{1}{b}\left[\left(\frac{a+b}{2}\right)^k - \left(\frac{a-b}{2}\right)^k\right],$$

where $a = 1 - \beta z$ and $b = \sqrt{a^2 - 4\alpha\gamma z^2}$. Again, $|\mathbf{A}_k(z)|$ can be expressed by means of the Fibonacci polynomials as

$$|\mathbf{A}_k(z)| = (1-\beta z)^{k-1} F_k\left(\frac{\alpha\gamma z^2}{(1-\beta z)^2}\right).$$

By Cramer's rule, we get

$$\Psi_{h,m,\ell}(z) = (\alpha\gamma)^{\frac{|\ell|}{2}} \rho^{\frac{\ell}{2}} z^{|\ell|} \frac{|\mathbf{A}_{m-(|\ell|-\ell)/2}(z)|\,|\mathbf{A}_{h-(|\ell|+\ell)/2}(z)|}{|\mathbf{A}_{h+m}(z)|}. \tag{4.11}$$

The substitution

$$z = g(v) = \frac{v}{\alpha v^2 + \beta v + \gamma}$$

is now the counterpart of (4.1). Applying this substitution to (4.11) again results in a considerable simplification as

$$\Psi_{h,m,\ell}(z) = \frac{\rho^{\frac{|\ell|+\ell}{2}}}{\gamma}\, v^{|\ell|}\, \frac{\alpha v^2+\beta v+\gamma}{1-\rho v^2}\, \frac{(1-(\rho v^2)^{m-\frac{|\ell|-\ell}{2}})(1-(\rho v^2)^{h-\frac{|\ell|+\ell}{2}})}{1-(\rho v^2)^{h+m}}. \qquad (4.12)$$

As before, the generating function also comprises the one-sided cases, viz.

$$\Psi_{\infty,m,\ell}(z) = \frac{\rho^{\frac{|\ell|+\ell}{2}}}{\gamma}\, v^{|\ell|}(\alpha v^2+\beta v+\gamma)\, \frac{1-(\rho v^2)^{m-\frac{|\ell|-\ell}{2}}}{1-\rho v^2},$$

$$\Psi_{h,\infty,\ell}(z) = \frac{\rho^{\frac{|\ell|+\ell}{2}}}{\gamma}\, v^{|\ell|}(\alpha v^2+\beta v+\gamma)\, \frac{1-(\rho v^2)^{h-\frac{|\ell|+\ell}{2}}}{1-\rho v^2},$$

and the unrestricted case

$$\Psi_{\infty,\infty,\ell}(z) = \frac{\rho^{\frac{|\ell|+\ell}{2}}}{\gamma}\, v^{|\ell|}\, \frac{\alpha v^2+\beta v+\gamma}{1-\rho v^2}.$$

Again, these generating functions prove very useful in deriving appropriate generating functions for more intricate problems connected with simple random walks, as shall be illustrated by the following example. As direct results, of course, they furnish explicit expressions for the probabilities $p(h, m, \ell, n)$ by an application of Cauchy's integral formula as

$$p(h,m,\ell,n) = \frac{1}{2\pi i}\oint \frac{\Psi_{h,m,\ell}(z)}{z^{n+1}}\, dz. \qquad (4.13)$$

Since $v \approx z$ when $|z| \ll 1$, we may change variables in (4.13). Since

$$\frac{dz}{z^{n+1}} = \frac{g'(v)}{g^{n+1}(v)} = \gamma\, \frac{1-\rho v^2}{v^{n+1}}\, (\alpha v^2+\beta v+\gamma)^{n-1},$$

we have

$$p(h,m,\ell,n) = \frac{1}{2\pi i}\oint \frac{\Psi_{h,m,\ell}(g(v))}{g^{n+1}(v)}\, g'(v)\, dv.$$

Technically, this means that $p(h, m, \ell, n)$ is

$$[v^n]\left\{ \rho^{\frac{|\ell|+\ell}{2}} v^{|\ell|}(\alpha v^2+\beta v+\gamma)^n \frac{\left(1-(\rho v^2)^{m-\frac{|\ell|-\ell}{2}}\right)\left(1-(\rho v^2)^{h-\frac{|\ell|+\ell}{2}}\right)}{1-(\rho v^2)^{h+m}} \right\}.$$

Consequently,

$$p(h,m,\ell,n) = \sum_{j=0,\pm1,\ldots} \rho^{-jd}\left[\binom{n;\alpha,\beta,\gamma}{n+\ell+2jd} - \rho^{h}\binom{n;\alpha,\beta,\gamma}{n+\ell-2h+2jd}\right], \tag{4.14}$$

$$p(\infty,m,\ell,n) = \binom{n;\alpha,\beta,\gamma}{n+\ell} - \rho^{-m}\binom{n;\alpha,\beta,\gamma}{n+\ell+2m}, \tag{4.15}$$

$$p(h,\infty,\ell,n) = \binom{n;\alpha,\beta,\gamma}{n+\ell} - \rho^{h}\binom{n;\alpha,\beta,\gamma}{n+\ell-2h}, \tag{4.16}$$

$$p(\infty,\infty,\ell,n) = \binom{n;\alpha,\beta,\gamma}{n+\ell}, \tag{4.17}$$

where $d = h + m$ and where generalized trinomial coefficients (GTC) are used. They have the generating function $(\alpha v^2+\beta v+\gamma)^n$, i.e.

$$\binom{n;\alpha,\beta,\gamma}{k} = [v^k](\alpha v^2+\beta v+\gamma)^n,$$

which, of course, entails

$$\sum_{k=-n}^{n} \binom{n;\alpha,\beta,\gamma}{n+k} = 1. \tag{4.18}$$

of course. GTC are quasi-symmetric, i.e.,

$$\binom{n;\alpha,\beta,\gamma}{n-k} = \rho^{-k}\binom{n;\alpha,\beta,\gamma}{n+k} \tag{4.19}$$

and comprise binomial coefficients as a special case, viz.

$$\binom{n;1/2,0,1/2}{2m} = \binom{n}{m}2^{-n}. \tag{4.20}$$

They are connected to ordinary trinomial coefficients by the relation

$$\binom{n;\alpha,\beta,\gamma}{n+k} = \sum_{\substack{a,b,c\geq 0\\ a+b+c=n\\ a-c=k}} \binom{n}{a,b,c}\alpha^a\beta^b\gamma^c,$$

which allows the following representation as a hypergeometric function:

$$\binom{n;\alpha,\beta,\gamma}{n+k} = \alpha^k\beta^{n-k}\binom{n}{k}\,{}_2F_1\Big(-\frac{n-k}{2},-\frac{n-k-1}{2};k+1;\frac{4\alpha\gamma}{\beta^2}\Big).$$

An integral representation is

$$\binom{n;\alpha,\beta,\gamma}{n+k} = \frac{(\alpha/\gamma)^{k/2}}{\pi}\int_0^\pi \cos k\theta\,(\beta+2\sqrt{\alpha\gamma}\cos\theta)^n\,d\theta\,, \tag{4.21}$$

which corresponds to the well-known integral representation of binomial coefficients [see Gradshteyn and Ryzhik (1980, p. 374)]:

$$\binom{n}{\frac{n+k}{2}} = 2^n\frac{1}{\pi}\int_0^\pi \cos k\theta\,(\cos\theta)^n\,d\theta\,.$$

Both integral representations can easily be verified by means of the residue theorem.

We would like to mention that the formulas (4.14)–(4.17) can all be obtained by reflection arguments as well. At first sight, one would think that these arguments are not applicable to the problem treated here, since the paths are no longer symmetrical. However, the classical reflection approach can be modified by interchanging the probabilities α and γ in the reflected parts of the paths [see Panny (1984)].

In accordance with the preceding section, we would like to mention here an alternative representation of $p(h,m,\ell,n)$, based on partial fraction expansion of $\Psi_{h,m,\ell}(z)$, which can conveniently be derived from (4.12), as

$$p(h,m,\ell,n) = \frac{2\rho^{\frac{\ell}{2}}}{h+m}\sum_{j=1}^{h+m-1} \sin h\theta_j\,\sin(h-\ell)\theta_j\,(\beta+2\sqrt{\alpha\gamma}\cos\theta_j)^n,$$

where $\theta_j = j\pi/(h+m)$. Hence, the asymptotic behavior of $p(h,m,\ell,n)$ as $n\to\infty$ can be described by

$$p(h,m,\ell,n) \sim \frac{2\Delta(\beta)\rho^{\frac{\ell}{2}}}{h+m}\sin\frac{h}{h+m}\pi\,\sin\frac{h-\ell}{h+m}\pi\left(\beta+2\sqrt{\alpha\gamma}\cos\frac{\pi}{h+m}\right)^n,$$

where the indicator function

$$\Delta(\beta) = \begin{cases} 1 & \beta>0 \\ 2 & \beta=0 \end{cases}$$

is necessary, since for the case $\beta=0$ the contribution of $\theta_{h+m-1} = \pi - \frac{\pi}{h+m}$ has to be taken into account, too.

Example 4.3.1 In the following, we generalize Example 4.2.1 for simple random walks. Let D_n^+ denote the maximum of the random walk, i.e., $D_n^+ = \max_{0\le k\le n}\{S_k\}$ and let Q_n denote the number of times that the maximum is reached. We are interested in the probability $\Pr[D_n^+ = k,\ Q_n = r]$ for a random walk with $S_0 = 0$ and $S_n = \ell$. Of course, one always has $k \ge 0$, $k \ge \ell$ and $r > 0$. However, regarding the possibility of horizontal steps, the definition of Q_n must be properly adapted: The maximum is achieved if $S_k = S_{k+1} = S_{k+2} = \ldots = S_{k+m} = D_n^+$ and $S_{k-1}, S_{k+m+1} < D_n^+$, $0 \le k \le k+m \le n$. If there should be one or more consecutive horizontal steps coinciding with the line $y = D_n^+$ (i.e. $m > 0$), this counts only as a single maximum. By definition, if $S_0 = D_n^+$ the path starts with a maximum; similarly, if $S_n = D_n^+$ the path ends with a maximum.

Let us first consider the case $k > 0$, $k > \ell$. The following figure shows an appropriately decomposed sample path (with $Q_n = 3$).

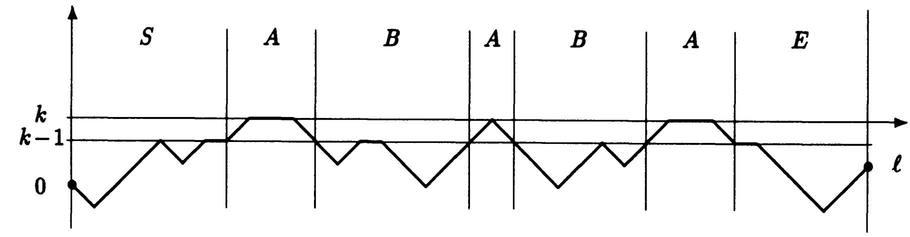


Figure 4.4: Number of times where the maximum is achieved

A path corresponding to the event $D_n^+ = k$, $Q_n = r$ can symbolically be written as $SA(BA)^{r-1}E$. The probability generating function for the individual segments are

$$\begin{aligned} \Upsilon_S(z) &= \Psi_{k,\infty,k-1}(z), \\ \Upsilon_A(z) &= \alpha\gamma z^2\, \Psi_{1,1,0}(z), \\ \Upsilon_B(z) &= \Psi_{1,\infty,0}(z), \\ \Upsilon_E(z) &= \Psi_{1,\infty,\ell-k+1}(z). \end{aligned}$$

Hence, the probability generating function $\phi_{k,\ell,r}(z)$ for $\Pr[D_n^+ = k,\ Q_n = r]$, $k > 0$, $k > \ell$, is

$$\Upsilon_S(z)\Upsilon_A^r(z)\Upsilon_B^{r-1}(z)\Upsilon_E(z),$$

which yields

$$\phi_{k,\ell,r}(z) = \frac{1}{\gamma}\,\frac{(\alpha v^2 + \beta v + \gamma)\,(\rho v^2)^{r+k-1}}{v^\ell\,(1+\rho v^2)^r}.$$

A separate investigation of the remaining cases shows that the last formula in fact covers all cases, i.e., $k \ge 0$, $\ell \le k$. Extracting

$$[v^n]\left\{\frac{(\alpha v^2 + \beta v + \gamma)^n\,(1-\rho v^2)\,(\rho v^2)^{r+k-1}}{v^\ell\,(1+\rho v^2)^r}\right\}$$

yields the following expression for $\Pr[D_n^+ = k, Q_n = r]$:

$$\rho^{k+r-1}\sum_{j\ge 0}\rho^j\binom{-r}{j}\left[\binom{n;\alpha,\beta,\gamma}{n+\ell-2k-2r-2j+2} - \rho\binom{n;\alpha,\beta,\gamma}{n+\ell-2k-2r-2j}\right], \quad (4.22)$$

or, equivalently,

$$\Pr[D_n^+ > k, Q_n = r] = \rho^{k+r} \sum_{j\geq 0} \rho^j \binom{-r}{j} \binom{n;\alpha,\beta,\gamma}{n+\ell-2k-2r-2j}. \qquad (4.23)$$

Since the range of all possible values of k is given by $k \geq \max\{0, \ell\}$, the last formula yields at once

$$\Pr[Q_n = r] = \rho^{r-1+\frac{|\ell|+\ell}{2}} \sum_{j\geq 0} \rho^j \binom{-r}{j} \binom{n;\alpha,\beta,\gamma}{n-|\ell|-2r-2j+2}. \qquad (4.24)$$

Clearly, formulae (4.22), (4.23), and (4.24) translate to the corresponding formulae (4.8), (4.9), and (4.10) after an application of the identity in (4.20) and Vandermonde's convolution formula. The last two formulae also comprise the pertaining results [viz. VIII(a) and VIII(b)] due to Dwass (1967) as special cases. This can be checked by specializing on $\alpha = \gamma = 1/2$, $\ell = 0$, substituting $2n$ for n and dividing by the probability of the conditioning event, viz. $\Pr[S_{2n} = 0 \mid S_0 = 0] = 2^{-2n}\binom{2n}{n}$.

It has been shown by Katzenbeisser and Panny (1996) that this method allows to generalize all of Dwass's results by considering arbitrary endpoints (n, ℓ), introducing horizontal steps as third step type and assigning arbitrary probabilities α, β, γ to the three step types. Regarding these extensions, Dwass's rank order statistics in fact are extended to *simple random walk statistics.*

4.4 Randomized Random Walks

In the present section, it will be shown as to how the results on simple random walks can be translated to randomized random walks by means of a limiting process. In the following, our randomization approach will be put forward. However, we confine ourselves to a presentation of the basic ideas and skim over the details. The reader interested in a rigorous proof is referred to the papers Böhm and Mohanty (1994) and Böhm and Panny (1996).

Usually, the x-axis of a random walk is interpreted as time. Accordingly, the simple random walk corresponds to discrete time. This may be visualized by dividing a time interval of lenght $t = 1$, say, in n time slots, each of which has length $1/n$. Each slot j has an associated random variable X_j (and S_j). If $X_j = 1$ or $X_j = -1$, we have a jump (up or down). Of course, the number of jumps follows a binomial distribution given by

$$\Pr[\text{number of jumps} = k] = \binom{n}{k} (\alpha+\gamma)^k (1-\alpha-\gamma)^{n-k}.$$

It has been shown in the preceding section that for an unrestricted simple random walk, we have

$$\Pr[S_n = k] = \binom{n; \alpha, 1-\alpha-\gamma, \gamma}{n+k}.$$

Let us now consider the limiting process $n \to \infty$, $\alpha \to 0$, $\gamma \to 0$, where the expectation of the number of jumps within the interval of length t is kept constant, i.e., $\alpha n = \lambda t$ and $\gamma n = \mu t$. Intuitively, this means that the division of the interval becomes finer and finer, whereby the proportion of time slots with no jumps tends to 1. In the limit, we arrive at

$$\begin{aligned} \Pr[\text{number of jumps} = k] &= \lim \binom{n}{k} (\alpha+\gamma)^k (1-\alpha-\gamma)^{n-k} \\ &= \frac{(\lambda+\mu)^k\, t^k}{k!}\, e^{-(\lambda+\mu)t} \end{aligned}$$

or equivalently

$$\Pr[\text{time between two consecutive jumps} \le t] = 1 - \frac{1}{\lambda+\mu} e^{-(\lambda+\mu)t}.$$

In other words, this limiting process takes us from a simple random walk S_k in discrete time $k = 0, 1, \ldots, n$ to a randomized random walk $S(t)$ in continuous time $t \ge 0$, where the term *randomized random walk* is used in the sense of Feller (1971, p. 58).

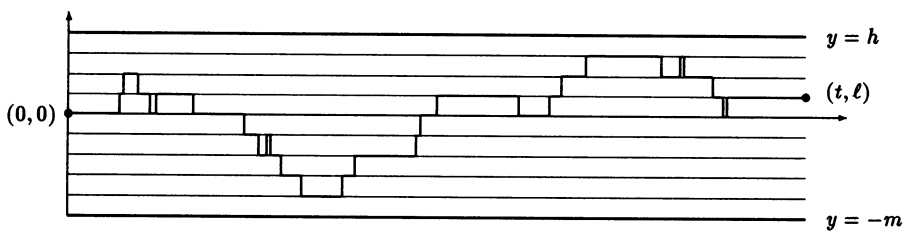


Figure 4.5: Sample path of a randomized random walk

It is well-known that for a randomized random walk, we have

$$\Pr[S(t) = k] = \rho^{k/2} e^{-(\lambda+\mu)t} I_k(2t\sqrt{\lambda\mu}),$$

where $I_k(x)$ denotes the modified Bessel-function of order k and $\rho = \alpha/\gamma = \lambda/\mu$. This suggests us to apply the same limiting process to the generalized trinomial coefficients. It turns out that this limit is well-defined and, in fact, we get

$$\lim_{\substack{n\to\infty \\ \alpha\to 0, \alpha n = \lambda t \\ \gamma \to 0, \gamma n = \mu t}} \binom{n; \alpha, 1-\alpha-\gamma, \gamma}{n+k} = \rho^{k/2} e^{-(\lambda+\mu)t} I_k(2t\sqrt{\lambda\mu}). \tag{4.25}$$

This relation is most conveniently established by using the integral representation (4.21) for the generalized trinomial coefficients given by

$$\binom{n;\alpha,1-\alpha-\gamma,\gamma}{n+k} = \frac{(\alpha/\gamma)^{k/2}}{\pi}\int_0^\pi \cos k\theta\,(1-\alpha-\gamma+2\sqrt{\alpha\gamma}\cos\theta)^n\,d\theta\,.$$

The sequence of functions $f_n(\theta) = (\cos k\theta)(1-\alpha-\gamma+2\sqrt{\alpha\gamma}\cos\theta)^n$ converges uniformly for all real θ. Hence, the order of taking the limit and integration may be interchanged. Putting $\alpha = \lambda t/n$ and $\gamma = \mu t/n$ results in

$$1-\alpha-\gamma+2\sqrt{\alpha\gamma}\cos\theta = 1+\frac{-(\lambda+\mu)t+2t\sqrt{\lambda\mu}\cos\theta}{n},$$

which eliminates the dependent variables. Since

$$\lim_{n\to\infty}\left(1+\frac{-(\lambda+\mu)t+2t\sqrt{\lambda\mu}\cos\theta}{n}\right)^n = e^{-(\lambda+\mu)t}e^{2t\sqrt{\lambda\mu}\cos\theta},$$

we have to determine

$$\frac{\rho^{k/2}e^{-(\lambda+\mu)t}}{\pi}\int_0^\pi (\cos k\theta)e^{2t\sqrt{\lambda\mu}\cos\theta}\,d\theta.$$

But, it is well-known that the modified Bessel function has the integral representation [Spanier and Oldham (1987, p. 481)]

$$I_k(z) = \frac{1}{\pi}\int_0^\pi (\cos k\theta)e^{z\cos\theta}\,d\theta.$$

The above proof has first been given by Mohanty and Panny (1990). An alternative proof of (4.25) based on the Taylor series expansion of the modified Bessel function has been given by Mohanty and Panny (1989).

It should be noted that (4.25) also covers the cases $\lambda = 0$ or $\mu = 0$. This can be shown by considering the asymptotic behavior of $I_k(z)$ as $z \to 0$, viz.

$$I_k(z) \sim \frac{(z/2)^k}{k!} \quad \text{as} \quad z\to 0$$

[Spanier and Oldham (1987, p. 495)]. If $\mu \to 0$, the right-hand side of (4.25) consequently equals

$$\begin{cases} \dfrac{(\lambda t)^k}{k!}e^{-\lambda t} & k \ge 0 \\ 0 & k < 0\,. \end{cases}$$

Correspondingly, if $\lambda \to 0$, the right-hand side of (4.25) becomes

$$\begin{cases} \dfrac{(\mu t)^{|k|}}{|k|!}e^{-\mu t} & k \le 0 \\ 0 & k > 0\,. \end{cases}$$

Also, the quasi-symmetry property (4.19) of the GTC is reflected by the modified Bessel functions, since $I_{-k}(z) = I_k(z)$ whenever the order k is an integer. Moreover, it is well-known [see Spanier and Oldham (1987, p. 479)] that

$$e^{-(\lambda+\mu)t} \sum_{k=-\infty}^{+\infty} \rho^{k/2} I_k\left(2t\sqrt{\lambda\mu}\right) = 1\,,$$

which corresponds to property (4.18).

Hence, the above limiting process in fact constitutes an alternative approach to Feller's randomization technique. A nice point about this approach is that the discrete time results can be taken as a starting point to derive their continuous time counterparts through a limiting process. In particular, the generalized trinomial coefficients make it possible to express the discrete time results in quite a natural and simple way. Hence, Eq. (4.25) is fundamental for our randomization approach since it links together the discrete and continuous time results and in many cases allows to derive the continuous time results by more or less mechanically translating the pertaining discrete time results, as will be shown in the following examples. Proceeding this way, we perfectly conform to a suggestion expressed by Meisling (1958) in his last remark: "It is even conceivable that some continuous-time problems could be solved more simply by first considering the discrete-time case and then obtaining the continuous-time result by a limiting process." Recently, some interesting problems in queueing theory have been solved by adopting similar approaches; see, for example, Mohanty and Panny (1989, 1990), Böhm and Mohanty (1990, 1993, 1994), Kanwar Sen, Jain and Gupta (1993), Jain and Gupta (1993), and Mohanty, Parthasarathy and Sharaf Ali (1990).

Example 4.4.1 In the following, we want to extend Example 4.3.1 for randomized random walks. We first have to adapt the definitions of D^+ and Q in the following way: Let D_t^+ denote the maximum of the randomized random walk, i.e., $D_t^+ = \max_{0\le\tau\le t}\{S(\tau)\}$ and let Q_t denote the number of times that the maximum is reached. The maximum is achieved if there is an interval $I = [a, b)$, $0 \le a < b \le t$, such that $S(I) = D_t^+$ and the interval is maximal with respect to this property. Accordingly, Q_t counts the number of such intervals in $[0, t]$. The following figure shows a sample path with $Q_t = 3$.

We are interested in the probability $\Pr[D_t^+ > k,\, Q_t = r]$ for a randomized random walk with $S(0) = 0$ and $S(t) = \ell$. As before, the possible cases are characterized by $k \ge 0$, $k \ge \ell$ and $r > 0$. The corresponding result (4.23) for simple random walks reads

$$\Pr[D_n^+ > k, Q_n = r] = \rho^{k+r} \sum_{j\ge 0} \rho^j \binom{-r}{j} \binom{n;\alpha,\beta,\gamma}{n+\ell-2k-2r-2j}.$$

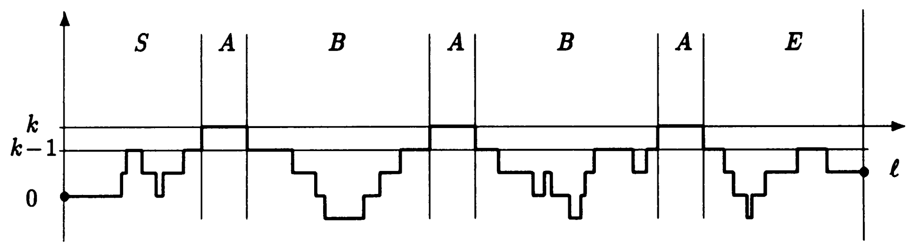


Figure 4.6: Number of times where the maximum is achieved

Applying the limiting process in this case boils down to an application of (4.25), which furnishes at once

$$\Pr[D_t^+ > k, Q_t = r] = \rho^{\ell/2} e^{-(\lambda+\mu)t} \sum_{j \geq 0} \binom{-r}{j} I_{\ell-2k-2r-2j}(2t\sqrt{\lambda\mu}).$$

The same can be done for $\Pr[Q = r]$. In the preceding section, we have derived the discrete time result [*cf.* (4.24)]

$$\Pr[Q_n = r] = \rho^{r-1+\frac{|\ell|+\ell}{2}} \sum_{j \geq 0} \rho^j \binom{-r}{j} \binom{n; \alpha, \beta, \gamma}{n - |\ell| - 2r - 2j + 2}$$

which, through an application of (4.25), translates to

$$\Pr[Q_t = r] = \rho^{\ell/2} e^{-(\lambda+\mu)t} \sum_{j \geq 0} \binom{-r}{j} I_{|\ell|+2r+2j-2}(2t\sqrt{\lambda\mu}).$$

The usefulness of this approach has been demonstrated by Böhm and Panny (1996) by considering various statistics for randomized random walks and by deriving the pertaining distributional results by means of the above randomization procedure.

References

1. Barton, D. E. and Mallows, C. L. (1965). Some aspects of the random sequence, *Annals of Mathematical Statistics*, **36**, 236–260.

2. Böhm, W. and Mohanty, S. G. (1990). Discretization of Markovian queueing systems, *Technical Report*, Department of Mathematical Statistics, University of Economics, Vienna.

3. Böhm, W. and Mohanty, S. G. (1993). The transient solution of M/M/1 queues under (M,N)-policy. A combinatorial approach, *Journal of Statistical Planning and Inference*, **34**, 23–33.

4. Böhm, W. and Mohanty, S. G. (1994). On discrete time Markovian N-policy queues involving batches, *Sankhyā, Series A*, **56**, 1–20.

5. Böhm, W. and Panny, W. (1996). Simple random walk statistics. Part II: Continuous time results, *Journal of Applied Probability*, **33**, 331–339.

6. Cox, D. R. and Miller, H. D. (1965). *The Theory of Stochastic Processes*, London: Methuen.

7. Dwass, M. (1967). Simple random walk and rank order statistics, *Annals of Mathematical Statistics*, **38**, 1042–1053.

8. Feller, W. (1971). *An Introduction to Probability Theory and its Applications*, Vol. 2, Second edition, New York: John Wiley & Sons.

9. Gradshteyn, I. S. and Ryzhik, I. M. (1980). *Table of Integrals, Series, and Products*, Orlando, FL: Academic Press.

10. Kanwar Sen, Jain, J. L. and Gupta, J. M. (1993). Lattice path approach to transient solution of M/M/1 with (0,K) control policy, *Journal of Statistical Planning and Inference*, **34**, 259–268.

11. Katzenbeisser, W. and Panny, W. (1984). Asymptotic results on the maximal deviation of simple random walks, *Stochastic Processes and Their Applications*, **18**, 263–275.

12. Katzenbeisser, W. and Panny, W. (1996). Simple random walk statistics. Part I: Discrete time results, *Journal of Applied Probability*, **33**, 311–330.

13. Jain, J. L. and Gupta, J. M. (1993). Combinatorial approach to Markovian queues, *Journal of Statistical Planning and Inference*, **34**, 269–280.

14. Meisling, T. (1958). Discrete-time queueing theory, *Journal of the Operations Research Society of America*, **6**, 96–105.

15. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

16. Mohanty, S. G. and Panny, W. (1989). A discrete-time analogue of the M/M/1 queue and the transient solution: An analytic approach, In *Limit Theorems in Probability and Statistics*, Pécs (Hungary), *Colloqu. Math. Society Janos Bólyai*, **57**, 417–424.

17. Mohanty, S. G. and Panny, W. (1990). A discrete-time analogue of the M/M/1 queue and the transient solution: A geometric approach, *Sankhyā, Series A*, **52**, 364–370.

18. Mohanty, S. G., Parthasarathy, P. R. and Sharaf Ali, M. (1990). On the transient solution of a discrete queue, *Technical Report*, Department of Mathematics and Statistics, McMaster University, Hamilton.

19. Panny, W. (1984). *The Maximal Deviation of Lattice Paths*, Athenäum/ Hain/DiscretionaryHanstein, Königstein/Ts.

20. Spanier, J. and Oldham, K. B. (1987). *An Atlas of Functions*, Berlin: Springer-Verlag.

5

Combinatorial Identities: A Generalization of Dougall's Identity

Erik Sparre Andersen and Mogens Esrom Larsen

University of Copenhagen, Denmark

Abstract: In this paper, I will discuss combinatorial identities as a tool for individuals working with combinatorial problems. I will also present a generalization of Dougall's (1907) identity. In the notations of this paper, the general combinatorial identity established is the following:

If $p = b + c + d + e - n + 1$ is a non-negative integer then we have:

$$\begin{aligned}
&\sum_{k=0}^{n} \binom{n}{k} [n+2a]_k [b+a]_k [c+a]_k [d+a]_k \\
&\qquad \times [n-2a]_{n-k} [b-a]_{n-k} [c-a]_{n-k} [d-a]_{n-k} [e-a]_{n-k} (n+2a-2k) \\
&= (-1)^n [n+2a]_{2n+1} [b+d-p]_{n-p} [b+e-p]_{n-p} [d+e-p]_{n-p} \\
&\qquad \times \sum_{j=0}^{p} \binom{p}{j} [n]_j [c+a]_j [c-a]_j [b+e-n]_{p-j} [d+e-n]_{p-j} [d+e-n]_{p-j},
\end{aligned}$$

where $[x]_k$ denotes the descending factorial. Dougall's identity, which is usually written in terms of a hypergeometric series, corresponds to the case $p = 0$.

Keywords and phrases: Combinatorial identities, Dougall's identity, Zeilberger's algorithm, Pfaff-Saalschütz's identity

5.1 Introduction

We shall discuss combinatorial identities as a tool for individuals working on combinatorial problems. It is important to mention here that we are well aware of Zeilberger's algorithm for proving combinatorial identities for sums of hypergeometric type. The generalization of Dougall's (1907) identity, which we shall prove, however, cannot be proved using Zeilberger's algorithm. This may change in the future, but the version described by Petkovšek, Wilf and Zeilberger (1996) does not work in this case.

Combinatorial sums may be written in many ways. It is therefore important to define – if possible – a standard form of a sum. This is not always possible, but for sums of hypergeometric type it is possible. The quotient between two consecutive terms of a hypergeometric sum may be written as the quotient between two polynomials in the summation variable k. This has resulted in the adoption of finite hypergeometric series as a standard form. We prefer to write finite hypergeometric combinatorial sums using descending factorials

$$[x]_n = x(x-1)...(x-n+1).$$

If the sum has arbitrary natural limits m and n and the quotient between the $(k+1)$th term and kth term has the form

$$\frac{(n-k)(a_1+m-k)...(a_{p-1}+m-k)}{(m-1-k)(n-1-b_1-k)...(n-1-b_{q-1}-k)}\, z, \tag{5.1}$$

then we use as standard form of the sum:

$$\sum_{k=m}^{n} \binom{n-m}{k-m} [a_1]_{k-m} \cdots [a_{p-1}]_{k-m} [b_1]_{n-k} \cdots [b_{q-1}]_{n-k} (-1)^{q(k-m)} z^{k-m}. \tag{5.2}$$

If a combinatorial sum has (5.2) as standard form, we shall say that it is of *Type* $II(p,q,z)N$, where II represents hypergeometric sums and N stands for natural limits. The descending factorial is more natural for combinatorial problems than the ascending factorial used in connection with hypergeometric series. It is also an advantage to avoid division. It may be mentioned here that all the results in this paper are valid (and with essentially the same proof) if the parameters in (5.2),

$$a_1, ..., a_{p-1}, b_1, ..., b_{q-1}, z, \tag{5.3}$$

all belong to a commutative ring which contains the natural numbers as elements.

Except for rearrangements of the a's and the b's, there is a one–to–one correspondence between the sum (5.2) with $m = 0$ and the hypergeometric series

$${}_pF_{q-1}(-n, -a_1, ..., -a_{p-1}; b_1 - n + 1, ..., b_{q-1} - n + 1; (-1)^{p+q} z).$$

We shall say that the sum (5.2) is balanced if $p = q$ and there exist a number a such that, with a suitable ordering of the parameters $a_1, ..., a_{p-1}, b_1, ..., b_{p-1}$, we have:

$$\begin{aligned} a_{p-1} &= n + 2a \\ b_{p-1} &= n - 2a \\ a_j &= b_j + 2a \quad \text{for } j = 1, ..., p-2. \end{aligned}$$

A balanced sum when written as a finite hypergeometric series is well–poised according to the terminology introduced by Whipple (1926). In a balanced sum, we prefer to introduce

$$c_j = \frac{1}{2}(a_j + b_j) \quad \text{for } j = 1, ..., p-2.$$

The sum (5.2) may then be written as

$$\sum_{k=m}^{n} \binom{n-m}{k-m} [n-m+2a]_{k-m} [c_1 + a]_{k-m} \cdots [c_{p-2} + a]_{k-m} \\ \times [n-m-2a]_{n-k} [c_1 - a]_{n-k} \cdots [c_{p-2} - a]_{n-k} (-1)^{p(k-m)} z^{k-m}. \tag{5.4}$$

If, furthermore $c_{p-2} = \frac{n-m}{2} - 1$, we shall say that the sum (5.2) is *well–balanced.* In this case, we have

$$\begin{aligned} & [c_{p-2} + a]_{k-m} [c_{p-2} - a]_{n-k} \\ = \ & \left[\frac{n-m}{2} + a - 1\right]_{k-m} \left[\frac{n-m}{2} - a - 1\right]_{n-k} \\ = \ & \left[\frac{n-m}{2} + a - 1\right]_{n-m-1} (-1)^{n-k} \left(\frac{n+m}{2} + a - k\right) \end{aligned}$$

The sum (5.4) multiplied by

$$\frac{2(-1)^{n-m}}{\left[\frac{n-m}{2} + a - 1\right]_{n-m-1}}$$

may then be written as

$$\sum_{k=m}^{n} \binom{n-m}{k-m} [n-m+2a]_{k-m} [c_1 + a]_{k-m} \cdots [c_{p-3} + a]_{k-m} [n-m-2a]_{n-k} \\ \times [c_1 - a]_{n-k} \cdots [c_{p-3} - a]_{n-k} (n+m+2a-2k)(-1)^{(p-1)(k-m)} z^{k-m}. \tag{5.5}$$

In this paper, we prefer the form (5.5) to the standard form (5.2) for a well-balanced sum.

The basic tool in this paper is the Chu–Vandermonde convolution given by

$$\sum_{k=m}^{n} \binom{n-m}{k-m} [x]_{k-m} [y]_{n-k} = [x+y]_{n-m} \tag{5.6}$$

which we shall use without proof.

5.2 The Generalized Pfaff–Saalschütz Formula

For a proof of the generalized Pfaff–Saalschütz formula due to H. M. Srivastava (1989), we need the following lemma.

Lemma 5.2.1 *Let*

$$S_n(a_1,a_2,b_1,b_2)=\sum_{k=0}^{n}\binom{n}{k}[a_1]_k\,[a_2]_k\,[b_1]_{n-k}\,[b_2]_{n-k}\,(-1)^k. \tag{5.7}$$

Then

$$S_n(a_1,a_2,b_1,b_2)=(-1)^n S_n(n-a_1-b_1-1,a_2,b_1,n-a_2-b_2-1), \tag{5.8}$$

and

$$S_n(a_1,a_2,b_1,b_2)=S_n(a_2,a_1,b_1,b_2)=S_n(a_1,a_2,b_2,b_1)=S_n(a_2,a_1,b_2,b_1). \tag{5.9}$$

PROOF. To prove (5.8), we use the fact that $[a_1]_k\,(-1)^k=[-a_1+k-1]_k$ and apply the Chu–Vandermonde convolution to write $[-a_1+k-1]_k$ as

$$\sum_{j=0}^{k}\binom{k}{j}[n-a_1-b_1-1]_j\,[b_1-n+k]_{k-j}\,. \tag{5.10}$$

Replacing $[a_1]_k\,(-1)^k$ in the right-hand side of (5.7) by (5.10) and interchanging the order of summation, and after using the identities

$$[b_1]_{n-k}\,[b_1-n+k]_{k-j}=[b_1]_{n-j}$$

and

$$[a_2]_k=[a_2]_j\,[a_2-j]_{k-j}\,,$$

we obtain

$$\begin{aligned} &S_n(a_1,a_2,b_1,b_2)\\ &\quad=\sum_{j=0}^{n}\binom{n}{j}[n-a_1-b_1-1]_j\,[a_2]_j\,[b_1]_{n-j}\sum_{k=j}^{n}\binom{n-j}{k-j}[a_2-j]_{k-j}\,[b_2]_{n-k}\,. \end{aligned} \tag{5.11}$$

The inner sum in (5.11) is a Chu–Vandermonde-sum and equals

$$[a_2+b_2-j]_{n-j}=[n-a_2-b_2-1]_{n-j}\,(-1)^{n-j}. \tag{5.12}$$

Using this, we obtain from (5.11) that

$$\begin{aligned}
&S_n(a_1, a_2, b_1, b_2) \\
&= (-1)^n \sum_{j=0}^{n} \binom{n}{j} [n - a_1 - b_1 - 1]_j [a_2]_j [b_1]_{n-j} [n - a_2 - b_2 - 1]_{n-j} (-1)^j \\
&= (-1)^n S_n(n - a_1 - b_1 - 1, a_2, b_1, n - a_2 - b_2 - 1). \qquad (5.13)
\end{aligned}$$

The formula (5.9) is now obvious. ■

We are now prepared to prove the generalized Pfaff–Saalschütz identity.

Theorem 5.2.1 *If* $a_1 + a_2 + b_1 + b_2 - n + m + 1 = p \in \{0, 1, 2, ...\}$, *then*

$$\begin{aligned}
&\sum_{k=m}^{n} \binom{n-m}{k-m} [a_1]_{k-m} [a_2]_{k-m} [b_1]_{n-k} [b_2]_{n-k} (-1)^{k-m} \\
&= \sum_{k=m}^{n} \binom{n-m}{k-m} [a_1]_{k-m} [a_2]_{k-m} [b_1]_{n-k} \\
&\quad \times [n - m + p - a_1 - a_2 - b_1 - 1]_{n-k} (-1)^{k-m} \\
&= \sum_{k=0}^{p} \binom{p}{k} [n-m]_k [a_2]_k [a_1 + b_1 - p]_{n-m-k} \\
&\quad \times [n - m - a_2 - b_1 - 1]_{n-m-k} (-1)^k \\
&= [a_1 + b_1 - p]_{n-m-p} [n - m - a_2 - b_1 - 1]_{n-m-p} \\
&\quad \times \sum_{k=0}^{p} \binom{p}{k} [n-m]_k [a_2]_k [a_1 + b_1 - n + m]_{p-k} [p - a_2 - b_1 - 1]_{p-k} (-1)^k. \qquad (5.14)
\end{aligned}$$

We shall call p as the *excess* of the generalized Pfaff–Saalschütz sum.

Proof. It is sufficient to prove Theorem 5.2.1 for $m = 0$. Using the notations in Lemma 5.2.1, we may write the left-hand side of (5.14) for $m = 0$ as $S_n(a_1, a_2, b_1, b_2)$. From Lemma 5.2.1, it then equals

$$\begin{aligned}
&(-1)^n S_n(n - a_1 - b_1 - 1, a_2, b_1, n - a_2 - b_2 - 1) \\
&= (-1)^n S_n(n - a_1 - b_1 - 1, a_2, n - a_2 - b_2 - 1, b_1). \qquad (5.15)
\end{aligned}$$

We now apply the Lemma 5.2.1 to the right-hand side of (5.15) to obtain

$$\begin{aligned}
&S_n(a_1, a_2, b_1, b_2) \\
&= (-1)^n S_n(n - a_1 - b_1 - 1, a_2, n - a_2 - b_2 - 1, b_1) \\
&= S_n(a_1 + a_2 + b_1 + b_2 - n + 1, a_2, n - a_2 - b_2 - 1, n - a_2 - b_1 - 1) \\
&= S_n(p, a_2, a_1 + b_1 - p, n - a_2 - b_1 - 1). \qquad (5.16)
\end{aligned}$$

In the last step, we have used the condition in the theorem twice.

Using the fact that $\binom{n}{k}[p]_k = \binom{p}{k}[n]_k$, we obtain the middle expression in (5.14) with $m = 0$. The right-hand side of (5.14) with $m = 0$ is obtained upon using

$$[a_1 + b_1 - p]_{n-k} = [a_1 + b_1 - p]_{n-p}\,[a_1 + b_1 - n]_{p-k}$$

and

$$[n - a_2 - b_1 - 1]_{n-k} = [n - a_2 - b_1 - 1]_{n-p}\,[p - a_2 - b_1 - 1]_{p-k}\,. \qquad \blacksquare$$

Note that the right-hand side of (5.14) is the product of two factorials both of length $n - m - p$ and a polynomial in $n - m$ of degree p.

5.3 A Modified Pfaff–Saalschütz Sum of Type $II(4,4,1)N$

For the proof of the generalized Dougall's identity, we shall use the following result.

Theorem 5.3.1 *If $a_1 + a_2 + b_1 + b_2 - n + 1 = p \in \{0, 1, 2, \ldots\}$, then*

$$\begin{aligned}
&\sum_{k=0}^{n} \binom{n}{k} [a_1]_k\,[a_2]_k\,[b_1]_{n-k}\,[b_2]_{n-k}\,[c+k]_p\,(-1)^k \\
&\quad = \sum_{k=0}^{p} \binom{p}{k} [n]_k\,[a_2]_k\,[p - a_1 - c - 1]_k \\
&\qquad\qquad \times [a_1 + b_1 - p]_{n-k}\,[n - a_2 - b_1 - 1]_{n-k}\,[c]_{p-k}\,. \qquad (5.17)
\end{aligned}$$

The sum on the left-hand side of (5.17) is a generalized Pfaff–Saalschütz sum, where the terms have the extra factor $[c+k]_p$. For $p = 0$, (5.17) reduces to the Pfaff–Saalschütz identity.

Proof. We use the Chu–Vandermonde identity to write $[c+k]_p$ in the left-hand side of (5.17) as $\sum_{j=0}^{p} \binom{p}{j} [k]_j\,[c]_{p-j}$ to obtain a double sum. We change the order of summation and use the identities

$$\begin{aligned}
\binom{n}{k} [k]_j &= [n]_j \binom{n-j}{k-j}, \\
[a_1]_k &= [a_1]_j\,[a_1 - j]_{k-j}\,, \\
[a_2]_k &= [a_2]_j\,[a_2 - j]_{k-j}\,,
\end{aligned}$$

to obtain

$$\sum_{j=0}^{p} \binom{p}{j} [n]_j [a_1]_j [a_2]_j [c]_{p-j} (-1)^j \\ \times \sum_{k=j}^{n} \binom{n-j}{k-j} [a_1 - j]_{k-j} [a_2 - j]_{k-j} [b_1]_{n-k} [b_2]_{n-k} (-1)^{k-j}. \tag{5.18}$$

The inner sum is a generalized Pfaff–Saalschütz sum with excess $p - j$. Using this, we obtain for the left-hand side of (5.17) the expression

$$\sum_{j=0}^{p} \binom{p}{j} [n]_j [a_1]_j [a_2]_j [c]_{p-j} (-1)^j \\ \times \sum_{h=0}^{p-j} \binom{p-j}{h} [n-j]_h [a_2 - j]_h [a_1 + b_1 - p]_{n-j-h} [a_2 + b_2 - p]_{n-j-h} (-1)^h. \tag{5.19}$$

We now use $[n]_j [n-j]_h = [n]_{j+h}$ and $[a_2]_j [a_2 - j]_h = [a_2]_{j+h}$ and substitute $k - j$ for h. We next change the order of summation to obtain

$$\sum_{k=0}^{p} \binom{p}{k} [n]_k [a_2]_k [a_1 + b_1 - p]_{n-k} [a_2 + b_2 - p]_{n-k} (-1)^k [c]_{p-k} \\ \times \sum_{j=0}^{k} \binom{k}{j} [a_1]_j [c - p + k]_{k-j}. \tag{5.20}$$

In (5.20), the inner sum is a Chu–Vandermonde-sum and using the identity, we then obtain

$$\sum_{k=0}^{p} \binom{p}{k} [n]_k [a_2]_k [a_1 + b_1 - p]_{n-k} [a_1 + b_2 - p]_{n-k} \\ \times [c]_{p-k} [a_1 + c - p + k]_k (-1)^k. \tag{5.21}$$

Since $[a_1 + c - p + k]_k (-1)^k = [p - a_1 - c - 1]_k$, we finally obtain the right-hand side of (5.17). This proves the theorem. ∎

5.4 A Well–Balanced $II(5,5,1)N$ Identity

The following theorem is a special case of an evaluation of an infinite well–poised hypergeometric series found in Slater (1966, p. 56).

Theorem 5.4.1 *For arbitrary a, b, c and integers m and n with $m \le n$, we have*

$$\begin{aligned}
&\sum_{k=m}^{n} \binom{n-m}{k-m} [n-m+2a]_{k-m} [b+a]_{k-m} [c+a]_{k-m} \\
&\qquad \times [b-a]_{n-k} [c-a]_{n-k} [n-m-2a]_{n-k} (n+m+2a-2k) \\
&= [n-m+2a]_{2n-2m+1} [b+c]_{n-m} . \qquad (5.22)
\end{aligned}$$

PROOF. Once again, we may assume that $m = 0$. We use the original Pfaff–Saalschütz identity to replace $[b+a]_k [c+a]_k$ in the left-hand side of (5.22) by

$$\begin{aligned}
&\sum_{j=0}^{k} \binom{k}{j} [n+2a-k]_j [n-b-c-1]_j [b-a-n+k]_{k-j} \\
&\qquad \times [c-a-n+k]_{k-j} (-1)^j. \qquad (5.23)
\end{aligned}$$

Upon using the identities

$$\binom{n}{k}\binom{k}{j} = \binom{n}{j}\binom{n-j}{k-j},$$

$$\begin{aligned}
[n+2a]_k [n+2a-k]_j &= [n+2a]_{2j} [n+2a-2j]_{k-j} , \\
[b-a]_{n-k} [b-a-n+k]_{k-j} &= [b-a]_{n-j} , \\
[c-a]_{n-k} [c-a-n+k]_{k-j} &= [c-a]_{n-j} ,
\end{aligned}$$

and interchanging the order of summation, we obtain the following double sum for the left-hand side of (5.22):

$$\begin{aligned}
&\sum_{j=0}^{n} \binom{n}{j} [n+2a]_{2j} [n-b-c-1]_j [b-a]_{n-j} [c-a]_{n-j} (-1)^j \\
&\qquad \times \sum_{k=j}^{n} \binom{n-j}{k-j} [n+2a-2j]_{k-j} [n-2a]_{n-k} (n+2a-2k). \qquad (5.24)
\end{aligned}$$

Writing $(n+2a-2k)$ as $(n+2a-2j)-2(k-j)$, the inner sum may be written as

$$\begin{aligned}
&\sum_{k=j}^{n} \binom{n-j}{k-j} [n+2a-2j]_{k-j} [n-2a]_{n-k} (n+2a-2j) \\
&\qquad - \sum_{k=j}^{n} \binom{n-j}{k-j} [n+2a-2j]_{k-j} [n-2a]_{n-k} 2(k-j) \\
&= (n+2a-2j) \sum_{k=j}^{n} \binom{n-j}{k-j} [n+2a-2j]_{k-j} [n-2a]_{n-k}
\end{aligned}$$

$$- 2(n-j)(n+2a-2j) \sum_{k=j+1}^{n} \binom{n-j-1}{k-j-1} \times [n+2a-2j-1]_{k-j-1} [n-2a]_{n-k} . \tag{5.25}$$

In the right-hand side of (5.25), both sums are Chu–Vandermonde sums and, hence, we obtain for the inner sum the expression

$$(n+2a-2j) \left([2n-2j]_{n-j} - 2(n-j) [2n-2j-1]_{n-j-1} \right) . \tag{5.26}$$

It is easily seen that (5.26) vanishes for $j = 0, 1, ..., n-1$. For $j = n$, (5.26) reduces to $(n+2a-2n)(1-0) = 2a-n$. The sum (5.24), therefore, reduces to a single term given by

$$(-1)^n [n+2a]_{2n+1} [n-b-c-1]_n = [n+2a]_{2n+1} [b+c]_n .$$

This completes the proof, which might be replaced by a proof using Zeilberger's algorithm. ∎

5.5 A Generalization of Dougall's Well–Balanced $II(7,7,1)N$ Identity

For $p = 0$, this result was proved by Dougall (1907).

Theorem 5.5.1 *Let*

$$\begin{aligned} S &= \sum_{k=0}^{n} \binom{n}{k} [n+2a]_k [b+a]_k [c+a]_k [d+a]_k [e+a]_k \\ &\quad \times [n-2a]_{n-k} [b-a]_{n-k} [c-a]_{n-k} [d-a]_{n-k} [e-a]_{n-k} (n+2a-2k). \end{aligned} \tag{5.27}$$

If

$$b+c+d+e-n+1 = p \in \{0, 1, 2, ...\}, \tag{5.28}$$

then

$$\begin{aligned} S &= (-1)^n [n+2a]_{2n+1} \\ &\quad \times \sum_{k=0}^{p} \binom{p}{k} [n]_k [c+a]_k [c-a]_k [b+d-p]_{n-k} [b+e-p]_{n-k} [d+e-p]_{n-k} \\ &= (-1)^n [n+2a]_{2n+1} [b+d-p]_{n-p} [b+e-p]_{n-p} [d+e-p]_{n-p} \\ &\quad \times \sum_{k=0}^{p} \binom{p}{k} [n]_k [c+a]_k [c-a]_k [b+d-n]_{p-k} [b+e-n]_{p-k} [d+e-n]_{p-k} . \end{aligned} \tag{5.29}$$

PROOF. From the Pfaff–Saalschütz identity, it follows that

$$\begin{aligned}[d+a]_k\,[e+a]_k &= \sum_{j=0}^{k}\binom{k}{j}[n+2a-k]_j\,[n-d-e-1]_j \\ &\quad\times[d-a-n+k]_{k-j}\,[e-a-n+k]_{k-j}\,(-1)^j.\end{aligned}\tag{5.30}$$

Replacing $[d+a]_k\,[e+a]_k$ in S by the right-hand side of (5.30) and simplifying as we did in the earlier proofs, we obtain

$$\begin{aligned}S &= \sum_{j=0}^{n}\binom{n}{j}[n+2a]_{2j}\,[b+a]_j\,[c+a]_j\,[n-d-e-1]_j \\ &\quad\times[d-a]_{n-j}\,[e-a]_{n-j}\,(-1)^j \\ &\quad\times\sum_{k=j}^{n}\binom{n-j}{k-j}[n+2a-2j]_{k-j}\,[b+a-j]_{k-j}\,[c+a-j]_{k-j} \\ &\quad\times[n-2a]_{n-k}\,[b-a]_{n-k}\,[c-a]_{n-k}\,(n+2a-2k).\end{aligned}\tag{5.31}$$

The inner sum is well–balanced and we can use Theorem 5.2.1 to observe that this sum reduces to

$$[n-2j+2a]_{2n-2j+1}\,[b+c-j]_{n-j}\,.\tag{5.32}$$

Replacing the inner sum in (5.31) by (5.32) and simplifying, we find that

$$\begin{aligned}S &= [n+2a]_{2n+1}\sum_{j=0}^{n}\binom{n}{j}[b+a]_j\,[c+a]_j\,[n-d-e-1]_j \\ &\quad\times[d-a]_{n-j}\,[e-a]_{n-j}\,[b+c-j]_{n-j}\,(-1)^j.\end{aligned}\tag{5.33}$$

We now use condition (5.28) and find that

$$\begin{aligned}&[n-d-e-1]_j\,[b+c-j]_{n-j} \\ &\quad= [n-d-e-1]_{n-p}\,[b+c-j]_p \\ &\quad= [n-d-e-1]_{n-p}\,[p-b-c-1+j]_p\,(-1)^p.\end{aligned}\tag{5.34}$$

Using (5.34), we may rewrite (5.33) as

$$\begin{aligned}S &= [n+2a]_{2n+1}\,[n-d-e-1]_{n-p} \\ &\quad\times\sum_{j=0}^{n}\binom{n}{j}[b+a]_j\,[c+a]_j\,[d-a]_{n-j}\,[e-a]_{n-j} \\ &\quad\times[p-b-c-1+j]_p\,(-1)^{p-j}.\end{aligned}\tag{5.35}$$

The sum in the formula (5.35) may be simplified upon using Theorem 5.3.1 since condition (5.28) ensures the applicability of this theorem. We then obtain

$$\begin{aligned} S &= (-1)^p [n+2a]_{2n+1} [n-d-e-1]_{n-p} \\ &\quad \times \sum_{k=0}^{p} \binom{p}{k} [n]_k [c+a]_k [c-a]_k [b+d-p]_{n-k} \\ &\quad \times [n-c-d-1]_{n-k} [p-b-c-1]_{p-k} . \end{aligned} \tag{5.36}$$

From condition (5.28), we also have

$$[n-c-d-1]_{n-k} = [b+e-p]_{n-k}$$

and

$$[p-b-c-1]_{p-k} = [d+e-n]_{p-k} .$$

Using now the simplification

$$\begin{aligned} (-1)^p [n-d-e-1]_{n-p} [d+e-n]_{p-k} &= (-1)^n [d+e-p]_{n-p} [d+e-n]_{p-k} \\ &= (-1)^n [d+e-p]_{n-k} , \end{aligned}$$

we obtain the first expression in (5.29). The second expression in (5.29) is readily obtained upon using the identities

$$[b+d-p]_{n-k} = [b+d-p]_{n-p} [b+d-n]_{p-k} ,$$

$$[b+e-p]_{n-k} = [b+e-p]_{n-p} [b+e-n]_{p-k}$$

and

$$[d+e-p]_{n-k} = [d+e-p]_{n-p} [d+e-n]_{p-k} .$$

If e is eliminated using condition (5.28), then the right-hand side of (5.29) takes the form of a product of four factorials and a polynomial in n of degree at most p. ■

References

1. Dougall, J. (1907). On Vandermonde's theorem and some more general expansions, *Proceedings of the Edinburgh Mathematical Society*, **25**, 114–132.

2. Petkovšek, M., Wilf, H. S. and Zeilberger, D (1996). $A = B$ (Ed., A. K. Peters), Massachusetts: Wellesley.

3. Slater, L. J. (1966). *Generalized Hypergeometric Functions*, Cambridge: Cambridge University Press.

4. Srivastava, H. M. (1989). An extension of the q-Saalschützian theorem, *Acta Mathematica Hungarica*, **53**, 115–118.

5. Whipple, F. J. W. (1926). On well–poised series, generalized hypergeometric series having parameters in pairs, each pair with the same sum, *Proceedings of the London Mathematical Society*, **24**, 247–263.

6

A Comparison Of Two Methods For Random Labelling of Balls by Vectors of Integers

Doron Zeilberger

Temple University, Philadelphia, PA

Abstract: Kirk (1993) raised the question of comparing the following two ways for labelling balls. Given r pre-determined positive integers n_i $(1 \leq i \leq r)$, and given N balls (N large), consider two ways to randomly assign $r-$component vectors of integers $(a_1, \dots, a_r)$ to them, such that $1 \leq a_i \leq n_i$. We will call these vectors *labels*. Of course, altogether there are $\prod_{i=1}^{r} n_i$ possible labels.

Keywords and phrases: Urn models, combinatorial methods for finding moments

6.1 First Way

You put all the balls in one big pot. For $i = 1, \dots, r$, at the i-th iteration, line up n_i smaller pots, each with capacity N/n_i balls, and labeled with labels 1 through n_i, and, uniformly at random, distribute them into these smaller pots. Assign the i-th component of the vector-label of each ball, a_i, to be the label of the pot in which it was dropped. Having done that, you dump all the balls back into the big pot, and go on to the next iteration.

6.2 Second Way

Do the same as above for $i = 1$, except that at the end of the first iteration you do *not* dump back the balls into the large ball but proceed as follows. For $i = 2, \dots, r$, assuming that the balls have already received their first $i - 1$ components, leaving the balls in their pots from the $(i - 1)$-th iteration, you line-up n_i new pots, each with a capacity of N/n_i balls, and labeled with labels 1 through n_i. For each of the n_{i-1} pots from the previous iteration, individually, we uniformly at random, distribute their contents into the new pots, each of the n_i new pots getting exactly $N/(n_{i-1}n_i)$ balls from each of the n_{i-1} pots from the previous $(i-1)$-th iteration.

Note that in the First Way, assuming that we can reuse the pots, we need $1+\max(n_1,\ldots,n_r)$ pots, one of which should have a capacity of N balls, while in the Second Way, we need $\max(1+n_1, n_1+n_2,\ldots,n_{r-1}+n_r)$ pots.

The goal is to maximize the 'equal representation' of all the possible $\prod_{i=1}^r n_i$ vector-labels. It is obvious that, with either way, the probability of a ball to be assigned any given label is $\prod_{i=1}^r n_i^{-1}$, and hence that the expected number of balls to be given label v, for each of $v \in \prod_{i=1}^r [1, n_i]$, is $N\prod_{i=1}^r n_i^{-1}$.

It is intuitively obvious that in the Second Way the 'spread' in the distribution is less than in the First Way. In fact, when $r = 2$, the Second Way gives a *perfect* way of equi-distribution. We are *guaranteed* that the number of balls given any particular label (a_1, a_2) is *exactly* $N/(n_1n_2)$.

Throughout this note we assume that N is divisible by $lcm(n_1n_2, n_2n_3, \ldots, n_{r-1}n_r)$. For any statement P, $\chi(P)$ is 1 or 0 according as whether P is true or false, respectively.

The way to quantify 'spread' is via *standard deviation*, or its square, the *variance*. By symmetry, it is enough to pick any one *fixed* label v, say $v = (1,\ldots,1)$.

The 'random variable' on a given 'experiment' is the 'number of balls labelled v'. To compute its variance, we will use an old trick, described beautifully in Section 8.2 of the modern classic by Graham, Knuth and Patashnik (1993). This trick can also be used to find the average (i.e., first moment), in which case it is even easier to use, and higher moments, in which case it is (usually) harder to use.

Let S denote the set of all possible outcomes of the 'labelling experiment'. The total number of outcomes in the First Way is

$$|S| = \prod_{i=1}^{r} \frac{N!}{(N/n_i)!^{n_i}} .$$

For each outcome s, let $\alpha(s)$ be the quantity 'number of balls that receive the (fixed) label v'.

Let us first compute the average of this quantity (even though we know the answer, just as a warm-up for the calculation of the variance, that would follow). We have

$$\sum_{s\in S} \alpha(s) = \sum_{s\in S}\sum_{j=1}^{N} \chi(\text{the } j\text{-th ball is labelled } v)$$

$$= \sum_{j=1}^{N} \sum_{s \in S_j} 1, \tag{6.1}$$

where the inner sum extends over the set of outcomes, say S_j, of $s \in S$ for which the j-th ball was labelled v. By symmetry, this inner sum is independent of j, and equals

$$\prod_{i=1}^{r} \frac{(N-1)!}{((N/n_i)-1)!(N/n_i)!^{n_i-1}} ,$$

since at each iteration one of the balls (the j-th) is committed to lend in one of the pots (Pot v_i in the i-th iteration.)

Hence the sum in (6.1) equals

$$N \prod_{i=1}^{r} \frac{(N-1)!}{((N/n_i)-1)!(N/n_i)!^{n_i-1}} ,$$

and hence the average is

$$\begin{aligned} ave. &= N \prod_{i=1}^{r} \frac{\frac{(N-1)!}{((N/n_i)-1)!(N/n_i)!^{n_i-1}}}{\frac{(N)!}{(N/n_i)!^{n_i}}} \\ &= N \prod_{i=1}^{r} \frac{1}{n_i} , \end{aligned}$$

as expected.

6.3 Variance and Standard Deviation

Let us recall a few elementary facts about *variance*. The *standard deviation* is defined to be the square root of the variance. Suppose that we have a finite set S, and there is some numerical attribute (*random variable*) $X(s)$ for every element $s \in S$. Then the variance, $V(X)$, is the 'average of the squares of the deviation from the average', i.e.

$$V(X) = \frac{\sum_{s \in S}(X(s) - ave.)^2}{|S|} ,$$

where $|S|$ is the number of elements of S.

It is easier to compute the related quantity

$$W(X) = \frac{\sum_{s \in S} \binom{X(s)}{2}}{|S|} .$$

Simple algebra shows that

$$V(X) = 2W(X) + ave. - ave.^2 .$$

Now we are ready to compute $W(\alpha)$.

We have

$$\begin{aligned} W(\alpha) &= \frac{1}{|S|}\sum_{s\in S}\binom{\alpha(s)}{2} \\ &= \frac{1}{|S|}\sum_{s\in S}\sum_{1\le i<j\le N}\chi(\text{the } i\text{-th and } j\text{-th balls are both labelled } v) \\ &= \frac{1}{|S|}\sum_{1\le i<j\le N}[\text{Number of outcomes with the } i\text{-th and } j\text{-th} \\ &\qquad\qquad \text{balls both labelled } v]. \end{aligned} \tag{6.2}$$

By symmetry, the summand is independent of (i, j) and is easily seen to be equal to

$$\prod_{i=1}^{r}\frac{(N-2)!}{((N/n_i)-2)!(N/n_i)!^{n_i-1}}$$

since, at each of the r iterations, two balls are committed to lend at a predetermined pot (the v_i-th pot at the i-th iteration.)

Simple algebra yields

$$W(\alpha) = \binom{N}{2}\prod_{i=1}^{r} n_i^{-2}\frac{(1-n_i/N)}{(1-1/N)} .$$

It then follows that

$$V(\alpha) = ave.-ave.^2+2W(\alpha) = \frac{N}{\prod_{i=1}^{r} n_i}-\frac{N^2}{\prod_{i=1}^{r} n_i^2}+N(N-1)\prod_{i=1}^{r} n_i^{-2}\frac{(1-n_i/N)}{(1-1/N)} .$$

Assuming that N is large, so that $1/N$ is small, and using the approximation $1/(1-x) = 1+x+O(x^2)$, we get the following proposition.

Proposition 6.3.1 *The average number of occurrences of any given vector v as a label, in the First Way, is $N/\prod_{i=1}^{r} n_i$, and its variance is*

$$\frac{N}{\prod_{i=1}^{r} n_i}-\frac{N}{\prod_{i=1}^{r} n_i^2}\left(1+\sum_{i=1}^{r}(n_i-1)\right)+O(1) .$$

6.4 Analysis of the Second Way

In this case, the total number of outcomes is

$$|S| = \frac{N!}{(N/n_1)!^{n_1}}\prod_{i=2}^{r}\left[\frac{(N/(n_{i-1})!}{(N/(n_{i-1}n_i))!^{n_i}}\right]^{n_{i-1}} .$$

Using an analogous argument as before, the number of outcomes with the i-th and j-th balls labelled v equals

$$\frac{(N-2)!}{((N/n_1)-2)!(N/n_1)!^{n_1-1}} \prod_{i=2}^{r} \frac{((N/n_{i-1})-2)!}{((N/n_{i-1}n_i)-2)!(N/(n_{i-1}n_i))!^{n_i-1}} \times \left[\frac{(N/n_{i-1})!}{(N/(n_{i-1}n_i))!^{n_i}}\right]^{n_{i-1}-1} .$$

Simple algebra then yields that

$$W(\alpha) = \binom{N}{2} \prod_{i=1}^{r} n_i^{-2} \, \frac{(1-n_1/N)}{(1-1/N)} \prod_{i=2}^{r} \frac{(1-n_{i-1}n_i/N)}{(1-n_{i-1}/N)} ,$$

which, as before, leads to the following proposition.

Proposition 6.4.1 *The average number of occurrences of any given vector v as a label, in the Second Way, is $N/\prod_{i=1}^{r} n_i$, and its variance is*

$$\frac{N}{\prod_{i=1}^{r} n_i} - \frac{N}{\prod_{i=1}^{r} n_i^2}\left(n_1 + \sum_{i=2}^{r}(n_i-1)n_{i-1}\right) + O(1) ,$$

which is slightly smaller.

References

1. Graham, R. L., Knuth, D. E. and Patashnik, O. (1993). *Concrete Mathematics*, Second edition, Reading, Massachusetts: Addison-Wesley.

2. Kirk, G. (1993). *Private communication* (In 'Jeff's Bagels Coffee Bar', Rocky Hill, New Jersey, November 1993).

PART II

Applications to Probability Problems

7

On The Ballot Theorems

Lajos Takács

Case Western Reserve University, Cleveland, OH

Abstract: The discovery of various ballot theorems has had great impact on several areas of combinatorics and probability theory. This paper deals with the historical background and the development of these theorems, analyzes various proofs and gives some applications.

Keywords and phrases: Classical ballot theorem, general ballot theorem, historical development, applications, combinatorial identities

7.1 Introduction

In this paper we discuss the historical development of various ballot theorems, provide several proofs for these theorems, and give some applications. It is very surprising that the various ballot theorems have so many useful applications in many areas of mathematics, such as combinatorics, the theory of random walks, queuing theory, order statistics, and the theory of graphs. The simplicity and the generality of the ballot theorems might explain their wide range of uses.

7.2 The Classical Ballot Theorem

The following theorem is usually called the classical ballot theorem.

Theorem 7.2.1 *If in a ballot candidate A scores a votes and candidate B scores b votes where $a \geq b\mu$ and μ is a positive integer, then the probability that throughout the counting the number of votes registered for A is always greater* *than μ times the number of votes registered for B is given by*

$$P(a,b,\mu) = \frac{a-b\mu}{a+b}, \tag{7.1}$$

provided that all the possible voting records are equally probable.

PROOF. Every voting record can be represented by a sequence of a letters A and b letters B, where an A stands for a vote for A and a B stands for a vote for B. The number of possible voting records in which A scores a votes and B scores b votes is

$$\binom{a+b}{b} = \frac{(a+b)!}{a!b!}. \tag{7.2}$$

Let us denote by $N(a,b,\mu)$ the number of favourable voting records, that is, voting records in which throughout the counting the number of votes registered for A is always greater than μ times the number of votes registered for B. Then

$$P(a,b,\mu) = N(a,b,\mu) \Big/ \binom{a+b}{b}. \tag{7.3}$$

If we take into consideration that the last vote is registered for either A or B, then we obtain that

$$N(a,b,\mu) = N(a-1,b,\mu) + N(a,b-1,\mu) \tag{7.4}$$

for $a > b\mu$ and $b \geq 1$. Furthermore, we have $N(b\mu, b, \mu) = 0$ for $b \geq 1$, and $N(a, 0, \mu) = 1$ for $a \geq 1$. The recurrence formula (7.4) makes it possible to calculate $N(a,b,\mu)$ for $a > b\mu$. Table 7.1 contains $N(a,b,1)$ for $0 \leq b \leq a \leq 6$.

Table 7.1: $N(a,b,1)$

$a\backslash b$	0	1	2	3	4	5	6
0	-						
1	1	0					
2	1	1	0				
3	1	2	2	0			
4	1	3	5	5	0		
5	1	4	9	14	14	0	
6	1	5	14	28	42	42	0

After the publication of Blaise Pascal's famous *Treatise on the Arithmetic Triangle* in 1665 [Pascal (1908a,b)], it has became generally known that the binomial coefficients

$$F(a,b) = \binom{a+b-1}{b} = \frac{(a+b-1)!}{(a-1)!b!} \tag{7.5}$$

defined for $a \geq 1$ and $b \geq 0$ can be calculated by the recurrence formula

$$F(a,b) = F(a-1,b) + F(a,b-1) \tag{7.6}$$

where $a \geq 1$, $b \geq 1$, $F(a,0) = 1$ for $a \geq 0$, and $F(0,b) = 0$ for $b \geq 1$. In other words, (7.5) is uniquely determined by (7.6) and by the boundary conditions. See Table 7.2 for $F(a,b)$ for $a \leq 6$ and $b \leq 6$.

Table 7.2: $F(a,b)$

$a\backslash b$	0	1	2	3	4	5	6
0	1	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	1	2	3	4	5	6	7
3	1	3	6	10	15	21	28
4	1	4	10	20	35	56	84
5	1	5	15	35	70	126	210
6	1	6	21	56	126	252	462

We observe that both $F(a,b)$ and $F(b,a)$ satisfy (7.4). Therefore, if $F(a,b)$ is defined by (7.5), then

$$N(a,b,\mu) = F(a,b) - \mu F(b,a) \tag{7.7}$$

also satisfies (7.4) whenever $a > b\mu > 0$. Moreover, (7.7) satisfies the boundary conditions $N(b\mu, b, \mu) = 0$ for $b \geq 1$ and $N(a,0,\mu) = 1$ for $a \geq 1$. Accordingly, by (7.5) we obtain that

$$N(a,b,\mu) = \binom{a+b-1}{b} - \mu\binom{a+b-1}{a} \tag{7.8}$$

if $a \geq b\mu$. Finally, by (7.3) we get

$$P(a,b,\mu) = (a - b\mu)/(a+b) \tag{7.9}$$

if $a \geq b\mu$. This proves (7.1) for any positive integer μ. ■

In the particular case when $\mu = 1$, formula (7.1) was discovered by Bertrand (1887), and was proved in the same year by André (1887). Also Barbier (1887) noticed that if μ is a positive integer, then (7.1) holds. However, Barbier did not prove (7.1). Its proof was given only in Aeppli (1923, 1924). Aeppli's proof is in his dissertation which he wrote under the supervision of Professor György Pólya. Thanks to Professor Pólya, I have a copy of Aeppli's dissertation and I believe that this is the first paper which gives an account of Aeppli's remarkable proof. As we have shown, Theorem 7.2.1 can be proved simply by making use of only a fundamental property of the arithmetic triangle. Although the classical ballot theorem attracted considerable attention at the time, it required 37 years to accomplish the task of proving it.

The numbers $F(a,b)$ defined in (7.5) are known as *figurate numbers* or *binomial coefficients*. Printed tables for $F(a,b)$ were already available in the sixteenth and seventeenth centuries. See, for example, Apianus (1527) for $a+b \leq 10$, Cardano (1570, p. 135) for $a+b \leq 12$, Mersenne (1635-1636, Libr. VII, p. 134) for $a \leq 25$ and $b \leq 12$, and Pascal (1908, p. 446) for $a+b \leq 10$. It has been known for a long time that the numbers $F(a,b)$ have also combinatorial interpretations. In particular, $F(a,b)$ is the number of different arrangements of $a-1$ letters A and b letters B in a row.

7.3 The Original Proofs of Theorem 7.2.1

In the particular case of $\mu = 1$, Theorem 7.2.1 was proved by André (1887). His proof was highly appreciated. In his book, Bertrand (1889, pp. 18-20) presented André's proof and praised André for his ingenious demonstration. Poincaré (1912, pp. 21-26) also included André's proof in his book.

PROOF OF BERTRAND'S THEOREM BY D. ANDRÉ. In what follows we describe the original proof of André (1887). He demonstrated that

$$N(a,b,1) = \binom{a+b}{b} - 2\binom{a+b-1}{a} \tag{7.10}$$

for $a \geq b \geq 1$.

His reasoning is as follows: Every voting record can be described by a sequence of a letters A and b letters B if an A stands for a vote for A and a B stands for a vote for B. The total number of voting records is given by (7.2). André showed that the number of unfavourable voting records is

$$2\binom{a+b-1}{a}. \tag{7.11}$$

Consequently (7.10) is true, and (7.10) implies (7.1).

To prove (7.11) let us observe that the set of unfavourable voting records can be divided into two disjoint subsets: The first subset contains all the voting records in which the first letter is B and in addition there are a letters A and $b-1$ letters B. The second subset contains all the unfavourable voting records in which the first letter is A and in addition there are $a-1$ letters A and b letters B.

There is a one-to-one correspondence between the voting records in these two subsets. This can be seen as follows: If a voting record belongs to the second subset, then counting the letters from left to right, there is a shortest subsequence which contains an equal number of letters A and B. The last letter in this shortest sequence is necessarily B. In this shortest sequence, let us remove all the letters except the last B and put them at the end of the voting record in the same order. Then we obtain a voting record which belongs to the first subset.

Conversely, if a voting record belongs to the first subset, then counting letters from right to left, there is a shortest subsequence which contains one more letter A than B. The first letter in this shortest sequence is necessarily A. Let us remove all the letters in this shortest sequence and put them in the same order at the beginning of the voting record. Then we obtain a voting record which belongs to the second subset.

It is evident that this mapping is one-to-one, and therefore both subsets contain $\binom{a+b-1}{a}$ voting records. Thus the total number of unfavourable voting records is given by (7.11). ■

For $\mu \geq 1$, Theorem 7.2.1 was proved by Aeppli (1923, 1924).

Proof of Theorem 7.2.1 by A. Aeppli. This proof is a somewhat modified version of the original proof of Aeppli (1924). Among the first r votes recorded, denote by α_r the number of votes for A and by β_r the number of votes for B. Then (7.1) can also be expressed as

$$P(a,b,\mu) = \Pr[\alpha_r > \beta_r\mu \text{ for } r = 1,2,\ldots,a+b] = \frac{(a-b\mu)}{(a+b)} \tag{7.12}$$

if $a \geq b\mu$. To prove (7.12), define $\gamma_r = \alpha_r - \beta_r\mu$ for $r = 1,2,\ldots,a+b$. Then

$$\begin{aligned} P(a,b,\mu) &= \Pr[\gamma_r > 0 \text{ for all } r = 1,2,\ldots,a+b] \\ &= 1 - \Pr[\gamma_r \leq 0 \text{ for some } r = 1,2,\ldots,a+b]. \end{aligned} \tag{7.13}$$

We have

$$\begin{aligned} &\Pr[\gamma_r \leq 0 \text{ for some } r = 1,2,\ldots,a+b] \\ &= \Pr[\gamma_1 = -\mu] + \sum_{\ell=0}^{\mu-1} \Pr[\gamma_r > 0 \text{ for } 0 < r < s,\ \gamma_s = -\ell \text{ for some } s \geq 2]. \end{aligned} \tag{7.14}$$

In the sum, each term is equal to

$$\Pr[\gamma_1 = -\mu] = \binom{a+b-1}{b-1} \Big/ \binom{a+b}{b} = b/(a+b). \tag{7.15}$$

This can be seen as follows: If for a fixed $\ell = 0,1,\ldots,\mu-1$, we consider the set of voting records in which $\gamma_r > 0$ for $0 < r < s$ and $\gamma_s = -\ell$ for some

$s \geq 2$, and if in each voting record we reverse the order of the first s votes, then we obtain a voting record which belongs to the set of voting records in which $\gamma_1 = -\mu$. Conversely, if a voting record belongs to the set of voting records in which $\gamma_1 = -\mu$, then for each fixed $\ell = 0, 1, \ldots, \mu - 1$, there exists a smallest $s \geq 2$ such that $\gamma_s = -\ell$. Let us reverse the order of the first s votes in such a voting record. Then we obtain a voting record for which $\gamma_r > 0$ for $0 < r < s$ and $\gamma_s = -\ell$. There is a one-to-one correspondence between the voting records in the two sets. Thus

$$P(a, b, \mu) = 1 - (\mu + 1)\Pr[\gamma_1 = -\mu] = (a - \mu b)/(a + b) \tag{7.16}$$

for $a \geq b\mu$. This proves (7.12). ∎

7.4 Historical Background

The origin of the classical ballot theorem can be traced back to a problem in games of chance. In 1708, De Moivre (1711, pp. 262-263 and 1984, pp. 260-261) solved the following problem of games of chance: Two players A and B agree to play a series of games. In each game, independently of the others, either A wins a counter from B with probability p or B wins a counter from A with probability q where $p > 0, q > 0$ and $p + q = 1$. Let us suppose that initially A has an unlimited number of counters and B has only k counters where k is a positive integer. If B is ruined, that is, if B loses all of his counters, the series ends. Denote by $\rho(k)$ the duration of the games, that is, the number of games played until B is ruined. If B is never ruined, then $\rho(k) = \infty$. The problem is to determine the distribution of $\rho(k)$. De Moivre (1711, Problem XXV, p. 262 and 1984, Problem 25, p. 260) discovered that

$$\Pr[\rho(k) \leq n] = \sum_{k \leq j < (n+k)/2} \binom{n}{j-k} p^j q^{n-j} + \sum_{(n+k)/2 \leq j \leq n} \binom{n}{j} p^j q^{n-j} \tag{7.17}$$

for $1 \leq k \leq n$; see also Hald (1984). De Moivre (1718, Problem XL, pp. 119-122, 1738, Problem LXIV, pp. 179-181 and 1756, Problem LXV, pp. 208-210) also expressed (7.17) in the following form

$$\Pr[\rho(k) \leq n] = \sum_{0 \leq j \leq (n-k)/2} \frac{k}{k+2j} \binom{k+2j}{j} p^{k+j} q^j \tag{7.18}$$

for $1 \leq k \leq n$. From (7.18), it follows that

$$\Pr[\rho(k) = k + 2j] = \frac{k}{k+2j} \binom{k+2j}{j} p^{k+j} q^j \tag{7.19}$$

for $j \geq 0$ and $k \geq 1$. We note that $\Pr[\rho(k) < \infty] = 1$ if $p \geq q$. If $p < q$, then $\Pr[\rho(k) < \infty] = (p/q)^k$.

De Moivre stated (7.17) and (7.18) without proof. Formula (7.18) was proved only in 1773 by Laplace (1776, pp. 188-193; 1812, p. 235, and 1814, p. 238). Both (7.17) and (7.18) were proved by Lagrange (1777, pp. 230-238). It is interesting to recall Ampère (1802, p. 9) who describes formula (7.19) as remarkable for its simplicity and elegance. See also Takács (1969).

The probability that in $k + 2j$ games A wins $k + j$ games and B wins j games is

$$\binom{k+2j}{j} p^{k+j} q^j. \tag{7.20}$$

The conditional probability that B will be ruined at the $(k+2j)$th game, given that in the $k + 2j$ games A wins $k + j$ games and B wins j games, is

$$Q(k+j, j) = k/(k+2j). \tag{7.21}$$

Let us imagine that two candidates A and B play a series of games of chance. Suppose that A has an unlimited number of counters and B has only $a - b \geq 0$ counters and A wins a games and B wins b games until B is ruined. Let us consider the $a + b$ games in reverse order and suppose that a win for A corresponds to a vote for A, and a win for B corresponds to a vote for B. Then we can see immediately that $P(a, b, 1) = Q(a, b) = (a - b)/(a + b)$.

Although De Moivre's books were widely known, it escaped the attention of contemporary mathematicians that De Moivre's results can be used to solve the ballot problem for $\mu = 1$. It also escaped attention that for $\mu = 1$, Theorem 7.2.1 can also be deduced from some results of Whitworth (1879) for random walks.

The method of reflection is widely used in the theory of random walks, and it seems interesting to mention how Theorem 7.2.1 can be proved for $\mu = 1$ simply by using the reflection principle.

PROOF OF BERTRAND'S THEOREM BY THE REFLECTION PRINCIPLE. A voting record is favourable if the first vote is for A and if in the course of counting no tie occurs. Let us consider the set of voting records in which the first vote is for A. The number of such voting records is

$$\binom{a+b-1}{b} = \frac{(a+b-1)!}{(a-1)!b!}. \tag{7.22}$$

Among these voting records,

$$\binom{a+b-1}{a} = \frac{(a+b-1)!}{a!(b-1)!} \tag{7.23}$$

are not favourable. To prove this, let us consider the first tie in an unfavourable voting record. After the first tie, let us change every vote into opposite. Then we obtain a voting record which contains a votes for B and b votes for A and the first vote is registered for A. The number of such voting records is given by (7.23). Conversely, if we consider a voting record of the latter type, and if after the first tie in this voting record we change every subsequent vote into opposite, then we obtain an unfavourable voting record. There is a one-to-one correspondence between the voting records in these two sets. Consequently,

$$N(a,b,1) = \binom{a+b-1}{b} - \binom{a+b-1}{a} \tag{7.24}$$

which is in agreement with (7.8). As we have already seen, (7.1) follows from (7.8). In the above proof, when after the first tie we changed each subsequent vote into its opposite, we actually applied the reflection principle.

Now we have several different proofs for Theorem 7.2.1 if $\mu \geq 1$. In 1947 Dvoretzky and Motzkin (1947) observed that if we consider any voting record which contains a votes for A and b votes for B where $a > b\mu$ and if we form all the $a+b$ cyclic permutations of this voting record, then there are exactly $a - b\mu$ cyclic permutations which are favourable, that is, throughout the counting the number of votes recorded for A is always greater than μ times the number of votes recorded for B. Hence, (7.1) immediately follows. For a geometric interpretation of this proof, see Grossman (1950), Yaglom and Yaglom (1954, pp. 172-175 and p. 184) and Mohanty (1966).

7.5 The General Ballot Theorem

The following theorem which is a generalization of Theorem 7.2.1 is usually called the *general ballot theorem*.

Theorem 7.5.1 *Let us suppose that a box contains n cards marked with non-negative integers $k_1, k_2, \ldots, k_n$ respectively, where $k_1 + k_2 + \cdots + k_n = k \leq n$. All the n cards are drawn without replacement from the box. Denote by ν_r the number obtained at the r-th drawing ($r = 1, 2, \ldots, n$). Then*

$$\Pr[\nu_1 + \nu_2 + \cdots + \nu_r < r \text{ for } r = 1, 2, \ldots, n] = (n-k)/n, \tag{7.25}$$

provided that all the possible results are equally probable.

To demonstrate that Theorem 7.2.1 is a particular case of Theorem 7.5.1, let us consider a box which contains a cards marked 0 and b cards marked $\mu+1$. We draw all the $a+b$ cards from the box without replacement, assuming that all the possible results are equally probable. Let us suppose that a card marked 0 corresponds to a vote for A, and a card marked $\mu+1$ corresponds to a vote for B. If among the first r drawings there are α_r cards marked 0 and β_r cards marked $\mu+1$, then $\alpha_r 0 + \beta_r(\mu+1) < r = \alpha_r + \beta_r$ holds if and only if $\alpha_r > \beta_r \mu$. Now, $\alpha_1 + \alpha_2 + \cdots + \alpha_{a+b} = a$ and $\beta_1 + \beta_2 + \cdots + \beta_{a+b} = b$. If $a \geq b\mu$ and if in (7.25) we put $n = a + b$ and $k = b(\mu + 1)$, we obtain

$$\Pr[\alpha_r > \beta_r \mu \text{ for } r = 1, 2, \ldots, a+b] = (a - b\mu)/(a+b) \tag{7.26}$$

which is in agreement with (7.1). See Takács (1961, 1962).

We can formulate Theorem 7.5.1 in the following equivalent way:

Theorem 7.5.2 *Let us suppose that n cards are marked with non-negative integers $k_1, k_2, \ldots, k_n$ respectively, where $k_1 + k_2 + \cdots + k_n = k \leq n$. Among the $n!$ permutations of the n cards, there are exactly*

$$S(n, k) = (n-k)(n-1)! \tag{7.27}$$

permutations in which the sum of the numbers on the first r cards is less than r for every $r = 1, 2, \ldots, n$.

PROOF. We can prove by mathematical induction that $S(n, k)$ does not depend individually on $k_1, k_2, \ldots, k_n$, it depends only on their sum k and their number n, and is given by (7.27). Obviously, $S(1, 0) = 1$ and $S(1, 1) = 0$. Let us suppose that $S(m, k) = (m-k)(m-1)!$ for $0 \leq k \leq m \leq n-1$ where $n \geq 2$. If we take into consideration that the last card in the $n!$ permutations of the n cards may be marked $k_1, k_2, \ldots, k_n$, then we can write down that

$$S(n, k) = \sum_{i=1}^{n} S(n-1, k-k_i) \tag{7.28}$$

for $k < n$ and $S(n, n) = 0$. If $k < n$, then by the induction hypothesis

$$S(n, k) = \sum_{i=1}^{n} (n - 1 - k + k_i)(n-2)! = (n-k)(n-1)!. \tag{7.29}$$

Consequently, (7.27) is true for all $n = 1, 2, \ldots$ and $0 \leq k \leq n$. ■

If in Theorem 7.5.2 we replace the $n!$ permutations by n cyclic permutations, we obtain the following result.

Theorem 7.5.3 *Let us suppose that n cards are marked with non-negative integers $k_1, k_2, \ldots, k_n$ respectively, where $k_1 + k_2 + \cdots + k_n = k \leq n$. Among the n cyclic permutations of the n cards, there are exactly $n - k$ in which the sum of the numbers on the first r cards is less than r for every $r = 1, 2, \ldots, n$.*

PROOF. Let $k_{r+n} = k_r$ for $r = 1, 2, \ldots$ and set $\varphi_r = k_1 + k_2 + \cdots + k_r$ for $r = 1, 2, \ldots; \varphi_0 = 0$. Define

$$\delta_r = \begin{cases} 1 & \text{if } i - \varphi_i > r - \varphi_r \text{ for } i > r, \\ 0 & \text{otherwise,} \end{cases} \tag{7.30}$$

and

$$\psi_r = \inf\{i - \varphi_i \text{ for } i \geq r\} \tag{7.31}$$

for $r = 0, 1, 2, \ldots$. Evidently, $\delta_r = \psi_{r+1} - \psi_r$. Since $\varphi_{r+n} = \varphi_r + \varphi_n$, we have $\delta_{r+n} = \delta_r$ and $\psi_{r+n} = \psi_r + n - k$ for $r = 0, 1, 2, \ldots$. Therefore, among the n cyclic permutations of $(k_1, k_2, \ldots, k_n)$, there are exactly

$$\sum_{r=1}^{n} \delta_r = \psi_{n+1} - \psi_1 = n - k \tag{7.32}$$

for which the sum of the first r elements is less than r for $r = 1, 2, \ldots, n$. ∎

Obviously, Theorem 7.5.3 implies Theorem 7.5.2. Theorem 7.5.3 can also be formulated in the following more general way.

Theorem 7.5.4 *Let $\nu_1, \nu_2, \ldots, \nu_n$ be interchangeable or cyclically interchangeable discrete random variables which take on non-negative integers only. Write $N_r = \nu_1 + \nu_2 + \cdots + \nu_r$ for $r = 1, 2, \ldots, n$ and $N_0 = 0$. We have*

$$\Pr[N_r < r \text{ for } 1 \leq r \leq n \text{ and } N_n = n - i] = \frac{i}{n} \Pr[N_n = n - i] \tag{7.33}$$

for $0 \leq i \leq n$ and $n = 1, 2, \ldots$.

The following result is a Corollary of Theorem 7.5.4.

Theorem 7.5.5 *Let $\nu_1, \nu_2, \ldots, \nu_n$ be interchangeable discrete random variables which take on non-negative integers only. Write $N_r = \nu_1 + \nu_2 + \cdots + \nu_r$ for $r = 1, 2, \ldots, n$ and $N_0 = 0$. We have*

$$\Pr[N_r < r \text{ for at least one } r = 1, 2, \ldots, n] = \sum_{i=1}^{n} \frac{1}{i} \Pr[N_i = i - 1] \tag{7.34}$$

for $n = 1, 2, \ldots$.

PROOF. The event that $N_r < r$ for some $r = 1, 2, \ldots, n$ can occur in several mutually exclusive ways: there is an $i = 1, 2, \ldots, n$ such that $N_i = i - 1$ and $N_i - N_r < i - r$ for $0 \leq r < i$. Since $\nu_1, \nu_2, \ldots, \nu_n$ are interchangeable random variables, we have

$$\begin{aligned}
&\Pr[N_r < r \text{ for at least one } r = 1, 2, \ldots, n] \\
&\quad = \sum_{i=1}^{n} \Pr[N_i - N_r < i - r \text{ for } 0 \leq r < i \text{ and } N_i = i - 1] \\
&\quad = \sum_{i=1}^{n} \Pr[N_r < r \text{ for } 1 \leq r \leq i \text{ and } N_i = i - 1]
\end{aligned} \tag{7.35}$$

and by (7.33)

$$\Pr[N_r < r \text{ for } 1 \leq r \leq i \text{ and } N_i = i - 1] = \frac{1}{i} \Pr[N_i = i - 1]. \tag{7.36}$$

for $1 \leq i \leq n$. This proves (7.34). ∎

7.6 Some Combinatorial Identities

First Passage Time

By using Theorem 7.5.4, we can prove the following general result.

Theorem 7.6.1 *Let $\nu_1, \nu_2, \ldots, \nu_r, \ldots$ be interchangeable discrete random variables which take on non-negative integers only. Write $N_r = \nu_1 + \nu_2 + \cdots + \nu_r$ for $r \geq 1$ and $N_0 = 0$. Let $S_r = r - N_r$ for $r \geq 0$ and define*

$$\rho(k) = \inf\{r : S_r = k, r \geq 0\} \tag{7.37}$$

for $k = 0, 1, 2, \ldots$. If $S_r < k$ for all $r \geq 0$, then $\rho(k) = \infty$. We have

$$\Pr[\rho(k) = n] = \frac{k}{n} \Pr[S_n = k] \tag{7.38}$$

for $n \geq 1$ and $k \geq 0$.

PROOF. If $k > n$, then both sides of (7.38) are 0. If $0 \leq k \leq n$, and $n \geq 1$, then by (7.33)

$$\begin{aligned}
\Pr[\rho(k) = n] &= \Pr[r - N_r < k \text{ for } 0 \leq r < n \text{ and } N_n = n - k] \\
&= \Pr[N_n - N_r < n - r \text{ for } 0 \leq r < n \text{ and } N_n = n - k] \\
&= \Pr[N_i < i \text{ for } 1 \leq i \leq n \text{ and } N_n = n - k] \\
&= \frac{k}{n} \Pr[N_n = n - k].
\end{aligned} \tag{7.39}$$

∎

Two Identities

If in Theorem 7.6.1, we assume that $\nu_1, \nu_2, \ldots, \nu_r, \ldots$ are independent and identically distributed discrete random variables which take on non-negative integers only, then we have the following identities:

$$\Pr[\rho(k + l) = n] = \sum_{j=0}^{n} \Pr[\rho(k) = j] \Pr[\rho(l) = n - j] \tag{7.40}$$

and

$$\Pr[S_n = k + l] = \sum_{j=0}^{n} \Pr[\rho(k) = j] \Pr[S_{n-j} = l] \tag{7.41}$$

for $k \geq 0, l \geq 0$ and $n \geq 0$.

The first identity is valid, because

$$\rho(k + l) = \rho(k) + [\rho(k + l) - \rho(k)] \tag{7.42}$$

where $\rho(k)$ and $\rho(k+l) - \rho(k)$ are independent and $\rho(k+l) - \rho(k)$ has the same distribution as $\rho(l)$. The second identity is valid, because $S_n - S_j$ is independent of S_j and has the same distribution as S_{n-j}.

The Identities of Rothe and Hagen

If we use the notation

$$A_m(\alpha,\beta) = \binom{\alpha+\beta m}{m}\frac{\alpha}{\alpha+\beta m}\,, \tag{7.43}$$

then by the results of Rothe (1793), Schläfli (1847), Hagen (1891, pp. 64-68), Gould (1956b, 1957) and Blackwell and Dubins (1966), we have

$$\sum_{i=0}^{m} A_i(\alpha,\beta)A_{m-i}(\gamma,\beta) = A_m(\alpha+\gamma,\beta), \tag{7.44}$$

and by the results of Hagen (1891) and Gould (1956b, 1957),

$$\sum_{i=0}^{m} iA_i(\alpha,\beta)A_{m-i}(\gamma,\beta) = \frac{m\alpha}{\alpha+\gamma}A_m(\alpha+\gamma,\beta) \tag{7.45}$$

for $m = 1, 2, \ldots$ and arbitrary α, β and γ. If $\alpha = \gamma$, then (7.44) implies (7.45) and conversely.

Both (7.44) and (7.45) can be proved by using the general ballot theorem. As a matter of fact, (7.44) is a particular case of (7.40), and (7.45) is a particular case of (7.41). To demonstrate this let us assume that

$$\Pr[\nu_r = b] = p \text{ and } \Pr[\nu_r = 0] = q \tag{7.46}$$

where $p > 0, q > 0, p + q = 1$, and b is a positive integer. Then $\{S_n, n \geq 0\}$ is a random walk. We have

$$\Pr[S_{mb+k} = k] = \binom{mb+k}{m}p^m q^{m(b-1)+k}, \tag{7.47}$$

and by (7.38),

$$\Pr[\rho(k) = mb+k] = \frac{k}{mb+k}\ \Pr[S_{mb+k} = k] = A_m(k,b)p^m q^{m(b-1)+k} \tag{7.48}$$

for $k \geq 0$ and $m \geq 0$.

Now by (7.40),

$$\Pr[\rho(k+l) = mb+k+l] = \sum_{i=0}^{m} \Pr[\rho(k) = ib+k]\Pr[\rho(l) = (m-i)b+l]. \tag{7.49}$$

This proves (7.44) for $\alpha = k, \gamma = l$ and $\beta = b$. By (7.41),

$$\Pr[S_{mb+k+l} = k+l] = \sum_{i=0}^{m} \Pr[\rho(k) = ib+k]\Pr[S_{(m-i)b+l} = l]. \tag{7.50}$$

This proves (7.45) for $\alpha = k, \gamma = l$ and $\beta = b$. Since (7.44) and (7.45) are polynomials in α, β and γ, the identities (7.44) and (7.45) are valid for any real or complex α, β and γ.

An Identity of Chung

In 1946 K. L. Chung proposed a problem for solution in the *American Mathematical Monthly*. His problem can be restated as follows: Prove that

$$\binom{bn}{n} = \sum_{k=1}^{n} \frac{b-1}{bk-1}\binom{bk}{k}\binom{bn-bk}{n-k} \tag{7.51}$$

for $b = 2, 3, \ldots$ and $n = 1, 2, \ldots$. If $n = 1$ or $n = 2$, then (7.51) is trivially true. If $n > 2$, then it is easy to check that the identity (7.51) follows from either (7.44) or (7.45) where $m = n-2$, $\alpha = \gamma = b-1$ and $\beta = b$.

The problem of Chung (1946) was solved by Gould (1956a,b). See also Guy (1984). Gould (1956a,b) proved also (7.44) and (7.45), and generalized these formulas. For further extensions of (7.44) and (7.45), see Gould and Kaucký (1966) and Knuth (1992).

7.7 Another Extension of The Classical Ballot Theorem

Let us suppose again that in a ballot, candidate A scores a votes and candidate B scores b votes and all the possible $\binom{a+b}{b}$ voting records are equally probable. Denote by α_r and β_r the number of votes registered for A and B respectively among the first r votes counted. Let μ be a positive real number and define

$$P_j(a,b,\mu) = \Pr[\alpha_r > \beta_r\mu \text{ for } j \text{ subscripts } r = 1, 2, \ldots, a+b] \tag{7.52}$$

for $j = 0, 1, 2, \ldots, a+b$. We can write that

$$P_j(a,b,\mu) = N_j(a,b,\mu)\Big/\binom{a+b}{b} \tag{7.53}$$

for $j = 0, 1, 2, \ldots, a+b$. In what follows, we discuss the problem of finding $P_j(a,b,\mu)$ for $j = 0, 1, 2, \ldots, a+b$. For a survey of this topic, see Chao and Severo (1991) and Takács (1967).

If $P_0(a,b,\mu)$ and $P_{a+b}(a,b,\mu)$ are known for $a \geq 0$ and $b \geq 0$, then we can determine $P_j(a,b,\mu)$ for $j = 1, 2, \ldots, a+b-1$ by the following equation

$$N_j(a,b,\mu) = \sum_{0\leq s\leq j} N_j(j-s,s,\mu)N_0(a+s-j,b-s,\mu), \tag{7.54}$$

or by the equivalent formula

$$P_j(a,b,\mu) = \sum_{0\leq s\leq j} \Pr[\beta_j = s]P_j(j-s,s,\mu)P_0(a+s-j,b-s,\mu), \tag{7.55}$$

where

$$\Pr[\beta_j = s] = \frac{\binom{j}{s}\binom{a+b-j}{b-s}}{\binom{a+b}{b}} = \frac{\binom{a}{j-s}\binom{b}{s}}{\binom{a+b}{j}} \tag{7.56}$$

whenever $0 \leq s \leq j$ and $j - a \leq s \leq b$.

Formula (7.55) can be proved by making use of the following auxiliary theorem.

Theorem 7.7.1 *Let $\xi_1, \xi_2, \ldots, \xi_n$ be interchangeable real random variables. Define $\zeta_r = \xi_1 + \xi_2 + \cdots + \xi_r$ for $r = 1, 2, \ldots, n$ and $\zeta_0 = 0$. Denote by Δ_n the number of subscripts $r = 1, 2, \ldots, n$ for which $\zeta_r > 0$. Then,*

$$\Pr[\Delta_n = j] = \Pr[\zeta_r < \zeta_j \text{ for } 0 \leq r < j \text{ and } \zeta_r \leq \zeta_j \text{ for } j \leq r \leq n]. \tag{7.57}$$

PROOF. Formula (7.57) was proved by Andersen (1954). In Feller (1959), he deduced (7.57) from a simple elementary combinatorial theorem. ∎

Since the random variables $\xi_1, \xi_2, \ldots, \xi_n$ are interchangeable, we can express (7.57) in the following equivalent form:

$$\Pr[\Delta_n = j] = \Pr\left[\min_{1 \leq r < j} \zeta_r > 0 \text{ and } \max_{j \leq r \leq n} (\zeta_r - \zeta_j) \leq 0\right]. \tag{7.58}$$

If in the ballot problem, we define $\zeta_r = \alpha_r - \beta_r\mu = r - \beta_r(\mu + 1)$ for $r = 1, 2, \ldots, a + b$ and $\zeta_0 = 0$, then Theorem 7.7.1 can be applied to the random variables ζ_r $(0 \leq r \leq a + b)$. Under the condition that $\beta_j = s$, that is, $\zeta_j = j - s(\mu + 1)$, where $0 \leq s \leq j$, we obtain that $\min_{1 \leq r < j} \zeta_r > 0$ is satisfied if and only if $\alpha_r > \beta_r\mu$ for $1 \leq r < j$, and $\alpha_j = j - s$ and $\beta_j = s$, and also $\max_{j \leq r \leq n}(\zeta_r - \zeta_j) \leq 0$ is satisfied if and only if $\alpha_r - \alpha_j \leq (\beta_r - \beta_j)\mu$ for $j \leq r \leq a + b$, and $\alpha_{a+b} - \alpha_j = a + s - j$ and $\beta_{a+b} - \beta_j = b - s$. Consequently, in this case (7.57) proves (7.55).

If, in particular, μ is a positive integer, then by Theorem 7.2.1

$$P_{a+b}(a, b, \mu) = (a - b\mu)/(a + b) \tag{7.59}$$

for $a > b\mu$ and $P_{a+b}(a, b, \mu) = 0$ if $a \leq b\mu$. Thus, in formula (7.55), we have

$$P_j(j - s, s, \mu) = (j - s\mu - s)/j \tag{7.60}$$

if $0 \leq s < j/(\mu + 1)$ and $P_j(j - s, s, \mu) = 0$ otherwise.

If, in particular, μ is a positive integer, then by Theorem 7.5.5 we can prove that

$$P_0(a, b, \mu) = N_0(a, b, \mu) \Big/ \binom{a+b}{b}, \tag{7.61}$$

where

$$N_0(a,b,\mu) = \binom{a+b}{b} - \sum_{0\le s\le (a+b-1)/(\mu+1)} \binom{s\mu+s+1}{s} \times \binom{a+b-s\mu-s-1}{b-s} \frac{1}{(s\mu+s+1)} \tag{7.62}$$

if $a \le b\mu$ and $N_0(a,b,\mu) = 0$ if $a > b\mu$. To obtain (7.62), let us suppose that in Theorem 7.5.5, $n = a+b$ and define the random variables $\nu_1, \nu_2, \ldots, \nu_{a+b}$ in the following way: $\nu_r = 0$ if the rth vote is cast for A and $\nu_r = \mu + 1$ if the rth vote is cast for B. Then $N_r = \beta_r(\mu+1)$ and $\alpha_r > \beta_r\mu$ if and only if $r > N_r$. Thus, by Theorem 7.5.5,

$$P_0(a,b,\mu) = 1 - \sum_{i=1}^{a+b} \frac{1}{i} \Pr[N_i = i-1] \tag{7.63}$$

where now

$$\Pr[N_i = i-1] = \Pr[\beta_i(\mu+1) = i-1] \tag{7.64}$$

and the distribution of β_i is determined by (7.56). In (7.64), necessarily $i = s(\mu+1)+1$ where $0 \le s \le (a+b-1)/(\mu+1)$. By substituting $a+s-j$ for a and $b-s$ for b in (7.62), we obtain $P_0(a+s-j, b-s, \mu)$ in (7.55).

The solutions of the above mentioned ballot problems make it possible, for example, to find the distributions of the local times and the sojourn times for various stochastic processes.

References

1. Aeppli, A. (1923). A propos de l'interprétation géométrique du problème du scrutin, *L'Enseignement Mathématique*, **23**, 328–329.

2. Aeppli, A. (1924). *Zur Theorie verketteter Wahrscheinlichkeiten. Markoffsche Ketten höherer Ordnung*, Dissertation, Eidgenössische Technische Hochschule, Zürich.

3. Ampère, A. M. (1802). *Considérations sur la Théorie Mathématique de Jeu*, Lyon et Paris.

4. Andersen, E. S. (1954). On fluctuations of sums of random variables. II, *Mathematica Scandinavica*, **2**, 195–223.

5. André, D. (1887). Solution directe du problème résolu par M. Bertrand, *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, **105**, 436–437.

6. Apianus, P. (1527). *Ein newe und wolgegründte Underweisung aller Kauffmannß Rechnung in dreien Büchern*, Ingolstadt.

7. Barbier, É. (1887). Généralisation du problème résolu par M.J. Bertrand, *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, **105**, 407 and 440 (errata).

8. Bertrand, J. (1887). Solution d'un problème, *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences (Paris)*, **105**, 369.

9. Bertrand, J. (1889). *Calcul des Probabilités*, Paris, Second edition, 1907. [Reprinted by Chelsea, NY, 1973.]

10. Blackwell, D. and Dubins, L. (1966). An elementary proof of an identity of Gould's, *Boletin de la Sociedad Matématica Mexicana*, **11**, 108–110.

11. Cardano, H. (1570). *Opus Novum de Proportionibus Numerorum*, Basel.

12. Chao, C. C. and Severo, N. C. (1991). Distributions of ballot problem random variables, *Advances in Applied Probability*, **23**, 586–597.

13. Chung, K. L. (1946). Problem 4211, *The American Mathematical Monthly*, **53**, 397.

14. De Moivre, A. (1711). De mensura sortis, seu, de probabilitate eventuum in ludis a casu fortuito pendentibus, *Philosophical Transactions*, **27**, 213–264.

15. De Moivre, A. (1718). *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*, First edition, London.

16. De Moivre, A. (1738). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, Second edition, London. [Reprinted by Frank Cass, London, 1967.]

17. De Moivre, A. (1756). *The Doctrine of Chances: or, A Method of Calculating the Probabilities of Events in Play*, Third edition, London. [Reprinted by Chelsea, New York, 1967.]

18. De Moivre, A. (1984). On the measurement of chance, or, on the probability of events in games depending upon fortuitous chance, *International Statistical Review*, **52**, 237–262. [English translation of A. De Moivre (1711) by B. McClintock.]

19. Dvoretzky, A. and Motzkin, Th. (1947). A problem of arrangements, *Duke Mathematical Journal*, **14**, 305–313.

20. Feller, W. (1959). On combinatorial methods in fluctuation theory, *Probability & Statistics: The Harald Cramér Volume* (Ed., U. Grenander), pp. 75–91, Stockholm: Almqvist and Wiksell.

21. Gould, H. W. (1956a). A finite sum. Solution of Problem 4211, *The American Mathematical Monthly*, **63**, 126–127.

22. Gould, H. W. (1956b). Some generalizations of Vandermonde's convolution, *The American Mathematical Monthly*, **63**, 84–91.

23. Gould, H. W. (1957). Final analysis of Vandermonde's convolution, *The American Mathematical Monthly*, **64**, 409–415.

24. Gould, H. W. and Kaucký, J. (1966). Evaluation of a class of binomial coefficient summations, *Journal of Combinatorial Theory*, **1** 233–247. [Errata *Journal of Combinatorial Theory, Series A*, **12** (1972) 309–310.]

25. Grossman, H. D. (1950). Another extension of the ballot problem, *Scripta Mathematica*, **16**, 119–124.

26. Guy, R. K. (1984). A pentagonal pot-pouri of perplexing problems, primarily probabilistic, *The American Mathematical Monthly*, **91**, 559–563.

27. Hagen, J. G. (1891). *Synopsis der höheren Mathematik*, Volume 1, Berlin.

28. Hald, A. (1984). A. de Moivre: 'De mensura sortis' or 'On the measurement of chance', *International Statistical Review*, **52**, 229–236.

29. Knuth, D. E. (1992). Convolution polynomials, *The Mathematica Journal*, **2**, 67–78.

30. Lagrange, J. L. (1777). Recherches sur les suites récurrentes dont les termes varient de plusieurs manières différentes, ou sur l'intégration des équations linéaires aux différences finies et partielles; et sur l'usage de ces équations dans la théorie des hasards, *Nouveaux Mémoires de l'Académie Royale des Sciences et Belles-Lettres de Berlin*, année 1775, 183–272. [*Oeuvres de Lagrange*, IV. Gauthier-Villars, Paris, 1869, pp. 151–251.]

31. Laplace, P. S. (1776). Recherches sur l'intégration des équations différentielles aux différences finies et sur leur usage dans la théorie des hasards, *Mémoires de l'Académie Royale des Sciences de Paris*, année 1773, **7**, 43–163. [*Oeuvres Complètes de Laplace.* Tome VIII. Gauthier-Villars, Paris, 1891, pp. 69–197.]

32. Laplace, P. S. (1812). *Théorie Analytique des Probabilités*, Courcier, Paris. [Reprinted by *Culture et Civilisation*, Bruxelles, 1967.]

33. Laplace, P. S. (1814, 1820). *Théorie Analytique des Probabilités*, Courcier, Paris. [*Oeuvres Complètes de Laplace.* Tome VII. Gauthier-Villars, Paris, 1886.]

34. Mersenne, M. (1635–1636). *Harmonicorum Libri*, Paris.

35. Mohanty, S. G. (1966). An urn problem related to the ballot problem, *The American Mathematical Monthly*, **73**, 526–528.

36. Pascal, B. (1665). *Traité du Triangle Arithmétique, avec quelques autres petits traitez sur la mesme matiere*, Paris. [B. Pascal (1908) T. III. Paris, 1908, pp. 433–593.]

37. Pascal, B. (1908–1925). *Oeuvres*, Tomes I–XIV (Edited by L. Brunschwicg, P. Boutroux and F. Gazier), Hachette, Paris. [T. III (1908).]

38. Poincaré, H. (1896). *Calcul des Probabilités*, Paris: Gauthier-Villars. Second edition, 1912.

39. Rothe, H. A. (1793). *Formulae de serierum reversione demonstratio universalis signis localibus combinatorio-analyticorum vicariis exhibita*, Dissertation, Leipzig.

40. Schläfli, L. (1847). Bemerkungen über die Lambertsche Reihe, *Archiv der Mathematik und Physik*, **10**, 332–340. [Reprinted in *Ludwig Schläfli* (1814–1895) *Gesammelte Mathematische Abhandlungen*, Band I. Birkhäuser, Basel, 1950, pp. 38–45.]

41. Takács, L. (1961). The probability law of the busy period for two types of queuing processes, *Operations Research*, **9**, 402–407.

42. Takács, L. (1962). A generalization of the ballot problem and its application in the theory of queues, *Journal of the American Statistical Association*, **57**, 327–337.

43. Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*, New York: John Wiley & Sons.

44. Takács, L. (1969). On the classical ruin problems, *Journal of the American Statistical Association*, **64**, 889–906.

45. Whitworth, W. A. (1879). Arrangements of m things of one sort and n things of another sort, under certain conditions of priority, *Messenger of Mathematics*, **8**, 105–114.

46. Yaglom, A. M. and Yaglom, I. M. (1964). *Challenging Mathematical Problems with Elementary Solutions*, Volume I, *Combinatorial Analysis and Probability Theory*, San Francisco: Holden-Day. [English translation of the Russian original published in Moscow, 1954.]

8

Some Results for Two-Dimensional Random Walk

Endre Csáki

Mathematical Institute of the Hungarian Academy of Sciences, Budapest, Hungary

Abstract: We present some results for simple symmetric two-dimensional random walk. Our treatment is based on results concerning independent simple symmetric one-dimensional random walks.

Keywords and phrases: Linear random walk, planar random walk

8.1 Introduction

Though the treatment of the two-dimensional random walk has a long history and goes back to Pólya (1921), McCrea and Whipple (1940), Dvoretzky and Erdős (1950), Erdős and Taylor (1960), Spitzer (1964), and others, there are some recent results of combinatorial nature like those of DeTemple and Robertson (1984), Csáki, Mohanty and Saran (1990), Breckenridge *et al.* (1991), Guy, Krattenthaler and Sagan (1992), Barcucci and Verri (1992), Kreweras (1992) and Saran and Rani (1994) concerning simple symmetric two-dimensional random walk. Further results can be found in Gupta and Sen (1977, 1979), Révész (1990) and Di Crescenzo, Giorno and Nobile (1992).

In this paper we present a treatment of two-dimensional (planar) random walk problems based on stochastically independent one-dimensional (linear) random walks. In fact, there are two ways of doing this. Let

$$\left\{X_n = \left(X_n^{(1)}, X_n^{(2)}\right)\right\}_{n=1}^{\infty} \tag{8.1}$$

be a sequence of i.i.d. random vectors with the distribution

$$\begin{aligned}\Pr[X_i = (0,1)] &= \Pr[X_i = (0,-1)] = \Pr[X_i = (1,0)] \\ &= \Pr[X_i = (-1,0)] = 1/4.\end{aligned} \tag{8.2}$$

Then $T_0 = 0$, $T_n = \sum_{i=1}^n X_i$, $n = 1, 2, \dots$ is called a *two-dimensional simple symmetric*, or *planar random walk* (PRW). Let $T_n = (T_n^{(1)}, T_n^{(2)})$, i.e. $T_n^{(j)} = \sum_{i=1}^n X_i^{(j)}$, $j = 1, 2$. Then we have the following result [*cf.* Spitzer (1964)].

Proposition 8.1.1

$$S_n^{(1)} = T_n^{(1)} + T_n^{(2)}, \qquad n = 1, 2, \dots$$

$$S_n^{(2)} = T_n^{(1)} - T_n^{(2)}, \qquad n = 1, 2, \dots$$

are two stochastically independent one-dimensional simple symmetric, or linear random walks (LRW).

Thus if $A = A_1 \cap A_2$, where A_j are events measurable with respect to $S^{(j)}$, $j = 1, 2$, then

$$\Pr[A] = \Pr[A_1 \cap A_2] = \Pr[A_1] \Pr[A_2]. \tag{8.3}$$

The other way is to consider the coordinates $T_n^{(1)}$ and $T_n^{(2)}$ as one-dimensional random walks with possible steps $-1, 0, 1$. By eliminating the zero steps, we obtain two LRW-s, $H_1, H_2, \dots$ (from $T^{(1)}$) and $V_1, V_2, \dots$ (from $T^{(2)}$) with ± 1 steps. Let ν_n denote the number of non-zero steps in $T_1^{(2)}, \dots, T_n^{(2)}$ (number of vertical steps in the first n steps of PRW). Obviously, $n - \nu_n$ is the number of non-zero steps in $T_1^{(2)}, \dots, T_n^{(2)}$ (number of horizontal steps in the first n steps of PRW). Given $\nu_n = k$, the LRW-s $H_1, \dots, H_{n-k}$ and $V_1, \dots, V_k$ are conditionally independent. Clearly,

$$\Pr[\nu_n = k] = \frac{1}{2^n} \binom{n}{k}. \tag{8.4}$$

Therefore, if $B = B_1 \cap B_2$, where B_1 is an event measurable with respect to $H_1, \dots, H_{n-\nu_n}$ and B_2 is an event measurable with respect to $V_1, \dots, V_{\nu_n}$, then

$$\Pr[B] = \sum_{k=0}^{n} \frac{1}{2^n} \binom{n}{k} \Pr[B_1 | \nu_n = k] \Pr[B_2 | \nu_n = k]. \tag{8.5}$$

We illustrate by simple examples how the basic identities (8.3) and (8.5) will be used. Let $A = \{T_n = (c, d)\}$, i.e., the PRW path ends at the point (c, d). We have $\{T_n = (c, d)\} = \{S_n^{(1)} = c + d\} \cap \{S_n^{(2)} = c - d\}$, and so by (8.3) we obtain the well-known formula

$$\Pr[T_n = (c, d)] = \frac{1}{4^n} \binom{n}{\frac{n+c+d}{2}} \binom{n}{\frac{n+c-d}{2}}, \tag{8.6}$$

where the binomial coefficient $\binom{n}{k}$ is meant to be zero if k is not an integer satisfying $0 \le k \le n$. On the other hand, it is easy to see that under the condition $\{\nu_n = k\}$ we have $\{T_n = (c,d)\} = \{H_{n-k} = c\} \cap \{V_k = d\}$, and hence we obtain

$$\Pr[T_n = (c,d)] = \frac{1}{4^n} \sum_{k=0}^{n} \binom{n}{k} \binom{k}{\frac{k+d}{2}} \binom{n-k}{\frac{n-k+c}{2}}. \tag{8.7}$$

Comparing (8.6) with (8.7) we obtain the binomial identity

$$\sum_{k=0}^{n} \binom{n}{k} \binom{k}{\frac{k+d}{2}} \binom{n-k}{\frac{n-k+c}{2}} = \binom{n}{\frac{n+c+d}{2}} \binom{n}{\frac{n+c-d}{2}}, \tag{8.8}$$

which can also be obtained from Vandermonde convolution [*cf.* Gould (1972)].

The identity (8.3) was used in Csáki, Mohanty and Saran (1990) to derive some distributions concerning the boundaries $y = x + a$ and $y = -x + b$. For further (joint) distributions concerning these boundaries, one may refer to Saran and Rani (1994).

In Breckenridge *et al.* (1991), Guy, Krattenthaler and Sagan (1992) and Barcucci and Verri (1992), bijections between PRW and LRW paths were given to derive certain results concerning the boundaries $x = 0$, $y = 0$. It was shown among others that for PRW paths not crossing the x-axis, we have

$$\Pr[T_i^{(2)} \geq 0,\ i = 1, \ldots, n] = \frac{1}{4^n} \binom{2n+1}{n} \tag{8.9}$$

[see also Sands (1990)]. In view of Proposition 8.1.1, this is equivalent to

$$\Pr[S_i^{(1)} \geq S_i^{(2)},\ i = 1, \ldots, n] = \frac{1}{4^n} \binom{2n+1}{n}. \tag{8.10}$$

Pairs of LRW paths were studied in Karlin and McGregor (1959), Raifaizen (1972), Shapiro (1976a) and Karlin (1988). In Raifaizen (1972), a bijection is given between a pair of LRW paths each of length n and one LRW path of length $2n$ as follows: Let $Y_{2i} = Y_i^{(1)}$ and $Y_{2i-1} = Y_i^{(2)}$, $i = 1, \ldots, n$, where $Y_i^{(j)}$ is the i-th step of $S^{(j)}$. Then $Z_k = Y_1 + Y_2 + \ldots + Y_k$, $k = 1, \ldots, 2n$, is a LRW path in which steps are taken from $S^{(1)}$ and $S^{(2)}$ alternatively. This bijection combined with Proposition 8.1.1 gives a direct bijection between PRW paths of length n and LRW paths of length $2n$. Similar bijections were given in Breckenridge *et al.* (1991), Guy, Krattenthaler and Sagan (1992) and Barcucci and Verri (1992). For example, a PRW path satisfying $T_i^{(2)} \geq 0$, $i = 1, \ldots, n$, can be transformed into a LRW path satisfying $Z_{2k} \geq 0$, $k = 1, \ldots, n$, or equivalently $Z_k \geq -1$, $k = 1, \ldots, 2n$, giving (8.10).

8.2 Identities and Distributions

First we put together known results to derive certain binomial identities. It follows from Guy, Krattenthaler and Sagan (1992) that

$$\Pr[T_n^{(2)} = r] = \frac{1}{4^n}\binom{2n}{n+r}. \tag{8.11}$$

On the other hand, $\{T_n^{(2)} = r\} = \{V_{\nu_n} = r\}$ and so by (8.5) we have the identity

$$\frac{1}{4^n}\binom{2n}{n+r} = \sum_{k=0}^{n} \frac{1}{2^n}\binom{n}{k}\frac{1}{2^k}\binom{k}{\frac{k+r}{2}} \tag{8.12}$$

[see also 3.22 in Gould (1972)]. We note that the identity (8.12) is equivalent to

$$\frac{r}{n}\binom{2n}{n+r} = \sum_{k=1}^{n}\binom{n-1}{k-1}2^{n-k-1}\frac{r}{k}\binom{k}{\frac{k+r}{2}}, \tag{8.13}$$

giving an identity for the ballot numbers $B_{r,n} = \frac{r}{n}\binom{2n}{n+r}$. Putting $r = 2$ into (8.13), we obtain an identity of Touchard (1928) that

$$\sum_{(k)}\binom{n-1}{2k}2^{n-1-2k}C_k = C_n, \tag{8.14}$$

where $C_n = (n+1)^{-1}\binom{2n}{n}$ is the n-th Catalan number; see also Shapiro (1976b), Breckenridge *et al.* (1991) and Barcucci and Verri (1992). For other extensions of Touchard's identity, one may refer to Gould (1977).

Now we determine the probability that a PRW path stays in a strip $-a < T_i^{(2)} < b$ and ends on the x-axis after n steps. For $a > 0,\ b > 0$, we can write by (8.5)

$$\begin{aligned} &\Pr[-a < T_i^{(2)} < b,\ i = 1,\ldots,n,\ T_n^{(2)} = 0] \\ &= \sum_{(k)} \frac{1}{2^n}\binom{n}{2k}\Pr[-a < V_i < b,\ i = 1,\ldots,2k,\ V_{2k} = 0]. \end{aligned} \tag{8.15}$$

It is well-known that the latter probability can be given as [*cf.* Mohanty (1979)]

$$\begin{aligned} &\Pr[-a < V_i < b,\ i = 1,\ldots,2k,\ V_{2k} = 0] \\ &= \frac{1}{2^{2k}}\sum_{j=-\infty}^{\infty}\left\{\binom{2k}{k+j(a+b)} - \binom{2k}{k+a+j(a+b)}\right\}. \end{aligned} \tag{8.16}$$

Using this in (8.15) and applying the identity (8.12), we get

$$\begin{aligned}&\Pr[-a < T_i^{(2)} < b,\ i = 1, \ldots, n,\ T_n^{(2)} = 0] \\ &= \frac{1}{4^n} \sum_{j=-\infty}^{\infty} \left\{ \binom{2n}{n + 2j(a+b)} - \binom{2n}{n + 2a + 2j(a+b)} \right\}. \qquad (8.17)\end{aligned}$$

Next we determine the distribution of the number of times a PRW path crosses the x-axis. We say that a crossing occurs at i if either $T_i^{(2)} > 0,\ T_{i+1}^{(2)} = \ldots T_{i+j-1}^{(2)} = 0,\ T_{i+j}^{(2)} < 0$ or $T_i^{(2)} < 0,\ T_{i+1}^{(2)} = \ldots T_{i+j-1}^{(2)} = 0,\ T_{i+j}^{(2)} > 0$ for some $j > 1$. Let λ_n denote the number of such crossings completed before the n-th step. Then from (8.5), we have

$$\Pr[\lambda_n = \ell - 1,\ T_n^{(2)} = 0] = \sum_{(k)} \frac{1}{2^n} \binom{n}{2k} \Pr[\lambda_{2k}^{(2)} = \ell - 1,\ V_{2k} = 0], \qquad (8.18)$$

where $\lambda_{2k}^{(2)}$ is the number of times the LRW path $V_1, \ldots, V_{2k}$ crosses 0. Using the result

$$\Pr[\lambda_{2k}^{(2)} = \ell - 1,\ V_{2k} = 0] = \frac{1}{2^{2k}} \frac{2\ell}{k} \binom{2k}{k + \ell} \qquad (8.19)$$

from Csáki and Vincze (1961) and the identity (8.13), we get

$$\Pr[\lambda_n = \ell - 1,\ T_n^{(2)} = 0] = \frac{1}{4^n} \frac{4\ell + 2}{n + 1} \binom{2n + 2}{n + 2\ell + 2}. \qquad (8.20)$$

Now let ρ_k denote the time of k-th visit to the x-axis by a PRW path, i.e., $\rho_0 = 0$ and

$$\rho_k = \min\{i : \ i > \rho_{k-1},\ T_i^{(2)} = 0\}. \qquad (8.21)$$

Then by Proposition 8.1.1, the event $\{\rho_k = n\}$ is the same as $\{S^{(1)}$ and $S^{(2)}$ meet k-th time at $n\}$. Its probability is given in Raifaizen (1972) from which

$$\Pr[\rho_k = n] = \frac{1}{2^{2n-k}} \frac{k}{2n - k} \binom{2n - k}{n}. \qquad (8.22)$$

This distribution is the same as that of the k-th visit to zero by a LRW path (replacing n by $2n$).

Let ξ_n denote the number of visits to the x-axis by a PRW path, i.e.,

$$\xi_n = \#\{i : \ 1 \le i \le n,\ T_i^{(2)} = 0\}. \qquad (8.23)$$

Then (8.22) also gives

$$\Pr[\xi_n = k,\ T_n^{(2)} = 0] = \frac{1}{2^{2n-k}} \frac{k}{2n - k} \binom{2n - k}{n}. \qquad (8.24)$$

Moreover, it follows that the distribution of ξ_n is the same as that of the number of visits to zero of a LRW up to time $2n$. Hence we have [*cf.* Feller (1968)]

$$\Pr[\xi_n = k] = \frac{1}{2^{2n-k}}\binom{2n-k}{n}, \qquad k = 0, 1, \ldots, n. \tag{8.25}$$

One can similarly see that for τ_n, the last visit to the x-axis before the n-th step we have the arcsine law

$$\Pr[\tau_n = k] = \frac{1}{2^{2n}}\binom{2k}{k}\binom{2n-2k}{n-k}, \qquad k = 0, 1, \ldots, n. \tag{8.26}$$

Other distributions can be derived similarly.

8.3 Pairs of LRW Paths

In this section, we study certain properties of pairs of independent LRW-s and apply the results for PRW. Let $\{S_i^{(1)}\}_{i=1}^{\infty}$ and $\{S_i^{(2)}\}_{i=1}^{\infty}$ be two independent LRW-s. First, consider the probability

$$\Pr[S_0^{(1)} = 2a,\ S_0^{(2)} = 0,\ S_i^{(1)} > S_i^{(2)},\ i = 1, \ldots, 2n,\ S_{2n}^{(1)} = 2a + 2\ell,\ S_{2n}^{(2)} = 2k],$$

where $a > 0$, $a + \ell > k$, i.e., the probability that the two paths with given starting and terminating points do not meet in the first $2n$ steps. For determining the probability of the complement, i.e. that the two paths meet somewhere in the first $2n$ steps, we use a version of the reflection principle due to Karlin and McGregor (1959): Let κ be the smallest i for which $S_i^{(1)} = S_i^{(2)}$ and decompose both paths into two parts: $(S_0^{(j)}, S_1^{(j)}, \ldots, S_\kappa^{(j)})$, $(S_{\kappa+1}^{(j)}, \ldots S_{2n}^{(j)})$, $j = 1, 2$. Leaving the first parts as they are and interchanging the second parts, we get two new paths as $(S_0^{(1)}, S_1^{(1)}, \ldots, S_\kappa^{(1)}, S_{\kappa+1}^{(2)}, \ldots, S_{2n}^{(2)})$ and $(S_0^{(2)}, S_1^{(2)}, \ldots, S_\kappa^{(2)}, S_{\kappa+1}^{(1)}, \ldots, S_{2n}^{(1)})$. Now the endpoints of the two paths interchange and hence the new paths should meet somewhere. Hence, this is a bijection showing that

$$\begin{aligned}
\Pr[S_0^{(1)} &= 2a,\ S_0^{(2)} = 0,\ S_i^{(1)} > S_i^{(2)},\ i = 1, \ldots, 2n, \\
&\qquad S_{2n}^{(1)} = 2a + 2\ell,\ S_{2n}^{(2)} = 2k] \\
&= \Pr[S_0^{(1)} = 2a,\ S_0^{(2)} = 0,\ S_{2n}^{(1)} = 2a + 2\ell,\ S_{2n}^{(2)} = 2k] \\
&\quad - \Pr[S_0^{(1)} = 2a,\ S_0^{(2)} = 0,\ S_{2n}^{(1)} = 2k,\ S_{2n}^{(2)} = 2a + 2\ell] \\
&= \frac{1}{4^{2n}}\left\{\binom{2n}{n+k}\binom{2n}{n+\ell} - \binom{2n}{n+a-k}\binom{2n}{n+a+\ell}\right\}.
\end{aligned} \tag{8.27}$$

This is a particular case of the determinant formula of Karlin and McGregor (1959). From this, we can find the probability

$$\begin{aligned}&\Pr[S_0^{(1)} = S_0^{(2)} = 0,\ S_i^{(1)} \geq S_i^{(2)},\ i = 1, \ldots, 2n,\ S_{2n}^{(1)} = S_{2n}^{(2)} = 0]\\ &= \frac{1}{4^{2n}}\left\{\binom{2n}{n}^2 - \binom{2n}{n+1}^2\right\} = \frac{C_n}{4^{2n}}\binom{2n+1}{n}\end{aligned} \tag{8.28}$$

by shifting the path of $S^{(1)}$ two units upwards and putting $a = 2,\ k = \ell = 0$ into (8.27).

By putting $a = 2,\ k = 0$ into (8.27) and summing up for ℓ we obtain

$$\begin{aligned}&\Pr[S_0^{(1)} = S_0^{(2)} = 0,\ S_i^{(1)} \geq S_i^{(2)},\ i = 1, \ldots, 2n,\ S_{2n}^{(2)} = 0]\\ &= \frac{C_n}{2^{2n+1}} + \frac{1}{2^{4n+1}}\binom{2n}{n}\binom{2n+1}{n}.\end{aligned} \tag{8.29}$$

In view of Proposition 8.1.1, (8.28) also gives the probability that a PRW path does not cross the x-axis and returns to the origin after $2n$ steps and (8.29) gives the probability that a PRW does not cross the x-axis and ends on the half-diagonal $x = y \geq 0$. So we have

$$\Pr[T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n} = (0,0)] = \frac{C_n}{4^{2n}}\binom{2n+1}{n} \tag{8.30}$$

and

$$\Pr[T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)}] = \frac{C_n}{2^{2n+1}} + \frac{1}{2^{4n+1}}\binom{2n}{n}\binom{2n+1}{n}. \tag{8.31}$$

(8.30) and (8.31) also give

$$\begin{aligned}&\Pr[T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = \pm T_{2n}^{(2)}]\\ &= \Pr[T_i^{(1)}T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)}]\\ &= \frac{C_n}{2^{2n}} + \frac{1}{4^{2n}}\binom{2n}{n+1}\binom{2n+1}{n}.\end{aligned} \tag{8.32}$$

Next, consider PRW paths not crossing both the x- and y- axes. Guy, Krattenthaler and Sagan (1992), using reflection principle, give the formula

$$\begin{aligned}&\Pr[T_i^{(1)} \geq 0,\ T_i^{(2)} \geq 0,\ i = 1, \ldots, n,\ T_n^{(1)} = c,\ T_n^{(2)} = d]\\ &= \frac{1}{4^n}\left\{\binom{n}{\frac{n+c+d}{2}}\binom{n}{\frac{n+c-d}{2}} - \binom{n}{\frac{n+c+d}{2}+1}\binom{n}{\frac{n+c-d}{2}-1}\right.\\ &\left.\quad - \binom{n}{\frac{n+c+d}{2}+1}\binom{n}{\frac{n+c-d}{2}+1} + \binom{n}{\frac{n+c+d}{2}+2}\binom{n}{\frac{n+c-d}{2}}\right\}\end{aligned}$$

$$= \frac{1}{4^n} \left\{ \binom{n}{\frac{n+c-d}{2}} \binom{n+2}{\frac{n-c-d}{2}} - \binom{n+2}{\frac{n+c-d}{2}+1} \binom{n}{\frac{n-c-d}{2}-1} \right\}. \tag{8.33}$$

They note also that for $c = d = 0$, replacing n by $2n$, we obtain

$$\Pr[T_i^{(1)} \geq 0,\ T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)} = 0] = \frac{1}{4^{2n}}\, C_n C_{n+1}. \tag{8.34}$$

By Proposition 8.1.1, these are equivalent to

$$\begin{aligned} &\Pr[S_i^{(1)} \geq S_i^{(2)} \geq -S_i^{(1)},\ i = 1, \ldots, n,\ S_n^{(1)} = c + d,\ S_n^{(2)} = c - d] \\ &= \frac{1}{4^n} \left\{ \binom{n}{\frac{n+c-d}{2}} \binom{n+2}{\frac{n-c-d}{2}} - \binom{n+2}{\frac{n+c-d}{2}+1} \binom{n}{\frac{n-c-d}{2}-1} \right\} \end{aligned} \tag{8.35}$$

and

$$\Pr[S_i^{(1)} \geq S_i^{(2)} \geq -S_i^{(1)},\ i = 1, \ldots, 2n,\ S_{2n}^{(1)} = S_{2n}^{(2)} = 0] = \frac{1}{4^{2n}}\, C_n C_{n+1}. \tag{8.36}$$

Now putting $d = c$ into (8.35), replacing n by $2n$ and summing up for c, we obtain for the probability that a PRW path remains in the first quadrant $x \geq 0,\ y \geq 0$ and terminates on the diagonal $x = y$ after $2n$ steps:

$$\begin{aligned} &\Pr[T_i^{(1)} \geq 0,\ T_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)}] \\ &= \Pr[S_i^{(1)} \geq S_i^{(2)} \geq -S_i^{(1)},\ i = 1, \ldots, 2n,\ S_{2n}^{(2)} = 0] = \frac{C_n}{2^{2n}}. \end{aligned} \tag{8.37}$$

From the independence of $S^{(1)}$ and $S^{(2)}$, Proposition 8.1.1 and well-known properties of Catalan numbers, we also have

$$\begin{aligned} &\Pr[S_i^{(1)} \geq -1,\ S_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ S_{2n}^{(1)} = S_{2n}^{(2)} = 0] \\ &= \Pr[T_i^{(1)} \geq T_i^{(2)} \geq -T_i^{(1)} - 1,\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)} = 0] \\ &= \frac{1}{4^{2n}}\, C_n C_{n+1} \end{aligned} \tag{8.38}$$

and

$$\begin{aligned} &\Pr[S_i^{(2)} \geq 0,\ i = 1, \ldots, 2n,\ S_{2n}^{(2)} = 0] \\ &= \Pr[T_i^{(1)} \geq T_i^{(2)},\ i = 1, \ldots, 2n,\ T_{2n}^{(1)} = T_{2n}^{(2)}] = \frac{C_n}{2^{2n}}. \end{aligned} \tag{8.39}$$

It would be interesting to give a direct bijective proof for the equivalence of (8.36) and (8.38), and also for that of (8.37) and (8.39).

Acknowledgements. This research was supported by the Hungarian National Foundation for Scientific Research, Grant Nos. T016384 and T019346.

References

1. Barcucci, E. and Verri, M. C. (1992). Some more properties of Catalan numbers, *Discrete Mathematics*, **102**, 229–237.

2. Breckenridge, W., Gastineau-Hills, H., Nelson, A., Bos, P., Calvert, G. and Wehrhahn, K. (1991). Lattice paths and Catalan numbers, *Bulletin of the Institute of Combinatorics and Applications*, **1**, 41–55.

3. Csáki, E., Mohanty, S. G. and Saran, J. (1990). On random walks in a plane, *Ars Combinatoria*, **29**, 309–318.

4. Csáki, E. and Vincze, I. (1961). On some problems connected with the Galton test, *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, **6**, 97–109.

5. DeTemple, D. W., Jones, C. H. and Robertson, J. M. (1988). A correction for a lattice path counting formula, *Ars Combinatoria*, **25**, 167–170.

6. DeTemple, D. W. and Robertson, J. M. (1984). Equally likely fixed length paths in graphs, *Ars Combinatoria*, **17**, 243–254.

7. Di Crescenzo, A., Giorno, V. and Nobile, A. G. (1992). On some first crossing-time probabilities for a two-dimensional random walk with correlated components, *Advances in Applied Probability*, **24**, 441–454.

8. Dvoretzky, A. and Erdős, P. (1950). Some problems on random walk in space, *Proceedings of the Second Berkeley Symposium*, pp. 353–368, Berkeley: University of California Press.

9. Erdős, P. and Taylor, S. J. (1960). Some problems concerning the structure of random walk paths, *Acta Mathematica, Academy of Sciences of Hungary*, **11**, 137–162.

10. Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume I*, Third edition, New York: John Wiley & Sons.

11. Gould, H. W. (1972). *Combinatorial Identities*, Morgantown, West Virgina: Morgantown Printing and Binding Co.

12. Gould, H. W. (1977). Generalization of a formula of Touchard for Catalan numbers, *Journal of Combinatorial Theory, Series A*, **23**, 351–353.

13. Gupta, A. and Sen, K. (1977). A two-dimensional random walk in the presence of a multiple-function barrier, *Metron*, **35**, 189–203.

14. Gupta, A. and Sen, K. (1979). Returns and crossings in a two-dimensional random walk with unspecified terminus, *Metron*, **37**, 137–148.

15. Guy, R. K., Krattenthaler, C. and Sagan, B. E. (1992). Lattice paths, reflections, & dimension-changing bijections, *Ars Combinatoria*, **34**, 3–15.

16. Karlin, S. (1988). Coincident probabilities and applications in combinatorics, *A Celebration of Applied Probability, Journal of Applied Probability, Special Volume*, **25A**, 185–200.

17. Karlin, S. and McGregor, J. L. (1959). Coincidence Probabilities, *Pacific Journal of Mathematics*, **9**, 1141–1164.

18. Kreweras, G. (1992). About Catalan-like lattice paths, *Bulletin of the Institute of Combinatorics and Applications*, **4**, 63–64.

19. McCrea, W. H. and Whipple, F. J. W. (1940). Random paths in two and three dimension, *Proceedings of the Royal Society Edinburgh*, **60**, 281–298.

20. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

21. Pólya, G. (1921). Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Strassennetz, *Mathematische Annalen*, **84**, 149–160.

22. Raifaizen, C. H. (1972). Pairs of one dimensional random walk paths, *Annals of Mathematical Statistics*, **43**, 2095–2098.

23. Révész, P. (1990). *Random Walk in Random and Non-random Environments*, Singapore: World Scientific.

24. Sands, B. (1990). Problem 1517*, *Crux Mathematicorum*, **16**, 44.

25. Saran, J. and Rani, S. (1994). Some joint distributions concerning random walk in a plane, *Ars Combinatoria*, **38**, 33–45.

26. Shapiro, L. W. (1976a). A Catalan triangle, *Discrete Mathematics*, **14**, 83–90.

27. Shapiro, L. W. (1976b). A short proof of an identity of Touchard's concerning Catalan numbers, *Journal of Combinatorial Theory, Series A*, **20**, 375–376.

28. Spitzer, F. (1964). *Principles of Random Walk*, Princeton: Van Nostrand.

29. Touchard, J. (1928). Sur certaines équations fonctionnelles, *Proceedings of the International Congress on Matematics, Toronto*, 465–472.

9

Random Walks on $SL(2, F_2)$ and Jacobi Symbols of Quadratic Residues

Toshihiro Watanabe

Gifu University, Gifu, Japan

Abstract: The Euclidean algorithm with respect to the modulus 4 is given as random walks on the group $SL(2, F_2)$. The quadratic reciprocity law and a simple part of Zolotareff's Theorem are proved in terms of values on each step in the walks.

Keywords and phrases: Euclidean algorithm with respect to the modulus 4, quadratic reciprocity law

9.1 Introduction

In the theory of quadratic residues, the number of proofs already exceeds fifty since Gauss gave seven distinct proofs for the famous quadratic reciprocity law of the Legendre symbol (m/M)

$$\left(\frac{m}{M}\right)\left(\frac{M}{m}\right) = (-1)^{\frac{1}{4}(M-1)(m-1)} \quad \text{for odd } M \text{ and } m > 0, \tag{9.1}$$

[*cf.* Bachmann (1921), Hasse (1980), Frobenius (1914), Takagi (1903), Rousseau (1994) and Zolotareff (1872)]. The reciprocity law also holds for the Jacobi symbol, which is a generalization of the Legendre symbol. In their papers, Zolotareff (1872), Lerch (1896) and Riesz (1953) gave a relation between the Jacobi symbols (m/M) and the character $\chi(\sigma_{M,m})$ of the permutation groups defined by

$$\sigma_{M,m}(k) \equiv km \pmod{M}.$$

Zolotareff's Theorem. *For coprime integers M and m,*

$$\chi(\sigma_{M,m}) = \begin{cases} \left(\frac{m}{M}\right), & \text{if } M \text{ is odd,} \\ (-1)^{\frac{1}{2}(m-1)(m'-1)}, & \text{if } M = 2m'. \end{cases} \tag{9.2}$$

This note gives a proof of the quadratic reciprocity law and the exponent $\frac{1}{2}(m-1)(m'-1)$ in Zolotareff's Theorem, through a unified representation of characters $\chi(\sigma_{M,m})$, $\chi(\sigma_{M-m,m})$ and $\chi(\sigma_{m,M-m})$, which may be called a correlation of random walks on the group $SL(2, F_2)$.

In Section 9.2, we take the Euclidean algorithm for coprime integers M and m as a random walk on the Cayley digraph [*cf.* Coxeter and Moser (1972)] of the group $SL(2, F_2)$, shortly called a *random walk on the group* $SL(2, F_2)$. Proposition 9.2.1 shows that each of the integers M and m with respect to the modulus 4 becomes a sum of values on the steps belonging to the corresponding coset in the group $SL(2, F_2)$.

In Section 9.3, the exponents of characters $\chi(\sigma_{M,m})$, $\chi(\sigma_{M-m,m})$ and $\chi(\sigma_{m,M-m})$ are given a unified representation as a sum of products of two values [*cf.* Theorem 9.3.1]. In the case of even M, the exponent of character $\chi(\sigma_{M,m})$ becomes a product of two sums of values on cosets in the group $SL(2, F_2)$ [*cf.* Proposition 9.3.1], which yields the exponent $\frac{1}{2}(m-1)(m'-1)$ in Zolotareff's Theorem [*cf.* Corollary 9.3.1]. A kind of expectation on the random walks also becomes a similar product [*cf.* Proposition 9.3.1], which yields the exponent $\frac{1}{4}(M-1)(m-1)$ in the quadratic reciprocity law (9.1).

9.2 Preliminaries

Let M and m be any fixed coprime integers satisfying $M > m > 0$. Then, the Euclidean algorithm for M and m gives

$$\begin{pmatrix} M \\ m \end{pmatrix} = \begin{pmatrix} \alpha_n & 1 \\ 1 & 0 \end{pmatrix} \cdots \begin{pmatrix} \alpha_0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \alpha_i : \text{ integer } > 0, \; i = 0, \dots, n. \tag{9.3}$$

In this section, we shall get, from a calculation with respect to the modulus 4 of (9.3), a random walk on the group $SL(2, F_2)$ and a value on each step in this walk. These will be used later in Section 9.3.

First, let us introduce the notations and terminology. Set in (9.3), the following:

$$\alpha_k \equiv a_k + 2b_k \pmod 4, \qquad a_k, b_k \in \{0, 1\}; \tag{9.4}$$

$$J_{a_k} = \begin{pmatrix} a_k & 1 \\ 1 & 0 \end{pmatrix}. \tag{9.5}$$

Let

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Then, the matrices J_0 and J_1 satisfy the relation

$$J_0^2 \equiv J_1^3 \equiv (J_0J_1)^2 \equiv I \pmod 2,$$

and generate the group $SL(2,F_2)$, where F_2 is a field with two elements.

In the 2-dimensional Euclidean space over the field F_2, the inner product is denoted by the symbol $\langle , \rangle$, and the vectors e_i $(i = 0,\ 1)$ are the unit vectors

$$e_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad e_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The symbol $\equiv$ is a congruence with respect to the modulus 2 unless the modulus is given. Let the group H be an isotropy subgroup $\{I, J_0J_1^2\}$ of the group $SL(2,F_2)$:

$$H = \{g \in SL(2,F_2) : \quad ge_0 = e_0\}.$$

From (9.3), we use the following for $0 \le j \le n$:

$$\begin{aligned}
g_j &= J_{a_j} \dots J_{a_0}; \text{ an element in the group } SL(2,F_2). \\
\omega_j &= \begin{pmatrix} a_j \\ b_j \end{pmatrix} \dots \begin{pmatrix} a_0 \\ b_0 \end{pmatrix}; \text{ a biword, i.e., a sequence of two letters.} \\
\lambda(\omega_j) &\equiv \left\langle \begin{pmatrix} a_j \\ b_j \end{pmatrix}, g_je_0 \right\rangle + a_j; \text{ a function from the above biwords} \\
& \text{into the field } F_2. \\
\Lambda_j(HJ_1^k) &\equiv \sum_{g_i \in HJ_1^k}^{0 \le i \le j} \lambda(\omega_i),\ k = 1, 2; \text{ a sum on the coset } HJ_1^k \\
& \text{in the field } F_2. \\
\Lambda_j &= \begin{pmatrix} \Lambda_j(HJ_1) \\ \Lambda_j(HJ_1^2) \end{pmatrix}; \text{ a vector in the 2-dimensional Euclidean} \\
& \text{space over the field } F_2.
\end{aligned}$$

We note that

$$\lambda(\omega_k) \equiv 0 \quad \text{if} \quad g_k \in H. \tag{9.6}$$

For coprime integers N and n, the permutation $\sigma_{N,n}$ on a set $\{1, \dots, N\}$ is defined by

$$\sigma_{N,n}(k) \equiv kn \pmod N.$$

We also use the usual notations and terminology in number theory [see, for example, Hasse (1980)].

The Euclidean algorithm (9.3) with respect to the modulus 2 gives a random walk on the group $SL(2, F_2)$, and the algorithm with respect to the modulus 4 is determined by each sum of values $\lambda(\omega_k)$ on cosets HJ_1^i, $i = 1, 2$, in the group $SL(2, F_2)$ as follows:

Proposition 9.2.1 *In (9.3), set*

$$\begin{aligned} M &\equiv c_0 + 2d_0 \pmod 4, \qquad c_0, d_0 \in \{0,1\}; \\ m &\equiv c_1 + 2d_1 \pmod 4, \qquad c_1, d_1 \in \{0,1\}. \end{aligned}$$

Then

$$\begin{pmatrix} c_0 \\ c_1 \end{pmatrix} \equiv g_n e_0; \tag{9.7}$$

$$\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} \equiv g_n J_1^2 \Lambda_n. \tag{9.8}$$

PROOF. Eq. (9.7) clearly holds. Let us now prove (9.8). In (9.3), set

$$\begin{pmatrix} M \\ m \end{pmatrix} = \begin{pmatrix} \alpha_n & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} m \\ m' \end{pmatrix}; \tag{9.9}$$

$$m' \equiv c_2 + 2d_2 \pmod 4, \qquad c_2, d_2 \in \{0,1\}. \tag{9.10}$$

Thus, we have

$$\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} \equiv J_{a_n} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + e_0(a_n c_1 c_2 + b_n c_1). \tag{9.11}$$

Since the integers m and m' are coprime, it follows that

$$c_1 c_2 + c_1 + c_2 + 1 \equiv 0.$$

Hence,

$$\begin{aligned} a_n c_1 c_2 + b_n c_1 &\equiv (a_n + b_n)c_1 + a_n c_2 + a_n \\ &\equiv \left\langle \begin{pmatrix} a_n \\ b_n \end{pmatrix}, J_{a_n} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\rangle + a_n. \end{aligned}$$

Eq. (9.11), therefore, becomes

$$\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} \equiv J_{a_n} \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} + e_0 \lambda(\omega_n). \tag{9.12}$$

Using (9.12) successively, we get

$$\begin{aligned}\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} &\equiv \sum_{k=0}^{n} J_{a_n} \dots J_{a_{k+1}} e_0 \lambda(\omega_k) \\ &\equiv g_n \sum_{k=0}^{n} g_k^{-1} e_0 \lambda(\omega_k). \end{aligned} \tag{9.13}$$

Since

$$g^{-1} e_0 \equiv J_1^{-i} e_0 \text{ for any } g \in H J_1^i, \quad i = 0, 1, 2,$$

we obtain (9.8) from (9.13) ∎

9.3 A Calculation of the Character $\chi(\sigma_{M,m})$ and Its Relation

In this section, we give two relations for the numbers $\lambda(\omega_k)$ over the field F_2, and as a result we have the simpler part in Zolotareff's Theorem and the quadratic reciprocity law (9.1) of quadratic residues.

The characters $\chi(\sigma)$ of permutations σ are given by the number $I(\sigma)$ of inversions of the permutation σ as follows:

$$\chi(\sigma) = (-1)^{I(\sigma)}$$

[see Berge (1971)].

For the coprime integers M and m in (9.3), set

$$i_0(\omega_n) \equiv I(\sigma_{m,M-m}); \tag{9.14}$$
$$i_1(\omega_n) \equiv I(\sigma_{M-m,m}); \tag{9.15}$$
$$i_2(\omega_n) \equiv I(\sigma_{M,m}). \tag{9.16}$$

Then, we have the following theorem.

Theorem 9.3.1

$$\begin{aligned} & i_j(\omega_n) \\ \equiv & \sum_{n \geq k > l \geq 0} \left\{ \begin{matrix} g_k \\ g_l \end{matrix} \right\} \left(\left\langle \begin{pmatrix} a_k \\ b_k \end{pmatrix}, g_k g_n^{-1} J_1^j e_0 \right\rangle + a_k \right) \left(\left\langle \begin{pmatrix} a_l \\ b_l \end{pmatrix}, g_l e_0 \right\rangle + a_l \right), \\ & j = 0, 1, 2, \end{aligned} \tag{9.17}$$

where

$$\left\{ \begin{matrix} g_k \\ g_l \end{matrix} \right\} = \begin{cases} 0 & \textit{if } g_k g_l^{-1} \in H, \\ 1 & \textit{otherwise.} \end{cases} \tag{9.18}$$

We note the following:

(i)

$$\left\{ \begin{array}{c} g_k \\ g_l \end{array} \right\} \equiv \langle e_1, g_k g_l^{-1} e_0 \rangle$$
$$\equiv [a_{k-1}, \ldots, a_{l+1}],$$

where the symbol [ ] is the Gaussian Klammer;

(ii) We have for d_1 in Proposition 9.2.1

$$\begin{aligned} \sum_{n \geq k \geq 0} \left\{ \begin{array}{c} g_n \\ g_k \end{array} \right\} \lambda(\omega_k) &\equiv \sum_{n \geq k \geq 0} \langle e_1, g_n g_k^{-1} e_0 \rangle \lambda(\omega_k) \text{ (by (i) above)} \\ &\equiv d_1 \text{ (by (9.13))} \\ &\equiv \langle e_1, g_n J_1^2 \Lambda_n \rangle \text{ (by Proposition 9.2.1)} \\ &\equiv \langle g_n^{-1} e_0, J_0 J_1^2 \Lambda_n \rangle, \end{aligned} \tag{9.19}$$

where, in the last equation, we use the relation

$$J_0 \, {}^t g \; J_0 = g^{-1} \text{ for any } g \in SL(2, F_2).$$

Since we prove this theorem by a substitution of words or an automaton and needs preliminaries, we will give the proof in another paper.

The following proposition is the main result of this paper.

Proposition 9.3.1 *The following equations hold:*

$$\sum_{n \geq k > j \geq 0} \left\{ \begin{array}{c} g_k \\ g_j \end{array} \right\} \lambda(\omega_k) \lambda(\omega_j) \equiv \Lambda_n(HJ_1)\Lambda_n(HJ_1^2); \tag{9.20}$$

$$\sum_{n \geq k > j \geq 0} a_k \left\{ \begin{array}{c} g_k \\ g_j \end{array} \right\} \lambda(\omega_j) \equiv \langle ((J_1^2 g_n)^{-1} + I) e_0, J_0 J_1^2 \Lambda_n \rangle. \tag{9.21}$$

PROOF. Since the statements that $\left\{ \begin{array}{c} g_k \\ g_j \end{array} \right\} = 1$, and g_k and g_j are not in the same coset HJ_1^i are equivalent, we get (9.20) by (9.6).

For a proof of (9.21), let us first prove the following:

$$\begin{aligned} \langle ((J_1^2 g_k)^{-1} + I) e_0, J_0 J_1^2 \Lambda_k \rangle &\equiv \langle ((J_1^2 g_k)^{-1} + I) e_0, J_0 J_1^2 \Lambda_{k+1} \rangle \\ &\qquad \text{for any } k \geq 0. \end{aligned} \tag{9.22}$$

For any j, we have

$$HJ_1^j = \{J_1^j, J_0 J_1^{j-1}\};$$

$$g_{k+1} \in HJ_1^{j+1}(HJ_1^{j-1}) \text{ if } g_k = J_1^j(J_0J_1^{j-1}).$$

Thus, we have

$$\begin{cases} \Lambda_{k+1}(HJ_1^j) \equiv \Lambda_k(HJ_1^j) & \text{if} \quad g_k \in HJ_1^j; \\ \Lambda_{k+1}(HJ_1^{j-1}) \equiv \Lambda_k(HJ_1^{j-1}) & \text{if} \quad g_k = J_1^j; \\ \Lambda_{k+1}(HJ_1^{j+1}) \equiv \Lambda_k(HJ_1^{j+1}) & \text{if} \quad g_k = J_0J_1^{j-1}. \end{cases} \tag{9.23}$$

Using

$$I + J_1 + J_1^2 \equiv 0, \tag{9.24}$$

we have the following:

(*i*) $$J_1^2J_0((J_1^2J_1^j)^{-1} + I)e_0 \equiv \begin{cases} 0 & \text{if} \quad j = 1, \\ J_1^{2-j}e_0 & \text{if} \quad j \neq 1; \end{cases}$$

(*ii*) $$J_1^2J_0((J_1^2J_0J_1^{j-1})^{-1} + I)e_0 \equiv \begin{cases} 0 & \text{if} \quad j = 2, \\ J_1^{-j}e_0 & \text{if} \quad j \neq 2; \end{cases}$$

Using (i), (ii) and (9.23), we obtain for any $k \geq 0$,

$$\langle J_1^2J_0((J_1^2g_k)^{-1} + I)e_0, \Lambda_k\rangle \equiv \langle J_1^2J_0((J_1^2g_k)^{-1} + I)e_0, \Lambda_{k+1}\rangle; \tag{9.25}$$

that is, Eq. (9.22).

Since

$$\begin{aligned} ((J_1^2g_{k+1})^{-1} + I)e_0 + ((J_1^2g_k)^{-1} + I)e_0 &\equiv g_{k+1}^{-1}(I + J_{a_{k+1}})J_1e_0 \\ &\equiv a_{k+1}g_{k+1}^{-1}e_0 \text{ (by (9.24))}, \end{aligned}$$

we have from (9.22)

$$\begin{aligned} &a_{k+1}\langle g_{k+1}^{-1}e_0, J_0J_1^2\Lambda_{k+1}\rangle \\ &\quad\equiv \langle((J_1^2g_{k+1})^{-1} + I)e_0, J_0J_1^2\Lambda_{k+1}\rangle + \langle((J_1^2g_k)^{-1} + I)e_0, J_0J_1^2\Lambda_k\rangle. \end{aligned} \tag{9.26}$$

Since, by (9.19), the left hand side of (9.21) is

$$\sum_{n \geq k \geq 0} a_k\langle g_k^{-1}e_0, J_0J_1^2\Lambda_k\rangle,$$

by using (9.26) successively, we obtain (9.21). ∎

In the case of even M, Proposition 9.2.1 gives

$$\langle e_0, g_ne_0\rangle \equiv 0, \qquad \text{that is, } g_n = J_0 \text{ or } J_1^2.$$

So,

$$d_0 \equiv \Lambda_n(HJ_1) + \Lambda_n(HJ_1^2) \quad \text{for } g_n \in \{J_0, J_1^2\}$$

and

$$d_1 \equiv \begin{cases} \Lambda_n(HJ_1^2) & \text{if} \quad g_n = J_0, \\ \Lambda_n(HJ_1) & \text{if} \quad g_n = J_1^2. \end{cases}$$

Thus, we have

$$(d_0 + 1)d_1 \equiv \Lambda_n(HJ_1)\Lambda_n(HJ_1^2) \quad \text{for } g_n \in \{J_0, J_1^2\}. \tag{9.27}$$

Since

$$\left\langle \begin{pmatrix} a_k \\ b_k \end{pmatrix}, g_k g_n^{-1} J_1^2 e_0 \right\rangle + a_k \equiv \lambda(\omega_k) \quad \text{for } g_n \in \{J_0, J_1^2\},$$

from Theorem 9.3.1, Proposition 9.3.1 and Eq. (9.27), we have

$$\begin{aligned} I(\sigma_{M,m}) &\equiv i_2(\omega_n) \\ &\equiv \Lambda_n(HJ_1)\Lambda_n(HJ_1^2) \\ &\equiv (d_0+1)d_1 \\ &\equiv \frac{1}{2}\,(m-1)(m'-1). \end{aligned}$$

where $M = 2m'$. Thus, we obtain the simple part of Zolotareff's Theorem.

Corollary 9.3.1

$$\chi(\sigma_{M,m}) = (-1)^{\frac{1}{2}\,(m-1)(m'-1)} \qquad \textit{for } M = 2m'.$$

By (9.24), Eq. (9.21) in Proposition 9.3.1 gives the following.

Corollary 9.3.2

$$\sum_{j=0}^{2} i_j(\omega_n) \equiv \langle ((J_1^2 g_n)^{-1} + I)e_0, J_0 J_1^2 \Lambda_n \rangle. \tag{9.28}$$

In the case of odd M and m, Proposition 9.2.1 gives

$$g_n e_0 = e_0 + e_1, \qquad \text{that is,} \quad g_n = J_1 \text{ or } J_0 J_1.$$

So,

$$\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} \equiv \Lambda_n \qquad \text{for } g_n = J_1,$$

and

$$\begin{pmatrix} d_0 \\ d_1 \end{pmatrix} \equiv J_0 \Lambda_n \qquad \text{for } g_n = J_0 J_1.$$

Thus, for g_n in $\{J_1, J_0J_1\}$, we have

$$d_0d_1 \equiv \Lambda_n(HJ_1)\Lambda(HJ_1^2). \tag{9.29}$$

Since, for g_n in $\{J_1, J_0J_1\}$,

$$((J_1^2g_n)^{-1} + I)e_0 \equiv 0 \text{ and } \left\langle \begin{pmatrix} a_k \\ b_k \end{pmatrix}, g_kg_n^{-1}J_1e_0 \right\rangle + a_k \equiv \lambda(\omega_k),$$

from Proposition 9.3.1, Corollary 9.3.2 and Eq. (9.29), we have

$$\begin{aligned} i_1(\omega_n) \equiv \Lambda_n(HJ_1)\Lambda_n(HJ_1^2) &\equiv d_0d_1; \\ i_0(\omega_n) + i_2(\omega_n) &\equiv d_0d_1. \end{aligned}$$

Hence,

$$\begin{aligned} I(\sigma_{m,M-m}) + I(\sigma_{M,m}) &\equiv i_0(\omega_n) + i_2(\omega_n) \equiv d_0d_1 \\ &\equiv \frac{1}{4}(M-1)(m-1). \end{aligned} \tag{9.30}$$

Using the known property $(M/m) = (M - m/m)$ of the Jacobi symbol [see Hasse (1980)], we have

$$\left(\frac{M}{m}\right) = (-1)^{i_0(\omega_n)}.$$

Therefore, from Eq. (9.30), we obtain the quadratic reciprocity law.

Corollary 9.3.3 *For odd M and m,*

$$\left(\frac{m}{M}\right)\left(\frac{M}{m}\right) = (-1)^{\frac{1}{4}(M-1)(m-1)}.$$

Remark. When the exponents $i_j(\omega_n)$, $j = 0, 1, 2$, of characters of groups $\sigma_{M,m}$, $\sigma_{M-m,m}$ and $\sigma_{m,M-m}$ are put in a unified form (9.17), and we take it in terms of random walks, we can easily deduce the simple part of Zolotareff's Theorem and the quadratic reciprocity law of quadratic residues. Shanks (1985) made some comments on the quadratic reciprocity law while Kubota (1992) raised some questions in the class field theory. It is important to note that the unified form in (9.17) has these (and possibly some other) relations.

References

1. Bachmann, P. (1921). *Niedere Zahlentheorie I*, Leipzig: B. G. Teubner.

2. Berge, C. (1971). *Principles of Combinatorics*, New York: Academic Press.

3. Coxeter, H. S. M. and Moser, W. O. J. (1972). *Generators and Relations for Discrete Groups*, New York: Springer-Verlag.

4. Frobenius, F. G. (1914). *Über das quadratische Reziprozitätsgesetz*, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, pp. 335–349; reprinted in *F. G. Frobenius Gesanmelte Abhandlungen III*, 1968, pp. 628–642, Berlin: Springer-Verlag.

5. Frobenius, F. G. (1914). *Über das quadratische Reziprozitätsgesetz II*, Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, pp. 484–488; reprinted in *F. G. Frobenius Gesanmelte Abhandlungen III*, 1968, pp. 643–647, Berlin: Springer-Verlag.

6. Hasse, H. (1980). *Number Theory*, English edition, Berlin: Springer-Verlag. (German edition, 1969)

7. Kubota, T. (1987). Geometry of numbers and class field theory, *Japan Journal of Mathematics*, **13**, 235–275.

8. Kubota, T. (1992). A foundation of class field theory in terms of space figures, *Sugaku*, **41**, 1–12. (in Japanese)

9. Lehmer, D. H. (1976). The character of linear permutations, *Linear and Multilinear Algebra*, **4**, 1–16.

10. Lerch, M. (1896). Sur un théorème arithmétique de Zolotarev, *Ceska Akad., Prague, Bull. Inter. Cl. Math.*, **3**, 34–37.

11. Riesz, M. (1953). Sur le lemme de Zolotareff et sur loi de réciprocité des restes quadratiques, *Mathematica Scandinavica*, **1**, 159–169.

12. Rousseau, G. (1994). On the Jacobi symbol, *Journal of Number Theory*, **48**, 109–111.

13. Shanks, D. (1985). *Solved and Unsolved Problems in Number Theory*, Third edition, New York: Chelsea.

14. Takagi, T. (1903). A simple proof of the law of reciprocity for quadratic residues, *Proc. Phys.-Math. Soc. Japan*, II, **2**, 74–78; reprinted in *The Collected Papers of Teiji Takagi*, pp. 10–12, 1973, Tokyo: Iwanami Shoten.

15. Zolotareff, M. (1872). Nouvelle démonstration de la loi de réciprocité de Legendre, *Nouvelle Annales de Math.*, **11**, 354–363.

10

Rank Order Statistics Related to a Generalized Random Walk

Jagdish Saran and Sarita Rani

University of Delhi, Delhi, India

Abstract: This paper deals with the derivation of the joint and marginal distributions of certain rank order statistics related to the generalized random walk with steps $+1$ and $-\mu$ by using the extended Dwass technique evolved by Mohanty and Handa (1970). These generalize and extend the results of Saran and Rani (1991a,b).

Keywords and phrases: Extended Dwass technique; generalized random walk; rank order statistics - upcrossing of height a, upward crossing of height a, positive reflection at height a; run of upcrossings of height a; run of upward crossings of height a; run of positive reflections at height a $(a > 0)$.

10.1 Introduction

Let $X_1, X_2, \ldots, X_{\mu n}$ and $Y_1, Y_2, \ldots, Y_n$ be two independent random samples of sizes μn and n (where μ is a positive integer) from the same population having continuous distribution function. Let $F_{\mu n}(x)$ and $G_n(x)$ be the corresponding empirical distribution functions of the two samples. Define the rank order indicator of $\{X_1, X_2, \ldots, X_{\mu n}, Y_1, Y_2, \ldots, Y_n\}$ as a vector $(Z_1, Z_2, \ldots, Z_{(\mu+1)n})$ such that

$$Z_j = \begin{cases} +1 & \text{if the } j\text{-th minimum among } \{X_1, \ldots, X_{\mu n}, Y_1, \ldots, Y_n\} \\ & \text{is } X_t \text{ for some } t \in \{1, 2, \ldots, \mu n\} \\ -\mu & \text{if the } j\text{-th minimum among } \{X_1, \ldots, X_{\mu n}, Y_1, \ldots, Y_n\} \\ & \text{is } Y_t \text{ for some } t \in \{1, 2, \ldots, n\}, \end{cases}$$

$j = 1, 2, \ldots, (\mu + 1)n$. Obviously $(Z_1, Z_2, \ldots, Z_{(\mu+1)n})$ is a sequence of μn $(+1)$'s and n $(-\mu)$'s which we call a *sequence of rank order indicators*. Under the assumption, the $\binom{(\mu+1)n}{n}$ possible sequences of rank order indicators are equally likely. Any random variable defined on the rank order indicator $(Z_1, Z_2, \ldots, Z_{(\mu+1)n})$ is called a *rank order statistic.* Defining

$$H_{\mu,n}(u) = n\mu \left[F_{\mu n}(u) - G_n(u)\right], \qquad -\infty < u < \infty,$$

we note that statistics defined through $H_{\mu,n}(u)$ can be treated as rank order statistics.

Dwass (1967) developed a new technique (other than the combinatorial one) based on the simple random walk with independent steps, in order to derive the distributions of some rank order statistics for the case of equal sample sizes (i.e., for $\mu = 1$) which are defined on $H_{1,n}(u)$. Mohanty and Handa (1970) extended the technique of Dwass (1967) to the case when one sample size is an integer multiple of the other and derived the distributions of a few rank order statistics. For this purpose, they considered the generalized random walk $\{S_j : S_j = \sum_{i=1}^{j} W_i,\ S_0 = W_0 = 0\}$ generated by a sequence $\{W_i\}$ of independent random variables with common probability distribution

$$\Pr[W_i = +1] = p,\ \Pr[W_i = -\mu] = q = 1 - p,\ 1 \le i < \infty.$$

Further, Saran and Sen (1979), Kaul (1982, Ch. IV), Pratap (1982, Ch. IV), Sen and Saran (1983), Sen and Kaul (1985) and Saran and Rani (1990, 1991b) have derived the joint and marginal distributions of some rank order statistics related to the generalized random walk $\{S_i\}$ with steps $+1$ and $-\mu$. In this paper, we consider the above mentioned generalized random walk with steps $+1$ and $-\mu$ and derive the joint distributions of the number of upcrossings of height a and their runs, the number of positive reflections at height a and their runs, and the number of upward crossings of height a and their runs ($a > 0$), by employing the extended Dwass technique given by Mohanty and Handa (1970). These generalize and extend the earlier work by Saran and Rani (1991b) in which the above mentioned distributions have been derived for the special case $a = 0$.

10.2 Some Auxiliary Results

The basic results needed in the sequel are quoted from Mohanty and Handa (1970) and Sen and Saran (1983); see also Saran and Rani (1991b).

(i) For any α and β,

$$\sum_{k=0}^{\infty} A_k(\alpha, \beta)\theta^k = x^\alpha, \tag{10.1}$$

where

$$A_k(\alpha,\beta) = \frac{\alpha}{\alpha+k\beta}\binom{\alpha+k\beta}{k}, \quad \theta = (x-1)/x^{\beta}$$
$$\text{and } |\theta| < |(\beta-1)^{\beta-1}/\beta^{\beta}|,$$

the last inequality assuring the convergence of the series.

(ii) The probability generating function (pgf) for the first return to the origin in the generalized random walk with steps $+1$ and $-\mu$ is

$$F(t) = (\mu+1)p^{\mu}qx^{\mu}t^{\mu+1}, \tag{10.2}$$

where

$$p^{\mu}qt^{\mu+1} = (x-1)/x^{\mu+1} \text{ and } |t|^{\mu+1}p^{\mu}q < \mu^{\mu}/(\mu+1)^{\mu+1}.$$

(iii) The probability of never returning to the origin is

$$\delta = 1 - F(1) = 1 - (\mu+1)p^{\mu}qy^{\mu}, \tag{10.3}$$

where y is the value of x when $t = 1$.

(iv) The probability of ever reaching k is

$$G(1,k) = (py)^k, \qquad k = 1, 2, \ldots\ . \tag{10.4}$$

(v) The probability of ever returning to the origin with $S_1 = -\mu$ is given by

$$F^{-}(1) = p^{\mu}qy^{\mu}. \tag{10.5}$$

(vi) The probability of ever returning to the origin with $S_1 = +1$ and having one crossing of the origin at a non-lattice point in a generalized random walk with steps $+1$ and $-\mu$ is given by

$$F_1^{+}(1) = (\mu-1)p^{\mu}qy^{\mu}. \tag{10.6}$$

(vii) The probability of ever returning to the origin with $S_1 = +1$ and without crossing the origin before is given by

$$F_2^{+}(1) = p^{\mu}qy^{\mu}. \tag{10.7}$$

(viii) The probability of a particle starting from the origin with a positive step and returning to the origin with a positive step with the condition that

(a) it crosses the origin only once, and

(b) it is allowed to reach the origin before the crossing and it is not allowed to reach the origin after the crossing except at the end,

is given by

$$p^{\mu}qy^{\mu}(\mu - 1 + p^{\mu}qy^{\mu})/(1 - p^{\mu}qy^{\mu}). \tag{10.8}$$

(ix) The following power series expansion is also useful:

$$\frac{p^k}{1-(\mu+1)p^{\mu}qy^{\mu}} = \sum_{n=\langle k/\mu \rangle}^{\infty} \binom{(\mu+1)n-k}{n} (p^{\mu}q)^n, \; k > 0, \tag{10.9}$$

where $\langle x \rangle$ is the smallest integer greater than or equal to x.

10.3 The Technique

The main theorem of Mohanty and Handa (1970) which plays a vital role for finding the distributions of rank order statistics is presented below; see also Saran and Rani (1991b).

Theorem 10.3.1 *Suppose $V_{\mu,n}$ is a rank order statistic for every n and V_{μ} is the corresponding function defined on the random walk which is completely determined by $W_1, W_2, \ldots, W_T$ and does not depend on $W_{T+1}, W_{T+2}, \ldots$, whenever $T > 0$ (where T is the time for the last return to zero in the random walk).*

Define

$$h(p) = E(V_{\mu}), \qquad p < \mu/(\mu+1). \tag{10.10}$$

Then we have the following power series (in powers of $p^{\mu}q$) expansion:

$$\frac{h(p)}{1-(\mu+1)p^{\mu}qy^{\mu}} = \sum_{n=0}^{\infty} E(V_{\mu,n}) \binom{(\mu+1)n}{n} (p^{\mu}q)^n, \tag{10.11}$$

where y is as in (10.2) and (10.3).

10.4 Definitions of Rank Order Statistics

The following is the list of rank order statistics whose distributions will be derived. In what follows, we shall use the dual notation V_μ, $V_{\mu,n}$ for these rank order statistics as mentioned in Section 10.3.

I. $N^{+*}_{\mu,n}(a)$ = the number of upcrossings of height a
= the number of indices i for which $H_{\mu,n}(Z_i) = a+1$ and $H_{\mu,n}(Z_{i-1}) = a$, $i = 1, 2, \dots$.

II. $\Lambda^{+}_{\mu,n}(a)$ = the number of positive reflections at height a
= the number of indices i for which $H_{\mu,n}(Z_i) = a$, $H_{\mu,n}(Z_{i-1}) = a + \mu$ and $H_{\mu,n}(Z_{i+1}) = a+1$, $i = 1, 2, \dots$.

III. $N^{*}_{\mu,n}(a)$ = the number of upward crossings of height a
= the number of indices i for which $H_{\mu,n}(Z_i) = a$, $H_{\mu,n}(Z_{i-1}) = a - 1$ and $H_{\mu,n}(Z_{i+1}) = a+1$, $i = 1, 2, \dots$.

IV. $R^{+*}_{\mu,n}(a)$ = the number of runs of upcrossings of height a of type I whose number is $N^{+*}_{\mu,n}(a)$
= the number of sequences of (consecutive) upcrossings of height a with indices increasing by $\mu+1$. A sequence of upcrossing indices $i_k, i_{k+1}, \dots, i_c$ will be said to form a run of upcrossings if
(i) $i_j - i_{j-1} = \mu + 1$, $j = k+1, k+2, \dots, c$,
(ii) $i_k > i_{k-1} + \mu + 1$ and
(iii) $i_{c+1} > i_c + \mu + 1$, $c = 1, 2, \dots$.

V. $Rf^{+}_{\mu,n}(a)$ = the number of runs of positive reflections of height a of type II whose number is $\Lambda^{+}_{\mu,n}(a)$
= the definition IV with 'positive reflection' in place of 'upcrossing'.

VI. $R^{*}_{\mu,n}(a)$ = the number of runs of upward crossings of height a of type III whose number is $N^{*}_{\mu,n}(a)$
= the definition IV with 'upward crossing' in place of 'upcrossing'.

10.5 Distributions of $N^{+*}_{\mu,n}(a)$ and $R^{+*}_{\mu,n}(a)$

Theorem 10.5.1

$$
\begin{aligned}
&\binom{(\mu+1)n}{n} \Pr\left[N^{+*}_{\mu,n}(a)=c,\ R^{+*}_{\mu,n}(a)=k\right] \\
&= \binom{c-1}{k-1}\mu^{c-1}\sum_{i=0}^{k-1}\binom{k-1}{i}(-1)^i\Bigg\{\sum_{r=\langle(a+1)/\mu\rangle}^{n-c+1}\binom{(\mu+1)r-a-1}{r} \\
&\qquad \times A_{n-(c-1)-r}(\mu(k-i-1)+a+k-i,\mu+1) \\
&\quad -\sum_{s=\langle(a+2)/\mu\rangle}^{n-c+1}\binom{(\mu+1)s-a-2}{s} \\
&\qquad \times A_{n-(c-1)-s}(\mu(k-i-1)+a+k-i+1,\mu+1) \\
&\quad +\sum_{j=1}^{\mu-1}\sum_{t=\langle(a+1+j-\mu)/\mu\rangle}^{n-c}\binom{(\mu+1)t-a-1-j+\mu}{t} \\
&\qquad \times A_{n-c-t}(\mu(k-i-1)+k-i+a+j+1,\mu+1)\Bigg\}. \qquad (10.12)
\end{aligned}
$$

Proof. To establish (10.12), let $0P_1P_2\ldots P_cD$ (Figure 10.1) be a generalized random walk path with $N^{+*}_{\mu}(a)=c$, $R^{+*}_{\mu}(a)=k$ as stipulated in the theorem, where $P_1, P_2, \ldots, P_c$ are the upcrossing points of height a and D is the point

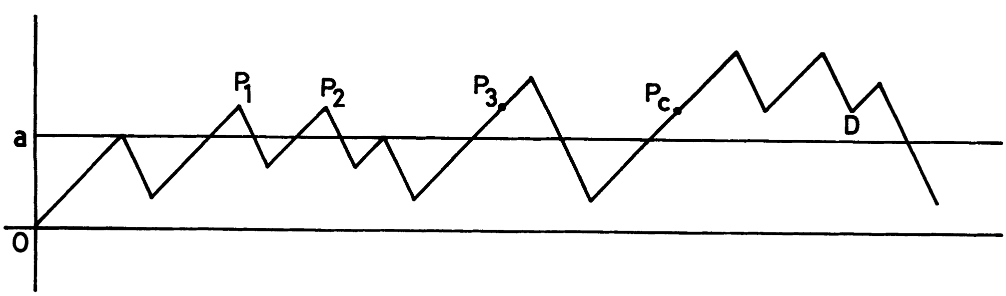


Figure 10.1: A sample path for the event $N^{+*}_{\mu}(a)=c$, $R^{+*}_{\mu}(a)=k$

where the particle reaches height $a+1$ for the last time. If P_c itself is the point of last return to height $a+1$, then the point D will coincide with P_c and the segment P_cD as shown in Figure 10.1 will not exist. The path is thus divided into $c+2$ segments (see Figure 10.1) by the c upcrossings of height a as follows:

(a) One segment in the beginning up to the first upcrossing of height a, i.e., $0P_1$ and it may be of any length and it occurs with probability $(py)^{a+1}$, by (10.4);

(b) $c-k$ segments, each of length $\mu+1$ (like P_1P_2 in Figure 10.1) and each with probability $\mu p^\mu q$;

(c) $k-1$ segments, each of length $>\mu+1$ (like P_2P_3 in Figure 10.1) and each with probability

$$p^\mu qy^\mu - p^\mu q + \frac{p^\mu qy^\mu(\mu-1+p^\mu qy^\mu)}{1-p^\mu qy^\mu} - (\mu-1)p^\mu q,$$

by (10.5) and (10.8);

(d) one segment from the last upcrossing point to D, i.e., P_cD and it may be of any length and will occur with probability

$$\sum_{r=0}^{\infty}[F_2^+(1)]^r = \frac{1}{1-p^\mu qy^\mu} = y;$$

(e) the last segment from D to ∞ is such that the particle crosses height $a+1$ only once and thereafter it does not reach height $a+1$. This segment has the following two contingencies:

  (i) when the last crossing takes place at a lattice point,

  (ii) when the last crossing takes place at a non-lattice point.

In case (i), the last segment from D to ∞ occurs with probability

$$q\{1-(py)^\mu\}$$

and in case (ii), it occurs with probability

$$\sum_{j=1}^{\mu-1}(py)^j q\{1-(py)^{\mu-j}\} = \sum_{j=1}^{\mu-1} p^j qy^j - (\mu-1)p^\mu qy^\mu,$$

by using an argument similar to the one used by Sen and Saran (1983, Lemma 3). Thus, the probability of the last segment from D to ∞ equals

$$q\{1-(py)^\mu\} + \sum_{j=1}^{\mu-1} p^j qy^j - (\mu-1)p^\mu qy^\mu.$$

Now $(c-k)$ segments each of length $\mu+1$ are to be combined with $k-1$ segments each of length $>\mu+1$ so as to form k runs of total c upcrossings which is possible in $\binom{c-k+k-1}{c-k}=\binom{c-1}{k-1}$ ways. Thus,

$$\begin{aligned}
&\Pr[N_\mu^{+*}(a)=c, R_\mu^{+*}(a)=k] \\
&= \binom{c-1}{k-1}\mu^{c-1}(p^\mu q)^{c-1}(py)^{a+1}(y^{\mu+1}-1)^{k-1} \\
&\quad\times\left\{1-py-(\mu-1)p^\mu q y^{\mu+1}+\sum_{j=1}^{\mu-1}p^j q y^{j+1}\right\}. \qquad (10.13)
\end{aligned}$$

Hence, identifying $h(p)$ as $\Pr\left[N_\mu^{+*}(a)=c,\ R_\mu^{+*}(a)=k\right]$, we have

$$\begin{aligned}
h(p)/\delta &= \binom{c-1}{k-1}\mu^{c-1}(p^\mu q)^c\sum_{i=0}^{k-1}\binom{k-1}{i}(-1)^i\, p^{a+1}y^{(\mu+1)(k-i-1)+a+1} \\
&\quad\times\left\{1-py-(\mu-1)p^\mu q y^{\mu+1}+\sum_{j=1}^{\mu-1}p^j q y^{j+1}\right\}\Big/\delta.
\end{aligned}$$

On comparing the coefficient of $(p^\mu q)^n$ on both sides, and using Theorem 10.3.1, we get the desired result in (10.12). ■

Deductions

(A) Putting $\mu=1$ in (10.13), we get

$$h(p)/(1-2p)=\binom{c-1}{k-1}(pq)^{c-2k-a-1}p^{3k+2a}(2-p)^{k-1}, \qquad (10.14)$$

which is in agreement with Saran and Rani (1991a).

(B) Summing (10.13) over k, we get

$$\begin{aligned}
h_1(p)/\delta &= \Pr\left[N_\mu^{+*}(a)=c\right]/\delta \\
&= (\mu p^\mu q)^{c-1}y^{a+1+(\mu+1)(c-1)}p^{a+1} \\
&\quad\times\left\{1-py-(\mu-1)p^\mu q y^{\mu+1}+\sum_{j=1}^{\mu-1}p^j q y^{j+1}\right\}\Big/\delta
\end{aligned}$$

(10.15)

in which the coefficient of $(p^\mu q)^n$ gives

$$
\begin{aligned}
&\binom{(\mu+1)n}{n} \Pr[N^{+*}_{\mu,n}(a) = c] \\
= \;& \mu^{c-1}\Bigg\{ \sum_{g=\langle(a+1)/\mu\rangle}^{n-c+1} \binom{(\mu+1)g-a-1}{g} \\
& \quad \times A_{n-g-c+1}(a+1+(\mu+1)(c-1), \mu+1) \\
& - \sum_{h=\langle(a+2)/\mu\rangle}^{n-c+1} \binom{(\mu+1)h-a-2}{h} \\
& \quad \times A_{n-h-c+1}(a+2+(\mu+1)(c-1), \mu+1) \\
& - (\mu-1) \sum_{g=\langle(a+1)/\mu\rangle}^{n-c} \binom{(\mu+1)g-a-1}{g} \\
& \quad \times A_{n-g-c}(a+1+(\mu+1)c, \mu+1) \\
& + \sum_{j=1}^{\mu-1} \sum_{m=\langle(a+j-\mu+1)/\mu\rangle}^{n-c} \binom{(\mu+1)m-a-j+\mu-1}{m} \\
& \quad \times A_{n-m-c}(a+j+2+(\mu+1)(c-1), \mu+1) \Bigg\},
\end{aligned}
$$

which is equivalent to the result of Kaul (1982).

(C) Summing (10.13) over c, we have

$$
\begin{aligned}
h_2(p)/\delta &= \Pr\left[R^{+*}_{\mu}(a) = k\right]/\delta \\
&= (py)^{a+1}(y^{\mu+1}-1)^{k-1} \sum_{c=k}^{\infty} \binom{c-1}{k-1} (\mu p^{\mu} q)^{c-1} \\
&\quad \times \left\{ 1 - py - (\mu-1)p^{\mu} q y^{\mu+1} + \sum_{j=1}^{\mu-1} p^j q y^{j+1} \right\} \Big/ \delta \\
&= (py)^{a+1}(y^{\mu+1}-1)^{k-1} \sum_{r=0}^{\infty} \binom{k+r-1}{r} (\mu p^{\mu} q)^{k+r-1} \\
&\quad \times \left\{ 1 - py - (\mu-1)p^{\mu} q y^{\mu+1} + \sum_{j=1}^{\mu-1} p^j q y^{j+1} \right\} \Big/ \delta \\
&= (py)^{a+1}(y^{\mu+1}-1)^{k-1}(1-\mu p^{\mu} q)^{-k} (\mu p^{\mu} q)^{k-1} \\
&\quad \times \left\{ 1 - py - (\mu-1)p^{\mu} q y^{\mu+1} + \sum_{j=1}^{\mu-1} p^j q y^{j+1} \right\} \Big/ \delta
\end{aligned}
\tag{10.16}
$$

in which the coefficient of $(p^\mu q)^n$ gives

$$
\begin{aligned}
&\binom{(\mu+1)n}{n} \Pr\left[R^{+*}_{\mu,n}(a)=k\right] \\
&= \sum_{i=0}^{k-1}\sum_{r=0}^{\infty}\binom{k-1}{i}\binom{k+r-1}{r}(-1)^i\mu^{k+r-1} \\
&\quad\times\Bigg\{\sum_{s=\langle(a+1)/\mu\rangle}^{n-k-r+1}\binom{(\mu+1)s-a-1}{s} \\
&\qquad\times A_{n-k-r+1-s}(\mu(k-i-1)+a+k-i,\mu+1) \\
&\quad-\sum_{t=\langle(a+2)/\mu\rangle}^{n-k-r+1}\binom{(\mu+1)t-a-2}{t} \\
&\qquad\times A_{n-k-r+1-t}(\mu(k-i-1)+a+k-i+1,\mu+1) \\
&\quad-(\mu-1)\sum_{s=\langle(a+1)/\mu\rangle}^{n-k-r}\binom{(\mu+1)s-a-1}{s} \\
&\qquad\times A_{n-k-r-s}(\mu(k-i)+a+k-i+1,\mu+1) \\
&\quad+\sum_{j=1}^{\mu-1}\sum_{g=\langle(a+1+j-\mu)/\mu\rangle}^{n-k-r}\binom{(\mu+1)g-a-1-j+\mu}{g} \\
&\qquad\times A_{n-k-r-g}(\mu(k-i-1)+k-i+a+j+1,\mu+1)\Bigg\}.
\end{aligned}
$$

10.6 Distributions of $\Lambda^+_{\mu,n}(a)$ and $Rf^+_{\mu,n}(a)$

Theorem 10.6.1

$$
\begin{aligned}
&\binom{(\mu+1)n}{n}\Pr\left[\Lambda^+_{\mu,n}(a)=r,\ Rf^+_{\mu,n}(a)=k\right] \\
&=\sum_{i=0}^{k-1}\sum_{s=0}^{\infty}\sum_{g=0}^{s}\binom{r-1}{k-1}\binom{k-1}{i}\binom{s+i+1}{s}\binom{s}{g}(-1)^{k-1-i+s-g}\mu^g \\
&\quad\times\Bigg\{\sum_{b=\langle a/\mu\rangle}^{\infty}\binom{(\mu+1)b-a}{b} \\
&\qquad\times A_{n-b-r-s-1}(\mu(s+i+2)+a+g+1,\mu+1) \\
&\quad-\mu\sum_{c=\langle(a+1)/\mu\rangle}^{\infty}\binom{(\mu+1)c-a-1}{c}
\end{aligned}
$$

$$
\begin{aligned}
&\times A_{n-c-r-s-1}(\mu(s+i+2)+a+2+g,\mu+1)\\
&+\sum_{j=1}^{\mu-1}\sum_{d=\langle(a+j-\mu)/\mu\rangle}^{\infty}\binom{(\mu+1)d-a-j+\mu}{d}\\
&\times A_{n-d-r-s-1}(\mu(s+1+i)+a+g+j+1,\mu+1)\Bigg\}. \qquad (10.17)
\end{aligned}
$$

PROOF. Let $0P_1P_2\dots P_r$ be a generalized random walk path with $\Lambda_\mu^+(a)=r$, $Rf_\mu^+(a)=k$ as stipulated in the theorem, P_i $(i=1,2,\dots,r)$ being the positive reflection points at height a. The path is thus divided into $r+1$ segments by these r positive reflections. Of these $r+1$ segments, there will be

(i) one segment in the beginning from the origin to the first positive reflection at height a, i.e., $0P_1$ and it may be of any length and occurs with probability

$$
\begin{aligned}
&(py)^a\sum_{i=0}^{\infty}(p^\mu qy^\mu)^i\sum_{j=0}^{\infty}\Bigg\{(\mu-1)p^\mu qy^\mu\sum_{i=0}^{\infty}(p^\mu qy^\mu)^i\\
&\qquad+p^\mu qy^\mu\sum_{i=1}^{\infty}(p^\mu qy^\mu)^i\Bigg\}^j p^\mu qy^\mu \qquad \text{by (10.4), (10.5) and (10.6)}\\
&=(py)^a p^\mu qy^{\mu+1}/(1+p^\mu qy^\mu-\mu p^\mu qy^{\mu+1}).
\end{aligned}
$$

(ii) $r-k$ segments, each of length $\mu+1$ and each with probability $p^\mu q$.

(iii) $k-1$ segments, each of length $>\mu+1$ and each with probability

$$
\begin{aligned}
&\sum_{j=0}^{\infty}\left\{\left[(\mu-1)p^\mu qy^\mu+(p^\mu qy^\mu)^2\right]\sum_{i=0}^{\infty}(p^\mu qy^\mu)^i\right\}^j p^\mu qy^\mu-p^\mu q\\
&\qquad=\frac{p^\mu q\left[y^\mu-(1+p^\mu qy^\mu-\mu p^\mu qy^{\mu+1})\right]}{1+p^\mu qy^\mu-\mu p^\mu qy^{\mu+1}},
\end{aligned}
$$

by (10.5), (10.6) and (ii) above.

(iv) One segment at the end from P_r to ∞ (i.e., from the last positive reflection at height a to ∞) and it may be of any length and with probability

$$
\begin{aligned}
&\sum_{j=0}^{\infty}\left\{(\mu-1)p^\mu qy^\mu\sum_{i=0}^{\infty}(p^\mu qy^\mu)^i+p^\mu qy^\mu\sum_{i=1}^{\infty}(p^\mu qy^\mu)^i\right\}^j\\
&\qquad\times\left\{p^\mu qy^\mu(1-py)+\sum_{j=1}^{\mu-1}(py)^j q\{1-(py)^{\mu-j+1}\}\right\}\\
&\qquad=\frac{p^\mu qy^\mu(1-py)+\sum_{j=1}^{\mu-1}(py)^j q\{1-(py)^{\mu-j+1}\}}{1-\mu p^\mu qy^{\mu+1}+p^\mu qy^\mu}.
\end{aligned}
$$

Further, r positive reflections will form k runs in $\binom{r-1}{k-1}$ ways. Thus, on identifying $h(p)$ as $\Pr\left[\Lambda_\mu^+(a)=r,\ Rf_\mu^+(a)=k\right]$, we have

$$\begin{aligned}
h(p)/\delta &= \binom{r-1}{k-1}(py)^a \frac{p^\mu q y^{\mu+1}}{1+p^\mu q y^\mu - \mu p^\mu q y^{\mu+1}} (p^\mu q)^{r-k} \\
&\quad \times \left\{\frac{p^\mu q[y^\mu - (1+p^\mu q y^\mu - \mu p^\mu q y^{\mu+1})]}{1+p^\mu q y^\mu - \mu p^\mu q y^{\mu+1}}\right\}^{k-1} \\
&\quad \times \frac{1}{1-\mu p^\mu q y^{\mu+1} + p^\mu q y^\mu} \\
&\quad \times \left\{p^\mu q y^\mu (1-py) + \sum_{j=1}^{\mu-1} (py)^j q\{1-(py)^{\mu-j+1}\}\right\} \Big/ \delta \\
&= \binom{r-1}{k-1}(p^\mu q)^r (py)^a y^{\mu+1}\left[y^\mu - 1 - p^\mu q y^\mu + \mu p^\mu q y^{\mu+1}\right]^{k-1} \\
&\quad \times \left\{1+p^\mu q y^\mu - \mu p^\mu q y^{\mu+1}\right\}^{-(k+1)} \\
&\quad \times \left\{p^\mu q y^\mu (1-py) + \sum_{j=1}^{\mu-1} (py)^j q\{1-(py)^{\mu-j+1}\}\right\} \Big/ \delta \qquad (10.18)
\end{aligned}$$

in which the coefficient of $(p^\mu q)^n$, by Theorem 10.3.1, gives the desired result in (10.17). ■

Deductions

(A) Putting $\mu = 1$ in (10.18), we get

$$\begin{aligned}
h(p)/(1-2p) &= \binom{r-1}{k-1}(pq)^{r-a-k-2} p^{2a+2k+2}(1+p)^{k-1} \\
&\quad \times (1-p^2/q)^{-k+1},
\end{aligned}$$

which is in agreement with Saran and Rani (1991a).

(B) Summing (10.18) over k, we get

$$\begin{aligned}
h_1(p)/\delta &= \Pr[\Lambda_\mu^+(a)=r]/\delta \\
&= (p^\mu q)^r p^a y^{\mu r+a+1}\left\{1+p^\mu q y^\mu - \mu p^\mu q y^{\mu+1}\right\}^{-(r+1)} \\
&\quad \times \left\{p^\mu q y^\mu (1-py) + \sum_{j=1}^{\mu-1} (py)^j q - (\mu-1) q (py)^{\mu+1}\right\} \Big/ \delta \\
&= \sum_{s=0}^{\infty}\sum_{g=0}^{s} \binom{s+r}{s}\binom{s}{g}(-1)^{s-g}\mu^g (p^\mu q)^{r+s} p^a\ y^{\mu(r+s)+a+g+1}
\end{aligned}$$

$$\times \left\{ p^\mu q y^\mu (1-py) + \sum_{j=1}^{\mu-1} (py)^j q\{1-(py)^{\mu-j+1}\} \right\} \Big/ \delta \tag{10.19}$$

in which the coefficient of $(p^\mu q)^n$ gives

$$\begin{aligned}
&\binom{(\mu+1)n}{n} \Pr\left[\Lambda^+_{\mu,n}(a) = r\right] \\
&= \sum_{s=0}^{\infty} \sum_{g=0}^{s} \binom{s+r}{s} \binom{s}{g} (-1)^{s-g} \mu^g \\
&\times \Bigg\{ \sum_{\lambda=\langle a/\mu \rangle}^{\infty} \binom{(\mu+1)\lambda - a}{\lambda} \\
&\qquad \times A_{n-r-s-1-\lambda}(\mu(r+s+1)+a+g+1, \mu+1) \\
&- \sum_{\lambda_1=\langle (a+1)/\mu \rangle}^{\infty} \binom{(\mu+1)\lambda_1 - a - 1}{\lambda_1} \\
&\qquad \times A_{n-r-s-1-\lambda_1}(\mu(r+s+1)+a+g+1, \mu+1) \\
&+ \sum_{j=1}^{\mu-1} \sum_{\lambda_2=\langle (a+1-j)/\mu \rangle}^{\infty} \binom{(\mu+1)\lambda_2 - a - j + \mu}{\lambda_2} \\
&\qquad \times A_{n-r-s-1-\lambda_2}(\mu(r+s)+a+g+1, \mu+1) \\
&- (\mu-1) \sum_{\lambda_1=\langle (a+1)/\mu \rangle}^{\infty} \binom{(\mu+1)\lambda_1 - a - 1}{\lambda_1} \\
&\qquad \times A_{n-r-s-1-\lambda_1}(\mu(r+1+s)+a+g+2, \mu+1) \Bigg\}.
\end{aligned}$$

(C) Summing (10.18) over r, we get

$$\begin{aligned}
h_2(p)/\delta &= \Pr\left[Rf^+_\mu(a) = k\right]/\delta \\
&= (py)^a y^{\mu+1} \{p^\mu q/(1-p^\mu q)\}^k \left\{y^r - 1 - p^\mu q y^\mu + \mu p^\mu q y^{\mu+1}\right\}^{k-1} \\
&\quad \times \left\{1 + p^\mu q y^\mu - \mu p^\mu q y^{\mu+1}\right\}^{-(k+1)} \\
&\quad \times \left\{ p^\mu q y^\mu (1-py) + \sum_{j=1}^{\mu-1} (py)^j q\{1-(py)^{\mu-j+1}\} \right\} \Big/ \delta
\end{aligned} \tag{10.20}$$

in which the coefficient of $(p^\mu q)^n$ gives

$$
\begin{aligned}
&\binom{(\mu+1)n}{n} \Pr\left[Rf^{+}_{\mu,n}(a)=k\right] \\
&= \sum_{t=0}^{\infty}\sum_{i=0}^{k-1}\sum_{s=0}^{\infty}\sum_{g=0}^{s}\binom{k+t-1}{t}\binom{k-1}{i}\binom{s+i+1}{s}\binom{s}{g} \\
&\quad \times (-1)^{k-1-i+s-g}\mu^{g}\Bigg\{\sum_{b=\langle a/\mu\rangle}^{\infty}\binom{(\mu+1)b-a}{a} \\
&\qquad \times A_{n-b-k-t-s-1}(\mu(s+i+2)+a+g+1,\mu+1) \\
&\quad -\mu\sum_{c=\langle (a+1)/\mu\rangle}^{\infty}\binom{(\mu+1)c-a-1}{c} \\
&\qquad \times A_{n-c-k-t-s-1}(\mu(s+i+2)+a+g+2,\mu+1) \\
&\quad +\sum_{j=1}^{\mu-1}\sum_{d=\langle (a+j-\mu)/\mu\rangle}^{\infty}\binom{(\mu+1)d-a-j+\mu}{d} \\
&\qquad \times A_{n-d-k-t-s-1}(\mu(s+1+i)+a+g+j+1,\mu+1)\Bigg\}.
\end{aligned}
$$

10.7 Distributions of $N^{*}_{\mu,n}(a)$ and $R^{*}_{\mu,n}(a)$

Theorem 10.7.1

$$
\begin{aligned}
&\binom{(\mu+1)n}{n}\Pr\left[N^{*}_{\mu,n}(a)=c,\ R^{*}_{\mu,n}(a)=k\right] \\
&= \binom{c-1}{k-1}\sum_{i=0}^{k-1}\sum_{s=0}^{i}\binom{k-1}{i}\binom{i}{s}(-1)^{k-1-s}(\mu-1)^{c-1-i}\mu^{s} \\
&\quad \times\Bigg\{\sum_{j=0}^{\mu-1}\sum_{r=\langle (2\mu+a-j)/\mu\rangle}^{\infty}\binom{(\mu+1)r-2\mu-a+j}{r} \\
&\qquad \times A_{n-r-c}(\mu(i+1)+a+s+i-j+2,\mu+1) \\
&\quad -\mu\sum_{t=\langle (2\mu+a+1)/\mu\rangle}^{\infty}\binom{(\mu+1)t-2\mu-a-1}{t} \\
&\qquad \times A_{n-t-c}(\mu(i+1)+a+s+i+3,\mu+1)\Bigg\}.
\end{aligned}
\tag{10.21}
$$

PROOF. A path contributing to (10.21) comprises $c+1$ independent segments as follows:

(a) The first segment from the origin to the first upward crossing of height a and will be with probability

$$(py)^a \sum_{i=0}^{\infty} (p^\mu q y^\mu)^i = p^a y^{a+1} \qquad \text{by (10.4) and (10.5);}$$

(b) $c-k$ segments each of length $\mu+1$ and each with probability $(\mu-1)p^\mu q$;

(c) $k-1$ segments each of length $> \mu+1$ and each with probability

$$\begin{aligned}&\sum_{i=0}^{\infty}(p^\mu q y^\mu)^i(\mu-1)p^\mu q y^\mu \sum_{j=1}^{\infty}(p^\mu q y^\mu)^j \\ &\quad + \sum_{i=1}^{\infty}(p^\mu q y^\mu)^i \sum_{j=1}^{\infty}(p^\mu q y^\mu)^j - (\mu-1)p^\mu q \\ &\quad = [(\mu-1+p^\mu q y^\mu)p^\mu q y^\mu - (1-p^\mu q y^\mu)^2(\mu-1)p^\mu q]y^2;\end{aligned}$$

(d) the last segment from the last upward crossing to ∞ with probability

$$\sum_{i=0}^{\infty}(p^\mu q y^\mu)^i q \left\{\sum_{j=0}^{\mu-1}(py)^{\mu-j} - \mu(py)^{\mu+1}\right\} = qy\left\{\sum_{j=0}^{\mu-1}(py)^{\mu-j} - \mu(py)^{\mu+1}\right\}.$$

Now $(c-k)$ segments each of length $\mu+1$ are to be combined with $k-1$ segments each of length $> \mu+1$ so as to form k runs of total c upward crossings, which is possible in $\binom{c-1}{k-1}$ ways. Thus, on identifying $h(p)$ as $\Pr[N_\mu^*(a) = c,\ R_\mu^*(a) = k]$, we have

$$\begin{aligned} h(p)/\delta &= \binom{c-1}{k-1} p^a y^{a+1}\{(\mu-1)p^\mu q\}^{c-k} \\ &\quad \times \{(\mu-1+p^\mu q y^\mu)p^\mu q y^\mu - (1-p^\mu q y^\mu)^2(\mu-1)p^\mu q\}^{k-1} y^{2k-2} \\ &\quad \times qy\left\{\sum_{j=0}^{\mu-1}(py)^{\mu-j} - \mu(py)^{\mu+1}\right\}\Big/\delta \\ &= \binom{c-1}{k-1}(\mu-1)^{c-k}(p^\mu q)^{c-1}p^a y^{a+1}\,\{(\mu y-1)y^{\mu+1}-(\mu-1)\}^{k-1} \\ &\quad \times qy\left\{\sum_{j=0}^{\mu-1}(py)^{\mu-j} - \mu(py)^{\mu+1}\right\}\Big/\delta \\ &= \binom{c-1}{k-1}\sum_{i=0}^{k-1}\sum_{s=0}^{i}\binom{k-1}{i}\binom{i}{s}(-1)^{k-1-s}(\mu-1)^{c-i-1}\mu^s(p^\mu q)^c p^{\mu+a} \\ &\quad \times y^{\mu i+a+s+i+2}\left\{\sum_{j=0}^{\mu-1}(py)^{\mu-j} - \mu(py)^{\mu+1}\right\}\Big/\delta \end{aligned} \tag{10.22}$$

in which the coefficient of $(p^\mu q)^n$ gives the desired result in (10.21). ■

Deductions

(A) Summing (10.22) over k, we have

$$\begin{aligned} h_1(p)/\delta &= \Pr[N^*_\mu(a) = c]/\delta \\ &= (p^\mu q)^{c-1} p^a y^{a+1+(c-1)(\mu+1)} (\mu y - 1)^{c-1} qy \\ &\quad \times \left\{ \sum_{j=0}^{\mu-1} (py)^{\mu-j} - \mu (py)^{\mu+1} \right\} \Big/ \delta \end{aligned} \tag{10.23}$$

in which the coefficient of $(p^\mu q)^n$ gives

$$\begin{aligned} &\binom{(\mu+1)n}{n} \Pr\left[N^*_{\mu,n}(a) = c\right] \\ &= \sum_{i=0}^{c-1} \binom{c-1}{i} (-1)^{c-1-i} \mu^i \Bigg\{ \sum_{j=0}^{\mu-1} \sum_{r=\langle (a-j)/\mu \rangle}^{\infty} \binom{(\mu+1)r - a + j}{r} \\ &\quad \times A_{n-c-r}(\mu c + a + c + i - j + 1, \mu + 1) \\ &\quad - \mu \sum_{s=\langle (a+1)/\mu \rangle}^{\infty} \binom{(\mu+1)s - a - 1}{s} \\ &\quad \times A_{n-c-s}(\mu c + a + c + i + 2, \mu + 1) \Bigg\}. \end{aligned}$$

(B) Summing (10.22) over c, we get

$$\begin{aligned} h_2(p)/\delta &= \Pr[R^*_\mu(a) = k]/\delta \\ &= \sum_{g=0}^{\infty} \binom{k+g-1}{g} (\mu-1)^{k+g-1} (p^\mu q)^{k+g-1} p^a y^{a+1} \\ &\quad \times \{(\mu y - 1) y^{\mu+1} - (\mu - 1)\}^{k-1} qy \\ &\quad \times \left\{ \sum_{j=0}^{\mu-1} (py)^{\mu-j} - \mu (py)^{\mu+1} \right\} \Big/ \delta \end{aligned} \tag{10.24}$$

in which the coefficient of $(p^\mu q)^n$ gives

$$\begin{aligned} &\binom{(\mu+1)n}{n} \Pr[R^*_{\mu,n}(a) = k] \\ &= \sum_{g=0}^{\infty} \sum_{i=0}^{k-1} \sum_{s=0}^{i} \binom{k-1}{i} \binom{i}{s} \binom{k+g-1}{g} \\ &\quad \times (-1)^{k-1-s} (\mu-1)^{k+g-i-1} \mu^s \\ &\quad \times \Bigg\{ \sum_{j=0}^{\mu-1} \sum_{r=\langle (2\mu+a-j)/\mu \rangle}^{\infty} \binom{(\mu+1)r - 2\mu - a + j}{r} \end{aligned}$$

$$
\begin{aligned}
&\times A_{n-r-k-g}(\mu(i+1)+a+s+i-j+2,\mu+1) \\
-\mu &\sum_{t=\langle(2\mu+a+1)/\mu\rangle}^{\infty} \binom{(\mu+1)t-2\mu-a-1}{t} \\
&\times A_{n-t-k-g}(\mu(i+1)+a+s+i+3,\mu+1) \Big\}.
\end{aligned}
$$

References

1. Dwass, M. (1967). Simple random walk and rank order statistics, *Annals of Mathematical Statistics*, **38**, 1042–1053.

2. Kaul, C. L. (1982). Contribution in rank order statistics, *Ph.D. Thesis*, University of Delhi, Delhi, India.

3. Mohanty, S. G. and Handa, B. R. (1970). Rank order statistics related to a generalized random walk, *Studia Scientiarum Mathematicarum Hungarica*, **5**, 267–276.

4. Pratap, M. (1982). Contribution to random walk, *Ph.D. Thesis*, Garhwal University, Srinagar (Garhwal), India.

5. Saran, J. and Rani, S. (1990). Generalized random walk and distributions of some rank order statistics, *Journal of Statistical Planning and Inference*, **24**, 119–133.

6. Saran, J. and Rani, S. (1991a). On joint distributions of rank order statistics for equal sample sizes, *Statistics*, **22**, 299–312.

7. Saran, J. and Rani, S. (1991b). Distributions based on upcrossings, upward crossings, positive reflections and their runs in a generalized random walk, *Communications in Statistics—Theory and Methods*, **20**, 3389–3404.

8. Saran, J. and Sen, K. (1979). Distributions of rank order statistics related to a generalized random walk, *Metron*, **37**, 121–135.

9. Sen, K. and Kaul, C. L. (1985). Distributions of rank order statistics related to a generalized random walk, *Journal of Statistical Planning and Inference*, **12**, 177–191.

10. Sen, K. and Saran, J. (1983). On the distribution of crossings in a generalized random walk, *Journal of Statistical Planning and Inference*, **7**, 371–385.

11

On a Subset Sum Algorithm and Its Probabilistic and Other Applications

V. G. Voinov and M. S. Nikulin

Institute of Pure and Applied Mathematics, Almaty, Kazakhstan
Université Bordeaux 2, Bordeaux, France

Abstract: An algorithm for constructing partitions of an integer by arbitrary positive integers has been considered. The algorithm helps in introducing a class of discrete probability distributions which are useful in a sampling survey of populations, in constructing probability models describing texture images, etc. It can also be used in integer programming, cryptography, and some other problems.

Keywords and phrases: Diophantine equation, discrete probability distribution, compositions, cryptography, generating function, knapsack problem, partitions, subset sum problem

11.1 Introduction

Some problems of the construction of usual partitions and their use in probability and statistics have been considered by Voinov and Nikulin (1994, 1995, 1996). In this note, we emphasize problems relating to partitions of integers by an arbitrary given set of positive integers. One such problem, called the *knapsack problem*, attracted the attention of mathematicians for many years due to its implication in public-key cryptography. Diffie (1988) wrote: "Given a cargo vector of integers $\mathbf{a} = (a_1, a_2, \ldots, a_n)$, it is easy to add up the elements of any specified subvector. Presented with an integer S, however, it is not easy to find a subvector of $\mathbf{a}$ whose elements sum to S, even if such a subvector is known to exist. This knapsack problem is well known in combinatorics and is believed to be extremely difficult in general. It belongs to the class of NP-complete problems, problems thought not to be solvable in polynomial time on any deterministic computer."

Actually, there exists a well-known and simple algorithm solving this problem in exponential time. In this note, we describe a possibly new approach for this algorithm derivation and point out some of its applications.

In Section 11.2, we derive an algorithm for constructing such partitions and discuss its potential applications in genetics, integer programming, cryptography, etc. In Section 11.3, we introduce a class of discrete probability distributions relating to such partitions. These distributions turn out to be useful in a sampling survey of populations, in constructing probability models describing texture images, radioactive contamination of lands, etc. Finally, in Section 11.4, we consider some approaches to summation procedures used in the construction of partitions.

11.2 A Derivation of the Algorithm

Consider the problem of representing a positive integer n as a sum of at most $M \leq n$ given arbitrary positive integers $a_1, a_2, \ldots, a_l$, $l \in \mathbf{Z}^+$, the set of positive integers. In other words, we would like to consider all integral representations of n as

$$s_1 a_1 + s_2 a_2 + \cdots + s_l a_l = n, \tag{11.1}$$

where $s_1 + s_2 + \cdots + s_l \leq M$ and $s_i, i = 1, 2, \ldots, n$, are non-negative integers. The generating function for the number $R_n(M, l)$ of compositions of n such as in (11.1) is [see, for example, Voinov and Nikulin (1995)]

$$\Psi_{\mathbf{a}}(z) = (1 + z^{a_1} + \cdots + z^{a_{l-1}} + z^{a_l})^M = \sum_{n=0}^{M \max_{1 \leq i \leq l}\{a_i\}} R_n(M, l) z^n. \tag{11.2}$$

Here, by compositions, we mean partitions taking order of summands into account.

Writing $\Psi_{\mathbf{a}}(z)$ as

$$\Psi_{\mathbf{a}}(z) = [(1 + z^{a_1} + \cdots + z^{a_{l-1}}) + z^{a_l}]^M$$

and applying the binomial formula, we get

$$\begin{aligned}
\Psi_{\mathbf{a}}(z) &= \sum_{k=0}^{M} \binom{M}{k} z^{a_l k} (1 + z^{a_1} + \cdots + z^{a_{l-1}})^{M-k} \\
&= \sum_{k=0}^{M} \binom{M}{k} z^{a_l k} \sum_{t=0}^{(M-k)\max_{1 \leq i \leq l-1}\{a_i\}} R_t(M-k, l-1) z^t \\
&= \sum_{n=0}^{M \max_{1 \leq i \leq l}\{a_i\}} \Big\{ \sum_{s_l=0}^{[\frac{n}{a_l}]} \binom{M}{s_l} R_{n - s_l a_l}(M - s_l, l-1) \Big\} z^n
\end{aligned}$$

$$= \sum_{n=0}^{M \max_{1\le i\le l}\{a_i\}} R_n(M,l) z^n,$$

where $[x]$ denotes the greatest integer part of x. We have also used the fact that

$$R_{n-a_l s_l}(M-s_l, l-1) = 0 \qquad \text{if } n - a_l s_l > (M-s_l) \max_{1\le i\le l-1}\{a_i\}.$$

Strictly speaking, the upper limit of the summation over s_l is $\min\{M, [n/a_l]\}$. Since, by definition, $\binom{M}{s_l} = 0$ for $s_l > M$, we prefer to use for simplicity $[n/a_l]$ as the upper limit.

From the above, we obtain the recurrence relation

$$R_n(M,l) = \sum_{s_l=0}^{[\frac{n}{a_l}]} \binom{M}{s_l} R_{n-s_l a_l}(M-s_l, l-1). \tag{11.3}$$

This recurrence relation gives

$$\begin{aligned} R_n(M,l) &= \sum_{s_l=0}^{[\frac{n}{a_l}]} \sum_{s_{l-1}=0}^{[\frac{n-s_l a_l}{a_{l-1}}]} \cdots \sum_{s_2=0}^{[\frac{n-s_l a_l-\cdots-s_3 a_3}{a_2}]} \binom{M}{s_l} \\ &\quad \times \binom{M-s_l}{s_{l-1}} \cdots \binom{M-s_l-\cdots-s_3}{s_2} \\ &\quad \times R_{n-s_l a_l-\cdots-s_2 a_2}(M-s_l-\cdots-s_2, 1). \end{aligned} \tag{11.4}$$

Evidently,

$$(1+z^{a_1})^M = \sum_{k=0}^{M} \binom{M}{k} z^{a_1 k} = \sum_{n=0}^{M a_1} R_n(M,1) z^n = \sum_{k=0}^{n} R_{a_1 k}(M,1) z^{a_1 k},$$

where

$$R_{a_1 k}(M,1) = \begin{cases} \binom{M}{k} & \text{if } k = \frac{n}{a_1} \text{ is a non-negative integer,} \\ 0 & \text{otherwise.} \end{cases} \tag{11.5}$$

Hence,

$$\begin{aligned} &R_{n-s_l a_l-\cdots-s_2 a_2}(M-s_l-\cdots-s_2, 1) = R_{s_1 a_1}(M-s_l-\cdots-s_2, 1) \\ &= \begin{cases} \binom{M-s_l-\cdots-s_2}{s_1} & \text{if } s_1 = \frac{n-s_l a_l-\cdots-s_2 a_2}{a_1} \text{ is non-negative integer,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Since

$$\begin{aligned} &\binom{M}{s_l}\binom{M-s_l}{s_{l-1}} \cdots \binom{M-s_l-\cdots-s_3}{s_2}\binom{M-s_l-\cdots-s_2}{s_1} \\ &= \frac{M!}{(M-s_l-\cdots-s_1)! s_1! s_2! \cdots s_l!}, \end{aligned} \tag{11.6}$$

from Eqs. (11.4), (11.5) and (11.6) we obtain

$$R_n(M,l) = \sum_{s_l=0}^{[\frac{n}{a_l}]} \sum_{s_{l-1}=0}^{[\frac{n-s_l a_l}{a_{l-1}}]} \cdots \sum_{s_2=0}^{[\frac{n-s_l a_l-\cdots-s_3 a_3}{a_2}]} \frac{M!}{(M-s_1-\cdots-s_l)!s_1!s_2!\cdots s_l!} \tag{11.7}$$

if $s_1+s_2+\cdots+s_l \le M$, $s_1=(n-s_l a_l-\cdots-s_2 a_2)/a_1$ being a non-negative integer, and is zero otherwise.

From this, we see that for all sets $\{s_2,s_3,\ldots,s_l\}$ defined by sums in (11.7) such that $s_1+\cdots+s_l \le M$, the number of parts in (11.1) is less than or equal M and

$$a_1s_1+a_2s_2+\cdots+a_ls_l = a_1\frac{n-s_la_l-\cdots-s_2a_2}{a_1}+a_2s_2+\cdots+a_ls_l = n.$$

Hence, sets $\{s_2,s_3,\cdots,s_l\}$ define all partitions of n such as in (11.1) and terms

$$\frac{M!}{(M-s_1-\cdots-s_l)!s_1!s_2!\cdots s_l!}$$

count all compositions of n for fixed $s_1,s_2,\ldots,s_l$.

Hence, the partitions may be written down in the form

$$\{0^{M-s_1-\cdots-s_l}, a_1^{s_1},\ldots,a_l^{s_l}\}, \tag{11.8}$$

where $\{s_2,\ldots,s_l\}$ are sets of summation indices of (11.7) and $s_1=(n-s_la_l-\cdots-s_2a_2)/a_1$ is a non-negative integer. Notation (11.8) means that in each partition there will be $M-s_1-\cdots-s_l$ zeros, s_1 terms will be a_1, s_2 terms will be a_2, and so on.

Example 11.2.1 Let $a_1=2$, $a_2=5$, $a_3=3$, $M=5$ and $n=17$. By formula (11.7), for $l=3$ we have

$$R_{17}(5,3) = \sum_{s_3=0}^{5} \sum_{s_2=0}^{[\frac{17-3s_3}{5}]} \frac{5!}{(5-s_1-s_2-s_3)!s_1!s_2!s_3!},$$

where $s_1=(17-5s_2-3s_3)/2$ and $s_1+s_2+s_3 \le 5$.

For this example, there are 15 sets $\{s_1,s_2,s_3\}$ but only 3 of them satisfy the conditions that s_1 is a non-negative integer and that $s_1+s_2+s_3 \le 5$; these are $\{1,3,0\},\{2,2,1\}$ and $\{0,1,4\}$. Using (11.8), we then obtain three partitions of $n=17$ (with at most 5 parts) as

$$\begin{aligned}
\{0^1,2^1,5^3,3^0\} &= 1\cdot2+3\cdot5=17,\\
\{0^0,2^2,5^2,3^1\} &= 2\cdot2+2\cdot5+1\cdot3=17,\\
\{0^0,2^0,5^1,3^4\} &= 1\cdot5+4\cdot3=17.
\end{aligned}$$

Formula (11.7) can be easily transformed for the construction of partitions of n with exactly M parts. Since the generating function in this case is

$$\begin{aligned}\Psi_{\mathbf{a}}(z) &= (z^{a_1}+\cdots+z^{a_l})^M = z^{Ma_1}(1+z^{a_2-a_1}+\cdots+z^{a_l-a_1})^M \\ &= z^{Ma_1}\sum_{n=0}^{M\max\{a_i-a_1\}} R_n(M,l-1)z^n = \sum_{n=Ma_1}^{M\max\{a_i\}} R_{n-Ma_1}(M,l-1)z^n,\end{aligned}$$

using (11.7) we obtain

$$\begin{aligned}R_{n-Ma_1}(M,l-1) &= \sum_{s_l=0}^{[\frac{n-Ma_1}{a_l-a_1}]}\ \sum_{s_{l-1}=0}^{[\frac{n-Ma_1-s_l(a_l-a_1)}{a_{l-1}-a_1}]} \cdots \sum_{s_3=0}^{[\frac{n-Ma_1-s_l(a_l-a_1)-\cdots-s_4(a_4-a_1)}{a_3-a_1}]} \\ &\quad \frac{M!}{(M-s_l-s_{l-1}-\cdots-s_2)!s_2!s_3!\cdots s_l!},\end{aligned} \tag{11.9}$$

where $s_2+s_3+\cdots+s_l \le M$ and

$$s_2 = \frac{n-Ma_1-s_l(a_l-a_1)-\cdots-s_3(a_3-a_1)}{a_2-a_1}$$

is a non-negative integer, $l \ge 3$.

The partitions $n = a_1s_1+\cdots+a_ls_l$ may then be written as

$$\{a_1^{M-s_2-\cdots-s_l}, a_2^{s_2}, a_3^{s_3}, \ldots, a_l^{s_l}\}. \tag{11.10}$$

Example 11.2.2 Let $a_1 = 2$, $a_2 = 4$, $a_3 = 3$, $a_4 = 6$, $M = 5$ and $n = 26$. By formula (11.9), for $l = 4$ we have

$$R_{16}(5,3) = \sum_{s_4=0}^{4}\ \sum_{s_3=0}^{16-4s_4} \frac{5!}{(5-s_2-s_3-s_4)!s_2!s_3!s_4!},$$

where $s_2 = (16-4s_4-s_3)/s_2$ and $s_2+s_3+s_4 \le 5$.

There are 45 sets of $\{s_2, s_3, s_4\}$ and only two of them satisfy the conditions that s_2 is a non-negative integer and that $s_2+s_3+s_4 \le 5$; these are $\{2,0,3\}$ and $\{0,0,4\}$ which give the two partitions

$$\begin{aligned}\{2^0,4^2,3^0,6^3\} &= 2\cdot 4+3\cdot 6 = 26, \\ \{2^1,4^0,3^0,6^4\} &= 1\cdot 2+4\cdot 6 = 26,\end{aligned}$$

of $n = 26$ on exactly 5 parts.

The algorithm for the construction of partitions defined by (11.7) and (11.9) shows that the subset sum problem is solvable at least in principle.

A variant of the subset sum or knapsack problem considered above has found applications in public-key cryptography [Diffie (1988)].

This problem is usually posed as follows [Brickell (1985)]. Let a set of positive integers $a_1, a_2, \ldots, a_n$ be an unordered knapsack. To cryptanalyze the system, we are to solve the subset sum problem for any sum S, i.e. to find a vector $(s_1, \ldots, s_n)$, with $s_i \in \{0,1\}$, such that

$$s_1 a_1 + s_2 a_2 + \cdots + s_n a_n = S$$

if such a vector exists. The density of a set of weights $a_1, \ldots, a_n$ is defined by

$$d = \frac{n}{\log_2 \max_{1 \le i \le n}\{a_i\}}.$$

Since there will be in general many subsets of weights with a sum S when $d > 1$, only the case $d \le 1$ is used in cryptography [Coster *et al.* (1992)].

Evidently, for every d this problem is solved in principle by formula (11.7), which in this case becomes

$$R_S(n,n) = \sum_{s_n=0}^{\min(1,[\frac{S}{a_n}])} \sum_{s_{n-1}=0}^{\min(1,[\frac{S-a_n s_n}{a_{n-1}}])} \cdots \sum_{s_2=0}^{\min(1,[\frac{S-a_n s_n-\cdots-a_3 s_3}{a_2}])} \frac{n!}{(n-s_1-\cdots-s_n)!s_1!\cdots s_n!} \tag{11.11}$$

for $s_1+\cdots+s_n \le n$ and $s_1 = \frac{S-s_n a_n-\cdots-s_2 a_2}{a_1}$ being 0 or 1, and is zero otherwise.

Example 11.2.3 Let $a_1 = 30$, $a_2 = 29$, $a_3 = 32$, $a_4 = 31$, $a_5 = 33$, $n = 5$ and $S = 90$. The density in this case is $d = 5/\log_2 32 = 1$. From (11.11), we have

$$R_{90}(5,5) = \sum_{s_5=0}^{1} \sum_{s_4=0}^{\min(1,[\frac{90-33s_5}{31}])} \sum_{s_3=0}^{\min(1,[\frac{90-33s_5-31s_4}{32}])} \sum_{s_2=0}^{\min(1,[\frac{90-33s_5-31s_4-32s_3}{29}])} \frac{5!}{(5-s_1-\cdots-s_5)!s_1!\cdots s_5!},$$

where $s_1 + \cdots + s_5 \le 5$ and

$$s_1 = \frac{90-33s_5-31s_4-32s_3-29s_2}{30}.$$

From these, we find 11 sets of $\{s_1, \ldots, s_5\}$ and only the set $\{1,1,0,1,0\}$ satisfies the conditions that $s_1+\cdots+s_5 \le 5$ and s_1 is 0 or 1. From (11.8), with $M = n = 5$ we have the following unique solution of this knapsack

$$\{0^2, 30^1, 29^1, 32^0, 31^1, 33^0\} = 30 + 29 + 31 = 90.$$

An algorithm given by formula (11.11) is inapplicable in cryptography since its computational complexity is of the order of 2^{n-1}. Nevertheless, it can be used in combination with other integer programming techniques. At present, some algorithms are known [Coster *et al.* (1992)] which solve almost all subset sum problems of density $d < 0.9408\ldots$ in polynomial time.

An interesting application of the subset sum problem arises while constructing a mathematical model of universal genetic code [Shcherbak (1994-1996)]. The model is described by a system of 22 linear Diophantine equations of 25 non-negative variables x_j, $j = 1, 2, ..., 25$, and some inequalities imposed on x_j describing the nucleon sums of the 23 amino acids. The problem is to find all the solutions of this system.

Many techniques are known for solving such systems; see, for example, Smith (1861), McClellan (1973), Phrumkin (1976) and Votyakov and Phrumkin (1976). The algorithm of Votyakov and Phrumkin (1976) gives the general discrete solution of a system of linear equations in polynomial time, but the problem of enumerating all the solutions still remains open. In view of this, the exponential algorithm given by (11.7) and (11.8) which enumerates all solutions of Eq. (11.1) naturally becomes useful for solving the above problem, since its solution is defined by the intersection of sets of solutions of every equation of the system which satisfies inequalities imposed on some variables.

11.3 A Class of Discrete Probability Distributions

Suppose that an urn contains balls. The balls bear fixed positive numbers $a_1, a_2, ..., a_l$, $l \in \mathbf{Z}^+$. Let p_i be the probability that a ball bearing the number a_i will be drawn ($i = 1, 2, ..., l$) with $\sum_{i=1}^{l} p_i = 1$. Let the random variable X take the value r if, of n balls drawn with replacement, r_1 bear the number a_1, r_2 bear the number a_2, and so on, and $\sum_{i=1}^{l} a_i r_i = r$ with $\sum_{i=1}^{l} r_i = n$. The probability that the summation of numbers on balls drawn is r ($n\tilde{a}_1 \le r \le n\tilde{a}_2, \tilde{a}_1 = \min_{1\le i\le l}\{a_i\}, \tilde{a}_2 = \max_{1\le i\le l}\{a_i\}$) is

$$\Pr[X = r] = \sum_{\sum_{i=1}^{l} a_i r_i = r} \binom{n}{r_1, r_2, ..., r_l} \prod_{i=1}^{l} p_i^{r_i}, \tag{11.12}$$

where

$$\binom{n}{r_1, r_2, ..., r_l} = \frac{n!}{r_1! \ldots r_{l-1}!(n - \sum_{i=1}^{l-1} r_i)!}$$

and is zero if $\sum_{i=1}^{l-1} r_i > n$. If can be easily shown, using the arguments of Panaretos and Xekalaki (1986), that (11.12) is a proper probability distribution; see also Johnson, Kotz and Balakrishnan (1997).

Its probability generating function $G(s)$ is

$$G(s) = \left(\sum_{i=1}^{l} p_i s^{a_i}\right)^n . \tag{11.13}$$

All $n(\tilde{a}_2 - \tilde{a}_1) + 1$ probabilities of (11.12) can naturally be defined by (11.13); for example,

$$\Pr[X = n\tilde{a}_1] = \frac{1}{(n\tilde{a}_1)!} \frac{\partial^{n\tilde{a}_1} G(s)}{\partial s^{n\tilde{a}_1}} |_{s=0}.$$

To simplify the evaluation of probabilities in (11.12), one may use the algorithm derived in the last section. Using Eqs. (11.9) and (11.10), with $n = r$, $M = n$ the distribution in (11.12) becomes

$$\begin{aligned}\Pr[X = r] \quad = \quad & \sum_{s_l=0}^{[\frac{r-na_1}{a_l-a_1}]} \sum_{s_{l-1}=0}^{[\frac{r-na_1-s_l(a_l-a_1)}{a_{l-1}-a_1}]} \cdots \sum_{s_3=0}^{[\frac{r-na_1-s_l(a_l-a_1)-\cdots-s_4(a_4-a_1)}{a_3-a_1}]} \\ & \frac{n!}{(n - s_2 - s_3 - \cdots - s_l)! s_2! s_3! \cdots s_l!} p_1^{n-s_2-\cdots-s_l} p_2^{s_2} \cdots p_l^{s_l},\end{aligned} \tag{11.14}$$

where $s_2 + s_3 + \cdots + s_l \le n$, $l \ge 3$ and

$$s_2 = \frac{r - na_1 - s_l(a_l - a_1) - \cdots - s_3(a_3 - a_1)}{a_2 - a_1}$$

is a non-negative integer.

This probability model is suitable if a set of possible numbers on balls is confined to a small ordered or disordered set of positive integers, even excluding zero. Suppose one has to plan a cloth production. Having obtained sample estimates of probabilities $p_1, \ldots, p_l$ of heights $a_1, \ldots, a_l$, he/she will now be able to evaluate the probability that the average height of individuals belongs to a prescribed interval.

The model is also applicable for describing radioactive contamination of lands where, due to the natural background, digitized levels of measured radioactivity belong to a set with a similar property as above. Due to the stochastic nature of radioactive fields, probabilities of summed levels can describe these fields more adequately.

11.4 A Remark on a Summation Procedure When Constructing Partitions

Suppose we have to construct partitions of an integer k on at most n parts with each part less than or equal l. This problem is a particular case of (11.1) with

$a_1 = 1, a_2 = 2, \ldots, a_l = l$ and $\Psi(z)$, the generating function for the number $a_k(n, l)$ of partitions, being

$$\Psi(z) = (1 + z + z^2 + \cdots + z^l)^n = \sum_{k=0}^{nl} a_k(n, l) z^k.$$

Using (11.7), we then have

$$a_k(n,l) = \sum_{s_l=0}^{[\frac{k}{l}]} \sum_{s_{l-1}=0}^{[\frac{k-ls_l}{l-1}]} \cdots \sum_{s_2=0}^{[\frac{k-ls_l-\cdots-3s_3}{2}]} \frac{n!}{(n-s_1-\cdots-s_l)!s_1!s_2!\cdots s_l!}, \tag{11.15}$$

where

$$s_1 = k - 2s_2 - \cdots - ls_l \quad \text{and} \quad s_1 + s_2 + \cdots + s_l \leq n. \tag{11.16}$$

Partitions can be represented in this case as

$$\{0^{n-s_1-\cdots-s_l}, 1^{s_1}, 2^{s_2}, \cdots, l^{s_l}\}. \tag{11.17}$$

The way of summation by partitions [summation by $s_l, s_{l-1}, \ldots, s_2$ in (11.15)] is well-known. An alternate way of summation by the same partitions has been proposed by Voinov and Nikulin (1994, 1995) and can be written as follows [see, for example, Voinov and Nikulin (1995, Formula 7)]

$$a_k(n,l) = \sum_{l_1=k-n}^{[\frac{(l-1)k}{l}]} \sum_{l_2=(2l_1-k)_+}^{[\frac{(l-2)l_1}{l-1}]} \cdots \sum_{l_{l-1}=(2l_{l-2}-l_{l-3})_+}^{[\frac{l_{l-2}}{2}]} \frac{n!}{(n-k+l_1)!(k-2l_1+l_2)!(l_1-2l_2+l_3)!\cdots(l_{l-2}-2l_{l-1})!l_{l-1}!}, \tag{11.18}$$

with partitions being

$$\{0^{n-k+l_1}, 1^{k-2l_1+l_2}, \cdots, (l-1)^{l_{l-2}-2l_{l-1}}, l^{l_{l-1}}\}. \tag{11.19}$$

The way of summation in (11.18) is more suitable than in (11.15) and (11.16) in the sense that it saves considerable computing time. Let, for example, $l = 3$, $n = 6$ and $k = 14$. In this case, formula (11.15) enumerates 24 sets $\{l_1, l_2, l_3\}$ but only four of them, viz. $\{0,4,2\}, \{1,2,3\}, \{2,0,4\}$ and $\{0,1,4\}$, satisfy the condition in (11.16). Formula (11.18) does not have a condition like (11.16) and enumerates for this example exactly four sets $\{l_1, l_2\}$, viz. $\{8,2\}$,$\{8,3\}$,$\{8,4\}$

and $\{9, 4\}$. Naturally, we obtain the same partitions in both cases, which by formulas (11.17) and (11.19) are

$$\{0^0, 1^0, 2^4, 3^2\}, \{0^0, 1^1, 2^2, 3^3\}, \{0^0, 1^2, 2^0, 3^4\}, \{0^1, 1^0, 2^1, 3^4\}.$$

Acknowledgements. The authors would like to thank Professors N. Balakrishnan, S. Kotz and A. M. Odlyzko for their helpful discussions.

References

1. Brickell, E. F. (1985). Breaking iterated knapsacks, In *Advances in Cryptology: Proceedings of Crypto'84* (Eds., G. R. Blakley and D. Chaum), pp. 342–357, Berlin: Springer-Verlag.

2. Coster, M. J., Joux, A., LaMacchia, B. A., Odlyzko, A. M., Schnorr, C. P. and Stern, J. (1992). Improved low-density subset sum algorithms, *Journal of Computational Complexity*, **2**, 111–128.

3. Diffie, W. (1988). The first ten years of public-key cryptography, *Proceedings of the IEEE*, **76**, 560–577.

4. Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

5. McClellan, M. T. (1973). The exact solution of systems of linear equations, *Journal of the Association of Computing Machinery*, **20**, 563–588.

6. Panaretos, J. and Xekalaki, E. (1986). On generalized binomial and multinomial distributions and their relation to generalized distributions, *Annals of the Institute of Statistical Mathematics*, **38**, 223–231.

7. Phrumkin, M. A. (1976). An algorithm solving systems of linear equations in integers, In *Investigations on Discrete Optimization*, pp. 97–127, "Nauka", Moscow. (in Russian)

8. Shcherbak, V. I. (1994). Sixty-four triplets and 20 canonical amino acids of the genetic code: the arithmetical regularities, *Journal of Theoretical Biology*, **166**, 475–477.

9. Shcherbak, V. I. (1995). A mathematical model of the universal genetic code, *Ph.D. Thesis* (N 0495 RK 00500), Kazakh State Institute of Scientific and Technical Information, Almaty, Kazakhstan.

10. Shcherbak, V. I. (1996). An analytical formula of the genetic code, *Journal of Theoretical Biology* (submitted for publication).

11. Smith, H. (1861). On systems of linear indeterminate equations and congruences, *Philosophical Transactions*, **151**.

12. Voinov, V. G. and Nikulin, M. S. (1994). On power series, Bell polynomials, Hardy-Ramanujan-Rademacher problem and its statistical applications, *Kybernetika*, **30**, 343–358.

13. Voinov, V. G. and Nikulin, M. S. (1995). Generating functions, problems of additive number theory and some statistical applications, *Rev. Roumaine Math. Pures Appl.*, **40**, 107–147.

14. Voinov, V. G. and Nikulin, M. S. (1996). *Unbiased Estimators and Their Applications. Volume 2: Multivariate Case*, Dordrecht: Kluwer Academic Publishers.

15. Votyakov, A. A. and Phrumkin, M. A. (1976). An algorithm giving a general integer solution of a system of linear equations, In *Investigations on Discrete Optimization*, pp. 128–141, "Nauka", Moscow. (in Russian)

12

I and J Polynomials in a Potpourri of Probability Problems

Milton Sobel

University of California, Santa Barbara, CA

Abstract: Some new methodology is developed for Network Reliability problems and for random paths on finite lattices. In terms of stopping sets which define different (random) ways of reaching a goal in a geometrical setting, certain I and J polynomials are developed which give rise to the probability distribution (and its moments) of the waiting time (WT) needed to reach the preassigned goal. These new techniques have many different applications from Network Reliability to Recreational problems of tic-tac-toe and attacking all the squares on a chess board with randomly placed rooks or knights or queens, etc. Failure probabilities need not be equal and random sampling can be carried out with replacement, without replacement or by Pólya sampling schemes.

Keywords and phrases: Network reliability, waiting time problems in a geometrical setting, Dirichlet methodology, random paths on lattices

12.1 Introduction

In a recent paper by Boehme, Kossow and Preuss (1992), the concept of system reliability for consecutive k-out-of-n: F systems was generalized and applied to linear and circular lattice networks. The consecutive idea was extended to any consecutive set connected by bonds in a preassigned linear or circular lattice network. In the present paper, one of the goals is to continue with this approach without restricting ourselves to any special classes of lattice networks. Actually, the central theme of this paper is waiting time (WT) problems and we regard reliability as an important application of these problems. We introduce I and J polynomials which reduce a large class of inverse sampling WT problems to a small finite linear combination of solvable fixed-sample-size problems with sample-size parameter n. Basically the argument depends on inclusion-exclusion, but Dirichlet methodology, introduced in Sobel, Uppuluri and Frankowski (1977, 1985), is useful in the final steps to get numerical solutions. With a little extra work, it is shown that the method of I and J polynomials can also handle reliability problems with unequal failure probabilities for the individual units. If a problem can be regarded as a sampling WT problem, then we give three different expectation $E[\mathrm{WT}]$ answers for each problem depending on whether the sampling is carried out without replacement (H) or with replacement (M) or by Polya sampling (P), i.e., putting back two (or more) of the same for each item removed. All of these results, including the variance $\sigma^2(\mathrm{WT})$ and $\Pr[\mathrm{WT} \geq n+1]$, are derived from one appropriate I or J polynomial. The reason for considering both the I and the J polynomial is that each pair of I and J polynomials defines a pair of dual problems with related results. This problem-duality is not the same as graphic-duality, which defines equivalent problems; both dualities are illustrated by several examples in the paper. The I and/or J polynomial method can also be used in path problems and percolation problems on a finite lattice network. The results are exact in all cases; no asymptotic results have yet been obtained for very large lattice sizes.

Aside from the basic references Sobel, Uppuluri and Frankowski (1977, 1985) for Dirichlet methodology, there are two places in the literature where the above method of inclusion-exclusion was successfully employed; one is Sobel and Uppuluri (1974), where we wait for X-rays to hit each of the 4 cells in a local 2×2 structure within a larger structure. The other is Gleser *et al.* (1989), where a single die is marked on its 6 sides with the pairs (1,2), (1,3), (1,4), (2,3), (2,4) and (3,4); we are interested in the waiting time to see all the four digits 1, 2, 3, 4, each at least once.

12.2 Guide to the Problems of this Paper

Problems 12.1, 12.2 and 12.3, are explained with Tables 12.1, 12.2 and 12.3, respectively. Each illustrates the concept of the problem-dual; thus Problems 12.1A and 12.1B are dual Problems as are 12.2A and 12.2B and also 12.3A and 12.3B. For each pair, under H-sampling the expected waiting times (WT) add to one more than the original number of sampling elements and the variances of WT are equal for H-sampling. This problem-duality is not to be confused with graph-duality which is also included in the tables. For example, under graph-duality the 6 faces and 8 vertices of a cube are interchanged with the 6 vertices and 8 faces of a regular octahedron; the number of edges is 12 in both cases.

The explanation of the derivation of the I- and J- polynomials is given step-by-step in Problem 12.4A, which deals with a triangular array problem. The reason for considering such an array was to show that we are not limited to the consideration of only linear (i.e., rectangular) and circular (or cylindrical) arrays of nodes as in Boehme, Kossow and Preuss (1992). Problem 12.4 is then extended to show that the method of I- and J-polynomials "works" in a multivariate setting with unequal parameters. However, different sampling methods no longer yield intuitively comparable results.

Problem 12.5, dealing with the edges of a square with diagonals that do not intersect, illustrates the fact that the concept of problem duality can be extended to different levels ($d = 1, 2, \ldots$).

More examples of problem-duality are given in Problems 12.6, 12.7 and 12.8.

Problems 12.9, 12.10, 12.11 and 12.12 are problems of percolation type. They all deal with a rectangular or square lattice and in each case the I- and/or J- polynomials gives exact answers for the expected waiting time (in terms of number of observations) needed to complete any one of the specified class of paths. Variance of WT and percentiles (which can be treated as upper confidence limits of WT) are all obtained from either the I- or the J-polynomial.

Pólya sampling with $c = 1$ in Tables 12.1, 12.2 and 12.3 means that you put back 2 items (or 1 extra) for each item removed for sampling.

It should be noted in Tables 12.1, 12.2 and 12.3 for each pair of IJ-dual problems (like 12.1A and 12.1B) that if the sampling set is the same (as it is in all 3 Tables), then for H-sampling (i.e., without replacement) we have (i) the variances $\sigma^2(\mathrm{WT})$ are the same for A and B, and (ii) the sum of the expectations $E(\mathrm{WT})$ for A and B is $V + 1$, $E + 1$, $F + 1$ or $C + 1$ depending on whether the common sampling set is a set of vertices, edges, faces or cells; the capital letter denotes the size of this sampling set.

In Tables 12.2 and 12.3, system reliability (for sampling without replacement) is included. If we place J^α by p^α for each α (where $p = 1 - q =$ unit reliability), then we obtain the system reliability in terms of the unit reliability. If we replace J^α by $\binom{E-\alpha}{n}/\binom{E}{n}$ for each α, then we obtain $\Pr[\mathrm{WT} \geq n + 1]$, where WT is the waiting time until system failure; here, $E = 12$ (edges) in both tables. By summing the latter on n, we obtain the E[Reliability] or the expected number of failures the system will survive if we are sampling without replacement.

Table 12.1: Sampling the $V = 8$ vertices of a regular cube; dual problems One-at-a-time sampling method used

Graph-dual problems are grouped together below	W/O replacement (H)	W/replacement (M)	Polya W/$c = 1$ (P)
Problem 12.1A: SVC: Sample vertices of a cube until 1 complete face is obtained. Graph-dual problem: Sample the faces of an octahedron until you obtain all 4 associated with any one vertex			
E[WT]	$197/35 = 5.628571$	9.000000	20.600000
Mode (WT), Pr{mode}	6, $1/7 = 0.142857$	7, 0.129956	8, 0.066201
Min(WT), Max(WT)	4, 7	4, ∞	4, ∞
σ^2(WT)	$846/1225 = 0.690612$*	13.688889	∞
I-polynomial	$6I^4 - 12I^6 + 8I^7 - I^8$ (Sum of Coeff's = 1) $= \Sigma_\alpha a_\alpha I^\alpha$		
J-polynomial	$4J^2 + 8J^3 - 36J^4 + 40J^5 - 16J^6 + J^8$ (Sum of Coeff's = 1) $= \Sigma_\alpha b_\alpha J^\alpha$		
(Dual) **Problem 12.1B:** SVC: Sample the vertices of a cube until 1 vertex is obtained on each face. Graph-Dual Problem: Sample the faces of an octahedron until you obtain at least 1 face associated with each vertex			
E[WT]	$118/35 = 3.371429$	4.142857	5.553333
Mode (WT), Pr{Mode}	3, $3/7 = 0.428571$	3, 0.248047	3, 0.222222
Min(WT), Max(WT)	2, 5	2, ∞	2, ∞
σ^2(WT)	$846/1225 = 0.690612$*	2.925172	14.648889
Sum of expect. for A, B	$V + 1 = 9$*	–	–
I-polynomial	$4I^2 + 8I^3 - 36I^4 + 40I^5 - 16I^6 + I^8$		
J-polynomial	$6J^4 - 12J^6 + 8J^7 - J^8$		
* Note also that the variances are equal under H-sampling			

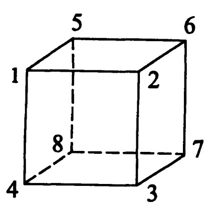
Common diagram	Stopping sets for Problem 12.1A	Pr[WT $\geq n + 1$] under Polya ($c = 1$) Sampling
	(1, 2, 3, 4) (1, 2, 5, 6) (2, 3, 6, 7) (3, 4, 7, 8) (1, 4, 5, 8) (5, 6, 7, 8) 6 stopping sets	$P_A = \sum_{\alpha=2}^{7} b_\alpha \binom{7}{\alpha}/\binom{n+7}{\alpha} + \delta_{no}$, $P_B = \sum_{\alpha=4}^{7} a_\alpha \binom{7}{\alpha}/\binom{n+7}{\alpha} - \delta_{no}$.

Table 12.2: Sampling the $E = 12$ edges of a cube to get a square face
One-at-a-time sampling method used

Graph-dual problems are grouped together below	W/O replacement (H)	W/replacement (M)	Polya W/$c = 1$ (P)
Problem 12.2A: SEC: Sample the edges of a cube until a complete square face is obtained. Graph-dual problem: Sample the edges of an octahedron until you obtain 4 associated with any one vertex			
E[WT]	$505/56 = 7.651515$	11.752814	22.147619
Mode (WT), Pr{Mode}	8, 0.312121	10, 0.106282	12, 0.049368
Min(WT), Max(WT)	4, 10	4, ∞	4, ∞
σ^2(WT)	1.608861	18.304525	406.287735
I-polynomial	$6I^4 - 12I^7 - 3I^8 + 8I^9 + 12I^{10} - 12I^{11} + 2I^{12} = \Sigma_\alpha a_\alpha\, I^\alpha$		
J-polynomial	$8J^3 + 66J^4 - 348J^5 + 672J^6 - 684J^7 + 381J^8 - 92J^9$ $-12J^{10} + 12J^{11} - 2J^{12} = \Sigma_\alpha b_\alpha\, J^\alpha$		
System reliability	$1 - 6q^4 + 12q^7 + 3q^8 - 8q^9 - 12q^{10} + 12q^{11} - 2q^{12}$; $q = \Pr\{\text{failed item or edge}\} = 1 - p$ (System fails if each edge in any square face fails.)		
(Dual) **Problem 12.2B:** SEC: Sample the edges of a cube until you have at least 1 edge from each face. Graph-Dual Problem: Sample the edges of an octahedron until you obtain at least 1 edge associated with each vertex			
E[WT]	5.348484	6.904329	9.752381
Mode (WT), Pr{Mode}	5, 0.312121	5, 0.190008	5, 0.125824
Min(WT), Max(WT)	3, 9	3, ∞	3, ∞
σ^2(WT)	1.608861	7.837761	54.186303
Sum of expect. for A, B	$E + 1 = 13$	–	–
I-polynomial	$8I^3 + 66I^4 - 348I^5 + 672I^6 - 684I^7 + 381I^8 - 92I^9$ $-12I^{10} + 12I^{11} - 2I^{12}$		
J-polynomial	$6J^4 - 12J^7 - 3J^8 + 8J^9 + 12J^{10} - 12J^{11} + 2J^{12}$		

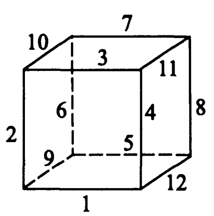
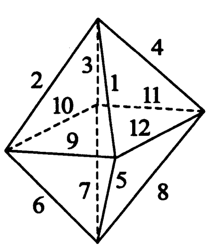
Common diagram	Stopping sets for Problem 12.2A	$\Pr[\text{WT} \geq n + 1]$ under Polya $(c = 1)$ Sampling
	(1, 2, 3, 4) (1, 5, 9, 12) (2, 6, 9, 10) (3, 7, 10, 11) (4, 8, 11, 12) (5, 6, 7, 8) 6 stopping sets	$P_A = \sum_{\alpha=3}^{11} b_\alpha \binom{11}{\alpha} / \binom{n+11}{\alpha} - 2\delta_{on}$, $P_B = \sum_{\alpha=4}^{11} a_\alpha \binom{11}{\alpha} / \binom{n+11}{\alpha} + 2\delta_{on}$.

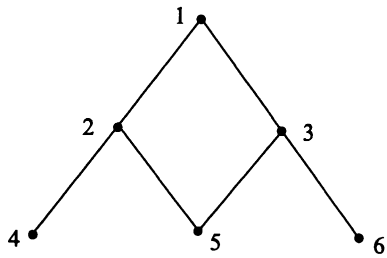
Table 12.3: Sampling the $E = 12$ edges of an octahedron to get a triangular face; One-at-a-time sampling method used

Graph-dual problems are grouped together below	W/O replacement (H)	W/replacement (M)	Polya W/$c = 1$ (P)
Problem 12.3A: SEO: Sample the edges of an octahedron until a complete triangular face is obtained. Graph-dual problem: Sample the edges of a cube until you obtain 3 edges associated with any one vertex			
E[WT]	6.005195	8.118182	12.271424
Mode(WT), Pr{Mode}	6, 0.270563	7, 0.140900	7, 0.086080
Min(WT), Max(WT)	3, 9	3, ∞	3, ∞
σ^2(WT)	1.84760	14.857948	83.045486
I-polynomial	$8I^3 - 12I^5 - 16I^6 + 24I^7 + 18I^8 - 24I^9 - 6I^{10} + 12^{11}$ $-3I^{12} = \Sigma_\alpha a_\alpha I^\alpha$		
J-polynomial	$9J^4 + 36J^5 - 152J^6 + 168J^7 - 27J^8 - 84J^9 + 72J^{10}$ $-24J^{11} + 3J^{12} = \Sigma_\alpha b_\alpha J^\alpha$		
System reliability	$1 - 8q^3 + 12q^5 + 16q^6 - 24q^7 - 18q^8 + 24q^9 + 6q^{10}$ $-12q^{11} + 3q^{12}$; $q = \Pr\{\text{failed edge}\} = 1 - p$ (System fails if each edge in any triangular face fails.)		
(Dual) **Problem 12.3B:** SEO: Sample the edges of a rectangular octahedron until you obtain at least 1 edge from each face. Graph-Dual Problem: Sample the edges of the cube until you obtain at least 1 edge associated with each vertex			
E[WT]	6.994808	10.233756	17.952381
Mode(WT), Pr{Mode}	7, 0.270563	8, 0.124446	9, 0.067437
Min(WT), Max(WT)	4, 10	4, ∞	4, ∞
σ^2(WT)	1.84760	16.806357	318.759635
Sum of expect. for A, B	$E + 1 = 13$	–	–
I-polynomial	$9I^4 + 36I^5 - 152I^6 + 168I^7 - 27I^8 - 84I^9 + 72I^{10}$ $-24I^{11} + 3I^{12}$		
J-polynomial	$8J^3 - 12J^5 - 16J^6 + 24J^7 + 18J^8 - 24J^9 - 6J^{10}$ $+12J^{11} - 3J^{12}$		

Common diagram	Stopping sets for Problem 12.3A	Pr[WT $\geq n + 1$] under Polya ($c = 1$) Sampling
	(1, 2, 9) (1, 4, 12) (2, 3, 10) (3, 4, 11) (5, 6, 9) (5, 8, 12) (6, 7, 10) (7, 8, 11) 8 stopping sets	$P_A = \sum_{\alpha=4}^{11} b_\alpha \binom{11}{\alpha} / \binom{n+11}{\alpha} + 3\delta_{on}$, $P_B = \sum_{\alpha=3}^{11} a_\alpha \binom{11}{\alpha} / \binom{n+11}{\alpha} - 3\delta_{on}$.

12.3 Triangular Network with Common Failure Probability q for Each Unit

Problem 12.4: Consider the given diagram with 6 nodes representing independent units.



The system fails if any connected (by the bonds shown) subset of size $s = 3$ has only failures. The eight possible stopping sets of size $s = 3$ are (1, 2, 3), (1, 2, 4), (1, 2, 5), (1, 3, 5), (1, 3, 6), (2, 3, 5), (2, 4, 5) and (3, 5, 6). The principal interest is for H-sampling, but if we do M or P-sampling, then we stop if any of these triples has had at least one failure at each of its three nodes. Equal probability of failure and independence of the units are still assumed; we simply assume that the node numbers are marked on balls put into an urn and under P-sampling there can be more than one ball with the same number.

Unions of these 8 triples j at-a-time ($j = 1, 2, \ldots, 8$) gives rise to the Prelude to Inclusion-Exclusion Table on the next page, where the last row is obtained by alternately adding and subtracting the items in each column. This gives our I-polynomial, namely,

$$P_1(I) = 8I^3 - 11I^4 + 4I^5. \tag{12.1}$$

If we replace I^α by $(1-J)^\alpha$ ($\alpha = 3, 4, 5$) and take a complement, we obtain the J-polynomial, namely,

$$P_2(J) = 2J^2 + 4J^3 - 9J^4 + 4J^5. \tag{12.2}$$

In both (12.1) and (12.2), the sum of the coefficients is unity. $P_1(I)$ represents the probability of stopping in at most n observations, if the sampling parameters are properly appended to it; $P_2(J)$ represents the complement, i.e., $\Pr[\mathrm{WT} \geq n+1]$ with the sampling parameters put in. We only use the J-polynomial at present. Each J^α on the right side of (12.2) is under H-sampling a standard probability that if we start with $N = 6$ balls (one at each node) and take a sample of size n (without replacement and at random, of course) we will miss α specified balls, which is $C(N-\alpha, n)/C(N, n)$, where C denotes the usual

Prelude

Size of the union					
# at-a-time	3	4	5	6	Total Frequency
1	8	0	0	0	$\binom{8}{1} = 8$
2	0	16	10	2	$\binom{8}{2} = 28$
3	0	6	32	18	$\binom{8}{3} = 56$
4	0	1	28	41	$\binom{8}{4} = 70$
5	0	0	12	44	$\binom{8}{5} = 56$
6	0	0	2	26	$\binom{8}{6} = 28$
7	0	0	0	8	$\binom{8}{7} = 8$
8	0	0	0	1	$\binom{8}{8} = 1$
	8	−11	4	0	Sum of Coeff's = 1

binomial coefficient, or can be obtained by an algorithm developed in Sobel, Uppuluri and Frankowski (1977) where the notation used is $HJ_{1,N}^{(\alpha)}(1,n)$. For M-sampling we can write the answer as a multinomial with common probability $1/N$ for each unit or we can use an algorithm developed in Sobel, Uppuluri and Frankowski (1985), where the notation used is $J_{1,N}^{(\alpha)}(1,n)$. For P-sampling, the same probability is $C(N-\alpha+n-1,n)/C(N+n-1,n)$ and we can use the notation $PJ_{1,N}^{(\alpha)}(1,n)$. Since all three represent $\Pr[\mathrm{WT} \geq n+1]$ and all three can be summed either directly or by the results in the references cited, we can obtain $E[\mathrm{WT}]$. We also can use the variance formula for non-negative random variables

$$\begin{aligned} E[(\mathrm{WT})^2] &= \sum_{n=0}^{\infty}(2n+1)\Pr[\mathrm{WT} \geq n+1]; \\ \sigma^2(\mathrm{WT}) &= E[\mathrm{WT}]^2 - [E\{\mathrm{WT}\}]^2. \end{aligned} \qquad (12.3)$$

The results in tabular form are as follows:

Table 12.4A:

Problem 12.4	One-at-a-time sampling method used		
	W/O replacement (H)	W/replacement (M)	Polya ($c = 1$) (P)
E[WT]	3.73333	5.300000	8.000000
Mode(WT), Pr{Mode}	4, 0.466667	4, 0.240741	4, 0.150794
Min(WT), Max(WT)	3, 5	3, ∞	3, ∞
$\sigma^2(WT)$	0.462222	5.630000	∞

For the distribution of WT and in particular the Mode (WT), in Table 12.4A, we use the three formulas for $n \geq 0$:

$$\begin{aligned} P_H[\mathrm{WT} \geq n+1] &= 2\,\frac{\binom{4}{n}}{\binom{6}{n}} + 4\,\frac{\binom{3}{n}}{\binom{6}{n}} - 9\,\frac{\binom{2}{n}}{\binom{6}{n}} + 4\,\frac{\binom{1}{n}}{\binom{6}{n}} \\ &= \begin{cases} \left(1, 1, 1, \frac{3}{5}, \frac{2}{15}\right) & \text{for } n = 0, 1, 2, 3, 4 \\ 0 & \text{for } n \geq 5, \end{cases} \end{aligned} \tag{12.4}$$

$$\begin{aligned} P_M[\mathrm{WT} \geq n+1] &= 2J^{(2)}_{1/6}(1, n) + 4J^{(3)}_{1/6}(1, n) - 9J^{(4)}_{1/6}(1, n) + 4J^{(5)}_{1/6}(1, n) \\ &= 2\left(\frac{2}{3}\right)^n + 4\left(\frac{1}{2}\right)^n - 9\left(\frac{1}{3}\right)^n + 4\left(\frac{1}{6}\right)^n, \end{aligned} \tag{12.5}$$

$$P_P[\mathrm{WT} \geq n+1] = \frac{2\binom{5}{2}}{\binom{n+5}{2}} + \frac{4\binom{5}{3}}{\binom{n+5}{3}} - \frac{9\binom{5}{4}}{\binom{n+5}{4}} + \frac{4\binom{5}{5}}{\binom{n+5}{5}}\,. \tag{12.6}$$

Note that the answers in Table 12.4A and in Eqs. (12.4), (12.5) and (12.6) are all obtained from the one J-polynomial in (12.2). If we use q for the common probability of any unit failure to conform with the notations used in Boehme, Kossow and Preuss (1992), then we can also make use of the I polynomial in (12.1). Thus, if we replace I^α by q^α ($\alpha = 3, 4, 5$) and take a complement, we can write the reliability function for the system that fails when any 3 connected units fail as

$$R(q) = 1 - 8q^3 + 11q^4 - 4q^5; \tag{12.7}$$

this concept makes more sense under H-sampling, i.e., without replacement, and we will assume H-sampling when using it. Here, q can be obtained from a common distribution of unit life-time by inserting some fixed time point t_0. The subscript 1/6 on each J in (12.5) corresponds to the q in (12.7). Although this is often denoted by p as in Sobel, Uppuluri and Frankowski (1985), we are trying to conform with the notations in Boehme, Kossow and Preuss (1992) by using q in the reliability application.

Next, we wish to show with the same example how the I- and J-polynomial method also "works" for unequal unit failure probabilities. We again start with the sampling aspect and sampling any unit will correspond to a failure for that unit.

Suppose that in the same triangular network as above there are two different sampling (or failure) probabilities: q for nodes 1, 2 and 3 and Q for nodes 4, 5 and 6. This presents little or no difficulty for M-sampling; we merely assume that $3q + 3Q = 1$. In order to continue to apply simple random sampling from an urn, we assume there are r units at each of nodes 1, 2 and 3 and s units at each of nodes 4, 5 and 6. At each node, the units are in parallel so that for node 1 (respectively, 2, 3, 4, 5, 6) to fail we need (at least) r (respectively, r, r, s, s, s) failures, regardless of whether the sampling is H or M or P. In general, we could have B_1 original units at each of nodes 1, 2 and 3 and B_2 original units at each of nodes 4, 5 and 6 for the hypergeometric (H) model. For system failure, we need 3 connected nodes to fail as before. Note that q and Q are now probabilities of node failures, not unit failures and the two types of nodes will be referred to as Type 1 (nodes 1, 2, 3) and Type 2 (nodes 4, 5, 6). For H-sampling, we consider a special case with $B_1 = 1$ and $B_2 = 2$ (call it H_0) so that $N = 3\ B_1 + 3B_2 = 9$. The prelude table based on the same 8 triples as above now has the same 8 rows and for $r = 1$ and $s = 2$ it has 9 columns with column headings (q^3, q^2Q, qQ^2, q^3Q, q^2Q^2, qQ^3, q^3Q^2, q^2Q^3, q^3Q^3). Using exactly the same method as before, we obtain an "I-polynomial" with powers replaced by 2-vectors, namely

$$P_1(I) = I^{(3,0)} + 5I^{(2,1)} + 2I^{(1,2)} - 5I^{(3,1)} - 6I^{(2,2)} + 4I^{(3,2)}, \qquad (12.8)$$

the remaining 3 terms having cancelled by the plus and minus addition. These I-functions, written out for general r and s, take the Dirichlet form

$$\begin{aligned} I^{(3,0)} &= I_{q,Q}^{(3,0)}(r,s,n) = I_q^{(3)}(r,n), \\ I^{(2,1)} &= I_{q,Q}^{(2,1)}(r,s,n), \\ I^{(1,2)} &= I_{q,Q}^{(1,2)}(r,s,n), \text{ etc.} \end{aligned} \qquad (12.9)$$

and are the same multinomial functions studied in Sobel, Uppuluri and Frankowski (1985) for which algorithms can be made available. If the nodes are operating independently and $r = 1$, $s = 2$, then the coefficients of (12.8) give us the system reliability $R(q,Q)$ for the H_0-sampling model in terms of q and Q, namely,

$$R(q, Q \mid H_0) = 1 - [q^3 + 5q^2Q + 2qQ^2 - 5q^3Q - 6q^2Q^2 + 4q^3Q^2]. \qquad (12.10)$$

Since in our new H_0-model we altered the original set-up by setting $B_2 = 2$ (i.e., by putting 2 items in parallel at each of nodes 4, 5 and 6), it is no longer clear which of several multinomial models is comparable with H_0. We consider three different possibilities, M_1, M_2 and M_3. Under M_1, we have $N = 9$ units (1 marked α for $\alpha = 1, 2, 3$ and 2 marked α for $\alpha = 4, 5, 6$), $r = 1$, $s = 2$ and $q = Q = 1/9$. Under M_2, we have $N = 6$ units with $r = 1$, $s = 2$, $q = 1/9$ and $Q = 2/9$. Under M_3, we have $N = 6$ units with $r = s = 1$, $q = 1/9$, and $Q = 2/9$. The aim in doing this is to show that the derivation of a single *J-polynomial*, which we now derive, enables us to get formulas and numerical results for all 4 models, H_0, M_1, M_2 and M_3.

To obtain the *J-polynomial* corresponding to (12.8), which is based on our rule for system failure, we use inclusion-exclusion to change each I in (12.8) into J-form; the first 3 terms (without coefficients) are transformed as follows:

$$\begin{aligned}
I^{(3,0)} &= I^3 = (1-J)^3 = 1 - 3J^{(1,0)} + 3J^{(2,0)} - J^{(3,0)}, \\
I^{(2,1)} &= 1 - 2J^{(1,0)} - J^{(0,1)} + J^{(2,0)} + 2J^{(1,1)} - J^{(2,1)}, \\
I^{(1,2)} &= I - J^{(1,0)} - 2J^{(0,1)} + J^{(0,2)} + 2J^{(1,1)} - J^{(1,2)}, \text{ etc.} \qquad (12.11)
\end{aligned}$$

Then, using the coefficients in (12.8) to linearly combine these J-expressions and taking the complement of the result gives our *J-polynomial*, namely

$$P_2(J) = J^{(2,0)} + J^{(1,1)} + 2J^{(2,1)} + 2J^{(1,2)} - 3J^{(3,1)} - 6J^{(2,2)} + 4J^{(3,2)}, \qquad (12.12)$$

the remaining terms having cancelled. We use (12.12) to obtain Pr[WT $\geq n+1$ | Model] for all 4 models: H_0, M_1, M_2, M_3 after inserting the appropriate values of r, s; for the H-model, we also need B_1, B_2, N and for the M-models we also need q and Q. $J^{(b,c)}$, under H_0, is the probability that in a sample of size n we get less than one (or zero) from b specified cells and less than 2 from a disjoint set of c specified cells. Hence, we obtain for $n \geq 0$ under H_0

$$\Pr[\mathrm{WT} \geq n+1] = \left\{ \begin{array}{l} \binom{7}{n} + \left[\binom{6}{n} + 2\binom{6}{n-1}\right] + 2\left[\binom{5}{n} + 2\binom{5}{n-1}\right] \\ +2\left[\binom{4}{n} + 4\binom{4}{n-1} + 4\binom{4}{n-2}\right] - 3\left[\binom{4}{n} + 2\binom{4}{n-1}\right] \\ -6\left[\binom{3}{n} + 4\binom{3}{n-1} + 4\binom{3}{n-2}\right] + 4\left[\binom{2}{n} + 4\binom{2}{n-1} + 4\binom{2}{n-2}\right] \end{array} \right\} \Big/ \binom{9}{n}. \qquad (12.13)$$

The I and J notation is the same as that used in Sobel, Uppuluri and Frankowski (1977, 1985), except that, in the latter, HI and HJ are used for the H-models and, in the former, I and J are used for the M-models; since we only wrote superscripts in (12.8) and (12.12), we were able to use I and J for both H and

M models. For M_1, we first write the answer in terms of general q showing the coefficients $(1, 1, 2, 2, -3, -6, 4)$ from (12.12) and then insert $q = 1/9$ for the final result in terms of n $(n = 0, 1, \ldots)$, obtaining

$$\begin{aligned}
&\Pr[\mathrm{WT} \geq n+1 \mid M_1] \\
&= (1-2q)^n + \Big[(1-2q)^n + nq(1-2q)^{n-1}\Big] + 2\Big[(1-3q)^n + nq(1-3q)^{n-1}\Big] \\
&\quad + 2\Big[(1-3q)^n + 2nq(1-3q)^{n-1} + n(n-1)q^2(1-3q)^{n-2}\Big] \\
&\quad - 3\Big[(1-4q)^n + nq(1-4q)^{n-1}\Big] \\
&\quad - 6\Big(1-4q)^n + 2nq(1-4q)^{n-1} + n(n-1)q^2(1-4q)^{n-2}\Big] \\
&\quad + 4\Big[(1-5q)^n + 2nq(1-5q)^{n-1} + n(n-1)q^2(1-5q)^{n-2}\Big] \\
&= 2\left(\frac{7}{9}\right)^n + \frac{n}{9}\left(\frac{7}{9}\right)^{n-1} + 4\left(\frac{6}{9}\right)^n + \frac{2n}{3}\left(\frac{6}{9}\right)^{n-1} + \frac{2n(n-1)}{81}\left(\frac{6}{9}\right)^{n-2} \\
&\quad - 9\left(\frac{5}{9}\right)^n - \frac{5}{3}n\left(\frac{5}{9}\right)^{n-1} - \frac{2n(n-1)}{27}\left(\frac{5}{9}\right)^{n-2} + 4\left(\frac{4}{9}\right)^n \\
&\quad + \frac{8}{9}n\left(\frac{4}{9}\right)^{n-1} + \frac{4n(n-1)}{81}\left(\frac{4}{9}\right)^{n-2}.
\end{aligned} \tag{12.14}$$

Summing on n from 0 to ∞ gives by straightforward algebra

$$E_{M_1}[\mathrm{WT}] = 10.864333. \tag{12.15}$$

If we multiply (12.14) by $2n+1$ before summing, then by similar algebra, we obtain

$$E_{M_1}[(\mathrm{WT})^2] = 141.668717, \qquad \sigma^2_{M_1}(\mathrm{WT}) = 23.634986. \tag{12.16}$$

Using (12.12) again for models M_2 and M_3 we obtain, corresponding to (12.14),

$$\begin{aligned}
&\Pr[\mathrm{WT} \geq n+1 \mid M_2] \\
&= \left(\frac{7}{9}\right)^n + \left(\frac{6}{9}\right)^n + \frac{2n}{9}\left(\frac{6}{9}\right)^{n-1} + 2\left(\frac{5}{9}\right)^n + \frac{4n}{9}\left(\frac{5}{9}\right)^{n-1} \\
&\quad + 2\left(\frac{4}{9}\right)^n + 4n\left(\frac{2}{9}\right)\left(\frac{4}{9}\right)^{n-1} \\
&\quad + 2n(n-1)\left(\frac{2}{9}\right)^2\left(\frac{4}{9}\right)^{n-2} - 3\left(\frac{4}{9}\right)^n - 3n\left(\frac{2}{9}\right)\left(\frac{4}{9}\right)^{n-1}
\end{aligned}$$

$$
\begin{aligned}
&-6\left(\frac{3}{9}\right)^n - 12n\left(\frac{2}{9}\right)\left(\frac{3}{9}\right)^{n-1} - 6n(n-1)\left(\frac{2}{9}\right)^2\left(\frac{3}{9}\right)^{n-2} \\
&+4\left(\frac{2}{9}\right)^n + 8n\left(\frac{2}{9}\right)\left(\frac{2}{9}\right)^{n-1} + 4n(n-1)\left(\frac{2}{9}\right)^2\left(\frac{2}{9}\right)^{n-2}, \quad (12.17)
\end{aligned}
$$

$$
\begin{aligned}
\Pr[\mathrm{WT} \geq n+1 \mid M_3] &= \left(\frac{7}{9}\right)^n + \left(\frac{6}{9}\right)^n + 2\left(\frac{5}{9}\right)^n + 2\left(\frac{4}{9}\right)^n \\
&\quad - 3\left(\frac{4}{9}\right)^n - 6\left(\frac{3}{9}\right)^n + 4\left(\frac{2}{9}\right)^n. \quad (12.18)
\end{aligned}
$$

We present the numerical results for all 4 models in tabular form as follows.

Table 12.4B: Triangular array problem with unequal parameters, 4 models One-at-a-time sampling method used

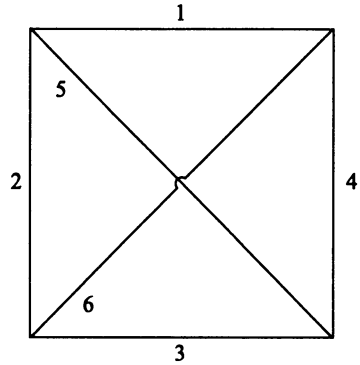
	W/O replacement (H_0)	W/replacement (M_1)	W/replacement (M_2)	W/replacement (M_3)
	$N=9,\ r=1,$ $s=2,$ $B_1=1,\ B_2=2$	$N=9,\ r=1,$ $s=2,$ $q=Q=1/9$	$N=6,\ r=1,$ $s=2,$ $q=1/9,$ $Q=2/9$	$N=6,$ $r=s=1,$ $q=1/9,$ $Q=2/9$
E[WT]	6.007937	10.864333	8.243288	6.342857
Mode(WT), Pr{Mode}	6, 0.325397	7, 0.099028	6, 0.158569	4, 0.192958
Min(WT), Max(WT)	3, 8	3, ∞	3, ∞	3, ∞
σ^2(WT)	1.436439	23.634986	11.766017	11.920339

Note that the value of N is included for the 3 multinomial models to explain the contents of the urn being sampled. Note also for multinomial sampling that if we had 9 equiprobable items with $r = s = 1$, with 2 marked α for each of nodes α =4, 5 and 6 and one for each of the other 3 nodes, then the result would be the same as for Model M_3 in the last column of the table. In contrast to these three M-sampling solutions, the original M-sampling solution of Problem 12.4 had $N = 6$, $r = s = 1$ and $q = Q = 1/9$.

12.4 Duality Levels in a Square with Diagonals That Do Not Intersect: Problem 12.5

This square has $V = 4$ vertices and $E = 6$ edges associated with it as shown in the following diagram. Define the set $S(v_1)$ associated with vertex v_1 as (1, 2,

5) and similarly (2, 3, 6), (3, 4, 5) and (1, 4, 6) represent the sets $S(v_\alpha)$, $\alpha = 2, 3, 4$, respectively. Let d denote the level of duality, so that $d = 1,\ 2$, etc. We define the problem duals A_d and B_d for this set-up.



Problem 12.5A$_d$: Sample the 6 edges until you get d complete sets from the sets $S(v_\alpha)$ $(\alpha = 1, 2, 3, 4)$.

Problem 12.5B$_d$: Sample the 6 edges until you get at least 1 edge for each vertex in any set of $V + 1 - d = 5 - d$ different vertices.

Thus, for level $d = 1$, the stopping sets for Problem 12.5A$_1$ are the 4 triples mentioned above. Note that for Problem 12.5B$_1$, *any set* becomes *the set* and we have to add the pair (5, 6) to these same 4 triples for the stopping set; however, we don't use this since we obtain results for Problem 12.5B$_1$ by duality. It is also interesting to note that for $d = 2$ under H-sampling the expectation is 5 for Problem 12.5A$_2$ and 2 for Problem 12.5B$_2$ without any variance (i.e., $\sigma_H^2(\mathrm{WT}) = 0$), but the duality still holds, namely,

$$E_H[\mathrm{WT} \mid A_2] + E_H[\mathrm{WT} \mid B_2] = E + 1 \quad \text{and} \quad \sigma_H^2(\mathrm{WT} \mid A_2) = \sigma_H^2(\mathrm{WT} \mid B_2).$$

For level $d = 1$, the I- and J-polynomials are easily shown to be

$$P_1(I) = 4I^3 - 6I^5 + 3I^6; \quad P_2(J) = 3J^2 + 4J^3 - 15J^4 + 12J^5 - 3J^6, \tag{12.19}$$

$$P_H[\mathrm{WT} \geq n+1] = \left[3\binom{4}{n} + 4\binom{3}{n} - 15\binom{2}{n} + 12\binom{1}{n} - 3\delta_{0n}\right] \Big/ \binom{6}{n}, \tag{12.20}$$

$$P_M[\mathrm{WT} \geq n+1] = 3\left(\frac{4}{6}\right)^n + 4\left(\frac{3}{6}\right)^n - 15\left(\frac{2}{6}\right)^n + 12\left(\frac{1}{6}\right)^n - 3\delta_{0n}, \tag{12.21}$$

$$P_{P(c=1)}[\mathrm{WT} \geq n+1] = \frac{3\binom{5}{2}}{\binom{n+5}{2}} + \frac{4\binom{5}{3}}{\binom{n+5}{3}} - \frac{15\binom{5}{4}}{\binom{n+5}{4}} + \frac{12\binom{5}{5}}{\binom{n+5}{5}} - 3\delta_{0n}. \tag{12.22}$$

For Problem 12.5B$_1$, the results are the same except that the coefficients are taken from the I-polynomial instead of the J-polynomial in (12.19). In tabular form, the numerical results for Problems 12.5A$_1$ and 12.5B$_1$ are as follows.

Table 12.5: Sampling the edges of a square with diagonals that do not intersect; One-at-a-time sampling method used

Problem 12.5A$_1$: Stop with any one of the 4 vertex sets $S(v_\alpha)$ (α = 1, 2, 3, 4)	W/O replacement (H_0)	W/replacement (M_1)	Polya W/$c = 1$ (P)
E[WT]	4.000000	5.900000	12.000000
Mode(WT), Pr{Mode}	4, 0.600000	4, 0.222222	(3,4)*, 0.119048
Min(WT), Max(WT)	3, 5	3, ∞	3, ∞
σ^2(WT)	0.400000	6.350000	∞ **
I-polynomial:	$4I^3 - 6I^5 + 3I^6$		
J-polynomial:	$3J^2 + 4J^3 - 15J^4 + 12J^5 - 3J^6$		
Problem 12.5B$_1$: Stop with at least 1 edge associated with each of the 4 vertices	(For J-polynomial, use $4J^3 - 6J^5 + 3J^6$)		
E[WT]	3.000000	3.800000	5.500000
Mode(WT), Pr{Mode}	3, 0.600000	3, 0.361111	3, 0.250000
Min(WT), Max(WT)	2, 4	2, ∞	2, ∞
σ^2(WT)	0.400000	2.480000	11.250000 **
Sum of Expectations A,B	$E + 1 = 7$	–	–

* double mode

** The power 2 on J leads to an infinite variance, but not the power 3.

Some other problem dualities are stated but not solved here; recall that this duality is only for H-sampling. Consider a regular polyhedron (RP) with F faces.

Problem 12.6A: Sample the faces of RP until you see f different faces, each at least once.

Problem 12.6B: Sample the faces of RP until you see $F+1-f$ different faces, each at least once.

Here, the sum of the expectations of WT is $F+1$ and the variances are equal. Suppose that the RP has f_v faces associated with each vertex. Then another pair of dual problems is

Problem 12.7A: Sample the faces until you get a complete set of f_v faces for any one vertex.

Problem 12.7B: Sample the faces until you get at least 1 face associated with each vertex.

Here, the sum of the expectations of WT is again $F+1$ and the variances are equal. Suppose the RP has E edges and F faces. Can we generalize Problem 12.2 to the following dual pair?

Problem 12.8A: Sample the E edges of RP until you get f completed faces.

Problem 12.8B: Sample the E edges of RP until you get at least 1 edge for $F+1-f$ different faces.

Problem 12.9: Random solo tic-tac-toe with one or two players.

The method of I- and/or J-polynomials has a broad application. Suppose a single player wishes to randomly put X symbols on a square tic-tac-toe board of size $s \times s$ ($s = 3, 4, 5$). We number the squares in a systematic manner (for convenience only) and write the same s^2 integers on otherwise indistinguishable balls in an urn. Balls are then drawn one-at-a-time using H- or M- or $P(c=1)$-sampling. We are interested in the expectation and variance of the WT until a row, column or diagonal of X's is obtained. The following table of numerical results is based on the J-polynomial which is derived in the usual manner. Thus, for any s there are $2s+2$ stopping s-triples and for $s = 3$ with systematic numbering, these are (1, 2, 3), (4, 5, 6), (7, 8, 9), (1, 4, 7), (2, 5, 8), (3, 6, 9), (1, 5, 9) and (3, 5, 7).

The "Prelude Table" for $s = 3$ is on the next page. From this table, the I- and J-polynomials are easy to obtain and they are given in Table 12.6. Table 12.6 also contains results for two-player problems, provided each player plays on his own tic-tac-toe board. Then, the goal is to find the sooner of the two waiting times under independence.

The method of I- and J-polynomials has also been applied to some more challenging finite lattice problems and gives exact answers to path problems and percolation-type problems. The only limitation is that the process of forming unions has to be accurate and once this is put on the computer, many more and different problems can be solved exactly. The tables that follow indicate at least 4 different problems (with variations on each one) and give the J-polynomial for each problem.

Distribution of sizes of the unions for all possible combinations

# at-a-time	3	4	5	6	7	8	9	Total Row Frequency
1	8	0	0	0	0	0	0	$\binom{8}{1} = 8$
2	0	0	22	6	0	0	0	$\binom{8}{2} = 28$
3	0	0	0	16	38	0	2	$\binom{8}{3} = 56$
4	0	0	0	0	22	37	11	$\binom{8}{4} = 70$
5	0	0	0	0	2	24	30	$\binom{8}{5} = 56$
6	0	0	0	0	0	4	24	$\binom{8}{6} = 28$
7	0	0	0	0	0	0	8	$\binom{8}{7} = 8$
8	0	0	0	0	0	0	1	$\binom{8}{8} = 1$

Problem 12.10: Boundary-connection problems.

For a finite rectangular lattice of size ($m + 1$ by $n + 1$), the total number $2mn + m + n$ of unit segments consists of $2(m + n)$ boundary segments and $N = 2mn - (m+n)$ internal segments. The former are "free" and the latter are numbered 1 through N and balls with these same numbers are drawn one-at-a-time from the urn with H- or M- or $P(c = 1)$-sampling. Any internal segment can be used (to form a path) only after its number is observed on a ball taken from the urn. The stopping rule is to continue sampling until we have obtained for each of the $(m-1)(n-1)$ internal vertices at least one path to any boundary point; the paths need not be disjoint. This has been done for several different sizes of rectangular lattices. We do depend on a clearly defined set of vertices each of which is one unit away from the boundary but do not require that the lattice be rectangular in shape.

Problem 12.11: Center-connection problems.

Using the same set-up as in Problem 12.10, if m and n are both odd, then the lattice has a center and stop sampling as soon as we have obtained a path from the center point to any boundary point.

Table 12.6: Expected waiting time for random solo tic-tac-toe: Problem 12.9

Square Lattice Size	Sampling Method	E[WT]	σ^2(WT)	1-I and/or J-polynomial $= \Pr[\text{WT} \geq n+1]$
One-player problems				
3×3	w/o repl. (H)	4.769841	1.107352	$1 - [8I^3 - 22I^5 + 10I^6 + 18I^7 - 17I^8$
	w/repl. (M)	6.45357	6.40328	$+4I^9] = 2J^3 + 16J^4 - 32J^5 + 4J^6$
	Polya (P)	10.04444	77.89010	$+26J^7 - 19J^8 + 4J^9$
4×4	H	9.00318	2.88891	$1 - [10I^4 - 32I^7 - 13I^8 + 24I^9 + 88I^{10}$
	M	13.02964	21.04164	$-56I^{11} - 78I^{12} + 48I^{13} + 36I^{14} - 32I^{15}$
	P	21.49950	219.00857	$+6I^{16}] = 10J^4 + 112J^5 - 328J^6 - 328J^7$
				$+2503J^8 - 4464J^9 + 3960J^{10} - 1616J^{11}$
				$-182J^{12} + 552J^{13} - 276J^{14}$
				$+64J^{15} - 6J^{16}$
5×5	H	14.89717	6.45994	$48J^5 + 1216J^6 - 5136J^7 - 7846J^8 + 96261J^9$
	M	22.64750	74.14346	$-282980J^{10} + 409240J^{11} - 176660J^{12}$
	P	39.34026	544.44663	$-508124J^{13} + 1291728J^{14} - 1652824J^{15}$
				$+1412693J^{16} - 848996J^{17} + 347932J^{18}$
				$-80248J^{19} - 3700J^{20} + 11036J^{21} - 4528J^{22}$
				$+1008J^{23} - 126J^{24} + 7J^{25}$
Two-player problems (Each on their separate boards, playing alternately and waiting for the first winner)				
3×3	H	4.16786	0.77133	Square the numerical results from
	M	5.11576	2.39427	above for 3×3 to get
	P	6.51813	10.07817	$\Pr[\text{WT} \geq n+1]$
4×4	H	8.05527	2.29823	Square the numerical results from
	M	10.51772	9.70097	above for 4×4 to get
	P	14.66697	46.24823	$\Pr[\text{WT} \geq n+1]$
5×5	H	13.47674	5.40445	Square the numerical results from
	M	18.66348	26.49544	above for 5×5 to get
	P	27.86007	145.67330	$\Pr[\text{WT} \geq n+1]$

Problem 12.12: Top to bottom-connection problems.

In this case, only the top and bottom boundaries are free; the rest are all put into the urn. We stop sampling as soon as we have obtained a path from any point on the top boundary to any point on the bottom boundary.

Problem 12.13: Corner to corner-connection problems.

In this case, there are no free boundaries; all segments go into the urn. We stop sampling as soon as we have obtained a path from the lower left corner to the upper right corner of the rectangular lattice.

For each of Problems 12.10—12.13, in addition to numerical results and the J-polynomial for each problem, percentiles of the WT needed are given corresponding to 5 selected probability levels. These can be regarded as confidence limits that the goal will be reached in the indicated (decimal) number of observations, the decimals being obtained by linear interpolation between two successive integers.

We also consider some modifications of Problems 12.10—12.13. The symbol TS stands for *True Sponge* and indicates that initially you sample only among segments that have one endpoint on the boundary. Subsequently, you can sample among segments that are attached to those already observed. Being a restriction, the TS model answer should be a little larger than the unrestricted answer in many cases, but this may not be true in general. If we allow all internal segments from the outset but are forced to replace those not attached to a boundary point without using them to form a path but only in calculating WT, then we have a clear upper bound and it is denoted by UB in the table.

Acknowledgements: The author wishes to acknowledge with thanks some conversations about percolation-type problems with Tom Fuller and Professor Walter Kohn of the Department of Physics at UCSB. He also wishes to thank Hokwon Cho of the Department of Statistics and Applied Probability at UCSB for providing help with some of the numerical computations.

References

1. Boehme, T. K., Kossow, A. and Preuss, W. (1992). A generalization of consecutive k-out-of-n: F systems, *IEEE Transactions on Reliability*, **41**, 451–457. [See also the more complete reference list at the end of Chapters 5 and 6 in *New Trends in System Reliability Evaluation* (Ed., K. B. Misra), 1993, Elsevier Science Publishers.]

2. Gleser, L. J., Perlman, M. D., Press, S. J. and Sampson, A. R. (1989). *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin*, p. 112, New York: Springer-Verlag.

3. Sobel, M. and Uppuluri, V. R. R. (1974). Sparse and crowded cells and Dirichlet distributions, *Annals of Statistics*, **2**, 977–987.

4. Sobel, M., Uppuluri, V. R. R. and Frankowski, K. (1977). Dirichlet distribution–Type 1, *Selected Tables in Mathematical Statistics—Vol. 4*, Providence, RI: American Mathematical Society.

5. Sobel, M., Uppuluri, V. R. R. and Frankowski, K. (1985). Dirichlet integrals of Type 2 and their applications, *Selected Tables in Mathematical Statistics—Vol. 9*, Providence, RI: American Mathematical Society.

Tables 12.7A, 12.8A, 12.9A and 12.10A: Exact (finite lattice) results for hypergeometric and multinomial percolation-type problems (Unrestricted model)

Lattice Size	Sampling Type	$\underline{n}$, $\bar{n}$, N	E{WT}	σ^2(WT)	Mode and its Probability	E{WT} for Two Modified Models, TS and UB, for the same goals.	
Boundary-Connection Problems						TS: True Spong (Inward)	UB: Upper Bound
2 × 2	w/o Repl. (H)		1.00000	0	1,(1.000000)	1.00000	1.00000
	w/ Repl. (M)	1, 1, 4	1.00000	0	1,(1.000000)	1.00000	1.00000
2 × 3	w/o Repl. (H)		2.34286	0.33959	2,(.714286)	2.40000	2.66667
	w/ Repl. (M)	2, 4, 7	2.66667	1.19444	2,(.612245)	2.75000	2.91667
2 × 4	w/o Repl. (H)		3.84286	0.95150	3,(.466667)	3.97143	4.65625
	w/ Repl. (M)	3, 7, 10	4.67460	3.88155	3,(.336000)	4.93750	5.31250
3 × 3	w/o Repl. (H)		5.08918	1.31817	4,(.387879)		
	w/ Repl. (M)	4, 9, 12	6.43420	6.41954	5,(.229167)		
2 × 5	w/o Repl. (H)		5.44328	1.79573	4,(.295882)		6.83458
	w/ Repl. (M)	4, 10, 13	6.91295	8.15306	5,(.204130)		8.00042
3 × 4	w/o Repl. (H)		8.02947	2.79713	7,(.251028)		
	w/ Repl. (M)	6, 14, 17	10.74160	10.98445	8,(.139145)		
4 × 4	w/o Repl. (H)		12.47031	5.42662	11,(.171931)		
	w/ Repl. (M)	9, 21, 24	17.59274	36.84103	14,(.086465)		
3 × 3 × 3	w/o Repl. (H)		13.15964	11.89919	11,(.128178)		
	w/ Repl. (M)	8, 31, 36	16.55286	41.16427	13,(.086163)		
Center-Connection Problem						**TS: Inward**	**TS: Outward**
2 × 4	w/o Repl. (H)		2.61508	1.57485	2,(.311111)	2.61786	1.66667
	w/ Repl. (M)	3, 7, 10	3.00397	3.60678	2,(.280000)	3.20313	1.68750
4 × 4	w/o Repl. (H)		7.96578	9.82154	7,(135792)		
	w/ Repl. (M)	2, 21, 24	9.88069	26.10307	7,(.098105)		
Top to Bottom - Connection Problems						**TS: Downward**	**TS: Downward**
2 × 2	w/o Repl. (H)		3.73571	0.95158	4,(.407143)	3.52698	4.46032
	w/ Repl. (M)	2, 6, 8	4.80000	4.71360	4,(.238769)	4.15697	4.85326
2 × 3	w/o Repl. (H)		4.35584	1.58507	5,(.287446)		
	w/ Repl. (M)	2, 8, 11	5.41151	5.69986	4,(.191517)		
2 × 4	w/o Repl. (H)		4.90643	2.28262	5,(.250750)		
	w/ Repl. (M)	2, 10, 14	5.96115	6.75415	5,(.172268)		
3 × 3	w/o Repl. (H)		8.50039	4.50027	9,(.194826)		
	w/ Repl. (M)	3,15,18	11.51828	22.55859	10,(.103220)		
2 × 5	w/o Repl. (H)		5.40836	3.03378	5,(.208985)		
	w/ Repl. (M)	2, 12, 17	6.46594	7.86899	6,(.151460)		
Corner to Corner - Connection Problems						**TS: One Direction**	**TS: Two Directions**
2 × 2	w/o Repl. (H)		7.93932	2.27028	8,(.255051)		
	w/ Repl. (M)	4, 11, 12	12.79670	30.16251	10,(.091205)		

Tables 12.7B, 12.8B, 12.9B and 12.10B: Percentiles of the waiting time (WT): The confidence level is P^* that the number of observations needed is less than the value shown
(Unrestricted model)

Lattice Size	Sampling Type	P* = .50	P* = .75	P* = .90	P* = .95	P* = .99
Boundary-Connection Problems						
2 × 2	H	1.00000	1.00000	1.00000	1.00000	1.00000
	M	1.00000	1.00000	1.00000	1.00000	1.00000
2 × 3	H	1.70000	2.15625	2.81250	3.12500	3.82500
	M	1.81667	2.60577	3.65270	4.45971	6.33874
2 × 4	H	3.10769	3.91538	4.81250	5.37500	6.30000
	M	3.63863	4.96324	6.76867	8.06921	11.21840
3 × 3	H	4.36275	5.29878	6.30000	6.88729	7.94063
	M	5.27541	6.98649	9.24201	10.90407	14.83740
2 × 5	H	4.70194	5.78069	6.87941	7.61818	8.80312
	M	5.69224	7.67413	10.14035	11.98640	16.36269
3 × 4	H	7.25068	8.58464	9.91049	11.78116	13.25833
	M	9.29569	12.02328	15.46567	17.98360	23.91981
4 × 4	H	11.64363	13.48504	15.30625	16.40300	18.22871
	M	15.75442	19.96957	25.08410	28.81947	37.25682
3 × 3 × 3	H	12.07150	14.61777	17.35385	19.18453	23.15504
	M	14.60759	18.84424	24.09082	28.13257	38.50038
Center-Connection Problem						
2 × 4	H	1.96429	2.87755	3.85965	4.54167	5.72500
	M	2.08929	3.35385	4.91261	6.15270	9.13039
4 × 4	H	7.24814	9.31000	11.43566	12.83264	15.63887
	M	8.46074	11.77240	15.79323	18.88406	26.52547
Top to Bottom - Connection Problems						
2 × 2	H	3.26316	3.87719	4.60870	4.91304	5.72000
	M	3.86708	5.33419	7.06918	8.45505	11.62887
2 × 3	H	3.86170	4.73268	5.56333	5.94833	6.89419
	M	4.52978	6.12508	7.97715	9.39799	12.59344
2 × 4	H	4.40837	5.48607	6.42285	6.92003	7.95970
	M	5.08600	6.86681	8.85227	10.27671	13.51386
3 × 3	H	7.99830	9.37885	10.71273	11.63000	13.52522
	M	10.25040	13.32707	16.94746	19.76802	26.77742
2 × 5	H	4.89870	6.12798	7.18860	7.83618	8.97493
	M	5.60185	7.56581	9.66704	11.09655	14.40471
Corner to Corner - Connection Problem						
2 × 2	H	7.36302	8.56452	9.43902	9.84146	10.67000
	M	11.24066	14.95234	19.38435	22.64875	30.38680

Tables 12.7C, 12.8C, 12.9C and 12.10C: Dirichlet generating functions * (Unrestricted model)

Lattice Size

Boundary-Connection Problems

2 × 2	J^4
2 × 3	$2J^4 + J^6 - 2J^7$
2 × 4	$3J^4 + 2J^6 - 4J^7 - 4J^9 + 4J^{10}$
3 × 3	$4J^4 + 4J^6 - 8J^7 + 3J^8 - 16J^9 + 4J^{10} + 24J^{11} - 14J^{12}$
2 × 5	$4J^4 + 3J^6 - 6J^7 - J^8 - 8J^9 + 7J^{10} - 2J^{11} + 12J^{12} - 8J^{13}$
3 × 4	$6J^4 + 7J^6 - 14J^7 + 4J^8 - 40J^9 + 21J^{10} + 8J^{11} + 7J^{12} + 126J^{13} - 76J^{14} - 270J^{15} + 320J^{16} - 98J^{17}$
4 × 4	$5J^4 + 44J^5 - 176J^6 + 436J^7 - 798J^8 + 964J^9 - 744J^{10} - 332J^{11} + 1682J^{12} - 1616J^{13} + 1112J^{14} - 964J^{15} + 2935J^{16} - 9052J^{17} + 8888J^{18} - 1012J^{19} + 7105J^{20} - 24612J^{21} + 26024J^{22} - 11984J^{23} + 2096J^{24}$
3 × 3 × 3	$80J^6 - 648J^7 + 2588J^8 - 6044J^9 + 9132J^{10} - 9304J^{11} + 6612J^{12} - 3888J^{13} + 2684J^{14} - 2144J^{15} + 1392J^{16} - 972J^{17} + 1268J^{18} - 1608J^{19} + 1818J^{20} - 1224J^{21} - 456J^{22} + 2292J^{23} - 2591J^{24} - 36J^{25} + 4368J^{26} - 8437J^{27} + 7353J^{28} + 6408J^{29} - 32864J^{30} + 57018J^{31} - 61284J^{32} + 47412J^{33} - 28164J^{34} + 11415J^{35} - 2175J^{36}$

Center-Connection Problems

2 × 4	$J^4 + 2J^6 - 2J^7 + J^8 - 2J^9 + J^{10}$
4 × 4	$J^4 + 4J^6 + 4J^7 - 30J^8 + 92J^9 - 130J^{10} + 120J^{11} - 272J^{12} + 508J^{13} - 662J^{14} + 516J^{15} + 440J^{16} - 1352J^{17} + 1290J^{18} - 1844J^{19} + 3527J^{20} - 3936J^{21} + 2370J^{22} - 740J^{23} + 95J^{24}$

Top to Bottom - Connection Problems

2 × 2	$2J^3 + 4J^4 - 2J^5 - 13J^6 + 14J^7 - 4J^8$
2 × 3	$2J^4 + 6J^5 + 2J^6 - 18J^7 - 15J^8 + 52J^9 - 36J^{10} + 8J^{11}$
2 × 4	$2J^5 + 8J^6 + 8J^7 - 20J^8 - 44J^9 + 31J^{10} + 116J^{11} - 172J^{12} + 88J^{13} - 16J^{14}$
3 × 3	$15J^4 - 76J^5 + 270J^6 - 360J^7 - 135J^8 + 882J^9 - 1481J^{10} + 2046J^{11} - 993J^{12} - 3028J^{13} + 6764J^{14} - 6390J^{15} + 3282J^{16} - 898J^{17} + 103J^{18}$
2 × 5	$2J^6 + 10J^7 + 16J^8 - 16J^9 - 80J^{10} - 20J^{11} + 199J^{12} + 86J^{13} - 532J^{14} + 512J^{15} - 208J^{16} + 32J^{17}$

Corner to Corner - Connection Problems

2 × 2	$2J^2 + 14J^3 - 31J^4 + 2J^5 + 20J^6 + 50J^7 - 145J^8 + 150J^9 - 82J^{10} + 24J^{11} - 3J^{12}$

*For each goal (and each lattice) one polynomial in J gives rise to all the answers for both types of sampling, H and M.

Table 12.11: Expected waiting time for selected card and die combinations

(The original is the usual deck of 52 cards)	One-at-a-time Sampling Procedure Used		
	W/O Replacement (H)	W/Replacement (M)	Polya W/c = 1 (P)
1. Any black card	53/27 = 1.962963	2.000000	51/25 = 2.040000
2. An Ace or a Spade	53/17 = 3.117647	3.250000	51/15 = 3.400000
3. Any Spade	53/14 = 3.785714	4.000000	51/12 = 4.250000
4. Any Picture Card	53/13 = 4.076923	4.333333	51/11 = 4.636364
5. An Ace or a King	53/9 = 5.888889	6.500000	51/7 = 7.285714
6. Any Ace	53/5 = 10.600000	13.000000	51/3 = 17.000000
7. Either One of the Two Black Aces	53/3 = 17.666667	26.000000	51/1 = 51.000000
8. The Ace of Spades (or any specified card)	53/2 = 26.500000	52.000000	Finite or Infinite??
9. A Black Ace and a Red Ace	371/15 = 24.733333	39.000000	102.000000
10. Any Two Aces			
(i) different suits	106/5 = 21.200000	30.333333	39.666667
(ii) same suit	---	41.843751	105.764286
(iii) same or different	21.2000000	26.000000	29.750000
11. Four Aces			
(i) different suits	42.400000	108.333333	Finite or Infinite??
(ii) same suit		113.497039	1311.441274
(iii) same or different	42.40000	52.000000	98.175000
12. Fifty one different cards	51.000000	183.978285	2601.000000
(The six numbers (1,2,...,6) from a die are put on balls in an urn at the outset; then M-sampling is the usual tossing of a die).			
13. Three different Numbers	3.000000	3.700000	20/3 = 666667
14. Four different Numbers	4.000000	5.700000	236/21 = 11.238095
15. Five different Numbers	5.000000	8.700000	66.302438
16. All Six different Numbers	6.000000	14.700000	Finite or Infinite??
17. Six same Numbers	---	19.737384	Finite or Infinite?*
18. Seven same Numbers	---	24.224483	Finite or Infinite?*
*Hard Computations; Need the distribution of Max Cell Frequency for Polya Sampling.			

13

Stirling Numbers and Records

N. Balakrishnan and V. B. Nevzorov

McMaster University, Hamilton, Ontario, Canada
St.-Petersburg State University, St.-Petersburg, Russia

Abstract: Stirling numbers and generalized Stirling numbers and their properties are briefly described first. Then some relationships between Stirling numbers and record times are presented. Finally, we show that generalized Stirling numbers of the first kind describe distributions of some record statistics in the so-called F^{α}-scheme .

Keywords and phrases: Stirling numbers of the first kind, Stirling numbers of the second kind, Generalized Stirling numbers, Records, Record times, F^{α}-scheme

13.1 Stirling Numbers

In this section, we first present definitions and some basic properties of Stirling numbers of the first and second kinds; for more details, see Goldberg, Neiman and Heinsvort (1964), and the excellent review article by Charalambides and Singh (1988) concerning Stirling numbers and their generalizations.

Stirling numbers of the first kind $s(n,k)$ and Stirling numbers of the second kind $S(n,k)$ are given by the following equalities:

$$x(x-1)\dots(x-n+1)=\sum_{k=0}^{n} s(n,k)x^k \tag{13.1}$$

and

$$x^n=\sum_{k=0}^{n} S(n,k)x(x-1)\dots(x-k+1), \qquad n=0,1,2,\dots. \tag{13.2}$$

It is easy to show that

$$S(n,k)=\sum_{\ell=0}^{k}(-1)^{k-\ell}\binom{k}{\ell}\frac{\ell^n}{k!}=\sum_{\ell=0}^{k}(-1)^{k-\ell}\frac{\ell^n}{\ell!(k-\ell)!}$$

and

$$s(n,k) = \sum_{\ell=0}^{n-k} (-1)^{\ell} \binom{n-1+\ell}{n-k+\ell} \binom{2n-k}{n-k-\ell} S(n-k+\ell,\ell).$$

In particular, we have

$$\begin{aligned} S(n,1) &= 1, \quad S(n,n) = 1, \quad S(n,n-1) = \binom{n}{2}, \\ s(n,1) &= (-1)^{n-1}(n-1)!, \ s(n,n) = 1, \ s(n,n-1) = -\binom{n}{2}. \end{aligned}$$

Furthermore, we have the generating functions of these numbers to be

$$\{log(1+x)\}^k/k! = \sum_{n=k}^{\infty} s(n,k)x^n/n!, \qquad |x| < 1, \tag{13.3}$$

$$(e^x - 1)^k/k! = \sum_{n=k}^{\infty} S(n,k)x^n/n!, \tag{13.4}$$

and

$$\frac{1}{(1-x)(1-2x)\dots(1-kx)} = \sum_{n=k}^{\infty} S(n,k)x^{n-k}, \quad |x| < 1/k. \tag{13.5}$$

From (13.1) and (13.2), we also readily have the recurrence relations

$$\begin{aligned} s(n+1,k) &= s(n,k-1) - n\ s(n,k), \ n \geq k \geq 1, \\ S(n+1,k) &= k\ S(n,k) + S(n,k-1), \ n \geq k \geq 1. \end{aligned}$$

Finally, we have the following asymptotic (as $n \to \infty$) expressions:

$$|s(n,k)| \sim (n-1)!\{log\ n\}^{k-1}/(k-1)! \ and \ S(n,k) \sim k^n/n!.$$

13.2 Generalized Stirling Numbers

Let $\mathbf{a}$ be any sequence $(a_0, a_1, \dots)$. Let

$$p_0(x, \mathbf{a}) = 1 \ and \ p_k(x, \mathbf{a}) = (x - a_0)(x - a_1)\dots(x - a_{k-1}), k = 1, 2, \dots.$$

Generalized Stirling numbers of the first kind $s(n, k, \mathbf{a})$ and generalized Stirling numbers of the second kind $S(n, k, \mathbf{a})$ are then defined as follows:

$$p_n(x, \mathbf{a}) = \sum_{k=0}^{n} s(n, k, \mathbf{a}) x^k, \qquad n = 0, 1, \ldots, \tag{13.6}$$

and

$$x^n = \sum_{k=0}^{n} S(n, k, \mathbf{a}) p_k(x, \mathbf{a}). \tag{13.7}$$

For $\mathbf{a} = (0, 1, 2, \ldots)$, definitions in (13.6) and (13.7) coincide with those in (13.1) and (13.2), respectively. It is easy to observe that

$$\begin{aligned} s(n, 0, \mathbf{a}) &= (-1)^n a_0 a_1 \ldots a_{n-1}, \\ s(n, 1, \mathbf{a}) &= (-1)^{n-1} a_0 a_1 \ldots a_{n-1} \sum_{k=0}^{n-1} 1/a_k, \\ s(n, n-1, \mathbf{a}) &= -\sum_{k=0}^{n-1} a_k, \\ s(n, n, \mathbf{a}) &= 1; \end{aligned} \tag{13.8}$$

$$\begin{aligned} S(n, 0, \mathbf{a}) &= a_0^n, \\ S(n, 1, \mathbf{a}) &= (a_1^n - a_0^n)/(a_1 - a_0), \\ S(n, n-1, \mathbf{a}) &= \sum_{k=0}^{n-1} a_k, \\ S(n, n, \mathbf{a}) &= 1. \end{aligned} \tag{13.9}$$

The following recurrence relations are also satisfied by these generalized Stirling numbers:

$$s(n+1, k, \mathbf{a}) = s(n, k-1, \mathbf{a}) - a_n\, s(n, k, \mathbf{a}) \tag{13.10}$$

$$S(n+1, k, \mathbf{a}) = S(n, k-1, \mathbf{a}) + a_k S(n, k, \mathbf{a}). \tag{13.11}$$

For more elaborate details on these generalized Stirling numbers of the first and second kinds, interested readers may refer to the review article by Charalambides and Singh (1988).

Remark 13.2.1 One can see from (13.6) that numbers $s(n, n, \mathbf{a})$, $s(n, n-1, \mathbf{a}), \ldots$ change signs and $(-1)^{n-m} s(n, m, \mathbf{a}) > 0$ if elements a_k of the vector $\mathbf{a}$ are all positive. In this case, the following equality holds:

$$(x + a_0)(x + a_1) \ldots (x + a_{n-1}) = \sum_{k=0}^{n} |s(n, k, \mathbf{a})| x^k \tag{13.12}$$

and, consequently,

$$\sum_{k=0}^{n} |s(n,k,\mathbf{a})| = (1+a_0)(1+a_1)\dots(1+a_{n-1}).$$

For the classical Stirling numbers of the first kind, relation (13.12) can be rewritten as

$$x(x+1)\dots(x+n-1) = \sum_{k=0}^{n} |s(n,k)|x^k.$$

Let us introduce random variables $X(n)$, $n = 1, 2, \dots$, such that

$$\begin{aligned} Pr[X(n) = k] &= \frac{|s(n,k,\mathbf{a})|}{\sum_{k=0}^{n} |s(n,k,\mathbf{a})|} \\ &= \frac{|s(n,k,\mathbf{a})|}{(1+a_0)(1+a_1)\dots(1+a_{n-1})}, \quad k = 0, \dots, n. \end{aligned}$$

It follows from (13.12) that

$$E[s^{X(n)}] = \frac{(s+a_0)(s+a_1)\dots(s+a_{n-1})}{(1+a_0)(1+a_1)\dots(1+a_{n-1})}.$$

Hence, if a_k $(k = 0, 1, 2, \dots)$ are nonnegative, then $X(n)$ can be represented as a sum of n independent random variables $Y_1, Y_2, \dots, Y_n$ each of them taking values 0 and 1 with

$$Pr[Y_k = 1] = 1 - Pr[Y_k = 0] = \frac{1}{1+a_{k-1}}, \qquad k = 1, 2, \dots.$$

As a result, we readily have

$$E[X(n)] = \sum_{k=0}^{n-1} \frac{1}{1+a_k} \quad \text{and} \quad Var(X(n)) = \sum_{k=0}^{n-1} \frac{a_k}{(1+a_k)^2}.$$

Rusinski and Voigt (1990) established that, if $a_0 \geq 0$ and $a_k = a_0 + r$ with r_0 and $a_0 + r > 0$, an analogous result holds for generalized Stirling numbers of the second kind $S(n,k,\mathbf{a})$. In this situation, all the roots $x_{1,n}, x_{2,n}, \dots, x_{n,n}$ of the equation

$$\sum_{k=0}^{n} S(n,k,\mathbf{a})x^k = 0$$

are all real and nonpositive. Therefore, random variables Z_n $(n = 1,2,...)$ such that

$$Pr[Z_n = k] = \frac{S(n,k,\mathbf{a})}{\sum_{k=0}^{n} S(n,k,\mathbf{a})}, k = 0, 1, \dots, n,$$

also can be represented as a sum $V_1 + \cdots + V_n$ of independent random variables taking values 0 and 1 with probabilities

$$Pr[V_k = 1] = 1 - Pr[V_k = 0] = \frac{1}{1 - x_{k,n}}, \qquad k = 1, 2, \ldots, n.$$

13.3 Stirling Numbers and Records

Stirling numbers of the first kind are very closely connected with records. In the recent years, considerable amount of research has been done on records and related statistics; see, for example, Nevzorov (1987), Nagaraja (1988), Arnold and Balakrishnan (1989), Arnold, Balakrishnan and Nagaraja (1992), Ahsanullah (1995), and Arnold, Balakrishnan and Nagaraja (1997).

Let us introduce the concept of records. For a sequence of random variables $X_1, X_2, \ldots$, the record times $L(n)$ and record values $X(n)$ are defined as

$$L(1) = 1, \; L(n+1) = \min\{j : X_j > X_{L(n)}\}, \qquad n = 1, 2, \ldots,$$

and

$$X(n) = X_{L(n)},$$

respectively. Also, let us introduce the record indicators

$$\xi_k = I\{X_n \text{ is a record value}\}$$

for $k = 1, 2, \ldots$, where $I(\cdot)$ is the indicator function. Then, the variable

$$N(n) = \xi_1 + \cdots + \xi_n$$

denotes the number of records amongst the variables $X_1, X_2, \ldots, X_n$, for $n = 1, 2, \ldots$.

Renyi (1962) then obtained the following important result.

Lemma 13.3.1 *If $X_1, X_2, \ldots$ are independent random variables with a common continuous distribution function $F(\cdot)$, the record indicators $\xi_1, \xi_2, \ldots$ are all independent and*

$$Pr[\xi_k = 1] = 1 - Pr[\xi_k = 0] = \frac{1}{k}, \qquad k = 1, 2, \ldots.$$

Since the random variables $N(n)$ and $L(n)$ have the relationships

$$Pr[L(n) > k] = Pr[N(k) < n]$$

and

$$Pr[L(n) = k] = Pr[N(k-1) = n-1, \xi_k = 1],$$

the distributions of record times $L(n)$ can be expressed in terms of independent random variables as follows:

$$Pr[L(n) > k] = Pr[\xi_1 + ... + \xi_k \ < \ n] \tag{13.13}$$

and

$$\begin{aligned} Pr[L(n) = k] &= \frac{1}{k} Pr[N(k-1) = n-1] \\ &= \frac{1}{k} Pr[\xi_1 + ... + \xi_{k-1} = n-1]. \end{aligned} \tag{13.14}$$

Using the independence of record indicators presented in Lemma 13.3.1, Renyi (1962) then derived the probability generating function of the random variable $N(n)$ as

$$E[s^{N(n)}] = s(s+1)\dots(s+n-1)/n!. \tag{13.15}$$

This readily implies that

$$Pr[N(n) = k] = |s(n,k)|/n! \tag{13.16}$$

and, as $n \to \infty$,

$$Pr[N(n) = k] \sim \frac{(n-1)!(log\ n)^{k-1}}{n!(k-1)!} = \frac{(log\ n)^{k-1}}{n(k-1)!}.$$

Relations (13.14) and (13.16) immediately yields

$$Pr[L(n) = k] = |s(k-1, n-1)|/k! \tag{13.17}$$

and, as $k \to \infty$,

$$Pr[L(n) = k] \sim \frac{(log\ k)^{n-2}}{(n-2)!k^2}.$$

In a similar vein, Shorrock (1972) obtained the following result:

$$Pr[L(n) = k | X_1, X_2, \dots, X_n] = \frac{(n-1)!}{k! \tau_n^{n-1}} (1 - e^{-\tau_n})^{k-1} |s(k-1, n-1)|, \tag{13.18}$$

where $\tau_n = -log\{1 - F(X_n)\}$. In order to prove (13.18), Shorrock used the generating function of Stirling numbers of the first kind presented in (13.3).

13.4 Generalized Stirling Numbers and Records in the F^α-scheme

In the last section, we observed that the Stirling numbers of the first kind are closely connected with the distribution of record times in the classical record scheme (viz., when the underlying X_i's are i.i.d.). In this section, we will show how the generalized Stirling numbers of the first kind come into the distribution theory of records arising from an F^α-scheme. Here, we use the conventional notation *F^α-scheme* to denote the case when the underlying independent random variables $X_1, X_2, \ldots$ have distribution functions $F_1, F_2, \ldots$, such that $F_n = F^{\alpha(n)}$, $n = 1, 2, \ldots$, where F is any continuous distribution function and the coefficients $\alpha(1), \alpha(2), \ldots$ are arbitrary positive numbers. The case of equal values for $\alpha(1), \alpha(2), \ldots$ corresponds to the classical record scheme discussed in the last section. Note that we can take $F = F_1$ and, therefore, without loss of any generality, we can take $\alpha(1) = 1$. This F^α-scheme, which is a generalization of Yang's (1975) record model, was suggested and discussed by Nevzorov (1984, 1985, 1986). A variety of results concerning the records arising from this F^α-scheme have been developed by Pfeifer (1989, 1991), Deheuvels and Nevzorov (1993), Nevzorov (1990, 1993, 1995), and Nevzorova, Nevzorov and Balakrishnan (1997); see also Arnold, Balakrishnan and Nagaraja (1997) for a concise review of these developments. It needs to be mentioned here that Ballerini and Resnick (1987) and Deheuvels and Nevzorov (1994) have suggested some further generalizations of this F^α-scheme.

Let us denote $A(0) = 0$, $A(n) = \alpha(1) + \cdots + \alpha(n)$, $n \geq 1$. Furthermore, as in the last section, let $\xi_1, \xi_2, \ldots$ be the record indicators and $N(n) = \xi_1 + \xi_2 + \cdots + \xi_n$ denote the number of records amongst $(X_1, X_2, \cdots, X_n)$. Then, Nevzorov (1984) has proved that the independence of the record indicators (see Renyi's Lemma 13.3.1) continues to hold in this F^α-scheme.

Lemma 13.4.1 *In the F^α-scheme, the record indicators $\xi_1, \xi_2, \ldots$ are all statistically independent with*

$$p_n = Pr[\xi_n = 1] = 1 - Pr[\xi_n = 0] = \frac{\alpha(n)}{\alpha(1) + \cdots + \alpha(n)}, \quad n = 1, 2, \ldots.$$

Lemma 13.4.1 immediately yields the probability generating function of $N(n)$ as

$$\begin{aligned} E[s^{N(n)}] &= \prod_{r=1}^{n} E[s^{\xi_r}] = \prod_{r=1}^{n} (1 - p_r + p_r s) \\ &= \prod_{r=1}^{n} p_r \prod_{r=1}^{n} \left(s + \frac{A(r-1)}{\alpha(r)}\right) = \prod_{r=1}^{n} p_r \prod_{r=0}^{n-1} (s + a_r), \end{aligned}$$

where

$$a_r = \frac{A(r)}{\alpha(r+1)} = \frac{A(r)}{A(r+1) - A(r)}, \quad r = 0, 1, \ldots.$$

Since the elements of the vector $\mathbf{a} = (a_1, a_2, \ldots)$ are all nonnegative and that

$$\prod_{r=0}^{n-1} (s + a_r) = \sum_{k=0}^{n} |s(n, k, \mathbf{a})| x^k$$

[from (13.12)], we have

$$E[s^{N(n)}] = \prod_{r=1}^{n} p_r \sum_{k=0}^{n} |s(n, k, \mathbf{a})| x^k$$

and

$$\begin{aligned} Pr[N(n) = k] &= |s(n, k, \mathbf{a})| \prod_{r=1}^{n} p_r \\ &= |s(n, k, \mathbf{a})| \prod_{r=1}^{n} \left(1 - \frac{A(r-1)}{A(r)}\right), \quad 1 \leq k \leq n, \end{aligned} \tag{13.19}$$

where the Stirling numbers $s(n, k, \mathbf{a})$ are generated by the sequence

$$a_0 = 0, \quad a_r = \frac{A(r)}{A(r+1) - A(r)}, \quad r = 1, 2, \ldots.$$

Independence of the record indicators in the F^α-scheme and (13.19) readily imply that

$$\begin{aligned} Pr[L(n) = k] &= Pr[N(k-1) = n-1] \; Pr[\xi_k = 1] \\ &= |s(n-1, m-1, \mathbf{a})| \prod_{r=1}^{n} \left(1 - \frac{A(r-1)}{A(r)}\right), \quad k \geq n. \end{aligned} \tag{13.20}$$

Since $a_r = \frac{A(r)}{A(r+1) - A(r)}$ and $1 - \frac{A(r-1)}{A(r)} = \frac{1}{1 + a_{r-1}}$, we can write the right-hand sides of (13.19) and (13.20) in terms of coefficients a_r as follows:

$$Pr[N(n) = k] = \frac{|s(n, k, \mathbf{a})|}{\prod_{r=1}^{n} (1 + a_{r-1})} \tag{13.21}$$

and

$$Pr[L(n) = k] = \frac{|s(n-1, k-1, \mathbf{a})|}{\prod_{r=1}^{n} (1 + a_{r-1})}. \tag{13.22}$$

Thus, for any coefficients $\alpha(1), \alpha(2), \ldots$ of the F^α-scheme, we have shown that the generalized Stirling numbers of the first kind, $s(n, k, \mathbf{a})$, which were used in (13.21) and (13.22), can be generated by the sequence

$$a_0 = 0, \quad a_r = \frac{\alpha(1) + \cdots + \alpha(r)}{\alpha(r+1)}, \quad r = 1, 2, \ldots.$$

On the other hand, if we take the generalized Stirling numbers of the first kind, $s(n, k, \mathbf{a})$, for any given vector $\mathbf{a} = (a_0, a_1, \ldots)$ with nonnegative elements, we have also shown that it is possible to construct the corresponding F^α-scheme for which equations (13.21) and (13.22) are satisfied. In this case, the coefficients of the F^α-scheme must satisfy the following equalities:

$$\alpha(1) = 1 \quad \text{and} \quad \alpha(n) = \frac{1}{a_{n-1}} \prod_{r=1}^{n-2} \left(1 + \frac{1}{a_r}\right), \quad n = 2, 3, \ldots. \tag{13.23}$$

In fact, from

$$a_r = \frac{1}{\alpha(r+1)} \{\alpha(1) + \cdots + \alpha(r)\}, \quad r = 1, 2, \ldots,$$

we obtain

$$\begin{aligned}
A(r-1) &= \alpha(1) + \cdots + \alpha(r-1) = a_{r-1}\alpha(r), \\
A(r) &= A(r-1) + \alpha(r) = A(r-1)\left(1 + \frac{1}{a_{r-1}}\right) \\
&= A(1)\left(1 + \frac{1}{a_{r-1}}\right)\left(1 + \frac{1}{a_{r-2}}\right) \cdots \left(1 + \frac{1}{a_1}\right) \\
&= \left(1 + \frac{1}{a_{r-1}}\right)\left(1 + \frac{1}{a_{r-2}}\right) \cdots \left(1 + \frac{1}{a_1}\right)
\end{aligned}$$

and

$$\alpha(r) = A(r) - A(r-1) = \frac{1}{a_{r-1}}\left(1 + \frac{1}{a_{r-2}}\right) \ldots \left(1 + \frac{1}{a_1}\right).$$

13.5 Record Values from Discrete Distributions and Generalized Stirling Numbers

Let $Y_1, Y_2, \ldots$ be a sequence of independent and identically distributed discrete random variables taking on values $2, 3, \ldots$ with positive probabilities $p_2, p_3, \ldots$. Let the record values generated by these variables be denoted by $Y(1), Y(2), \ldots$.

For the sake of simplicity, let us assume that $Y(1) = 1$. Now, let the sequence of random variables $X_1, X_2, \ldots$ form an F^α-scheme with coefficients

$$\alpha(1) = 1, \quad \alpha(n) = \frac{1}{Pr[Y > n]} - \frac{1}{Pr[Y > n-1]}, \quad n = 2, 3, \ldots,$$

and let $L(n)$ denote the record times in this sequence. Nevzorov (1985) [see also Deheuvels and Nevzorov (1993)] has then proved that, for any $n = 1, 2, \ldots,$

$$\{Y(1), Y(2), \ldots, Y(n)\} \stackrel{d}{=} \{L(1), L(2), \ldots, L(n)\}. \tag{13.24}$$

It, therefore, follows from Eqs. (13.22), (13.23) and (13.24) that

$$Pr[Y(n) = k] = \frac{|s(n-1, k-1, \mathbf{a})|}{\prod_{r=1}^{n}(1 + a_{r-1})}, \qquad k = 2, 3, \ldots,$$

where

$$a_0 = 0, a_r = \frac{\alpha(1) + \cdots + \alpha(r)}{\alpha(r+1)} = \frac{Pr[Y > r+1]}{Pr[Y = r+1]}, \; r = 1, 2, \ldots, \tag{13.25}$$

and $s(n, k, \mathbf{a})$, $k = 1, 2, \ldots$, are the generalized Stirling numbers of the first kind corresponding to the sequence $\mathbf{a} = (a_0, a_1, \ldots)$ as given in (13.25).

Acknowledgements. The work of the first author was supported by the Natural Sciences and Engineering Research Council of Canada while that of the second author was supported by the Russian Foundation of Fundamental Research.

References

1. Ahsanullah, M. (1995). *Record Statistics*, Commack, NY: Nova Science Publishers.

2. Arnold, B. C. and Balakrishnan, N. (1989). *Relations, Bounds and Approximations for Order Statistics*, Lecture Notes in Statistics - **53**, New York: Springer-Verlag.

3. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1992). *A First Course In Order Statistics*, New York: John Wiley & Sons.

4. Arnold, B. C., Balakrishnan, N. and Nagaraja, H. N. (1997). *Records*, New York: John Wiley & Sons (to appear).

5. Ballerini, R. and Resnick, S. (1987). Embedding sequences of successive maxima in extremal processes, with applications, *Journal of Applied Probability*, **24**, 827-837.

6. Charalambides, C. A. and Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications, *Communications in Statistics - Theory and Methods*, **17**, 2533-2595.

7. Deheuvels, P. and Nevzorov, V. B. (1993). Records in F^{α}-scheme. I. Martingale properties, *Zapiski Naucnyh Seminarov Leningrad. Otdel. Mat. Inst. Steklova*, **207**, 19-36 (in Russian; to be translated in *Journal of Mathematical Sciences*).

8. Deheuvels, P. and Nevzorov, V. B. (1994). Records in F^{α}-scheme. II. Limit theorems, *Zapiski Naucnyh Seminarov Leningrad. Otdel. Mat. Inst. Steklova*, **216**, 42-51 (in Russian; to be translated in *Journal of Mathematical Sciences*).

9. Goldberg, K., Neiman, M. and Heinsvort, E. (1964). Combinatorial analysis, In *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables* (Eds., M. Abramovitz and I. A. Stegun), National Bureau of Standards, Applied Mathematics Series, No. 55.

10. Nagaraja, H. N. (1988). Record values and related statistics-A review, *Communications in Statistics—Theory and Methods*, **17**, 2223-2238.

11. Nevzorov, V. B. (1984). Record times in the case of nonidentically distributed random variables, *Teoriya veroyatnostey i ee Primenenija*, **29**, 808-809 (in Russian; translated in *Theory of Probability and Its Applications*).

12. Nevzorov, V. B. (1985). Records and inter-record times for sequences of nonidentically distributed random variables, *Zapiski Naucnyh Seminarov Leningrad. Otdel. Mat. Inst. Steklova*, **142**, 109-118 (in Russian; translated in *Journal of Soviet Mathematics*, **36** (1987), 510-516.)

13. Nevzorov, V. B. (1986). Two characterizations using records, In *Stability Problems for Stochastic Models* (Eds., V. V. Kalashnikov, B. Penkov and V. M. Zolotarev), pp. 79-85, Lecture Notes in Mathematics **1233**, Berlin: Springer-Verlag.

14. Nevzorov, V. B. (1987). Records, *Teoriya veroyatnostey i ee Primenenija*, **32**, No.2, 219-251 (in Russian; translated in *Theory of Probability and Its Applications*, **32** (1988), 201-228).

15. Nevzorov, V. B. (1990). Records for nonidentically distributed random variables, *Proceedings of the 5th Vilnius Conference on Probability & Statistics*, pp. 227-233, Amsterdam: VSP Publishers.

16. Nevzorov, V. B. (1993). Characterizations of certain nonstationary sequences by properties of maxima and records, In *Rings and Modules:*

Limit Theorems of Probability Theory - 3 (Eds., Z. I. Borevich and V. V. Petrov), pp. 188-197, St.-Petersburg: St.-Petersburg State University (in Russian).

17. Nevzorov, V. B. (1995). Asymptotic distributions of records in nonstationary schemes, *Journal of Statistical Planning and Inference*, **44**, 261-273.

18. Nevzorova, L. N., Nevzorov, V. and Balakrishnan, N. (1997). Characterizations of distributions by extremes and records in Archimedean copula processes, *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz* (Eds., N. L. Johnson and N. Balakrishnan), pp. 475-484, New York: John Wiley & Sons.

19. Pfeifer, D. (1989). Extremal processes, secretary problems and the 1/e law, *Journal of Applied Probability*, **26**, 722-733.

20. Pfeifer, D. (1991). Some remarks on Nevzorov's record model, *Advances in Applied Probability*, **23**, 823-834.

21. Renyi, A. (1962). Theorie des elements saillants d'une suite d'observations, *Colloquium on Combinatorial Methods in Probability Theory* (August 1-10, 1962), Mathematical Institute, Aarhus University, Aarhus, Denmark, pp. 104-117. See also: On the extreme elements of observations, *Selected Papers of Alfred Renyi, Vol. 3* (1976), pp. 50-65, Budapest: Akademiai Kiado.

22. Rusinski, A. and Voigt, B. (1990). A local limit theorem for generalized Stirling numbers, *Revue roumaine de mathematiques pures et appliquees*, **35**, 161-172.

23. Shorrock, R. W. (1972). On record values and record times, *Journal of Applied Probability*, **9**, 316-326.

24. Yang, M. C. K. (1975). On the distribution of the inter-record times in an increasing population, *Journal of Applied Probability*, **7**, 432-439.

PART III

Applications to Urn Models

14

Advances in Urn Models During the Past Two Decades

Samuel Kotz and N. Balakrishnan
University of Maryland, College Park, MD
McMaster University, Hamilton, Ontario, Canada

Abstract: This paper surveys some key developments that have occurred pertaining to urn models in probabilistic, statistical and biological literatures during the past two decades. It should be regarded as a compact sequel to the book *Urn Models and Their Applications* by Johnson and Kotz (1977) and as a natural companion to the two chapters in the book *Discrete Multivariate Distributions* by Johnson, Kotz and Balakrishnan (1997) dealing with multivariate Pólya-Eggenberger distributions and multivariate Ewens distributions, respectively.

Keywords and Phrases: Urn models, Pólya-Eggenberger distributions, Multivariate Ewens distributions, Classical occupancy models, Ehrenfest urn models, Pólya urn model with a continuum of colors, Urn models with random drawings, Urn models with indistinguishable balls, Compartmental urn models, Applications

14.1 Introduction

Since the publication of the book *Urn Models and Their Applications* by Johnson and Kotz (1977), the theory as well as applications of urn models received increased attention and intensive research from probabilists, statisticians and applied scientists alike. As a result, numerous theoretical results as well as new applications were discovered during the last two decades. A complete bibliography of these developments, if compiled, would include around 800 papers and a booklength account will be necessary to elaborate on all the pertinent details. In this paper, we restrain ourselves to discussing a selected subset of these advancements. We sincerely hope that this admittedly biased subset will reveal to researchers and students the flavor and directions of current research in the area of urn models and will stimulate further advances in this interesting field of research. We have knowingly given somewhat more attention to less accessible papers in this survey. We have included a somewhat selected bibliography in this paper while a more comprehensive bibliography can be obtained by writing to either one of us.

We begin with some brief historical remarks. The concept of urn models dates back to biblical times [see Rabinovitch (1973)] and the ancient Greek period [see Sambursky (1956)]. The first explicit mention of urn models seems to have been made by James Bernoulli (1713) who, in the third book of his *Ars Conjectandi*, discusses the problem of drawing "calculi" out of urns. In a letter dated 3 December 1703, Leibniz questioned Bernoulli on his use of *a posteriori* probabilities and forced him to clarify the assumptions which in his opinion legitimated the mathematics of posterior probabilities. Leibniz's criticisms implicitly questioned the adequacy of Bernoulli's urn model of probability and causation. Since urn models have since become a staple of the literature on probability, a stock way of conceptualizing more intricate problems involving chance, it is important to note that Bernoulli's use of the now familiar urn example to model the relation between underlying causes and observed effects was perhaps the first quantitative attempt to construe a chance mechanism metaphorically. Thereafter, probabilists and statisticians have treated lotteries, dice games, and coin tosses at the immediate level of practical problems, not as analogues for more general processes in nature. Bernoulli's appropriation of the urn example to describe the processes linking inaccessible causes to observed effects expanded not only the domain of problems upon which probabilists and statisticians might test their skills, but also the conceptual tools for extending the range of the theory's applications still further [Daston (1988)] as this review article will hopefully amply testify. Bernoulli also used the term "urn models" in his *Meditationes* which was republished in 1975 as Vol. 3 of *Die Werke von Jakob Bernoulli*. In the third book of *Ars Conjectandi*, Bernoulli formulated the first and the third problems in the language of urn problems. Also, Problem 6 is a reformulation of Problem 4 at the end of the classical 1657 treatise of Huygens [see, for example, Maistrov (1980) for more details]. Huygens, however, did not use the term "urns"; hence, the priority in this respect belongs to J. Bernoulli. An important point to be made in this connection is that Bernoulli used the term "urn model" as equivalent to models involving a die or an "abstract" die introduced by de Moivre, cards, etc. The essential idea behind this is that these models assure "equipossibility" which was the basic supposition for the notions of chance and probability in late seventeenth century. Later on, Laplace, who owes much to Bernoulli's earlier contributions, provided the required prominence to urn models. The formal definition of urn models—random allocation of balls into urns—does not of course capture the variety and richness of the concept. Stigler (1986, p. 124) has elaborated on beautifully this role that Bernoulli played with regard to urns and also explained how Bayes's structure differed from that of Bernoulli in order to treat the problem of binomial directly.

An urn model proposed by the English mathematician, Augustus De Morgan in 1838, is closely related to the Laplace rule of succession and the Ewens sampling formula described later in Section 14.11 of this paper. An excellent analysis of De Morgan's urn model in connection with these problems and the problem of random partitions has been given by Zabell (1992).

It is possible to assign urn models to a majority of chance experiments, particularly those with a countable sample space. The urn model idea plays a fundamental role in many problems for the following main reasons:

- It is an efficient way to describe the concept of "random choice" which can be tested a posteriori but which is in principle inaccessible to an absolute mathematical definition;

- Urns and chance experiments can be compounded into new ones—hyperurns, in Pólya's (1963) terminology—corresponding to compounded experiments. As Pólya ably displayed over thirty years ago, it allows one to "simulate" such complex random (chance) processes as the course and pattern of weather as a sequence of urns;

- The term "simulation" can be interpreted as a statistical equivalent to the basic mathematical concept of *isomorphism* which is intrinsically associated with urn models.

As emphasized by Karlin and Leung (1991) and by Holst on numerous occasions [Blom and Holst (1991) capture the flavor of Holst's technique], many ball-in-urn distributional problems can be handled expeditiously via an embedding into an appropriate system of independent Poisson processes or equivalently to distributing a Poisson distributed number of balls in the urns and from this it is easy to obtain the corresponding results for a fixed number of balls from. This technique, quite an old one, has been explained in the book by Johnson and Kotz (1977). It is well known that the embedding into Poisson processes is equivalent to distributing a Poisson number of balls into the urns, and it is easy to obtain the corresponding results for a fixed number of balls from this fact. We shall return to this point later in Section 14.2 when dealing with generalizations of the Pólya-Eggenberger urn model.

As mentioned earlier, we will not discuss multivariate urn problems in this paper; interested readers may refer to the recent book by Johnson, Kotz and Balakrishnan (1997) for a detailed discussion on this topic. Another related topic that has not been covered in this article is the generalized ballot problem investigated by Mohanty (1979) [see also Watanabe (1986)] and closely related to the so-called *Takacs' urn model* [Takács (1962, 1967)]. The review article of this topic by Takacs, published in this volume, provides an excellent up-to-date treatment to this topic.

14.2 Pólya-Eggenberger Urns and Their Generalizations and Modifications

The Pólya-Eggenberger urn model was originally applied to problems dealing with the spread of contagious diseases; see Eggenberger and Pólya (1923). In this classical model, an urn initially contains a total of t_0 balls of which w_0 are white and the remaining $t_0 - w_0$ are black. This model has been generalized in the following manner:

> A ball is drawn at random from the urn and its color is noted and returned back into the urn. If the color is white, a white balls and b black balls are added to the urn; if it is black, c white balls and d black balls are added to the urn. It should be noted that a, b, c and d can take on negative values indicating that some balls can be thrown away from the urn. Now, denote by W_n the number of white balls in the urn after n draws.

The probability distribution of W_n is known for some special cases; see Sections 4.3 and 6.3 of Johnson and Kotz (1977).

In an interesting paper dealing with applications of these urn models in computer data structures, Bagchi and Pal (1985) defined a *tenable Pólya-Eggenberger model* as one described above satisfying the following conditions:

(i) $a + b = c + d = s \geq 1$ - that is, the same number of balls is added to the urn at every stage;

(ii) $t_0 \geq 1, w_0 \geq 1$;

(iii) $a \neq c$ - that is, W_n is non-deterministic;

(iv) $b > 0, c > 0$,

and a somewhat artifical condition

(v) If $a < 0$, then a divides c and also w_0; if $d < 0$, then d divides b and also $t_0 - w_0$.

Assumption (iv) is a natural one in the sense that if $b < 0$ (say) there may not be any more black balls left in the urn to throw away in which case the model gets into difficulty. Assumption (v) protects against the possibility, when a white ball is drawn from the urn, of having fewer than $-a$ white balls in the urn.

Observe that $b = c = 0$ corresponds to the classical (simple) Pólya-Eggenberger model which has been studied extensively. The case $\min(b, c) = 0 < \max(b, c)$ presents some technical difficulties and in this case the asymptotic distribution of W_n has been discussed by Gouet (1989, 1993).

The tenable Pólya-Eggenberger model represents the process of *random* insertions of "keys" into a 2-3 tree (a rooted oriented tree in which each internal node has 2 or 3 sons and every path from the root to a leaf has the same length). Such trees are widely used data structures for storage organizations in computers and consequently Pólya-Eggenberger models could be used for estimation of memory requirements; see Bagchi and Pal (1985) for details. The replacements in this model are controlled by a deterministic matrix $\mathbf{R} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Aldous, Flannery and Palacios (1988) have modelled 2-3 trees using an urn scheme with non-negative $\mathbf{R}$. It turns out that the distribution of the standardized random variable corresponding to W_n converges asymptotically (as $n \to \infty$) to standard normal distribution whenever $q \equiv a - c \leq s/2$. Bagchi and Pal (1985) have proved this result by using the method of moments. Their main idea was to determine the higher-order moments of the variable

$$Y_n = W_n - \left(\frac{c}{b+c}\right) t_n, \text{ where } t_n = t_0 + ns, \tag{14.1}$$

with the help of which the asymptotic values of the higher-order central moments of W_n and hence of

$$Z_n = \frac{W_n - E(W_n)}{\sigma(W_n)} \tag{14.2}$$

can be computed. Note that

$$Z_n = \frac{W_n - E(W_n)}{\sigma(W_n)} = \frac{Y_n - E(Y_n)}{\sigma(Y_n)}, \tag{14.3}$$

where $\sigma(W_n)$ and $\sigma(Y_n)$ denote respectively the standard deviations of the random variables W_n and Y_n. Bagchi and Pal (1985) have shown, in particular, that for even $r \geq 2$

$$E(Y_n^r) = e_r\, t_n^{r/2} + o(t_n^{r/2}), \tag{14.4}$$

where

$$e_r = 1 \cdot 3 \cdots (r-1)\, e_2^{r/2} \text{ and } e_2 = bc(\frac{q}{b+c})^2/(s-2q), \tag{14.5}$$

and for odd $r \geq 1$

$$E(Y_n^r) = o(t_n^{r/2}); \tag{14.6}$$

further, $\sigma(W_n) \sim C_1 t_n^{1/2}$ for $q < s/2$ and $\sim C_2(t_n \ln(t_n))^{1/2}$ for $q = s/2$. Thus, the moments $E(Z_n^r) \to 1 \cdot 3 \cdots (r-1)$ for r even and $\to 0$ for r odd. This property of the central moments, of course, uniquely characterizes the standard normal distribution [see Johnson, Kotz and Balakrishnan (1994)].

Gouet (1989) corrected some of the statements made by Bagchi and Pal (1985) and showed in particular that for all tenable Pólya-Eggenberger models with the replacement matrix $\mathbf{R} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that $max(b, c) > 0$

$$\frac{W_n}{T_n} \rightarrow \frac{c}{b+c} \quad a.s. \text{ (strong convergence)}, \tag{14.7}$$

where $T_n = B_n + W_n$. [Recall that for the classical Pólya model ($a = d = \alpha > 0$, $b = c = \beta = 0$), $\frac{W_n}{T_n}$ converges strongly to a beta random variable; for example, see Blom and Holst (1991)]. Gouet's approach is based on martingale arguments and does not require moment expressions. The strong convergence $\frac{W_n}{T_n} \rightarrow 0$ (for c=0) can be improved to $W_n/T_n^{a/d} \rightarrow Z$, where Z is a non-degenerate random variable. In a more recent paper, Gouet (1993) established the asymptotic normality of the standardized W_n using martingale theory and extended the convergence result to a functional limit theorem. The normalizing constant and the limiting Gaussian process have been shown to depend on the ratio of the eigenvalues of the replacement matrix $\mathbf{R}$ given above, and more specifically on $\rho = \frac{a-c}{d}$ and the product bc.

A typical result established by Gouet (1993) is as follows:

> If (W_n) is a tenable urn process such that $\mathbf{R} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ where $d > a > 0$ with $bc = 0$ and $\max(b, c) > 0$, then
>
> $$n^{-\rho/2}(W_{[nt^{1/\rho}]} - n^\rho t Z) \rightarrow W \bullet \phi(t), \tag{14.8}$$
>
> where $W \bullet \phi(t)$ denotes the continuous Gaussian martingale with covariance function $\phi(\min(s, t))$ with $\phi(t) = aZt$ and Z as a non-degenerate positive random variable independent of W.

The case of the classical Pólya urn model ($b = c = 0$) was studied earlier by Heyde (1977) and Gouet (1993) in fact used Heyde's methodology in his proof (in particular, the idea that yields a functional central limit theorem for the tail of an *a.s.* convergent martingale.

Pemantle (1990a) has generalized the original Pólya-Eggenberger urn process (in which only one ball of the color initially selected is added to the urn) by assuming that the number of extra balls added of the color drawn is a function of time. He has suggested this model for the American presidential primary election procedure. To this end, let us assume an initial amount of popular support for each candidate that dictates that candidate's chance of winning the first primary and then assume that the support increases proportionally to the size of the states won by the candidate in each primary.

Formally, let $F : Z \geq 0 \to (0, \infty)$ be any function. Let $\nu_1, \nu_2, \ldots$ denote the successive proportions of red balls in an urn that begins with R red balls and B black balls and evolves as follows: at discrete times $n = 1, 2, \ldots$, a ball is drawn and replaced in the urn along with $F(n)$ balls of the same color. (Allow F to take non-integral values by defining the probability of drawing a red ball still to be proportional to the total mass of red balls in the urn.) The usual Pólya urn scheme is the case where $F(n) = 1$ for all n. Pemantle (1990a) has then shown that ν_n must converge for any F and that the limit has no atoms except possibly at 0 and 1. Here, ν_n is the proportion of red balls at time n and $\delta_n = \frac{F(n)}{R+B+\Sigma_{i=0}^{n-1} F(i)}$ are fractional additions. Rigorously stated, the limit ν is such that $\Pr[\nu = 0] = 1 - \Pr[\nu = 1] = \frac{B}{R+B}$ iff $\sum_{n=1}^{\infty} \delta_n^2 = \infty$ and that the distribution of ν has no atoms in (0,1). As a counterexample, let us consider $R = B = 1$ and $F(n) = n$; in this case, clearly the probability that all draws result in the same color is $\frac{2}{3} \cdot \frac{6}{7} \cdot \frac{15}{16} \cdots > 0$, but in view of the above stated result is not entirely concentrated on {0,1}.

Hill, Lane and Sudderth (1980) discussed the following related generalization of the Pólya-Eggenberger urn model:

> An urn containing red and black balls has a given initial composition. At each new time, a ball is drawn from the urn and replaced along with another ball of the same color. The draws are not exactly representative of the contents of the urn but are determined by the contents in the following manner. Let the number of red and black balls at time n be R_n and B_n, respectively, and let $\nu_n = \frac{R_n}{R_n+B_n}$. Instead of drawing a red ball with probability ν_n, draw a red ball with probability $f(\nu_n)$, where f is any function mapping $[0, 1]$ into itself.

Hill, Lane and Sudderth (1980) have then shown that, under a condition on the discontinuities of f, ν_n converges almost surely to a random variable ν for which $f(\nu) = \nu$. On the other hand, if $f(p) = p$ and is a point satisfying $f(x) < x$ for $x < p$ and $f(x) > x$ for $x > p$ in some neighborhood of p (i.e., p is an "upcrossing" for f), then $\Pr[\nu_n \to p] = 0$.

A generalization of this model to urns of more than two colors has been discussed by Arthur, Ermoliev and Kaniovskii (1983). As Pemantle (1990b) correctly pointed out, this generalization considered by Arthur, Ermoliev and Kaniovskii (1983) was already present in a somewhat disguised form in the book of Nevel'son and Hasminskii (1973) in the language of stochastic approximation rather than urn models. The recent work of Benaim and Hirsch (1995) concerning the dynamics of Morse-Smale urn processes needs to be mentioned here as it explains the generic behavior of three-color urn models.

Finally, another interesting modification of the Pólya urn model is the *cannibal model* discussed originally by Green (1980) and more recently by Pittel (1987). In this model, we have an urn which initially contains n balls of which

r are red and the remaining $w = n - r$ are white. At each step, a white ball is removed and one more ball is selected at random from the remaining balls; this ball is then painted red if it was white and then put back into the urn. After at most w draws, all balls become red at which time the process terminates. Here, red balls are interpreted as cannibals and painting a white ball red means that one more member of the population has become a cannibal with the removed white ball representing a victim. The interest lies in this case in the random variable X_{nr} denoting the terminal number of red balls (remember that at each stage a white ball is removed from the run). This model differs from other generalizations and modifications of the Pólya-Eggenberger scheme because the two balls at each draw are not selected at random and also because the total number of draws is random and not fixed.

Closed-form expressions of the distribution of X_{nr} and its moments are not available yet. However, Green's (1980) conjecture that in the case when $r = 1$ (i.e., the urn initially contains only one red ball) the limiting distribution of X_{nr} (as $n \to \infty$) is normal with mean $n\,e^{-1}$ and variance $n\{3e^{-2} - e^{-1}\}$ was verified to be true by Pittel (1987). Specifically, Pittel (1987) has shown that, as $n \to \infty$ and r is such that $\rho = r/n$ is bounded away from 1, the random variable $\{X_{nr} - n\ \phi(\rho)\}/\sqrt{\{n\psi(\rho)\}}$ converges in distribution to the standard normal variables; here, $\phi(x) = e^{x-1}$ and $\psi(x) = e^{2(x-1)}(x^2 - 3x + 3 - e^{1-x})$. Note that if r is close to n (i.e., ρ is close to 1), $E(X_{nr})$ should be close to n and the variance of X_{nr} should be small. Actually, we have $\phi(1) = 1$ and $\psi(1) = 0$. However, the function $\psi(x)$ attains its maximum when $x \simeq 0.259$ which means that the variance of X_{nr} is maximum when the cannibals constitute initially 26% of the whole population. Pittel's proof of the result, which involves Laplace transforms, is based on the observation that the first step of the process results in the number of red balls present in the urn either being the same or increasing by one with the corresponding probabilities $r/(n-1)$ and $1-r/(n-1)$, respectively. The final step involves the use of a theorem of Curtiss (1942) on the moment generating function, and then noting that $exp\{u^2\phi(\rho)/2\}$ is indeed the Laplace transform of a normal distribution with mean 0 and variance $\phi(\rho)$.

For a Pólya urn model containing n balls all of different colors in which balls are drawn randomly one at a time and each drawn ball is replaced together with one more of the same color, one of the most elegant results is due to Holst, Kennedy and Quine (1988) which is based on an earlier seminal paper by Holst (1979). For further details, see Section 14.7 dealing with an unified approach for limit theorems. In this model, let us use $X_i(r)$ to denote the number of balls of color i obtained from r draws, and let N_r be the number of colors not obtained, i.e., $N_r = \sum_{i=1}^{n} I(X_i(r) = 0)$, where $I(\cdot)$ denotes the indicator function. Then, a combinatorial argument yields

$$\Pr[N_r = k] = \frac{\binom{n}{k}\binom{r-1}{n-k-1}}{\binom{n+r-1}{n-1}} = \frac{\binom{n}{k}\binom{r-1}{n-k-1}}{\binom{n+r-1}{r}} \tag{14.9}$$

and

$$\Pr[X_i(r) = 0] = \frac{n-1}{n+r-1}. \tag{14.10}$$

Now, with $E(N_r) = \frac{n(n-1)}{n+r-1} \to \lambda > 0$ (i.e, $r \sim n^2/\lambda$), $N_r \to P(\lambda)$ in distribution. The rate of convergence has been shown by Holst, Kennedy and Quine (1988) to be as follows:

$$d(N_r, N_r^*) \leq C\ (n^{-1}\log n)^{1/2}, \tag{14.11}$$

where

$$d(N_r, N_r^*) = \sup_A |\Pr[N_n \in A] - \Pr[N_n^* \in A]|, \tag{14.12}$$

N_r^* is Poisson with parameter $\frac{n(n-1)}{n+r-1}$ and $C = C(\lambda) > 0$. Their proof of this result is based on the classical methodology of imbedding Pólya drawings into n independent birth processes with intensities $1, 2, 3, \ldots$.

Ivchenko and Ivanov (1995) have investigated the following model: An urn initially contains a given number of balls of N colors. At each trial, a ball is randomly selected from the urn independently of other trials. For a ball drawn of a specific color, before the next trial, the number of balls of the same color in the urn is changed according to a certain rule. This is done by determining a level ν_j for each color j ($j = 1, 2, \ldots, N$) where the quantities $\nu_1, \ldots, \nu_N$ are integer-valued random variables. The procedure then stops when the numbers of balls of arbitrary K colors attain or exceed the corresponding levels for the first time. In other words, $\nu_m(N, k)$ is the number of balls required to be thrown into N urns in order to get for the first time k urns containing no less than m balls each. Several papers have been published about this waiting time variable $\nu_m(N, k)$; see, for example, Ivanov, Ivchenko and Medvedev (1985) and Ivchenko (1993) and the references therein.

In particular, the generalized urn scheme defined by Ivchenko and Ivanov (1995) involves an urn containing initially $a_{j,0}$ balls of color A_j ($j = 1, 2, \ldots, N$) so that $a = \sum_{j=1}^N a_{j,0}$ is the total number of balls in the urn at the beginning of the procedure. Balls are then drawn at random from the urn with each ball having the same probability of being selected. After each draw, the contents of the urn is changed so that if a ball of color A_j is drawn for the n-th time, the number of balls of color A_j is changed from $a_{j,n-1}$ to $a_{j,n}$, $n = 1, 2, \ldots$, $j = 1, 2, \ldots, N$. Note that if the color A_j has already been drawn k_j times, $j = 1, 2, \ldots, N$, the vector of probabilities of drawing a ball of the corresponding color at the next trial is

$$(a_{1k_1}, \ldots, a_{Nk_N}) \frac{1}{\sum_{j=1}^N a_{jk_j}}. \tag{14.13}$$

For the stopping time, a level ν_j for color A_j is determined before the starting of trials and, as mentioned earlier, ν_j ($j = 1, 2, \ldots, N$) are all integers. The trials stop when the frequencies of k unspecified colors attain or exceed the corresponding level for the first time, where k $(1 \leq k \leq N)$ is a given parameter of the stopping time. Now, let $\eta_j(n)$ denote the observed frequency of balls of color A_j after n trials, $j = 1, 2, \ldots, N$. Evidently, $\eta_j(0) = 0$. The stopping time is defined here as

$$\nu(N, k) = \min\{n : \sum_{j=1}^{N} I(\eta_j(n) \geq \nu_j) \geq k\}, \tag{14.14}$$

where $I(\cdot)$ once again denotes the indicator function. Let $\eta_j = \eta_j(\nu(N, k)), j = 1, 2, \ldots, N$, be the frequencies of the corresponding colors at the stopping time. Then, under this very general set-up, Ivchenko and Ivanov (1995) have studied the so-called *decomposable statistics*, a term introduced earlier by Medvedev (1970), defined as

$$L_{Nk} = \sum_{j=1}^{N} g_j(\eta_j). \tag{14.15}$$

Observe that there are three components involved in the construction of decomposable statistics: (i) the parameters $\{a_{jn}\}$ reflecting the generality of the generalized urn scheme, (ii) the stopping rule expressed by the distribution of $\{\nu_j\}$ and the parameter k, and (iii) the specific characteristic of the decomposable statistic expressed by the functions $\{g_j\}$. Examples which are particular cases include:

- *Sampling with replacement* when $a_{j,n} = a_{j,0}$, $n = 1, 2, \ldots$, $j = 1, 2, \ldots, N$. Here, $p_j = a_{j,0}/a$, $j = 1, 2, \ldots, N$ and the vector $(\eta_1(n), \eta_2(n), \ldots, \eta_N(n))$ has a multinomial distribution; see Chapter 35 of Johnson, Kotz and Balakrishnan (1997).

- *Sampling without replacement* when $a_{j,n} = a_{j,n-1} - 1 = a_{j,0} - n$ if $n < a_{j,0}$ and $a_{j,n} = 0$ if $n \geq a_{j,0}$, $j = 1, 2, \ldots, N$. If the observed frequencies of the colors are denoted by $k_1, \ldots, k_N (k_j \leq a_{j,0}$, $j = 1, 2, \ldots, N)$, then provided $\sum_{i=1}^{N} k_i = n < a$, the probability of drawing a ball of color A_j in the next trial is $(a_{j,0} - k_j)/(a - n), j = 1, 2, \ldots, N)$, and

$$\Pr[\eta_j(n) = k_j,\ j = 1, 2, \ldots, N] = \binom{a}{n}^{-1} \prod_{j=1}^{N} \binom{a_{j,0}}{k_j}, \text{ with } \sum_{j=1}^{N} k_j = n, \tag{14.16}$$

  is the multivariate hypergeometric distribution; see Chapter 39 of Johnson, Kotz and Balakrishnan (1997).

- *Pólya sampling* when $a_{j,n} = a_{j,n-1} + s = a_{j,0} + sn$, $n = 1, 2, \ldots$, $j = 1, 2, \cdots, N$. The probability of drawing a ball of color A_j in the next trial is $(a_{j,0} + sk_j)/(a + sn)$, $j = 1, 2, \ldots, N$, where $n = \sum_{i=1}^{N} k_i$. In this case,

$$\begin{aligned}\Pr[\eta_j(n) = k_j,\ j = 1, 2, \dots, N] \\ = \binom{(a/s) + n - 1}{n}^{-1} \prod_{j=1}^{N} \binom{(a_{j,0}/s) + k_j - 1}{k_j},\end{aligned} \tag{14.17}$$

which is the generalized Pólya distribution; see Chapter 40 of Johnson, Kotz and Balakrishnan (1997).

- *The waiting time* is a special case of the decomposable statistic L_{Nk} when $g_j(x) = x,\ j = 1, 2, \dots, N$.

Ivchenko and Ivanov (1995) have noted that the colors $A_1, A_2, \cdots, A_N$ are independent pure birth processes with intensities $\{a_{j,n},\ n \geq 0\}$ for A_j-process. When an event occurs in an A_j-process, $j = 1, 2, \dots, N$, it can be interpreted as a ball of the corresponding color is drawn from the urn and the urn scheme is thus embedded in the process $\mathbf{A} = (A_1, A_2, \cdots, A_N)$. The stopping time is then the k-th order statistic $T_{(k)}$ of

$$T_j = \min\{t : \psi_j(t) \geq \nu_j\},\ j = 1, 2, \dots, N, \tag{14.18}$$

where $\psi_j(t)$ is the number of births in the A_j-process in the interval $(0, t]$ ($\psi_j(0) = 0$). This interpretation allows to determine $E(e^{i\tau L_{Nk}})$ which is the characteristic function of the decomposable statistic L_{Nk}. The binomial, beta and mixture distributions appear prominently in various representations discussed by Ivchenko and Ivanov (1995). Their work is closely related to the seminal paper in this topic by Holst and Hüsler (1985) who were successful in obtaining exact as well as asymptotic results.

Ling (1993) succeeded in deriving the probability distribution function of the waiting time variable under frequency quota defined in a Pólya-Eggenberger urn model. In his terminology, let $W_S(\mathbf{K}, \boldsymbol{\theta}, c)$ be the minimum of drawings until a frequency of K_1 white balls or a frequency of K_2 black balls has been drawn whichever comes first ("sooner" in Ling's terminology); the urn initially contains w white balls and b black balls, and c balls are added after each draw. Here, $\boldsymbol{\theta} = (w, b)$ and $\mathbf{K} = (K_1, K_2)$. A similar set-up with "sooner" replaced by "later" introduced the waiting time variable $W_L(\mathbf{K}, \boldsymbol{\theta}, c)$. Observe that $W_S((K_1, \infty), \boldsymbol{\theta}, c)$ and $W_L((K_1, 0), \boldsymbol{\theta}, c)$ are identical, representing the waiting time for a frequency of K_1 white balls to occur. Explicit expressions for the distributions of the variables W_S and W_L and the corresponding expected values of W_S and W_L are provided only in some special cases. For example, if $K_1 = K_2 = 2$, then $W_S(\mathbf{K}, \boldsymbol{\theta}, c)$ is shown by Ling (1993) to have the same distribution as $2 + X$, where X is a Bernoulli random variable with parameter $\frac{2wb}{(w+b)(w+b+c)}$. It has also been shown that

$$\Pr[W_L((2, 2), (1, 1), 1) = n] = \frac{4}{n(n+1)} \text{ for } n = 4, 5, 6, \dots, \tag{14.19}$$

which readily shows that the mean of $W_L((2,2),(1,1),1)$ does not exist. Ling has additionally presented some partial results on the distributions of $W_S(\mathbf{K},\mathbf{r},c)$ and $W_L(\mathbf{K},\mathbf{r},c)$ by starting with a Pólya-Eggenberger model initially containing r_j (>0) balls of color A_j $(j=1,2,\ldots,m)$, where $\mathbf{r}=(r_1,r_2,\ldots,r_m)$ and $\mathbf{K}=(K_1,K_2,\ldots,K_m)$ are non-negative integers. Additional research is needed in this area.

An elementary but illuminating exposition of the general urn model with multicolor balls with specified drawing procedure with a quota stopping has been presented by Blom and Holst (1991). Specifically, a population of M balls is considered of which m_i are of color i $(i=1,2,\ldots,s)$ with $\sum_{i=1}^{s} m_i = M$. Balls are drawn from this urn one by one according to some prescribed procedure. We assign a stopping rule as follows: With $q_1,q_2,\ldots,q_s$ denoting given positive integers, when q_i balls of color i have been obtained, the color i has reached its quota. The drawings terminate when K arbitrary colors have reached their corresponding quotas. In this case, the distribution of the number of drawings until the stopping (N) is of interest. Blom and Holst (1991) have then discussed the cases of drawings with replacement, drawings without replacement, and the Pólya-Eggenberger urn model. They have utilized embedding techniques by embedding the drawings into either Poisson process, order statistics from uniform distribution, or Yule process. The essential part of this methodology is to draw at random times instead of drawing at fixed times $1,2,3,\ldots$. This technique is then used to derive the first two moments of N in a very elegant manner. The key step is to show that the sequence of colors generated by the superposed process is probabilistically equivalent to the sequence of colors generated by the deterministically specified drawing procedure.

Shur (1984) has discussed the probability distribution that arises from the Pólya-Eggenberger urn model with just one change—viz., the s additional balls that are added after each drawing are of the opposite color instead of being of the same color, thus producing a negative contagion model.

Johnson and Kotz (1991) observed that the standard Pólya-Eggenberger urn model (with w white balls and $n+1-w$ black balls with a single ball of the chosen color added) produces the distribution of the number of white balls T in the course of m drawings with replacements, given by

$$\Pr[T=t] = \binom{m}{t}\frac{q^{[t]}(n+1-q)^{[m-t]}}{(n+1)^{[m]}} \text{ for } t=0,1,\ldots,n. \tag{14.20}$$

This is equivalent to Matveychuk and Petunin's (1990) model which is as follows:

> Let $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ and $\mathbf{Y}=(Y_1,Y_2,\ldots,Y_m)$ be random samples from absolutely continuous distributions with cumulative distribution functions $F_X(x)$ and $F_Y(y)$, respectively. Let $X^{(1)} \le X^{(2)} \le \ldots \le X^{(n)}$ be the order statistics of $\mathbf{X}$, and let $J_{i,q} \equiv$

> $(X^{(i)}, X^{(i+q)})$. Under the assumption that $F_X(x) = F_Y(y)$, let T denote the number of Y_i's falling in the interval $J_{i,q}$.

Then, the distribution of T is indeed the Pólya-Eggenberger distribution in (14.20). Special cases of this statistic arise in connection with several standard rank-order tests in nonparametric statistical inference.

Wei (1979) discussed the following generalized Pólya-Eggenberger urn design in the course of developing a nonparametric treatment assignment in comparing $k(\geq 2)$ treatments in medical trails. This type of a scheme tends to put more patients on better treatments:

> An urn has balls of K different colors. We start with w_i balls of color i, $i = 1, 2, \ldots, K$. When an eligible patient arrives at the experimental site, a ball is selected at random from the urn. We observe its color i and return the ball to the urn. Treatment i is then assigned to this patient. When the response of a previous patient to treatment i is available, we perform one of the following operations: (i) if the response is a success, we add α (> 0) balls of color i, and (ii) if the response is a failure, we add β (> 0) balls of each color j, where $j = 1, 2, \ldots, K$ and $j \neq i$.

This treatment assignment rule is called a *generalized Pólya's urn design* and is denoted by $GPUD(\mathbf{W}, \alpha, \beta)$, where $\mathbf{W} = (w_1, w_2, \ldots, w_K)$. Obviously, this design is applicable when we have delayed responses from patients to the treatments.

Robbins and Whitehead (1979) considered a sequence of i.i.d. random variables $X_1, X_2, \ldots$ with common distribution function $F(\cdot)$. After m observations have been made, fix attention to the i-th smallest and denote its value by ℓ, which is simply the i-th order statistic of $(X_1, X_2, \ldots, X_m)$. Let R_n be its rank when n random variables have been seen where $n \geq m$. Thus, $R_m = i$ and

$$R_n = i + Y_1 + Y_2 + \cdots + Y_{n-m}, \quad n = m, m+1, \ldots, \tag{14.21}$$

where $Y_j = 1$ if $X_{m+j} \leq \ell$ and $= 0$ otherwise, for $j = 1, 2, \ldots$. For any $n \geq m$, the observations $(X_1, X_2, \ldots, X_n)$ divide the real line into $(n+1)$ intervals, and the next observation, X_{n+1}, has an equal probability of falling into any one of these intervals. If it falls into one of the first R_n, then $R_{n+1} = R_n + 1$, while if it falls into one of the remaining $(n + 1 - R_n)$ intervals, then $R_{n+1} = R_n$. Hence, by identifying intervals with balls, this process can be modelled by a Pólya urn scheme. Observe the similarity of this model with the Matveychuk-Petunin model described earlier. The limiting distribution of the variable R_n/n has been studied, among many others, by Gumbel (1958) in the order statistics literature; see also Johnson and Kotz (1977). The limiting variable is a beta random variable with parameters $(i, m-i+1)$. It should be mentioned here that Robbins and Whitehead (1979) studied the probability that the limiting proportion of black balls ever exceeds α, i.e., the probability that the observation of interest **ever** leaves the bottom $100\alpha\%$ of observations.

Multinomial allocations of r balls into n boxes with probability p_k of hitting the kth box and the accuracy of Poisson approximations to the distribution of the number of boxes with a given content have been discussed in some detail by Barbour, Holst, and Janson (1992).

14.3 Generalizations of the Classical Occupancy Model

Maindonald (1990) revisited the classical occupancy model of placing balls into urns (or birthdays into 365 days of the year) expressing it in the language of egg laying procedure carried out by a certain species of parasitic wasps. In this case, the urns are housefly pupae in which a female parasitic wasp places one or more eggs. It turns out that the wasp tends to avoid pupae already parasited leading to *avoidance-modified urn models.*

Specifically, assuming the availability of H urns (pupae), let $p(x, r_0)$ be the probability that after depositing x balls (eggs), r_0 urns remain empty. Each new ball of placed with probability $1/H$ in any one of the H available urns. Using the notation $\rho(r_0) = r_0/H$, Maindonald (1990) established the recurrence relation

$$p(x+1, r_0) = p(x, r_0)\{1 - \rho(r_0)\} + p(x, r_0 + 1)\rho(r_0 + 1). \tag{14.22}$$

This is easily observed by noting that after depositing the $(x+1)$th ball, r_0 urns will remain empty if either $r_0 + 1$ urns had previously been empty, and with probability $\rho(r_0 + 1) = (r_0 + 1)/H$ the $(x + 1)$th ball was placed in an unoccupied urn, or r_0 urns had been empty prior to this $(x+1)$th ball, and with probability $1 - \rho(r_0) = 1 - r_0/H$ the $(x + 1)$th ball was placed in an already occupied urn.

Maindonald (1990) then noted that the recurrence relation in (14.22) can be easily adapted to avoidance-modified urn model situation. Assume that the wasp chooses pupae at random but lays an egg with certainty only if the pupa is "unparasitized." Otherwise, she lays with probability $\delta < 1$. Then, $\rho(r_0)$ above should be interpreted as the probability that the next oviposition will eventually occur in one of the r_0 unparasitized pupae. We thus have the generalization

$$p(x+1, r_0, \delta) = p(x, r_0)\{1 - \rho(r_0)\} + p(x, r_0 + 1)\rho(r_0 + 1), \tag{14.23}$$

with

$$\rho(r_0) = \frac{r_0}{H} + (1 - \frac{r_0}{H})(1 - \delta)\rho(r_0),$$

that is,

$$\rho(r_0) = \frac{r_0}{r_0 + \delta(H - r_0)}.$$

Maindonald (1990) has presented an extensive numerical analysis of this recurrence relation.

Next, he generalized this model by assuming that the encounters with pupae occur according to a Poisson process with parameter λ. [The model (14.23) corresponds to stopping the process when a total of x ovipositions had occurred.] This is the so-called *Poisson embedding* popularized among others by Holst (1986) (as already mentioned in the last section) for deriving theoretical results for classical urn problems. Maindonald (1990) arrived at an estimator of λ to be $\hat{\lambda} = -\ln(r_0/H)$.

Maindonald's recurrence relation in (14.22) corresponds to Harkness's "falling through model" (when a ball falls through with some probability $1 - \phi$ and is unavailable to fill the urn) if we take $\rho(r) = \phi r/H$. The relation

$$\rho(r) = 1 - \left(1 - \frac{r}{H}\right)^r \tag{14.24}$$

may correspond to the situation in which up to the $(r-1)$-th successive encounter with an already parasitized pupae the wasp will not lay an egg, while on the r-th such encounter she will lay an egg. Other choices of $\rho(r)$ as well as of $p(x, r_0, r_1)$—the probability that after depositing x balls, r_0 urns are unoccupied and r_1 urns are singly occupied—are discussed in this elementary and fruitful paper.

A somewhat different generalization of occupancy models is discussed by Fang (1982). He tackles the so-called *restricted occupancy problem* when m urns and n balls are given and each urn consists of k cells. The balls are assigned to the urn in such a manner that each cell will have at most one ball. Denote by M_t the number of urns containing exactly t balls $(t = 0, 1, 2, \ldots, k)$. Several cases depending on whether empty urns are permitted or not, and whether the cells, urns and balls are distinguishable or not, have been provided by Fang (1982). The distribution of M_t depends on $T(n, m, k)$—the number of ways of distributing the n balls among m urns under the restricted model, and $Q(n, m, k, r, t)$—the number of ways of distributing the n balls among the m urns so that exactly r urns have exactly t balls. Evidently, $\Pr[M_t = r] = Q/T$. Fang (1982) expressed $\Pr[M_t = r]$ in terms of T and the function $g(n, k, t, a)$. For example, for the restricted model when empty urns are allowed and the urns are distinguishable but the balls and the cells are not (the well-known Bose-Einstein system—see Section 14.10), by making use of the inclusion-exclusion principle, we have

$$T(n, m, k) = \sum_{j=0}^{m} (-1)^j \binom{m}{j} \binom{n + m - j(k+1) - 1}{m - 1}, \tag{14.25}$$

and in this case $g(n,k,t,a) = 1$. Also,

$$\Pr[M_t = r] = \binom{m}{r} \sum_{j=0}^{m-r} (-1)^j \binom{m-r}{j} T(n-(r+j)t, m-(r+j), k)\, g(n,k,t,r+j), \tag{14.26}$$

once again by applying the inclusion-exclusion principle. Further, expressions for moments and the moment generating function [generalizing the old result of Freund and Pozner (1956)] have also been provided by Fang (1982).

In a remarkable paper, Nishimura and Sibuya (1988) extended the classical occupancy problem to the case when two types of balls are placed. Given two types of balls, say, n_1 white balls and n_2 red balls, the balls are drawn at random and independently into one of m urns with probability $1/m$. Let S denote the number of urns with balls of both colors. For the famous *birthday problem* (n_1 boys and n_2 girls in the sample), the event $S > 0$ means that there is at least one birthday which a boy and a girl have in common. Nishimura and Sibuya (1988) provided an interesting interpretation of this urn model in terms of security evaluation of authentication procedures in electronic communication networks. In their model, each ball enters an urn with or without balls of the different color. In the former case, however, we have a collision between the two different colors. The number of balls that collide with white but not with red will be denoted by $n_1 - Y_1$, and that collide with red but not with white by $n_2 - Y_2$. Nishimura and Sibuya (1988) have then shown that the joint probability mass function $\Pr[(S, R_1, R_2, Y_1, Y_2) = (s, r_1, r_2, y_1, y_2); m, n_1, n_2]$, where R_1 is the number of urns containing white balls but no red balls, and R_2 is the number of urns containing red balls but no white balls, is given by

$$\frac{1}{m^{n_1+n_2}} \binom{n_1}{y_1} \binom{n_2}{y_2} \binom{y_1}{s} \binom{n_1 - y_1}{r_1} \binom{y_2}{s} \binom{n_2 - y_2}{r_2} \frac{m!s!}{(m - r_1 - r_2 - s)!} \,. \tag{14.27}$$

Note also that $T_i = S + R_i$ (i=1,2) (with $1 \le T_i \le \min(n_i, m_i)$, $i = 1, 2$) has the classical occupancy distribution

$$\Pr[T_i = t] = \binom{n_i}{t} \frac{m^{(t)}}{m^{n_i}}, \; 1 \le t \le \min(m, n_i) \tag{14.28}$$

and, similarly,

$$\Pr[S + R_1 + R_2 = u] = \binom{n_1 + n_2}{u} \frac{m^{(u)}}{m^{n_1+n_2}}, \; 1 \le u \le \min(m, n_1 + n_2) \tag{14.29}$$

and the crucial point is to realize that, given $T_1 = t_1$, Y_2 follows a binomial distribution with parameters $n_2, t_1/m$.

Upon using the exponential generating function of $\binom{n}{m}$, $n = m, m+1, \ldots$

$$\sum_n \binom{n}{m} \frac{z^n}{n!} = \frac{(e^z - 1)^m}{m!}, \; m = 1, 2, \ldots, \tag{14.30}$$

one can derive

$$E(S^{(\ell)}) = \frac{1}{m^{n_1+n_2}} m^{(\ell)} \nabla^\ell m^{n_1} \nabla^\ell m^{n_2}, \tag{14.31}$$

where ∇ is the backward difference operator. Similarly, upon noting that the event $S = 0$ is equivalent to the event $Y_1 = 0$ or $Y_2 = 0$, we have

$$\Pr[S = 0; m, n_1, n_2] = \frac{1}{m^{n_1+n_2}} \sum_{t_1} \sum_{t_2} \binom{n_1}{t_1} \binom{n_2}{t_2} m^{(t_1+t_2)}, \tag{14.32}$$

where t_i denotes the number of urns randomly occupied by n_i balls ($i = 1, 2$). In particular, $\Pr[S = 0;\ m, n_1, n_2]$, with m and $n_1 + n_2$ fixed, decreases as $|n_1 - n_2|$ decreases. With N_1 and N_2 denoting the number of balls of each type thrown one by one when the first collision between the two types of balls occurs in one of the m urns, Nishimura and Sibuya (1988) have then shown that

$$\Pr[\frac{N_1 N_2}{m} \le w] \to 1 - e^{-w} \text{ for } 0 < w < M \tag{14.33}$$

for any positive $M > 0$ as $m \to \infty$. It should be noted that a "rule of choice" according to which the white and the red balls are thrown one by one into the urns is determined by n_{1j} and n_{2j}—the number of white and red balls thrown up to the j-th step. The assumption is that $n_{1j} n_{2j} > 0$ after some finite number of steps. However, if white and red balls are thrown alternatively, the distribution of $N_1/\sqrt{m}$ or $N_2/\sqrt{m}$ is asymptotically Rayleigh with probability density function $2we^{-w^2}$, $w > 0$ [see Chapter 18 of Johnson, Kotz and Balakrishnan (1994)].

14.4 Ehrenfest Urn Model

The Ehrenfest urn model and its modifications as described in Johnson and Kotz (1977) (actually a model involving an exchange between two urns) have continued to fascinate researchers in combinatorial probability theory during the past two decades. This model [1], appropriately designated as "An Urn Problem of Paul and Tatiana Ehrenfest" by Takács (1979) among other authors, was originally proposed for resolving the apparent discrepancy between irreversibility and recurrence in Ludwig Boltzmann's theory of gases [Kac (1947) and Takács (1979) supply interesting details in this connection.] Briefly, there are m balls numbered 1 to m distributed between two urns (I and II). Choose an integer between 1 to m (the integers are assumed to be equi-probable) and transfer the

[1] which provides a simple model of heat exchange between two isolated bodies of unequal temperatures. Temperatures correspond to the number of balls in the urn while heat exchange corresponds to transference of balls.

ball with this number to the other box. Let $P(n, m, i, k)$ denote the probability that box I, originally containing i balls, contains k balls after n transferences. Note that $\frac{i}{m}$ is the probability that a ball is transferred from urn I to urn II when urn I contains i balls and the probability of transferences from urn II to urn I is $\frac{m-i}{m}$. Thus, the system is equivalent to a *random walk* in which the state of the system is the number of balls in urn I and the single-stage transition probabilities from state i to state $i-1$ $(i+1)$ is $\frac{i}{m}$ $(\frac{m-i}{m})$. The n-stage transition probability from state i to state k is given by $P(n, m, i, k)$. Takács (1979) has provided a novel short derivation of an expression for $P(n, m, i, k)$, thus simplifying Kac's (1947) classical derivation. Takács ingenious proof of Kac's classical result is based on the notion of a Poisson process and by connecting the transition probabilities of the homogeneous Markov chain generated by the Ehrenfest urn model and a homogeneous Markov process representing on the time interval $[0, \infty)$ machines which operate independently with working and idle periods being mutually independent exponential random variables with cumulative distribution function

$$F(x) = 1 - e^{-x}, \qquad x \geq 0.$$

Goulden and Jackson (1986) have provided an elegant combinatorial derivation of the Kac-Takács form for $P(n, m, i, k)$ as

$$P(n, m, i, k) = \frac{1}{2^m} \sum_{j=0}^{m} a_{ij} a_{jk} \left(1 - \frac{2j}{m}\right)^n, \tag{14.34}$$

where

$$\sum_{j=0}^{m} a_{ij} z^j = (1-z)(1+z)^{m-i}.$$

The mathematical tool used is continued fractions to enumerate lattice weighted paths and combinatorial bijections between them and permutations [see, for example, Françon and Viennot (1979)] and pairings [a concept due to Read (1979) and independently by Flajolet (1980)].

In an obscure (for a non-Scandinavian reader) Swedish journal by name *Elementa*, Matsoms (1988) conjectured that when $m = 2N$ the "average time" (provided the balls are transferred at each time-point $t = 0, 1, 2, \ldots$ and the process starts at state E_0) t_k required for transition from E_0 to E_k (k balls are in urn II, $k = 0, 1, \ldots, m$) satisfies

$$t_N = N \sum_{j=1}^{N-1} \frac{1}{2j+1}\,. \tag{14.35}$$

Blom (1989) provided a general representation for t_k in the form of a definite integral in terms of a type of beta function (applying a recursive formula for the expectation of a one-step transition) as

$$t_k = \frac{m}{2} \int_0^1 (1-y)^{m-k}[\{(1+y)^k - (1-y)^k\}/y]\, dy. \tag{14.36}$$

The above integral reduces to $\sum_{j=0}^{N-1} \frac{1}{2j+1}$ for the case $m = 2N$ showing that Matsoms' conjecture is indeed correct, and also leading to the neat expression

$$t_m = m \sum_{j=0}^{m-1} \frac{2j}{j+1} \tag{14.37}$$

for the case $k = m$. Observe that for $m = 10$, $t_5 = 8.9$ while $t_{10} = 1186.5$. Evidently, the final terms are the dominant ones in the expression for t_m.

It needs to be mentioned here that an alternative formula for the mean passage times for the Ehrenfest urn model was derived by Kemperman (1961) in terms of Krawtchouk polynomials

$$K_n(x, p, N) = {}_2F_1(-n, -N; -x; 1/p) = \sum_{j=0}^{n} \frac{(-n)_j(-x)_j}{j!(-N)_j} \left(\frac{1}{p}\right)^j . \tag{14.38}$$

where $(-x)_j$ is as defined and used by Johnson, Kotz, and Kemp (1992). Voit (1996) has recently discussed the asymptotic distributions for the Ehrenfest urn model and some related random walks. His method, based on Krawtchouk polynomials and a version of the Diaconis-Shahshahani upper bound lemma, also gives exact asymptotic error rate.

Motivated by the works of Kemperman and Blom, Krafft and Schaefer (1993) defined a generalized two parameter Ehrenfest model. As earlier, there are n balls distributed in two urns I and II. Consider a Markov chain which is in state i when there are i balls in urn I. At each time point, one ball is chosen with equal probability. If it is in urn I, it is then placed in urn II with probability t and returned to urn I with probability $1 - t$. If the selected ball is in urn II, it is then placed in urn I with probability s and returned to urn II with probability $1 - s$. (Visualize—if you will—a vessel separated by a diaphragm which has a different permeability from either side). For the case $s = t = 1$, we get the classical Ehrenfest urn model, while the case $s + t = 1$ corresponds to the one-parameter Ehrenfest urn model discussed by Karlin (1968) in his classical text. Krafft and Schaefer (1993) have then shown that the above described generalized two-parameter Ehrenfest urn model corresponds to a homogeneous Markov chain with state space $(0, \ldots, n)$ and transition probabilities

$$\begin{aligned} p_{i,j} &= \left(1 - \tfrac{i}{n}\right) s && \text{if } j = i+1 \\ &= \tfrac{i}{n} t && \text{if } j = i-1 \\ &= \tfrac{(1-s)n + i(s-t)}{n} && \text{if } j = i \\ &= 0 && \text{otherwise,} \end{aligned}$$

and derived general formulas for the expected first passage times and transitional probabilities of quite a different structure as compared to those of Blom or Kemperman.

Yet another generalization of the Ehrenfest urn model was given earlier by van Beek and Stam (1987) wherein the model is modified by drawing r (> 1) balls at a time. The authors have then shown that the stationary distribution on $E = \{0,1\}^N$ of all sequences $x = (x_1, \ldots, x_N)$, where $x_i = 1$ means that the ball i is in urn I and $x_i = 0$ means that the ball is in urn II, is uniform on E—exactly the same as in the classical case corresponding to $r = 1$.

Dette (1994) generalized the Krafft-Schaefer model by allowing the transition probabilities $p_{i,i-1}$ and $p_{i,i+1}$ to be quadratic functions of the current state i. It turns out that the results obtained by Krafft and Schaefer (1993) also hold for Dette's model, with the change that the Krawtchouk polynomials in the two parameter Ehrenfest urn model replaced by the so-called Hahn discrete orthogonal polynomials defined in terms of ${}_3F_2$ functions [see Johnson, Kotz, and Kemp (1992) for a definition of ${}_3F_2$ function]. The transition probabilities in Dette's generalization are given by

$$\begin{aligned} p_{i,j} &= \left(1 - \tfrac{i}{N}\right) \tfrac{a+1+i}{N+a+b+2}\nu & \text{if } j = i+1 \\ &= \tfrac{i}{N} \tfrac{N+b+1-i}{N+a+b+2}\nu & \text{if } j = i-1 \\ &= 1 - \left(1 - \tfrac{i}{N}\right) \tfrac{a+1+i}{N+a+b+2}\nu - \tfrac{i}{N} \tfrac{N+b+1-i}{N+a+b+2}\nu & \text{if } j = i \\ &= 0 & \text{otherwise,} \end{aligned}$$

where a, b are real numbers such that either a, b >-1 or a, b <-N, and $\nu > 0$ is arbitrary assuring that $p_{i,j} \in [0,1]$ for all $i, j \in \{0, 1, \ldots, N\}$. Setting $a = \frac{su}{s+t}$, $b = \frac{tu}{s+t}$ and $\nu = s + t$ and taking the limit as $u \to \infty$, we arrive at the Krafft-Schaefer generalization of the Ehrenfest urn model.

Palacios (1994) observed that the evolution of the Ehrenfest urn model could be viewed as a simple random walk on a cube. Using a "full description" with 2^n states representing the possible configurations by n-tuples $\mathbf{i} = (i_1, \ldots, i_n)$ where $i_k = 1$ or 0 if ball k is in urn I or II, one may identify the states with the vertices of the n-cube. Palacios utilized an electric approach [see, for example, Doyle and Snell (1984)] to random walks on graphs to compute the hitting times (or first passage times). Using this approach, in particular, the expected time to move all balls from urn II to urn I can be shown to be

$$2^{n-1} \sum_{k=0}^{n-1} \frac{1}{\binom{n-1}{k}} \,. \tag{14.39}$$

Another related result shows that the expected time for an urn full but for one ball to get the very last ball to be

$$E_{n-1}T_n = 2^n - 1 \tag{14.40}$$

thus showing that it is indeed very difficult for an urn to get the very last ball. In fact, more generally,

$$E_{k-1}T_k = \frac{\sum_{j=0}^{k-1}\binom{n}{j}}{\binom{n-1}{k-1}}. \tag{14.41}$$

Note that $E_{k-1}T_k$—the expected time to increase by one (from $k-1$) the count of balls in urn I—forms a strictly increasing sequence in k which follows easily from the inequality

$$\frac{\binom{n}{s}}{\binom{n-1}{k-1}} < \frac{\binom{n}{s+1}}{\binom{n-1}{k}} \quad \text{for} \quad 0 \le s \le k-1. \tag{14.42}$$

Inequality (14.42) becomes an equality for $s = k-1$. Thus, the increase is slow for $1 \le k \le n/2$ [since both numerator and denominator in (14.41) increase] and is rapid for $k > n/2$ [since numerator increases but denominator decreases in (14.41)]. This means that $E_0 T_{n/2}$ is much smaller than $E_0 T_k$ for $k > n/2$ as observed earlier by Blom (1989) for $k = n$.

In an interesting paper, Bingham (1991) interpreted the Ehrenfest urn model as a random walk on the unit cube in n dimension and focussed his attention on the "fluctuation theory" of model (behavior on unusual states) and, in particular, on the first passage time to the opposite vertex (T_{0n}) and its continuous time analog. Bingham noted the importance of the latter concept in reliability theory [see Keilson (1979) and Takács (1979)] and in genetics [Donnelly (1983)].

Bingham's results include:

- The probability generating function of T_{0n} is

$$E[s^{T_{0n}}] = \frac{\sum_{j=0}^{n}(-1)^j\binom{n}{j}\Big/\left[1 - s\left(1 - \frac{2j}{n}\right)\right]}{\sum_{j=0}^{n}\binom{n}{j}\Big/\left[1 - s\left(1 - \frac{2j}{n}\right)\right]}; \tag{14.43}$$

- The fact that $2^{-n}T_{0n} \xrightarrow{d} \mathit{exponential}(1)$ as $n \to \infty$.

It is important to note that the Ehrenfest urn model can be described by a Markov chain at two levels: "full description" with 2^n states (all possible configurations by n-tuples of 0's and 1's) and a "reduced description" with $n+1$ states counting the number of balls in urn I. In the reduced description, the Ehrenfest matrix $\mathbf{P} = ((p_{i,j}))$ is given by

$$\begin{aligned} p_{i,j} &= \tfrac{n-i}{n} && \text{if } j = i+1 \\ &= \tfrac{i}{n} && \text{if } j = i-1 \\ &= 0 && \text{otherwise} \end{aligned}$$

and has eigenvalues $\lambda_j = 1 - \frac{2j}{n}$, $j = 0, 1, \ldots, n$. In particular, for $j = n$ we have the eigenvalue -1 reflecting periodicity of period 2, as well as the Perron-Frobenius eigenvalue equal to 1 for $j = 0$.

Continuous time formulations (not discussed here for brevity) yield different limiting results; refer to Bingham (1991) for details.

In a not so widely known paper, Uppuluri and Wright (1987) discussed the following extension of the Ehrenfest urn model. Given an urn with w_0 white balls and b_0 black balls with the total number of balls $N = w_0 + b_0$ being constant, a ball is selected at random in the first trial—if it is white, it is replaced by a black (white) ball with probability $\alpha_1(1 - \alpha_1)$; if it is black, it is replaced by a white (black) ball with probability $\alpha_2(1 - \alpha_2)$. An analogous procedure is carried out at the i-th trial. Let $W_n(B_n)$ denote the number of white (black) balls in the urn after the n-th trial ($B_n = N - W_n$). Then the expected number of white and black balls that we can expect to have in the urn after n trials is given by

$$\boldsymbol{\mu}_n = E\,(\,W\,)_n\,B_n = \left[\mathbf{I} + \frac{1}{N}\mathbf{A}\right]^n \boldsymbol{\mu}_0, \tag{14.44}$$

where $\mathbf{A} = \begin{pmatrix} -\alpha_1 & \alpha_2 \\ \alpha_1 & -\alpha_2 \end{pmatrix}$ and $\boldsymbol{\mu}_0 = \begin{pmatrix} w_0 \\ b_0 \end{pmatrix}$. Evidently, this result can be easily derived from Krafft and Schaefer's (1993) model by reformulating the problem in terms of two urns, but the latter authors apparently were not aware of Uppuluri and Wright's work. After receiving a copy of Uppuluri and Wright's (1987) note from the authors of this survery, Krafft and Schaefer (1996) (in a yet unpublished manuscript) provided more tractable formulas for the expected value and the variance of the number of white balls after n trials in their framework of Markov chains (by utilizing the generating functions of Krawtchouk polynomials).

Uppuluri and Wright (1987) viewed their generalization of the Ehrenfest urn model as a flexible sampling scheme which leaves the choice of replacement at each trial to chance and at the same time achieves a predetermined goal about the desirable proportion of white balls. An explicit expression for $E[W_n]$ is easy to derive by using Blatz's (1968) result which asserts that for any matrix $\mathbf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$,

$$\mathbf{M}^n = \frac{\lambda_1^n - \lambda_2^n}{\lambda_1 - \lambda_2}\mathbf{M} - \frac{\lambda_1^{n-1} - \lambda_2^{n-1}}{\lambda_1 - \lambda_2}\lambda_1\lambda_2\mathbf{I}, \tag{14.45}$$

where λ_1 and λ_2 are the eigenvalues of $\mathbf{M}$ and $\mathbf{I}$ is a (2×2) identity matrix. Using this expression, one could then easily determine α_1 and α_2 which would on an average yield a predetermined "goal" value for $E[W_n]$ after a predetermined number of trials n in a particular urn. Some numerical values for (α_1, α_2) are given by Uppuluri and Wright (1987).

14.5 Pólya Urn Model with a Continuum of Colors

Pólya urn model with a continuum of colors was introduced by Blackwell and MacQueen (1973). However, only after 20 years later, an important investigation of this model was carried out by Yamato (1993). This model, briefly stated, is as follows [for motivation and more details, one may refer to Johnson and Kotz (1977)]: A color is initially selected from a continuous probability distribution Q (on a d-dimensional space R^d) and r balls of this color are thrown into an empty urn. Next, after n draws with probability $\frac{M}{M+nr}$ a color is selected from the distribution Q and r balls of this color are thrown into the urn **or** with probability $\frac{nr}{M+nr}$ a ball is drawn from the urn and returned to it with r balls of the same color. We thus arrive at $X_1, X_2, \ldots, X_n, \ldots$, a sequence of chosen colors. The important characteristic of this model is $C(m_1, \ldots, m_n; n)$ which denotes the set of the first n trials in which m_1 colors appear once, m_2 colors appear twice, and in general m_i colors appear i times, where $\sum_{i=1}^n i\ m_i = n$.

Let $(X_{i,1}, \ldots, X_{i,m_i})$ be the m_i colors appearing i times given that $(X_1, \ldots, X_n)$ belongs to $C(m_1, \ldots, m_n; n)$. A formal definition of the sequence of colors $X_1, \ldots, X_n$ is as follows:

- $\Pr[X_1 \in A] = Q(A)$, where Q is a continuous probability distribution on R^d and the Borel set $A \subseteq R^d$;
- $$\Pr[X_{n+1} \in A | X_1 = x_1, \ldots, X_n = x_n] = \frac{MQ(A) + r\sum_{i=1}^n I_{X_i}(A)}{M+nr}, \quad (14.46)$$

  where $I_x(A)$ is the indicator function taking on the value 1 if $x \in A$ and 0 otherwise.

It is clear that, by introducing the parameter $M^* = M/r$, r can be chosen to be 1 without loss of any generality. If we now introduce the variable D_n for the number of distinct colors among the sequence $X_1, \ldots, X_n$, it is evident that $D_n = \sum_{i=1}^n m_i$ if $(X_1, \ldots, X_n)$ belongs to $C(m_1, \ldots, m_n; n)$. The important point here is to realize that the event $\{D_n = k\}$ is equivalent to the union of the events $\{(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n; n)\}$ over $(m_1, \ldots, m_n)$ satisfying the conditions $\sum_{i=1}^n im_i = n$ and $\sum_{i=1}^n m_i = k$. This readily leads to the Stirling numbers of the first kind $s(n, k)$ and a result of the type

$$\Pr[D_n = k] = |s(n,k)| \frac{M^k}{M^{[n]}} \qquad (14.47)$$

as shown by Yamato (1993).

Another sequence of interest in this model is $Y_1, Y_2, \ldots$ corresponding to the sequence of new colors. Clearly, $Y_1, Y_2, \ldots, Y_k$ are the distinct colors amongst $X_1, X_2, \ldots, X_n$ if $(X_1, \ldots, X_n)$ belongs to $C(m_1, \ldots, m_n; n)$ under the condition

$\sum_{i=1}^{n} m_i = k$. A very interesting result in this model, due to Yamato (1993), is that the random variables $Y_1, Y_2, \ldots$ are all independent and identically distributed random variables with the distribution Q. Yamato (1993), as a matter of fact, applied this result to estimate the parameters M and Q. In particular, D_n is a complete sufficient statistic for the parameter M. The maximum likelihood estimator $\hat{M}$ of M is obtained by maximizing

$$L(M) = \frac{M^{D_n}}{M^{[n]}} \, . \tag{14.48}$$

When $D_n = 1$, $\hat{M} = 0$ and when $D_n = n$, $\hat{M} = \infty$ which are intuitively clear. Also, since $(X_1, \ldots, X_n) \in C(m_1, \ldots, m_n; n)$ does not depend on the continuous distribution Q, the empirical distribution function based on the distinct colors X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m_i$, is the "best" estimator of Q.

14.6 Stopping Problems in Urns

A typical stopping problem in urns was considered by Simons (1987) which is described below exactly in his formulation. Suppose an urn contains m (-1)'s and p $(+1)$'s. (Note how each author uses different terminology which often causes confusion.) We draw at random without replacement and we are free to stop at anytime. The objective is to maximize the sum. The values m and p are pre-assigned and we are free not to carry out any drawings at all. The problem that Simons tried to solve (and succeeded only partially) is to determine whether the expected return $R(m,p)$ under optimal stopping is strictly positive. Starting with $R(m,0) = 0$ $(m = 0, 1, \ldots)$ and $R(0,p) = p$ $(p = 0, 1, \ldots)$, the recurrence relation

$$R(m,p) = \max\left(0, \frac{m}{m+p} R(m-1,p) + \frac{p}{m+p} R(m,p-1) - \frac{m-p}{m+p}\right) \tag{14.49}$$

will enable the determination of $R(m,p)$ easily. Simons (1987) observed that $R(m,p)$ can be positive even when $m > p$. [Obviously, $R(m,p) > 0$ when $p > m$.] Since the sampling is without replacement, it may be desirable to proceed with a policy more sophisticated than making a fixed number of draws that depends upon the outcomes of the draws. Combining earlier results of Shepp (1969) and Boyce (1973), Simons (1987) suggested the following first (second)-order asymptotic stopping rules: Let $n = m + p$ and $\alpha \approx 0.83992$ be the unique solution of the equation

$$(1 - \alpha^2) \int_0^\infty e^{\alpha x - x^2/2} dx = \alpha. \tag{14.50}$$

Then, stop as soon as the current values of k and n satisfy $k \geq \alpha\sqrt{n}$ (first order rule), or as soon as $k \geq \alpha\sqrt{n} - 0.5$ (second order rule), or as soon as $k \geq \alpha\sqrt{n} - 0.5 + \beta c(n)$ (third order rule), where $c(n)$ goes to zero with n and β is an empirically determined constant which depends on $c(n)$.

Simons (1987) determined all favorable urns (satisfying $R(m,p) > 0$ under the optimal stopping rule) for sizes $n \leq 54,000$ (some 196 urns). There are about 1458,081,000 urns of size $n \leq 54,000$. The procedure took about 100 hours of computing time on an IBM-AT. Simons has cautiously recommended the choice $c(n) = n^{-1/2}(\log n)^2$ with corresponding β satisfying the inequality $0.008890 \leq \beta \leq 0.008976$.

Samuel-Cahn (1993) discussed the following more refined problem. An urn contains N balls, labelled $1, 2, \ldots, N$. The balls are drawn one at a time. Let M_k be the maximal label seen by time k. The payoff function $f(k,m)$ (i.e., reward when stopping after k draws and largest number seen by then is m) is assumed throughout to be nondecreasing in m for each k. The problem discussed is about the optimal stopping rules for sampling with or without replacement. Let $n \leq N$ denote the total permissible "horizon", i.e, the latest draw by which one must stop. It has been shown *inter alia* that for any horizon under optimal stopping, sampling without replacement yields a larger expected value than under sampling with replacement. Samuel-Cahn also investigated the conditions under which the optimal rule has the simple form

$$t = \inf\{k : M_k \geq q_k\}$$

for some constant q_k (the so-called *threshold rule*). A sufficient (but not necessary) rule for sampling with replacement is

$$\frac{m}{N} \triangle\,(k,m) \leq \triangle(k-1,m) \text{ for } k = 2,\ldots,n,\ \ m = 1,\ldots,N-1, \tag{14.51}$$

where $\triangle(k,m) = f(k,m+1) - f(k,m)$. Evidently, this condition holds when $\triangle(k,m)$ is nonincreasing in k for every fixed m. Samuel-Cahn (1993) has also briefly investigated the limiting behavior (as $N \to \infty$), generalizing the earlier result of Chen and Starr (1980).

14.7 Limit Theorems for Urns with Random Drawings

Shortly after the publication of the volume by Johnson and Kotz (1977), Holst (1979) provided a unified approach for proving limit theorems for a variety of urn models.

Consider an urn containing balls of N different colors numbered $1, 2, \ldots, N$. Balls are drawn at random one at a time. There are three schemes considered:

the chosen ball is returned (scheme M), the chosen ball is not returned (scheme H), and the chosen ball is returned together with s balls of the same color (scheme P). The urn initially contains A balls. After n drawings, let us denote the number of selected balls of different colors by $X_1, X_2, \ldots, X_N$.

For a given function $f(\cdot)$ and integer $M \le N$, Holst (1979) considered the random variable

$$Z_M = \sum_{k=1}^{M} f(X_k).$$

If $f(0) = 1$ and $f(j) = 0$ otherwise, then Z_M represents the number of colors amongst the first M that did not occur in the n drawings—that is, we have an occupancy problem, with the classical occupancy problem being the special case when $M = N$, $f(0) = 1$ and $f(j) = 0$ otherwise under scheme M. Holst then derived an expression for the characteristic function of Z_M, $E[e^{ivZ_M}]$, in terms of $E[e^{iu(Y-n/N)}]$, where (i) Y is a Poisson random variable in the case of scheme M, (ii) Y is a binomial random variable in the case of scheme H, and (iii) Y is a negative binomial random variable in the case of scheme P. He also presented a limit theorem for Z_M (as M, N and $n \to \infty$) by using a general method originally devised by Le Cam (1958). Holst then illustrated his result for a variety of classical urn models. Since then, most of the limit theorems in the literature mainly dealt with cases which can not be proved from Holst's theorem. For example, limit theorems for sequential occupancy and for infinite urn models are two such cases. These are discussed in the next two sections.

14.8 Limit Theorems for Sequential Occupancy

Limit theorems for sequential occupancy [see Johnson and Kotz (1977, p. 353)] are based on the classical result of Erdös and Renyi (1961). Balls are successively thrown, independently and uniformly, in n given urns labeled $1, 2, \ldots, n$. Let $N_{n,m}$, $1 \le n, m < \infty$, be the number of throws required to obtain at least m balls in each urn, in which case the urns are said to be covered m times. Erdös and Renyi (1961) proved the following limit law:

Let $N_{n,m} = n \log n + (m-1) n \log\log n + n X_{n,m}$. *Then,*

$$\lim_{n\to\infty} \Pr[X_{n,m} \le x] = e^{-e^{-x}/(m-1)!}.$$

Observe that the above given limiting distribution of $X_{n,m}$ is an extreme value distribution; for example, see Chapter 22 of Johnson, Kotz and Balakrishnan (1995). Newman and Shepp (1960) derived the asymptotic behavior of the expectation $E[N_{n,m}]$ as

$$E[N_{n,m}] = n \log n + (m-1) n \log\log n + n C_m + o(n) \text{ as } n \to \infty,$$

where $C_m = \gamma - \log(m-1)!$, γ being Euler's constant. This is also immediately evident from the limiting distribution of $X_{n,m}$ being an extreme value distribution.

The above results reveal at once the rather surprising feature that, up to first order terms, it takes $n \log n$ throws for the first cover, each subsequent cover requiring only an additional *nloglogn* throws. To obtain a better understanding of this phenomenon, Flatto (1982) derived limiting theorems for $N'_{n,m}$, $N''_{n,m}$ conditioned on $N_{n,m}$, where $N'_{n,m}$, $N''_{n,m}$ are respectively defined to be the number of urns containing precisely m balls upon completion of the $N_{n,m}$-th throw, and the number of throws past the $N_{n,m}$-th one required to obtain at least one more ball in each of $N'_{n,m}$ urns. Thus, $N_{n,m+1} = N_{n,m} + N''_{n,m}$ for all n, m.

Flatto (1982) established that, given $N_{n,m} = [n \log n + (m-1) n \log\log n + nx]$, $[x]$ denoting the largest integer contained in x,

$$N'_{n,m,r} \sim e^{-x}(logn)^{r-m+1}/r! \text{ in probability }, \tag{14.52}$$

where $N'_{n,m,r}$, $r \geq m$, equal the number of urns containing exactly r balls upon completion of the $N_{n,m}$-th throw. More precisely, let $N = N(n, m, x)$ balls be thrown into n urns. The probability of hitting a specific urn is $1/n$. As $n \to \infty$, the number of balls in a given urn becomes Poisson distributed with parameter $\lambda = N/n$. Hence, given $N_{n,m} = N(n, m, x)$, the number of urns with r balls should be $\sim e^{-x}(\log n)^{r-m+1}/r!$. Flatto (1982) also proved that $N''_{n,m} \sim n \log\log n$ in probability, and established the following result:

> Let $1 \leq k \leq n$, $-a \leq x \leq a$, where $a > 0$. Let N_n^k be the number of throws necessary to obtain at least one ball in each of the urns $1, 2, \ldots, k$, the balls being thrown independently and uniformly into the urns $1, 2, \ldots, n$. Then,
>
> $$\lim_{k\to\infty} |\Pr[N_n^k \leq n \log k + nx] - e^{-e^{-x}}| = 0 \text{ uniformly in } n \text{ and } x\ . \tag{14.53}$$

Finally, Flatto (1982) also provided a heuristic explanation of the Erdös-Renyi (1961) result as follows: Let $m > 1$. Since $N'_{n,1,m-1}$ will be much larger than $N'_{n,1,r}, 1 \leq r < m-1, N_{n,m}$ should roughly equal $N_{n,1} + N''_{n,1,m-1}$, where $N''_{n,1,m-1}$ is the number of throws past the $N_{n,1}$-th one required to obtain at least one addtional ball in each of the $N'_{n,1,m-1}$ urns. Then, upon using the asymptotic result for $N'_{n,m,r}$ presented above, Flatto (1982) in fact arrived at a generalization of the Erdös-Renyi theorem.

Flatto's limit theorem clarified and provided an insight into the limiting behavior of classical sequential occupancy models.

14.9 Limit Theorems for Infinite Urn Models

Infinite urn models were studied in the classical paper of Karlin (1967) where in covergence in distribution to the standard normal was established. Among recent results generalizing Karlin's contributions, Dutko's (1989) results seem to be noteworthy.

n balls are placed independently in an **infinite** set of urns and each ball has probability $p_k > 0$ of being assigned to the kth urn, $k = 2, 3, \ldots$. It is assumed, without loss of any generality, that the urns are arranged in decreasing order so that $p_k \geq p_{k+1}$ and $\sum_{k=1}^{\infty} p_k = 1$. Let $X_{n,k}$ be the number of balls in the kth urn after n throws. If the number of throws is a Poisson random variable with mean n denoted by $N(n)$, then $X_{N(n),k}$ is the number of balls in the kth urn after $N(n)$ throws. Note that the random variables $\{X_{n,k}\}$ where the sample size n is fixed and k varies are not independent, while $\{X_{N(n),k}\}$, $k = 1, 2, 3, \ldots$ are mutually independent Poisson random variables with respective means $\{np_k\}$. Finally, let $Z_n = \sum_{k=1}^{\infty} \varphi(X_{n,k})$, where $\varphi(u)$ equals 0 if $u = 0$ and equals 1 if $u > 0$, and analogously $Z_{N(n)} = \sum_{k=1}^{\infty} \varphi(X_{N(n),k})$. That is, Z_n $(Z_{N(n)})$ is the number of occupied urns after n $(N(n))$ balls have been thrown. Let us denote $E[Z_n] = \mu_n$, $\mathrm{Var}(Z_n) = \sigma_n^2$; similarly, $E[Z_{N(n)}] = \mu(n)$ and $\mathrm{Var}(Z_{N(n)}) = \sigma^2(n)$.

One of Karlin's (1967) results is that $\frac{Z_n - \mu_n}{b_n} \to N(0, 1)$ in distribution for all $\{p_k\} \in A = \{\{p_k\} | \alpha(x) = x^{\gamma} L(x), 0 < \gamma < 1\}$, where $\alpha(x) = \max\{k | p_k \geq \frac{1}{x}\}$ and $\frac{L(cx)}{L(x)} \to 1$ as $x \to \infty$ for any fixed $c > 0$ (that is, $\alpha(x)$ is of regular variation in the Karamata sense) and $b_n \to \infty$ and $b_n \sim \sigma_n$ as $n \to \infty$; see Karlin (1967, p. 386) for an explicit formula.

Dutko succeeded to show that $\frac{Z_n - \mu_n}{\sigma(n)} \to N(0, 1)$ in distribution for all

$$\{p_k\} \in B = \{\{p_k\} \mid \lim_{n \to \infty} \sigma^2(n) = \infty\}.$$

The class A is wider than B. Roughly speaking, the class A contains sequences $\{p_k\}$ where $\sigma^2(n) \to \infty$ for $n \to \infty$, but the smoothness condition $\alpha(x) = x^{\gamma} L(x)$, $0 < \gamma < 1$, does not hold. The fact that for $p_k \in B, \sigma^2(n) \to \infty$ as $n \to \infty$ had been established earlier by Karlin (1967). Normalization by means of $\sigma^2(n)$ instead of by σ_n^2 is done in order to take advantage of the fact that $X_{N(n),k}$ (instead of $X_{n,k}$) are independent. Dutko's (1989) proof involves two steps: First one shows that $\frac{Z_{N(n)} - \mu(n)}{\sigma(n)} \to N(0, 1)$ in probability as $n \to \infty$ for all $\{p_k\} \in A$ essentially via Lindeberg's criterion for convergence, and the second step shows that $lim_{n \to \infty}[\mu_n - \mu(n)] = 0$ for appropriately defined sequences $\{p_k\}$.

14.10 Urn Models with Indistinguishable Balls (Bose-Einstein Statistics)

Indira and Menon (1988a,b) and Menon and Indira (1990) investigated the limiting distribution of the number of cells containing k balls each in the case when n indistinguishable balls are distributed into m cells. Their work is a continuation of the unpublished work of Park (1976) as described in Johnson and Kotz (1977); they have used in part the general methodology of proving limit theorems of this kind developed by Holst (1979) (described earlier) and an elementary method of Menon (1973) relating to the maximum of Stirling numbers of the second kind.

Two closely related models have been investigated:

Model I: The n indistinguishable balls are distributed into m cells such that all the $\binom{m+n-1}{m-1}$ possible distinguishable arrangements are equi-probable. Let $M_k \equiv M_k(n, m)$ denote the number of cells containing exactly k balls each $(k \geq 0)$ (the occupancy variable).

Model II: The n distinguishable balls are distributed into m cells such that no cell remains empty and all the $\binom{n-1}{m-1}$ possible distinguishable arrangements are equally likely. Let $M_k^* \equiv M_k^*(n, m)$ denote the number of cells containing exactly k balls each $(k \geq 1)$.

Evidently, $M_k^*(n, m) = M_{k-1}(n - m, n)$ have the same probability distributions. Let $\alpha = n/m$, where we regard n as a function of m.

These models are referred to in the classical literature [for example, see Feller (1968) and Johnson and Kotz (1977)] as *Bose-Einstein statistics*. Indira and Menon (1988a) have shown that the possible limit laws for the occupancy variables M_k and M_k^* (as $m \to \infty$) are normal, Poisson or degenerate. More precisely, the possible limit forms for the sequence $\alpha = \alpha(m)$ as $m \to \infty$ are $0, \infty$ or $\alpha_0 \in (0, \infty)$ in Model I, and $1, \infty$ or $\alpha_0 \in (1, \infty)$ in Model II. If $\alpha \to \alpha_0$, the variables M_k and M_k^* are asymptotically normal and in the other two cases, the variables are either Poisson or degenerate in the limit. A distinction between the Poisson or degenerate limit for Model I is given by the condition $\alpha \to 0, m\alpha^2 = n^2/m \to 0$ yielding the degenerate distribution at 0, while the condition $\alpha \to 0$, $m\alpha^2 \to \lambda$ results in a Poisson limit law; and analogously, the condition $\alpha \to \infty$, $m/\alpha \to 0$ results in a degenerate distribution, while the condition $\alpha \to \infty$, $m/\alpha \to \lambda$ results in a Poisson distribution. Similar results are obtained for M_k^*.

Indira and Menon (1988b) have also investigated in great detail the local central limit theorem and expansions related to this central limit theorem for the random variable M_k and the random variable N_k, where $N_k = N_k(n,m)$ denotes the number of urns containing at least k balls ($k \geq 2$). Asymptotic normality of M_k and N_k follow from the general result of Holst (1979) described earlier. For example, they have shown that, as $m \to \infty$ and $n/m \to \alpha_0 > 0$,

$$\Pr[M_k \leq ma_m + x\sqrt{m}B_m] = \Phi(x) + R, \tag{14.54}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function, $R = O((1 + x^2)/\sqrt{m})$ and $\frac{x^3}{\sqrt{m}} \to 0$ as $m \to \infty$. Here, a_m and B_m are functions of $\alpha = n/m$ and k which are given by

$$a_m = \frac{\alpha^k}{(1+\alpha)^{k+1}}, \; B_m^2 = a_m\left\{1 - \left\{1 + \frac{(k-\alpha)^2}{\alpha(1+\alpha)}\right\} a_m\right\}, \tag{14.55}$$

with $ma_m + x\sqrt{m}B_m$ being an integer. As a corollarly, they also obtained the result that

$$\Pr[M_k = ma_m + x\sqrt{m}B_m] \sim \frac{1}{\sqrt{m}B_m}\phi(x), \tag{14.56}$$

where $\phi(\cdot)$ denotes the standard normal density function. They have also recommended a continuity correction of the form $\frac{h}{2}$ and provided tables revealing that the approximations they proposed are quite satisfactory even for $M_2(120, 30)$ in which case the mean and variance 3.84 and 3.24, respectively (almost the Poisson case as the mean and variance are almost equal).

Menon and Indira (1990) also examined the approximation of $\Pr[M_k = r]$ and $Pr[M - k \leq r]$ by the corresponding quantities obtained from the limiting Poisson distribution. One of their useful results is as follows:

> Let $n \to \infty, m \to \infty$, $\alpha = n/m \to \infty$, and $\lambda^2/m \to 0$, where $\lambda = m^2/(m+n)$; then, their Poisson approximation to M_0 is given by
>
> $$\Pr[M_0 = r] - p(r;\lambda) = \frac{1}{m}\{\lambda - (r-\lambda)^2\} + O(\frac{\lambda^4}{m^2})p(r;\lambda), \tag{14.57}$$
>
> where $p(r;\lambda)$ denotes the Poisson (with mean λ) probability mass function at r. Prasad and Menon (1985), realizing that there is a Poisson limit for the distribution of M_0 when $\frac{n}{m} \to \infty$ in the case of the randomized occupancy model, have discussed some approximations to the distribution of M_0 when m is large but finite.

Similar results for M_k^* have also been developed by these authors.

For example, the following two tables present a few numerical comparisons for the case when $m = 30$, $n = 300$, $\alpha = 10$, $\lambda = 2.5$:

r	$\Pr[M_0 = r]$	$p(r;\lambda)$
0	0.071	0.082
3	0.231	0.214
6	0.018	0.028

r	$\Pr[M_2 = r]$	$p(r;\lambda)$
0	0.072	0.083
3	0.228	0.213
6	0.021	0.027

14.11 Ewens Sampling Formula and Coalescent Urn Models

One of the most famous and celebrated applications of urn models that has emerged during the last two decades is the *Ewens Sampling Formula*, due to Ewens (1972). Indeed, the qualifier "celebrated" is almost universally used whenever the term Ewens sampling formula is mentioned. Ewens laid the foundation for the use of urn models in population genetics in his treatment of the infinite-alleles model at equilibrium [Fisher, Corbet and Williams (1943)]. In this section, we shall concentrate on a very small sample of representative papers that appeared during the last two decades rather than trying to exhaust the variety of very ingenious and mathematically intricate contributions that have appeared in the literature. As a matter of fact, the literature on this topic is overwhelming and interested readers may refer to Chapter 41 of Johnson, Kotz and Balakrishnan (1997) for a concise review.

Ewens sampling formula was developed by Ewens (1972) and has been described by the originator in the book by Johnson and Kotz (1977). It is associated with a partition structure of population genetics. Stated in biological terms, it asserts that [see, for example, Donnelly (1986)]:

> If n gametes in a sample from a biological population are classified according to the gene at a particular locus, then under suitable conditions the probability that a_1 alleles will be represented once, a_2 twice, etc., is given by
>
> $$P_n(a_1, a_2, \ldots, a_n) = \frac{n!}{\theta(\theta+1)\ldots(\theta+n-1)} \prod_{j=1}^{n} \left(\frac{\theta^{a_j}}{j^{a_j} a_j!} \right) \quad (14.58)$$
>
> for some parameter $\theta > 0$.

The remarkable fact about the above given Ewens sampling formula is that it has been shown to apply to a wide variety of models for reproduction provided only that the population size N is large compared to n, mutation is nonrecurrent (the infinite alleles models), all alleles are selectively neutral, and the population is in equilibrium. The parameter θ is typically related to the mutation rate.

The main result that has emerged in the recent past concerning the Ewens sampling formula is that it may be generated by a Pólya-like urn model. Specifically: At the start, the urn contains one black ball of mass $\theta > 0$. Successively balls are drawn from the urn at random (i.e., in proportion to their masses). After every draw, the selected ball is returned to the urn together with an additional ball of mass 1, whose color is the same except when the black ball has been drawn. In this case, the new ball is painted with a color not yet present in the urn. The natural numbers are used to label the colors in order of their need. [This is in essence the Hoppe-Donnelly urn model; see Hoppe (1984, 1987) and Donnelly (1986); see also Zabell (1992) for illuminating historical remarks.]

For $n \in \mathbf{N}$, let X_n be the number of the color of the n-th ball added. Then, $(X_n)_{n\in\mathbf{N}}$ is a stochastic process on $\mathbf{N}$ with $X_1 \equiv 1$ and

$$\begin{aligned}\Pr[X_{n+1}=k|X_n=x_n,\ldots,X_1=x_1] &= \frac{\ell_k(n)}{\theta+n},\ k\le \ell(n)\\ &= \frac{\theta}{\theta+n},\ k=\ell(n)+1\\ &= 0,\ \text{otherwise,}\end{aligned}$$

for every realization $x_1,\ldots,x_n$ of $X_1,\ldots,X_n$, $n\in\mathbf{N}$, and $k\in\mathbf{N}$, where

$$\ell(n)=\max\{x_i|1\le i\le n\}\quad\text{and}\quad \ell_k(n)=\#\{i|x_i=k,1\le i\le n\}.$$

The colors of the balls in the urn after n trials induce a random partition

$$\Pi_n=(c_1,\ldots,c_n), \tag{14.59}$$

where c_i is the number of colors belonging to exactly i balls, $1\le i\le n$. Then, $(\Pi_n)_{n\in\mathbf{N}}$ is Markovian, and every Π_n is distributed according to the Ewens sampling formula, i.e.,

$$\Pr[\Pi_n=\pi]=\frac{n!}{[\theta]_n}\prod_{j=1}^{n}\frac{\theta^{a_j}}{j^{a_j}a_j!} \tag{14.60}$$

for all partitions $\pi=(a_1,\ldots,a_n)$ of m, where $[\theta]_n=\theta(\theta+1)\ldots(\theta+n-1)$, $n\in\mathbf{N}$. We thus equate the labelling of balls in the urn to the partition by age of alleles in the sample. Moreover, as has been shown by Donnelly (1986), the Ewens sampling formula is the **only** partition structure which can be constructed via the urn described above.

In late seventies, *partition structures* were introduced by Kingman (1977, 1978). Denote by Ω_n the set of all partitions of an integer n and $\mathbf{P}_n$ the set of all distributions on Ω_n. A sample of size n is taken from a biological population consisting of k different types, and the corresponding sample model consists of specifying a probability distribution $P_n\in\mathbf{P}_n$ for the observed partition of n. Since there is nothing special about a particular value of the sample size n, a sequence of *consistent* family of distributions $(P_1,P_2,\ldots)$ is known as a *partition structure*, introduced by Kingman (1977, 1978). The main result of Kingman is that all partition structures may be constructed via a generalized "paintbox". Specifically, with each partition structure we associate a unique *representing measure* μ on the space ∇_0 of sequences $x=(x_n;\ n=0,1,2,\ldots)$ satisfying

$$x_0\ge 0,\ x_n\ge x_{n+1}\ge 0\ (n\ge 1),\ \sum_{n=0}^{\infty}x_n=1. \tag{14.61}$$

For $x \in \nabla_0$ and let ξ_r ($r = 1, 2, \ldots$) be independent random variables with distributions

$$\Pr[\xi_r = n] = x_n \ (n \geq 1), \ \Pr[\xi_r = -r] = x_0. \tag{14.62}$$

For $n \geq 1$, $\pi = (a_1, a_2, \ldots, a_n) \in \Omega_n$, define $\varphi_\pi^n(x)$ to be the probability that among the values of the random variables $\xi_1, \xi_2, \ldots, \xi_n$, exactly a_j integers are represented j times ($j = 1, 2, \ldots, n$). We may then write

$$P_n(\pi) = \int_{\nabla_0} \varphi_\pi^n \, d\mu. \tag{14.63}$$

All known partition structures of biological interest have representing measures concentrated on the subset ∇ of ∇_0 corresponding to those sequences described in (14.63) with $x_0 = 0$. Kingman (1977) termed such partition structures *representable.*

Now, as Donnelly (1986) and Hoppe (1987) have shown, the Pólya urn model generating the Ewens sampling formula parallels the construction of Kingman using a Poisson-Dirichlet distribution "paintbox". Specifically, the paintbox is constructed as follows: Consider a representable partition structure. Choose a point $x = (x_1, x_2, x_3, \ldots) \in \nabla$ according to the distribution μ and imagine a hypothetical infinite population in which one type has frequency x_1, a second type has frequency x_2, and so on. When averaged over the distribution μ of x, the partition of n induced by a sample of size n taken from the infinite population will have distribution P_n. Equivalently, one can think of coloring balls by dipping a paintbrush at random into a paintbox of which a fraction x_1 of the paint is of one color, x_2 of a second color, and so on. The partition induced by the colors of the first n balls will have distribution P_n (when averaged over the distribution μ of the colors).

The Poisson-Dirichlet distribution can be obtained from a symmetric Dirichlet $D(\alpha; k)$ (see below for its definition) as follows: Arrange the population frequencies in decreasing order as $P_{(1)} \geq P_{(2)} \geq \ldots P_{(k)}$. It can be shown [see, for example, Kingman (1975)] that for each fixed j, $(P_1^*, P_2^*, \ldots, P_j^*)$ as $k \to \infty$, $\alpha \to 0$, $k\alpha \to \theta$, and this vector is the j-th joint marginal of a random probability $\mathbf{P}^*$, termed by Kingman (1975) as the Poisson-Dirichlet distribution. It was actually Watterson (1976) who first associated the Ewens sampling formula with the Poisson-Dirichlet distribution. He actually conjectured that the Ewens sampling formula could be derived by directly sampling from the population described by $D(\alpha, k)$—asymmetric Dirichlet distribution with density

$$\frac{\Gamma(k\alpha)}{\Gamma^k(\alpha)} \left\{1 - \sum_{i=1}^{k-1} p_i\right\}^{\alpha - 1} \prod_{i=1}^{k-1} p_i^{\alpha_i - 1}, \tag{14.64}$$

where $\Gamma(\cdot)$ is the complete gamma function—over the simplex $\{(p_1, \ldots, p_{k-1}) : p_i \geq 0 \text{ and } \sum_{i=1}^{k-1} p_i \leq 1\}$ rather than proceeding indirectly by letting $k \to \infty$. A proof of this conjecture was provided subsequently by Kingman (1977).

A related urn model is an urn model with coalescent, which was introduced by Kingman (1982a,b,c) motivated by mimicing the procedure of tracing a sample's ancestry backwards in time and noting the appearance of common ancestors or new mutants (until one reaches a single common ancestor). An extension of this model has been provided recently by Branson (1994) as follows:

> Suppose an urn originally contains m balls of the same size but each of a different color (not black) representing a sample of size m taken from the population. We remove the balls from the urn in m stages, which we label successively as stage m, stage $m-1$, $\ldots$, stage 1 (so that stage i commences with i balls in the urn). In stage m, we remove one ball chosen at random and designate it an "offspring" ball. We replace it with a black ball whose size is such that the relative probabilities of picking the black ball or any particular colored ball are in the ratio $\theta : 1$. We now remove a second ball: if it is black, stage m ends (the interpretation will be that the "offspring" was a mutant); (if it is colored, we designate it a "parent" ball, replace it in the urn, by removing the black interpreted as a coalescence). This completes the initial stage m.
>
> We continue in a similar fashion. In stage i, the urn contains i balls, and the probability that any particular ball is chosen as "offspring", followed by the black ball is
>
> $$\frac{1}{i} \times \frac{\theta}{\theta + i - 1} . \tag{14.65}$$
>
> On the other hand, the probability that any particular ball is chosen as "offsping" and any other particular ball as "parent" is
>
> $$\frac{1}{i} \times \frac{1}{\theta + i - 1} . \tag{14.66}$$
>
> If any stage results in a coalescence, we recolor with the "parent's" color not only the "offspring" at that stage but also all previously removed balls of the same color (which are "offspring" of "offspring"...). For example, at stage m, we may remove the blue ball as offspring and the red ball as parent, so we replace the red ball but change the color of the offspring ball from blue to red. At a later stage, we may remove the red ball as offspring and the green ball as parent. We replace the green ball but change to green the color of both balls that are now red (that is not only green's "child" but also its "grandchild" originally colored blue). Thus, at any stage, we have painted with the same color a parent (in the urn) and all its progeny of succeeding generations (outside the urn).
>
> After stage 1, the urn will be empty. Outside the urn there will be m balls of ℓ (say) different colors if the black ball was chosen on

ℓ occasions. The balls carrying the same color represent a single family descended from an originating mutant ancestor.

While Hoppe-Donnelly model mimics the drawing of a sample at some fixed time from a Poisson-Dirichlet distribution and also is equivalent to the reversed jump chain of the coalescent, Branson's (1992) model directly mimics the coalescent itself (rather than its time reversal). Further, it provides a computational tool in those cases where, in tracing backwards through time, the coalescent does not run its full course and so there are some old "genes" remaining—viz., when the urn is being only partly emptied. Branson (1992) also provided a direct proof for the relationship between the Ewens sampling formula and the Poisson-Dirichlet distribution.

In an interesting paper, Trieb (1992) discussed the concept of n-coalescent with mutation (which is a continuous-time Markov chain taking values in the set E_n of all equivalence relations on $\{1, 2, \ldots, n\}$; two individuals are in the same equivalence class at time t if and only if they have a common ancestor not more than t time units in the past and if their lines of descent go back to this common ancestor without intervening mutations) and its connection to the Hoppe-Donnelly urn and the Ewens sampling formula. He found some inconsistencies in Hoppe's (1987) arguments relating to genealogy of the coalescent with mutation.

14.12 Reinforcement-Depletion (Compartmental) Urn Models

A study conducted by Bernard of the Health and Safety Division at the Oak Ridge National Laboratory [Bernard (1977)] generated a flood of papers and, as a result, provided a number of useful and far-reaching generalizations from theoretical as well as applied points of view. It involved a single compartment model with bulk arrivals and departures, known in the literature as *reinforcement-depletion urn model* or *replenishment-depletion urn model.* It was originally introduced in connection with radioactive atoms and stable atoms in humans, to model, for example, the uptake of radioactive iodine by the thyroid in humans.

The original basic model of Bernard (1977) is as follows: Initially, there are b black balls and w white balls in the urn. At each stage, a fixed number (r) of black balls is added to the urn (reinforcement) and $(b + w + r)$ balls are uniformly mixed. Next, a random sample of r balls is then removed (depletion) from the urn. The main interest centers on the number of white balls at each stage and the time (stage) when all the white balls have first been removed. Note that we are considering here a constant reinforcement-depletion size. The reinforcement-depletion cycle mimics a birth and death process, and in the simplest case the depletion always equals the reinforcement cycle by cycle. As it was pointed out by Bowman, Shenton and Bernard (1985), in urn models related to the daily ingestion of radioiodine in humans, reinforcements and depletions are of order 10^{12} or so, and Shenton (1983) has taken reinforcement at the j-th cycle to be $r_j = 10^{21} + K(j-1)10^{17}$, $j = 1, 2, \ldots, 7300$ with $K \approx 2$. These sizes make it extremely difficult to connect the possible configurations at a cycle to those at a previous cycle, which makes the recursive approach an attractive one. In models for the uptake of radioactive iodine atoms by the thyroids of humans at different ages, the uptake of iodine is usually taken to be about twice the depletion, and the problem when the daily depletion is significantly less than the reinforcement is of interest (which leads to a natural generalization of Bernard's urn model). Finally, we note that reinforcement-depletion are sometimes assumed to occur in discrete time (as in the case of Bernard's original model) while at other times at the instances of a time-homogeneous Poisson process [see, for example, Donnelly and Whitt (1989).]

Denoting by W_n the number of white balls remaining in the urn after n stages, it is relatively easy to derive simple expressions for the mean and variance of the variable W_n. In fact,

$$\begin{aligned} E[W_n] &= w(1-\rho)^n, \\ \mathrm{Var}(W_n) &= w\{(1-\rho)^n - (1-\alpha\rho)^n\} + w^2\{(1-\alpha\rho)^n - (1-\rho)^{2n}\}, \end{aligned}$$

where

$$\rho = \frac{r}{w+b+r} \quad \text{and} \quad \alpha = 1 + \frac{w+b}{w+b+r-1};$$

see, for example, Leitnaker and Purdue (1985) and Donnelly and Whitt (1989). The distribution of W_n has been derived by Shenton (1981) and Leitnaker and Purdue (1985) and is given by

$$\begin{aligned} \Pr[W_n = k] &= \binom{w}{k} \sum_{i=0}^{w-k} (-1)^i \binom{w-k}{i} \\ &\quad \times \left\{ \binom{b+w+r-k-i}{r} \Big/ \binom{b+w+r}{r} \right\}^n, \qquad (14.67) \end{aligned}$$

for $k = 0, 1, \ldots, w$. Shenton (1981) used generating function approach to arrive at the form of the distribution in (14.69) while Leitnaker and Purdue (1985) used a more elementary and direct approach by utilizing an indicator function

$$\begin{aligned} I_i(n) &= 1 \text{ if ball } i \text{ is in the urn after the } n\text{-th R-D cycle} \\ &= 0 \text{ otherwise} \qquad (14.68) \end{aligned}$$

in order to derive the distribution in (14.69). Analogous, but more complicated expressions, are available in the case of random independent and identically distributed reinforcements; see, for example, Leitnaker and Purdue (1985).

In a more general case, as it was mentioned earlier, reinforcement sizes will depend on the stage n; let these sizes be denoted by $\{R_n : n \geq 1\}$. Assume that the reinforcement size vectors $(R_1, \ldots, R_n)$ have independent marginals and means $(r_1, \ldots, r_n)$. It can then be shown that $E(W_n^m)$ is minimized by using the deterministic reinforcement sizes $(r_1, \ldots, r_n)$. Moreover, among all the deterministic reinforcement sequences $(r_1, \ldots, r_n)$ with $r_1 + \cdots + r_n = nr$ for some integer r, $E(W_n^m)$ is is minimized by using the constant deterministic sequence $(\underline{r}, \ldots, \underline{r})$; see Donnelly and Whitt (1989). By allowing r, the number of particles being transferred, to be random with

$$\Pr[R_i = r_j] = p_j \quad \text{where} \quad \sum_j p_j = 1,$$

Leitnaker and Purdue (1985) showed that

$$\begin{aligned}\Pr[W_n = k] &= \binom{w}{k} \sum_{i=0}^{k-1} (-1)^i \binom{w-k}{i} \\ &\quad \times \left\{ \frac{(b+w)!}{(b+w-k-i)!} \right\}^n E\left\{ \frac{(b+w+R_1-k-i)!}{b+w+R_1)!} \right\}. \end{aligned} \tag{14.69}$$

Assuming that the sequence $\{R_n : n \geq 1\}$ are i.i.d. random variables, the following results have been obtained:

$$\begin{aligned} E[W_n] &= w \left\{ \sum_{i=0}^{\infty} (-1)^i \frac{E[R_1^i]}{(b+w)^i} \right\}^n, \\ \mathrm{Var}(W_n) &= V_1 + V_2 - V_3, \end{aligned}$$

where

$$\begin{aligned} V_1 &= w \left\{ \sum_{i=0}^{\infty} (-1)^i \frac{E[R_1^i]}{(b+w)^i} \right\}^n, \\ V_2 &= w(w-1) \left\{ \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} (-1)^{i+k} \frac{E[R_1^{i+k}]}{(b+w)^i (b+w-1)^k} \right\}^n, \\ V_3 &= w^2 \left\{ \sum_{i=0}^{\infty} (-1)^i \frac{E[R_1^i]}{(b+w)^i} \right\}^{2n}. \end{aligned}$$

The w^2 term in the above expression for the variance of W_n can be made large by increasing the variance of R_i. It needs to be mentioned here that Purdue (1981) had discussed earlier a model under which the times at which the R-D cycles occur are determined by a Poisson process with rate λ.

Some asymptotic results are very revealing here. If the distribution of R_1 is independent of w and if $w \to \infty$ so that $\frac{w}{b+w} \to p > 0$, then $E[W_n] - w \sim -pnE[R_1]$.

In many applications, the original Bernard model underrepresents the variability in the observed data; hence, attempts were made to modify the model to increase the variability as measured by the squared coefficient of variation. In the model with $\{R_n : n \geq 1\}$ being i.i.d. described above, the squared coefficient of variation of the number of white balls remaining is asymptotically negligible as $w \to \infty$, while the squared coefficient of variation of the number of white balls removed $\left(\frac{Var(w-W_n)}{\{E[w-W_n]\}^2}\right)$ is not so small for any fixed n. As a matter of fact, the number of white balls removed $(w - W_n)$ approaches the number of successes in a random number of Bernoulli trials with parameter p. Note that, in the cases just discussed, the distribution of R_1 remains fixed as $w \to \infty$. If, on the other hand, the distribution of R_1 grows with w—that is, if $R_j = wX_j$ where X_j $(j \geq 1)$ are i.i.d. integer-valued random variables independent of w—a significantly greater variability is obtained. For more details, one may refer to Donnelly and Whitt (1989).

Ball and Donnelly (1988) emphasized that in the original Bernard model with a fixed r, W_n can be expressed as

$$W_n = \sum_{i=1}^{w} \chi_i(n), \quad n = 0, 1, \ldots, \tag{14.70}$$

where $\chi_i(n)$ is just an indicator variable taking the value 1 if white ball i is present in the urn after state n and the value 0 otherwise (compare this with $I_i(n)$ defined earlier). Although the variability of the model will be increasing with $Cov(\chi_1(n), \chi_2(n))$, it is intuitively clear and also straightforward to verify in this case that $\text{Cov}(\chi_1(n), \chi_2(n))$ is negative and thus the model displays less variation than a model with "independent" balls.

In the case when the size of the reinforcement-depletion at the n-th stage is given by a random variable R_n, with $R_1, R_2, \ldots$ being an i.i.d. sequence,

$$\text{Cov}(\chi_1(n), \chi_2(n)) = \left\{E\left[\frac{a(a-1)}{(a+R_1)(a+R_1-1)}\right]\right\}^n - \left\{E\left[\frac{a}{a+R_1}\right]\right\}^{2n}, \tag{14.71}$$

where $a = b + w$; in this case, the covariance may be positive (for example, when R has distribution concentrated at zero and some large integer).

Shenton (1981, 1983), followed by a series of papers by Bowman and Shenton (1986a,b) and by Shenton and Bowman (1985, 1996a,b), have all provided a deep and comprehensive discussion of the distributional properties of the Bernard R-D urn. Of particular interest amongst these is the paper by Shenton and Bowman (1985) in which the authors have discussed the case of an urn containing balls of three colors (red, white and blue). At the j-th cycle, the reinforcement consists of n_j balls: r_j red, w_j white and b_j blue. The urn then randomly depletes of d_j balls (if $d_j > 0$ exceeds the number of balls in the urn, the cycles cease). Note that, in this case, the size of each total reinforcement is fixed.

Formulas for the multivariate factorial moment generating function (*fmgf*) for the state of the urn at the mth cycle were one of the main foci of Shenton and Bowman's investigations. Let us first consider the elementary joint probability function $p(x_1, x_2)$ for depletion at the first cycle:

$$p(x_1, x_2) = \frac{\binom{r_1}{x_1}\binom{w_1}{x_2}\binom{n_1 - r_1 - w_1}{d_1 - x_1 - x_2}}{\binom{n_1}{d_1}}. \tag{14.72}$$

Now, the *fmgf* for α_1 red balls and α_2 white balls remaining in the urn after the first cycle is

$$f_1(\alpha_1, \alpha_2) = \sum_{x_1}\sum_{x_2} \alpha_1^{r_1 - x_1}\alpha_2^{w_1 - x_2} p(x_1, x_2), \tag{14.73}$$

and

$$\frac{\partial^{i+j}}{\partial\alpha_1^i \partial\alpha_2^j} f_1(\alpha_1, \alpha_2) \mid_{\alpha_1 = \alpha_2 = 0} = \frac{r_1^{(i)} r_2^{(j)} (n_1 - i - j)^{(d_1 - i - j)}}{n_1^{(d_1)}}. \tag{14.74}$$

An alternate useful expression is

$$f_1(\alpha_1, \alpha_2) = \left(1 + \frac{\alpha_1}{E_1}\right)^{r_1} \left(1 + \frac{\alpha_2}{E_2}\right) \frac{Y_1^{(d_1)}}{t_1^{d_1}} \mid_{Y_1 = t_1} \tag{14.75}$$

with $t_1 = n_1 = r_1 + w_1 + b_1$ and $E_i \equiv E_{Y_i} \equiv Y_i + 1$ (the shift operator), and $x^{(a)} = x(x-1)\cdots(x-a+1)$ (the a-th descending factorial of x).

An useful but rather complicated recurrence relation (in an obvious notation) for the *fmgf* at the j-th cycle is

$$f_j(\alpha_1, \alpha_2) = I_j\left(\frac{\alpha_1}{E_j}, \frac{\alpha_2}{E_j}\right) f_{j-1}\left(\frac{\alpha_1}{E_j}, \frac{\alpha_2}{E_j}\right) \frac{Y_j^{(d_j)}}{t_j^{(d_j)}} \mid_{Y_j = t_j}, \tag{14.76}$$

with $t_j = \sum_{k=1}^{j} n_k - \sum_{k=1}^{j-1} d_k$ ($d_0 = 0$) and $f_0(\alpha_1, \alpha_2) = E[(1+\alpha_1)^{r_0}(1+\alpha_2)^{w_0}]$. Also,

$$I_j(\alpha_1, \alpha_2) = E[(1+\alpha_1)^{r_j}(1+\alpha_2)^{w_j}] \ (j = 1, 2, \ldots)$$

is the reinforcement *fmgf* at the j-th cycle, α_1, α_2 are the numbers of red and white balls respectively (the third color is omitted since $r_j + w_j + b_j = n_j$ and $I_j(\alpha_1, \alpha_2)$ will have n_j as the highest coefficient in α_1 and α_2).

Bowman and Shenton (1986a) observed that in the two-color model when the urn initially has σ red balls and ω white balls and from which at the jth cycle only white balls receive increment w_j and d_j balls are randomly removed and the contiguity condition $w_j = d_{j-1} + d_j$ ($j = 2, 3, \ldots, m$) is present, then the distribution of red balls in this model is hypergeometric. This may also serve as an approximation to the "general" reinforcement-depletion two-color urn model and especially so when the reinforcements of white balls are about double the random depletions. Approximations of this kind suggest that the half-life distribution of red balls is nearly normal $N\left(\frac{\sigma}{2}, \frac{\sigma}{8}\right)$ provided that $M = \omega + w_1 - d_1$ is fixed and the initial number is large.

Shenton and Bowman (1996b) discussed the two-color case when replenishments are positive random variables with given factorial moments. This results in a variety of discrete distributions including the hypergeometric. In their most recent paper, Shenton and Bowman (1996b) investigated the two-color urn in equilibrium: The $R - D$ phase may result in a closed system; thus $R = D$ at a cycle, so that the total number of balls in the system at the completion of any cycle is constant. This is the urn in equilibrium ultimately. Under what circumstances do simple solutions exist? Initially, there are 0 balls of color C_1, and k balls of color C_2. At the jth cycle, the replenishments are p balls of C_1, and q balls of C_2, with depletion $p+q$. In the equilibrium urn ($p = q$), as k (the initial number of C_2 balls) becomes large, the following numerical values were obtained by Shenton and Bowman (1996b) for the mean, standard deviation and the coefficient of kurtosis (the coefficient of skewness is 0 in all cases).

k	p	Mean	Std. Dev.	Kurt.
16	3	8	1.5422	2.9432
	5	8	1.5927	2.9416
	25	8	1.8053	2.9167
	50	8	1.8820	2.9021
50	3	25	2.5820	2.9811
	5	25	2.6231	2.9814
	25	25	2.8964	2.9800
	50	25	3.0695	2.9766

Shenton and Bowman (1996b) also observed a rather "curious similarity" of the probability mass function of the number of balls of color C_1 given by

$$\Pr[N_1 = s] = \binom{k}{s}(k+p)^{(s)}(k+p)^{(k-s)}/(2k+2p)^{(k)} \quad (s = 0, 1, \ldots, k) \quad (14.77)$$

corresponding to the symmetric urn distribution in equilibrium, with that of the Pólya-Eggenberger probability mass function given by

$$\Pr[X = x] = \binom{n}{x}\alpha^{[x]}\beta^{[n-x]}/(\alpha+\beta)^{[n]},$$

where $\alpha^{[x]} = \alpha(\alpha+1)\cdots(\alpha+x-1)$. This is so eventhough the basic structures of the two distributions are quite different.

A summary of their main findings is as follows: A new distribution in equilibrium arises when there are two colors, each reinforced by the same amount with $R = D$. With $R = D$, there is asymptotic normality ($k > p, k \to \infty$) and asymptotic binomiality ($p = q \to \infty, k$ fixed). In general, the symmetric urn is such that the distribution of balls of color C_1 has skewness zero and kurtosis a little less than three (but being very close to three suggests close proximity to the normal distribution).

In concluding this section, we mention about the relationship with the *Consul urn models* discussed by Consul (1974), Consul and Mittal (1975, 1977) and Famoye and Consul (1989) in which a strategy (or decision) of individuals alter the structure of the urn model and the ultimate probability process. The most recent paper by Consul (1995) deals with the two-color case in this framework. This class of urns generate numerous discrete distributions as well. Mishra and Sen (1987) modified Consul's urn model in such a way that, in the resulting quasi-binomial distribution, the probability of success may even decrease with the number of successes. Such a model may be applied to a situation where a particular type of species (say, S_1) of living beings migrates to a place where they reside with another type of species (say, S_2). Naturally, S_2 would protest against the migrating group while other S_1 species living already there might welcome.

14.13 Urn Models for Interpretation of Mathematical and Probabilistic Concepts and Engineering and Statistical Applications

We describe five urn models in this short section which, of course, is a very small sample of papers dealing with various applications of urn models in interpretation of basic mathematical and probabilistic concepts and also in engineering and statistical applications.

1. Zaman (1981, 1984) used urn models to describe the concept of *Markov exchangeability*. A probability of finite strings of letters is said to be Markov exchangeable if it assigns the same probability to strings which have the same initial letter and the same transition counts (for example, *abbaab*, *abbaabb*, *aabbab*, or *aababb*). Diaconis and Freedman (1980) considered the problem of expressing the extreme points of the set of Markov exchangeable probability measures. The general solution was posed as an unsolved problem, though they gave an urn model for a two letter alphabet. A solution to the general alphabet was given by Zaman (1981) in terms of the following urn model:

   Let V be a finite set and $\{U(u)\}_{u \in V}$ be a collection of urns. Each urn $U(u)$ contains a total of $\sum_{v \in V} a_{uv}$ balls, with a_{uv}

of them labeled v. Choose a fixed $X_1 \in V$ and construct a random sequence $X_1, X_2, \ldots, X_n$ by letting X_{i+1} be the label on a ball drawn from the urn $U(X_i)$. It is clear that the resulting sequence is a Markov chain if the draws are done at random with replacement. When the draws are done without replacement, after some draw X_n the urn $U(X_n)$ will be empty, so that the ball X_{n+1} cannot be drawn. The probability distribution on these finite random sequences, with some modification (viz., conditioning on the event that $n = 1 + \sum a_{uv}$, i.e. all balls from all urns are used), is an example of a Markov exchangeable distribution.

Zaman (1984) then noted that the set of Markov exchangeable measures forms a convex set and provided rather complicated and contorted construction of an urn model to represent extremal Markov exchangeable measures. These representations, in fact, reveal that Markov chains and Markov exchangeability are indeed fundamental but quite different concepts. He then concluded his paper by stating that the classical i.i.d. condition cannot be universally replaced by exchangeability (as claimed by some researchers working on foundations of probability and statistics) eventhough the i.i.d. condition is sometimes overused in cases where exchangeability is more natural.

2. Paik (1983) used an urn model construction to clarify some paradoxes associated with the concept of infinity.
For example, let us consider the following problem: Suppose, on day 1, you put in an urn 10 balls numbered 1-10 and withdraw the ball numbered n_1. On day 2, 10 more balls numbered consecutively 11-20 are put in the urn with the ball numbered n_2 being withdrawn at the same time, and so on. Then, how many balls are there in the urn in the limit? Infinitely many or none?
The paradox is that in the first experiment with $n_i = 10i$, the answer is infinitely many, while in the second experiment with $n_i = i$, the answer is none. The two different answers are disturbing since the number of balls remaining in the urn increases steadily and exactly in the same manner in each of the two situations. Paik (1983) then pointed out that the problem here is due to a singular behavior of a function at ∞ similar to sequence of functions on the positive real line

$$\begin{aligned} f_n(x) &= 1 \ \text{ if } x \in [n+1, 10n] \\ &= 0 \ \text{ otherwise .} \end{aligned} \tag{14.78}$$

Hence, the limit of the integral $\int_0^\infty f_n(x)dx$ is ∞, but the integral of the limit, $f(x) = lim_{n \to \infty} f_n(x)$, is 0.

However, as aptly displayed by Paik (1983), an urn model interpretation provides a more vivid description of the breakdown which occurs at ∞.

3. An interesting urn model with applications to modeling outliers in the context of robustness was discussed by Small (1985). This is somewhat similar to the Shur modification of the Pólya-Eggenberger urn model in order to achieve negative contagion. Small's urn model is as follows: Suppose an urn initially contains c white and c black balls. If, on trial m, a white ball is selected, set $U_m = 1$. If it was black, set $U_m = 2$. Before the $(m+1)$th selection, a ball of opposite color to that of trial m is put in the urn. The sequence $U_1, U_2, \ldots$ can be continued indefinitely. Now, let $S_{ni} = \#\{m : U_m = i, m \leq n\}$ for $i = 1, 2$. Small (1985) then showed that such urn models can be used to model outliers in data. The occurrence of an outlier in the collection of data can also be allowed to increase or decrease the probability of outliers on subsequent observations. Actually, the behavior of outliers is governed by the quantities S_{ni}.

4. An ingenious single urn model, motivated by an imperfect debugging scheme, in which the balls represent flaws in a system has been described by Siegrist (1987). To be specific, let us consider a single urn model in which the number of balls in the urn at time $n+1$ is determined as follows:

   - First, each ball in the urn at time n, independently of all others, is removed with probability $1 - p_{n+1}$ $(0 \leq p_{n+1} \leq 1)$. The balls are first removed with probability that depends on the time value.
   - Next, a random number of new balls are added to the urn, independently of the number of balls remaining after the first step.

   A similar perfect debugging model (with no new balls added and a constant p_n) was studied earlier by Siegrist (1986a,b).

   In the general model described above, Siegrist (1987) concentrated on the distribution and moments of the number of balls in the urn at time n and the asymptotic behavior as $n \to \infty$. The key equation is

$$X_{n+1} = (X_n - U_{n+1}) + V_{n+1}, \quad n = 0, 1, \ldots,$$

   where X_n is the number of balls in the urn at time n $(n = 0, 1, \ldots)$ and U_{n+1} (V_{n+1}) is the number of balls removed (added) at time $n+1$. Evidently, $X_0 = V_0$ (the number of balls in the urn initially). Moreover, the distribution of $X_n - U_{n+1}$, given X_n, is binomial with parameters X_n and p_{n+1}, and V_{n+1} is independent of $X_n - U_{n+1}$ for $n = 0, 1, \ldots$. This leads to the equation

$$F_n(s) = \prod_{k=0}^{n} G_k\left(1 - (1-s)\prod_{i=k+1}^{n} p_i\right), \tag{14.79}$$

where $F_n(s) = E[s^{X_n}]$ and $G_n(s) = E[s^{V_n}]$ are the probability generating functions of X_n and V_n, respectively.

Denoting by X_{nk} the number of balls added at time k that are still present in the urn at time n $(n \geq k)$, it is easy to show that

$$E[X_{nk}] = E[V_k] \prod_{i=k+1}^{n} p_i \tag{14.80}$$

(obviously, the conditional distribution of X_{nk}, given V_k, is binomial with parameters V_k and $\prod_{i=k+1}^{n} p_i$) and

$$\operatorname{Var}(X_{nk}) = \{\operatorname{Var}(V_k) - E[V_k]\} \prod_{i=k+1}^{n} p_i^2 + E[V_k] \prod_{i=k+1}^{n} p_i. \tag{14.81}$$

Thus, we get

$$E[X_n] = \sum_{k=0}^{n} E[V_k] \prod_{i=k+1}^{n} p_i, \tag{14.82}$$

and

$$\operatorname{Var}(X_n) = \sum_{k=0}^{n} \left\{ (\operatorname{Var}(V_k) - E[V_k]) \prod_{i=k+1}^{n} p_i^2 + E[V_k] \prod_{i=k+1}^{n} p_i \right\}. \tag{14.83}$$

Note that $X_n = \sum_{i=0}^{n} X_{ni}$. Limiting distributions of X_n depend on the limiting behavior of the sequences V_n $(n = 0, 1, \ldots)$ and p_n $(n = 1, 2, \ldots)$. Siegrist (1987) discussed the following two cases:

- $\sum_{k=0}^{\infty} V_k$ converges weakly to a proper random variable;
- $\sum_{k=0}^{\infty} V_k$ diverges weakly to ∞.

The first case corresponds to $\prod_{k=0}^{\infty} G_k(s) > 0$ for $s \in (0,1)$ while the second case corresponds to $\prod_{k=0}^{\infty} G_k(s) = 0$ for $s \in (0,1)$.

Of particular interest is the result that, when the limiting distribution of V_n is Poisson with parameter λ and $p_n \to p < 1$ with $E[V_n]$ bounded in n, the limiting distribution of X_n is Poisson with parameter $\frac{\lambda}{1-p}$. This result could be anticipated based on intuitive grounds. In addition to this result, Siegrist (1987) also discussed the following interesting finite cases:

- If V_k's are Poisson with parameters $\lambda_k > 0$ for $k = 0, 1, 2, \ldots$, then X_n is Poisson with parameter $\mu_n = \sum_{k=0}^{n} \prod_{i=k+1}^{n} p_i$ and if μ_n converges to μ, the limiting distribution of X_n is Poisson(μ);
- If V_k's are binomial with parameters M_k and α_k (with M_k being an integer and $\alpha_k \in [0,1]$) for $k = 0, 1, 2, \ldots$, there are no explicit forms

available for distributions; in this case, the results are given in terms of moment generating functions of the type

$$\prod_{k=0}^{\infty}\left\{1-(1-s)\alpha_k \prod_{i=k+1}^{\infty} p_k\right\}^{M_k}. \tag{14.84}$$

There is a need for additional studies in this direction.

The case $V_k = 0$ for $k = 1, 2, \ldots$ corresponds to a special case of a subcritical Galton-Watson process in varying environments studied earlier by Jagers (1974).

5. Urn model techniques play a natural role in the analysis of a technique, called *perfect hashing*, for organizing data in computer files. Here, considertable attention has been paid to the computation of probability of no overflows. To be specific, let us consider a traditional urn model. There are n balls to be randomly distributed into m urns, each urn having a capacity of at most b balls. Let each ball be randomly tossed into an urn so that the probability of a ball falling into a particular urn is $\frac{1}{m}$ and independent of the outcome of other tossings. If an urn already contains b balls, any subsequent ball tossed into the urn is said to *overflow*. Let $P(n, m, b)$ denote the probability of a random distribution of n balls into m urns of size b resulting in no overflows. Let X be the random variable denoting the number of balls in the urn (or urns) containing the maximum number of balls. Then, it is evident that $\Pr[X \leq b] = P(n, m, b)$. Barton and David (1959) and David and Barton (1962) discussed a combinatorial extreme-value problem in this case, while Kolchin, Sevast'yanov and Chistyakov (1978) have examined the asymptotic behavior of $P(n, m, b)$. The exact computation of this probability distribution has been discussed by Ramakrishna (1987) and Monahan (1987).

References

1. Aldous, D., Flannery, B. and Palacios, J. L. (1988). Two applications of urn processes. The fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains, *Probability in Engineering and Information Sciences*, **2**, 193-207.

2. Arthur, B., Ermoliev, Y. and Kaniovskii, Y. (1983). A generalized urn problem and its applications, *Cybernetics*, **19**, 61-71.

3. Bagchi, A. and Pal, A. K. (1985). Asymptotic normality in the generalized Pólya-Eggenberger urn model, with an application to computer data structures, *SIAM Journal of Algebraic Discrete Methods*, **6**, 394-405.

4. Ball, F. and Donnelly, P. (1988). A unified approach to variability in compartmental models, *Biometrics*, **44**, 685-694.

5. Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson Approximations*, Oxford Studies in Probability **2**, Oxford, U.K.: Clarendon Press.

6. Barton, D. E. and David, F. N. (1959). Combinatorial extreme value distributions, *Mathematika*, **6**, 63-76.

7. Benaim, M. and Hirsch, M. W. (1995). Dynamics of Morse-Smale urn processes, *Ergodic Theory & Dynamical Systems*, **15**, 1005-1030.

8. Bernard, S. R. (1977). An urn model study of variability within a compartment, *Bulletin of Mathematical Biology*, **39**, 463-470.

9. Bernoulli, J. (1713). *Ars Conjectandi*, Fratran, Basileue: Impensis Thurnisiorm.

10. Bingham, N. H. (1991). Fluctuation theory for the Ehrenfest urn, *Advances in Applied Probability*, **23**, 598-611.

11. Blackwell, D. and MacQueen, J. (1973). Ferguson distribution via Pólya urn schemes, *Annals of Statistics*, **1**, 353-355.

12. Blatz, P. J. (1968). On the arbitrary power of an arbitrary 2×2 matrix, *American Mathematical Monthly*, **75**, 57-58.

13. Blom, G. (1989). Mean transition times for the Ehrenfest urn model, *Advances in Applied Probability*, **21**, 479-480.

14. Blom, G. and Holst, L. (1991). Embedding procedures for discrete problems in probability, *The Mathematical Scientist*, **16**, 29-40.

15. Bowman, K. O. and Shenton, L. R. (1986a). The two-color urn model: recursions for the moments, *Communications in Statistics—Simulation and Computation*, **15**, 787-799.

16. Bowman, K. O. and Shenton, L. R. (1986b). A reinforcement-depletion model: a contiguity case, *Annals of the Institute of Statistical Mathematics*, **38**, 233-243.

17. Bowman, K. O., Shenton, L. R. and Bernard, S. R. (1985). Study of age dependent half-life of iodine in man: a reinforcement-depletion urn model, *Bulletin of Mathematical Biology*, **47**, 205-213.

18. Boyce, W. (1973). On a simple optimal stopping problem, *Discrete Mathematics*, **5**, 297-312.

19. Branson, D. (1992). An urn model and the coalescent in neutral infinite-alleles genetic process, In *Selected Proceedings of the Sheffield Symposium on Applied Probability* (Eds., I.V. Basawa and R. Taylor), pp. 174-192, Hayward, CA: Institute of Mathematical Statistics.

20. Chen, W. C. and Starr, N. (1980). Optimal stopping in an urn, *Annals of Probability*, **8**, 451-464.

21. Consul, P. C. (1974). A simple urn model dependent upon predetermined strategy, *Sankhyā, Series B*, **36**, 391-399.

22. Consul, P. C. (1995). A new generalized stochastic urn model *preprint.*

23. Consul, P. C. and Mittal, S. P. (1975). A new urn model with predetermined strategy, *Biometrische Zeitschrift*, **17**, 67-75.

24. Consul, P. C. and Mittal, S. P. (1977). Some discrete multinomial probability models with predetermined strategy, *Biometrical Journal*, **19**, 161-173.

25. Curtiss, I. H. (1942). A note on the theory of moment generating functions, *Annals of Mathematical Statistics*, **13**, 430-433.

26. Daston, L. (1988). *Classical Probability in the Enlightenment*, Princeton, NJ: Princeton University Press.

27. David, F. N. and Barton, D. E. (1962). *Combinatorial Chance*, London: Griffin.

28. Dette, H. (1994). On a generalization of the Ehrenfest urn model, *Journal of Applied Probability*, **31**, 930-939.

29. Diaconis, P. and Freedman, D. (1980). de Finetti's generalizations of exchangeability, In *Studies in Inductive Logic and Probability* (Ed., R. Jeffrey), Berkeley, CA: University of California Press.

30. Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent, *Theoretical Population Biology*, **23**, 34-63.

31. Donnelly, P. (1986). Partition structures, Polya urns, the Ewens sampling formula, and the ages of alleles, *Theoretical Population Biology*, **30**, 271-288.

32. Donnelly, P. and Whitt, W. (1989). On reinforcement-depletion compartmental urn models, *Journal of Applied Probability*, **26**, 477-489.

33. Doyle, P. G. and Snell, J. L. (1984). *Random Walks and Electric Networks*, Washington, DC: Mathematical Association of America.

34. Dutko, M. (1989). Central limit theorems for infinite urn models, *Annals of Probability*, **17**, 1255-1263.

35. Eggenberger, F. and Pólya, G. (1923). Über die Statistik verketetter Vorgänge, *Zeitschrift für Angewandte Mathematik und Mechanik*, **1**, 279-289.

36. Erdös, P. and Renyi, A. (1961). On a classical problem in probability theory, *Magyar Tud. Akad. Kutato Int. Kozl.*, **6**, 215-219.

37. Ewens, W. J. (1972). The sampling theory of selectively neutral alleles, *Theoretical Population Biology*, **3**, 87-112.

38. Famoye, F. and Consul, P. C. (1989). A stochastic urn model for the generalized negative binomial distribution, *Statistics*, **20**, 607-613.

39. Fang, K. T. (1982). A restricted occupancy problem, *Journal of Applied Probability*, **19**, 707-711.

40. Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Third edition, New York: John Wiley & Sons.

41. Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample from an animal population, *Journal of Animal Ecology*, **12**, 42-58.

42. Flajolet, P. (1980). Combinatorial aspects of continued fractions, *Discrete Mathematics*, **32**, 125-162.

43. Flatto, L. (1982). Limit theorems for some random variables associated with urn models, *Annals of Probability*, **10**, 927-934.

44. Françon, J. and Viennot, G. (1979). Permutations selon les pics, creuk, doubles montées, doubles descentes, nombres d'Euleur et nombres de Genocchi, *Discrete Mathematics*, **28**, 21-35.

45. Freund, J. E. and Pozner, A. N. (1956). Some results on restricted occupancy theory, *Annals of Mathematical Statistics*, **27**, 537-540.

46. Gouet, R. (1989). A martingale approach to strong convergence in a generalized Pólya-Eggenberger urn model, *Statistics & Probability Letters*, **8**, 225-228.

47. Gouet, R. (1993). Martingale functional central limit theorems for generalized Pólya urn, *Annals of Probability*, **21**, 1624-1639.

48. Goulden, I. P. and Jackson, D. M. (1986). Distributions, continued fractions, and the Ehrenfest urn model, *Journal of Combinatorial Theory, Series A*, **41**, 21-31.

49. Green, R. F. (1980). The cannibals' urn, *Unpublished manuscript.*

50. Gumbel, E. J. (1958). *Statistics of Extremes*, New York: Columbia University Press.

51. Heyde, C. C. (1977). On the central limit and iterated logarithm supplements to the martingale convergence theorem, *Journal of Applied Probability*, **14**, 758-775.

52. Hill, B., Lane, D. and Sudderth, W. (1980). A strong law for some generalized urn processes, *Annals of Probability*, **8**, 214-226.

53. Holst, L. (1979). A unified approach to limit theorems for urn models, *Journal of Applied Probability*, **16**, 154-162.

54. Holst, L. (1986). On birthday, collectors, occupancy and and other classical urn problems, *International Statistical Review*, **54**, 15-27.

55. Holst, L. and Hüsler, J. (1985). Sequential urn schemes and birth processes, *Advances in Applied Probability*, **17**, 257-279.

56. Holst, L., Kennedy, J. E., and Quine, M. P. (1988). Rates of Poisson convergence for some coverage and urn problems using coupling, *Journal of Applied Probability*, **25**, 717-724.

57. Hoppe, F. M. (1984). Polya-like urns and the Ewens' sampling formula, *Journal of Mathematical Biology*, **20**, 91-94.

58. Hoppe, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics, *Journal of Mathematical Biology*, **25**, 123-159.

59. Indira, N. K. and Menon, V. V. (1988a). Some limit theorems in urn models with indistinguishable balls, *Scandinavian Actuarial Journal*, 117-123.

60. Indira, N. K. and Menon, V. V. (1988b). Normal approximation in an urn model with indistinguishable balls, *Communications in Statistics—Theory and Methods*, **17**, 3389-3402.

61. Ivanov, V. A., Ivchenko, G. I., and Medvedev, Yu. I. (1985). Discrete problems in probability theory, *Journal of Soviet Mathematics*, **31(2)**.

62. Ivchenko, G. I. (1993). The waiting time and related statistics in the multinomial scheme: a survey, *Discrete Mathematics and Applications*, **3**, 451-482.

63. Ivchenko, G. I. and Ivanov, A. V. (1995). Decomposable statistics in inverse urn problems, *Discrete Mathematics and Applications*, **5**, 159-172.

64. Jagers, P. (1974). Galton-Watson process in varying environments, *Journal of Applied Probability*, **11**, 174-178.

65. Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications*, New York: John Wiley & Sons.

66. Johnson, N. L. and Kotz, S. (1991). Some generalizations of Bernoulli and Pólya-Eggenberger contagion models, *Statistical Papers*, **32**, 1-17.

67. Johnson, N. L., Kotz, S., and Balakrishnan, N. (1994). *Continuous Univariate Distributions - Volume 1*, Second edition, New York: John Wiley & Sons.

68. Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions - Volume 2*, Second edition, New York: John Wiley & Sons.

69. Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

70. Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second edition, New York: John Wiley & Sons.

71. Kac, M. (1947). Random walk and the theory of Brownian motion, *American Mathematical Monthly*, **54**, 369-391.

72. Karlin, S. (1967). Central limit theorems for certain infinite urn schemes, *Journal of Mathematical Mechanics*, **17**, 373-401.

73. Karlin, S. (1968). *A First Course in Stochastic Processes*, New York: Academic Press.

74. Karlin, S. and Leung, M.-Y. (1991). Some limit theorems on distributional patterns of balls in urns, *Annals of Applied Probability*, **1**, 513-538.

75. Keilson, J. (1979). *Markov Chain Models - Rarity and Exponentiality*, New York: Springer-Verlag.

76. Kemperman, J. H. B. (1961). *The Passage Problem for a Stationary Markov Chain*, Chicago, IL: Chicago University Press.

77. Kingman, J. F. C. (1975). Random discrete distributions, *Journal of the Royal Statistical Society, Series B*, **37**, 1-22.

78. Kingman, J. F. C. (1977). The population structure associated with Ewens' sampling formula, *Theoretical Population Biology*, **11**, 274-283.

79. Kingman, J. F. C. (1978). Random partitions in population genetics, *Proceedings of the Royal Society of London, Series A*, **361**, 1-20.

80. Kingman, J. F. C. (1982a). On the genealogy of large populations, *Journal of Applied Probability*, **19**, 27-43.

81. Kingman, J. F. C. (1982b). Exchangeability and the evolution of large populations, In *Exchangeability in Probability and Statistics* (Eds., G. Koch and F. Spizzichino), pp. 91-112, Amsterdam: North-Holland.

82. Kingman, J. F. C. (1982c). The coalescent, *Stochastic Processes and Their Applications*, **13**, 235-248.

83. Kolchin, V. F., Sevast'yanov, B. A., and Chistyakov, V. P. (1978). *Random Allocations*, New York: John Wiley & Sons.

84. Krafft, O. and Schaefer, M. (1993). Mean passage times for triangular transition matrices and a two parameter Ehrenfest urn model, *Journal of Applied Probability*, **30**, 964-970.

85. Krafft, O. and Schaefer, M. (1996). On the moments of the random walk associated with the two-parameter Ehrenfest urn model, *Unpublished manuscript*, Institute of Statistics, Technical University of Aachen, Germany.

86. Le Cam, L. (1958). Un théorème sur la division d'un intervalle par, *Publications of the Institute of Statistics, University of Paris*, **7**, 7-16.

87. Leitnaker, M. G. and Purdue, P. (1985). *Bulletin of Mathematical Biology*, **47**, 193-204.

88. Ling, K. D. (1993). Sooner and later waiting time distributions for frequency quota defined on a Pólya-Eggenberger urn model, *Soochow Journal of Mathematics*, **19**, 139-151.

89. Maindonald, J. H. (1990). Avoidance-modified urn models and parasitic wasps, *The Mathematical Scientist*, **15**, 128-136.

90. Maistrov, L. (1980). *The History of Probability Theory*, New York: Academic Press; Translated from Russian and annotated by S. Kotz.

91. Matsoms, P. (1988). Hoppande löss, *Elementa*, **71**, 95-99 (in Swedish).

92. Matveychuk, S. A. and Petunin, Yu. I. (1990). A generalization of the Bernoulli model arising in order statistics, *Ukrainian Mathematical Journal*, **42**, 518-528 (In Russian). (See also Abstract CP-103: " A new approach for the construction of statistical criteria for testing the uniformity of samples", *Bulletin of the Institute of Mathematical Statistics*, **19**, 355.)

93. Medvedev, Yu. I. (1970). Some theorems on asymptotic distribution of the chi-squared statistic, *Soviet Mathematical Doklady*, **192**, 987-989.

94. Menon, V. V. (1973). On the maximum Stirling numbers of the second kind, *Journal of Combinatorial Theory*, **15**, 11-24.

95. Menon, V. V. and Indira, N. K. (1990). A Poisson approximation in an urn model with indistinguishable balls, *Journal of Statistical Planning and Inference*, **26**, 93-101.

96. Mishra, A. and Sen, K. (1987). A modified Consul's urn model, *Gujarat Statistical Review*, **14**, 39-42.

97. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

98. Monahan, J. F. (1987). An alternative method for computing overflow probabilities, *Communications in Statistics – Theory and Methods*, **16**, 3355-3357.

99. Nevel'son, M. and Hasminskii, R. (1973). Stochastic approximation and recursive estimation, *Trans. Math. Monographs*, **47**, Providence, RI: American Mathematical Society.

100. Newman, D. J. and Shepp, L. (1960). The double dixie cup problem, *American Mathematical Monthly*, **67**, 58-61.

101. Nishimura, K. and Sibuya, M. (1988). Occupancy with two types of balls, *Annals of the Institute of Statistical Mathematics*, **40**, 77-91.

102. Paik, M. K. (1983). Empty or infinitely full, *Mathematics Magazine*, **56**, 221-223.

103. Palacios, J. L. (1994). Another look at the Ehrenfest urn via electric networks, *Advances in Applied Probability*, **26**, 820-824.

104. Park, C. J. (1976). A note on the asymptotic normality for the distribution of the number of empty cells under Bose-Einstein statistics, *Unpublished manuscript*, San Diego State University.

105. Pemantle, R. (1990a). A time-dependent version of Pólya's urn, *Journal of Theoretical Probability*, **3**, 627-637.

106. Pemantle, R. (1990b). Nonconvergence to unstable points in urn models and stochastic approximations, *Annals of Probability*, **18**, 698-712.

107. Pittel, B. (1987). An urn model for cannibal behavior, *Journal of Applied Probability*, **24**, 522-526.

108. Pólya, G. (1954). *Mathematics and Plausible Reasoning*, Princeton, NJ: Princeton University Press. Second edition (1963).

109. Prasad, B. and Menon, V. V. (1985). Poisson approximation to the distribution of empty cells in randomized occupancy, *Communications in Statistics—Theory and Methods*, **14**, 41-48.

110. Purdue, P. (1981). Variability in a single compartment system: A note on S. R. Bernard's model, *Bulletin of Mathematical Biology*, **43**, 111-116.

111. Rabinovitch, N. L. (1973). Studies in the history of probability and statistics. XXII. Probability in the Talmud, *Biometrika*, **56**, 437-441.

112. Ramakrishna, M. V. (1987). Computing the probability of hash table/urn overflow, *Communications in Statistics - Theory and Methods*, **16**, 3343-3353.

113. Read, R. C. (1979). The chord intersection problem, *Annals of the New York Academy of Sciences*, **319**, 444-454.

114. Robbins, H. and Whitehead, J. (1979). Sequential rank and the Pólya urn, *Journal of Applied Probability*, **16**, 213-219.

115. Sambursky, S. (1956). On the possible and probable in ancient Greece, *Osiris*, **12**, 35-48.

116. Samuel-Cahn, E. (1993). Optimal stopping in urn models with pay-off depending on maximal observed element, *Statistics & Probability Letters*, **16**, 361-368.

117. Shenton, L. R. (1981). A reinforcement-depletion urn problem-I: Basic theory, *Bulletin of Mathematical Biology*, **43**, 327-340.

118. Shenton, L. R. (1983). A reinforcement-depletion urn problem-II: Application and generalization, *Bulletin of Mathematical Biology*, **45**, 1-9.

119. Shenton, L. R. and Bowman, K. O. (1985). Distributions arising from reinforcement-depletion urn models, *Scandinavian Journal of Statistics*, **12**, 293-299.

120. Shenton, L. R. and Bowman, K. O. (1996a). Moment solution to an urn model, *Annals of the Institute of Statistical Mathematics*, **48**, 169-184.

121. Shenton, L. R. and Bowman, K. O. (1996b). Replenishment-Depletion: Urn in equilibrium, *Annals of the Institute of Statistical Mathematics* (to appear).

122. Shepp, L. A. (1969). Explicit solutions to some problems of optimal stopping, *Annals of Mathematical Statistics*, **40**, 993-1010.

123. Shur, W. (1984). The negative contagion reflection of the Pólya-Eggenberger distribution, *Communications in Statistics – Theory and Methods*, **13**, 877-885.

124. Siegrist, K. (1986a). Estimation and optimal stopping in a debugging model with masking, *Journal of Applied Probability*, **22**, 336-345.

125. Siegrist, K. (1986b). A debugging model with independent flows and random masking, *Operations Research*, **34**, 177-178.

126. Siegrist, K. (1987). An urn model with Bernoulli removals and independent additions, *Stochastic Processes and Their Applications*, **25**, 315-324.

127. Simons, G. (1987). Easily determining which urns are 'favorable', *Statistics & Probability Letters*, **5**, 43-48.

128. Small, C.G. (1985). Decomposition of models whose marginal distributions are mixtures, *Canadian Journal of Statistics*, **13**, 131-136.

129. Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*, Cambridge, MA: Harvard University Press.

130. Takács, L. (1962). A generalization of the ballot problem and its application in the theory of queues, *Journal of the American Statistical Association*, **57**, 327-337.

131. Takács, L. (1967). *Combinatorial Methods in the Theory of Stochastic Processes*, New York: John Wiley & Sons.

132. Takács, L. (1979). On an urn problem of Paul and Tatiana Ehrenfest, *Mathematical Proceedings of the Cambridge Philosophical Society*, **86**, 127-130.

133. Trieb, G. (1992). A Pólya urn model and the coalescent, *Journal of Applied Probability*, **29**, 1-10.

134. Uppuluri, V. R. R. and Wright, T. (1987). A note on a further generalization of the Ehrenfest urn model, *Proceedings of the American Statistical Association, Sampling Survey Section*, pp. 564-569.

135. van Beek, K. W. H. and Stam, A. J. (1987). A variant of the Ehrenfest model, *Advances in Applied Probability*, **19**, 995-996.

136. Voit, M. (1996). Asymptotic distributions for the Ehrenfest urn and related random walks, *Journal of Applied Probability*, **33**, 340-356.

137. Watanabe, T. (1986). On a generalization of polynomials in the ballot problem, *Journal of Statistical Planning and Inference*, **14**, 143-152.

138. Watterson, G. A. (1976). The stationary distribution of an infinitely many neutral alleles diffusion model, *Journal of Applied Probability*, **13**, 639-651.

139. Wei, L. J. (1979). The generalized Pólya's urn design for sequential medical trials, *Annals of Statistics*, **7**, 291-296.

140. Yamato, H. (1993). A Pólya urn model with a continuum of colors, *Annals of the Institute of Statistical Mathematics*, **45**, 453-458.

141. Zabell, S. L. (1992). Predicting the unpredictable, *Synthese*, **90**, 205-230.

142. Zaman, A. (1981). *An Approximation Theorem for Finite Markov Exchangeability*, Ph.D. Thesis, Stanford University.

143. Zaman, A. (1984). Urn models for Markov exchangeability, *Annals of Probability*, **12**, 222-229.

15

A Unified Derivation of Occupancy and Sequential Occupancy Distributions

Ch. A. Charalambides

University of Athens, Panepistemiopolis, Athens, Greece

Abstract: Consider a supply of balls randomly distributed in $n+r$ distinguishable urns and assume that the number X of balls distributed in any specific urn is a random variable with probability function $\Pr[X=j]=q_j$, $j=0,1,2,\ldots$. The probability function and factorial moments of the number K_i of urns occupied by i balls each, among n specified urns, given that a total of $S_{n+r}=m$ balls are distributed in the $n+r$ urns, are expressed in terms of finite differences of the u-fold convolution of $q_j, j=0,1,2,\ldots$. As a particular case, the probability function and factorial moments of the number $K=n-K_0$ of occupied urns (by at least one ball), among n specified urns, given that $S_{n+r}=m$, are deduced. Further, when balls are sequentially distributed, the probability function and ascending factorial moments of the number W_k of balls required until a predetermined number k of urns, among n specified urns, are occupied, are also expressed in terms of finite differences of the u-fold convolution of $q_j, j=0,1,2,\ldots$. Finally, the conditional probability function $\Pr[W_{k+1}-W_k=j|W_k=m]$, $j=1,2,\ldots$, is derived. Illustrating these results, the cases with $q_j, j=0,1,2,\ldots$, the Poisson, geometric, binomial and negative binomial distributions are presented.

Keywords and phrases: Finite differences, non-central Stirling numbers, non-central C-numbers, random occupancy models, urn models

15.1 Introduction

Barton and David (1959a), considering a supply of balls randomly distributed in n distinguishable urns and assuming that the number X of balls distributed in any specific urn is a random variable obeying a Poisson, binomial or negative binomial law, derived the probability function and factorial moments of the number K of occupied urns (by at least one ball), in these three cases, given that a total of $S_n = m$ balls are distributed in the n urns; these distributions are the classical, restricted and pseudo-contagious occupancy, respectively. In sequential occupancy, Barton and David (1959b), derived in the preceding three cases, the probability function and cumulants or factorial moments of the number W_k of balls required until a predetermined number k of urns are occupied. Charalambides (1986) derived in the general case $\Pr[X = j] = q_j$, $j = 0, 1, 2, \ldots$ the probability function and factorial moments of K given that $S_n = m$.

In the present paper a unified derivation of occupancy and sequential occupancy distributions is presented. More precisely, considering a supply of balls randomly distributed in $n + r$ distinguishable urns, the probability function and factorial moments of the number K_i of urns occupied by i balls each, among n specified urns, given that $S_{n+r} = m$, are obtained in the general case $\Pr[X = j] = q_j$, $j = 0, 1, 2, \ldots$. Then the occupancy probability function and factorial moments of the number K of occupied urns, among n specified urns, given that $S_{n+r} = m$, are deduced. Further, the sequential occupancy probability function and ascending factorial moments of the number W_k of balls required until a predetermined number k of urns, among n specified urns, are occupied, are obtained. Finally, the conditional probability function $\Pr[W_{k+1} - W_k = j | W_k = m]$, $j = 1, 2, \ldots$, is derived.

15.2 Occupancy Distributions

Consider a supply of balls randomly distributed in $n + r$ distinguishable urns among which n are specified. Suppose that the number X of balls distributed in any specific urn is a random variable with known probability function $\Pr[X = j] = q_j$, $j = 0, 1, 2, \ldots$. Let S_{n+r} be the total number of balls distributed in the $n + r$ urns with $\Pr[S_{n+r} = m] = q_m(n + r)$, $m = 0, 1, 2, \ldots$. Assuming that the occupancy of each urn is independent of the others, the probability $q_m(n + r)$, $m = 0, 1, 2, \ldots$, is given by the sum

$$q_m(n + r) = \sum q_{j_1} q_{j_2} \cdots q_{j_{n+r}}, \tag{15.1}$$

where the summation is extended over all integers $j_s \geq 0$, $s = 1, 2, \ldots, n + r$, such that $j_1 + j_2 + \ldots + j_{n+r} = m$. Notice that the conditional joint probability $q(j_1, j_2, \ldots, j_{n+r}; m)$ that the s-th urn contains $j_s \geq 0$ balls, $s = 1, 2, \ldots, n + r$, given that m balls are distributed in the $n + r$ urns, is given by

$$q(j_1, j_2, \ldots, j_{n+r}; m) = \frac{q_{j_1} q_{j_2} \cdots q_{j_{n+r}}}{q_m(n + r)} . \tag{15.2}$$

Under this model, let K_i be the number of urns, among the n specified urns, occupied by i balls each and

$$p_k(m,n,r,i) = \Pr[K_i = k|S_{n+r} = m], \ k = 0,1,2,\ldots,n. \tag{15.3}$$

With $(u)_j = u(u-1)\ldots(u-j+1)$, the (descending) factorial of u of degree j, let us denote by

$$\mu_{(j)}(m,n,r,i) = E[(K_i)_j|S_{n+r} = m], j = 1,2,\ldots, \tag{15.4}$$

the j-th factorial moment of the conditional probability function (15.3).

An expression of the probability function (15.3) in terms of finite differences of the convolutions (15.1) is obtained in the following theorem. As regards this expression notice that the difference operator with unit increment, denoted by Δ, is defined by $\Delta a(u) = a(u+1) - a(u)$; its powers are defined recursively by $\Delta^n a(u) = \Delta[\Delta^{n-1}a(u)]$, $n = 2,3\ldots$. The shift operator with unit increment, denoted by E, is defined by $E^n a(u) = a(u+n)$, $n = 1,2,\ldots$. Therefore, $\Delta = E - I$ and

$$\begin{aligned}[\Delta^n a(u)]_{u=r} &= \sum_{j=0}^{n}(-1)^{n-j}\binom{n}{j}[E^j a(u)]_{u=r} \\ &= \sum_{j=0}^{n}(-1)^{n-j}\binom{n}{j}a(r+j).\end{aligned} \tag{15.5}$$

In the presence of more than one variable the symbols Δ_u and E_u are used, with the subscript indicating the variable with respect to which the corresponding operator is performed.

Theorem 15.2.1 *(a) The conditional probability $p_k(m,n,r,i)$ that k urns, among the n specified urns, are occupied by i balls each, given that m balls are distributed in the $n+r$ urns, $k = 0,1,2,\ldots,n$, is given by*

$$p_k(m,n,r,i) = \frac{\binom{n}{k}}{q_m(n+r)}[\Delta_u^{n-k} q_i^{n+r-u} q_{m-in-ir+iu}(u)]_{u=r}. \tag{15.6}$$

(b) The j-th factorial moment $\mu_{(j)}(m,n,r,i)$, $j = 1,2,\ldots$, of the probability function $p_k(m,n,r,i)$, $k = 0,1,2,\ldots,n$, is given by

$$\mu_{(j)}(m,n,r,i) = \frac{(n)_j q_i^j q_{m-ij}(n+r-j)}{q_m(n+r)}. \tag{15.7}$$

PROOF. (a) The conditional probability $p_k(m,n,r,i)$, on using (15.2), may be written as

$$p_k(m,n,r,i) = \frac{\binom{n}{k} q_i^k}{q_m(n+r)} \sum q_{j_1} q_{j_2} \cdots q_{j_{n-k}} q_{j_{n+1}} q_{j_{n+2}} \cdots q_{j_{n+r}},$$

where the summation is extended over all nonnegative integers $j_s \neq i$, $s = 1, 2, \ldots, n-k$ and j_s, $s = n+1, n+2, \ldots, n+r$ such that $j_1 + j_2 + \cdots + j_{n-k} + j_{n+1} + j_{n+2} + \cdots + j_{n+r} = m - ik$. Considering the events: A_s that the s-th urn contains i balls, $s = 1, 2, \ldots, n-k$, when $m - ik$ balls are distributed in the $n + r - k$ urns, then, applying the inclusion-exclusion principle, this sum may be expressed as

$$\begin{aligned}
p_k(m,n,r,i) &= \frac{\binom{n}{k} q_i^k}{q_m(n+r)} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} P(A_{s_1} A_{s_2} \ldots A_{s_j}) \\
&= \frac{\binom{n}{k}}{q_m(n+r)} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} q_i^{k+j} q_{m-ik-ij}(n+r-k-j) \\
&= \frac{\binom{n}{k}}{q_m(n+r)} \sum_{j=0}^{n-k} (-1)^{n-k-j} \binom{n-k}{j} q_i^{n-j} q_{m-in+ij}(r+j).
\end{aligned}$$

The last expression, by virtue of (15.5), implies (15.6).

(b) The j-th binomial moment $b_j(m,n,r,i) = \mu_{(j)}(m,n,r,i)/j!$ is equal to the sum of the conditional probabilities that any j urns, among the n specified urns, are occupied by i balls each, given that m balls are distributed in the $n + r$ urns. Hence,

$$b_j(m,n,r,i) = \binom{n}{j} \frac{q_i^j q_{m-ij}(n+r-j)}{q_m(n+r)}$$

and since $\mu_{(j)}(m,n,r,i) = j! b_j(m,n,r,i)$, (15.7) is deduced. ■

The generating function of the bivariate probability function

$$\begin{aligned}
p_{k,m}(n,r,i) &= \Pr[K_i = k, S_{n+r} = m] \\
&= \binom{n}{k} [\Delta_u^{n-k} q_i^{n+r-u} q_{m-in-ir+iu}(u)]_{u=r}, \\
& \quad k = 0,1,2,\ldots,\min\{n,[m/i]\},\ m = 0,1,2,..., \qquad (15.8)
\end{aligned}$$

is derived in the next Theorem.

Theorem 15.2.2 *The double generating function of the sequence $p_{k,m}(n,r,i)$, $k = 0,1,2,\ldots,s$, $s = \min\{n,[m/i]\}$, $m = 0,1,2,\ldots$ is given by*

$$g_{n,r,i}(t,u) = \sum_{m=0}^{\infty} \sum_{k=0}^{s} p_{k,m}(n,r,i) t^k u^m = [g(u)]^r [g(u) + q_i u^i (t-1)]^n, \quad (15.9)$$

where $g(u) = \sum_{j=0}^{\infty} q_j u^j$.

PROOF. Since

$$\sum_{k=0}^{s} p_k(m,n,r,i)t^k = \sum_{j=0}^{s} \mu_{(j)}(m,n,r,i)(t-1)^j/j!$$

and $p_{k,m}(n,r,i) = p_k(m,n,r,i)q_m(n+r)$, it follows from (15.7) that

$$\sum_{k=0}^{s} p_{k,m}(n,r,i)t^k = \sum_{j=0}^{s} \binom{n}{j} q_i^j q_{m-ij}(n+r-j)(t-1)^j.$$

Hence,

$$\begin{aligned} g_{n,r,i}(t,u) &= \sum_{m=0}^{\infty}\sum_{j=0}^{s} \binom{n}{j} q_i^j q_{m-ij}(n+r-j)(t-1)^j u^m \\ &= \sum_{j=0}^{m} \binom{n}{j} [q_i u^i(t-1)]^j \sum_{m=ij}^{\infty} q_{m-ij}(n+r-j)u^{m-ij} \\ &= \sum_{j=0}^{n} \binom{n}{j} [q_i u^i(t-1)]^j [g(u)]^{n+r-j} \\ &= [g(u)]^r[g(u)+q_i u^i(t-1)]^n \end{aligned}$$

which proves the theorem. ■

Consider now the number K of urns, among the n specified urns, occupied by at least one ball each. Since $K = n - K_0$, the conditional probability function

$$p_k(m,n,r) = \Pr[K=k|S_{n+r}=m],\ k=0,1,2,\ldots,n,$$

and its factorial moments

$$v_{(j)}(m,n,r) = E[(K)_j|S_{n+r}=m],\ j=1,2,\ldots$$

may be deduced from (15.6) and (15.7) respectively.

Corollary 15.2.1 *(a) The conditional probability $p_k(m,n,r)$ that k urns, among the n specified urns, are occupied (by at least one ball) given that m balls are distributed in the $n+r$ urns, $k=0,1,2,...,n$, is given by*

$$p_k(m,n,r) = \frac{\binom{n}{k}}{q_m(n+r)}[\Delta_u^k q_0^{n+r-u} q_m(u)]_{u=r}. \tag{15.10}$$

(b) The j-th factorial moment $v_{(j)}(m,n,r)$, $j=1,2,\ldots$, of the probability function $p_k(m,n,r)$, $k=0,1,2,\ldots,m$, is given by

$$v_{(j)}(m,n,r) = \frac{(n)_j}{q_m(n+r)}\ [\Delta_u^j q_0^{n+r-u} q_m(u)]_{u=n+r-j}. \tag{15.11}$$

PROOF. (a) The probability function (15.10) follows directly from (15.6) by putting $i = 0$ and replacing $n - k$ by k.

(b) The k-th factorial moment $\mu_{(k)}(m,n,r) = E[(K_0)_k | S_{n+r} = m]$ of the number K_0 of empty urns, among the n specified urns, by virtue of (15.7), is given by

$$\mu_{(k)}(m,n,r) = \frac{(n)_k q_0^k q_m(n+r-k)}{q_m(n+r)} .$$

Since $v_{(j)}(m,n,r) = E[(K)_j | S_{n+r} = m] = E[(n - K_0)_j | S_{n+r} = m]$ may be written as $v_{(j)}(m,n,r) = (-1)^j E([K_0 - n]_j | S_{n+r} = m)$, where $[u]_j = u(u+1)\dots(u+j-1)$, on using the expression [Charalambides (1986)]

$$E([K_0 - m]_j | S_{n+r} = m) = \sum_{k=0}^{j} (-1)^{j-k} \binom{n-k}{j-k} E[(K_0)_k | S_{n+r} = m],$$

it follows that

$$\begin{aligned} v_{(j)}(m,n,r) &= \frac{j!}{q_m(n+r)} \sum_{k=0}^{j} (-1)^k \binom{n}{k} \binom{n-k}{j-k} q_0^k q_m(n+r-k) \\ &= \frac{(n)_j}{q_m(n+r)} \sum_{k=0}^{j} (-1)^k \binom{j}{k} q_0^k q_m(n+r-k) \\ &= \frac{(n)_j}{q_m(n+r)} \sum_{k=0}^{j} (-1)^{j-k} \binom{j}{k} q_0^{n+r-(n+r-j+k)} \\ &\quad \times q_m(n+r-j+k). \end{aligned}$$

The last expression, by virtue of (15.5), yields (15.11). ■

Example 15.2.1 *Poisson probabilities.* Let the number X of balls allocated in any specific urn is a Poisson random variable with

$$q_j = \Pr[X = j] = e^{-\lambda}\lambda^j / j!, \ j = 0,1,2,\dots,(\lambda > 0).$$

Its u-fold convolution is again a Poisson random variable with

$$q_m(u) = \Pr[S_u = m] = e^{-u\lambda}(u\lambda)^m / m!, \ m = 0,1,2,\dots .$$

Notice that the conditional joint probability that the s-th urn contains $j_s \geq 0$ balls, $s = 1,2,\dots,n+r$, given that m balls are distributed in the $n+r$ urns, by virtue of (15.2), is given by

$$q(j_1, j_2, \dots, j_{n+r}; m) = \frac{m!}{j_1! j_2! \dots j_{n+r}!} \frac{1}{(n+r)^m} , \ j_1 + j_2 + \dots + j_{n+r} = m.$$

Therefore, this random occupancy model with the assumption of Poisson probabilities is equivalent to the classical occupancy model where a fixed number m of distinguishable balls are randomly distributed in $n+r$ distinguishable urns.

The conditional probability that k urns, among the n specified urns, are occupied by i balls each, given that m balls are distributed in the $n+r$ urns, by virtue of (15.6), is given by

$$\begin{aligned} p_k(m,n,r,i) &= \binom{n}{k}\frac{m!}{(n+r)^m}\left[\Delta_u^{n-k}\frac{u^{m-in-ir+iu}}{(i!)^{n+r-u}(m-in-ir+iu)!}\right]_{u=r} \\ &= \binom{n}{k}\frac{m!}{(n+r)^m}\sum_{j=0}^{n-k}(-1)^{n-k-j}\binom{n-k}{j} \\ &\quad\times\frac{(j+r)^{m-in+ij}}{(i!)^{n-j}(m-in+ij)!}\,. \end{aligned}$$

This probability, under the classical occupancy model, is given in the particular case $r=0$, by Feller (1968, p. 112).

The j-th factorial moment of the probability function $p_k(m,n,r,i)$, $k = 0,1,2,\ldots,n$, by (15.7), is given by

$$\mu_{(j)}(m,n,r,i) = \frac{(n)_j m!}{(i!)^j(m-ij)!}\,\frac{(n+r-j)^{m-ij}}{(n+r)^m}\,.$$

This moment, for $r=0$, is given by Riordan (1958, p. 101).

The conditional probability $p_k(m,n,r)$ that k urns, among the n specified urns, are occupied (by at least one ball) given that m balls are distributed in the $n+r$ urns, by virtue of (15.10), is given by

$$p_k(m,n,r) = (n)_k(n+r)^{-m}S(m,k,r),\ k=0,1,2,\ldots,n,$$

where

$$S(m,k,r) = \frac{1}{k!}\left[\Delta_u^k u^m\right]_{u=r} = \frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}(r+j)^m$$

is the non-central Stirling number of the second kind [Koutras (1982)]. This probability function was derived by Barton and David (1959a). For $r=0$, this is the very well-known classical occupancy distribution.

The j-th factorial moment of $p_k(m,n,r)$, $k=0,1,2,\ldots,n$, by (15.11), is

$$v_{(j)}(m,n,r) = (n)_j(n+r)^{-m}S(m,j,n+r-j),\ j=1,2,\ldots\,.$$

Example 15.2.2 *Geometric probabilities.* Assume that the number of balls allocated in any specific urn obeys a geometric distribution with

$$q_j = \Pr[X=j] = pq^j,\ j=0,1,2,\ldots\ (q=1-p, 0<p<1).$$

Its u-fold convolution obeys a negative binomial distribution with

$$q_m(u) = \Pr[S_u = m] = \binom{u+m-1}{m}p^u q^m,\ m=0,1,2,\ldots\,.$$

Notice that the conditional joint probability that the s-th urn contains $j_s \geq 0$ balls, $s = 1, 2, \ldots, n+r$, given that m balls are distributed in the $n+r$ urns, by virtue of (15.2), is given by

$$q(j_1, j_2, \ldots, j_{n+r}; m) = 1 \Big/ \binom{n+r+m-1}{m}, \quad j_1 + j_2 + \ldots + j_{n+r} = m.$$

Thus, this random occupancy model with the assumption of geometric probabilities is equivalent to the occupancy model where a fixed number m of indistinguishable (like) balls are randomly distributed in $n+r$ distinguishable urns.

The conditional probability that k urns, among the n specified urns, are occupied by i balls each, given that m balls are distributed in the $n+r$ urns, by virtue of (15.6), is given by

$$\begin{aligned} p_k(m,n,r,i) &= \frac{\binom{n}{k}}{\binom{n+r+m-1}{m}} \left[\Delta_u^{n-k} \binom{u+m-in-ir+iu-1}{u-1} \right]_{u=r} \\ &= \frac{\binom{n}{k}}{\binom{n+r+m-1}{m}} \sum_{j=0}^{n-k} (-1)^{n-k-j} \binom{n-k}{j} \\ &\quad \times \binom{n+r+m-j(i+1)-1}{m-ij}. \end{aligned}$$

Its j-th factorial moment, by (15.7), is given by

$$\mu_{(j)}(m,n,r,i) = (n)_j \binom{n+r+m-j(i+1)-1}{m-ij} \Big/ \binom{n+r+m-1}{m}.$$

The conditional probability that k urns, among the n specified urns, are occupied (by at least one ball), given that m balls are randomly distributed in the $n+r$ urns, by virtue of (15.10), is given by

$$\begin{aligned} p_k(m,n,r) &= \binom{n}{k} \left[\Delta_u^k \binom{u+m-1}{m} \right]_{u=r} \Big/ \binom{n+r+m-1}{m} \\ &= \binom{n}{k} \binom{r+m-1}{r+k-1} \Big/ \binom{n+r+m-1}{m}. \end{aligned}$$

Its j-th factorial moment, by (15.11), is

$$\begin{aligned} v_{(j)}(m,n,r) &= (n)_j \left[\Delta_u^j \binom{u+m-1}{m} \right]_{u=n+r-j} \Big/ \binom{n+r+m-1}{m} \\ &= (n)_j \binom{n+r+m-j-1}{n+r-1} \Big/ \binom{n+r+m-1}{m}. \end{aligned}$$

These probability functions, under the assumption that a fixed number m of like balls are randomly distributed in $n+r$ distinguishable balls, were given, in the case $r = 0$, in terms of certain bivariate generating functions by Riordan (1958, p. 103). These generating functions are closely related to the bivariate generating function (15.9).

Notice that the random occupancy model with the assumption of geometric probabilities truncated to the right at the point s, that is, with

$$q_j = (1 - q^{s+1})^{-1} pq^j, \qquad j = 0, 1, 2, \dots, s,$$

whence

$$q_m(u) = L(m, u, s)(1 - q^{s+1})^{-u} p^u q^m, \qquad m = 0, 1, 2, ..., su,$$

where

$$L(m, u, s) = \sum_{j=0}^{u} (-1)^j \binom{u}{j} \binom{u + m - j(s+1) - 1}{u - 1},$$

is equivalent to the occupancy model where a fixed number m of indistinguishable (like) balls are randomly distributed in $n+r$ distinguishable urns each with capacity limited to s balls. The corresponding occupancy probabilities can be similarly deduced from (15.6) and (15.10).

15.3 Sequential Occupancy Distributions

Suppose now that balls are sequentially distributed at random in the $n + r$ distinguishable urns until a predetermined number k of urns, among the n specified urns, are occupied (by at least one ball) and let W_k be the number of balls required. Consider the probability function

$$q_m(k, n, r) = \Pr[W_k = m], \ m = k, k+1, \dots \tag{15.12}$$

and its j-th ascending factorial moment

$$\mu_{[j]}(k, n, r) = E([W_k]_j), \qquad j = 1, 2, \dots, \tag{15.13}$$

where $[u]_j = u(u+1)\dots(u+j-1)$ is the ascending factorial of u of degree j. Notice that for a discrete waiting time distribution the ascending factorial moments, in general, can be computed more easily than any other moments. This computation is facilitated by the consideration of the conditional probability

$$P_m(u, n, r) = \frac{nq_0^{n+r-u-1} q_1 q_{m-1}(u)}{mq_m(n+r)}, \ m = 1, 2, \dots, \tag{15.14}$$

that one ball is distributed at the m-th allocation in an urn among the n specified urns and $m-1$ balls are distributed in u specific urns (among the other

$n+r-1$ urns), given that m balls are distributed in the $n+r$ urns. Further, consider the j-th ascending factorial moment of the sequence of probabilities $P_m(u,n,r)$, $m=1,2,\ldots$,

$$\alpha_{[j]}(u,n,r)=\sum_{m=1}^{\infty}[m]_j P_m(u,n,r),\ j=1,2,\ldots\ . \tag{15.15}$$

Notice that the sequence of probabilities $P_m(u,n,r)$, $m=1,2,\ldots$, does not necessarily add to unity.

In the next theorem the probabilities (15.12) and the factorial moments (15.13) are expressed in terms of finite differences of the convolution probabilities $q_m(u)$, $m=0,1,2,\ldots$, and the factorial moments (15.15), respectively.

Theorem 15.3.1 *(a) The probability $q_m(k,n,r)$ that m balls are sequentially distributed in the $n+r$ urns until k urns, among the n specified urns, are occupied (by at least one ball), $m=k,k+1,\ldots$, is given by*

$$q_m(k,n,r)=\binom{n-1}{k-1}\frac{nq_1}{mq_m(n+r)}\left[\Delta_u^{k-1}q_0^{n+r-u-1}q_{m-1}(u)\right]_{u=r}. \tag{15.16}$$

(b) If the conditional probability (15.14) is a polynomial in u of degree $m-1$, then the j-th ascending factorial moment $\mu_{[j]}(k,n,r)$, $j=1,2,\ldots$, of the probability function $q_m(k,n,r)$, $m=1,2,\ldots$, is given by

$$\mu_{[j]}(k,n,r)=\binom{n-1}{k-1}\left[\Delta_u^{k-1}a_{[j]}(u,n,r)\right]_{u=r}, \tag{15.17}$$

where $a_{[j]}(u,n,r)$ is given by (15.15).

PROOF. (a) The probability $q_m(k,n,r)$ that m balls are sequentially distributed in the $n+r$ urns until k urns, among the n specified urns, are occupied is equal to the product

$$q_m(k,n,r)=P_m(n,r)p_{k-1}(m-1,n-1,r),$$

where

$$P_m(n,r)\equiv P_m(n+r-1,n,r)=\frac{nq_1q_{m-1}(n+r-1)}{mq_m(n+r)}$$

is the conditional probability that one ball is distributed at the m-th allocation in an urn among the n specified urns and $m-1$ balls are distributed in the other $n+r-1$ urns, given that m balls are in the $n+r$ urns and

$$p_{k-1}(m-1,n-1,r)$$

is the conditional probability that $k-1$ urns among $n-1$ specified urns are occupied given that $m-1$ balls are allocated in the $n+r-1$ urns. Thus, on using (15.10), the expression (15.16) is deduced.

(b) From the definition of $\mu_{[j]}(k,n,r)$, on using (15.16) and (15.14), it follows that

$$\mu_{[j]}(k,n,r) = \binom{n-1}{k-1}\left[\Delta_u^{k-1}\sum_{m=k}^{\infty}[m]_j P_m(u,n,r)\right]_{u=r}.$$

Further, the assumption that the probability $P_m(u,n,r)$ is a polynomial in u of degree $m-1$ implies $\Delta_u^{k-1}P_m(u,n,r)=0$, $m=1,2,\ldots,k-1$ and hence

$$\mu_{[j]}(k,n,r) = \binom{n-1}{k-1}\left[\Delta_u^{k-1}\sum_{m=1}^{\infty}[m]_j P_m(u,n,r)\right]_{u=r}.$$

Thus, by (15.15), the expression (15.17) is deduced. Hence, the theorem. ■

The number W_k of balls required to be sequentially distributed in the $n+r$ urns until a predetermined number k of urns, among the n specified urns, are occupied may be expressed as a sum

$$W_k = Y_1 + Y_2 + \ldots + Y_k, \qquad k=1,2,\ldots, \tag{15.18}$$

where Y_i is the number of balls required to be distributed in the $n+r$ urns after $(i-1)$-st urn, among the n specified urns, empty before receives one ball and until the i-th urn, among the n specified urns, receives one ball, $i=1,2,\ldots$. Some light into the structure of the distribution of W_k can be shed by the conditional distribution of Y_{k+1} given W_k. In the next theorem, the conditional probability function

$$q_{j;m}(k,n,r) = \Pr[Y_{k+1}=j|W_k=m],\ j=1,2,\ldots \tag{15.19}$$

is expressed in terms of the convolution probabilities $q_m(u)$, $m=0,1,2,\ldots$.

Theorem 15.3.2 *The conditional probability $q_{j;m}(k,n,r)$ that j additional balls are sequentially distributed in the $n+r$ urns until the $(k+1)$-th urn, among the n specified urns, is occupied (by one ball) given that m balls are sequentially distributed in the $n+r$ urns until the k-th urn, among the n specified urns, is occupied (by one ball), $j=1,2,\ldots$, is given by*

$$q_{j;m}(k,n,r) = \frac{(n-k)q_1 q_{m+j-1}(k+r)/q_m(k+r)}{(m+j)q_0 q_{m+j}(n+r)/q_m(n+r)}. \tag{15.20}$$

PROOF. Notice first that the conditional probability $q_{j;m}(n,k,r)$ is equal to the conditional probability that $j-1$ additional balls are distributed in $k+r$ specific urns (which are the k occupied urns, among the n specified urns, and the r unspecified urns) and one ball is distributed at the j-th additional allocation in the remaining $n-k$ (empty) urns given the j additional balls are distributed in the $n+r$ specific urns. This conditional probability is, in turn, equal to the quotient

$$q_{j;m}(k,n,r) = P_{m+j}(k+r,n-k,r)/Q_m(k+r,n+r),$$

where $P_{m+j}(k+r, n-k, r)$ is the conditional probability that $m+j-1$ balls are distributed in $k+r$ specific urns (which are k specific urns among the n specified urns and the r unspecified urns) and one ball is distributed at the $(m+j)$-th allocation in the $n-k$ remaining urns given that $m+j$ balls are distributed in the $n+r$ urns and $Q_m(k+r, n+r)$ is the conditional probability that m balls are in $k+r$ specific urns given that m balls are distributed in the $n+r$ urns. Since

$$\begin{aligned} P_{m+j}(k+r, n-k, r) &= \frac{(n-k)q_0^{n-k-1}q_1 q_{m+j-1}(k+r)}{(m+j)q_{m+j}(n+r)}, \\ Q_m(k+r, n+r) &= \frac{q_0^{n-k}q_m(k+r)}{q_m(n+r)}, \end{aligned}$$

the expression (15.20) is deduced. Hence, the theorem. ■

Example 15.3.1 *Poisson probabilities.* As in Example 15.2.1, assume that

$$q_j = \Pr[X = j] = e^{-\lambda}\lambda^j/j!, \ j = 0, 1, 2, \ldots \ (\lambda > 0)$$

whence

$$q_m(u) = \Pr[S_u = m] = e^{-u\lambda}(u\lambda)^m/m!, \ m = 0, 1, 2, \ldots .$$

Then, by virtue of (15.16), the probability that m balls are sequentially distributed in the $n+r$ urns until k urns, among the n specified urns, are occupied (by at least one ball) is given by

$$q_m(k, n, r) = (n)_k(n+r)^{-m}S(m-1, k-1, r), \ m = k, k+1, \ldots,$$

where $S(m, k, r)$ is the non-central Stirling number of the second kind. For $r = 0$, this is the classical waiting time occupancy distribution [Barton and David (1959b)].

For the computation of the moments of the distribution $q_m(k, n, r)$, $m = k, k+1, \ldots$, note first that the probability (15.14) reduces to

$$P_m(u, n, r) = \frac{n}{n+r}\left(\frac{u}{n+r}\right)^{m-1}, \ m = 1, 2, \ldots,$$

which is a polynomial in u of degree $m-1$. Since, by (15.15),

$$\begin{aligned} a_{[j]}(u, n, r) &= \frac{n}{n+r}\sum_{m=1}^{\infty}[m]_j\left(\frac{u}{n+r}\right)^{m-1} \\ &= \frac{j!n}{n+r}\sum_{m=1}^{\infty}\binom{j+m-1}{m-1}\left(\frac{u}{n+r}\right)^{m-1} \\ &= \frac{j!n}{n+r}\left(1-\frac{u}{n+r}\right)^{-j-1}, \end{aligned}$$

it follows from (15.17) that

$$\begin{aligned}\mu_{[j]}(k,n,r) &= \frac{j!n}{n+r}\binom{n-1}{k-1}\left[\Delta_u^{k-1}(1-\frac{u}{n+r})^{-j-1}\right]_{u=r}\\ &= \frac{j!n}{n+r}\binom{n-1}{k-1}\sum_{i=0}^{k-1}(-1)^{k-i-1}\binom{k-1}{i}\left(1-\frac{i+r}{n+r}\right)^{-j-1}.\end{aligned}$$

The conditional probability function (15.19) reduces to

$$q_{i;m}(k,n,r) = \frac{n-k}{n+r}\left(\frac{k+r}{n+r}\right)^{j-1}, \ j=1,2,\dots .$$

Therefore W_k, according to (15.18), is a sum of independent geometric random variables with varying success probability [Johnson and Kotz (1977, p. 155)].

Example 15.3.2 *Binomial and negative binomial probabilities.* Assume that

$$q_j = \binom{s}{j}p^j q^{s-j}, \qquad j=0,1,2,\dots,$$

where s is a positive integer and $0<p<1$, $q=1-p$ (binomial distribution) or $s,p<0$, $q=1-p$ (negative binomial distribution). Its u-fold convolution is

$$q_m(u) = \binom{su}{m}p^m q^{su-m}, \ m=0,1,2,\dots .$$

Notice that the conditional joint probability that the i-th urn contains $j_i \geq 0$ balls, $i=1,2,...,n+r$, given that m balls are distributed in the $n+r$ urns, by virtue of (15.2), is given by

$$\begin{aligned}&q(j_1,j_2,...,j_{n+r};m)\\ &= \binom{s}{j_1}\binom{s}{j_2}\cdots\binom{s}{j_{n+r}}\Big/\binom{sn+sr}{m}, \ j_1+j_2+...+j_{n+r}=m.\end{aligned}$$

Therefore, this random occupancy model with the assumption of binomial or negative binomial probabilities is equivalent to the occupancy model where a fixed number m of like (indistinguishable) balls are randomly distributed in $n+r$ distinguishable cells each with s distinguishable compartments of capacity limited to one ball (binomial distribution) or with $-s$ distinguishable compartments of unlimited capacity (negative binomial distribution with s a negative integer). A particular case of this model for $s=-1$ is examined in Example 15.2.2.

The probability $q_m(k,n,r)$ that m balls are sequentially distributed in the $n+r$ urns until k urns, among the n specified urns, are occupied (by at least one ball), by virtue of (15.16), is given by

$$q_m(k,n,r) = \frac{s(n)_k}{(sn+sr)_m}\, C(m-1,k-1,s,rs), \ m=k,k+1,\dots,$$

where

$$\begin{aligned} C(m-1,k-1,s,rs) &= \frac{1}{(k-1)!}\left[\Delta_u^{k-1}(su)_{m-1}\right]_{u=r} \\ &= \frac{1}{(k-1)!}\sum_{j=0}^{k-1}(-1)^{k-j-1}\binom{k-1}{j}(sj+sr)_{m-1} \end{aligned}$$

is the non-central C-number [*cf.* Charalambides and Koutras (1983) where this distribution was also derived as a waiting time coupons collector distribution]. The expression of the probability function $q_m(k,n,r)$ in view of the expression of the non-central C-numbers conforms with the corresponding expression derived by Barton and David (1959b) in the case $r=0$.

For the computation of the moments of the distribution $q_m(k,n,r)$, $m = k, k+1, \ldots$, note first that the probability (15.14) reduces, in this case, to

$$P_m(u,n,r) = \frac{sn(su)_{m-1}}{(sn+sr)_m}\ , \quad m=1,2,\ldots,$$

which is a polynomial in u of degree $m-1$. Its j-th ascending factorial moment (15.15), on using the combinatorial identity

$$\sum_{i=0}^{\infty}\binom{k+i-1}{i}\frac{(x)_i}{(z-k)_i} = \frac{(z)_k}{(z-x)_k},$$

with $k=j+1$, $x=su$, $z=sn+sr+j$, reduces to

$$\begin{aligned} a_{[j]}(u,n,r) &= \frac{sn}{sn+sr}\sum_{m=1}^{\infty}[m]_j\frac{(su)_{m-1}}{(sn+sr-1)_{m-1}} \\ &= \frac{j!sn}{sn+sr}\sum_{m=1}^{\infty}\binom{j+m-1}{m-1}\frac{(su)_{m-1}}{(sn+sr-1)_{m-1}} \\ &= \frac{j!sn}{sn+sr}\sum_{i=0}^{\infty}\binom{j+i}{i}\frac{(su)_i}{(sn+sr-1)_i} \\ &= \frac{j!sn(sn+sr+j)_{j+1}}{(sn+sr)(sn+sr-su+j)_{j+1}} = \frac{j!sn(sn+sr+j)_j}{(sn+sr-su+j)_{j+1}}. \end{aligned}$$

Thus, from (15.17) it follows that

$$\begin{aligned} \mu_{[j]}(k,n,r) &= \binom{n-1}{k-1}\left[\Delta_u^{k-1}\frac{j!sn(sn+sr+j)_j}{(sn+sr-su+j)_{j+1}}\right]_{u=r} \\ &= \binom{n-1}{k-1}\sum_{i=0}^{k-1}(-1)^{k-i-1}\binom{k-1}{i}\frac{j!sn(sn+sr+j)_j}{(sn-si+j)_{j+1}}. \end{aligned}$$

The conditional probability function (15.19) reduces, in this case, to

$$q_{j;m}(k,n,r) = \frac{(sn-sk)(sk+sr-m)_{j-1}}{(sn+sr-m)_j}\ , \quad j=1,2,\ldots .$$

Note that this distribution for s a positive integer is a particular case of a waiting time hypergeometric with support $\{1, 2, ..., sk + sr - m + 1\}$ while for $s < 0$ is a Waring distribution with parameters $\theta = -s(n + r) + m > 0$ and $a = -s(k + r) + m > 0$ [Johnson and Kotz (1977, p. 88)].

Acknowledgements. This work was carried out while the author was visiting the University of Cyprus, on sabbatical leave from the University of Athens, Greece. The hospitality of the Department of Mathematics and Statistics is very much appreciated.

References

1. Barton, D. E. and David, F. N. (1959a). Contagious occupancy, *Journal of the Royal Statistical Society, Series B*, **21**, 120–133.

2. Barton, D. E. and David, F. N. (1959b). Sequential occupancy, *Biometrika*, **46**, 218–223.

3. Charalambides, Ch. A. (1986). Derivation of probabilities and moments of certain generalized discrete distributions via urn models, *Communications in Statistics—Theory and Methods*, **15**, 677–696.

4. Charalambides, Ch. A. and Koutras, M. (1983). On the differences of the generalized factorials at an arbitrary point and their combinatorial applications, *Discrete Mathematics*, **47**, 183–201.

5. Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, Volume 1, Third edition, New York: John Wiley & Sons.

6. Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications*, New York: John Wiley & Sons.

7. Koutras, M. (1982). Non-central Stirling numbers and some applications, *Discrete Mathematics*, **42**, 73–89.

8. Riordan, J. (1958). *An Introduction to Combinatorial Analysis,* New York: John Wiley & Sons.

16

Moments, Binomial Moments and Combinatorics

Janos Galambos

Temple University, Philadelphia, PA

Abstract: The relation of Bonferroni-type inequalities to combinatorial problems is demonstrated. An urn model in this setting leads to a statistical paradox as well as to an open problem concerning a statistical test of goodness of fit. An extension of Bonferroni-type inequalities to quadratic inequalities is discussed, which are then applied to the analysis of the structure of pairwisely independent events and of exchangeable events.

Keywords and phrases: Sequences of events, number of occurrences, moment, binomial moment, Bonferroni-type inequalities, methods for proof, relation to combinatorics, an urn model, goodness of fit test, quadratic inequalities, pairwise independence, exchangeable events

16.1 Basic Relations

Let $A_1, A_2, \ldots, A_n$ be events on a given probability space. Let $X = X_n = X_n(A)$ denote the number of those A_j which occur. Set

$$p_r = \Pr[X = r], \; S_k = E\left[\binom{X}{k}\right], \text{ and } m_k = E(X^k) \tag{16.1}$$

where $r \geq 0$ and $k \geq 0$ are integers. Another form for S_k is the sum

$$S_k = \sum\nolimits^* p(i_1, i_2, \ldots, i_k), \qquad k \geq 1, \tag{16.2}$$

where $\sum^*$ signifies summation over all subscripts $1 \leq i_1 < i_2 < \cdots < i_k \leq n$, and

$$p(i_1, i_2, \ldots, i_k) = \Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}]. \tag{16.3}$$

The equivalence of (16.1) and (16.2) for S_k easily follows by observing that, upon introducing the indicator variables

$$I_j = \begin{cases} 1 & \text{if } A_j \text{ occurs} \\ 0 & \text{if } A_j \text{ fails,} \end{cases}$$

and denoted by $I(i_1, i_2, \ldots, i_k) = I_{i_1} I_{i_2} \cdots I_{i_k}$ the indicator of the intersection on the right hand side of (16.3), the identity

$$\sum\nolimits^{*} I(i_1, i_2, \ldots, i_k) = \binom{X}{k} \tag{16.4}$$

holds. By taking expectations on both sides of (16.4), the equivalence of (16.1) and (16.2) follows.

Let us write the formulae at (16.1) in detail. We have

$$S_k = \sum_{r=k}^{n} \binom{r}{k} p_r \tag{16.5}$$

and

$$m_k = \sum_{r=0}^{n} r^k p_r. \tag{16.6}$$

We also have $S_1 = m_1$, $2S_2 + S_1 = m_2$, and in general, a sequential computation yields that the sequences $\{S_k\}_1^n$ and $\{m_k\}_1^n$ uniquely determine each other. An easy combinatorial argument also yields that the sequences $\{S_k\}$ and $\{p_r\}$, too, determine each other. Indeed, by the identities

$$\binom{k+u}{u}\binom{r}{k+u} = \binom{r}{u}\binom{r-u}{k}$$

and

$$\sum_{k=0}^{t} (-1)^k \binom{t}{k} = \begin{cases} 1 & \text{if } \ t = 0 \\ 0 & \text{if } \ t > 0, \end{cases} \tag{16.7}$$

we get

$$\begin{aligned} \sum_{k=0}^{n-u} (-1)^k \binom{k+u}{u} S_{k+u} &= \sum_{k=0}^{n-u} (-1)^k \binom{k+u}{u} \sum_{r=k+u}^{n} \binom{r}{k+u} p_r \\ &= \sum_{r=u}^{n} \binom{r}{u} p_r \sum_{k=0}^{r-u} (-1)^k \binom{r-u}{k} = p_u. \end{aligned} \tag{16.8}$$

The inversion formula (16.8) becomes an inequality if we sum on the extreme left hand side up to a number $d < n - u$. However, since we want to discuss more general inequalities than just those stemming from (16.8), we do not carry out the details of computation.

In addition to p_r, we shall discuss bounds on

$$q_r = \Pr[X \geq r] = p_r + p_{r+1} + \cdots + p_n.$$

In spite of the preceding close relation between p_r and q_r, the literature treated bounds on p_r and q_r as two separate problems. I shall quote a recent result of Galambos and Simonelli (1996b) that provides a tool for combining these two lines of research.

16.2 Linear Inequalities in S_k, p_r and q_r

Let $c_k = c_k(n, r)$ and $d_k = d_k(n, r)$, $1 \leq k \leq n$, $0 \leq r \leq n$, be two sequences of real constants. Let a and b be two coefficients taking one of the values 1, 0, or -1. Then, the linear inequalities

$$ap_r + \sum_{k=0}^{n} c_k S_k \geq 0 \tag{16.9}$$

and

$$bq_r + \sum_{k=0}^{n} d_k S_k \geq 0 \tag{16.10}$$

will be referred to as *Bonferroni-type inequalities* if they are valid on every probability space for an arbitrary choice of the underlying events A_j, $1 \leq j \leq n$. Note that, due to the choice of a and b, (16.9) and (16.10) cover both lower and upper bounds on p_r and q_r, respectively, or they may be just inequalities among binomial moments. It also has to be stressed that the coefficients c_k and d_k may take the value zero, so (16.9) and (16.10) may involve a few binomial moments only. Since the recent book of Galambos and Simonelli (1996a) discusses Bonferroni-type inequalities in great detail, I wish to limit my statements to those aspects which are relevant from a combinatorial point of view. When a correct inequality of the kind of (16.9) and (16.10) is set up, its actual proof can be done by turning to indicator variables: we apply (16.4), prove the resulting combinatorial inequality and take expectations. This method is known as the method of indicators, which we demonstrate on two simple inequalities. First, we probe

$$q_1 \geq S_1 - S_2. \tag{16.11}$$

When we turn to indicators, both sides of (16.11) become zero if $X = 0$, and thus we have to prove that

$$1 \geq X - \binom{X}{2} \qquad \text{if } X \geq 1. \tag{16.12}$$

But clearly, the right hand side of (16.12) equals 1 if $X = 1$ or 2, while

$$X - \binom{X}{2} = \frac{X(3-X)}{2} \leq 0 \qquad \text{if } X \geq 3.$$

This proves (16.12) and the expected value in (16.12) becomes (16.11). An improvement over (16.11) is provided by the parametric family of inequalities

$$q_1 \geq \frac{2}{u+1}\, S_1 - \frac{2}{u(u+1)}\, S_2, \qquad 1 \leq u < n \text{ integer}. \tag{16.13}$$

Note that (16.13) reduces to (16.11) if $u = 1$. In order to see that (16.13) is indeed valid, we once again turn to indicators. Since (16.13) is zero on both sides if we go to indicators and if $X = 0$, we again assume that $X \geq 1$, and then

$$\begin{aligned} \frac{2}{u+1}\, X - \frac{2}{u(u+1)}\binom{X}{2} &= \frac{X(2u+1-X)}{u(u+1)} \leq \max_{1 \leq X \leq n} \frac{X(2u+1-X)}{u(u+1)} \\ &= \frac{u(2u+1-u)}{u(n+1)} = 1, \end{aligned}$$

where the maximum in the penultimate equation is computed from the fact that $X(2u+1-X)$ is a parabola in X, whose maximum is taken at $X = u$ or $u+1$ (both X and U are integers). By taking expectations in the preceding inequalities, we get (16.13).

The fact that (16.13) contains the best lower bound on q_1 by means of S_1 and S_2 is not an easy argument [for a purely combinatorial argument for optimality, see Galambos (1977)]. Given that (16.13) contains the optimal lower bound, an elementary argument yields that the best bound is achieved by choosing $u = [2S_2/S_1] + 1$, where $[y]$ signifies the integer part of y.

We can see clearly from the preceding proofs that linear inequalities like (16.9) and (16.10) are actually nonprobabilistic ones. Another way of expressing this fact is to observe that (16.9) or (16.10) is valid on an arbitrary probability space if it is valid on the trivial probability space $(\Omega, \mathcal{A}, P)$, $\mathcal{A}$ containing Ω and the empty set $\emptyset$ only. Now, the trivial space is a subspace of every probability space, so an inequality like (16.9) and (16.10) is valid on an arbitrary space if it is valid on a single specific probability space. In particular, if (16.9) or (16.10) is valid on a space in which $\mathcal{A}$ is generated by a single sequence of n independent events A_j with $\Pr[A_j] = p$ for each j, on which space (16.9) and (16.10) become polynomial inequalities in p, we established that (16.9) and (16.10) are essentially polynomial inequalities. This simple logic already implies the quite surprising conclusion that (16.7), which is a combinatorial formula, (16.8) with $u = 0$, which is known as the method of inclusion and exclusion, and the algebraic formula

$$(1-p)^n = \sum_{k=0}^{n} (-1)^k \binom{n}{k} p^k \tag{16.14}$$

are equivalent. In particular, the fact that (16.7) implies (16.14) is quite surprising. This analysis can be carried further which led to the following general result [Galambos and Simonelli (1996b)].

Theorem 16.2.1 *Assume that*

$$ap_0 + \sum_{k=0}^{n} c_k(n) S_k \geq 0 \tag{16.15}$$

is a Bonferroni-type inequality. Then, for arbitrary $0 \leq r \leq n$,

$$ap_r + \sum_{k=0}^{n-r} c_k(n-r) \binom{k+r}{r} S_{k+r} \geq 0 \tag{16.16}$$

and

$$aq_r + \sum_{k=0}^{n-r} c_k(n-r) \binom{k+r-1}{r-1} S_{k+r} \geq 0 \tag{16.17}$$

are also valid on an arbitrary probability space, i.e., Bonferroni-type inequalities.

The significance of Theorem 16.2.1 is that it provides a uniform rule of generating bounds on p_r and q_r concurrently, utilizing the coefficients of (16.15). Just think it over for a variety of urn problems: if one can get good bounds by means of binomial moments on the probability of having no urn empty, then we automatically have bounds on the probabilities of having exactly r, or at least r urns empty. For the classical urn models, this does not provide new asymptotic results but it does simplify and shorten several proofs.

As an example, let us go back to (16.11) and write $p_0 = 1 - q_1$. This way, (16.11) is of the form (16.15) and thus we have

$$p_r \leq S_r - (r+1)S_{r+1} + \binom{r+2}{2} S_{r+2},$$

in which the first term S_r comes from $1 = S_0$. We also have

$$q_r \leq S_r - rS_{r+1} + \binom{r+1}{2} S_{r+2}.$$

The emphasis here is that the preceding two inequalities do not need proof; they are consequences of (16.11). As a consequence of (16.13), the just established inequalities remain valid if we multiply (in either one) S_{r+1} by $2/(u+1)$ and S_{r+2} by $2/u(u+1)$, $1 \le u < n$ integer.

16.3 A Statistical Paradox and an Urn Model with Applications

Let Y, Y_1 and Y_2 be three independent random variables with a common continuous distribution function $F(x)$. Then $\Pr[Y < Y_1] = \Pr[Y < Y_2] = 1/2$. However, if we place the variables Y, Y_1 and Y_2 onto the real line and we observe that $Y < Y_1$, then $\Pr[Y < Y_2] = 2/3$. Indeed, given that $Y < Y_1$, then Y_2 will fall with equal probabilities below Y, between Y and Y_1, and over Y_1 (if logic is not sufficiently convincing, condition on the variables $Y = y$ and $Y_1 = z$, $y < z$, and use the continuous total probability rule). I usually refer to this phenomenon as the magnet effect of Y_1 on Y_2: although Y_1 and Y_2 are independent, once Y_1 took up its position relative to Y, then Y_1 pulls Y_2 onto its own side of Y.

If we continue with dropping further independent copies of Y into the real line, an urn model develops: assume that $Y, Y_1, Y_2, \dots, Y_{n-1}$ have been placed (independently) on the real lines, and if their order statistics are $Y_{1:n} < Y_{2:n} < \cdots < Y_{n:n}$, then Y_n will fall with equal probabilities into the line segments $(Y_{j:n}, Y_{j+1:n})$, $0 \le j \le n$, where $Y_{0:n} = -\infty$ and $Y_{n+1:n} = +\infty$. That is, if the preceding line segments are urns, and the sequentially placed Y_j are balls, then the n-th ball is placed into the available $n+1$ urns with equal probabilities, i.e., into each urn with probability $1/(n+1)$. This leads to the Bose-Einstein statistic of physics for which a large number of results is available. However, instead of relating the model to physics, I wish to suggest to use the model for tests of significance. This is not entirely new since related urn models are used for such tests, several versions can be developed from this urn model and I want to point to one such possibility that is both simple and sensitive for the null hypothesis.

Take n independent variables from the population F (either observations or random numbers), and form the urns as described. Place another set of n independent observations $Z_1, Z_2, \dots, Z_n$, say, into the urns, and we want to test the null hypothesis that the Z_j also come from F. If their distribution is indeed F, then the last urn will remain empty with probability $1/2$. This is true regardless of the value of n (why not paired comparison then?). However, if the Z_j did not come from the population F, then the number of balls in the last urn is very sensitive to n. Namely, the last urn compares the maxima in the original population (which is F) and in the Z-sample, and extremes soon show divergence if the samples are from different populations. In particular, such a test should be very powerful when one wants to make a decision between two specific distributions for the Z_j (for example, normal versus Weibull, or deciding between two extreme value distributions). Since the computations involved are combinatorial (using the urn model), it appears appropriate to propose this model here even if the details are yet to be worked out.

16.4 Quadratic Inequalities

Even though the simple quadratic inequality

$$q_1(2S_2 + S_1) \geq S_1^2, \tag{16.18}$$

which is just the Cauchy-Schwarz inequality applied to $S_1^2 = E^2[X\ I(X \geq 1)]$ (recall that $2S_2 + S_1 = m_2$), has been applied in the literature for generating the best available Borel-Cantelli lemma, quadratic inequalities failed to induce an activity similar to Bonferroni-types. Since we would like to revive interest in this field, we included a section on such inequalities in Galambos and Simonelli (1996a). Here, I would like to include an application of quadratic inequalities different from the ones in the cited book.

Just as in the linear case, we want to establish inequalities which are quadratic in the sequences p_r, q_r and S_k and which are valid on every probability space and for all choices of the underlying events. The indicator method does not work and thus, contrary to Bonferroni-type inequalities, quadratic inequalities are probabilistic in nature. Yet, one does not have to prove them on an arbitrary probability space. Namely, the following result is true [Galambos (1969)].

Theorem 16.4.1 *If a quadratic inequality in the sequences p_r, q_r and S_k, $Q \geq 0$, is of the form*

$$Q = up^2 + vpq + wq^2 \text{ with } u \geq 0,\ v \geq 0,\ w \geq 0$$

on a probability space that carries at least one nontrivial event A with $\Pr[A] = p$ and $q = 1 - p = \Pr[A^c]$, $0 < p < 1$, whenever the underlying events A_j, $1 \leq j \leq n$, consist of at least one A or at least one A^c (the rest are Ω's and $\emptyset$'s), then $Q \geq 0$ on an arbitrary probability space and for every choice of the basic events A_j.

Just to demonstrate the application of Theorem 16.4.1, let us prove the following extension of (16.18)

$$q_k[(k+1)S_{k+1} + kS_k] \geq (S_1 - k + 1)S_k. \tag{16.19}$$

Let m of the A_j be chosen as A with $\Pr[A] = p$, s of the A_j as A^c, t of the A_j as Ω and the rest $R = n - m - s$ as $\emptyset$. We assume that $t + r < n$, and each of m, s, t and R is nonnegative. Now,

$$q_k = p\Delta_{m+t,k} + q\Delta_{s+t,k}$$

where $\Delta_{a,b} = 1$ if $a \geq b$, and it is zero otherwise. Furthermore,

$$S_j = \binom{m+t}{j} p + \binom{s+t}{j} q. \tag{16.20}$$

Hence, if Q is the difference of the left hand side and the right hand side of (16.19), then easy algebraic manipulation yields that Q is indeed of the required form of Theorem 16.4.1, so (16.19) is a universal inequality.

Evidently, (16.19) continues to hold if we replace q_k by 1. In this new form, we apply (16.19) to the first n terms of an infinite sequence of exchangeable events A_j, $j \geq 1$. Put

$$w_k = \Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}], \qquad k \geq 1.$$

By exchangeability, $S_k = \binom{n}{k} w_k$, and thus (16.19) entails

$$(k+1)\binom{n}{k+1} w_{k+1} + k\binom{n}{k} w_k \geq (nw_1 - k + 1)\binom{n}{k} w_k.$$

Dividing the preceding inequality by $(k!)^{-1} n^{k+1}$ and letting $n \to +\infty$, we get

$$w_{k+1} \geq w_1 w_k, \qquad k \geq 1. \tag{16.21}$$

In particular, $w_2 \geq w_1^2$ which is known as positive dependence of the indicator variables $I(A_j)$. While this is a well known property of infinite exchangeability, its usual proof depends on de Finetti's representation theorem for infinite exchangeable sequences.

A reversed form of (16.19) is the inequality [see Galambos and Simonelli (1996a, Chapter X)]

$$kS_k \leq S_1 S_{k-1} + 2\binom{n-2}{k-2}\left[S_2 - \binom{S_1}{2}\right], \qquad k \geq 2. \tag{16.22}$$

Its proof by Theorem 16.4.1 is a simple combinatorial calculation after utilizing (16.20). Now, if we divide (16.22) by $n^k/(k-1)!$ and let $n \to +\infty$, we get

$$w_k \leq w_1 w_{k-1} + (k-1)(w_2 - w_1^2), \qquad k \geq 2. \tag{16.23}$$

An immediate consequence of the combination of (16.21) and (16.23) is that pairwisely independent infinite sequences of exchangeable events are completely independent. This, once again, follows from de Finetti's representation theorem, but such an elementary proof for this property appears here for the first time.

It is interesting to look at (16.21) and (16.23) from a numerical point of view. It is a long-standing problem that, given a finite sequence of exchangeable events, whether they can be extended into an infinite sequence without violating exchangeability. We can now conclude from (16.21) and (16.23) that if $A_1, A_2, \ldots, A_n$, for some $n \geq 3$, are exchangeable with $\Pr[A_j] = 0.5$, $\Pr[A_i \cap A_j] = 0.3$ and $\Pr[A_i \cap A_j \cap A_k] = 0.26$ (i, j and k are distinct), then the sequence A_j, $1 \leq j \leq n$, cannot be extended into an infinite exchangeable sequence. Here, the inequalities of (16.21) are satisfied but (16.23) fails. In fact, we must have $w_1 w_2 = 0.15 \leq w_3 \leq w_1 w_2 + 2(w_2 - w_1^2) = 0.25$.

We also get an insight into pairwisely independent events via (16.19) and (16.22) even if the underlying events are not exchangeable. Let $A_1, A_2, \ldots$ be an infinite sequence of events with $\Pr[A_j] = p$ for each j and assume that the A_j are pairwisely independent. Then, by arguing as in the exchangeable case, we get from (16.19) and (16.22) that on a subsequence on which

$$\lim S_k \Big/ \binom{n}{k} = w_k, \qquad 1 \leq k \leq T, \text{ say},$$

exist (such subsequence always exists because the 'averages' on the left hand side are bounded), $w_k = p^k$, $1 \leq k \leq T$. Here, T can be arbitrarily large.

References

1. Galambos, J. (1969). Quadratic inequalities among probabilities, *Annals of the University of Budapest, Section Mathematics*, **12**, 11–16.

2. Galambos, J. (1977). Bonferroni inequalities, *Annals of Probability*, **5**, 577–581.

3. Galambos, J. (1995). *Advanced Probability Theory*, Second edition, New York: Marcel Dekker.

4. Galambos, J. and Simonelli, I. (1996a). *Bonferroni-type Inequalities with Applications*, New York: Springer-Verlag.

5. Galambos, J. and Simonelli, I. (1996b). An extension of the method of polynomials and a new reduction formula for Bonferroni-type inequalities, *Statistics & Probability Letters*, **28**, 147–151.

6. Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Applications*, New York: John Wiley & Sons.

7. Koch, G. and Spizzichino, F. (Eds.) (1982). *Exchangeability in Probability and Statistics*, Amsterdam: North-Holland.

8. Kolchin, V. F., Sevast'yanov, B. A. and Chistyakov, V. P. (1978). *Random Allocations*, New York: John Wiley & Sons.

9. Stoyanov, J. M. (1987). *Counterexamples in Probability*, New York: John Wiley & Sons.

PART IV

APPLICATIONS TO QUEUEING THEORY

17

Nonintersecting Paths and Applications in Queueing Theory

Walter Böhm

University of Economics, Augasse, Vienna

Abstract: In this paper, some applications of nonintersecting lattice paths to queueing problems are presented. In particular, we derive a determinant formula for non-coincidence probabilities of non-identically distributed Poisson processes from which, in an almost elementary way, zero-avoiding transition probabilities in a Markovian tandem queue can be found. Finally, we present a result about $D/M/1$ queues, where the arrival instances are not equally spaced.

Keywords and phrases: Nonintersecting paths, tandem queues, $D/M/1$ queues

17.1 Introduction

The correspondence between two-dimensional lattice paths and the sample paths of certain Markovian queueing systems is now a well established fact. Several authors have applied this correspondence together with the powerful and elegant tools of lattice path combinatorics successfully to the analysis of the time dependent behavior of such types of queueing systems. It was Professor Mohanty (1979) who initiated most of the research work in this field and he himself has contributed many valuable results.

However, it is less well known that, besides simple lattice path counting, counting results regarding sets of nonintersecting paths are also of some significance in the theory of queues.

In this paper, I will to present some examples of queueing problems which may be solved easily using representations in terms of nonintersecting lattice paths.

To clarify matters, let us first state a theorem, which is now regarded as a key result in modern path combinatorics.

Theorem 17.1.1 *Let $\mathcal{S}$ be a set of steps in $\mathbf{Z}^2$. Furthermore, let $\mathbf{A} = (A_1, A_2, .., A_n)$ and $\mathbf{E} = (E_1, E_2, \ldots, E_n)$ be sequences of lattice points in $\mathbf{Z}^2$ and consider the families $\mathbf{P} = (P_1, P_2, \ldots, P_n)$ of lattice paths, where P_i has steps in $\mathcal{S}$ and leads from A_i to $E_i, i = 1, 2, \ldots, n$. Assume that the sequences $\mathbf{A}$ and $\mathbf{E}$ are such that any permutation of the origins A_i other than the identity implies that at least two paths in $\mathbf{P}$ intersect each other, i.e. have a point in common. Then, the number of families $\mathbf{P}$ which are nonintersecting is given by the determinant*

$$\det_{1\leq i,j\leq n} \|L(A_j \to E_i)\|, \tag{17.1}$$

where $L(A_j \to E_i)$ denotes the number of lattice paths from A_j to E_i.

This theorem is due to Gessel and Viennot (1985). An analogue of (17.1) for sample paths of identically distributed strong Markov processes is due to Karlin and McGregor (1959); see also Karlin (1988).

Our plan is as follows. In Section 17.2, we will derive a determinant formula for non-coincidence probabilities of sets of dissimilar Bernoulli and Poisson processes by means of Theorem 17.1.1. This way, we extend the Karlin-McGregor theorem to a special case of non i.i.d. Markov processes. These results are then used in Section 17.3 to derive a formula for zero-avoiding transition probabilities in a simple Jackson network. In Section 17.4, we apply Theorem 17.1.1 to determine the probability that a stationary Poisson process avoids a finite set of taboo points. In the last section, we demonstrate how an interesting problem in connection with $D/M/1$ queues can be solved.

17.2 Dissimilar Bernoulli Processes

Consider $r \geq 2$ independent Bernoulli processes $S_1(n), S_2(n), \ldots, S_r(n)$, which start at time zero in positions $m_1 < m_2 < \cdots < m_r$ and terminate at time n in positions $k_1 < k_2 < \cdots < k_r$. The process $S_i(n)$ is assumed to have success probability $p_i, 1 \leq i \leq r$. Let $\mathbf{S}_n = (S_1(n), \ldots, S_r(n))$ and let

$$[\mathbf{S}, 0, \mathbf{m}] \to [\mathbf{S}, n, \mathbf{k}]$$

denote the event that the processes S_i move simultaneously from m_i to k_i, $i = 1, 2, \ldots, r$, in the time interval $(0, n)$, where $\mathbf{m} = (m_1, \ldots, m_r)$ and $\mathbf{k} = (k_1, \ldots, k_r)$.

We want to find a formula for

$$\Pr\Big[[\mathbf{S}, 0, \mathbf{m}] \to [\mathbf{S}, n, \mathbf{k}], \text{ without coincidence}\Big], \tag{17.2}$$

where *without coincidence* means that no two of the processes S_i occupy the same state at the same time. To determine this probability, it will be convenient to encode the sample paths of the processes $S_i(n)$ by lattice paths in the plane, which

(i) have step set $\mathcal{S} = \{(1,0),(1,1)\}$;

(ii) start at heights m_i and terminate after n steps at heights $k_i, i = 1, \ldots, r$.

The number of families of lattice paths satisfying (i) and (ii), which in addition do not touch each other, will be denoted by $N_r(\mathbf{m}, \mathbf{k}, n)$. Once this number is known, we also know (17.2), because a particular path of process S_i has probability $p_i^{k_i-m_i}(1-p_i)^{n-k_i+m_i}$. Hence, by independence, any family of paths has probability

$$\prod_{i=1}^{r} p_i^{k_i-m_i}(1-p_i)^{n-k_i+m_i}.$$

However, since the number of paths from m_i to k_j with n steps is trivially $\binom{n}{k_j-m_i}$, it follows from Theorem 17.1.1 that

$$N_r(\mathbf{m}, \mathbf{k}, n) = \det \left\| \binom{n}{k_j - m_i} \right\|_{r\times r}.$$

Hence,

$$\begin{aligned} \Pr\Big[[\mathbf{S}, 0, \mathbf{m}] &\to [\mathbf{S}, n, \mathbf{k}], \text{ without coincidence}\Big] \\ &= \det \left\| \binom{n}{k_j - m_i} \right\|_{r\times r} \prod_{\nu=1}^{r} p_\nu^{k_\nu - m_\nu}(1-p_\nu)^{n-k_\nu+m_\nu}. \end{aligned} \tag{17.3}$$

In what follows, it will be helpful to rewrite (17.3) in terms of the transition functions $P_i(m, k; n)$ of the processes S_i, viz.,

$$\begin{aligned} P_i(m,k;n) &= \Pr[S_i(n) = k | S_i(0) = m] \\ &= \binom{n}{k-m} p_i^{k-m}(1-p_i)^{n-k+m}. \end{aligned}$$

By applying elementary transformation rules for determinants, we obtain successively

$$\begin{aligned} \Pr\Big[[\mathbf{S}, 0, \mathbf{m}] &\to [\mathbf{S}, n, \mathbf{k}], \text{ without coincidence}\Big] \\ &= \det \left\| \binom{n}{k_j - m_i} \right\| \prod_{\nu=1}^{r} p_\nu^{n-k_\nu+m_\nu}(1-p_\nu)^{n-k_\nu+m_\nu} \end{aligned}$$

$$
\begin{aligned}
&= \det\left\|\binom{n}{k_j - m_i} p_j^{k_j-m_j}(1-p_j)^{n-k_j+m_j}\right\| \\
&= \det\left\|\binom{n}{k_j - m_i} p_j^{k_j-m_i}(1-p_j)^{n-k_j+m_i}\left(\frac{p_j}{1-p_j}\right)^{m_i-m_j}\right\| \\
&= \det\left\|P_j(m_i, k_j; n)\left(\frac{p_j}{1-p_j}\right)^{m_i-m_j}\right\|. \qquad (17.4)
\end{aligned}
$$

Similarly, we get a second, equivalent formula:

$$
\begin{aligned}
\Pr\Big[[\mathbf{S}, 0, \mathbf{m}] &\to [\mathbf{S}, n, \mathbf{k}], \text{ without coincidence}\Big] \\
&= \det\left\|P_i(m_i, k_j; n)\left(\frac{p_i}{1-p_i}\right)^{k_i-k_j}\right\|. \qquad (17.5)
\end{aligned}
$$

Observe that these are determinant formulas for the non-coincidence probabilities of a special class of dissimilar Markov processes.

From (17.4) and (17.5), we obtain with almost no additional efforts the transition probability function of a set of non-coincident and dissimilar Poisson processes.

In fact, put $p_i = \frac{\lambda_i t}{n}$ for some real number $\lambda_i > 0$ and let $n \to \infty$ while keeping t fixed. Setting $\Delta_i = p_i/(1-p_i)$, we may rewrite (17.4) as

$$
\begin{aligned}
\det\left\|P_j(m_i, k_j; n)\Delta_j^{m_i-m_j}\right\| &= \det\left\|P_j(m_i, k_j; n)\Delta_j^{m_i}\right\| \prod_{\nu=1}^{r} \Delta_\nu^{-m_\nu} \\
&= \det\left\|P_j(m_i, k_j; n)\left(\frac{\Delta_j}{\Delta_i}\right)^{m_i}\right\|. \qquad (17.6)
\end{aligned}
$$

Since this determinant is a finite sum of finite products, we may apply the limit to the generic component in (17.6).

By the Poisson limit theorem, we have, as $n \to \infty$,

$$
P_j(m_i, k_j; n) \to \frac{e^{-\lambda_j t}(\lambda_j t)^{k_j-m_i}}{(k_j - m_i)!},
$$

and

$$
\frac{\Delta_j}{\Delta_i} = \frac{p_j}{1-p_j}\,\frac{1-p_i}{p_i} = \frac{\frac{\lambda_j t}{n}\left(1 - \frac{\lambda_i t}{n}\right)}{\left(1 - \frac{\lambda_j t}{n}\right)\frac{\lambda_i t}{n}} \to \frac{\lambda_j}{\lambda_i}.
$$

Thus,

$$
\lim_{n\to\infty} P_j(m_i, k_j; n)\left(\frac{\Delta_j}{\Delta_i}\right)^{m_i} = \frac{e^{-\lambda_j t}(\lambda_j t)^{k_j-m_i}}{(k_j - m_i)!}\left(\frac{\lambda_j}{\lambda_i}\right)^{m_i}.
$$

Hence, we have proved the following theorem.

Theorem 17.2.1 *Let $\mathbf{N}(t) = (N_1(t), N_2(t), \ldots, N_r(t))$ be a vector of $r \geq 2$ independent Poisson processes, with $\mathbf{E}[N_i(t)] = \lambda_i t$. Then,*

$$\begin{aligned} \Pr[\mathbf{N}(t) &= \mathbf{k}, \textit{ without coincidence} | \mathbf{N}(0) = \mathbf{m}] \\ &= \det \left\| \frac{e^{-\lambda_j t}(\lambda_j t)^{k_j - m_i}}{(k_j - m_i)!} \left(\frac{\lambda_j}{\lambda_i}\right)^{m_i} \right\|_{r \times r} \qquad (17.7) \\ &= \det \left\| \frac{e^{-\lambda_i t}(\lambda_i t)^{k_j - m_i}}{(k_j - m_i)!} \left(\frac{\lambda_j}{\lambda_i}\right)^{k_j} \right\|_{r \times r}, \qquad (17.8) \end{aligned}$$

where (17.8) follows, if we perform the limit in (17.5).

17.3 The r-Node Series Jackson Network

Using the results of the last section, it is not difficult to analyze the transient behavior of a Markovian tandem queue during time periods where all service stations are continuously busy. By a tandem queue or feed-forward network, we mean an arrangement of service stations along a line, such that the output of one node is the input of the next node. There is one external arrival stream at the first node, and customers leaving the last node are leaving the system also. The system we are going to discuss has at each node a single exponential server, and we assume that the external arrival process is Poisson. A scheme of this system is given in Figure 17.1 below.

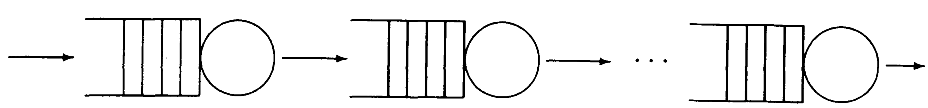


Figure 17.1: A tandem queue

Before we analyze this queueing system in more detail, it is instructive to have a look first at the simple $M/M/1$ model, which is, of course, a special case of the tandem system.

Let $A(t)$ denote the arrival process of an $M/M/1$ queueing system, i.e. $A(t)$ equals the accumulated number of arrivals in the interval $(0, t)$. Similarly, let

$D(t)$ denote the service process, the number of completed services in $(0, t)$. During time periods where the server is continuously busy, the processes are Poisson with rates λ and μ, respectively. Furthermore, let us define a random stopping time

$$T_m = \inf\{t : m + A(t) - D(t) = 0\},$$

where $m > 0$ is the number of customers waiting at time zero, and define $Q(t)$ to be the number of customers in the system at time t.

Our first task will be to rederive a well-known formula for the zero-avoiding transition probabilities of the process $Q(t)$,

$$P^0(m, k; t) = \Pr[Q(t) = k, T_m > t | Q(0) = m], \tag{17.9}$$

which play a fundamental role in the general transient analysis of this queueing system.

For this purpose, observe that the following identity holds:

$$\begin{aligned} &\{Q(t) = k, T_m > t | Q(0) = m\} \\ &\quad \equiv \{m + A(t) - D(t) = k, m + A(s) > D(s), 0 \leq s \leq t\}. \end{aligned} \tag{17.10}$$

Thus, the zero avoiding transition probabilities are just non-coincidence probabilities of the processes $m + A(t)$ and $D(t)$. Figure 17.2 illustrates this correspondence.

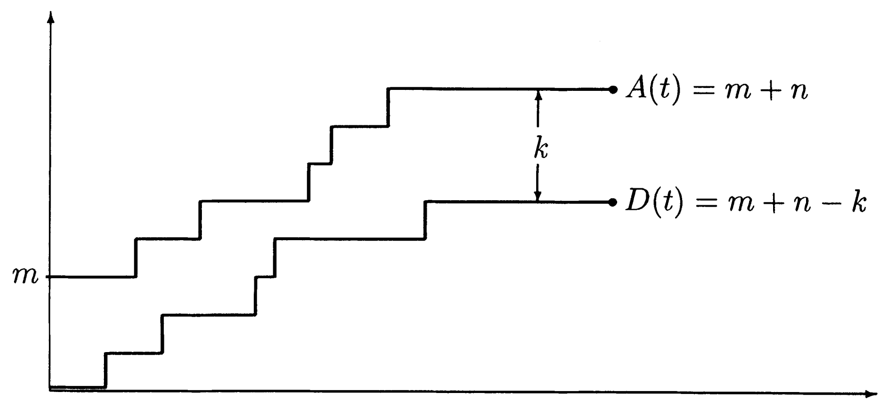


Figure 17.2: Two non-touching sample paths

Now assume that there are $n \geq 0$ arrivals in $(0, t)$, then the process $m + A(t)$ moves from m to $m + n$, and similarly, $D(t)$ moves from 0 to $m + n - k$. Hence, if we apply (17.7), we obtain

$$\begin{aligned} P^0(m, k; t) &= \sum_{n \geq 0} \det \begin{Vmatrix} \frac{e^{-\lambda t}(\lambda t)^n}{n!} & \frac{e^{-\mu t}(\mu t)^{n-k}}{(n-k)!} \left(\frac{\mu}{\lambda}\right)^m \\ \frac{e^{-\lambda t}(\lambda t)^{m+n}}{(m+n)!} & \frac{e^{-\mu t}(\mu t)^{m+n-k}}{(m+n-k)!} \end{Vmatrix} \\ &= e^{-(\lambda+\mu)t} \left[\sum_{n \geq 0} \frac{\lambda^n \mu^{m-k+n} t^{m-k+2n}}{n!(m-k+n)!} - \sum_{n \geq 0} \frac{\lambda^n \mu^{m-k+n} t^{m-k+2n}}{(n-k)!(m+n)!} \right]. \end{aligned}$$

Putting $\rho = \lambda/\mu$ (the traffic intensity of the $M/M/1$ system) and using the definition of the modified Bessel functions

$$I_k(z) = \sum_{n\geq 0} \frac{(z/2)^{k+2n}}{n!(n+k)!},$$

we arrive finally at

$$P^0(m,k;t) = e^{-(\lambda+\mu)t}\rho^{\frac{k-m}{2}}\left[I_{k-m}(2t\sqrt{\lambda\mu}) - I_{k+m}(2t\sqrt{\lambda\mu})\right], \tag{17.11}$$

a well-known formula; see, for example, Prabhu (1965, p. 15).

Exactly the same argument may be used to derive the multivariate analogue of (17.11) for the tandem system.

To fix notation, let $N_0(t)$ denote the arrival process at the first node and $N_i(t), i = 1,\ldots,r$, the accumulated number of services in $(0,t)$ at node i. During time periods where all servers are continuously busy, these processes are Poisson with rates $\lambda_i, i = 0,\ldots,r$. We assume that at time zero there are $m_i > 0$ customers already waiting at node i. Furthermore, let $Q_i(t)$ denote the number of customers at node i at time t and define (in full analogy to simple $M/M/1$) the random stopping times

$$\begin{aligned} T_i &= \inf\{t : Q_i(t) = 0 | Q_i(0) = m_i\} \\ &= \inf\{t : m_i + N_{i-1}(t) - N_i(t) = 0\}. \end{aligned}$$

Put $T = \min_{1\leq i\leq r} T_i$ and define zero-avoiding transition probabilities for the tandem system

$$P^0(\mathbf{m},\mathbf{k};t) = \Pr[\mathbf{Q}(t) = \mathbf{k}, T > t | \mathbf{Q}(0) = \mathbf{m}], \tag{17.12}$$

with

$$\mathbf{Q}(t) = (Q_1(t),\ldots,Q_r(t)),\ \mathbf{m} = (m_1,\ldots,m_r),\ \mathbf{k} = (k_1,\ldots,k_r),$$

and $k_i > 0, i = 1,\ldots,r$. We will show now that (17.12) is again a sum of non-coincidence probabilities, but this time, of the set of independent and dissimilar Poisson processes $N_0(t), N_1(t),\ldots,N_r(t)$. Figure 17.3 illustrates this correspondence for a simple two node network.

Let $N_r(t)$, the service process at the last node, start at height zero. At time zero, there are already m_r customers waiting at this node. Furthermore, the output process of node $r-1$ is the input process at node r. So, we let the service process at node $r-1$ start at height m_r. Since we require that the server at node r is continuously busy in $(0,t)$, the sample paths of the processes $N_{r-1}(t)$ and $N_r(t)$ must not touch. Similarly, let N_{r-2} start at height $m_r + m_{r-1}$, since there are m_{r-1} customers already waiting at node $r-1$. Again, we require that

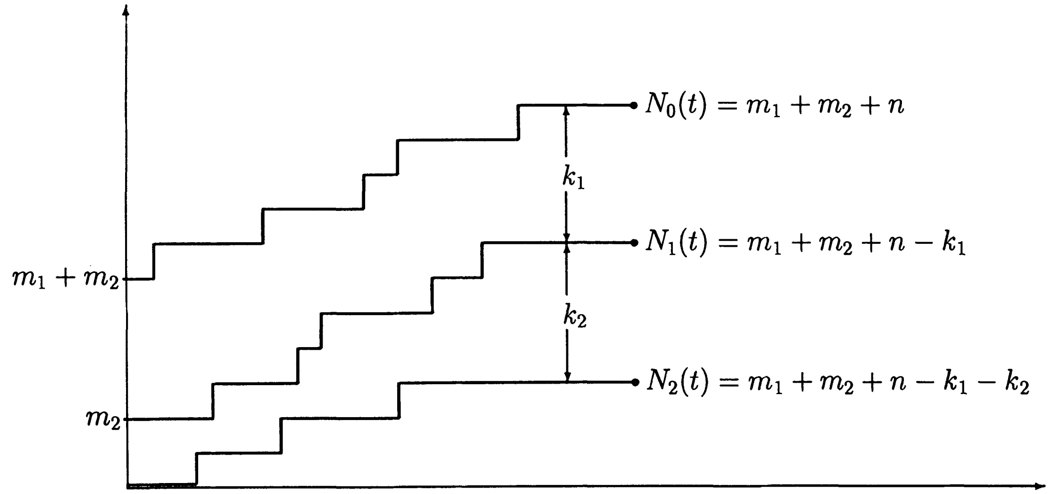


Figure 17.3: Zero-avoiding transitions of a two-node network

the processes $N_{r-2}(t)$ and $N_{r-1}(t)$ to be non-coincident in $(0, t)$. In general, the process $N_i(t)$ starts at height

$$\sum_{\nu=i+1}^{r} m_\nu, \tag{17.13}$$

where we agree to interpret an empty sum as zero.

Now assume that there are $n \geq 0$ arrivals at the first node during $(0, t)$. Thus, the process $N_0(t)$ starts at height $\sum_{\nu=1}^{r} m_\nu$ and terminates at height $n + \sum_{\nu=1}^{r} m_\nu$. Since there must be k_1 customers left at node 1 at time t, the service process at node 1, $N_1(t)$, must terminate at height $m + \sum_{\nu=1}^{r} m_\nu - k_1$, and in general, the process $N_i(t)$ terminates at height

$$n + \sum_{\nu=1}^{r} m_\nu - \sum_{\nu=1}^{i} k_\nu, \qquad i = 0, 1, \ldots, r. \tag{17.14}$$

Again, an empty sum is to be taken as zero.

Thus, we are left with a set of $r + 1$ independent and dissimilar Poisson processes which start at heights (17.13), terminate at heights (17.14), and which are not allowed to have coincidences. Using (17.8), we immediately obtain

$$P^0(\mathbf{m}, \mathbf{k}; t) = \sum_{n \geq 0} \det \left\| \frac{e^{-\lambda_i t} (\lambda_i t)^{b_j - a_i}}{(b_j - a_i)!} \left(\frac{\lambda_j}{\lambda_i} \right)^{b_j} \right\|_{(r+1) \times (r+1)}, \tag{17.15}$$

where

$$a_i = \sum_{\nu=i+1}^{r} m_\nu, \quad b_i = n + \sum_{\nu=1}^{r} m_\nu - \sum_{\nu=1}^{i} k_\nu, \quad i = 0, 1, \ldots, r,$$

which is the desired formula. Essentially, this formula has been derived by Böhm, Jain and Mohanty (1993) using the k-candidate ballot theorem. An equivalent formula in terms of lattice Bessel functions, the multivariate generalizations of the modified Bessel functions, has been given by Massey (1987).

17.4 The Dummy Path Lemma for Poisson Processes

A very useful result in path combinatorics is the *dummy path lemma*; see Krattenthaler and Mohanty (1992) and Stanley (1986, p. 84). It is an immediate consequence of Theorem 17.1.1.

Lemma 17.4.1 *Let $C_1, C_2, \ldots, C_n$ be pairwise distinct points in $\mathbf{Z}^2$. Then, the number of lattice paths with step set $\mathcal{S}$, which lead from (a,b) to point (c,d) and avoid points C_i, $i = 1, 2, \ldots, n$, is given by the determinant*

$$\det_{0\le i,j\le n} \|L(A_j \to E_i)\|, \tag{17.16}$$

where $L(A_j \to E_i)$ denotes the number of paths from point A_j to point E_i, and

$$\begin{aligned} A_0 &= (a,b), \quad A_i = C_i, \quad i = 1,2,\ldots,n \\ E_0 &= (c,d), \quad E_i = C_i, \quad i = 1,2,\ldots,n. \end{aligned}$$

Consider now a Bernoulli process $S(n)$ with success probability p, which starts at time zero in state zero and terminates after n steps in state $k \ge 0$. As before, we will represent the sample paths of $S(n)$ by lattice paths with step set $\mathcal{S} = \{(1,0),(1,1)\}$.

Let $C_i = (u_i, a_i), i = 1, 2, \ldots, r$, be a sequence of pairwise distinct points, such that

$$0 \le u_1 \le u_2 \le \ldots \le u_r \le n, \quad a_i \in \mathbf{N}_0, \quad i = 1, 2, \ldots, r,$$

and define $\mathcal{C}_r = \{C_1, C_2, \ldots, C_r\}$. By the dummy path lemma, the number of paths with step set $\mathcal{S}$, which lead from $(0,0)$ to (n,k) and avoid the points in $\mathcal{C}_r$, is given by the $(r+1) \times (r+1)$ determinant

$$D_{n,k} = \begin{Vmatrix} \binom{n}{k} & \binom{n-u_1}{k-a_1} & \binom{n-u_2}{k-a_2} & \ldots & \binom{n-u_r}{k-a_r} \\ \binom{u_1}{a_1} & 1 & 0 & \ldots & 0 \\ \binom{u_2}{a_2} & \binom{u_2-u_1}{a_2-a_1} & 1 & \ldots & 0 \\ \vdots & \vdots & & \ldots & \vdots \\ \binom{u_r}{a_r} & \binom{u_r-u_1}{a_r-a_1} & \binom{u_r-u_2}{a_r-a_2} & \ldots & 1 \end{Vmatrix}.$$

Any such path has probability $p^k(1-p)^{n-k}$, and thus

$$\Pr[S(n) = k, S(i) \notin \mathcal{C}_r, i = 1, \ldots, n] = D_{n,k} p^k (1-p)^{n-k}, \tag{17.17}$$

where $S(i) \notin \mathcal{C}_r$ denotes the event $(i, S(i)) \notin \mathcal{C}_r$. Absorbing the term $p^k(1-p)^{n-k}$ into the first row of $D_{n,k}$ and transforming the resulting determinant by applying elementary rules, we obtain

$$\Pr[S(n) = k, S(i) \notin \mathcal{C}_r, i = 1, \ldots, n]$$
$$= \left\| \begin{array}{cccc} P(n,k) & P(n-u_1, k-a_1) & \ldots & P(n-u_r, k-a_r) \\ P(u_1, a_1) & 1 & \ldots & 0 \\ P(u_2, a_2) & P(u_2-u_1, a_2-a_1) & \ldots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ P(u_r, a_r) & P(u_r-a_1, a_r-a_1) & \ldots & 1 \end{array} \right\|, \tag{17.18}$$

where

$$P(i,j) = \binom{i}{j} p^j (1-p)^{i-j}.$$

Now, put $\Delta t = t/n$ and $\lambda t/n = p$. Also, set $u_i \Delta t = s_i$, $i = 1, \ldots, r$ and let $n \to \infty$. To perform this limit, consider the generic component in the determinant (17.18), viz., $\lim_{n\to\infty} P(u_i - u_j, a_i - a_j)$. We have

$$\begin{aligned} &\lim_{n\to\infty} P(u_i - u_j, a_i - a_j) \\ &= \lim_{n\to\infty} \binom{u_i - u_j}{a_i - a_j} \frac{(\lambda t)^{a_i - a_j}}{n^{a_i - a_j}} \left(1 - \frac{\lambda t}{n}\right)^{u_i - u_j - a_i + a_j} \\ &= \lim_{n\to\infty} \frac{(u_i - u_j) \cdots (u_i - u_j - a_i + a_j + 1)}{(a_i - a_j)!} \frac{(\lambda t)^{a_i - a_j}}{n^{a_i - a_j}} \left(1 - \frac{\lambda t}{n}\right)^{u_i - u_j - a_i + a_j} \end{aligned}$$

Now, since $u_i = \frac{s_i}{\Delta t} = \frac{s_i n}{t}$, we get

$$\begin{aligned} &\frac{(u_i - u_j)(u_i - u_j - 1) \ldots (u_i - u_j - a_i + a_j + 1)}{(a_i - a_j)!} \frac{\lambda^{a_i - a_j} t^{a_i - a_j}}{n^{a_i - a_j}} \\ &= \frac{(n\frac{s_i}{t} - n\frac{s_j}{t})(n\frac{s_i}{t} - n\frac{s_j}{t} - 1) \ldots (n\frac{s_i}{t} - n\frac{s_j}{t} - a_i + a_j + 1) \lambda^{a_i - a_j} t^{a_i - a_j}}{(a_i - a_j)! n^{a_i - a_j}} \\ &\to \frac{(s_i - s_j)^{a_i - a_j}}{(a_i - a_j)!} \lambda^{a_i - a_j} \quad \text{as } n \to \infty. \end{aligned}$$

Furthermore,

$$\begin{aligned} &\left(1 - \frac{\lambda t}{n}\right)^{u_i - u_j - a_i + a_j} \\ &= \exp\left[(u_i - u_j - a_i + a_j) \ln(1 - \frac{\lambda t}{n})\right] \end{aligned}$$

$$\begin{aligned} &= \exp\left[-\frac{\lambda t}{n}(n\frac{s_i}{t} - n\frac{s_j}{t} - a_i + a_j) - \frac{\lambda^2 t^2}{n^2}(n\frac{s_i}{t} - n\frac{s_j}{t} - a_i + a_j)\ldots\right] \\ &\to e^{-\lambda(s_i - s_j)} \quad \text{as } n \to \infty. \end{aligned}$$

Thus,

$$\lim_{n\to\infty} P(u_i - u_j, a_i - a_j) = \frac{e^{-\lambda(s_i - s_j)}[\lambda(s_i - s_j)]^{a_i - a_j}}{(a_i - a_j)!} \tag{17.19}$$

which is simply the transition function of a Poisson process with rate λ, which moves from state a_j to a_i in the time interval (s_j, s_i).

Hence, we have the following interesting result for Poisson processes.

Theorem 17.4.1 *Let $N(t)$ be a Poisson process with rate $\lambda > 0$ and let*

$$\mathcal{C}_r = \{(s_i, a_i), i = 1, \ldots, r\}$$

be a sequence of pairwise distinct points such that

$$0 \leq s_1 \leq s_2 \leq \ldots \leq s_r \leq t, \quad a_i \in \mathbf{N}_0.$$

Then,

$$\Pr[N(t) = k, N(s) \notin \mathcal{C}_r] = e^{-\lambda t}\lambda^k \left\| \begin{matrix} \frac{t^k}{k!} & \frac{(t-s_1)^{k-a_1}}{(k-a_1)!} & \frac{(t-s_2)^{k-a_2}}{(k-a_2)!} & \cdots & \frac{(t-s_r)^{k-a_r}}{(k-a_r)!} \\ \frac{s_1^{a_1}}{a_1!} & 1 & 0 & \cdots & 0 \\ \frac{s_2^{a_2}}{a_2!} & \frac{(s_2-s_1)^{a_2-a_1}}{(a_2-a_1)!} & 1 & \cdots & 0 \\ \vdots & & & & \vdots \\ \frac{s_r^{a_r}}{a_r!} & \frac{(s_r-s_1)^{a_r-a_1}}{(a_r-a_1)!} & \frac{(s_r-s_2)^{a_r-a_2}}{(a_r-a_2)!} & \cdots & 1 \end{matrix} \right\|. \tag{17.20}$$

17.5 A Special Variant of $D/M/1$ Queues

Stadje (1995) considered the following queueing system: customers arrive at deterministic, but not necessarily equidistant arrival times $0 < a_1 < \ldots < a_n$ at a single exponential server with mean service time $1/\mu$. Stadje derived a determinant formula for the distribution of the number of customers served during a busy period initiated by $m > 0$ customers. In this section, we give a formula for zero-avoiding transition probabilities of this type of queueing system, from which Stadje's result follows as a special case.

Let $D(t)$ denote the number of services completed up to time t during a time interval where the server is continuously busy. Furthermore, let $Q(t)$ denote the number of customers in the system at time t and define the random stopping time

$$T_m = \inf\{t : Q(t) = 0 | Q(0) = m\},$$

the duration of a busy period, if there were $m > 0$ customers present at time zero. Assume that a customer arriving at time a_i joins the system at time a_i+ and define zero-avoiding transition probabilities

$$P^0(m, k; t) = \Pr[Q(t) = k, T_m > t | Q(0) = m].$$

Figure 17.4 shows that the sequence of arrival times gives rise to a sequence of taboo points, which must not be touched by the Poisson process $D(t)$.

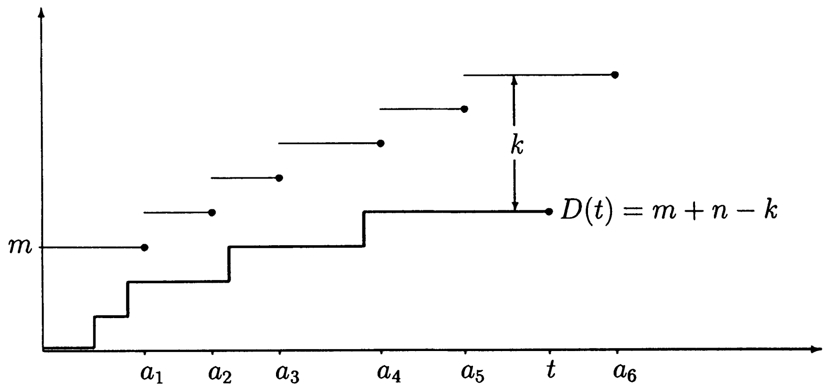


Figure 17.4: Zero-avoiding transitions of $D/M/1$

Thus, it follows immediately from Theorem 17.4.1 that

$$P^0(m, k; t) = \Pr[D(t) = m + n - k, D(s) \notin \mathcal{C}_n], \qquad a_n < t \leq a_{n+1}, \tag{17.21}$$

where

$$\mathcal{C}_n = \{(a_1, m), (a_2, m + 1), \ldots, (a_n, m + n - 1)\}.$$

Hence,

$$P^0(m, k; t) = e^{-\mu t} \mu^{m+n-k}$$
$$\times \begin{Vmatrix} \frac{t^{m+n-k}}{(m+n-k)!} & \frac{(t-a_1)^{n-k}}{(n-k)!} & \frac{(t-a_2)^{n-k-1}}{(n-k-1)!} & \ldots & \frac{(t-a_n)^{-k+1}}{(-k+1)!} \\ \frac{a_1^m}{m!} & 1 & 0 & \ldots & 0 \\ \frac{a_2^{m+1}}{(m+1)!} & \frac{a_2-a_1}{1!} & 1 & \ldots & 0 \\ \vdots & & & & \vdots \\ \frac{a_n^{m+n-1}}{(m+n-1)!} & \frac{(a_n-a_1)^{n-1}}{(n-1)!} & \frac{(a_n-a_2)^{n-2}}{(n-2)!} & \ldots & 1 \end{Vmatrix}. \tag{17.22}$$

Finally, it should be remarked here that exactly the same technique may be used to determine zero-avoiding transition probabilities in a $M/D/1$ system, if the service times are deterministic but not necessarily equally spaced.

References

1. Böhm, W., Jain, J. L. and Mohanty, S. G. (1993). On zero-avoiding transition probabilities of an r-node tandem queue: a combinatorial approach, *Journal of Applied Probability*, **30**, 737–741.

2. Gessel, I. and Viennot, G. (1985). Binomial determinants, paths and hook length formulae, *Advances in Mathematics*, **58**, 300–321.

3. Karlin, S. and McGregor, J. L. (1959). Coincidence probabilities, *Pacific Journal of Mathematics*, **9**, 1141–1164.

4. Karlin, S. (1988). Coincident probabilities and applications in combinatorics, *Journal of Applied Probability*, **25**, 185–200.

5. Krattenthaler, C. and Mohanty, S. G. (1992). Counting tableaux with row and column bounds, In *Proceedings of the Fourth Colloque de Series Formelles et Combinatoire Algebraique* (Eds., C. Reutenauer and P. Leroux), Montreal, Canada.

6. Massey, W. A. (1987). Calculating exit times for series Jackson networks, *Journal of Applied Probability*, **27**, 226–393.

7. Mohanty, S. G. (1979). *Lattice Path Counting and Applications*, New York: Academic Press.

8. Prabhu, N. U. (1965). *Queues and Inventories*, New York: John Wiley & Sons.

9. Stadje, W. (1995). The busy period of some queueing systems, *Stochastic Processes and Their Applications*, **55**, 159–167.

10. Stanley, R. P. (1986). *Enumerative Combinatorics*, Volume 1, Wadsworth & Brooks: New Jersey.

18

Transient Busy Period Analysis of Initially Non-Empty M/G/1 Queues—Lattice Path Approach

Kanwar Sen and Manju Agarwal

University of Delhi, Delhi, India

Abstract: This paper aims at transient busy period analysis of M/G/1 queueing systems starting initially with i customers through *lattice path approach.* The service time distribution is approximated by a 2-phase Cox distribution, C_2. Distributions having rational Laplace-Stieltjes transforms and square coefficient of variation lying in $\left[\frac{1}{2}, \infty\right)$ form a very wide class of distributions. As any distribution of this class can be approximated by a C_2, that has Markovian property, amenable to the application of lattice paths combinatorial analysis, the use of C_2 therefore has led us to achieve transient results applicable to almost any real life queueing system M/G/1.

Keywords and phrases: Transient analysis, busy period, M/G/1 queues, Cox distribution C_2, lattice path approach

18.1 Introduction

As of today, the study of time-dependent behaviour of non-Markovian queues has not been done much, though for real-life systems, such as computer systems and communication systems, time-dependent solutions are even more important than their steady-state solutions. This scarcity of time-dependent solutions is mainly due to the fact that the non-Markovian queues are harder to analyze for their time-dependent behaviour.

Takács (1962) and Benes (1963), in their classic works, analyze the time-dependent behaviour of the M/G/1 queue via two-dimensional transforms. Lucantoni, Choudhury and Whitt (1994) further extended Takács results for the more general arrival process – Batch Markovian Arrival Process (BMAP), i.e., analyzed BMAP/G/1 queue through two-dimensional transforms to numerically compute the transient queue length and waiting time distributions. Then, Logothetis, Mainkar and Trivedi (1996) also developed numerical computational algorithms for the time-dependent solutions of the queue length distribution for a class of such non-Markovian queueing systems for which the queue length process is Markov regenerative. For their purpose, they took general distributions to be deterministic and considered standard finite capacity queues M/G/1/K and GI/M/1/K. They also showed that their algorithms can as well be further extended to BMAP, and multiclass queueing systems. Dalen and Natvig (1980) derived the transient waiting times for a GI/M/1 priority queue. Besides, whatever other time-dependent solutions for the non-Markovian queues are available, they are simply for M/G/1 type queues and that too in terms of *Laplace-Stieltjes transforms* (LSTs) [see Takagi (1991, 1993a,b), Neuts (1989) and Böhm (1993)]. As such before one implements them in practice, their inversion is required. Finding their inversion reduces to the difficult task of locating the roots of a function, polynomial or transcendental. Naturally, this raises the question that what good is the model if its implementation is difficult?

In fact, all the well-known techniques applied to analyze non-Markovian systems reduce to solving a set of probability differential difference equations governing the system behaviour. The standard top-to-bottom approach [see Böhm and Mohanty (1994b)] of solving these equations runs into difficulties because of several types of constraints. Therefore, a bottom-to-top approach [see Böhm and Mohanty (1994a)] of splitting the process at suitable renewable epochs for analysis and combining them using *Lattice Paths* (LPs) combinatorial analysis is worthwhile. This is exactly what is being done here by representing a realization of the process by a LP. A distinctive feature of the LP approach is that it yields explicit transient solutions which are amenable to computations as well as probabilistic interpretations. Through the LP approach, this paper attempts in studying the transient behaviour of a non-Markovian queue, though recently, transient solutions of Markovian queueing models, viz. M/M/1 types, have been successfully obtained by Sen and his coauthors (1993, 1994, 1996).

For the purpose of analyzing M/G/1 queue through the LP approach, the service time distribution G can be approximated by a k-phase Cox distribution C_k, that has Markovian property. Moreover, the distributions C_k are highly versatile, as any probability distribution function having rational LST can be approximated as closely as one wishes by a C_k [Cox (1955)], and as such they generalize all well-known distributions; for example, generalized hyperexponential of order k (GHE$_k$), generalized Erlang of order k (GE$_k$), mixed-generalized Erlang of order k (MGE$_k$), PH (phase type) and k_k (distribution functions whose LSTs are reciprocals of polynomials of degree k) [Botta, Harris and Marchal (1987)]. The distribution C_k consists of k independent exponential phases with service rate μ_j at phase j $(1 \leq j \leq k)$ as shown below (Figure 18.1). After completing service in phase j, the unit either enters phase $j+1$ with probability α_j or completes service with probability β_j $(= 1 - \alpha_j)$.

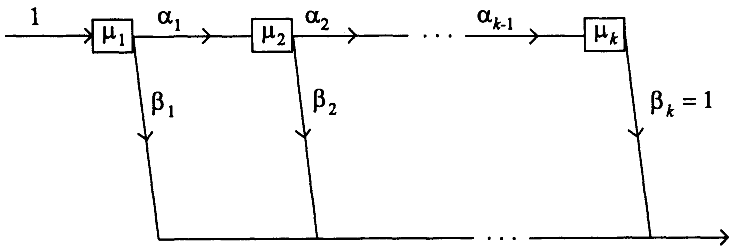


Figure 18.1: k-phase Cox distribution

Cox distributions cover multistage queueing processes where *feedback* may occur as in manufacturing processes where quality control inspections are performed after certain stages and parts that do not meet quality standards are sent back for reprocessing. Cox has shown that no further generality is introduced with feedback and feedforward concept over that of the system in the above figure.

In this paper, we are concerned mainly with the busy period analysis of M/C$_2$/1 queues. The 2-phase distributions C_2, which have 3 parameters only, are desirable both from a theoretical point of view due to ease of analysis as well as practical applications. Moreover, the complexity of the statistical estimation of parameters also reduces considerably [Khoshgoftaar and Perros (1987)]. Besides, they also cover a wide class of distributions that have CV^2, the square of coefficient of variation, lying in $\left[\frac{1}{2}, \infty\right)$ [Marie (1978, 1980)]. As such, the investigations carried out in this paper are fundamental and have a significant role in solving a wide variety of problems occurring in almost any real life situation. A distinctive feature of our paper, in relation to previous papers on transient behaviour of M/G/1 type queues, is that our results are explicit and computable.

It may be mentioned here that results for M/C$_k$/1 and C$_2$/M/1 queues are also under investigation and will be reported by Sen and Agarwal (1996a,b).

The remainder of the paper is organized as follows. In Section 18.2, we describe the LP approach. Section 18.3 contains the discretized model M/C$_2$/1 giving the transition probabilities and results on the LP counting. In Section 18.4, the busy period density for the continuous M/C$_2$/1 model is derived as a limiting case. In Section 18.5, some particular cases are discussed.

18.2 Lattice Path Approach

To conduct transient analysis, we first discretize the system time, i.e., segment the time interval $(0, t]$ into a sequence of t/h (an integer) time slots each of duration h (> 0). In a time slot, the following are the possible events:

(i) either arrival of a new customer

(ii) or a departure, after either phase 1 or phase 2 of service

(iii) or an entry into phase 2 of service

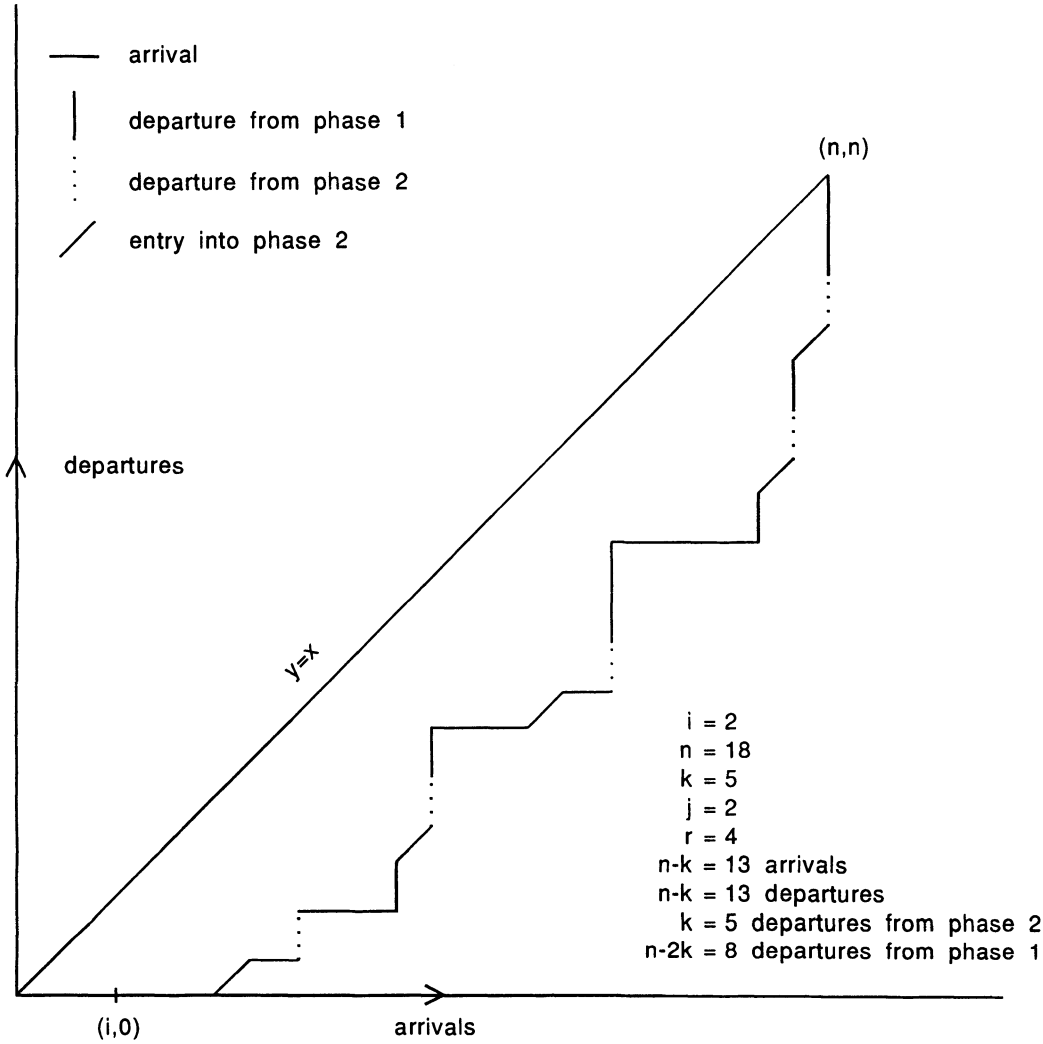
(iv) or none of these. This would be termed as a *stay.*

Obviously, due to the properties of exponential distribution, stochastic processes involving Cox distributions (whose branching probabilities are real) are Markovian. By discretizing the system-time, the process sample paths can be represented as 2-dimensional LPs enabling us to apply *LP Counting Principle.* Then, by a limiting process, the desired transient solution for continuous time can be obtained. The discretized system can be viewed as 2-dimensional LPs representing an arrival, departure, entry into phase 2 service of a customer and stay at a time slot by a horizontal, vertical, diagonal and a point on LP, respectively; see, for example, Figure 18.2. However, the counting of such LPs would depend on the skeleton path (see Figure 18.3) obtained by ignoring the diagonals in Figure 18.2. After the counting of stipulated LPs, appropriate transition probabilities are to be associated with them yielding the desired results for the discretized model. Further, their limiting forms provide corresponding continuous time results.

18.3 Discretized $M/C_2/1$ Model

18.3.1 Transition probabilities

If (x, y) $(x \geq y$ and $x \geq i)$ denotes a vertex, at the end of a time slot, of the LP representing the $M/C_2/1$ queueing process, then in the next time slot, the following transitions are possible:



Note: 1. Two or more consecutive diagonals will not occur.
2. Two or more consecutive dotted verticals will not occur.
3. In any runs of arrivals not more than one diagonal will occur.

Figure 18.2: Lattice path representation
(Illustration for busy period)

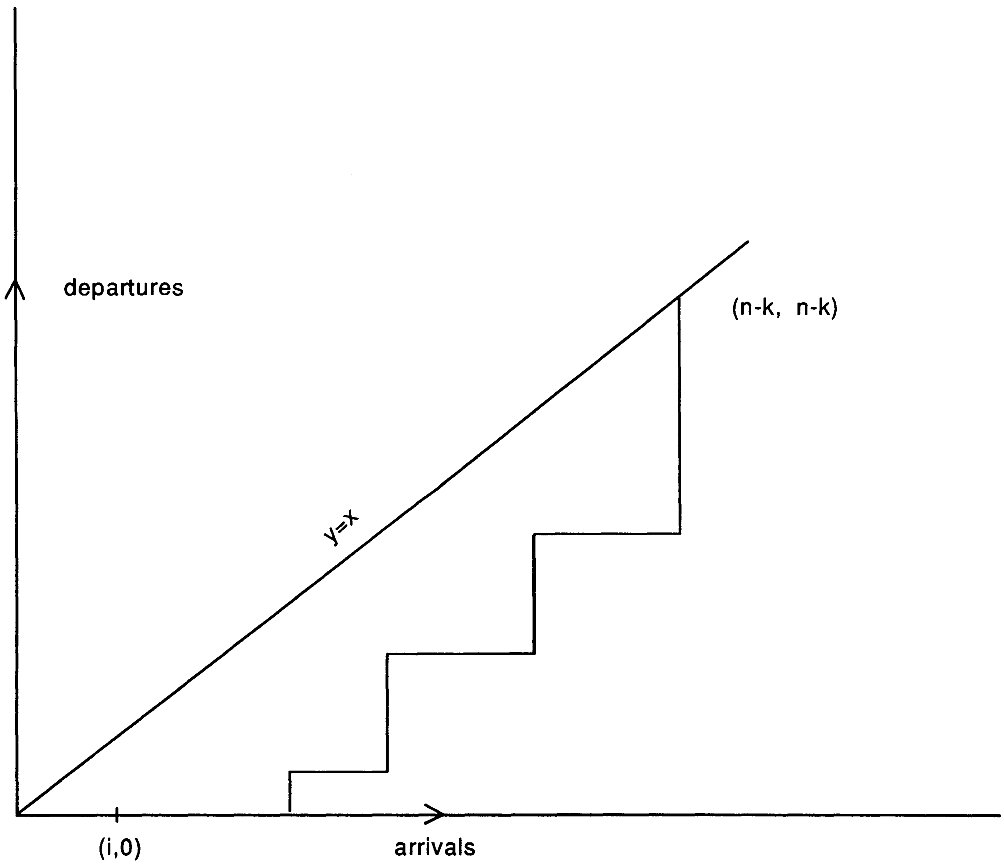


Figure 18.3: Skeleton lattice path

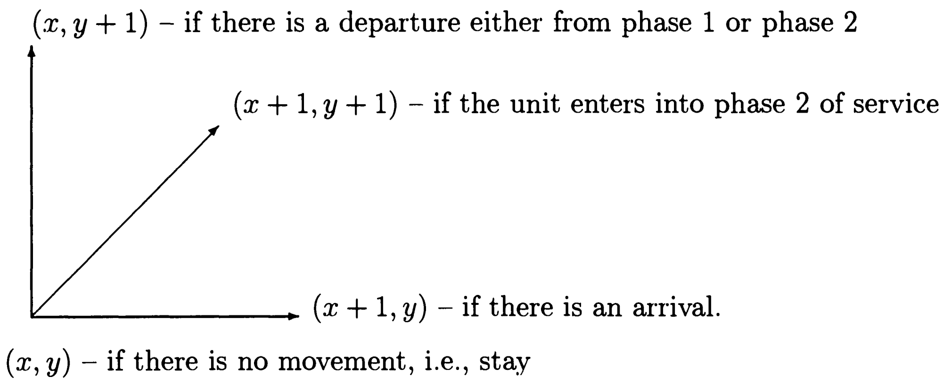


Figure 18.4: Possible transitions from the vertex (x, y)

Let

- λ: exponential inter-arrival rate of customers
- μ_1: exponential service rate in phase 1
- μ_2: exponential service rate in phase 2
- α: Pr[a customer goes to phase 2 after completing phase 1 service]
- i: number of customers initially in the system.

Therefore, the model assumptions lead to

$$
\begin{aligned}
&\text{(i)} \ \Pr[(x,y) \to (x+1,y)] = \Pr[\text{an arrival}] = \lambda h + O(h)\\
&\text{(ii)} \ \Pr[(x,y) \to (x,y+1) \text{ if departure from phase 1}] = \beta\mu_1 h + O(h)\\
&\text{(iii)} \ \Pr[(x,y) \to (x,y+1) \text{ if departure from phase 2}] = \mu_2 h + O(h)\\
&\text{(iv)} \ \Pr[(x,y) \to (x+1,y+1)] = \Pr[\text{entry into phase 2}] = \alpha\mu_1 h + O(h)\\
&\text{(v)} \ \Pr[(x,y) \to (x,y) \text{ if a customer is in phase 1 service}]\\
&\qquad = 1 - (\lambda + \mu_1)h + O(h)\\
&\text{(vi)} \ \Pr[(x,y) \to (x,y) \text{ if a customer is in phase 2 service}]\\
&\qquad = 1 - (\lambda + \mu_2)h + O(h).
\end{aligned}
\tag{18.1}
$$

Stays occurring in (v) and (vi) will be called as type 1 and type 2, respectively.

18.3.2 Counting of lattice paths

In the counting of LPs, concept of run and the following formula will be used:

Run: A sequence of consecutive horizontal (vertical) steps bounded on each side by a vertical (horizontal) step (see Figure 18.3) is called a horizontal (vertical) run of arrivals (departures), respectively. Moreover, the sequence of arrivals starting from the origin followed by the first vertical as well as the sequence of departures at the end following the last arrival, are also regarded as run of arrivals and run of departures, respectively.

Formula: The number of ways of distributing r similar balls into n cells is given by [Feller (1968, p. 38)]

$$\binom{r+n-1}{n-1}. \tag{18.2}$$

Theorem 18.3.1 *For nonnegative integers i, n; k, j; a, b; $r\ (\geq 1)$, $l_1, l_2, \ldots, l_r$; $L_1, L_2, \ldots, L_r$, let $(LP_{(i,n;k,j;r,L;a,b)})$, where $L = (l_1, l_2, \ldots, l_r; L_1, L_2, \ldots, L_r)$, denotes the number of LPs from $(i, 0)$ to (n, n) remaining below the line $y = x$, each comprising of $n-k$ horizontal steps (including those from $(0,0)$ to $(i,0)$), $n-k$ vertical steps and k diagonals, such that*

(a) $n-k$ horizontal steps form r runs of lengths $l_1, l_2, \ldots, l_r$, respectively, satisfying $l_1 \geq i$, $l_2, l_3, \ldots, l_r > 0$ and $\sum_{m=1}^{r} l_m = n-k$

(b) $n-k$ vertical steps form r runs of lengths $L_1, L_2, \ldots, L_r$, respectively, satisfying $L_1, L_2, \ldots, L_r > 0$ and $\sum_{m=1}^{r} L_m = n-k$

(c) $l_1 \geq \max(i, L_1+1)$, $\sum_{i=1}^{m} l_i > \sum_{i=1}^{m} L_i$, $m = 2, \ldots, r-1$, and $\sum_{i=1}^{r} l_i = \sum_{i=1}^{r} L_i = n-k$

(d) j diagonals are inserted one each in any j out of r horizontal runs (including the vertices at both ends of the runs)

(e) the remaining $k-j$ diagonals are inserted at vertices along the vertical runs

(f) 'b' stays of type 2 are distributed at vertices following the k diagonal steps and preceding the subsequent vertical steps

(g) 'a' stays of type 1 are distributed at the remaining vertices which follow the subsequent vertical steps mentioned in case (f) and preceding the next diagonal including those preceding the first diagonal as well as those following the last subsequent vertical, i.e., at vertices other than those used in case (f) for type 2 out of the total $2n-k-i$ vertices.

Then, for $r \geq 1$ and $k > 0$,

$$\left(LP_{(i,n;k,j;r,L;a,b)}\right)$$
$$= \sum_{R_7} \sum_{R_8} \binom{n-k-r}{k-j}$$

$$\times\binom{a+(2n-k-i)-\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)-(k-j)-1}{a}$$
$$\times\binom{b+\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)-1}{b} \tag{18.3}$$

where

$$\begin{aligned} R_7: &\quad \{(i_1,i_2,\ldots,i_j): 1\le i_1<i_2<\cdots<i_j\le r\},\\ R_8: &\quad \{(k_{i_1},k_{i_2},\ldots,k_{i_j}): 0\le k_{i_s}\le l_{i_s},\ s=1,2,\ldots,j\} \end{aligned}$$

and for $r\ge 1$ *and* $k=0$ $(\Rightarrow j=0)$

$$\left(LP_{(i,n;0,0;r,L;a,0)}\right)=\binom{a+(2n-i)-1}{a}. \tag{18.4}$$

PROOF. To prove (18.3), let us suppose that the skeleton path from $(i,0)$ to $(n-k,n-k)$ (see Figure 18.3) consists of r horizontal as well as r vertical runs of lengths l_i and L_i $(i=1,2,\ldots,r)$, respectively. It is obvious that there will be only one unique path with r runs of each type of given fixed lengths. Now, for the insertion of j diagonals, out of the total k, in any j horizontal runs, we suppose that these j diagonals are inserted into runs numbered $i_1,i_2,\ldots,i_j$, respectively, of lengths $l_{i_1},l_{i_2},\ldots,l_{i_j}$, at distances $k_{i_1},k_{i_2},\ldots,k_{i_j}$, from their extreme left end points. Further, as the remaining $k-j$ diagonal steps would therefore be inserted into any $k-j$ vertices out of the remaining $n-k-r$ vertices available, the number of possible ways of doing so is $\binom{n-k-r}{k-j}$. It is easy to see that 'b' stays of type 2, as stated in (f), would be distributed into $\left[\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)\right]$ vertices. By identifying stays with balls and vertices with cells and then using result (18.2), 'b' stays of type 2 can be distributed into $\left[\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)\right]$ vertices in

$$\binom{b+\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)-1}{b} \tag{18.5}$$

ways. Similarly, according to (g), 'a' stays of type 1 are to be distributed in $\left[(2n-k-i)-\left(\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)\right)\right]$ vertices in

$$\binom{a+(2n-k-i)-\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)-(k-j)-1}{a} \tag{18.6}$$

ways.

Multiplying $\binom{n-k-r}{k-j}$ with (18.5) and (18.6) and then summing over all possible j-tuples $(i_1,i_2,\ldots,i_j)$ and $(k_{i_1},k_{i_2},\ldots,k_{i_j})$, we get (18.3).

For (18.4) it is easily seen that when $k=0$, i.e., there is no diagonal in the LP, all 'a' stays of type 1 will be distributed into $(2n-i)$ vertices. Thus, the theorem is proved. ∎

The following corollary is an immediate consequence of Theorem 18.3.1.

Corollary 18.3.1 *Let $\left(LP_{(i,n;k;a,b)}\right)$ denote the number of LPs from $(i,0)$ to (n,n), remaining below the line $y=x$, each comprising of $n-k$ horizontal steps [including horizontal steps from $(0,0)$ to $(i,0)$], $n-k$ vertical steps and k diagonals satisfying conditions (f) and (g) of Theorem 18.3.1. Then, summing (18.3) over r, j and L, we have*

$$\left(LP_{(i,n;k;a,b)}\right) = \sum_{R_4}\sum_{R_5}\sum_{R_6}\left(LP_{(i,n;k,j;r,L;a,b)}\right), \tag{18.7}$$

where $R_4 = \{r : 1 \le r \le n-i\}$, $R_5 = \{j : 0 \le j \le \min(r,k)\}$ and

$$\begin{aligned} R_6 &= \left\{L : l_1 \ge \max(i, L_1+1) \bigcap_{m=2}^{r-1}\left(\sum_{i=1}^{m} l_i > \sum_{i=1}^{m} L_i\right)\right. \\ &\quad \left.\cap\left(\sum_{i=1}^{r} l_i = \sum_{i=1}^{r} L_i = n-k\right)\right\}, \end{aligned}$$

and summing (18.4) over r and L [Narayana (1959)], we get

$$\begin{aligned} \left(LP_{(i,n;0;a,0)}\right) &= \sum_{R_4}\sum_{R_6}\binom{a+(2n-i)-1}{a} \\ &= \binom{a+(2n-i)-1}{a}\frac{i}{2n-i}\binom{2n-i}{n}. \end{aligned} \tag{18.8}$$

18.3.3 Busy period probability

We now derive the main result, i.e., the probability that $\frac{t}{h}$ time slots elapse before the system with initially i customers becomes empty for the first time, i.e., busy period is of length t/h, in Theorem 18.3.3 below.

Theorem 18.3.2 *Let $P_{i,n;k}(t/h)$ denote the probability that the discretized $M/C_2/1$ system starting initially with i customers has busy period of length t/h encountering $n-k-i$ arrivals and $n-k$ departures of which k are through phase 2. Then, for $k > 0$,*

$$\begin{aligned} P_{i,n;k}(t/h) &= \sum_{R_3}\Big[\left(LP_{(i,n;k;a,b)}\right)(\lambda h)^{n-k-i}(\alpha\mu_1 h)^k(\beta\mu_1 h)^{n-2k}(\mu_2 h)^k \\ &\quad \times (1-(\lambda+\mu_1)h)^a(1-(\lambda+\mu_2)h)^b\Big] + O(h), \end{aligned} \tag{18.9}$$

where

$$a = t_1/h, \qquad b = t/h - (2n-k-i) - t_1/h$$

and

$$R_3 = \{t_1/h : 0 \leq t_1/h \leq t/h - \overline{2n-k-i}\}$$

and, for $k = 0$,

$$\begin{aligned} P_{i,n;0}(t/h) &= \binom{t/h-1}{(2n-i)-1} \frac{i}{2n-i} \binom{2n-i}{n} (\lambda h)^{n-i} \\ &\quad \times (\beta\mu_1 h)^n (1-(\lambda+\mu_1)h)^{t/h-2n+i} + O(h). \quad (18.10) \end{aligned}$$

PROOF. Here, the total number of time slots is t/h. Let t_1/h be the number of time slots, out of t/h slots, in which the system has type 1 stays. The total number of transitions in the remaining $t/h - t_1/h$ time slots is $2n-k-i$, leading obviously to $t/h - t_1/h - \overline{2n-k-i}$ stays of type 2. The number of LPs stipulated in the theorem is given by (18.7) for $k > 0$ and by (18.8) for $k = 0$. Therefore, from (18.1), for the case $k > 0$ the probability of occurrence of

(i) t_1/h type 1 stays is $(1-(\lambda+\mu_1)h)^{t_1/h} + O(h)$

(ii) $t/h - t_1/h - \overline{2n-k-i}$ type 2 stays is $(1-(\lambda+\mu_2)h)^{t/h-t_1/h-\overline{2n-k-i}} + O(h)$

(iii) $(n-k-i)$ arrivals is $(\lambda h)^{n-k-i} + O(h)$

(iv) $n-2k$ departures through phase 1 is $(\beta\mu_1 h)^{n-2k} + O(h)$

(v) k entries into phase 2 is $(\alpha\mu_1 h)^k + O(h)$

(vi) k departures through phase 2 is $(\mu_2 h)^k + O(h)$.

Thus, multiplying the number of stipulated LPs from (18.7) by the above transition probabilities, and then summing over t_1/h, we get (18.9). Similarly from (18.8), (18.10) easily follows for the case $k = 0$. ■

Theorem 18.3.3 *Let $f_i(t/h)$ denote the probability that the busy period is of length t/h units for the discretized $M/C_2/1$ model starting initially with i customers. Then, for the case when at least one customer receives service in both the phases*

$$\begin{aligned} f_i(t/h) &= \sum_{R_1}\sum_{R_2}\sum_{R_3}\sum_{R_4}\sum_{R_5}\sum_{R_6}\sum_{R_7}\sum_{R_8} \binom{n-k-r}{k-j} \\ &\quad \times \binom{t_1/h + (2n-k-i) - \sum_{s=1}^{j}(l_{i_s} - k_{i_s} + 1) - (k-j) - 1}{t_1/h} \\ &\quad \times \binom{t/h - (2n-k-i) - t_1/h + \sum_{s=1}^{j}(l_{i_s} - k_{i_s} + 1) + (k-j) - 1}{t/h - (2n-k-i) - t_1/h} \\ &\quad \times (\lambda h)^{n-k-i} (\alpha\mu_1 h)^k (\beta\mu_1 h)^{n-2k} (\mu_2 h)^k (1-(\lambda+\mu_1)h)^{t_1/h} \\ &\quad \times (1-(\lambda+\mu_2)h)^{t/h-(2n-k-i)-t_1/h} + O(h), \quad (18.11) \end{aligned}$$

and when all customers leave the system after phase 1 service

$$f_i\left(\frac{t}{h}\right) = \sum_{R_1} \frac{i}{2n-i}\binom{2n-i}{n}\binom{t/h-1}{(2n-i)-1}(\lambda h)^{n-i} \times (\beta\mu_1 h)^n(1-(\lambda+\mu_1)h)^{t/h-(2n-i)} + O(h), \quad (18.12)$$

where

$$R_1 = \left\{n : i \le n \le \left[\tfrac{1}{2}\,(t/h+i)\right]\right\}$$

and

$$R_2 = \left\{k : 1 \le k \le \left[\frac{n}{2}\right]\right\},$$

where $[x]$ denotes the largest integer in x.

PROOF. As envisaged in the theorem, the expression (18.11) for $f_i(t/h)$ is obtained by summing (18.9) over k and n and then using (18.7) and (18.3). Eq. (18.12) follows from (18.10) by summing it over n. ■

18.4 Continuous M/C$_2$/1 Model

On using a limiting process as $h \to 0$ [Meisling (1958)], we obtain the expression for the busy period density function as given in the following theorem.

Theorem 18.4.1 *The probability density function of the busy period for M/G/1 system starting initially with i customers is given by*

$$\begin{aligned} f_i(t) &= \frac{i}{t}\, e^{-(\lambda+\mu_1)t}\left(\frac{\lambda}{\beta\mu_1}\right)^{-i/2} I_i(2\sqrt{\lambda\beta\mu_1}t) \\ &+ \sum_{R_1}\sum_{R_2}\sum_{R_4}\sum_{R_5}\sum_{R_6}\sum_{R_7}\sum_{R_8}\Bigg[\binom{n-k-r}{k-j}\alpha^k\beta^{n-2k}(\lambda\mu_1)^{-k/2} \\ &\times (\lambda/\mu_1)^{-i/2}\mu_2^k(\mu_1/\mu_2)^{\frac{1}{2}[\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)]} \\ &\times \int_0^t \frac{e^{-(\lambda+\mu_1)t_1}\, e^{-(\lambda+\mu_2)(t-t_1)}(\sqrt{\lambda\mu_1}t_1)^{(2n-k-i)-\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)-(k-j)}}{t_1(t-t_1)\Gamma\left((2n-k-i)-\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)-(k-j)\right)} \\ &\times \frac{(\sqrt{\lambda\mu_2}(t-t_1))^{\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)}}{\Gamma\left(\sum_{s=1}^{j}(l_{i_s}-k_{i_s}+1)+(k-j)\right)}\, dt_1\Bigg] \end{aligned} \quad (18.13)$$

The first term in (18.13) corresponds to the case when no unit receives service in phase 2 and the second term in (18.13) corresponds to the case when at least one unit receives service in phase 2.

18.5 Particular Cases

(i) M/M/1: Taking $\alpha = 0$, $\beta = 1$, $\mu_1 = \mu$, $\mu_2 = 0$ and hence $k = j = 0$, (18.13) reduces to

$$f_i(t) = \frac{i}{t}\, e^{-(\lambda+\mu)t} \left(\frac{\lambda}{\mu}\right)^{-i/2} I_i(2\sqrt{\lambda\mu}t). \tag{18.14}$$

This result is the same as that obtained by Saaty (1961, p. 128) and Sen and Jain (1993).

(ii) Similarly, busy period density for M/GE$_2$/1 model can be obtained from (18.13) by taking $\alpha = 1$, $\beta = 0$.

(iii) For the M/HE$_2$/1 model, the busy period density can also be obtained from (18.13) by taking $\alpha = q(\mu_1 - \mu_2)/\mu_1$, where the pdf of HE$_2$ is given by $p\mu_1 e^{-\mu_1 x} + q\mu_2 e^{-\mu_2 x}$ $(p + q = 1)$, $x > 0$.

References

1. Benes, V. E. (1963). *General Stochastic Processes in the Theory of Queues*, Reading, MA: Addison-Wesley.

2. Böhm, W. M. (1993). *Markovian Queueing Systems in Discrete Time*, Frankfurt am Main: Antonhain.

3. Böhm, W. M. and Mohanty, S. G. (1994a). On discrete time Markovian N-policy queue involving batches, *Sankhyā, Series A*, **56**, 1–20.

4. Böhm, W. M. and Mohanty, S. G. (1994b). Transient analysis of queues with heterogeneous arrivals, *Queueing Systems*, **18**, 27–45.

5. Botta, R. F., Harris, C. M. and Marchal, W. G. (1987). Characterization of generalized hyperexponential distribution functions, *Communications in Statistics—Stochastic Models*, **3**, 115–148.

6. Cox, D. R. (1955). A use of complex probabilities in the theory of stochastic processes, *Proceedings of the Cambridge Philosophical Society*, **51**, 313–319.

7. Dalen, G. and Natvig, B. (1980). On the transient waiting times for a GI/M/1 priority queue, *Journal of Applied Probability*, **17**, 227–234.

8. Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Volume 1*, Third edition, New York: John Wiley & Sons.

9. Khoshgoftaar, T. M. and Perros, H. G. (1987). A comparison of three methods of estimation for approximating general distributions by a Coxian distribution, In *Modelling Techniques and Performance* (Eds., S. Fdida and G. Fujolle), Amsterdam: North-Holland.

10. Logothetis, D., Mainkar, V. and Trivedi, K. (1996). Transient analysis of non-Markovian queues via Markov regenerative processes, In *Probability Models and Statistics—J. Medhi Festschrift*, 109–131, New Delhi: New Age International Publishers.

11. Lucantoni, D. M., Choudhury, G. L. and Whitt, W. (1994). The transient BMAP/G/1 queue, *Communications in Statistics—Stochastic Models*, **10**, 145–182.

12. Marie, R. (1978). Methods iteratives de resolution de models mathematiques de systemes informatiques, *R.A.I.R.O. Informatique/Computer Sciences*, **12**, 107–122.

13. Marie, R. (1980). Calculating equilibrium probabilities for $\lambda^{(n)}/C_k/1/N$ queues, ACM SIGMETRICS, *Conference on Measurement and Modelling of Computer Systems*, 117–125.

14. Meisling, T. (1958). Discrete time queueing theory, *Journal Operations Research Society of America*, **6**, 96–105.

15. Narayana, T. V. (1959). A partial order and its applications to probability theory, *Sankhyā*, **21**, 91–98.

16. Neuts, M. F. (1989). *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, New York: McGraw-Hill.

17. Saaty, T. L. (1961). *Elements of Queueing Theory with Applications*, New York: McGraw-Hill.

18. Sen, K. and Agarwal, M. (1996a). Lattice path approach to transient busy period analysis of GI/M/1 queues using Cox distribution, (under preparation).

19. Sen, K. and Agarwal, M. (1996b). Transient solutions of $M/C_k/1$ queues, (under preparation).

20. Sen, K. and Gupta, R. (1994). Transient analysis of threshold T-policy M/M/1 queues with server control, *Sankhyā, Series B*, **56**, 39–51.

21. Sen, K. and Gupta, R. (1997). Transient solution of $M^b/M/1$ system under threshold control policies, *Journal of Statistical Research* (to appear).

22. Sen, K. and Gupta, R. (1997). Discrete time Markovian batch arrival queues with restricted vacations and non-exhaustive service, *Sankhyā, Series B* (submitted for publication).

23. Sen, K. and Jain, J. L. (1993). Combinatorial approach to Markovian queueing models, *Journal of Statistical Planning and Inference*, **34**, 269–279.

24. Sen, K., Jain, J. L. and Gupta, J. M. (1993). Lattice path approach to transient solution of M/M/1 with control policy, *Journal of Statistical Planning and Inference*, **34**, 259–267.

25. Takács, L. (1962). *Introduction to Theory of Queues*, New York: Oxford University Press.

26. Takagi, H. (1991). *Queueing Analysis, 1: Vacation and Priority System*, Amsterdam: North-Holland.

27. Takagi, H. (1993a). *Queueing Analysis, 2: Finite Systems*, Amsterdam: North-Holland.

28. Takagi, H. (1993b). *Queueing Analysis, 3: Discrete Time Systems*, Amsterdam: North-Holland.

19

Single Server Queueing System with Poisson Input: A Review of Some Recent Developments

J. Medhi

Institute of Advanced Study in Science & Technology, Guwahati, India

Abstract: The classical single-server queue with Poisson input has been extended to include several types of generalizations, to which attention has been paid by several researchers. It would be worthwhile to look into some sort of unified approach of the results of several models. It is also of interest to investigate a few important performance measures through heuristic 'mean value analysis' only without looking through the 'Laplacian Curtain'. The purpose of this paper is to make a comprehensive survey of many interesting results that have appeared recently with a view to unify the results and to derive some new ones. The review is based on recent contributions as listed in the References.

Keywords and phrases: N (threshold)-policy, vacations, residual distribution

19.1 Introduction

In this paper some extensions and generalizations of the classical $M/G/1$ queueing system are discussed. In a series of papers, Takagi (1991, 1992, 1993), Takine, Takagi and Hasegawa (1993) extended classical $M/G/1$ to cover cases of models with vacations as well as those with finite buffer or finite calling population or both. Lee *et al.* (1994, 1995, 1996) and Chae and Lee (1995) discussed $M^X/G/1$ systems with batch arrivals and with vacations. While they confined to continuous systems in steady state, in a series of papers, using combinatorial methods, Böhm and Mohanty (1993, 1994a,b), and Mohanty and Panny (1989, 1990) discussed transient analysis of Markovian systems in discrete time. They observed that the results for continuous time processes can be obtained from those of discrete time processes by appropriate limiting procedure. Böhm and

Mohanty assert that for certain systems at least, such as digital communication and data transmission, discrete time analysis is more appropriate while continuous time treatment is at best an approximation.

Here we shall be concerned mainly with systems in continuous time and in steady state, and also to some specific performance measures only. A survey of some results of finite systems as obtained by Takagi, Takine *et al.* and others is given in Medhi (1994).

The notation and terminology used are as described below. Units (customers/messages) arrive at a single server service facility in a Poisson stream with rate λ. In case of batch arrival we shall assume Poisson streams of arrivals at rate λ, with random group size X, that is, compound Poisson arrivals with rate $\lambda E(X)$. For arrival rate, we shall use $\lambda E(X)$ in place of λ in case of batch Poisson arrivals. Service time B is general having $B(\cdot)$ for its d.f. and $B^*(\cdot)$ as its LST. The mean service time is b $(= 1/\mu)$ and higher moments are $b^{(r)} = E(B^r)$, $r = 2, 3, \ldots$. The offered load is $a = \lambda b$, and the utilization factor (fraction of time the server is busy) is denoted by ρ. A busy period T is the period during which the server remains busy; ordinarily it commences with the start of service of a unit that arrives to an empty system and lasts till the system becomes empty again. This concept needs to be modified in case of vacation and/or control policy queues. The idle period (of the server) is denoted by I and the cycle time by C $(= T + I)$, the average waiting time in the queue (queueing time) of a unit by W_Q, and the average number in the queue by L_Q. P_k denotes the probability that the system (queue and service) contains k units at an arbitrary point of time. P_0 denotes the probability that the system is empty. Once W_Q is found, the average waiting time in the system W can be found; by employing Little's law, the average number in the queue and system can be obtained. One can thus have some of the important performance measures.

It may be recalled that for $M/G/1$ system, the following hold: $\rho = a$, $E(I) = 1/\lambda$, and

$$\begin{aligned} E(T) &= \frac{1}{\mu - \lambda} = \frac{b}{1-a} = \left(\frac{a}{1-a}\right) E(I) \\ &= \left(\frac{\rho}{1-\rho}\right) E(I) \qquad (\text{since } \rho = a); \end{aligned} \tag{19.1}$$

when the busy period starts with m (> 1) in the system the expected busy period equals $m\{b/(1-a)\}$.

The average waiting time in the queue (of a test customer) is easily seen to be

$$W_Q = L_Q E(B) + E(B_R) \Pr[A], \tag{19.2}$$

where $\Pr[A]$ is the probability that the server is busy and B_R is the residual service time of the unit in service as seen by the test customer on arrival. Now

$$\frac{E(B^2)}{2E(B)} = \frac{b^{(2)}}{2b}$$

is the expectation of the forward recurrence time of the service time (or residual service time of the unit in service). Using $L_Q = \lambda W_Q$, one gets

$$W_Q = aW_Q + \rho \, \frac{b^{(2)}}{2b}$$

or

$$W_Q = \frac{1}{1-a}\left(\frac{\lambda b^{(2)}}{2}\right) = \frac{\lambda b^{(2)}}{2(1-\rho)}, \tag{19.3}$$

which is the Pollackzek-Khinchine Formula.

19.2 Exceptional Service for the First Unit in Each Busy Period

Takagi (1991, 1992) examined this problem considered earlier by Welch. The service time of the first unit in each busy period is B_0 with mean b_0 and $E(B_0^r) = b_0^{(r)}$, $r = 2, 3, \ldots$, while the service time of the subsequent units is denoted by B [with $E(B^r) = b^{(r)}$, $r = 2, 3, \ldots$, $E(B) = b$]. While $E(I) = 1/\lambda$,

$$E(T) = b_0 + \frac{(\lambda b_0)b}{1-a} = \frac{b_0}{1-a}, \tag{19.4}$$

$$\rho = \frac{E(T)}{E(I) + E(T)} = \frac{\lambda b_0}{1 + \lambda b_0 - a} \tag{19.5}$$

and

$$P_0 = 1 - \rho = \frac{1-a}{1 + \lambda b_0 - a}; \tag{19.6}$$

further, $E(T) = \frac{\rho}{1-\rho}\, E(I)$ holds.

Denote

$B_1 \equiv$ generalized service time

$= B_0$ or B according as the unit is the first to be served in a busy period or not;

by conditioning, we get

$$\begin{aligned} E(B_1) &= b_0 P_0 + b(1 - P_0) \\ &= \frac{b_0}{1 + \lambda b_0 - a} = \frac{\rho}{\lambda}\,; \end{aligned} \tag{19.7}$$

that is, $\rho = \lambda E(B_1)$. Further,

$$\begin{aligned} E(B_1^2) &= b_0^{(2)} P_0 + b^{(2)}(1 - P_0) \\ &= \frac{(1-a) b_0^{(2)} + \lambda b_0 b^{(2)}}{1 + \lambda b_0 - a}. \end{aligned}$$

Thus the mean residual (generalized) service time is given by

$$E(B_{1R}) = \frac{E(B_1^2)}{2E(B_1)} = \frac{(1 - a_0) b_0^{(2)} + \lambda b_0 b^{(2)}}{2b_0} \tag{19.8}$$

(writing B_{1R} as the residual generalized service time). One gets as in (19.2)

$$W_Q = L_Q\, E(B) + \Pr[A] E(B_{1R})$$

or

$$W_Q = \frac{\rho}{1-a}\, E(B_{1R}) = \frac{\rho}{1-a}\, \frac{E(B_1^2)}{2E(B_1)}$$

which can be put in the form

$$W_Q = \frac{\lambda b^{(2)}}{2(1-a)} + \frac{\lambda (b_0^{(2)} - b^{(2)})}{2(1 + \lambda b_0 - a)}\,. \tag{19.9}$$

When $B_0 \equiv B$, then the second term vanishes and (19.9) reduces to (19.3).

Considering that the first unit of the busy period has service time $B + \Delta$, where $\{\Delta\}$ is a sequence of i.i.d. random variables independent of other random variables, that is, $B_0 = B + \Delta$, one gets the same result as above.

19.3 $M/G/1$ With Random Setup Time S

Here the first message (customer) in a busy period needs a setup time of random duration S before service is started. The expected idle and busy periods are, respectively,

$$E(I) = \frac{1}{\lambda} + E(S) \tag{19.10}$$

$$E(T) = \frac{\lambda E(I) b}{1-a} = \frac{\{1 + \lambda E(S)\} b}{1-a} = \frac{b + aE(S)}{1-a} \tag{19.11}$$

and the utilization factor $\rho = \frac{E(T)}{E(T)+E(I)} = a$, which equals the probability that the server is busy in a standard $M/G/1$ queue without setup time. The utilization factor is not affected by the setup time; further $E(T) = \left(\frac{\rho}{1-\rho}\right) E(I)$ holds. Consider that the delay cycle D starts as soon as a unit arrives to an empty system and continues till the system becomes empty again. It is generated by the setup period and the first service time. Thus

$$E(D) = \{E(S) + b\} + \frac{[\lambda\{E(S) + b\}]b}{1-\rho} = \frac{1}{1-\rho}\,\{E(S) + b\}.$$

The delay cycle D and the empty period (duration of state 0) L form an alternating renewal process. Thus

$$P_0 = \frac{E(L)}{E(L) + E(D)} = \frac{1-\rho}{1+\lambda E(S)} \tag{19.12}$$

so that $P_0 = 1 - \rho$, only when $E(S) \equiv 0$, that is, when there is no setup time. The expected waiting time in the queue W_Q is given by

$$\begin{aligned} W_Q \quad = \quad & L_Q E(B) + E(\text{Residual service time})\,\Pr[\text{server is busy}] \\ & + \{E(\text{setup time})\,\Pr[A] \\ & + E(\text{Residual setup time})\,\Pr[B]\}\,\Pr[\text{server is idle}], \end{aligned}$$

where

$$\begin{aligned} A \quad \equiv \quad & \text{event that the test customer is the first to arrive} \\ & \text{in the idle period,} \\ B \quad \equiv \quad & \text{event that the test customer arrives during the} \\ & \text{setup time.} \end{aligned}$$

We get

$$\begin{aligned} \Pr[A] \quad &= \quad \frac{1}{\text{Expected no. of arrivals in the idle period}} \\ &= \quad \frac{1}{1+\lambda E(S)} \end{aligned}$$

and

$$\Pr[B] = \frac{E(\text{setup period})}{E(\text{idle period})} = \frac{E(S)}{(1/\lambda) + E(S)} = \frac{\lambda E(S)}{1+\lambda E(S)}\ .$$

Thus

$$(1-a)W_Q = \rho\,\frac{b^{(2)}}{2b} + (1-\rho)\left[\frac{E(S)}{1+\lambda E(S)} + \frac{E(S^2)}{2E(S)}\,\frac{\lambda E(S)}{1+\lambda E(S)}\right]$$

and since $\rho = a$, we get

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{2E(S) + \lambda E(S^2)}{2\{1 + \lambda E(S)\}} . \tag{19.13}$$

The second term is the expected delay due to the setup time; it vanishes when there is no setup time.

19.4 $M/G/1$ System Under N-Policy

System with control (threshold) policy was first considered by Yadin and Naor (1963). Heyman (1968) investigated some of its optimal properties. Here, the server, after termination of a busy period, waits until there are a specified number $N \geq 1$ (a pre-assigned number called *threshold*) of arrivals; it starts service with the arrival of the N-th unit, and the busy period starts and lasts till none is left in the system (exhaustive service discipline). More general service disciplines have also been considered in the literature [Takagi (1991)]. We shall confine here to exhaustive service discipline.

We have

$$E(I) = N/\lambda \qquad \text{and} \qquad E(T) = \frac{Nb}{1-a} .$$

The utilization factor

$$\rho = \frac{E(T)}{E(T) + E(I)} = \lambda b = a,$$

is independent of N. Again $E(T) = \frac{\rho}{1-\rho} E(I)$ holds.

The expected waiting time is given by

$$W_Q = L_Q E(B) + E(B_R) \Pr[\text{server is busy}] + E(I_R) \Pr[\text{server is idle}], \tag{19.14}$$

where $B_R \equiv$ residual service period, and $I_R \equiv$ residual idle (buildup) period (that is, the duration from the instant of arrival of the test customer to the instant when N-th customer arrives). Conditioning on the order in which the test customer arrives, we get

$$E(I_R) = \left(\frac{N-1}{\lambda} + \frac{N-2}{\lambda} + \cdots + \frac{1}{\lambda} + 0\right)\frac{1}{N} = \frac{N-1}{2\lambda} . \tag{19.15}$$

The mean system size during an idle period equals $\sum_{i=0}^{N-1} i\left(\frac{1}{N}\right) = \frac{N-1}{2}$ so that

$$E(I_R) = (\text{mean system size during an idle period})/(\text{arrival rate}).$$

From (19.14) one gets

$$(1-a)W_Q = \rho \frac{b^{(2)}}{2b} + (1-\rho)\left(\frac{N-1}{2\lambda}\right)$$

or

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{N-1}{2\lambda}\ . \tag{19.16}$$

The second term is due to the threshold N; it vanishes when $N = 1$.

19.5 $M/G/1$ Under N-Policy and With Setup Time

Here, once the system becomes empty the server waits till exactly N (a preassigned number) units have accumulated: this is the buildup period which is followed by a setup period S. The buildup and setup periods together constitute the idle period. Thus

$$E(I) = (N/\lambda) + E(S) = \{N + \lambda E(S)\}/\lambda$$

and

$$E(T) = \frac{\{N + \lambda E(S)\}b}{1-a}\ , \tag{19.17}$$

since the busy period starts with an expected number of $N + \lambda E(S)$ customers. Further, $\rho = \frac{E(T)}{E(T)+E(I)} = a$, independent of threshold and setup time. That the utilization factor is independent of the buildup period and setup period was shown by Medhi and Templeton (1992). Thus, $E(T) = \left(\frac{\rho}{1-\rho}\right) E(I)$ holds.

The expected waiting time is given by

$$\begin{aligned} W_Q \quad = \quad & L_Q E(B) + E(\text{residual service time}) \Pr[\text{server is busy}] \\ & + [E(\text{residual buildup period plus the setup period}) \\ & \qquad \times \Pr[A \mid \text{server is idle}] \\ & + E(\text{residual setup period}) \Pr[B \mid \text{server is idle}]] \\ & \qquad \times \Pr[\text{server is idle}], \end{aligned}$$

where A is the event that the test customer arrives in the buildup period, and B is the event that the test customer arrives in the setup period. Now

$$\Pr[A \mid \text{server is idle}] \quad = \quad \frac{N/\lambda}{(N/\lambda) + E(S)} = \frac{N}{N + \lambda E(S)} \tag{19.18}$$

$$\Pr[B \mid \text{server is idle}] \quad = \quad \frac{\lambda E(S)}{N + \lambda E(S)}\ . \tag{19.19}$$

Thus, we get

$$
\begin{aligned}
(1-a)W_Q &= \frac{b^{(2)}}{2b}\,\rho + (1-\rho)\Bigg[\left\{\frac{N-1}{2\lambda} + E(S)\right\}\frac{N}{N+\lambda E(S)} \\
&\qquad + \frac{E(S^2)}{2E(S)}\,\frac{\lambda E(S)}{N+\lambda E(S)}\Bigg].
\end{aligned}
$$

Thus,

$$
W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{N(N-1) + 2\lambda N E(S) + \lambda^2 E(S^2)}{2\lambda[N+\lambda E(S)]} \tag{19.20}
$$

which is the result obtained by Takagi (1992); this result was obtained earlier by Yadin and Naor (1963).

Putting $N = 1$, one gets the corresponding result for the system without threshold in (19.13); putting $E(S) = E(S^2) = 0$, one gets the corresponding result for the system without setup time in (19.16).

19.6 Queues With Vacation: $M/G/1$ Queueing System With Vacation

Consider a situation in which the server, as soon as the system become empty, goes on a vacation of random duration V, where V is the generic random variable of the sequence of vacation periods $V_1, V_2, \ldots$, which are i.i.d. random variables. At the termination of the first vacation period the server returns to the service facility. First, consider the situation where the server, finding no one when he returns from the first vacation, goes for a second vacation, then a third vacation and so on till he finds one or more waiting at the end of a vacation, each vacation period being of the same random duration V. This is a case of multiple vacations. Secondly, if he finds none waiting at the termination of the first vacation, he does not take any further vacation but waits till an arrival occurs and begins service. This is the case of single vacation. An $M/G/1$ queueing system with multiple vacation is denoted by $M/G/1 - V_m$ and that with a single vacation by $M/G/1 - V_s$. In either case it is assumed that service is exhaustive, that is, the service will continue until none is left in the system. For a survey on vacation queue, refer to Doshi (1986) and for an exhaustive account to Takagi (1991, Vol. I).

19.7 $M/G/1 - V_m$ System

The server idle period I is the duration of time from the instant the server leaves for a vacation until the instant the next service begins.

We have

$$\begin{aligned} V^*(\lambda) &= \int_0^{\infty} e^{-\lambda t} dV(t) \\ &= \Pr[\text{no customer arrives during a vacation period } V] \end{aligned} \tag{19.21}$$

and $1-V^*(\lambda) = \Pr[\text{at least one arrives during the period } V]$. Thus I is the sum of a random number of vacation periods, that is, the sum of random number of i.i.d. random variables. The number is a geometric random variable with probability of success $1 - V^*(\lambda)$ and having mean $1/[1 - V^*(\lambda)]$. Thus,

$$E(I) = \frac{E(V)}{1 - V^*(\lambda)} \tag{19.22}$$

and the average number of arrivals during an idle period is $\lambda E(I)$. Thus, the expected busy period equals

$$E(T) = \frac{\lambda E(V)}{1 - V^*(\lambda)} \left(\frac{b}{1-a} \right) = \frac{aE(V)}{(1-a)[1 - V^*(\lambda)]} \tag{19.23}$$

and that $\rho = \frac{E(T)}{E(T)+E(I)} = a$, independent of the vacation policy. That $\rho = a$ holds is intuitively clear. Böhm and Mohanty (1994b) show that $\rho = a$ holds in the discrete time $M/M/1$ queue with vacation. It follows that $E(T) = \frac{\rho}{1-\rho} E(I)$ holds.

The expected waiting time is given by

$$\begin{aligned} W_Q &= L_Q E(B) + E(\text{residual service time}) \Pr[\text{server is busy}] \\ &\quad + E(\text{residual vacation time}) \Pr[\text{server is idle}]. \end{aligned}$$

Thus

$$\begin{aligned} W_Q &= \frac{1}{1-a} \left[\frac{b^{(2)}}{2b} \rho + \frac{E(V^2)}{2E(V)} (1-\rho) \right] \\ &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(V)} \, . \end{aligned} \tag{19.24}$$

The second term is due to the vacation policy.

19.8 $M/G/1 - V_m$ With Exceptional First Vacation

Lee (1988) considered such a model. Here the server, as soon as the system becomes empty, goes for a vacation of random duration V. If, on return from the first vacation, he finds some unit(s) waiting he starts service till the system becomes empty again; if however, he does not find any one waiting at the end of the first vacation, he goes for a second vacation and then for a third vacation and so on until he finds some unit(s) waiting on return from a vacation, the duration of the second and subsequent vacations being of random duration U. Here the first vacation is of exceptional length given by random variable V, and subsequent vacations of random duration given by i.i.d. random variable U. When $V \equiv U$, we get the case discussed in the preceding section. The model is applicable in a situation where at the end of a busy period, the server takes a vacation of random duration V and if he finds none waiting on return from the first vacation he goes for other jobs, each of which is of random duration U.

Since $V^*(\lambda)$ and $U^*(\lambda)$ give, respectively, the probability that no one arrives during the first vacation of duration V and subsequent vacations each of which is of duration U, we get, on conditioning

$$\begin{aligned} E(I) &= [1 - V^*(\lambda)]E(V) + V^*(\lambda)\left(E(V) + \frac{E(U)}{1 - U^*(\lambda)}\right) \\ &= E(V) + V^*(\lambda)\left(\frac{E(U)}{1 - U^*(\lambda)}\right). \end{aligned} \tag{19.25}$$

The expected number of arrivals during an idle period is $\lambda E(I)$ and thus the expected length of a busy period is given by

$$E(T) = \lambda E(I)\left(\frac{b}{1-a}\right) = \frac{a}{1-a}\,E(I). \tag{19.26}$$

Thus

$$\rho = \frac{E(T)}{E(T) + E(I)} = a,$$

which is independent of the vacation policy, and that

$$E(T) = \frac{\rho}{1-\rho}\,E(I)$$

holds.

The expected waiting time W_Q is given by

$$\begin{aligned} W_Q &= L_Q\,E(B) + E(\text{residual service time})\,\Pr[\text{server is busy}] \\ &\quad +\{E(\text{residual first vacation})\,\Pr[A|C] \\ &\quad +E(\text{residual vacation other that the first})\,\Pr[B|C]\} \\ &\qquad \times \Pr[\text{server is idle}], \end{aligned} \tag{19.27}$$

where

A is the event that server is on first vacation,
B is the event that server is on a vacation other than the first,
C is the event that server is idle.

We have

$$\Pr[A|C] = \frac{E(V)}{E(I)} = \frac{E(V)[1 - U^*(\lambda)]}{E(V)[1 - U^*(\lambda)] + V^*(\lambda)E(U)} \tag{19.28}$$

and

$$\Pr[B|C] = 1 - \Pr[A|C].$$

From (19.27), one gets

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{[1 - U^*(\lambda)]E(V^2) + V^*(\lambda)E(U^2)}{2[E(V)\{1 - U^*(\lambda)\} + V^*(\lambda)E(U)]} . \tag{19.29}$$

When $V \equiv U$, (19.29) reduces to (19.24).

19.9 $M/G/1 - V_s$ System

In a Poisson input system, where, the server, at the end of a busy period, goes for a single vacation of random duration V, the expected idle period of the server can be obtained, by conditioning, as

$$\begin{aligned} E(I) &= [1 - V^*(\lambda)]E(V) + V^*(\lambda)[E(V) + 1/\lambda] \\ &= E(V) + V^*(\lambda)/\lambda. \end{aligned} \tag{19.30}$$

The expected busy period is given by

$$E(T) = \lambda E(I) \, \frac{b}{1-a} = \frac{a}{1-a} \, E(I)$$

so that

$$\rho = \frac{E(T)}{E(T) + E(I)} = a$$

and

$$E(T) = \frac{\rho}{1-\rho} \, E(I).$$

The expected waiting time is given by

$$\begin{aligned} W_Q \;=\; & L_Q \, E(B) + E(\text{residual service time}) \Pr[\text{server is busy}] \\ & + E(\text{residual vacation time}) \\ & \quad \times \Pr[\text{server on vacation} \mid \text{server is idle}] \\ & \quad \times \Pr[\text{server is idle}]. \end{aligned} \tag{19.31}$$

Now

$$\Pr[\text{server on vacation} \mid \text{server is idle}] = \frac{E(V)}{E(I)} .$$

Thus,

$$(1-a)W_Q = \frac{b^{(2)}}{2b}\,\rho + \frac{E(V^2)}{2E(V)}\,\frac{E(V)}{E(I)}\,(1-\rho)$$

so that

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(I)} . \tag{19.32}$$

19.10 $M/G/1$ System With Vacation and Under N-Policy (With Threshold N)

The system has been considered by Kella (1989) and the more general case with batch arrivals by Lee *et al.* (1994) and recently by Chae and Lee (1995).

We examine the system $M/G/1 - V_m$ under N-policy. In the approach adopted in the two latter papers they consider *grand vacation* and *grand vacation process.* A grand vacation G starts at the instant the system becomes empty and the server leaves for the first vacation V_1 and then for a second vacation and so on until the server finds one or more customers after returning from a vacation. A *grand vacation* comprises of m vacations (the duration of these being i.i.d. random variable V) such that there is no arrival in the first $m-1$ vacations and there are one or more arrivals only in the m-th vacation, $m = 1, 2, \ldots$. Clearly, the number M of vacations comprised in a grand vacation is a geometric random variable such that

$$\begin{aligned}
\Pr[M=m] &= \alpha_0^{m-1}(1-\alpha_0), \qquad m = 1, 2, \ldots\ , \\
\alpha_n &= \int_0^\infty e^{-\lambda t}\,\frac{(\lambda t)^n}{n!}\,dF_V(t), \qquad n = 0, 1, 2, \ldots\ ,
\end{aligned} \tag{19.33}$$

$F_V(\cdot)$ being the d.f. of V and $\alpha_0 = V^*(\lambda)$, LST of $V(\cdot)$; α_n gives the probability that the number of arrivals during a vacation V (with d.f. $F_V(\cdot)$) is n, $n = 0, 1, 2, \ldots$. It follows that the expected length of a grand vacation is given by

$$E(G) = \frac{E(V)}{1-\alpha_0} = \frac{E(V)}{1-V^*(\lambda)} . \tag{19.34}$$

The probability that j customers arrive during a grand vacation is $\alpha_j/(1-\alpha_0)$. Let β_k be the probability that the grand vacation process passes through state k, with $k = 1, 2, \ldots, N-1$ and since it passes through state 0, $\beta_0 = 1$. Conditioning on the arrival size during the first grand vacation, one gets

$$\beta_k = \sum_{j=1}^{k} \beta_{k-j}[\alpha_k/(1-\alpha_0)], \qquad k = 1, 2, \ldots, N-1. \tag{19.35}$$

Thus, β_k's can be computed.

Lee *et al.* show that the expected number of grand vacations during an idle period is given by $\sum_{j=0}^{N-1} \beta_j$. Using Wald identity, the expected length of idle period I is seen to be

$$E(I) = \left[\sum_{j=0}^{N-1} \beta_j\right] \frac{E(V)}{1 - V^*(\lambda)} \,. \tag{19.36}$$

Now the expected number of customers at the initiation of a busy period is given by $\lambda E(I)$, so that the expected busy period is given by

$$E(T) = \frac{a}{1-a}\, E(I) \tag{19.37}$$

from which it follows that

$$\rho = \frac{E(T)}{E(T) + E(I)} = a,$$

independent of vacation policy and threshold N and that

$$E(T) = \frac{\rho}{1-\rho}\, E(I) \,.$$

The probability that there are j customers in the system at a vacation initiation point equals $\beta_j/[\sum_{i=0}^{N-1} \beta_i]$. The ratio

$$\sum_{j=0}^{N-1} j\beta_j \bigg/ \sum_{j=0}^{N-1} \beta_j \tag{19.38}$$

can be interpreted as the mean state of the grand vacation process during an idle period. We have seen in Section 19.4 (N-policy system without vacation) that the expected residual buildup (dormant) period equals

[the mean system size during an idle period]/[the arrival rate].

Here residual idle period comprises of residual vacation period and residual buildup period. Thus, we get that the mean residual idle period $E(I_R)$ in case of N-policy multiple vacation queue

$$E(I_R) = E(V_R) + \frac{1}{\lambda} \frac{\sum_{j=0}^{N-1} j\beta_j}{\sum_{j=0}^{N-1} \beta_j} \,. \tag{19.39}$$

Now from

$$\begin{aligned} W_Q &= L_Q\, E(B) + E(B_R)\,\Pr[\text{server is busy}] \\ &\quad + E(I_R)\,\Pr[\text{server is idle}], \end{aligned} \tag{19.40}$$

we get

$$(1-a)W_Q = \frac{b^{(2)}}{2b}\,\rho + \left[\frac{E(V^2)}{2E(V)} + \frac{1}{\lambda}\,\frac{\sum_{j=0}^{N-1} j\beta_j}{\sum_{j=0}^{N-1}\beta_j}\right](1-\rho)$$

or

$$W_Q = \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(V)} + \frac{1}{\lambda}\,\frac{\sum_{j=0}^{N-1} j\beta_j}{\sum_{j=0}^{N-1}\beta_j}\,. \tag{19.41}$$

The first term corresponds to expected waiting time in a standard $M/G/1$ queue, the second term involves vacation random variable, so that, the first two terms give W_Q for $M/G/1-V_m$ system; the third term is due to the N-policy: when $N=1$, this term vanishes.

Now we consider the $M/G/1-V_s$ system under N-policy with single vacation. Let γ_n be the probability that in a buildup (dormant) period after the completion of vacation period, the system size is n, $n=0,1,\ldots,N-1$. Then $\gamma_0 = \alpha_0 =$ probability that no one arrives in the vacation period; and if the number of arrivals in the vacation period is k ($< N$), then the buildup (dormant) period begins and the system at the stage is like one without vacation and under $(N-k)$-policy. Thus,

$$\gamma_n = \sum_{k=0}^{n} \alpha_k g_{n-k}, \tag{19.42}$$

where g_n is the probability that the system without vacation and under $(N-k)$-policy passes through the state n during the buildup period. It is shown that $\sum_{n=0}^{N-1}\gamma_n$ is the mean number of arrivals during a dormant period, and that the mean duration of dormant period is

$$= \frac{\text{mean number of arrivals during dormant period}}{\text{mean arrival rate}} = \sum_{n=0}^{N-1}\gamma_n/\lambda$$

so that the expected idle period is given by

$$E(I) = E(V) + \sum_{n=0}^{N-1}\gamma_n/\lambda. \tag{19.43}$$

The expected busy period equals

$$E(T) = \lambda E(I)\left(\frac{b}{1-a}\right) = \frac{a}{1-a}\,E(I)$$

so that the utilization factor is given by

$$\rho = \frac{E(T)}{E(T) + E(I)} = a,$$

independent of the vacation and the N-policy. Further,

$$\begin{aligned} \Pr[V|I] &\equiv \Pr[\text{server is on vacation} \mid \text{server is idle}] \\ &= \frac{E(V)}{E(I)}, \end{aligned} \tag{19.44}$$

and

$$\begin{aligned} \Pr[D|I] &\equiv \Pr[\text{system is in dormant period} \mid \text{server is idle}] \\ &= \frac{\sum_{n=0}^{N-1} \gamma_n/\lambda}{E(I)}. \end{aligned} \tag{19.45}$$

The mean system size during the dormant period equals

$$\sum_{n=0}^{N-1} n\gamma_n \Big/ \sum_{n=0}^{N-1} \gamma_n \tag{19.46}$$

and the meantime for mean system size to accumulate is

$$(1/\lambda) \sum_{n=0}^{N-1} n\gamma_n / \sum_{n=0}^{N-1} \gamma_n.$$

Thus, the mean residual idle period (after the arrival of the test customer) is given by

$$E(I_R) = E(V_R) \Pr[V|I] + (1/\lambda) \left[\sum_{n=0}^{N-1} n\gamma_n \Big/ \sum_{n=0}^{N-1} \gamma_n \right] \Pr[D|I].$$

Now

$$W_q = L_Q\, E(B) + \rho E(B_R) + (1-\rho) E(I_R)$$

which gives

$$\begin{aligned} W_Q &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{E(V^2)}{2E(V)} \frac{E(V)}{E(V) + \Sigma\gamma_n/\lambda} + \frac{1}{\lambda} \frac{\Sigma n\gamma_n}{\Sigma\gamma_n} \frac{\Sigma\gamma_n/\lambda}{E(V) + \Sigma\gamma_n/\lambda} \\ &= \frac{\lambda b^{(2)}}{2(1-\rho)} + \frac{\lambda E(V^2)}{2[\lambda E(V) + \Sigma\gamma_n]} + \frac{1}{\lambda} \frac{\Sigma n\gamma_n}{\lambda E(V) + \Sigma\gamma_n}. \end{aligned} \tag{19.47}$$

For details, see Chae and Lee (1995).

19.11 $M^X/G/1$ System With Batch Arrival

In case of compound Poisson arrivals, Poisson streams of customers arrive in batches and the test customer has to wait for completion of service of the unit in service, if any, and of those in the queue that he finds on arrival (in a batch), but also has to wait for completion of the service of those arriving in the same batch but are served ahead of him.

Thus, in place of (19.2) for single arrivals, in case of compound Poisson arrivals, the mean waiting time in the queue (of the test customer) is found to be

$$W_Q = [L_Q + E(X_R)]E(B) + E(B_R)\,\Pr[\text{server is busy}], \tag{19.48}$$

where X_R is the number of customers in the group in which the test customer in the group arrives and are to be served ahead of the test customer. $E(X_{;t})$, which is the mean residual group size, is given by

$$E(X_R) = \frac{1}{2}\left[\frac{E(X^2)}{E(X)} - 1\right]. \tag{19.49}$$

The mean delay due to this group is $bE(X_R)$ [see also Medhi (1991)].

With $\rho = \lambda E(X)E(B) = aE(X)$, $L_Q = \lambda E(X)W_Q$, one gets

$$\begin{aligned} W_Q &= \frac{1}{1-\rho}[E(X_R)E(B) + E(B_R)\rho] \\ &= \frac{\lambda E(X)b^{(2)}}{2(1-\rho)} + \frac{b}{2(1-\rho)}\left[\frac{E(X^2 - X)}{E(X)}\right]. \end{aligned} \tag{19.50}$$

The second term is due to the expected delay for the service of the residual group size; it does not occur in case of a single arrival system, since $E(X_R) = 0$ for single arrival case.

19.12 $M^X/G/1$ Under N-Policy

The system was considered by Lee and Srinivasan (1989) and also by Lee *et al.* (1994). The arrivals occur in a batch of size X at Poisson input instants with rate λ, and $h_i = \Pr[X = i]$; the arrival rate is $\lambda E(X)$. The server, as soon as the system becomes empty, is idle and waits till there are at least N arrivals. This duration is the idle period I. Let π_j be the probability that the number in the system is j in an idle period. Then $\pi_0 = 1$, since the idle period always begins at state 0; by conditioning on the first group size i, one gets $\pi_j = \sum_{i=1}^{j} \Pr[X = i]\pi_{j-i} = \sum_{i=1}^{j} h_i \pi_{j-i}$, $j = 1, 2, \ldots, N-1$. Thus since π_0 is known, π_j's can be obtained recursively. It can be seen that $\sum_{j=0}^{N-1} \pi_j$ gives the mean number of arrival groups during an idle period, and with λ as the arrival rate for groups, the mean duration for the arrival of these mean number of batches is $\sum_{j=0}^{N-1} \pi_j/\lambda$, which gives the mean idle period as

$$E(I) = (1/\lambda) \sum_{j=0}^{N-1} \pi_j. \tag{19.51}$$

The mean number of arrivals during an idle period is $E(I)[\lambda E(X)]$, so that the mean busy period equals

$$E(T) = \lambda E(X) E(I) \frac{1}{\mu - \lambda E(X)} = \frac{\rho}{1-\rho} E(I).$$

Now, the mean residual idle period, given that the server is idle, is

$$\begin{aligned} &= \frac{\text{mean system size during an idle period}}{\text{arrival rate}} \\ &= \frac{\sum_{n=0}^{N-1} n\pi_n / \sum_{n=0}^{N-1} \pi_n}{\lambda E(X)} . \end{aligned} \tag{19.52}$$

Thus the mean waiting time can be obtained from (19.50) by adding a term corresponding to the residual idle period. We have

$$\begin{aligned} W_Q &= \frac{1}{1-\rho}[\{\text{expected residual service time}\} \Pr[\text{server is busy}] \\ &\quad +\{\text{expected service time of the residual group in which} \\ &\qquad \text{the test customer arrives and who are served prior to him}\} \\ &\quad +\{\text{expected residual idle period}\} \Pr[\text{server is idle}]] \\ &= \frac{1}{1-\rho}\left[\rho \frac{b^{(2)}}{2b} + \frac{1}{2} \frac{E(X^2 - X)}{E(X)} b + (1-\rho) \frac{\Sigma n\pi_n/\Sigma \pi_n}{\lambda E(X)}\right]. \end{aligned} \tag{19.53}$$

When $\Pr[X = 1] = 1$ (for single arrivals), the second term becomes 0 and $\pi_j = 1$, $j = 0, 1, 2, \ldots, N-1$, so that the last term becomes

$$(1-\rho)\left(\frac{N-1}{2\lambda}\right)$$

and (19.53) reduces to (19.16).

19.13 $M^X/G/1 - V_m$ and $M^X/G/1 - V_s$

Here as in the case of $M/G/1 - V_m$ (Section 19.7), we have

$$E(I) = E(V)/[1 - V^*(\lambda)]$$

and

$$E(T) = \frac{\rho}{1-\rho} E(I).$$

In computing queueing time, we have also to include the delay for residual vacation time V_R, should the test customer arrive during a vacation time with the server lying idle. Making this adjustment one gets, from (19.50),

$$W_Q = \frac{1}{1-\rho}[E(B)E(X_R) + E(B_R)\rho + E(V_R)\Pr[V|I](1-\rho)]. \quad (19.54)$$

In case of $M/G/1 - V_s$ system,

$$\begin{aligned} E(I) &= E(V) + V^*(\lambda)/\lambda \\ E(T) &= \frac{\rho}{1-\rho} E(I). \end{aligned}$$

Here,

$$\begin{aligned} \Pr[V|I] &\equiv \Pr[\text{server on vacation} \mid \text{server is idle}] \\ &= \frac{E(V)}{E(I)}(1-\rho) = \frac{E(V)(1-\rho)}{E(V) + V^*(\lambda)/\lambda}. \end{aligned} \quad (19.55)$$

Taking this into consideration, one can get (as in Section 19.9)

$$W_Q = \frac{1}{2(1-\rho)} \lambda E(X) b^{(2)} + \frac{b}{1-\rho} E(X_R) + \frac{E(V^2)}{2E(I)}. \quad (19.56)$$

For single arrival, $E(X_R) = 0$, $E(X) = 1$; one then gets (19.32).

19.14 $M^X/G/1$ Vacation Queues Under N-Policy

The analysis is identical with that of the corresponding system with single arrivals. Writing $\lambda E(X)$ for λ in (19.41), $\rho = aE(X)$ and adding the term due to the delay for the residual group size (those arriving in the same group but served prior to the test customer) given by the last term in (19.50), one gets for N-policy $M^X/G/1 - V_m$ queue with vacation,

$$\begin{aligned} W_Q &= \frac{\lambda E(X) b^{(2)}}{2(1-\rho)} + \frac{b}{2E(X)(1-\rho)} E(X^2 - X) \\ &+ \frac{E(V^2)}{2E(V)} + \frac{1}{\lambda E(X)} \frac{\sum_{n=0}^{N-1} n\beta_n}{\sum_{n=0}^{N-1} \beta_n} . \end{aligned} \tag{19.57}$$

Consider now N-policy $M^X/G/1 - V_s$ queue with vacation.

We note that $\sum_{n=0}^{N-1} \gamma_n$ (defined in Section 19.10) gives the mean number of arrival groups so that mean staying time will still be $\Sigma\gamma_n/\lambda$ so that $E(I)$, $\Pr[V|I]$ and $\Pr[D|I]$ will remain unchanged [as in Eqs. (19.43), (19.44), (19.45)]. Writing $\lambda E(X)$ for λ in the first factor of the last term of (19.47) and adding the term due to the delay for the residual group size as given by the last term of (19.50), one gets

$$\begin{aligned} W_Q &= \frac{\lambda E(X) b^{(2)}}{2(1-\rho)} + \frac{b}{2E(X)(1-\rho)} E(X^2 - X) \\ &+ \frac{E(V^2)}{2E(V)} \frac{\lambda E(V)}{\lambda E(V) + \Sigma\gamma_n} \\ &+ \frac{1}{\lambda E(X)} \frac{\Sigma n\gamma_n}{\Sigma\gamma_n} \frac{\Sigma\gamma_n}{\lambda E(V) + \Sigma\gamma_n} \\ &= \frac{\lambda E(X) b^{(2)}}{2(1-\rho)} + \frac{b}{2E(X)(1-\rho)} E(X^2 - X) \\ &+ \frac{\lambda E(V^2)}{2[\lambda E(V) + \Sigma\gamma_n]} + \frac{1}{\lambda E(X)} \frac{\Sigma n\gamma_n}{\lambda E(V) + \Sigma\gamma_n} . \end{aligned} \tag{19.58}$$

The first two terms give the expected waiting time for $M^X/G/1$ system without threshold and vacation [as given in (19.50)]. The result for N-policy $M/G/1 - V_s$ system can be obtained by putting $E(X) = E(X^2) = 1$.

19.15 Concluding Remarks

Here we have used heuristic approach to find some important performance measures of an important class of queueing systems without going into decomposition property. This important property, first established by Fuhrmann and Cooper (1985) for $M/G/1$ system with vacation, has been extended for more general case of N-policy $M^X/G/1$ system with vacations, multiple as well as single, by Lee *et al.* (1994, 1995) to which reference may be made for detailed analysis. In a very recent paper, Lee *et al.* (1996) consider continuous diffusion approximation for more general $GI/G/1$ queue with batch arrival and

N-policy. In another, Lee *et al.* (1996) also consider fixed batch service queues with vacations. Further results on general batch service [Medhi (1991)] queues would be of interest.

References

1. Böhm, W. and Mohanty, S. G. (1993). The transient solution of $M/M/1$ queues under (M, N) policy. A combinatorial approach, *Journal of Statistical Planning and Inference*, **34**, 23–33.

2. Böhm, W. and Mohanty, S. G. (1994a). On discrete time Markovian N-policy queues involving batches, *Sankhyā, Series A*, **56**, 144–163.

3. Böhm, W. and Mohanty, S. G. (1994b). Transient analysis of $M/M/1$ queues in discrete time with general server vacation, *Journal of Applied Probability*, **31**, 115–129.

4. Chae, K. C. and Lee, H. W. (1995). $M^X/G/1$ vacation models with N-policy: Heuristic interpretation of the mean waiting time, *Journal of Operational Research Society*, **46**, 258–264.

5. Doshi, B. T. (1986). Queueing systems with vacations—A survey, *QUESTA*, **1**, 29–66.

6. Fuhrmann, S. W. and Cooper, R. B. (1985). Stochastic decompositions in $M/G/1$ queue with generalized vacations, *Operations Research*, **33**, 1117–1129.

7. Heyman, D. P. (1968). Optimal operating policies for $M/G/1$ queueing system, *Management Science*, **16**, 362–382.

8. Kella, O. (1989). The threshold policy in the $M/G/1$ queue with server vacations, *Naval Research Logistics Quarterly*, **36**, 111–123.

9. Lee, H. S. and Srinivasan, M. M. (1989). Control policies for $M^X/G/1$ queueing systems, *Management Science*, **35**, 708–721.

10. Lee, H. W. (1988). $M/G/1$ queue with exceptional first vacation, *Computers & Operations Research*, **15**, 441–445.

11. Lee, H. W., Lee, S. S. and Chae, K. C. (1994). Operating characteristics of $M^X/G/1$ queue with N-policy, *QUESTA*, **15**, 387–399.

12. Lee, H. W., Lee, S. S., Park, J. O. and Chae, K. C. (1994). Analysis of $M^X/G/1$ queue with N-policy and multiple vacations, *Journal of Applied Probability*, **31**, 476–496.

13. Lee, S. S., Lee, H. W., Yoon, S. H. and Chae, K. C. (1995). Batch arrival queue with N-policy and single vacation, *Computers & Operations Research*, **22**, 173–189.

14. Lee, H. W., Yoon, S. H. and Lee, S. S. (1996). A continuous approximation for batch arrival queues with threshold, *Computers & Operations Research*, **23**, 299–308.

15. Lee, H. W., Lee, S. S. and Chae, K. C. (1996). A fixed-size batch service queue with vacations, *Journal of Applied Mathematics & Stochastic Analysis* (to appear).

16. Medhi, J. (1991). *Stochastic Models in Queueing Theory*, San Diego: Academic Press.

17. Medhi, J. (1994). Extensions and generalizations of the classical single-server queueing system with Poisson input: A survey, *Journal of Assam Science Society*, **36**, 211–226.

18. Medhi, J. and Templeton, J. G. C. (1992). A Poisson input queue under N-policy and with a general startup time, *Computers & Operations Research*, **19**, 35–41.

19. Mohanty, S. G. and Panny, W. (1989). A discrete time analogue of the $M/M/1$ queue and the transient solution: An analytic approach, *Colloquia Mathematica Societatis János Bolyai*, **57**, 417–424.

20. Mohanty, S. G. and Panny W. (1990). A discrete time analogue of the $M/M/1$ queue and the transient solution: A geometric approach, *Sankhyā, Series A*, **52**, 364–370.

21. Takagi, H. (1991). *Queueing Analysis: A Foundation of Performance Evalation, Volume I—Vacation and Priority Systems*, Amsterdam: North-Holland.

22. Takagi, H. (1992). Analysis of an $M/G/1$ queue with multiple server vacations and its application to a polling model, *Journal of Operations Research Society of Japan*, **35**, 300–315.

23. Takagi, H. (1993). *Queueing Analysis: A Foundation of Performance Evalation, Volume II—Finite Systems*, Amsterdam: North-Holland.

24. Takagi, H. (1993). $M/G/1/K$ queues with N-policy and setup time, *QUESTA*, **14**, 79–98.

25. Takine, T., Takagi, H. and Hasegawa, T. (1993). Analysis of an $M/G/1/K/n$ queue, *Journal of Applied Probability*, **30**, 446–454.

26. Yadin, M. and Naor, P. (1963). Queueing systems with a removable service station, *Operations Research Quarterly*, **14**, 393–405.

20

Recent Advances in the Analysis of Polling Systems

Diwakar Gupta and Yavuz Günalay

McMaster University, Hamilton, Ontario, Canada

Abstract: This article summarizes some recent advances in the analysis of polling models in which the server uses system-state information to affect its behavior. Literature dealing with a special class of such models in which the server lies dormant upon finding the system empty, and reactivates only when the system is populated by a critical number of customers, is closely examined. Rather than provide a broad review of literature, the focus of this article is on providing insights, on building bridges with earlier literature, and on identifying common underlying principles.

Keywords and phrases: Polling models, state dependent server scheduling, threshold start-up, threshold setups, dormant/patient server, descendant sets, queueing theory

20.1 Introduction

Polling systems is the name given to queueing systems with multiple customer classes, in which each customer class forms its own separate queue (station). There is typically only one server that travels from one queueing station to the next in a cyclic fashion. A strictly *cyclic* protocol is one in which each station is visited exactly once in a complete tour (by the server) of all stations. Many variations of this basic routing policy are possible, for example, routing according to a polling table—a generalization of cyclic protocol in which some stations might be visited more than once in a cycle; and state-dependent routing— e.g., server travelling directly (jumping) to the longest queue, or to the nearest non-empty queue next.

Inter-station travel times are usually non-zero and are called *switchover* times. When the server arrives at a station, it is often required to complete a *setup* before it can start serving customers at that station. Once it sets up, the server satisfies a certain number of customers depending on the *service policy* in effect. A majority of service policies that have been studied in literature require only local queue information. Some common examples are *exhaustive*—where server empties the queue before moving on to another, *gated*—where it serves only those customers that it finds waiting upon completing the setup, *K-limited*—where it serves either exactly K customers or empties the queue, whichever occurs first, and *timer-limited*—where the server stays at a queue for a prespecified length of time or until it empties the queue, again, whichever occurs first. However, advances in computer technology permit modern systems to acquire increasing amounts of system information at negligible cost. In other words, information concerning status of all queues in the system can be collected at great speed (often much faster than the speed with which the server moves), and at negligible additional cost. Therefore, more information-hungry policies, i.e., policies based on the status of all queues in the system are becoming increasingly more practical. An example of this kind of service policies is the *globally gated* policy—where server attends to only those customers that were present at the queue when it finished setting up for a home station, the starting point of its tour [Borst (1995) and Boxma, Weststrate and Yechiali (1993)].

In addition to the various combinations of routing and service policies that are possible, the class of polling models is further enriched by novel features that come from new applications of these models. For example, when modelling manufacturing systems, the server (which would be a machine, a robot, or a transportation device such as an automated guided vehicle (AGV)) might turn itself off (lie dormant), upon finding no pending requests for service in the entire system, in order to conserve energy/cost. Furthermore, when faced with a one time startup cost, the server might not reactivate until the number of new arrivals to the system exceeds a certain threshold. Similarly, the server might decide to setup for the next class of customers only if the number of waiting customers of that class is more than a certain minimum. Otherwise, the server passes through that station without setting up and polls the next queue, at which the same protocol is applied once more. The minimum number of customers necessary to warrant a setup might be different at different stations. A feature that is common to all these models is the fact they use information about the state of the system[1] to affect server behavior.

Polling models emerged from the telecommunications field as tools for modeling and evaluating performance of computer and communications networks (for example, Local Area Networks). The term polling systems evolved from models of decentralized communication networks having a single multiple-access channel in which communication is provided by having the token (bus) *poll*

[1]State is described by the number of customers in each of the queues of the system.

(query/response) each terminal (node) and then transmit messages according to some prespecified protocol, e.g., exhaustive. In these situations, the speed with which the server (token) travels from station to station is extremely fast and polling is the only mechanism by which the server can ascertain the number/length of messages waiting to be transmitted. Naturally, polling models were developed with a variety of routing/service policies, but in these early models, the server did not utilize any system-state information to determine its behavior.

With the enlargement of the set of potential applications of polling models [see Benjaafar and Gupta (1996a,b), Bozer and Srinivasan (1991), Federgruen and Katalan (1996), Sarkar and Zangwill (1991) and Srinivasan and Gupta (1996) for some recent examples and Levi and Sidi (1990) for some standard ones], and availability of cheaper and faster information, it is now practical to consider state-dependent server scheduling rules [Ferguson (1986), Günalay and Gupta (1996a,b) and Gupta and Srinivasan (1996b)]. The objective of this article is to review some recent developments in the analysis of such models, focusing on providing insights, on building bridges with earlier literature, and on identifying common underlying principles. Clearly, this is a biased view and colored by the authors' own endeavors in this area of research.

This article is not a basic review of the analysis of polling models, which has been a very fertile area of research over the last several decades. It has spawned hundreds of research articles, and the pace has been getting faster over the last several years. Takagi (1986) first reviewed and summarized literature until 1986 and since then he has twice updated his review to include recent developments: Takagi (1990) until 1990, and Takagi (1994) for advances in 1990-1993. Practically, all of the studies which have been summarized in this article are more recent. Thus, this article assumes a rudimentary familiarity with basic polling models literature, with a focus on providing an in-depth analysis of significant recent trends. Readers who are not at all familiar with polling models literature will benefit greatly by consulting the previous reviews by Takagi (1986, 1990, 1994) before proceeding further.

The organization of the rest of this article is as follows. Notations and model details are provided in the next Section. This Section also contains an important decomposition result [from Fuhrmann and Cooper (1985)] which is used often in the analysis of polling models. The main results are presented in Section 20.3. The approach outlined in this article is quite versatile and extends easily to cover many different variations of the basic model described here. Some examples of such alternate models that are motivated by well-known applications are mentioned in Section 20.4. Section 20.5 is devoted to building bridges with other polling systems literature. The last Section of this article contains a few thoughts on important future directions for research in this field.

20.2 Notations and Preliminaries

A polling system with M stations (customer classes) is considered in which each station behaves like an $M/G/1$ queue. That is, station i has an independent Poisson arrival stream of rate λ_i, service time S_i, and busy period Θ_i. The duration of a setup at station i, if performed, is denoted by T_i, and the switchover time from station i to station $i+1$ is denoted by R_i. All time length random variables, i.e., S_i, T_i and R_i, are assumed to have finite means and variances and time-stationary distributions that are independent of each other and of the arrival process. Load at station i is denoted by $\rho_i = \lambda_i E[S_i]$, and the system load by $\rho = \sum_{j=1}^{M} \rho_j$. $R_T = \sum_{i=1}^{M} R_i$ and $\Lambda = \sum_{i=1}^{M} \lambda_i$ denote, respectively, the total switchover time per cycle and the total arrival rate. Since arrivals are Poisson, $p_i = \lambda_i/\Lambda$ is the probability that an arbitrary arrival to the system is a type i customer. The performance measure of primary interest is the mean station waiting time, denoted by $E[W_i]$ for station i.

The notational conventions used in this article are as follows. For a random variable A, notation $A(t)$, $A^*(s)$, $E[A]$ and $E[A^2]$ are used to denote the cumulative distribution function, the Laplace-Stieltjes transform (LST), the mean, and the second moment, respectively. If A is discrete, then $A(z) \triangleq E[z^A]$ denotes its probability generating function (PGF). Single and double prime notation is used to denote, respectively, first and second derivative with respect to z. The notation $\bar{n}$ (or $\bar{z}$) represents a $1 \times M$ vector of n_i's (or z_i's).

This article concentrates on polling models in which the server always moves along a fixed path and visits stations in a strict cyclic order. Also, it primarily deals with the exhaustive service policy. The analysis of the gated service regime is very similar to what is presented here, as discussed briefly in Section 20.4.3.

Exhaustive policy is a good straw policy to consider when fairness is not an issue; this is typically true in manufacturing environments wherein customers are not people but jobs waiting for machining, or other types of processing. Liu, Nain and Towsley (1992) have shown that the exhaustive policy minimizes the total unfinished work in system. Another reason for choosing strictly cyclic-exhaustive service model is that for this model, Altman, Konstantopoulos and Liu (1992) have shown $\rho < 1$ to be both necessary and sufficient condition for stability of the queueing system, i.e., for the existence of the stationary joint distribution of station queue lengths. This is true irrespective of the duration of setups/switchovers, of whether setups are done in a state-dependent or independent fashion, and whether or not the server stops upon finding the system empty. When either the server routing is done in a dynamic fashion (different from cyclic or polling table) and/or service policy is other than exhaustive/gated, the stability of the system requires additional conditions [see Altman, Konstantopoulos and Liu (1992) for details] that depend on the magnitude of setups/switchovers.

In many applications, switchover times are either negligible or unavoidable since the server needs to switch to ascertain the status of the next queue. Whenever a setup always follows switchover from the previous queue, setup and switchover times can be treated as one by defining a new switchover time which is the sum of R_i and T_{i+1}. What we get is a model with state-independent switchover/setup overheads, and this fact greatly simplifies the mathematical analysis. However, the server may postpone a setup if it does not find at least a critical number of waiting customers upon polling the station. Such models are particularly suitable for modelling manufacturing systems and have been studied in two recent papers [Günalay and Gupta (1996a) and Gupta and Srinivasan (1996b)]. However, their analysis is difficult, and mathematically exact analysis of systems with infinite buffers and $M > 2$ is not yet available.

Belonging to the set of cyclic-exhaustive polling models is another distinction of the server behavior that deserves special mention: *continuously roving* server—where the server never stops, and *patient* server—where the server stops upon finding the system empty [Srinivasan and Gupta (1996)]. Within the patient class of models, several variations are possible. The server could stop at the same station where it first observes the system empty, or move to a *preferred* station and stop only if the system is still empty. If the system is repopulated before the server has had a chance to reach its preferred station, the server ignores the fact that the system became momentarily empty. Typically, if we are modelling a robot that is attending to several tasks, we might wish it to stop immediately upon discovering that there are no more pending jobs. On the other hand, if we are modelling AGVs, it would make sense to have them stop only at the designated park area (station).

Additional variations could arise from differences in monitoring frequency while the system is idle, as well as in the criterion used to restart the server. For example, if in the dormant state the server attends to other (possibly lower priority) tasks, it might only be able to monitor the state periodically (or react to changes in system state periodically). Alternatively, it is also possible to have continuous monitoring. In each case, the server could either start as soon as there is at least one customer in the system, or until the system is populated by at least a critical number of customers. Also, it could either travel in the fixed cyclic path or jump directly to the first non-empty queue upon restarting [Eisenberg (1994)].

In this survey article, we shall take a close look at cyclic-exhaustive polling systems with state-independent setups and continuous monitoring. Thus, when the system restarts after being dormant for some time, it will always have the same number of customers, $N \geq 0$. This model is also known as the *threshold startups model* and N is called the *startup threshold* [Günalay and Gupta (1996a)]. Several special instances of this model are well known in queueing literature. When $N = 0$, we obtain a model with a continuously roving server [this model has been reviewed extensively by Takagi (1986, 1990, 1994)]; when $N = 1$, the corresponding model is called the patient/stopping server model [studied first by Eisenberg (1971) for $M = 2$, and later by Eisenberg (1994) and Srinivasan and Gupta (1996) for any general M]; and finally when there is a single customer class, i.e., $M = 1$, and $N > 0$, we obtain the $M/G/1$ queue with a N-policy [studied by Heyman and Sobel; see Heyman and Sobel (1982, pp. 445) for waiting time characteristics of this model, and Heyman and Sobel (1984, Section 7.2) for a discussion of optimality of this policy within the class of server control policies with a removable server]. Models with periodic monitoring, and those with state-dependent setups, are discussed in Section 20.4. We also show how these different models relate to the threshold startup model.

Lying dormant after the system becomes empty is an example of server control policies that allow the server to *idle*. More general forms of idling behavior are certainly possible, e.g., idling at an empty station even when system is not empty. It is intuitively straightforward to reason that idling at a non-empty station is never advisable [see Liu, Nain and Towsley (1992) for discussion and proof] so long as the objective is to minimize total work in the system. Analysis of polling models, in which server idling does not necessarily commence when the system is empty, is mathematically difficult and only some heuristic solutions have been reported to date [see, for example, Duenyas and Van Oyen (1996)]. We shall revisit idling issues in Section 20.5 when we discuss forced server idling and its sometimes paradoxical impact on system performance: forced idling can sometimes reduce mean customer waiting times.

Before closing this Section, it is useful to discuss the Fuhrmann-Cooper Stochastic Decomposition Theorem [Fuhrmann and Cooper (1985)] for $M/G/1$ queues and its relevance to the analysis of polling models. Fuhrmann and Cooper have shown that the decomposition property, i.e., the fact that the stationary number of customers present at a random observation epoch in $M/G/1$ queues with server vacations is a sum of two or more random variables, one of which is the number of customers present in the corresponding standard $M/G/1$ queue (with no vacations), holds for a large class of $M/G/1$ models including several polling models [see Cooper (1970) for the first discovery of this principle in the context of polling models]. Whenever applicable, this decomposition principle greatly simplifies the analysis of polling models. The amount of time that the server is away from a station (serving other stations, switching, setting up, or idling) is treated as a vacation from that station. Station performance measures can be obtained by imbedding Markov chains at the start and end of service at each station (which is also the end and start of a server vacation, respectively, from that station), rather than imbedding it at every service completion epoch.

Consider the Markov chains imbedded at the start- and the end-of-server-vacation from each station. These instances correspond, respectively, to the time epochs when the server completes service at a station to either move to the next station or to lie dormant at that same station, and the moment when it returns to a station, or reactivates at that station, to start a new round of service. Let $k_i(\bar{n})$ and $g_i(\bar{n})$ denote the probabilities that an arbitrary start/end of server vacation instant happens to be at station i and that there are $\bar{n}$ customers (with elements $n_i \geq 0 \; \forall \; i$) in the system. Similarly, $k_i(\bar{z})$ and $g_i(\bar{z})$ will denote the partial PGFs of queue lengths at start/end of server vacation epochs, i.e., $k_i(\bar{z}) \triangleq \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} k_i(\bar{n}) \prod_{j=1}^{M} z_j^{n_j}$ and $g_i(\bar{z}) \triangleq \sum_{n_1=0}^{\infty} \cdots \sum_{n_M=0}^{\infty} g_i(\bar{n}) \prod_{j=1}^{M} z_j^{n_j}$.

In what follows, we shall develop relationships to determine stationary queue length distribution at station 1. Setting station index to 1 causes no loss of generality and other stations can be treated in a like manner, or simply by renumbering stations such that the station we wish to analyze is always indexed 1. Let $\Pi_1(\bar{z})$ denote the PGF of stationary queue lengths at a departure instant of an arbitrary type-1 customer. Then, from Fuhrmann and Cooper's Proposition 2 [Fuhrmann and Cooper (1985)], we obtain [additional details can be found, for example, in Günalay and Gupta (1996a, Section 4.1)]

$$\Pi_1(z, 1, \ldots, 1) = \frac{k_1(\bar{1}) - k_1(z, 1, \ldots, 1)}{k_1'(\bar{1})} \; \frac{(1-\rho_1)S_1^*(\lambda_1 - \lambda_1 z)}{S_1^*(\lambda_1 - \lambda_1 z) - z}. \tag{20.1}$$

Differentiating (20.1) with respect to z and then setting $z = 1$, we get the expected queue length of station 1 at a customer departure instant, which is also the average queue length at station 1 at an arbitrary observation epoch in a large number of models (see Section 20.4.1 for a model for which this does not apply). Distributions of queues left behind by a departing customer and those observed at arbitrary observation epochs coincide due to two important properties [see, for example, Cooper (1990, pp. 186-188)]: PASTA (Poisson Arrivals See Time Averages) property and the single-arrival single-departure property (i.e., there can be at most one arrival or departure at any observation instant). In such cases, Little's Law can be used to obtain the mean waiting time of type-1 customers as

$$E[W_1] \;=\; \frac{k_1''(\bar{1})}{2\lambda_1 k_1'(\bar{1})} + \frac{\lambda_1 E[S_1^2]}{2(1-\rho_1)}\,. \tag{20.2}$$

The reader should note that the decomposition in (20.1) also applies to the waiting time distribution for certain models with $N \leq 1$ [see, for example, Srinivasan and Gupta (1996, Eq. 1, pp. 441)]. However, it is no longer true when $N > 1$. Details of this important observation are worked out in Exercise 12, Part h, pp. 222-223 of Cooper (1990).

Following the simplification made possible by Fuhrmann-Cooper decomposition, the major effort in recent articles involving steady state analysis of polling models has been on determining the PGFs, $k_i(\bar{z})$'s, and their moments.

20.3 Main Results

In this section, some important results and insights are presented for the M station polling system with a patient server that restarts only after exactly N new arrivals occur following the start of idle state of the server. It is assumed, in this central analysis, that the server always restarts at the station at which it idles and cycles around to the first non-empty station in case the idling station is empty at the moment it is reactivated. Various other models that could be analyzed using essentially this very framework are discussed in the next section.

The most promising method to emerge in recent years to compute $k_i(\bar{z})$'s and their moments is the *descendant sets* method. It was first applied to the standard polling model by Konheim, Levi and Srinivasan (1994). Since then, it has been successfully used to model many variations; for example, Duenyas, Gupta and Lennon (1995) used it to model tandem queues, Srinivasan and Gupta (1996) modelled patient server systems, and Günalay and Gupta (1996a,b) developed models with state-dependent setups and threshold startups.

The main idea of the descendant sets method is straightforward. It begins by recognizing two types of customers (within each customer class). *Original* customers are those that arrive either during a setup/switchover time, or at a time when the server is idle. *Non-original* customers, on the other hand, are those that arrive during the service time of another customer, and therefore belong to the *descendant* set of an original customer. If we now examine an arbitrary end-of-server-vacation epoch at station 1, which is called the *reference point*, all type 1 customers who are waiting in the queue at that time must either be original customers or descendants of past original customers. Since counting of the size of the descendant set is a simple branching process, the task of finding the PGF of the queue length reduces to that of computing *contributions* (subsets of the descendant sets of original customers) in a recursive fashion.

The concept of a *cycle* plays an important role in the descendant sets method. At station i, a server cycle, C_i, is defined as the amount of time that elapses between any two end-of-server-vacation instances that are preceded by a switch from station $i-1$, i.e., when neither one corresponds to a server startup instance at station i following an idle period there. A consequence of this definition is that the cycle index does not advance when the server reactivates following an idle period. Since the routing protocol is strictly cyclic, it is easy to see that $E[C_i] = E[C]$ is the same for all stations. This is, however, not true for higher moments of the cycle length.

Let $R_{i,c}$ denote the number of customers in queue 1 at reference point who have descended from original customers that arrived during a switchover from station i to station $i+1$, c cycles prior to the reference point. Then, $R_{i,c}(z)$, its PGF, can be written as

$$R_{i,c}(z) \triangleq R_i^* \left(\sum_{j=i+1}^{M} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right), \quad i = 1, \ldots, M, \ c \geq 0. \tag{20.3}$$

Eq. (20.3) follows from the fact that type $j \leq i$ arrivals are served in the next cycle, but type $j > i$ are served in the same cycle, i.e., the cycle indexed c. The term $L_{j,c}(z)$ denotes the PGF of the contribution to station 1 queue, at reference point, from a customer that is served at station j, c cycles ago. We define $L_{i,c}(z)$ as follows:

$$L_{i,c}(z) \triangleq \Theta_i^* \left(\sum_{j=i+1}^{M} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i-1} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right), \quad i = 1, \ldots, M, \ c \geq 0. \tag{20.4}$$

Alternatively, $L_{i,c}(z)$ can also be written as

$$L_{i,c}(z) = S_i^* \left(\sum_{j=i}^{M} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i-1} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right), \quad i = 1, \ldots, M, \ c \geq 0. \tag{20.5}$$

The difference between these two equivalent representations is that in the former, we treat the busy period as the amount of time that each customer engages the server. Hence, we do not need to account for the off spring that arrive during the service period of a customer. These offspring receive service in the same cycle (though, in reality, this may not happen immediately after the service of the parent). In the latter representation, every off spring of a customer is explicitly modeled. This method of counting contributions is particularly suitable when considering customer routing, which is discussed in Section 20.4.1.

The boundary conditions for starting recursive counting of contributions are: $L_{1,-1}(z) = z$ and $L_{i,-1}(z) = 1$, for all $i > 1$. This follows from the facts that the reference point is the start of service at station 1 in the cycle indexed -1, that all customers present at station 1 at this moment have a contribution of exactly 1, and that the customers present at other stations at the start of the -1-th cycle do not get a chance to contribute to the station 1 queue length.

Counting contributions going backward in time, we know that the total contributions would amount to $\prod_{c=0}^{\infty} \prod_{i=1}^{M} R_{i,c}(z)$ if the server never idles [see, for example, Konheim, Levi and Srinivasan (1994) for formal arguments]. In our model, the server stops, every time the system is empty. Since only contributions from that moment onwards are non-zero, this means that we only need to worry about counting contributions from the most recent moment (prior to reference point) at which the server became idle. Now, first conditioning on most recent server idling incident being at station i, c cycles before reference point, and then summing carefully over all such possibilities, we can calculate the total contribution, if the server is allowed to idle. These arguments lead to the following identity:

$$\begin{aligned} k_1(z,1,\ldots,1) \;=\; \Phi \Bigg[& \prod_{c=0}^{\infty}\prod_{i=1}^{M} R_{i,c}(z) - \sum_{c=0}^{\infty}\sum_{i=1}^{M} \vartheta_i \prod_{j=i}^{M} R_{j,c}(z) \prod_{l=0}^{c-1}\prod_{k=1}^{M} R_{k,l}(z) \\ & + \sum_{c=0}^{\infty}\sum_{i=1}^{M} \vartheta_i \sum_{\bar{n}\in U(N)} p(\bar{n}) u_{i,c}(\bar{n},z) \prod_{j=i}^{M} R_{j,c}(z) \prod_{l=0}^{c-1}\prod_{k=1}^{M} R_{k,l}(z) \\ & + \vartheta_1 \sum_{\bar{n}\in U_1(N)} p(\bar{n}) z^{n_1} \Bigg]. \end{aligned} \tag{20.6}$$

Although the preceding arguments are quite intuitive, several terms in (20.6) need to be explained further. Φ is simply the probability that an arbitrary end-of-server-vacation instant happens to be at station 1 and that it occurs after a switch from station M; ϑ_i is the ratio of $g_i(\bar{0})$ and Φ, and it represents the fraction of new cycles beginning at station 1 during which the system becomes empty at station i. $U(N)$ represents the set of all states with exactly N customers in the system. Note that this is the set of startup states, following a server idle period. $p(\bar{n})$ represents the probability of observing state $\bar{n} \in U(N)$ at a start-up instant, and $u_{i,c}(\bar{n},z)$ denotes the PGF of the total contribution to station 1 queue at reference point from this state when the server is at station i, c cycles prior to the reference point. $p(\bar{n})$ and $u_{i,c}(\bar{n},z)$ are defined as follows:

$$p(\bar{n}) = \frac{N!}{n_1!\cdots n_M!}\prod_{j=1}^{M} p_j^{n_j} \tag{20.7}$$

and

$$u_{i,c}(\bar{n},z) = \prod_{j=i}^{M}\{L_{j,c}(z)\}^{n_j}\prod_{j=1}^{i-1}\{L_{j,c-1}(z)\}^{n_j}, \qquad i=1,\ldots,M,\ \ c\geq 0. \tag{20.8}$$

The first two terms on the right hand side of (20.6) denote the contribution to station 1 queue at reference point, if the system never becomes empty (the second term is necessary to eliminate the possibility of the system becoming empty at a start-of-server-vacation instance at any one of the M stations in each previous cycle); the third term accounts for the contributions from the moment system starts again, and the fourth and last term accounts for the fact that the reference point could be a end-of-server-vacation in which the server restarts at station 1, following an idle period there. Next, we find the first two derivatives of $k_1(z,1,\cdots,1)$, substitute in (20.2), and simplify to obtain the mean waiting times. The ϑ_i's are determined from the fact that system is never empty at the end-of-server-vacation instances at each station, and Φ turns out to be a function of ϑ_i's only [see Günalay and Gupta (1996a) for details]. We have

$$\begin{aligned} E[W_1] \\ = \quad & \frac{1-\rho_1}{2E[C]}\left[\frac{E[R_T]}{1-\rho}\right]^2 + \frac{\lambda_1 E[S_1^2] + [Var(R_M) + N(N-1)\vartheta_1/\Lambda^2]/E[C]}{2(1-\rho_1)} \\ & + \sum_{i=1}^{M}\left(\frac{\Gamma_i}{\rho_i^2}\right)\frac{\lambda_i E[S_i^2] + [Var(R_{i-1}) + N(N-1)\vartheta_i/\Lambda^2]/E[C]}{2(1-\rho_1)} \\ & + \sum_{i=1}^{M}\frac{N\vartheta_i}{\Lambda E[C]}\sum_{c=0}^{\infty}\frac{\gamma_{i,c}t_{i,c}}{\rho_i(1-\rho_1)}, \end{aligned} \tag{20.9}$$

where $\gamma_{i,c} = (\lambda_i/\lambda_1)L'_{i,c}(1)$, $i=1,\ldots,M$, $c \geq -1$,

$$t_{i,c} = \sum_{j=i+1}^{M} E[R_{j-1}]\frac{\gamma_{j,c}}{\rho_j} + \sum_{l=0}^{c-1}\sum_{k=1}^{M} E[R_{k-1}]\frac{\gamma_{k,l}}{\rho_k} + E[R_M], \tag{20.10}$$

$$\Gamma_i \triangleq \sum_{c=0}^{\infty}\gamma_{i,c}^2, \tag{20.11}$$

and

$$E[C] \quad = \quad \frac{\sum_{i=1}^{M} E[R_i] + (N\vartheta_i)/\Lambda}{1-\rho}. \tag{20.12}$$

The $\gamma_{i,c}$ terms are obtained recursively from the following relationship [see Srinivasan and Gupta (1996) for details]:

$$\gamma_{i,c} \quad = \quad \rho_i\left(\sum_{j=i}^{M}\gamma_{j,c} + \sum_{j=1}^{i-1}\gamma_{j,c-1}\right), \qquad i=1,\ldots,M, \quad c \geq 0. \tag{20.13}$$

Finally, the pseudo-conservation law, is given as follows:

$$\begin{aligned} \sum_{j=1}^{M}\rho_j E[W_j] \quad = \quad & \frac{\rho}{2(1-\rho)}\sum_{i=1}^{M}\left(\lambda_i E[S_i^2] + \frac{N(N-1)\vartheta_i/\Lambda^2}{E[C]}\right) + \frac{\rho E[R_T^2]}{2(1-\rho)E[C]} \\ & + \frac{E[R_T]}{2(1-\rho)}\left(\rho^2 - \sum_{i=1}^{M}\rho_i^2\right) + \sum_{i=1}^{M}\frac{N\rho_i Y_i}{1-\rho}, \end{aligned} \tag{20.14}$$

where Y_i is defined as

$$Y_i = \sum_{j=1}^{i-1}\frac{E[R_j]}{\Lambda E[C]}\left(\sum_{k=i+1}^{M}\vartheta_k + \sum_{k=1}^{j}\vartheta_k\right) + \sum_{j=i+1}^{M}\frac{E[R_j]}{\Lambda E[C]}\sum_{k=i+1}^{j}\vartheta_k, \qquad i=1,\ldots,M. \tag{20.15}$$

20.4 Some Related Models

In this Section, some variations of server and customer behavior are discussed that can be treated by slight modifications to the procedure described in Section 20.3.

20.4.1 Customer routing

Suppose that the customers do not leave the polling system upon receiving a single service. Instead, with probability η_{ij}, type i customers join station j queue to receive a type j service. We assume that $\sum_{j=1}^{M} \eta_{i,j} < 1$ and therefore, the customer leaves the system with probability $1 - \sum_{j=1}^{M} \eta_{i,j}$. The effective mean arrival rate at station i is now $\hat{\lambda}_i = \lambda_i + \sum_{j=1}^{M} \eta_{j,i}\hat{\lambda}_j$ and the stability condition has to be defined with respect to this new arrival rate. Surprisingly, there is very little that needs to be changed in our methodology described in Section 20.3. Essentially, we redefine (20.5) to account for the fact that a type i customer changes into other types of customers (including another of its own kind) as follows:

$$
\begin{aligned}
L_{i,c}(z) \quad = \quad & S_i^* \left(\sum_{j=i}^{M} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i-1} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right) \\
& \times \left[\sum_{j=i}^{M} \eta_{i,j} L_{j,c}(z) + \sum_{j=1}^{i-1} \eta_{i,j} L_{j,c-1}(z) \right] . \qquad (20.16)
\end{aligned}
$$

Using the above definition in (20.6), we obtain the PGF of station 1 queue length at a end-of-server-vacation instant at that station. However, since arrivals are no longer Poisson, the decomposition relationship in (20.1) does not apply. Therefore, a different method is needed to obtain mean waiting times.

It is possible to find the mean waiting times using a variation of the method described in Sidi, Levi and Fuhrmann (1992, pp. 128). In this method, we first find the mean queue length at station 1, conditioning on the three possible server states that could be observed at an arbitrary observation epoch. The server could either be serving some type j customers, performing a switchover, or else idling at an arbitrary station. These conditional mean queue lengths can be expressed in terms of moments of $k_1(z_1, z_2, \cdots, z_M)$ and the latter can be found using an equation similar to (20.6), but with vector arguments, i.e., $\boldsymbol{z} = (z_1, z_2, \cdots, z_M)$. The start of $L_{i,c}(\boldsymbol{z})$ recursions occur with $L_{i,-1}(\boldsymbol{z}) = z_i$. Next, utilizing the fact that the start of a new cycle is a regeneration point, it is easy to see that the proportion of time server spends in each state during a cycle is also the long run proportion of time it is in that state. Therefore, upon unconditioning, we obtain the arbitrary time mean queue length at station 1. Finally, applying Little's law, we can also obtain the mean waiting time.

A special case of such customer routing when $\eta_{i,i+1} = 1$, for all $i = 1, 2, \cdots$, $M-1$, and external arrivals occur only at station 1, gives rise to the tandem queue model [see Duenyas, Gupta and Lennon (1995) for the model with $N = 1$]. The model with arbitrary customer routing, but where the server never stops, i.e., $N = 0$, has been studied by Sidi, Levi and Fuhrmann (1992).

20.4.2 Stopping only at a preferred station

Suppose that the server stops only if it finds no customers in the system when it is ready to begin a vacation from station 1. In this case, although the system can be empty when the server is at other stations, it simply ignores that information unless it happens to be at station 1. This situation can be modeled with a slight change in our model in Section 20.3. We simply set all ϑ_j $(j \neq 1)$ to zero in Eqs. (20.9)—(20.15). Thus, we make server behavior at stations 2, $\cdots$, M independent of whether the system is empty or not. Note that now system cannot be empty only at a end-of-server-vacation instant from station 1, i.e., it could very well be empty at end-of-server-vacation instances from other stations. In the end, we have one equation in one unknown which could be solved easily [Günalay and Gupta (1996a)].

The ideas described above are easy to extend to situations in which there is more than one preferred station.

20.4.3 Gated or mixed service policy

Suppose that at a subset of stations, the server practices a gated service strategy. It is assumed that it continues to render exhaustive service at the remaining stations. Although Fuhrmann and Cooper's Stochastic Decomposition Theorem [Fuhrmann and Cooper (1985)] still applies, the queue is not necessarily empty at the start of a server vacation from a station at which the gated regime is followed. Therefore, the PGF of the queue length at station i (with gated service regime) at a customer departure instant is somewhat different from (20.1), as shown below:

$$
\begin{aligned}
&\Pi_i(1, \dots, z_i, \dots, 1) \\
&\quad = \frac{[k_i(1, \dots, S_i^*(\lambda_i - \lambda_i z_i), \dots, 1) - k_i(1, \dots, z_i, \dots, 1)] S_i^*(\lambda_i - \lambda_i z_i)}{k_i'(\bar{1})[S_i^*(\lambda_i - \lambda_i z_i) - z_i]} .
\end{aligned}
\tag{20.17}
$$

As before, the mean waiting time of type i customers is derived using the mean queue length of station i, which is obtained from Eq. (20.17). The result is as follows:

$$E[W_i] = \frac{(1+\rho_i)k_i''(\bar{1})}{2\lambda_i k_i'(\bar{1})} . \tag{20.18}$$

Eq. (20.6) is also valid for stations with gated service. Therefore, we can calculate the first and second moments of queue lengths by differentiating (20.6). The main difference from previous approach is that $L_{i,c}(z)$, $c \geq 0$, are now defined as follows:

$$L_{i,c}(z) = S_i^* \left(\sum_{j=i+1}^{N} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right) . \tag{20.19}$$

For systems with mixed service strategies, i.e., some stations operating under exhaustive and others under gated policy, we simply use either (20.2) and (20.5) or (20.18) and (20.19), depending on which service regime is in effect. Details of similarities between exhaustive and gated systems can be found in Konheim, Levi and Srinivasan (1994), Srinivasan and Gupta (1996) and Günalay and Gupta (1996a).

20.4.4 State-dependent setups

Suppose we have a polling system in which switchover times are unavoidable, but setups can be postponed. In this model, after the server polls a station, it sets up only if that queue is not empty. For example, an AGV must travel along its fixed route, but it does not need to set up to load/unload at a station if there are no jobs to pick up/deliver. The server may move among stations either according to a continuously roving or patient server protocol. Although the framework of Section 20.3 applies to polling models with state-dependent setups as well, their analysis is much harder. The major difficulty lies in calculating queue length distributions at the end-of-server-vacation instances.

Before presenting the new $k_1(z, 1, \ldots, 1)$, we need to define the following new notation. Let $T_{i,c}$ denote the total contributions to station 1 queue at the reference point, from the arrivals during a type i setup period, c cycles prior to the reference point. Then, its PGF can be written as

$$T_{i,c}(z) = T_i^* \left(\sum_{j=i}^{N} [\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i-1} [\lambda_j - \lambda_j L_{j,c-1}(z)] \right) , \quad i = 1, \ldots, M, \; c \geq 0. \tag{20.20}$$

Also, $F_{i,c}^0(z)$ is defined as the PGF of the total contributions of all customers in the system at the instant the server polls station i, c cycles ago, and finds no customers waiting at that queue, i.e.,

$$F^0_{i,c}(z) \triangleq \frac{1}{\Phi}k_i(L_{1,c-1}(z), \ldots, L_{i-1,c-1}(z), 1, L_{i+1,c}(z), \ldots, L_{M,c}(z)),$$
$$i = 1, \ldots, M, \quad c \geq 0. \tag{20.21}$$

Similarly, $F^0_{1,-1}(z)$ denotes the probability that at the reference point, station 1 is empty but not the system, i.e., $F^0_{1,-1}(z) = k_1(0, 1, \ldots, 1)/\Phi$. Recall that Φ is the probability that an end-of-server-vacation instant happens to be at station 1 and that it is not a start-up instant at station 1.

Now we are ready to update $k_1(z, 1, \ldots, 1)$, given in (20.6), as follows:

$$\begin{aligned} &k_1(z, 1, \ldots, 1) \\ &= \Phi\Bigg[T_1^*(\lambda_1 - \lambda_1 z)\Bigg(\prod_{c=0}^{\infty}\prod_{i=1}^{M} T_{i,c}(z)R_{i,c}(z) - \sum_{c=0}^{\infty}\sum_{i=1}^{M} F^0_{i,c}(z)T_{i,c}(z)E_{i,c}(z) \\ &\quad - \sum_{c=0}^{\infty}\sum_{i=1}^{M}\vartheta_i E_{i,c}(z) + \sum_{c=0}^{\infty}\sum_{i=1}^{M}\vartheta_i \sum_{\bar{n}\in U(N)} p(\bar{n})u_{i,c}(\bar{n}, z)E_{i,c}(z) \\ &\quad + \sum_{c=0}^{\infty}\sum_{i=1}^{M} F^0_{i,c}(z)E_{i,c}(z)\Bigg) + (1 - T_1^*(\lambda_1 - \lambda_1 z))F^0_{1,-1}(z) \\ &\quad + \vartheta_1 \sum_{\bar{n}\in U_1(N)} p(\bar{n})z^{n_1}\Bigg], \end{aligned} \tag{20.22}$$

where

$$E_{i,c}(z) = R_{i,c}(z)\prod_{j=i+1}^{M} R_{j,c}(z)T_{j,c}(z)\prod_{l=0}^{c-1}\prod_{k=1}^{M} R_{k,l}(z)T_{k,l}(z). \tag{20.23}$$

Notice that relationship (20.22) is identical to (20.6), but for the second, fifth and sixth terms on the right hand side of (20.22). These extra terms are needed to account for the fact that the server skips empty stations, without performing a setup, but continues to switch from one station to the next, so long as the system is not empty. Eq. (20.22) subsumes several variations of polling models with state-dependent server behavior, e.g., by setting $F^0_{i,c}(z) = 0$ and $T_{i,c}(z) = 1$ (i.e., setting $T_i = 0$) for all $i = 1, \ldots, M$ and $c \geq -1$, we obtain the N-threshold startup model of Section 20.3, and by setting $N = 0$, or $N = 1$, or $N > 1$, we obtain different state-dependent setups models [see Bradlow and Byrd (1987), Eisenberg (1971), Ferguson (1986), Günalay and Gupta (1996a,b) and Gupta and Srinivasan (1996b)].

Notice that the partial PGF's, $F^0_{i,c}(z)$'s, are unknown and difficult to obtain in terms of system parameters. The only known results are for a two station ($M = 2$) polling system [see Eisenberg (1971) for patient server model, and

Gupta and Srinivasan (1996b) for continuously roving server model]. All of the remaining methods proposed in the literature are approximations [Ferguson (1986), Gupta and Srinivasan (1996b), Bradlow and Byrd (1987) and Günalay and Gupta (1996b)]. Günalay and Gupta (1996b) have provided a near-exact numerical algorithm to obtain mean customer waiting times in a patient server model with state-dependent setups.

Threshold setups model is a generalization of the state-dependent setups model. Here, the server sets up only if the number of waiting customers at the polled station exceeds a minimum threshold. This model is related to the Stochastic version of the Economic Lot Scheduling problem, which is known to be a hard problem even in the deterministic setting [see, for example, Federgruen and Katalan (1996)]. A framework for analyzing this model, and a near-exact numerical algorithm to calculate the mean waiting times can be found in Günalay (1996). Other studies dealing with the threshold setups polling model include Hofri and Ross (1987), who examine a two station polling system, and Coffman, Puhalskii and Reiman (1995a, 1995b), who use a heave-traffic approximation. The model is considerably more complex, and therefore it is not discussed in detail here.

20.4.5 Periodic monitoring during idle period

Suppose there are some lower priority jobs that the server attends to, whenever the polling system becomes empty. As a result, the server is unavailable for a period of V time units, every time the system becomes empty. As opposed to the *continuous monitoring* strategy, here the server checks the system state periodically, i.e., every V time units until it finds at least one customer in the system. Then, it reactivates at the same station where it resided before attending to the lower priority customers. Note that with this strategy the start-up population is not fixed. This is how the periodic monitoring model differs from the continuous monitoring model of Section 20.3 in which the server always reactivates with exactly N customers in the system. When the system is not empty, we assume that the server sets up in a state-independent fashion.

The following expression presents the PGF of the queue length at station 1 end-of-server-vacation instances for the polling model with periodic monitoring:

$$\begin{aligned} &k_1(z,1,\ldots,1) \\ &= \Phi\Bigg[\prod_{c=0}^{\infty}\prod_{i=1}^{M} R_{i,c}(z) - \sum_{c=0}^{\infty}\sum_{i=1}^{M} \vartheta_i \prod_{j=i}^{M} R_{j,c}(z) \prod_{l=0}^{c-1}\prod_{k=1}^{M} R_{k,l}(z) \\ &\qquad + \sum_{c=0}^{\infty}\sum_{i=1}^{M} \vartheta_i V_{i,c}(z) \prod_{j=i}^{M} R_{j,c}(z) \prod_{l=0}^{c}\prod_{k=1}^{M} R_{k,l}(z) + \vartheta_1 V_1^*(\lambda_1 - \lambda_1 z)\Bigg], \end{aligned} \tag{20.24}$$

where $V_{i,c}(z)$ is the PGF of the contributions to station 1 queue at reference point, which come from arrivals during the time server is away attending to lower priority customers. The server departs from station i, c cycles prior to the reference point, i.e.,

$$V_{i,c}(z) = V^*\left(\sum_{j=i}^{N}[\lambda_j - \lambda_j L_{j,c}(z)] + \sum_{j=1}^{i-1}[\lambda_j - \lambda_j L_{j,c-1}(z)]\right), \quad i = 1, \ldots, M, \; c \geq 0. \tag{20.25}$$

Using Eqs. (20.24) and (20.25) in Section 20.3, we can obtain the mean waiting times for this model. Fuhrmann and Moon (1990) studied polling models with an arbitrary start-up distribution, and presented periodic monitoring as an example of what might cause an arbitrary start-up distribution. However, they modelled a microprocessor based system and therefore, assumed switchover times to be zero (negligible). Therefore, the above formulation reduces to their model upon setting $R_{i,c}(z) = 1$ for all $i = 1, \ldots, M$, $c \geq 0$ in (20.24).

20.5 Insights

In Eqs. (20.9) and (20.14), by putting $N = 0$ we get, respectively, the mean waiting time and the pseudo-conservation law for the system in which the server never stops [Boxma and Groenendijk (1987)]. Similarly, when $N = 1$ is substituted, we get the corresponding results for the patient server model [Srinivasan and Gupta (1996)]. Note that, Srinivasan and Gupta (1996) have defined the switchover times differently: in their notation, station $i \to i+1$ switchover time is denoted by R_{i+1}.

Next, if we let $R_i = R$, $S_i = S$ and $\lambda_i = \lambda$, $i = 1, \ldots, M$, we get a symmetric polling system. Then, $\lambda_i = \Lambda/M$, $\rho_i = \rho/M$ and $p_i = 1/M$. Furthermore, the empty system probabilities, ϑ_i, are the same for all stations, and we denote them by $\vartheta(N)$, $N \geq 0$. Similarly, let $E[W(N)]$ denote the mean waiting time at an arbitrary station. The last two quantities have an argument N, to explicitly recognize the dependence of these parameters on the start-up threshold N. Now, we can greatly simplify expression (20.9) for the mean station waiting time to obtain

$$E[W(N)] = \frac{(N - 1 + (M-1)\Lambda E[R])N\vartheta(N) + \Lambda^2(E[R^2] + (M-1)E[R]^2)}{2\Lambda(N\vartheta(N) + \Lambda E[R])} + \frac{\Lambda E[S^2] + \rho(M-1)E[R]}{2(1-\rho)}. \tag{20.26}$$

If we had only one station, i.e., $E[R] = 0$ and $\vartheta(N) = 1$, we can further simplify (20.26) by setting M equal to 1. Upon performing these simplifications, we obtain

$$E[W(N)] \quad = \quad \frac{(N-1)}{2\Lambda} + \frac{\Lambda E[S^2]}{2(1-\rho)}. \tag{20.27}$$

Accounting for differences in notation, Eq. (20.27) is the same as Eq. (11-117a) of Heyman and Sobel (1984, Volume I, p. 445) for the mean waiting time in a $M/G/1$ queue operating under N-policy.

Next, consider a symmetric polling system in which $N = 1$ and the server is forced to idle while it is in the process of switching between every pair of stations. Forced idling described above adds to the server overhead and effectively amounts to an increase in switchover time R. Intuitively, we would expect the mean waiting times to increase as switchover times increase. Surprisingly, this is not always the case. While it is true that $\vartheta(1)$, the empty system probability, decreases as R increases, the combined effect of these changes on $E[W(1)]$ is not always monotonic. This can be seen when (20.26) is simplified for $N = 1$, to obtain

$$E[W(1)] = \frac{M-1}{2}E(R) + \frac{\Lambda E[R^2]}{2[\vartheta(1) + \Lambda E(R)]} + \frac{\Lambda E(S^2) + \rho(M-1)E(R)}{2(1-\rho)}. \tag{20.28}$$

It is easy to see that the first and the third terms in (20.28) are increasing in $E(R)$. However, since $[\vartheta(1) + \Lambda E(R)]$ could increase in $E(R)$, the second term does not change in a monotonic fashion when $E(R)$ is increased. For a more rigorous analysis, consider the derivative of $E[W(1)]$ with respect to $E[R]$ given by

$$E[W(1)]' = \frac{M-1}{2(1-\rho)} + \frac{\Lambda E[R^2]'}{2(\vartheta(1) + \Lambda E[R])} - \frac{\Lambda E[R^2](\Lambda + \vartheta(1)')}{2(\vartheta(1) + \Lambda E[R])^2}. \tag{20.29}$$

Since $\vartheta(1)' < 0$ and $E[R^2]' > 0$, the above relationship can be either positive or negative depending on the magnitude of $E[R^2]$ with respect to $E[R]$. Therefore, the change in mean waiting time is not monotonic when we increase the mean switchover times.

Similar observations also apply to the case when $N > 1$ and when the system is asymmetric. However, the relationships are not so transparent in that case and it is therefore not so easy to see this counter-intuitive effect of forced idling on those systems.

Observations similar to the ones reported above were first made by Sarkar and Zangwill (1991) in the context of a polling system with $N = 0$. In fact, Sarkar and Zangwill noticed that reducing setup times can increase mean waiting time. Later, similar results were also observed in systems with $N = 1$ by Srinivasan and Gupta (1996). Gupta and Srinivasan (1996a) called this phenomenon the *variance paradox* since it usually occurs when switchover time variance is high. If switchover times are constant, i.e., $E[R] = R$, $E[R^2] = R^2$, and $E[R^2]' = 2R$, Eq. (20.29) yields

$$E[W(1)]' = \frac{M-1}{2(1-\rho)} + \frac{\Lambda R(2\vartheta(1) + \Lambda R - R\vartheta(1)')}{2(\vartheta(1) + \Lambda R)^2}. \tag{20.30}$$

The right hand side of (20.30) is always positive, proving that forced idling only increases mean waiting times when switchover times are constant. Gupta and Srinivasan have also provided a detailed discussion of this phenomenon and its implications for production theory. From recent results of N-threshold start-up polling systems (with $N > 1$), is it also possible to show that the variance paradox exists even in models involving a threshold start-up.

20.6 Future Directions

Polling models continue to gain popularity on account of their new applications in modelling of manufacturing, transportation, and storage/retrieval systems. The need to develop analysis of new models that are suitable for these environments is growing. Also, in these applications, system optimization (choosing system parameters and control policies) is important. However, optimization—for example, in determining optimal start-up and setup thresholds in order to minimize mean unfinished work in system—remains a difficult and open problem. Similarly, among the various server control policies, server idling behavior needs to be further investigated. These aspects of polling systems are substantially under-researched and more vigorous research effort is likely to occur in the next several years.

References

1. Altman, E., Konstantopoulos, P. and Liu, Z. (1992). Stability, monotonicity and invariant quantities in general polling systems, *Queueing Systems*, **11**, 35–57.

2. Benjaafar, S. and Gupta, D. (1996a). Workload allocation in multi-product, multi-facility production systems with setup times, *IIE Transactions* (submitted for publication).

3. Benjaafar, S. and Gupta, D. (1996b). Scope versus focus: Issues of flexibility, capacity, and number of production facilities, *IIE Transactions* (submitted for publication).

4. Borst, S. C. (1995). A globally gated polling system with a dormant server, *Probability in the Engineering and Information Sciences,* **9**, 239–254.

5. Boxma, O. J. and Groenendijk, W. P. (1987). Pseudo-conservation laws in cyclic-service systems, *Journal of Applied Probability*, **24**, 949–964.

6. Boxma, O. J., Weststrate, J. A. and Yechiali, U. (1993). A globally gated polling system with server interruptions, and applications to the repairman problem, *Probability in the Engineering and Information Sciences,* **7**, 187–208.

7. Bozer, Y. and Srinivasan, M. M. (1991). Tandem configurations for automated guided vehicle systems and the analysis of single vehicle loops, *IIE Transactions*, **23**, 72–82.

8. Bradlow, H. S. and Byrd, H. F. (1987). Mean waiting time evaluation of packet switches for centrally controlled PBX's, *Performance Evaluation*, **7**, 309–327.

9. Coffman, E. G., Jr., Puhalskii, A. A. and Reiman M. I. (1995a) Polling Systems with Zero Switchover Times: A Heavy Traffic Averaging Principle, *Annals of Applied Probability,* **5**, 681–719.

10. Coffman, E. G., Jr., Puhalskii, A. A. and Reiman M. I. (1995b) Polling Systems in Heavy-Traffic: A Bessel Process Limit, Working Paper, AT&T Bell Labs.

11. Cooper, R. B. (1970). Queues served in cyclic order: Waiting times, *The Bell System Technical Journal,* **49**, 399–413.

12. Cooper, R. B. (1981). *Introduction to Queueing Theory*, Second edition, Amsterdam: North-Holland. Reprinted by CEE Press of The George Washington University, 1990.

13. Duenyas, I., Gupta, D. and Lennon, T. (1996). Control of a single-server tandem queueing system with setups, *Operations Research* (to appear).

14. Duenyas, I. and Van Oyen, M. P. (1996). Heuristic scheduling of parallel heterogeneous queues with set-ups, *Management Science*, **42**, 814–829.

15. Eisenberg, M. (1971). Two queues with changeover times, *Operations Research,* **19**, 386–401.

16. Eisenberg, M. (1994). The polling system with a stopping server, *QUESTA,* **18**, 387–431.

17. Federgruen, A. and Katalan, Z. (1996). The stochastic economic lot scheduling problem: Cyclical base-stock policies with idle times, *Management Science*, **42**, 783–796.

18. Ferguson, M. J. (1986). Mean waiting time for a token ring with station dependent overheads, In *Local Area and Multiple Access Networks* (Ed., R. L. Pickholtz), pp. 43–67, Computer Science Press.

19. Fuhrmann, S. W. and Cooper, R. B. (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations, *Operations Research*, **33**, 1117–1129.

20. Fuhrmann, S. W. and Moon, A. (1990). Queues served in cyclic order with an arbitrary start-up distribution, *Naval Research Logistics*, **37**, 123–133.

21. Günalay, Y. (1996). State dependent server scheduling rules in polling systems, *Ph.D. Thesis*, McMaster University, Hamilton, Ontario, Canada.

22. Günalay, Y. and Gupta, D. (1996a). Threshold startup control policy for polling systems, *Management Science* (submitted for publication).

23. Günalay, Y. and Gupta, D. (1996b). Polling systems with a patient server and state dependent setup times, *IIE Transactions* (to appear).

24. Gupta, D. and Srinivasan, M. M. (1996a). The variance paradox and its implications for Japanese production theory, *Interfaces*, **26**, 69–77.

25. Gupta, D. and Srinivasan, M. M. (1996b). Polling systems with state-dependent setup times, *Queueing Systems* (to appear).

26. Heyman, D. P. and Sobel, M. J. (1982). *Stochastic Models in Operations Research, Volume I: Stochastic Processes and Operating Characteristics*, New York: McGraw-Hill.

27. Heyman, D. P. and Sobel, M. J. (1984). *Stochastic Models in Operations Research, Volume II: Stochastic Optimization*, New York: McGraw-Hill.

28. Hofri, M. and Ross, K. W. (1987) On the Optimal Control of Two Queues with Server Setup Times and Its Analysis, *SIAM Journal on Computing*, **16**, 399–420.

29. Konheim, A. G., Levi, H. and Srinivasan, M. M. (1994). Descendant set: An efficient approach for the analysis of polling systems, *IEEE Transactions on Communications*, **42**, 1245–1253.

30. Levi, H. and Sidi, M. (1990). Polling systems: Applications, modelling and optimization, *IEEE Transactions on Communications*, **38**, 1750–1760.

31. Liu, Z., Nain, P. and Towsley, D. (1992). On optimal polling policies, *QUESTA,* **11**, 59–83.

32. Sarkar, D. and Zangwill, W. I. (1991). Variance effects in cyclic production systems, *Management Science*, **37**, 444–453.

33. Sidi, M., Levi, H. and Fuhrmann, S. W. (1992). A queueing network with a single cyclically roving server, *Queueing Systems*, **11**, 121–144.

34. Srinivasan, M. M. and Gupta, D. (1996). When should a server be patient?, *Management Science*, **42**, 437–451.

35. Takagi, H. (1986). *Analysis of Polling Systems,* Cambridge, MA: The MIT Press.

36. Takagi, H. (1990). Queueing analysis of polling models: An update, In *Stochastic Analysis of Computer and Communication Systems* (Ed., H. Takagi), pp. 267–318, Amsterdam: North-Holland.

37. Takagi, H. (1994). *Queueing Analysis of Polling Models: Progress in 1990–93*, Institute of Socio-Economic Planning, University of Tsukuba, Japan.

PART V

Applications to Waiting Time Problems

21

Waiting Times and Number of Appearances of Events in a Sequence of Discrete Random Variables

Markos V. Koutras

University of Athens, Athens, Greece

Abstract: In this article, the probability and moment generating functions of the number of appearances of a pattern $\mathcal{E}$ in a sequence of discrete random variables (repeated trials) are expressed in terms of the generating function of the waiting time for the r-th occurrence of $\mathcal{E}$. The special case of delayed recurrent events is also examined in some detail. Finally, the general theory is employed for a systematic investigation of success runs enumeration problems in a sequence of binary outcomes arising from a first-order Markov chain.

Keywords and phrases: Recurrent events, success runs, Markov dependence, distributions of order k, patterns, waiting time distributions

21.1 Introduction

Let $Z_0, Z_1, Z_2, \ldots$ be a sequence of repeated trials with possible outcomes E_j, $j = 1, 2, \ldots$. We suppose that it is feasible in principle to continue the trials indefinitely whereas there is no particular need to assume that they are independent (as a matter of fact, certain applications to Markov dependent outcomes will be of special interest in the sequel). The term *a pattern* $\mathcal{E}$ will refer to a specific string (succession of outcomes) composed of characters E_j, $j = 1, 2, \ldots$. Given a sequence of outcomes $E = E_{j_1} E_{j_2} \ldots E_{j_k}$, we shall say that the pattern $\mathcal{E}$ (or, allowing an abuse of the language, the event $\mathcal{E}$) has occurred in E, if there exists a substring of E that matches exactly with $\mathcal{E}$.

In the pattern matching area, there are two different classes of problems one could look upon. In the first, the interest focuses on the waiting time until the first, or in general the r-th occurrence of the event $\mathcal{E}$. In the second, the random variable of primary interest enumerates the appearances of the event $\mathcal{E}$ in a fixed number of trials.

The importance of those variables arises from their wide applicability in many scientific areas, which include meteorology, molecular biology, radar detection, quality control, psychology, reliability theory, computer science, etc. It is not therefore surprising that, during the last decades, an increasing research activity has been observed on related problems. Waiting times for appearances of patterns and runs (a special pattern consisting of alike symbols) have been studied by numerous authors; see, for example, Aki (1992), Blom and Thorburn (1982), Feller (1968), Gerber and Li (1981), Glaz (1983), Guibas and Odlyzko (1980, 1981), Koutras (1996a) and Solovev (1966) for the case of independent and identically distributed (i.i.d.) trials, and Aki (1992), Aki, Balakrishnan and Mohanty (1996), Balasubramanian, Viveros and Balakrishnan (1993), Banjevic (1988), Benevento (1984), Chryssaphinou and Papastavridis (1990), Koutras and Alexandrou (1997), Koutras (1996b), Mohanty (1994), Schwager (1983) and Viveros, Balasubramanian and Balakrishnan (1994) for Markov dependent trials. The second of the aforementioned problems, i.e., the investigation of the number of occurrences of a pattern or run in a fixed sequence of outcomes has also been treated by many authors including Charalambides (1994), Chen and Glaz (1997), Chryssaphinou and Papastavridis (1988), Dembo and Karlin (1992), Fu (1996), Godbole (1991), Godbole and Schaffner (1993), Hirano and Aki (1993) and Mohanty (1994). In an excellent monograph by Barbour, Holst and Janson (1992), several (Poisson) approximations to the distributions mentioned above are described. For a recent comprehensive list of publications on these and related problems, we refer to Godbole (1994). The upcoming book by Balakrishnan and Koutras (1997) presents a lucid and elaborate account of various developments relating to runs and patterns with applications.

In this paper, we investigate the interrelation between the distributions of the waiting times and the number of appearances of a pattern in a sequence of repeated trials. After the introduction of the necessary notations in Section 21.2, we proceed to the derivation of formulae associating the generating function of the waiting time for the r-th appearance of a pattern to the generating function of the number of appearances distribution (Section 21.3). In addition, the first and second order moment generating functions of the latter are expressed in terms of the double generating function of the waiting time distribution. In Section 21.4, we specialize to a wide class of patterns (delayed recurrent events) and establish formulae involving the single recurrence time generating functions. Section 21.5 serves as an illustration of how the general theory of Section 21.4 can be employed for deriving results relating to success runs enumeration in sequences of Markov dependent trails. Finally, in Section 21.6, some concluding remarks and possible directions for further research are discussed briefly.

21.2 Definitions and Notations

Let $Z_0, Z_1, Z_2, \ldots$ be a sequence of repeated trials with possible outcomes E_j, $j = 1, 2, \ldots$, and $\mathcal{E}$ be a specific pattern (single or composite) such that with probability 1, $\mathcal{E}$ occurs at least once in an indefinitely prolonged sequence of trials. Then, the waiting time T until the first occurrence of $\mathcal{E}$ is a random variable with probability mass function

$$h(n) = \Pr[T = n], \qquad n = 1, 2, \ldots .$$

Let us denote by T_r ($r = 1, 2, \ldots$) the waiting time for the r-th occurrence of $\mathcal{E}$ and its probability mass function by

$$h_r(n) = \Pr[T_r = n], \qquad n = 1, 2, \ldots .$$

Clearly, $T_1 = T$ and $h_1(n) = h(n)$. For our convenience and also for facilitating the derivation of more compact formulae, we impose the convention

$$h_r(0) = \Pr[T_r = 0] = \delta_{r,0}, \qquad r = 0, 1, \ldots \tag{21.1}$$

(δ_{ij} is the Kronecker's delta function) which makes $h_r(n)$ meaningful for all non-negative integers n and assigns T_0 a degenerate probability mass function. The single and double generating functions of T_r, $r = 0, 1, \ldots$, will be denoted by $H_r(z)$ and $H(z, w)$, respectively; i.e.,

$$\begin{aligned} H_r(z) = E[z^{T_r}] &= \sum_{n=0}^{\infty} \Pr[T_r = n] z^n = \sum_{n=0}^{\infty} h_r(n) z^n, \quad r \geq 1, \quad H_0(z) \equiv 1, \\ H(z, w) &= \sum_{r=0}^{\infty} H_r(z) w^r = \sum_{r=0}^{\infty} \sum_{n=0}^{\infty} \Pr[T_r = n] z^n w^r. \end{aligned}$$

Another random variable of primary interest in the present context is the number X_n of occurrences of event $\mathcal{E}$ in the first n trials. Its probability mass function will be defined by

$$g_n(x) = \Pr[X_n = x], \qquad x = 0, 1, \ldots$$

and the corresponding single and double generating functions by

$$\begin{aligned} G_n(w) = E[w^{X_n}] &= \sum_{x=0}^{\infty} \Pr[X_n = x] w^x = \sum_{x=0}^{\infty} g_n(x) w^x, \quad n \geq 1, \\ G(z, w) &= \sum_{n=0}^{\infty} G_n(w) z^n = \sum_{n=0}^{\infty} \sum_{x=0}^{\infty} \Pr[X_n = x] w^x z^n. \end{aligned}$$

We conventionally set $G_0(z) = 1$ for all w, which is equivalent to

$$\Pr[X_0 = x] = \delta_{x,0}, \qquad x = 0, 1, \dots . \tag{21.2}$$

21.3 General Results

In this section, we will investigate the connection between the distributions of T_r and X_n. Our basic results provide expressions for certain generating functions relating to the distribution of X_n in terms of $H_r(z)$ and $H(z, w)$. The key point for establishing these results is the obvious identity

$$\Pr[X_n \geq r] = \Pr[T_r \leq n], \qquad n, r \geq 1. \tag{21.3}$$

Theorem 21.3.1 *The single generating function of the sequence* $\{\Pr[X_n = x]\}_{n \geq 0}$ *is given by*

$$\sum_{n=0}^{\infty} \Pr[X_n = x] z^n = \frac{1}{1-z}\, [H_x(z) - H_{x+1}(z)], \qquad x = 0, 1, \dots . \tag{21.4}$$

Proof. For $x \geq 1$ we get, by virtue of (21.3),

$$\begin{aligned}
\Pr[X_n = x] &= \Pr[X_n \geq x] - \Pr[X_n \geq x+1] = \Pr[T_x \leq n] - \Pr[T_{x+1} \leq n] \\
&= \sum_{j=1}^{n} \Pr[T_x = j] - \sum_{j=1}^{n} \Pr[T_{x+1} = j], \quad n \geq 1,
\end{aligned}$$

and, therefore,

$$\sum_{n=1}^{\infty} \Pr[X_n = x] z^n = \sum_{n=1}^{\infty} \sum_{j=1}^{n} \Pr[T_x = j] z^n - \sum_{n=1}^{\infty} \sum_{j=1}^{n} \Pr[T_{x+1} = j] z^n. \tag{21.5}$$

Interchanging the orders of summation in the RHS, it is easy to verify that [see also (21.1)]

$$\begin{aligned}
\sum_{n=1}^{\infty} \sum_{j=1}^{n} \Pr[T_x = j] z^n &= \frac{1}{1-z} \sum_{j=1}^{\infty} \Pr[T_x = j] z^j = \frac{1}{1-z}\, H_x(z), \\
\sum_{n=1}^{\infty} \sum_{j=1}^{n} \Pr[T_{x+1} = j] z^n &= \frac{1}{1-z} \sum_{j=1}^{\infty} \Pr[T_{x+1} = j] z^j = \frac{1}{1-z}\, H_{x+1}(z).
\end{aligned}$$

We now substitute these expressions into (21.5) and use the convention in (21.2) to obtain the formula in (21.4) for $x \geq 1$. For the special case $x = 0$, we first notice that

$$\Pr[X_n = 0] = 1 - \Pr[T_1 \leq n] = 1 - \sum_{j=1}^{n} \Pr[T_1 = j], \quad n \geq 1,$$

which readily yields

$$\sum_{n=1}^{\infty} \Pr[X_n = 0]z^n = -1 + \frac{1}{1-z}\,[1 - H_1(z) + \Pr[T_1 = 0]]\,.$$

Taking into account conventions (21.1) and (21.2) to complete the series in the LHS and eliminating the last summand of the RHS, we obtain the formula in (21.4) for $x = 0$. ■

As an example, let event $\mathcal{E}$ stand as an abbreviation for "a success occurs" in a sequence of Bernoulli trials with success probability $p = 1 - q$. Then, T_r follows a negative binomial distribution with probability generating function

$$H_r(z) = \sum_{n=0}^{\infty} \Pr[T_r = n]z^n = \left(\frac{pz}{1-qz}\right)^r \tag{21.6}$$

and expanding

$$\frac{1}{1-z}\,[H_x(z) - H_{x+1}(z)] = (pz)^x(1-qz)^{-(x+1)},$$

we get, in view of Theorem 21.3.1,

$$\sum_{n=0}^{\infty} \Pr[X_n = x]z^n = \sum_{n=x}^{\infty} \binom{n}{x} p^x q^{n-x} z^n$$

—an identity reestablishing the well-known fact that the number X_n of successes in n trials follows a binomial distribution.

Theorem 21.3.2 *The double generating function $G(z,w)$ of the probability mass function $g_n(x)$ can be expressed in terms of the double generating function $H(z,w)$ of the probability mass function $h_r(n)$ as*

$$G(z,w) = \frac{(w-1)H(z,w)+1}{w(1-z)}\;. \tag{21.7}$$

PROOF. Multiplying (21.4) by w^x and summing over all $x = 0, 1, \ldots$ yields

$$\sum_{x=0}^{\infty}\sum_{n=0}^{\infty} \Pr[X_n = x]z^n w^x = \frac{1}{1-z}\left(\sum_{x=0}^{\infty} H_x(z)w^x - \sum_{x=0}^{\infty} H_{x+1}(z)w^x\right)$$

and since

$$\sum_{x=0}^{\infty} H_{x+1}(z)w^x = \frac{1}{w}\sum_{x=1}^{\infty} H_x(z)w^x = \Big(H(z,w) - 1\Big)/w,$$

the previous formula can be rewritten as

$$G(z,w) = \frac{1}{1-z}\left[H(z,w) - \frac{1}{w}\Big(H(z,w) - 1\Big)\right]$$

which proves the desired result. ■

It should be noted that the inversion of formula (21.7) produces an expression of $H(z,w)$ in terms of $G(z,w)$ as

$$H(z,w) = \frac{w(1-z)G(z,w)-1}{w-1}\ .$$

In view of Theorem 21.3.2, it is clear that, should the probability mass functions $h_r(n)$, $n = 0,1,\ldots$, be known for all $r = 0,1,\ldots$, the probability mass functions $g_n(x)$, $x = 0,1,\ldots$, can be identified through (21.7), and vice versa. Therefore, while investigating the occurrences of specific patterns $\mathcal{E}$ in a sequence of trials, it suffices to study either the waiting times for the r-th occurrence or the number of occurrences in a fixed number of trials. In most of the cases, the first problem offers greater simplicity; in addition, as will be demonstrated in the next section, for a wide class of repetitive patterns, the probability mass function $h_r(z)$ can be directly established through the distributions of waiting times for the first appearance of $\mathcal{E}$.

Before closing this section, let us give some formulae expressing the generating functions of the first two moments of T_r and X_n by means of the double generating function $H(z,w)$.

Theorem 21.3.3 *The generating functions of the means $E[T_r]$ and $E[X_n]$ are given by*

$$\begin{aligned}\sum_{r=0}^{\infty} E[T_r]w^r &= \left[\frac{\partial}{\partial z}\, H(z,w)\right]_{z=1},\\ \sum_{n=0}^{\infty} E[X_n]z^n &= \frac{H(z,1)-1}{1-z}\ .\end{aligned}$$

PROOF. It is simple to note that

$$\frac{\partial}{\partial z}\, H(z,w) = \sum_{r=0}^{\infty} H_r'(z)w^r \tag{21.8}$$

and the first formula of the theorem is readily ascertainable from the well-known identity

$$E[T_r] = H_r'(1).$$

In a similar way,

$$\sum_{n=0}^{\infty} E[X_n]z^n = \left[\frac{\partial}{\partial w}\, G(z,w)\right]_{w=1}$$

and the proof is completed upon differentiating (21.7) with respect to w and then substituting $w = 1$. ■

It is worth mentioning that the second formula of Theorem 21.3.3 could also be established without making use Theorem 21.3.2. To achieve this, it suffices to observe that

$$E[X_n] = \sum_{r=1}^{\infty} \Pr[X_n \geq r] = \sum_{r=1}^{\infty} \Pr[T_r \leq n]$$

which yields

$$E[X_n] - E[X_{n-1}] = \sum_{r=1}^{\infty} \Pr[T_r = n].$$

Multiplying by z^n and summing up for $n = 1, 2, \ldots$, we get

$$\sum_{n=1}^{\infty} [E(X_n) - E(X_{n-1})]z^n = \sum_{r=1}^{\infty} H_r(z) = H(z,1) - 1$$

or, equivalently,

$$(1-z) \sum_{n=0}^{\infty} E[X_n]z^n = H(z,1) - 1$$

which reestablishes the second part of Theorem 21.3.3.

Theorem 21.3.4 *The generating functions of the second order moments $E[T_r^2]$ and $E[X_n^2]$ are given by*

$$\begin{aligned}
\sum_{r=0}^{\infty} E[T_r^2]w^r &= \frac{\partial}{\partial z}\left[z \frac{\partial}{\partial z} H(z,w)\right]_{z=1} \\
\sum_{n=0}^{\infty} E[X_n^2]z^n &= \frac{1}{1-z}\left\{2\left[\frac{\partial}{\partial w} H(z,w)\right]_{w=1} - H(z,1) + 1\right\}.
\end{aligned}$$

PROOF. The first formula is an immediate consequence of the obvious identities

$$\begin{aligned}
\frac{\partial}{\partial z}\left[z \frac{\partial}{\partial z} H(z,w)\right] &= \sum_{r=1}^{\infty} \left(\frac{d}{dz} [zH'(z)]\right) w^r, \\
E[T_r^2] &= \frac{d}{dz} [zH'(z)]|_{z=1}.
\end{aligned}$$

Employing the same arguments for X_n, we get

$$\sum_{n=0}^{\infty} E[X_n^2]z^n = \frac{\partial}{\partial w}\left[w \frac{\partial}{\partial w} G(z,w)\right]_{w=1}$$

and the second part of the theorem is then easily established by noting from (21.7) that

$$w \frac{\partial}{\partial w} G(z,w) = \frac{1}{(1-z)w}\left[w(w-1) \frac{\partial}{\partial w} H(z,w) + H(z,w) - 1\right]$$

differentiating with respect to w and then setting $w = 1$. ■

As an application of Theorems 21.3.2—21.3.4, let us demonstrate how one can derive the probability mass function and the first two moments of the binomial distribution, through the double generating function of the negative binomial distribution. It is easily seen that in this special case [*cf.* (21.6)]

$$H(z,w) = \sum_{r=0}^{\infty} H_r(z) w^r = \frac{1-qz}{1-z(q+pw)} \tag{21.9}$$

and substituting this expression in (21.7), we get

$$G(z,w) = \frac{1}{1-z(q+pw)} = \sum_{n=0}^{\infty} (q+pw)^n z^n.$$

This expression simply implies that the generating function of X_n is

$$G_n(w) = (q+pw)^n = \sum_{x=0}^{\infty} \binom{n}{x} p^x q^{n-x} w^x$$

and, therefore,

$$\Pr[X_n = x] = \binom{n}{x} p^x q^{n-x}.$$

On the other hand, using (21.9) in Theorems 21.3.3 and 21.3.4, and then expanding the resulting expressions in power series, we get

$$\begin{aligned}
\sum_{n=0}^{\infty} E[X_n] z^n &= \frac{pz}{(1-z)^2} = \sum_{n=1}^{\infty} (np) z^n, \\
\sum_{n=0}^{\infty} E[X_n^2] z^n &= \frac{p(p-q)z^2 + pz}{(1-z)^3} = \sum_{n=1}^{\infty} \left\{ (np)^2 + npq \right\} z^n
\end{aligned}$$

which reveals the well-known formulae $E[X_n] = np$ and $E(X_n^2) = (np)^2 + npq$, $n \geq 1$.

21.4 Waiting Times and Number of Occurrences of Delayed Recurrent Events

The results presented in the last section are fairly general in the sense that no particular assumption has been made about the original trials $Z_0, Z_1, Z_2, \ldots$ (it is not necessary that they be independent or have the same distribution) or about the nature of the pattern $\mathcal{E}$. There are, however, naturally arising event sets (patterns) $\mathcal{E}$ for which a special mention would be worthwhile; one such category is the subject of this section.

A patter $\mathcal{E}$ is called *recurrent event* if after each occurrence of $\mathcal{E}$ the trials start from scratch, i.e., the trials following an occurrence of $\mathcal{E}$ form a replica of the whole experiment. It is understood that the waiting time between successive occurrences of $\mathcal{E}$ (recurrence times) are mutually independent random variables having the same distribution. Feller (1968) introduced a slight extension of the notion of recurrent events by allowing the first occurrence of $\mathcal{E}$ to have a distribution different from the distribution of the recurrence times of the subsequent appearances of $\mathcal{E}$. Such a pattern was named *delayed recurrent event* since now the definition of recurrent event applies only if the trials leading up to the first occurrence of $\mathcal{E}$ are disregarded.

It is clear that the waiting time T_r for the r-th appearance of a delayed recurrent event is the sum of r random variables (recurrence times) which are independent although the distribution of the first one might be different from the common distribution of the rest $r-1$. Accordingly, the probability generating function of T_r can be expressed as

$$H_r(z) = \sum_{n=0}^{\infty} \Pr[T_r = n] z^n = H(z) A^{r-1}(z), \qquad r \geq 1 \tag{21.10}$$

where $H(z)$ and $A(z)$ are proper generating functions. Under this assumption, the outcomes of the analysis discussed earlier take a more appealing form as it is illustrated in the following three theorems.

Theorem 21.4.1 *The single generating function of the sequence* $\{\Pr[X_n = x]\}_{n\geq 0}$ *for a pattern* $\mathcal{E}$ *satisfying (21.10) is given by*

$$\sum_{n=0}^{\infty} \Pr[X_n = x] z^n = \begin{cases} H(z) A^{x-1}(z)\, \frac{1-A(z)}{1-z} & \text{for } x \geq 1 \\ \frac{1-H(z)}{1-z} & \text{for } x = 0. \end{cases}$$

PROOF. The result follows instantly from (21.4) with the aid of (21.10). ■

It is worth noticing that for a pure (not delayed) recurrent event $\mathcal{E}$, we have $A(z) = H(z)$ and the aforementioned generating function then simplifies to

$$\sum_{n=0}^{\infty} \Pr[X_n = x] z^n = H^x(z)\, \frac{1 - H(z)}{1 - z}$$

[see also Feller (1968, p. 340)].

Theorem 21.4.2 *The double generating function* $G(z, w)$ *of the probability mass function* $g_n(x)$ *for a pattern* $\mathcal{E}$ *satisfying (21.10) is given by*

$$G(z, w) = \frac{1}{1-z} \left[1 - H(z)\, \frac{1 - w}{1 - wA(z)} \right]. \tag{21.11}$$

PROOF. From (21.10), we obtain

$$\begin{aligned} H(z,w) &= 1+\sum_{r=1}^{\infty} H_r(z)w^r = 1+H(z)w\sum_{r=1}^{\infty}[wA(z)]^{r-1} \\ &= 1+\frac{wH(z)}{1-wA(z)} \end{aligned} \tag{21.12}$$

which, when used in (21.7), gives

$$G(z,w) = \frac{1}{w(1-z)}\left[(w-1)\left(1+\frac{wH(z)}{1-wA(z)}\right)+1\right].$$

Eq. (21.11) is readily obtained after some algebraic simplification. ■

For a recurrent event $\mathcal{E}$, the double generating function $G(z,w)$ in (21.11) reduces to

$$G(z,w) = \frac{1-H(z)}{1-z}\,\frac{1}{1-wH(z)}\,.$$

Theorem 21.4.3 *The generating functions of the first two moments of X_n for a pattern $\mathcal{E}$ satisfying (21.10) are given by*

$$\sum_{n=0}^{\infty} E[X_n]z^n = \frac{H(z)}{(1-z)\{1-A(z)\}}\,, \tag{21.13}$$

$$\sum_{n=0}^{\infty} E[X_n^2]z^n = \frac{H(z)\{1+A(z)\}}{(1-z)\{1-A(z)\}^2}\,. \tag{21.14}$$

PROOF. The generating function in (21.13) follows immediately if we use the second part of Theorem 21.3.3 in conjunction with (21.12). To derive the generating function in (21.14), begin by differentiating (21.12) with respect to w to obtain

$$\frac{\partial}{\partial w}\,H(z,w) = \frac{H(z)}{(1-wA(z))^2}$$

and then substitute both $H(z,w)$ and its derivative (evaluated at $w=1$) in the second formula of Theorem 21.3.4. ■

Needless to say, both (21.13) and (21.14) could also be established by considering proper (partial) derivatives of (21.11) at $w=1$ (*cf.* proofs of Theorems 21.3.3 and 21.3.4).

Subtracting (21.13) from (21.14), one may easily produce an expression for the generating function of the second descending factorial moments of X_n as

$$\sum_{n=0}^{\infty} E[X_n(X_n-1)]z^n = \frac{2H(z)A(z)}{(1-z)\{1-A(z)\}^2}\,. \tag{21.15}$$

Similarly, the generating function of the second ascending factorial moments of X_n can be obtained as

$$\sum_{n=0}^{\infty} E[X_n(X_n+1)]z^n = \frac{2H(z)}{(1-z)\{1-A(z)\}^2} \tag{21.16}$$

which reveals that

$$\{1-A(z)\}\sum_{n=0}^{\infty} E[X_n(X_n+1)]z^n = 2\sum_{n=0}^{\infty} E[X_n]z^n. \tag{21.17}$$

Apparently, the last expression leads to simple recursive relations for the second ascending factorial moments of X_n.

In closing, we mention that, on setting $A(z) = H(z)$ we deduce the generating functions of the second order moments of X_n for a recurrent event $\mathcal{E}$ [see also Feller (1968, p. 341)].

21.5 Distribution of the Number of Success Runs in a Two-State Markov Chain

This section is designed to serve as an illustration of how the general theory, presented thus far, can be applied in a systematic manner to success runs enumeration problems in a sequence of Markov dependent trials.

Let $Z_0, Z_1, Z_2, \ldots$ be a time homogeneous two-state Markov chain with transition probabilities

$$p_{ij} = \Pr[Z_t = j \mid Z_{t-1} = i], \qquad t \geq 1,\ 0 \leq i, j \leq 1$$

and initial distribution $p_j = \Pr[Z_0 = j]$, $j = 0, 1$. Any uninterrupted sequence of k consecutive 1's (successes) will be called *success run of length k* (k is a positive integer). The classical scheme for enumerating runs of fixed length is the one proposed by Feller (1968). According to this, we start counting from scratch each time a succession of k consecutive 1's is observed (non-overlapping counting). Ling (1988) proposed an alternative enumeration technique in which an uninterrupted sequence of $l \geq k$ successes preceded and followed by a failure accounts for $l - k + 1$ runs (overlapping counting). Finally, a third enumeration procedure can be initiated by viewing a succession of at least k consecutive 1's as a single run, i.e., once the number of consecutive 1's exceeds k, we don't care about the actual length of the run.

The distributions of the number of success runs of length k in a sequence of independent Bernoulli trials have been recently termed as *binomial distributions of order k*, while the corresponding waiting time distributions as *geometric and* *negative binomial distributions of order k*. Clearly, the special case $k = 1$ leads to the classical binomial, geometric and negative binomial distributions. For more details on this subject, an authoritative reference is Johnson, Kotz and Kemp (1992).

Run related problems under Markovian dependence set-ups have been studied among others by Aki and Hirano (1993), Aki , Balakrishnan and Mohanty (1996), Balasubramanian, Viveros and Balakrishnan (1993), Hirano and Aki (1993), Koutras and Alexandrou (1997), Rajarshi (1974), Schwager (1983) and Uchida and Aki (1995). A combinatorial development to this problem was recently given by Mohanty (1994).

In a recent paper, Koutras (1996b), using the Markov chain imbedding technique introduced by Fu and Koutras (1994) and subsequently refined by Koutras and Alexandrou (1995), conducted a unified study of success runs waiting time problems in a sequence of Markov dependent trials. A careful inspection of the results given there reveals that for all three enumeration schemes mentioned above, the probability generating function of T_r can be expressed in the form (21.10). Consequently, one can make use of the results presented in Section 21.4 to study the distributions of the number of success runs in a fixed number of Markov dependent trials. We now proceed to a brief analysis of these problems from this point of view.

21.5.1 Non-overlapping success runs

In this case, the probability generating function of T_r is given by [Koutras (1996b)]

$$H_r(z) = H(z)A^{r-1}(z), \qquad r \geq 1, \tag{21.18}$$

where

$$H(z) = \frac{P(z)(p_{11}z)^{k-1}}{Q(z)}, \tag{21.19}$$

$$A(z) = \frac{(p_{01}z)(p_{11}z)^{k-1}}{Q(z)} \tag{21.20}$$

with

$$\begin{aligned} P(z) &= p_1 + (p_0p_{01} - p_1p_{00})z, \\ Q(z) &= 1 - p_{00}z - p_{01}p_{10}z^2\sum_{i=2}^{k}(p_{11}z)^{i-2}. \end{aligned} \tag{21.21}$$

Evidently, $H(z)$ is the probability generating function of T_1, while $A(z)$ is the probability generating function of the waiting time for the occurrence of a success run of length k in a sequence of Markov dependent trials with initial distribution

$$\Pr[Z_0 = 0] = 1, \quad \Pr[Z_0 = 1] = 0.$$

Applying Theorem 21.4.2, we may easily verify that the double generating function $G(z, w)$ of X_n, the number of non-overlapping success runs of length k in a sequence of n Markov dependent trials, can be expressed as

$$G(z,w) = \frac{1}{1-z}\left[1 - \frac{(1-w)(p_{11}z)^{k-1}P(z)}{Q(z) - w(p_{01}z)(p_{11}z)^{k-1}}\right].$$

For the special case of i.i.d. Bernoulli trials ($p_0 = 1$, $p_1 = 0$, $p_{01} = p_{11} = p$, $p_{10} = p_{00} = q$, $p + q = 1$), this formula reduces to

$$G(z,w) = \frac{1-(pz)^k}{1 - z + [qz - w(1-pz)](pz)^k}$$

and coincides with the one given by Koutras and Alexandrou (1995).

For the benefit of a practical minded reader, we use this occasion to develop some simple recurrence relations for $\mu_n = E[X_n]$. From Theorem 21.4.3, it follows that

$$\sum_{n=0}^{\infty} E[X_n]z^n = \frac{P(z)(p_{11}z)^{k-1}}{(1-z)[Q(z) - (p_{01}z)(p_{11}z)^{k-1}]}. \tag{21.22}$$

Replacing $P(z)$ and $Q(z)$ in (21.22) by the expressions in (21.21) and carrying out some algebra, we get

$$\sum_{n=0}^{\infty} \mu_n z^n = \frac{p_1 p_{11}^{k-1} z^{k-1} + (p_0 p_{01} - p_1 p_{00}) p_{11}^{k-1} z^k}{1 + \sum_{i=1}^{k+1} \alpha_i z^i}$$

where, for our convenience, we have set

$$\begin{aligned}
\alpha_1 &= -p_{00} - 1, \quad \alpha_2 = p_{00} - p_{01}p_{10}, \\
\alpha_i &= p_{01}p_{10}^2 p_{11}^{i-3}, \quad 3 \le i \le k-1, \\
\alpha_k &= p_{01}p_{11}^{k-3}(p_{10} - p_{11}), \quad \alpha_{k+1} = p_{01}p_{11}^{k-2}.
\end{aligned}$$

Next, we multiply both sides of the generating function by the denominator of the RHS and perform the classical analysis on the resulting power series to obtain the recurrence relation

$$\mu_n = -\sum_{i=1}^{k+1} \alpha_i \mu_{n-i}, \qquad n \ge k+1.$$

This relation can be used in conjunction with the initial conditions

$$\mu_n = \begin{cases} 0 & \text{if } 0 \le n \le k-2 \\ p_1 p_{11}^{k-1} & \text{if } n = k-1 \\ (p_1 + p_0 p_{01}) p_{11}^{k-1} & \text{if } n = k \end{cases} \tag{21.23}$$

to compute any $\mu_n = E[X_n]$ in a simple recursive manner. For an alternate recurrence relation, we could write $Q(z)$ as

$$Q(z) = 1 - p_{00}z - p_{01}p_{10}z^2 \, \frac{1-(p_{11}z)^{k-1}}{1-p_{11}z} \tag{21.24}$$

and substitute into (21.22) to get

$$\sum_{n=0}^{\infty} \mu_n z^n = \frac{[p_1 + (p_0p_{01} - p_1p_{00})z](1-p_{11}z)(p_{11}z)^{k-1}}{(1-z)^2[1+(p_{01}-p_{11})z - p_{01}p_{11}^{k-1}z^k]} \,.$$

On introducing the notation $b = p_{01} - p_{11}$, $\quad c = p_{01}p_{11}^{k-1}$, we obtain

$$\sum_{n=0}^{\infty} \mu_n z^n = \frac{p_1p_{11}^{k-1}z^{k-1} + [p_{01} - p_1(1+p_{11})]p_{11}^{k-1}z^k - (p_0p_{01} - p_1p_{00})p_{11}^k z^{k+1}}{1+(b-2)z+(1-2b)z^2+bz^3-cz^k+2cz^{k+1}-cz^{k+2}}$$

which yields the recurrence relation

$$\mu_n = (2-b)\mu_{n-1} + (2b-1)\mu_{n-2} - b\mu_{n-3} + c\mu_{n-k} - 2c\mu_{n-k-1} + c\mu_{n-k-2}, \quad n \geq k+2.$$

The initial conditions required are given again by (21.23) with the additional entry

$$\mu_{k+1} = p_{11}^{k-1}[1 + p_1 - p_{00}^2 - p_1p_{01}(p_{11}+p_{00})]. \tag{21.25}$$

21.5.2 Success runs of length at least k

As Koutras (1996b) indicated, the probability generating function of T_r is of the form (21.18) with

$$A(z) = \frac{p_{10}z}{1-p_{11}z} \, \frac{(p_{01}z)(p_{11}z)^{k-1}}{Q(z)},$$

where $H(z)$ and $Q(z)$ are as defined in (21.19) and (21.21), respectively. The double generating function $G(z,w)$ of X_n, the number of success runs of length at least k in a sequence of n Markov dependent trials, can be directly determined via Theorem 21.4.2 and can be easily shown to be consistent with Hirano and Aki's (1993) conditional generating function formulae. However, due to a different set-up used there, some slight discrepancies are observed which can be instantly adjusted by a proper shift of the numbering and modification of the initial distribution.

For the generating function of the means $\mu_n = E[X_n]$, we substitute $H(z)$ and $A(z)$ into Eq. (21.13) and use (21.24) to obtain the final expression

$$\sum_{n=0}^{\infty} \mu_n z^n = \frac{[p_1 + (p_0p_{01} - p_1p_{00})z](1-p_{11}z)(p_{11}z)^{k-1}}{(1-z)^2[1+(1-p_{00}-p_{11})z]} \,. \tag{21.26}$$

Proceeding as we did in **(a)**, we may easily verify that μ_n satisfies the recurrence relation

$$\mu_n = (\alpha+2)\mu_{n-1} - (2\alpha+1)\mu_{n-2} + \alpha\mu_{n-3}, \qquad n \geq k+2,$$

where $\alpha = p_{00} + p_{11} - 1$. The initial conditions required to conduct numerical computations through this recurrence relation are given again by (21.23) and (21.25).

Eq. (21.26) may also be used for the development of non-recursive expressions for μ_n. More specifically, using the expansion

$$(1-z)^{-2} = \sum_{j=0}^{\infty}(j+1)z^j$$

and the geometric series for $(1-\alpha z)^{-1}$, it is possible to write

$$\sum_{n=0}^{\infty} \mu_n z^n = [p_1 + (p_0 p_{01} - p_1 p_{00})z](1 - p_{11}z)(p_{11}z)^{k-1} \sum_{n=0}^{\infty} c_n z^n,$$

where

$$c_n = \sum_{j=0}^{n}(n-j+1)\alpha^j = \frac{(n+1) - (n+2)\alpha + \alpha^{n+2}}{(1-\alpha)^2}. \tag{21.27}$$

It now follows immediately that

$$\mu_n = \{p_1 c_{n-k+1} + [p_0 p_{01} - p_1(p_{00} + p_{11})]c_{n-k} - (p_0 p_{01} - p_1 p_{00})p_{11} c_{n-k-1}\} p_{11}^{k-1}$$

for all $n \geq k+2$; this expression holds true for $n \leq k+1$ as well, if we set $c_n = 0$ for all $n < 0$.

In the special case of i.i.d. Bernoulli trials ($p_0 = 1$, $p_1 = 0$, $p_{01} = p_{11} = p$, $p_{10} = p_{00} = q$, $p + q = 1$), we have $\alpha = 0$ and $c_n = n+1$, and in this case our formula reduces to

$$\mu_n = p^k[(n-k)q + 1]$$

which coincides with the one mentioned in Goldstein (1990) and Hirano and Aki (1993).

The second order moments of X_n can be investigated by making use of expressions (21.14)–(21.16). It is perhaps easier though to employ (21.17) instead, which in our case becomes

$$\begin{aligned}
&\{1 - (\alpha+1)z + \alpha z^2\} \sum_{n=0}^{\infty} E[X_n(X_n+1)]z^n \\
&= \; 2[1 - (\alpha+1)z + \alpha z^2 + p_{01}p_{10}p_{11}^{k-1}z^{k+1}] \sum_{n=0}^{\infty} \mu_n z^n
\end{aligned}$$

and produce the following recurrence relation for the second ascending factorial moments $\pi_n = E[X_n(X_n+1)]$

$$\begin{aligned}
&\pi_n - (\alpha+1)\pi_{n-1} + \alpha\pi_{n-2} \\
&= \quad 2\mu_n - 2(\alpha+1)\mu_{n-1} + 2\alpha\mu_{n-2} + 2p_{01}p_{10}p_{11}^{k-1}\mu_{n-k-1}, \qquad n \geq k+1.
\end{aligned}$$

From this, the forgone analysis for μ_n, and the initial conditions $\pi_n = 2\mu_n$, $0 \leq n \leq k$, the evaluation of all π_n's [and therefore $E[X_n^2]$ and $\mathrm{Var}(X_n)$] is easily achieved.

21.5.3 Overlapping success runs

Starting from the probability generating function of T_r given by

$$H_r(z) = H(z)\left\{p_{11}z + p_{10}z\,\frac{(p_{01}z)(p_{11}z)^{k-1}}{Q(z)}\right\}^{r-1}, \qquad r \geq 1$$

[see Koutras (1996b)], we can readily derive the double generating function $G(z,w)$ of X_n and its moment generating function as

$$G(z,w) = \frac{1}{1-z}\left[1 - \frac{(1-w)P(z)(p_{11}z)^{k-1}}{(1-wp_{11}z)Q(z) - wp_{10}p_{01}p_{11}^{k-1}z^{k+1}}\right],$$

$$\sum_{n=0}^{\infty} E[X_n]z^n = \frac{P(z)(p_{11}z)^{k-1}}{(1-z)^2(1-\alpha z)}$$

where $P(z)$, $Q(z)$ and α are as defined earlier. The first formula could also be established by combining the two conditional double generating functions given by Hirano and Aki (1993) (after some modifications to adjust for the different set-up used therein). The second formula immediately yields the following expression for the mean $\mu_n = E[X_n]$:

$$\mu_n = \{p_1c_{n-k+1} + (p_0p_{01} - p_1p_{00})c_{n-k}\}p_{11}^{k-1}$$

[c_n's are as given in (21.27)].

21.5.4 Number of non-overlapping windows of length at most k containing exactly 2 successes

This random variable, say X_n, is a special case ($s = 2$) of the so-called (discrete) s-scan statistic which has recently received considerable attention, due to its applications in DNA sequencing, quality control, reliability, queueing theory etc [see, Chao, Fu and Koutras (1995), Dembo and Karlin (1992), Glaz (1989), Glaz and Naus (1983), Koutras and Alexandrou (1995)]. Even in the simple i.i.d. case with $s = 2$, the derivation of generating functions for X_n is quite involved.

But, as Koutras (1996a) indicated, the respective waiting time problem for the first and r-th appearance of the event we are interested in (i.e., the occurrence of two successes separated by at most $k-2$ failures), can be easily analyzed by elementary techniques in both i.i.d. and Markov dependent cases with initial distribution

$$\Pr[Z_0 = 0] = 1, \quad \Pr[Z_0 = 1] = 0.$$

One can readily adjust the methodology put forward there to the slightly more general Markov dependent model considered in this section and prove that the probability generating function of T_r takes on the form

$$H_r(z) = P(z)(p_{01}z)^{r-1}B^r(z),$$

where

$$B(z) = \frac{p_{11}z + p_{01}p_{10}z^2 \sum_{i=3}^{k}(p_{00}z)^{i-3}}{1 - p_{00}z - p_{01}p_{10}p_{00}^{k-2}z^k}\ .$$

Clearly, $H_r(z)$ satisfies (21.10) with

$$H(z) = P(z)B(z), \quad A(z) = p_{01}zB(z)$$

and we are thus in possession of a machinery to analyze the respective enumerating random variable X_n. So, the double generating function $G(z, w)$ becomes, by virtue of Theorem 21.4.2,

$$G(z,w) = \frac{1}{1-z}\left[1 - \frac{(1-w)P(z)B(z)}{1 - p_{01}wzB(z)}\right]$$

whereas the generating function of $\mu_n = E[X_n]$ is readily computed to be

$$\sum_{n=0}^{\infty} \mu_n z^n = \frac{zP(z)}{(1-z)^2}\,\frac{p_{11} - \alpha z - p_{01}p_{10}p_{00}^{k-2}z^{k-1}}{1 + (p_{01} - p_{00})z - \alpha p_{01}z^2 - p_{01}p_{10}p_{00}^{k-2}z^k}$$

with $P(z)$ and α as defined before. The above expressions for $k = 2$ are in concordance with the respective formulae of the non-overlapping success runs of length 2. Moreover, the last generating function may be used for developing simple recurrence relations for the mean μ_n and second order moments of X_n [employing (21.14)–(21.17)]. The details are left to the reader.

21.6 Conclusions

The general formulae developed in Sections 21.3 and 21.4 offer very powerful tools for the study of one kind of problem, viz. distribution of the number of events in a fixed number of trials, when the exact solution for the corresponding waiting time problems is known. This was clearly illustrated in Section 21.4 wherein four success runs related problems were treated. It should be mentioned that the method presented here is applicable to more general problems and setups as well; the consideration of the special case of two state trials and first-order Markov dependence was for the convenience of presentation only.

We now proceed to review a few typical generalizations and variations which deserve special attention for further investigation.

First of all the assumption of two different outcomes in each trial might be relaxed by considering multistate trials and general patterns instead of runs. Some work in this direction has already been done by Fu (1996) and Schwager (1983). The elegant generating function approach taken by Aki (1992) and the results of Chryssaphinou and Papastavridis (1990), Guibas and Odlyzko (1981) and Solovev (1966) pertaining to waiting time problems, offer instrumental tools for advancing to the study of the number of occurrences analogues, by the technique introduced here. The special case of sooner waiting time problems, discussed earlier by Aki and Hirano (1993), Balasubramanian, Viveros and Balakrishnan (1993) and Koutras and Alexandrou (1997), is also of great importance, leading by our approach to results for the *total number of success and failure runs.*

Another possible direction for extension of the basic models is offered by considering higher-order Markov dependent trials, in the lines put forward recently by Aki, Balakrishnan and Mohanty (1996).

A final variation of some interest arises by placing the outcomes of the trials in a circular (instead of linear arrangement). For a few results in the special case of circular success runs, we refer to the recent works by Koutras, Papadopoulos and Papastavridis (1994, 1995) and the references therein.

Acknowledgements. This research was partially supported by the University of Athens Research Foundation under grant 70/4/2551.

References

1. Aki, S. (1992). Waiting time problems for a sequence of discrete random variables, *Annals of the Institute of Statistical Mathematics*, **44**, 363–378.

2. Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two-state Markov chain, In *Statistical Science and Data Analysis* (Eds., K. Matusita, M. L. Puri and T. Hayakawa), pp. 467–474, Amsterdam: VSP International Science Publishers.

3. Aki, S., Balakrishnan, N. and Mohanty, S. G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **48**, 773–787 .

4. Balakrishnan, N. and Koutras, M. V. (1997). *Runs and Patterns with Applications*, (under preparation).

5. Balasubramanian, K., Viveros, R. and Balakrishnan, N. (1993). Sooner and later waiting time problems for Markovian Bernoulli trials, *Statistics& Probability Letters*, **18**, 153–161.

6. Banjevic, D. (1988). On some statistics connected with runs in Markov chains, *Journal of Applied Probability*, **25**, 815–821.

7. Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation*, New York: Oxford University Press.

8. Benevento, R. (1984). The occurrence of sequence patterns in ergodic Markov chains, *Stochastic Processes and their Applications*, **17**, 369–373.

9. Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained? *Journal of Applied Probability*, **19**, 518–531.

10. Chao, M. T., Fu, J. C. and Koutras, M. V. (1995). Survey of the reliability studies of consecutive-k-out-of-n: F and related systems, *IEEE Transactions on Reliability*, **44**, 120–127.

11. Charalambides, Ch. A. (1994). Success runs in a circular sequence of independent Bernoulli trials, In *Runs and Patterns in Probability* (Eds., A. P. Godbole and S. G. Papastavridis), pp. 15–30, Amsterdam: Kluwer Academic Publishers.

12. Chen, J. and Glaz, J. (1997). Approximations and inequalities for the distribution of scan statistic for 0–1 Bernoulli trials, In *Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz* (Eds., N. L. Johnson and N. Balakrishnan), New York: John Wiley & Sons (to appear).

13. Chryssaphinou, O. and Papastavridis, S. (1988). A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials, *Journal of Applied Probability*, **25**, 428–431.

14. Chryssaphinou, O. and Papastavridis, S. (1990). The occurrence of sequence patterns in repeated dependent experiments, *Theory of Probability and its Applications*, **35**, 145–152.

15. Dembo, A. and Karlin, S. (1992). Poisson approximations for r-scan processes, *Annals of Applied Probability*, **2**, 329–357.

16. Feller, W. (1968). *An Introduction to Probability Theory and its Applications*, Volume I, Third edition, New York: John Wiley & Sons.

17. Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials, *Statistica Sinica*, **6** (to appear).

18. Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: a Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050–1058.

19. Gerber, H. U. and Li, S. R. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain, *Stochastic Processes and their Applications*, **11**, 101–108.

20. Glaz, J. (1983). Moving window detection for discrete data, *IEEE Transactions on Information Theory*, **29**, 457–462.

21. Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic, *Journal of the American Statistical Association*, **84**, 560–566.

22. Glaz, J. and Naus, J. (1983). Multiple clusters on the line, *Communications in Statistics—Theory and Methods*, **12**, 1961–1986.

23. Godbole, A. P. (1991). Poisson approximations for runs and patterns of rare events, *Advances in Applied Probability*, **23**, 851–865.

24. Godbole, A. P. (1994). Applications of the Stein-Chen method to the theory of patterns and runs: An annotated bibliography, In *Runs and Patterns in Probability* (Eds., A. P. Godbole and S. G. Papastavridis), pp. 253–261, Amsterdam: Kluwer Academic Publishers.

25. Godbole, A. P. and Schaffner, A. A. (1993). Improved Poisson approximations for word patterns, *Advances in Applied Probability*, **25**, 334–347.

26. Goldstein, L. (1990). Poisson approximation in DNA sequence matching, *Communications in Statistics—Theory and Methods*, **19**, 4167–4179.

27. Guibas, L. J. and Odlyzko, A. M. (1980). Long repetitive patterns in random sequences, *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete*, **53**, 241–262.

28. Guibas, L. J. and Odlyzko, A. M. (1981). String overlaps, patterns matching and nontransitive games, *Journal of Combinatorial Theory*, **30**, 183–208.

29. Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313–320.

30. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, New York: John Wiley & Sons.

31. Koutras, M. V. (1996a). On a waiting time distribution in a sequence of Bernoulli trials, *Annals of the Institute of Statistical Mathematics*, **48** (to appear).

32. Koutras, M. V. (1996b). Waiting time distributions associated with runs of fixed length in two-state Markov Chains, *Annals of the Institute of Statistical Mathematics*, **48** (to appear).

33. Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: A unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743–766.

34. Koutras, M. V. and Alexandrou, V. A. (1997). Sooner waiting time problems in a sequence of trinomial trials, *Journal of Applied Probability*, **33** (to appear).

35. Koutras, M. V., Papadopoulos, G. K. and Papastavridis, S. G. (1994). Circular overlapping success runs, In *Runs and Patterns in Probability* (Eds., A. P. Godbole and S. G. Papastavridis), pp. 287–305, Amsterdam: Kluwer Academic Publishers.

36. Koutras, M. V., Papadopoulos, G. K. and Papastavridis, S. G. (1995). Runs on a circle, *Journal of Applied Probability*, **32**, 396–404.

37. Ling, K. D. (1988). On binomial distributions of order k, *Statistics & Probability Letters*, **6**, 247–250.

38. Mohanty, S. G. (1994). Success runs of length-k in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **46**, 777–796.

39. Rajarshi, M. B. (1974). Success runs in a two-state Markov chain, *Journal of Applied Probability*, **11**, 190–194.

40. Schwager, S. (1983). Run probabilities in sequences of Markov dependent trials, *Journal of the American Statistical Association*, **78**, 168–175.

41. Solovev, A. D. (1966). A combinatorial identity and its application to the problems concerning the first occurrence of a rare event, *Theory of Probability and its Applications*, **11**, 276–282.

42. Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **47**, 415–433.

43. Viveros, R., Balasubramanian, K. and Balakrishnan, N. (1994). Binomial and negative binomial analogues under correlated Bernoulli trials, *The American Statistician*, **48**, 243–247.

22

On Sooner and Later Problems Between Success and Failure Runs

Sigeo Aki
Osaka University, Osaka, Japan

Abstract: Let $X_0, X_1, X_2, \ldots$ be a sequence of $\{0,1\}$-valued Markov chain. Let E_1 denote a run of "1" of length k and E_0 denote a run of "0" of length r. We observe which run comes sooner or later in the sequence $X_0, X_1, X_2, \ldots$. The exact distributions of the numbers of overlapping sooner runs and non-overlapping sooner runs until the later run occurs (for the first time) are derived. Let F_1 be a success run of length k or more and let F_0 be a failure run of length r or more. The exact distribution of the number of occurrences of the sooner event until the first occurrence of the later event between F_1 and F_0 is also studied. Further, when X's have more than two values, more general problems are discussed and the exact joint distribution of the numbers of occurrences of the first, the second, ..., and the j-th runs until the j-th run occurs (for the first time) is obtained in the case of independent trials.

Keywords and phrases: Probability generating function, geometric distribution, discrete distributions, Markov chain, waiting time, geometric distribution of order k

22.1 Introduction

Let $X_0, X_1, X_2, \ldots$ be a sequence of $\{0,1\}$-valued Markov chain with the following probabilities: for $i = 1, 2, \ldots$,

$$p_0 = \Pr[X_0 = 0],\ p_1 = \Pr[X_0 = 1]\ (= 1 - p_0) \text{ and}$$

$$p_{xy} = \Pr[X_i = y | X_{i-1} = x] \text{ for } x, y = 0, 1.$$

We assume that $0 < p_{xy} < 1$. We call X_n the n-th trial and we say success and failure for the outcomes "1" and "0", respectively. Let k and r be given positive integers greater than 1. We denote by E_1 a success run of length k and by E_0 a failure run of length r. We are interested in two kinds of runs. This problem was investigated in a paper by Ebneshahrashoob and Sobel (1990) wherein the distributions of the numbers of trials until the sooner and later events firstly occur were derived when $X_1, X_2 \ldots$ are independent identically distributed random variables. The waiting time problems have been generalized for dependent trials and some results were obtained by Aki (1992), Balasubramanian, Viveros and Balakrishnan (1993), Aki and Hirano (1993), Mohanty (1994) and Aki, Balakrishnan and Mohanty (1996).

In this paper we investigate the distribution of the number of occurrences of the sooner event until the later event occurs for the first time.

We now give a practical example for the problem.

Example 22.1.1 Suppose we have two machines (e.g., cars), say A and B. Only one of the machines is randomly selected every day and used (by customers). The machines A and B are supposed to be possibly damaged enough for us to admit preventive maintenance if they are used consecutively k and r days, respectively. Since it is convenient for us to make preventive maintenance on both machines simultaneously, our basic strategy is to wait until the occurrence of the later event between events E_A and E_B, where E_A is the event that the machine A is used k days consecutively and E_B is the event that the machine B is used r days consecutively. However, if the sooner event occurs many times until the later event occurs for the first time, the sooner machine may be completely damaged by then and have to be replaced at a great cost. Then, it is important to know the distribution of the number of occurrences of the sooner event until the first occurrence of the later event in this example.

It is well known that there are different ways of counting the numbers of runs such as overlapping counting, non-overlapping counting, etc. [*cf.* Hirano and Aki (1993) and Fu and Koutras (1994)].

This paper is organized as follows: In Section 22.2 we derive the probability generating functions (p.g.f.'s) of the distributions of numbers of overlapping and non-overlapping occurrences of the sooner run between E_1 and E_0 until the later run occurs for the first time in $X_0, X_1, X_2, \ldots$. Let F_1 be a success run of length k or more and let F_0 be a failure run of length r or more. The p.g.f. of the distribution of the number of occurrences of the sooner event until the first occurrence of the later event between F_1 and F_0 is also provided. In Section 22.3 more general problems are discussed when X's have more than two values. The p.g.f. of the joint distribution of the numbers of occurrences of the first, the second, ..., and the j-th runs until the j-th run occurs for the first time is obtained in the case of independent trials.

22.2 Number of Occurrences of the Sooner Event Until the Later Waiting Time

First, we derive the distribution of the number of overlapping occurrences of the sooner run. Let $\Phi(t)$ be the p.g.f. of the number of occurrences of the sooner run until the later run occurs for the first time in $X_0, X_1, X_2, \ldots$. For $x = 0, 1$, let $\phi_x(t)$ be the p.g.f. of the conditional distribution of the number of overlapping occurrences until the later run occurs for the first time in $X_0, X_1, X_2, \ldots$ given that $X_0 = x$. Let i be an integer such that $1 \le i \le k-1$ and let j be an integer such that $1 \le j \le r-1$. Suppose we have currently "1"-run of length i and the sooner event has not occurred yet. Then, we denote by $\phi_{1,i} = \phi_{1,i}(t)$ the p.g.f. of the number of occurrences of the sooner run from this time until the later run occurs for the first time. Suppose we have currently "0"-run of length j and the sooner event has not occurred yet. Then, we denote by $\phi_{0,j} = \phi_{0,j}(t)$ the p.g.f. of the number of occurrences of the sooner run from this time until the later run occurs for the first time. For $x = 1$ or 0, we consider the following cases. Suppose that we have currently "1"-run of length i, that the sooner event is E_x, and that E_x has already occurred. Then, we denote by $\phi_{1,i}^{(x)} = \phi_{1,i}^{(x)}(t)$ the p.g.f. of the conditional distribution of the number of overlapping occurrences of the sooner run from this time until the later run occurs for the first time. Suppose that we have currently "0"-run of length j, that the sooner event is E_x, and that E_x has already occurred. Then, we denote by $\phi_{0,j}^{(x)} = \phi_{0,j}^{(x)}(t)$ the p.g.f. of the conditional distribution of the number of overlapping occurrences of the sooner run from this time until the later run occurs for the first time.

From the definition, we see that $\phi_1(t) = \phi_{1,1}(t)$ and $\phi_0(t) = \phi_{0,1}(t)$. From the definitions of $\Phi(t)$, $\phi_{1,i}(t)$, $\phi_{0,j}(t)$, $\phi_{1,i}^{(x)}$ and $\phi_{0,j}^{(x)}(t)$, we have the following system of equations:

$$\Phi(t) = p_1\phi_{1,1} + p_0\phi_{0,1}, \tag{22.1}$$

$$\left\{\begin{array}{lcl} \phi_{1,1} & = & p_{11}\phi_{1,2} + p_{10}\phi_{0,1} \\ \phi_{1,2} & = & p_{11}\phi_{1,3} + p_{10}\phi_{0,1} \\ & & \cdots \\ \phi_{1,k-1} & = & p_{11}t\phi_{1,k-1}^{(1)} + p_{10}\phi_{0,1}, \end{array}\right. \tag{22.2}$$

$$\left\{\begin{array}{lcl} \phi_{0,1} & = & p_{01}\phi_{1,1} + p_{00}\phi_{0,2} \\ \phi_{0,2} & = & p_{01}\phi_{1,1} + p_{00}\phi_{0,3} \\ & & \cdots \\ \phi_{0,r-1} & = & p_{01}\phi_{1,1} + p_{00}t\phi_{0,r-1}^{(0)}, \end{array}\right. \tag{22.3}$$

$$\phi_{1,k-1}^{(1)} = p_{11}t\phi_{1,k-1}^{(1)} + p_{10}\phi_{0,1}^{(1)}, \tag{22.4}$$

$$\begin{cases} \phi_{0,1}^{(1)} &= p_{01}\phi_{1,1}^{(1)} + p_{00}\phi_{0,2}^{(1)} \\ \phi_{0,2}^{(1)} &= p_{01}\phi_{1,1}^{(1)} + p_{00}\phi_{0,3}^{(1)} \\ & \cdots \\ \phi_{0,r-1}^{(1)} &= p_{01}\phi_{1,1}^{(1)} + p_{00}, \end{cases} \tag{22.5}$$

$$\phi_{0,r-1}^{(0)} = p_{01}\phi_{1,1}^{(0)} + p_{00}t\phi_{0,r-1}^{(0)}, \tag{22.6}$$

$$\begin{cases} \phi_{1,1}^{(0)} &= p_{11}\phi_{1,2}^{(0)} + p_{10}\phi_{0,1}^{(0)} \\ \phi_{1,2}^{(0)} &= p_{11}\phi_{1,3}^{(0)} + p_{10}\phi_{0,1}^{(0)} \\ & \cdots \\ \phi_{1,k-1}^{(0)} &= p_{11} + p_{10}\phi_{0,1}^{(0)}, \end{cases} \tag{22.7}$$

$$\begin{cases} \phi_{1,1}^{(1)} &= p_{11}\phi_{1,2}^{(1)} + p_{10}\phi_{0,1}^{(1)} \\ \phi_{1,2}^{(1)} &= p_{11}\phi_{1,3}^{(1)} + p_{10}\phi_{0,1}^{(1)} \\ & \cdots \\ \phi_{1,k-1}^{(1)} &= p_{11}t\phi_{1,k-1}^{(1)} + p_{10}\phi_{0,1}^{(1)}, \end{cases} \tag{22.8}$$

and

$$\begin{cases} \phi_{0,1}^{(0)} &= p_{01}\phi_{1,1}^{(0)} + p_{00}\phi_{0,2}^{(0)} \\ \phi_{0,2}^{(0)} &= p_{01}\phi_{1,1}^{(0)} + p_{00}\phi_{0,3}^{(0)} \\ & \cdots \\ \phi_{0,r-1}^{(0)} &= p_{01}\phi_{1,1}^{(0)} + p_{00}t\phi_{0,r-1}^{(0)}. \end{cases} \tag{22.9}$$

From (22.8) and (22.9) we obtain

$$\phi_{1,1}^{(1)} = \phi_{0,1}^{(1)}(1 - p_{11}^{k-1}) + p_{11}^{k-1}t\phi_{1,k-1}^{(1)} \tag{22.10}$$

and

$$\phi_{0,1}^{(0)} = \phi_{1,1}^{(0)}(1 - p_{00}^{r-1}) + p_{00}^{r-1}t\phi_{0,r-1}^{(0)}. \tag{22.11}$$

From (22.4) and (22.6) we obtain

$$\phi_{1,k-1}^{(1)} = \frac{p_{10}}{1 - p_{11}t}\,\phi_{0,1}^{(1)}$$

and

$$\phi_{0,r-1}^{(0)} = \frac{p_{01}}{1 - p_{00}t}\,\phi_{1,1}^{(0)},$$

respectively. By substituting these expressions into (22.10) and (22.11), we get

$$\phi_{1,1}^{(1)} = \phi_{0,1}^{(1)}(1 - p_{11}^{k-1}) + \frac{p_{11}^{k-1}p_{10}t}{1 - p_{11}t}\,\phi_{0,1}^{(1)} \tag{22.12}$$

and

$$\phi_{0,1}^{(0)} = \phi_{1,1}^{(0)}(1 - p_{00}^{r-1}) + \frac{p_{00}^{r-1}p_{01}t}{1 - p_{00}t}\,\phi_{1,1}^{(0)}\ . \tag{22.13}$$

From (22.5) and (22.12) we obtain

$$\phi_{1,1}^{(1)} = \frac{p_{00}^{r-1}\left(1 - p_{11}^{k-1} + \frac{p_{11}^{k-1} p_{10} t}{1 - p_{11} t}\right)}{1 - (1 - p_{11}^{k-1})(1 - p_{00}^{r-1}) - \frac{p_{11}^{k-1} p_{10} t}{1 - p_{11} t}(1 - p_{00}^{r-1})} \, . \tag{22.14}$$

From (22.7) and (22.13), we similarly obtain

$$\phi_{0,1}^{(0)} = \frac{p_{11}^{k-1}\left(1 - p_{00}^{r-1} + \frac{p_{00}^{r-1} p_{01} t}{1 - p_{00} t}\right)}{1 - (1 - p_{11}^{k-1})(1 - p_{00}^{r-1}) - \frac{p_{00}^{r-1} p_{01} t}{1 - p_{00} t}(1 - p_{11}^{k-1})} \, . \tag{22.15}$$

Eqs. (22.4) and (22.5) imply

$$\phi_{1,k-1}^{(1)} = \frac{(1 - p_{00}^{r-1}) p_{10}}{1 - p_{11} t} \, \phi_{1,1}^{(1)} + \frac{p_{10} p_{00}^{r-1}}{1 - p_{11} t} \, . \tag{22.16}$$

Eqs. (22.6) and (22.7) similarly imply

$$\phi_{0,r-1}^{(0)} = \frac{(1 - p_{11}^{k-1}) p_{01}}{1 - p_{00} t} \, \phi_{0,1}^{(0)} + \frac{p_{01} p_{11}^{k-1}}{1 - p_{00} t} \, . \tag{22.17}$$

From (22.2) and (22.3), we now obtain

$$\phi_{0,1} = \frac{p_{11}^{k-1}(1 - p_{00}^{r-1}) t \phi_{1,k-1}^{(1)} + p_{00}^{r-1} t \phi_{0,r-1}^{(0)}}{1 - (1 - p_{11}^{k-1})(1 - p_{00}^{r-1})} \, , \tag{22.18}$$

and

$$\phi_{1,1} = \frac{p_{00}^{r-1}(1 - p_{11}^{k-1}) t \phi_{0,r-1}^{(0)} + p_{11}^{k-1} t \phi_{1,k-1}^{(1)}}{1 - (1 - p_{11}^{k-1})(1 - p_{00}^{r-1})} \, . \tag{22.19}$$

By substituting (22.14) into (22.16), we get

$$\begin{aligned} &\phi_{1,k-1}^{(1)} \\ &= \frac{p_{10} p_{00}^{r-1}(1 - p_{00}^{r-1})\left(1 - p_{11}^{k-1} + \frac{p_{11}^{k-1} p_{10} t}{1 - p_{11} t}\right)}{(1 - p_{11} t) - (1 - p_{11} t)(1 - p_{11}^{k-1})(1 - p_{00}^{r-1}) - (1 - p_{00}^{r-1}) p_{11}^{k-1} p_{10} t} \\ &\quad + \frac{p_{10} p_{00}^{r-1}}{1 - p_{11} t} \, . \end{aligned}$$

By substituting (22.15) into (22.17), we similarly get

$$\begin{aligned} &\phi_{0,r-1}^{(0)} \\ &= \frac{p_{01} p_{11}^{k-1}(1 - p_{11}^{k-1})\left(1 - p_{00}^{r-1} + \frac{p_{00}^{r-1} p_{01} t}{1 - p_{00} t}\right)}{(1 - p_{00} t) - (1 - p_{00} t)(1 - p_{11}^{k-1})(1 - p_{00}^{r-1}) - (1 - p_{11}^{k-1}) p_{00}^{r-1} p_{01} t} \\ &\quad + \frac{p_{01} p_{11}^{k-1}}{1 - p_{00} t} \, . \end{aligned}$$

Making use of these expressions in (22.18) and (22.19), we obtain the following result.

Theorem 22.2.1 *The p.g.f. of the distribution of the number of overlapping occurrences of the sooner run until the first occurrence of the later run in* $X_0, X_1, X_2, \ldots$ *is given by*

$$\Phi(t) = p_1\phi_{1,1} + p_0\phi_{0,1},$$

where

$$\begin{aligned}\phi_{0,1} &= \frac{1}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}\\ &\times \left\{ \frac{p_{11}^{k-1}p_{10}p_{00}^{r-1}(1-p_{00}^{r-1})^2 t(1-p_{11}^{k-1}+\frac{p_{11}^{k-1}p_{10}t}{1-p_{11}t})}{(1-p_{11}t)-(1-p_{11}t)(1-p_{11}^{k-1})(1-p_{00}^{r-1})-(1-p_{00}^{r-1})p_{11}^{k-1}p_{10}t} \right.\\ &\qquad + \frac{p_{11}^{k-1}p_{10}p_{00}^{r-1}(1-p_{00}^{r-1})t}{1-p_{11}t}\\ &+ \frac{p_{00}^{r-1}p_{01}p_{11}^{k-1}(1-p_{11}^{k-1})t(1-p_{00}^{r-1}+\frac{p_{00}^{r-1}p_{01}t}{1-p_{00}t})}{(1-p_{00}t)-(1-p_{00}t)(1-p_{11}^{k-1})(1-p_{00}^{r-1})-(1-p_{11}^{k-1})p_{00}^{r-1}p_{01}t}\\ &\qquad \left. + \frac{p_{00}^{r-1}p_{01}p_{11}^{k-1}t}{1-p_{00}t} \right\},\end{aligned}$$

and

$$\begin{aligned}\phi_{1,1} &= \frac{1}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}\\ &\times \left\{ \frac{p_{11}^{k-1}p_{10}p_{00}^{r-1}(1-p_{00}^{r-1})t(1-p_{11}^{k-1}+\frac{p_{11}^{k-1}p_{10}t}{1-p_{11}t})}{(1-p_{11}t)-(1-p_{11}t)(1-p_{11}^{k-1})(1-p_{00}^{r-1})-(1-p_{00}^{r-1})p_{11}^{k-1}p_{10}t} \right.\\ &\qquad + \frac{p_{11}^{k-1}p_{10}p_{00}^{r-1}t}{1-p_{11}t}\\ &+ \frac{p_{00}^{r-1}p_{01}p_{11}^{k-1}(1-p_{11}^{k-1})^2 t(1-p_{00}^{r-1}+\frac{p_{00}^{r-1}p_{01}t}{1-p_{00}t})}{(1-p_{00}t)-(1-p_{00}t)(1-p_{11}^{k-1})(1-p_{00}^{r-1})-(1-p_{11}^{k-1})p_{00}^{r-1}p_{01}t}\\ &\qquad \left. + \frac{p_{00}^{r-1}p_{01}p_{11}^{k-1}(1-p_{11}^{k-1})t}{1-p_{00}t} \right\}.\end{aligned}$$

Let $\overline{\Phi}(t)$ be the p.g.f. of the distribution of the number of overlapping occurrences of the sooner event until the first occurrence of the later run in $X_1, X_2, \ldots$. For $x = 0, 1$, let $\overline{\Phi}_x(t)$ be the p.g.f. of the conditional distribution of the number of overlapping occurrences of the sooner run until the later run occurs for the first time in $X_1, X_2, \ldots$ given that $X_0 = x$. Then, it is easy to see that

$$\begin{aligned}\overline{\Phi} &= p_1\overline{\Phi}_1 + p_0\overline{\Phi}_0,\\ \overline{\Phi}_1 &= p_{11}\phi_{1,1} + p_{10}\phi_{0,1},\end{aligned}$$

and

$$\overline{\Phi}_0 = p_{01}\phi_{1,1} + p_{00}\phi_{0,1}.$$

Corollary 22.2.1 *The p.g.f. of the distribution of the number of overlapping occurrences of the sooner run until the first occurrence of the later run in* $X_1, X_2, \ldots$ *is given by*

$$\overline{\Phi} = (p_1p_{11} + p_0p_{01})\phi_{1,1} + (p_1p_{10} + p_0p_{00})\phi_{0,1},$$

where $\phi_{1,1}$ *and* $\phi_{0,1}$ *are as given in Theorem 22.2.1.*

Next, we investigate the distribution of the number of non-overlapping occurrences of the sooner run.

Let $\Psi(t)$ be the p.g.f. of the number of non-overlapping occurrences of the sooner run until the later run occurs for the first time in $X_0, X_1, X_2, \ldots$. For $x = 0, 1$, let $\psi_x(t)$ be the p.g.f. of the conditional distribution of the number of non-overlapping occurrences of the sooner run until the later run occurs for the first time in $X_0, X_1, \ldots$ given that $X_0 = x$.

As done in the case of overlapping counting, we define the p.g.f.'s of the conditional distribution of the number of non-overlapping occurrences of the sooner run, $\psi_{1,i}(t)$, $\psi_{0,j}(t)$, $\psi_{1,i}^{(x)}$, $\psi_{0,j}^{(x)}$ for $i = 1, \ldots, k-1$; $j = 1, \ldots, r-1$, and $x = 0, 1$.

Further, we define $\psi^{(x)}$ for $x = 0, 1$ as follows: Suppose the sooner event is E_x and E_x has just occurred. Then we denote by $\psi^{(x)} = \psi^{(x)}(t)$ the p.g.f. of the conditional distribution of the number of non-overlapping occurrences from this time until the later run occurs for the first time.

In this case, we have the following system of equations:

$$\Psi(t) = p_1\psi_{1,1} + p_0\psi_{0,1}, \tag{22.20}$$

$$\begin{cases} \psi_{1,1} &= p_{11}\psi_{1,2} + p_{10}\psi_{0,1} \\ \psi_{1,2} &= p_{11}\psi_{1,3} + p_{10}\psi_{0,1} \\ & \cdots \\ \psi_{1,k-1} &= p_{11}t\psi^{(1)} + p_{10}\psi_{0,1}, \end{cases} \tag{22.21}$$

$$\begin{cases} \psi_{0,1} &= p_{01}\psi_{1,1} + p_{00}\psi_{0,2} \\ \psi_{0,2} &= p_{01}\psi_{1,1} + p_{00}\psi_{0,3} \\ & \cdots \\ \psi_{0,r-1} &= p_{01}\psi_{1,1} + p_{00}t\psi^{(0)}, \end{cases} \tag{22.22}$$

$$\psi^{(1)} = p_{11}\psi_{1,1}^{(1)} + p_{10}\psi_{0,1}^{(1)}, \tag{22.23}$$

$$\psi^{(0)} = p_{01}\psi_{1,1}^{(0)} + p_{00}\psi_{0,1}^{(0)}, \tag{22.24}$$

$$\left\{\begin{array}{lcl} \psi_{0,1}^{(1)} & = & p_{01}\psi_{1,1}^{(1)} + p_{00}\psi_{0,2}^{(1)} \\ \psi_{0,2}^{(1)} & = & p_{01}\psi_{1,1}^{(1)} + p_{00}\psi_{0,3}^{(1)} \\ & & \cdots \\ \psi_{0,r-1}^{(1)} & = & p_{01}\psi_{1,1}^{(1)} + p_{00}, \end{array}\right. \tag{22.25}$$

$$\left\{\begin{array}{lcl} \psi_{1,1}^{(0)} & = & p_{11}\psi_{1,2}^{(0)} + p_{10}\psi_{0,1}^{(0)} \\ \psi_{1,2}^{(0)} & = & p_{11}\psi_{1,3}^{(0)} + p_{10}\psi_{0,1}^{(0)} \\ & & \cdots \\ \psi_{1,k-1}^{(0)} & = & p_{11} + p_{10}\psi_{0,1}^{(0)}, \end{array}\right. \tag{22.26}$$

$$\left\{\begin{array}{lcl} \psi_{1,1}^{(1)} & = & p_{11}\psi_{1,2}^{(1)} + p_{10}\psi_{0,1}^{(1)} \\ \psi_{1,2}^{(1)} & = & p_{11}\psi_{1,3}^{(1)} + p_{10}\psi_{0,1}^{(1)} \\ & & \cdots \\ \psi_{1,k-1}^{(1)} & = & p_{11}t\psi^{(1)} + p_{10}\psi_{0,1}^{(1)}, \end{array}\right. \tag{22.27}$$

and

$$\left\{\begin{array}{lcl} \psi_{0,1}^{(0)} & = & p_{01}\psi_{1,1}^{(0)} + p_{00}\psi_{0,2}^{(0)} \\ \psi_{0,2}^{(0)} & = & p_{01}\psi_{1,1}^{(0)} + p_{00}\psi_{0,3}^{(0)} \\ & & \cdots \\ \psi_{0,r-1}^{(0)} & = & p_{01}\psi_{1,1}^{(0)} + p_{00}t\psi^{(0)}. \end{array}\right. \tag{22.28}$$

From (22.25) and (22.27), we obtain

$$\psi_{1,1}^{(1)} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1} + p_{11}^{k-1}t\psi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} \tag{22.29}$$

and

$$\psi_{0,1}^{(1)} = \frac{(1-p_{00}^{r-1})p_{11}^{k-1}t\psi^{(1)} + p_{00}^{r-1}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}\ . \tag{22.30}$$

Similarly, from (22.26) and (22.28), we obtain

$$\psi_{1,1}^{(0)} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1}t\psi^{(0)} + p_{11}^{k-1}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} \tag{22.31}$$

and

$$\psi_{0,1}^{(0)} = \frac{(1-p_{00}^{r-1})p_{11}^{k-1} + p_{00}^{r-1}t\psi^{(0)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}. \tag{22.32}$$

Eqs. (22.23), (22.29) and (22.30) imply

$$\psi^{(1)} = \frac{p_{00}^{r-1}(1-p_{11}^{k})}{p_{11}^{k-1} + p_{00}^{r-1} - p_{11}^{k-1}p_{00}^{r-1} - p_{11}^{k-1}(1-p_{10}p_{00}^{r-1})t}\ . \tag{22.33}$$

Similarly, Eqs. (22.24), (22.31) and (22.32) imply

$$\psi^{(0)} = \frac{p_{11}^{k-1}(1-p_{00}^{r})}{p_{11}^{k-1}+p_{00}^{r-1}-p_{11}^{k-1}p_{00}^{r-1}-p_{00}^{r-1}(1-p_{01}p_{11}^{k-1})t}\,. \tag{22.34}$$

By solving (22.21) and (22.22), we obtain

$$\psi_{1,1} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1}t\psi^{(0)}+p_{11}^{k-1}t\psi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} \tag{22.35}$$

and

$$\psi_{0,1} = \frac{(1-p_{00}^{r-1})p_{11}^{k-1}t\psi^{(1)}+p_{00}^{r-1}t\psi^{(0)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}\,. \tag{22.36}$$

Making use of the expressions in (22.35) and (22.36) in (22.20), we get the following result.

Theorem 22.2.2 *The p.g.f. of the distribution of the number of non-overlapping occurrences of the sooner run until the first occurrence of the later run in* $X_0, X_1, X_2, \ldots$ *is given by*

$$\Psi = p_1\psi_{1,1} + p_0\psi_{0,1},$$

where $\psi_{1,1}$ *and* $\psi_{0,1}$ *are as given in (22.35) and (22.36), and* $\psi^{(1)}$ *and* $\psi^{(0)}$ *are as given in (22.33) and (22.34), respectively.*

Similar to Corollary 22.2.1, we now have the following result.

Corollary 22.2.2 *The p.g.f. of the distribution of the number of non-overlapping occurrences of the sooner run until the first occurrence of the later run in* $X_1, X_2, \ldots$ *is given by*

$$\overline{\Psi} = (p_1p_{11}+p_0p_{01})\psi_{1,1} + (p_1p_{10}+p_0p_{00})\psi_{0,1}.$$

Remark 22.2.1 When we set $p_1 = p_{11} = p_{01} = p$ and $p_0 = p_{10} = p_{00} = q$, the sequence reduces to independent sequence with success probability p. Then, from Theorem 22.2.2 it is easy to see that

$$\Psi(t) = P_0P_1t\left(\frac{1}{1-P_1t}+\frac{1}{1-P_0t}\right),$$

where

$$P_1 = \frac{p^{k-1}(1-q^r)}{1-(1-p^{k-1})(1-q^{r-1})}$$

and

$$P_0 = \frac{q^{r-1}(1-p^k)}{1-(1-p^{k-1})(1-q^{r-1})}\,.$$

This formula can be proved directly by using the independence of the sequence. Aki and Hirano (1994, p. 196) have noted that the number of occurrences of E_0 until the first occurrence of E_1 follows a geometric distribution with parameter P_0, where

$$P_0 = \Pr[E_0 \text{ comes sooner}] = \frac{q^{r-1}(1-p^k)}{1-(1-p^{k-1})(1-q^{r-1})} .$$

By changing the roles of E_0 and E_1, we note that the number of occurrences of E_1 until the first occurrence of E_0 follows the geometric distribution with parameter P_1, where $P_1 = 1 - P_0$. We denote by ν_0 the number of occurrences of E_0 until the first occurrence of E_1, and by ν_1 the number of occurrences of E_1 until the first occurrence of E_0. Then, we have

$$\begin{aligned}
\Pr[N=n] &= \Pr[N=n, E_0 \text{ comes sooner}] + \Pr[N=n, E_1 \text{ comes sooner}] \\
&= \Pr[E_0 \text{ comes sooner}] \Pr[N=n | E_0 \text{ comes sooner}] \\
&\quad + \Pr[E_1 \text{ comes sooner}] \Pr[N=n | E_1 \text{ comes sooner}] \\
&= \Pr[E_0 \text{ comes sooner}] \Pr[\nu_0 = n | \nu_0 \geq 1] \\
&\quad + \Pr[E_1 \text{ comes sooner}] \Pr[\nu_1 = n | \nu_1 \geq 1] \\
&= \begin{cases} \Pr[\nu_0 = n] + \Pr[\nu_1 = n] & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \end{cases} \\
&= \begin{cases} P_1 P_0^n + P_1^n P_0 & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 . \end{cases}
\end{aligned}$$

Next, we study the distribution corresponding to the occurrences of "1"-run of length k or more and "0"-run of length r or more. It needs to be mentioned that many results have been derived based on overlapping as well as nonoverlapping counting [*cf.* Goldstein (1990), Hirano and Aki (1993), Fu and Koutras (1994) and Uchida and Aki (1995)].

We denote by F_1 a success run of length k or more and by F_0 a failure run of length r or more. The waiting time distributions for the sooner and later problems between F_1 and F_0 were obtained by Uchida and Aki (1995).

Now we study the distribution of the number of occurrences of the sooner event until the first occurrence of the later event.

Let $\Xi(t)$ be the p.g.f. of the number of (non-overlapping) occurrences of the sooner run until the later run occurs for the first time in $X_0, X_1, X_2, \ldots$. Similar to the previous two cases, we define the p.g.f.'s of the number of occurrences of the sooner run, $\xi_{1,j}$, $\xi_{0,j}$, $\xi_{1,i}^{(x)}$, $\xi_{0,j}^{(x)}$ for $i = 1, \ldots, k-1$, $j = 1, \ldots, r-1$, and $x = 0, 1$. Further, we define $\xi^{(x)}$ for $x = 0, 1$ as follows: Suppose the sooner event is F_1 and F_1 has already occurred (though it may not have finished yet) and we have currently "1"-run of length k or more. Then we denote by $\xi^{(1)} = \xi^{(1)}(t)$ the p.g.f. of the conditional distribution of the number of occurrences of F_1

from this time until the later run (F_0) occurs for the first time. Suppose the sooner event is F_0 and F_0 has already occurred and we have currently "0"-run of length r or more. Then, we denote by $\xi^{(0)} = \xi^{(0)}(t)$ the p.g.f. of the conditional distribution of the number of occurrences of F_0 from this time until the later run (F_1) occurs for the first time.

From the definitions of the p.g.f.'s of the conditional distributions, we have the following system of equations in this case:

$$\Xi(t) = p_1\xi_{1,1} + p_0\xi_{0,1} \tag{22.37}$$

$$\begin{cases} \xi_{1,1} &= p_{11}\xi_{1,2} + p_{10}\xi_{0,1} \\ \xi_{1,2} &= p_{11}\xi_{1,3} + p_{10}\xi_{0,1} \\ & \cdots \\ \xi_{1,k-1} &= p_{11}t\xi^{(1)} + p_{10}\xi_{0,1} \end{cases} \tag{22.38}$$

$$\begin{cases} \xi_{0,1} &= p_{01}\xi_{1,1} + p_{00}\xi_{0,2} \\ \xi_{0,2} &= p_{01}\xi_{1,1} + p_{00}\xi_{0,3} \\ & \cdots \\ \xi_{0,r-1} &= p_{01}\xi_{1,1} + p_{00}t\xi^{(0)} \end{cases} \tag{22.39}$$

$$\xi^{(1)} = p_{11}\xi^{(1)} + p_{10}\xi_{0,1}^{(1)} \tag{22.40}$$

$$\xi^{(0)} = p_{01}\xi_{1,1}^{(0)} + p_{00}\xi^{(0)} \tag{22.41}$$

$$\begin{cases} \xi_{0,1}^{(1)} &= p_{01}\xi_{1,1}^{(1)} + p_{00}\xi_{0,2}^{(1)} \\ \xi_{0,2}^{(1)} &= p_{01}\xi_{1,1}^{(1)} + p_{00}\xi_{0,3}^{(1)} \\ & \cdots \\ \xi_{0,r-1}^{(1)} &= p_{01}\xi_{1,1}^{(1)} + p_{00} \end{cases} \tag{22.42}$$

$$\begin{cases} \xi_{1,1}^{(0)} &= p_{11}\xi_{1,2}^{(0)} + p_{10}\xi_{0,1}^{(0)} \\ \xi_{1,2}^{(0)} &= p_{11}\xi_{1,3}^{(0)} + p_{10}\xi_{0,1}^{(0)} \\ & \cdots \\ \xi_{1,r-1}^{(0)} &= p_{11} + p_{10}\xi_{0,1}^{(0)} \end{cases} \tag{22.43}$$

$$\begin{cases} \xi_{1,1}^{(1)} &= p_{11}\xi_{1,2}^{(1)} + p_{10}\xi_{0,1}^{(1)} \\ \xi_{1,2}^{(1)} &= p_{11}\xi_{1,3}^{(1)} + p_{10}\xi_{0,1}^{(1)} \\ & \cdots \\ \xi_{1,r-1}^{(1)} &= p_{11}t\xi^{(1)} + p_{10}\xi_{0,1}^{(1)} \end{cases} \tag{22.44}$$

and

$$\begin{cases} \xi_{0,1}^{(0)} &= p_{01}\xi_{1,1}^{(0)} + p_{00}\xi_{0,2}^{(0)} \\ \xi_{0,2}^{(0)} &= p_{01}\xi_{1,1}^{(0)} + p_{00}\xi_{0,3}^{(0)} \\ & \cdots \\ \xi_{0,r-1}^{(0)} &= p_{01}\xi_{1,1}^{(0)} + p_{00}t\xi^{(0)}. \end{cases} \tag{22.45}$$

From (22.42) and (22.44), we obtain

$$\xi_{1,1}^{(1)} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1} + p_{11}^{k-1}t\xi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} \tag{22.46}$$

and

$$\xi^{(1)} = \xi_{0,1}^{(1)} = \frac{(1-p_{00}^{r-1})p_{11}^{k-1}t\xi^{(1)} + p_{00}^{r-1}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} . \tag{22.47}$$

From (22.43) and (22.45), we similarly obtain

$$\xi^{(0)} = \xi_{1,1}^{(0)} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1}t\xi^{(0)} + p_{11}^{k-1}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} \tag{22.48}$$

and

$$\xi_{0,1}^{(0)} = \frac{(1-p_{00}^{r-1})p_{11}^{k-1} + p_{00}^{r-1}t\xi^{(0)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} . \tag{22.49}$$

Eq. (22.47) implies

$$\xi^{(1)} = \frac{p_{00}^{r-1}}{p_{11}^{k-1} + p_{00}^{r-1} - p_{11}^{k-1}p_{00}^{r-1} - (1-p_{00}^{r-1})p_{11}^{k-1}t} . \tag{22.50}$$

Eq. (22.48) similarly implies

$$\xi^{(0)} = \frac{p_{11}^{k-1}}{p_{11}^{k-1} + p_{00}^{r-1} - p_{11}^{k-1}p_{00}^{r-1} - (1-p_{11}^{k-1})p_{00}^{r-1}t} . \tag{22.51}$$

By solving (22.38) and (22.39), we get

$$\xi_{1,1} = \frac{(1-p_{11}^{k-1})p_{00}^{r-1}t\xi^{(0)} + p_{11}^{k-1}t\xi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})}$$

and

$$\xi_{0,1} = \frac{p_{00}^{r-1}t\xi^{(0)} + p_{11}^{k-1}(1-p_{00}^{r-1})t\xi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} .$$

Making use of these expressions in (22.37), we get the following result.

Theorem 22.2.3 *The p.g.f. of the distribution of the number of occurrences of the sooner run until the later run occurs for the first time between F_1 and F_0 in $X_0, X_1, X_2, \ldots$ is given by*

$$\Xi(t) = \frac{p_{00}^{r-1}t(1-p_1p_{11}^{k-1})\xi^{(0)} + p_{11}^{k-1}t(1-p_0p_{00}^{r-1})\xi^{(1)}}{1-(1-p_{11}^{k-1})(1-p_{00}^{r-1})} ,$$

where $\xi^{(1)}$ and $\xi^{(0)}$ are as given in (22.50) and (22.51), respectively.

22.3 Joint Distribution of Numbers of Runs

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $\Pr[X_1 = i] = p_i$, $i = 1, 2, \ldots, m$, where $p_1 + \ldots + p_m = 1$. For $i = 1, \ldots, m$, we are interested in runs of i of length k_i. We denote by E_i a run of i of length k_i. Let τ_i be the waiting time (the number of trials) for the i-th event. For notational convenience, we let $\tau_0 = 0$. Fix a permutation $\sigma \in \mathcal{S}_m$, where $\mathcal{S}_m$ is the permutation group on $\{1, 2, \ldots, m\}$. When the first event is $E_{\sigma(1)}$ among $E_1, \ldots, E_m$, we denote by N_{11}^{σ} the number of occurrences of the first event $E_{\sigma(1)}$ in $(\tau_0, \tau_1]$. Similarly, when the first and the second events are $E_{\sigma(1)}$ and $E_{\sigma(2)}$, the numbers of occurrences of $E_{\sigma(1)}$ and $E_{\sigma(2)}$ in $(\tau_1, \tau_2]$ are denoted by N_{12}^{σ} and N_{22}^{σ}, respectively. More generally, when the first, the second, ..., and the j-th events are $E_{\sigma(1)}, E_{\sigma(2)}, \ldots, E_{\sigma(j)}$, the numbers of occurrences of $E_{\sigma(1)}, E_{\sigma(2)}, \ldots, E_{\sigma(j)}$ in $(\tau_{j-1}, \tau_j]$ are denoted by $N_{1j}^{\sigma}, \ldots, N_{jj}^{\sigma}$. Let $\phi_1^{\sigma}(t_1)$ be the (improper) p.g.f. of N_{11}^{σ} and let $\phi_j^{\sigma}(t_1, \ldots, t_j)$ be the (improper) conditional joint p.g.f. of $(N_{1j}^{\sigma}, \ldots, N_{jj}^{\sigma})$ given that the first event is $E_{\sigma(1)}$, ..., the $(j-1)$-th event is $E_{\sigma(j-1)}$.

Remark 22.3.1 N_{11}^{σ} is defined only when the first event is $E_{\sigma(1)}$. Hence, we see that $\phi_1^{\sigma}(1) = \Pr[\text{the first event is } E_{\sigma(1)}]$. Similarly, the conditional probability of $(N_{1j}^{\sigma}, \ldots, N_{jj}^{\sigma})$ can be considered only when $\{\text{the } j\text{-th event} = E_{\sigma(j)}\}$. Then, we have $\phi_j^{\sigma}(1, 1, \ldots, 1)$ to be the conditional probability that the j-th event is $E_{\sigma(j)}$ given that the first event is $E_{\sigma(1)}$, ..., the $(j-1)$-th event is $E_{\sigma(j-1)}$.

Proposition 22.3.1 *The above (improper) p.g.f.'s are given by*

$$\phi_1^{\sigma}(t_1) = P_{\sigma(1)} t_1,$$

and

$$\phi_j^{\sigma}(t_1, \ldots, t_j) = \frac{P_{\sigma(j)} t_j}{1 - \sum_{i=1}^{j-1} P_{\sigma(i)} t_i} \qquad (j \geq 2),$$

where

$$P_i = \frac{\frac{p_i^{k_i}(1-p_i)}{1-p_i}}{\sum_{j=1}^{m} \frac{p_j^{k_j}(1-p_j)}{1-p_j^{k_j}}}, \qquad i = 1, 2, \ldots, m.$$

Proof. $\phi_1^{\sigma}(t_1) = P_{\sigma(1)} t_1$ is clear, since the first event is $E_{\sigma(1)}$ and P_i is the probability that the first event is E_i. The first possible event after τ_{j-1} is one of the events $E_{\sigma(1)}, \ldots, E_{\sigma(j)}$ and they are mutually exclusive. Therefore, we have

$$\begin{aligned} \phi_j^{\sigma}(t_1, \ldots, t_j) &= P_{\sigma(1)} t_1 \cdot \phi_j^{\sigma}(t_1, \ldots, t_j) \\ &\quad + P_{\sigma(2)} t_2 \cdot \phi_j^{\sigma}(t_1, \ldots, t_j) + \ldots \\ &\quad + P_{\sigma(j-1)} t_{j-1} \cdot \phi_j^{\sigma}(t_1, \ldots, t_j) + P_{\sigma(j)} t_j. \end{aligned}$$

By solving the equation, we have the result. This completes the proof. ■

When the first event is $E_{\sigma(1)}$, ..., the j-th event is $E_{\sigma(j)}$, the number of occurrences of the i-th event $E_{\sigma(i)}$ until τ_j is $N^\sigma_{ii} + N^\sigma_{i,i+1} + \ldots + N^\sigma_{ij}$. Then we have the following result.

Theorem 22.3.1 *For $j = 1, 2, \ldots, m$, the p.g.f. of the joint distribution of the numbers of occurrences of the first, the second, ..., the j-th events until τ_j, is given by*

$$\phi(t_1, \ldots, t_j) = \sum_{\sigma \in \mathcal{S}^m_j} \phi^\sigma_1(t_1)\phi^\sigma_2(t_1, t_2) \ldots \phi^\sigma_{j-1}(t_1, \ldots, t_{j-1})\phi^\sigma_j(t_1, \ldots, t_j),$$

where $\mathcal{S}^m_j$ means the totality of permutations of j different letters in $\{1, 2, \ldots, m\}$.

In particular, by setting $j = m$, we get the following result.

Corollary 22.3.1 *The joint p.g.f. of the number of occurrences of the first, the second, ..., the $(m-1)$-th events until the last event occurs for the first time is given by*

$$\begin{aligned} &\phi(t_1, \ldots, t_{m-1}) \\ &\quad = \sum_{\sigma \in \mathcal{S}_m} \phi^\sigma_1(t_1)\phi^\sigma_2(t_1, t_2) \ldots \phi^\sigma_{m-1}(t_1, \ldots, t_{m-1})\phi^\sigma_m(t_1, \ldots, t_{m-1}, 1). \end{aligned}$$

Example 22.3.1 When we set $m = 2$, Corollary 22.3.1 gives the p.g.f. of the number N of occurrences of the sooner event until the later event occurs for the first time as

$$\begin{aligned} \phi(t_1) &= P_1 t_1 \frac{P_2}{1 - P_1 t_1} + P_2 t_1 \frac{P_1}{1 - P_2 t_1} \\ &= P_1 P_2 t_1 \left(\sum_{n=0}^{\infty} P_1^n t_1^n + \sum_{m=0}^{\infty} P_2^m t_1^m \right). \end{aligned}$$

By taking the coefficient of t_1^n, we have

$$\Pr[N = n] = P_1^n P_2 + P_1 P_2^n \qquad \text{for } n \geq 1.$$

See also Remark 22.2.1.

Example 22.3.2 By setting $m = 3$ in Corollary 22.3.1, we see that the joint p.g.f. of the numbers (N_1, N_2) of occurrences of the first and the second events until the third event occurs for the first time is given by

$$\begin{aligned} \phi(t_1, t_2) &= \sum_{\sigma \in \mathcal{S}_3} \phi^\sigma_1(t_1)\phi^\sigma_2(t_1, t_2)\phi^\sigma_3(t_1, t_2, 1) \\ &= \sum_{\sigma \in \mathcal{S}_3} P_{\sigma(1)} t_1 \cdot \frac{P_{\sigma(2)} t_2}{1 - P_{\sigma(1)} t_1} \cdot \frac{P_{\sigma(3)}}{1 - P_{\sigma(1)} t_1 - P_{\sigma(2)} t_2}. \end{aligned}$$

For simplicity, we take the typical term for $\sigma =$ identity. Note that

$$\begin{aligned} P_1 t_1 \cdot \frac{P_2 t_2}{1 - P_1 t_1} \cdot \frac{P_3}{1 - P_1 t_1 - P_2 t_2} \\ = P_1 P_2 P_3 t_1 t_2 \left(\sum_{n=0}^{\infty} (P_1 t_1)^n \right) \left(\sum_{m=0}^{\infty} (P_1 t_1 + P_2 t_2)^m \right). \end{aligned}$$

The coefficient of $t_1^x t_2^y$ in the above formula is the probability $\Pr[N_1 = x, N_2 = y,$ (1st, 2nd, 3rd) $= (E_1, E_2, E_3)]$. Then, the probability is given by

$$\begin{aligned} & P_1 P_2 P_3 \sum_{n=0}^{\infty} P_1^n \binom{x - n + y - 2}{y - 1} P_1^{x-n-1} P_2^{y-1} \\ = \; & P_1^x P_2^y P_3 \sum_{a=0}^{x+y-2} \binom{a}{y-1} = \binom{x+y-1}{y} P_1^x P_2^y P_3. \end{aligned}$$

For the last equality, we used the formula for binomial coefficients

$$\sum_{a=0}^{n} \binom{a}{b} = \binom{n+1}{b+1}.$$

Therefore, we have

$$\Pr[N_1 = x, N_2 = y] = \sum_{\sigma \in S_3} \binom{x+y-1}{x-1} P_{\sigma(1)}^x P_{\sigma(2)}^y P_{\sigma(3)}.$$

References

1. Aki, S. (1992). Waiting time problems for a sequence of discrete random variables, *Annals of the Institute of Statistical Mathematics*, **44**, 363–378.

2. Aki, S., Balakrishnan, N. and Mohanty, S. G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **48**, 773–787.

3. Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two-state Markov chain, In *Statistical Science & Data Analysis* (Eds., K. Matusita, M. L. Puri and T. Hayakawa), pp. 467–474, Utrecht: VSP Publishers.

4. Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes, *Annals of the Institute of Statistical Mathematics*, **46**, 193–202.

5. Balasubramanian, K., Viveros, R. and Balakrishnan, N. (1993). Sooner and later waiting time problems for Markovian Bernoulli trials, *Statistics & Probability Letters*, **18**, 153–161.

6. Ebneshahrashoob, M. and Sobel, M. (1990). Sooner and later problems for Bernoulli trials: frequency and run quotas, *Statistics & Probability Letters*, **9**, 5–11.

7. Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050–1058.

8. Goldstein, L. (1990). Poisson approximation and DNA sequence matching, *Communications in Statistics—Theory and Methods*, **19**, 4167–4179.

9. Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313–320.

10. Mohanty, S. G. (1994). Success runs of length k in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **46**, 777–796.

11. Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **47**, 415–433.

23

Distributions of Numbers of Success-Runs Until the First Consecutive k Successes in Higher Order Markov Dependent Trials

Katuomi Hirano, Sigeo Aki and Masayuki Uchida

The Institute of Statistical Mathematics, Tokyo, Japan
Osaka University, Osaka, Japan
The Institute of Statistical Mathematics, Tokyo, Japan

Abstract: The distributions of numbers of overlapping and non-overlapping occurrences of success-runs of length l, and the distributions of numbers of occurrences of success-runs of exact length l and of length l or more until the first occurrence of success-run of length k in the m-th order Markov dependent trials are studied. When $m \leq l < k$, the derived distributions do not depend on the initial distribution of the m-th order Markov chain and they are shown to be equal to the corresponding distributions considered in independent trials with a success probability whose value is given by the transition probability that a success occurs after a success-run of length m in the m-th order Markov chain. When $l < m$, the distributions depend on the initial distribution of the m-th order Markov chain and are not necessarily so simple as the above case. A method for deriving the probability generating function of the conditional distribution of the number of overlapping occurrences of success-runs of length l until the first occurrence of success-run of length k under each initial condition is given.

Keywords and phrases: Probability generating function, discrete distributions of order k, success and failure runs, higher order Markov chain

23.1 Introduction

Currently exact discrete distribution theory of runs has been developed. In the simplest case of independent and identically distributed Bernoulli trials with success probability p, the distribution of the number of trials until the first consecutive k successes is well known [see, for example, Feller (1968) and Johnson, Kotz and Kemp (1992)]. Philippou, Georghiou and Philippou (1983) called it the geometric distribution of order k. The binomial distribution of order k is the distribution of the number of occurrences of success-runs of length k until the n-th trial. Discrete distributions of order k such as the geometric and the binomial distributions of order k have been studied by many authors in more general situations. Some of the results on these distributions are closely related to problems on reliability of the consecutive-k-out-of n:F systems [Hirano (1994), Chao, Fu and Koutras (1995)], and on evaluation of the start-up demonstration test [Viveros and Balakrishnan (1993) and Balakrishnan, Balasubramanian and Viveros (1995)].

Further exact discrete distribution theory has been investigated even in the case of dependent trials such as the homogeneous Markov chain and the binary sequence of order k [Rajarshi (1974), Schwager (1983), Aki (1985), Aki and Hirano (1993), Balasubramanian, Viveros and Balakrishnan (1993), Hirano and Aki (1993), Mohanty (1994), Koutras and Alexandrou (1995), Uchida and Aki (1995) and Aki, Balakrishnan and Mohanty (1996)]. Multivariate distributions of order k and related issues have been reviewed recently by Johnson, Kotz and Balakrishnan (1997).

Aki and Hirano (1994) showed that, when $\{0,1\}$-sequence follows the first order Markov chain, the distribution of the number of overlapping occurrences of success-runs of length l until the first consecutive k successes is the shifted geometric distribution of order $k-l$ with the support $\{k-l+1, k-l+2, \dots\}$, where we usually regard the value 1 as success and the value 0 as failure. Further, Aki and Hirano (1995) extended the result and studied the joint distributions of the numbers of trials and of outcomes such as successes, failures and success-runs until the first consecutive k successes in the first order Markov dependent trials. Now suppose that $\{0,1\}$-sequence follows the m-th order Markov chain. Then, the purpose of this paper is to examine the distribution of number of occurrences of success-runs of length l until the first occurrence of success-run of length k.

Throughout the paper, let k, l and m be fixed positive integers such that $l,\ m < k$. We denote by $G_k(p)$ the geometric distribution of order k, and by $G_k(p,a)$ the shifted geometric distribution of order k so that its support begins with a. Here the geometric distribution, to be denoted by $G(p)$, is defined as the distribution of the number of failures preceding the first success. Note that $G(p) = G_1(p,0)$.

There are different ways of counting the numbers of success-runs of length k [see, for example, Fu and Koutras (1993)]. Feller (1968, Chapter XIII) defined a way of counting the number of runs exactly of length k as counting the number from scratch every time a run occurs. For example, the sequence $SSS \mid SFSSS \mid SSS \mid F$ contains 3 success-runs of length 3. In Goldstein (1990), the number of success-runs of length 3 or more in the sequence $SSSS \mid FSSSSSS \mid F$ is 2. Ling (1988) defined a way of counting the number of success-runs of length k by the overlapping way of counting. By Ling's way of counting, the sequence $SSSSFSSSSSSF$ contains 6 success-runs of length 3. In Mood (1940), the number of success-runs of exact length 3 in the above sequence $SSSSFSSSSSSF$ is 0.

In Section 23.2, when $m \leq l < k$, we investigate the distributions of numbers of overlapping and non-overlapping occurrences of success-runs of length l, the distributions of numbers of occurrences of success-runs of exact length l and of length l or more. In Section 23.3, when $l < m$, we give a method for deriving the probability generating function of the conditional distribution of the number of overlapping occurrences of success-runs of length l under each initial condition until the first occurrence of success-run of length k.

23.2 Numbers of Success-Runs in Higher Order Markov Chain

In this Section, suppose that $m \leq k$. Let $X_{-m+1}, X_{-m+2}, \ldots, X_0, X_1, X_2, \ldots$ be $\{0,1\}$-valued m-th order Markov chain with

$$\begin{aligned}
\pi_{x_1,\ldots,x_m} &= \Pr[X_{-m+1} = x_1, X_{-m+2} = x_2, \ldots, X_0 = x_m], \\
p_{x_1,\ldots,x_m} &= \Pr[X_i = 1 | X_{i-m} = x_1, X_{i-m+1} = x_2, \ldots, X_{i-1} = x_m], \\
&= 1 - q_{x_1,\ldots,x_m},
\end{aligned}$$

for $x_1, \ldots, x_m = 0, 1$ and $i = 1, 2, \ldots$. For $x_1, \ldots, x_m = 0, 1$, we assume that $0 < p_{x_1,\ldots,x_m},\ q_{x_1,\ldots,x_m} < 1$. We denote by τ the number of trials until the first consecutive k successes in $X_1, X_2, \ldots$.

First, we derive the distribution of the number of overlapping occurrences of "1"-runs of length l until τ. Let $\phi^{(x_1,\ldots,x_m)}(t)$ be the probability generating function (p.g.f.) of the conditional distribution of the number of overlapping occurrences of "1"-runs of length l until τ given that $X_{-m+1} = x_1, X_{-m+2} = x_2, \ldots, X_0 = x_m$. Suppose we have currently "1"-run of length i in $X_i, X_{i-1}, \ldots$. Then, we denote by $\phi_i(t)$ the p.g.f. of the conditional distribution of the number of overlapping occurrences of "1"-runs of length l from this time until τ.

Theorem 23.2.1 *If $m \leq l < k$, then the distribution of the number of overlapping occurrences of "1"-runs of length l until the first occurrence of the "1"-run of length k in $X_1, X_2, \ldots$ is the shifted geometric distribution of order $k-l$, $G_{k-l}(p_{11\ldots1}, k-l+1)$.*

PROOF. We are waiting for the first occurrence of "1"-run of length k and l is less than k. Then, starting from any initial state $(x_1, \ldots, x_m)$ we observe the first occurrence of "1"-run of length l somewhere in $X_l, X_{l+1}, \ldots$ with probability 1. By considering the m-th order Markov chain just after the first occurrence of "1"-run of length l, we see that

$$\phi^{(x_1,\dots,x_m)} = t\phi_l \quad \text{for each initial state } (x_1,\dots,x_m).$$

This shows that all the conditional distributions are equal to each other and they do not depend on their initial conditions. Hence, we denote $\phi \equiv t\phi_l = \phi^{(x_1,\dots,x_m)}$. Then, it is easy to see that

$$\begin{cases} \phi_l &= p_{11\dots1}t\phi_{l+1} + q_{11\dots1}\phi \\ \phi_{l+1} &= p_{11\dots1}t\phi_{l+2} + q_{11\dots1}\phi \\ & \dots \\ \phi_{k-1} &= p_{11\dots1}t + q_{11\dots1}\phi. \end{cases}$$

By solving this system of equations, we obtain

$$\phi_l = q_{11\dots1}\phi\,\frac{1-(p_{11\dots1}t)^{k-l}}{1-p_{11\dots1}t} + (p_{11\dots1}t)^{k-l}.$$

Then, $\phi = t\phi_l$ implies

$$\phi = \frac{{p_{11\dots1}}^{k-l}t^{k-l+1}(1-p_{11\dots1}t)}{1-t+q_{11\dots1}t(p_{11\dots1}t)^{k-l}}.$$

This completes the proof. ■

Similarly, we can derive the distribution of the number of non-overlapping occurrences of "1"-runs of length l until τ. Let $\psi^{(x_1,\dots,x_m)}(t)$ be the p.g.f. of the conditional distribution of the number of non-overlapping occurrences of "1"-runs of length l until τ given that $X_{-m+1} = x_1, X_{-m+2} = x_2, \dots, X_0 = x_m$. Suppose we have currently "1"-run of length i in $X_i, X_{i-1}, \dots$. Then, we denote by $\psi_i(t)$ the p.g.f. of the conditional distribution of the number of non-overlapping occurrences of "1"-runs of length l from this time until τ.

Theorem 23.2.2 *Suppose that $m \le l < k$. Let $k = \nu l + \mu$, where ν and μ are nonnegative integers and $0 \le \mu < l$. Then, the p.g.f. of the distribution of the number of non-overlapping occurrences of "1"-runs of length l until the first occurrence of "1"-run of length k in $X_1, X_2, \dots$ is given by*

$$\frac{p^{k-l}t^{\nu}}{1-t(1-p^l)\frac{1-(p^lt)^{\nu-1}}{1-p^lt} - p^{(\nu-1)l}(1-p^{\mu})t^{\nu}}, \tag{23.1}$$

where $p = p_{11\dots1}$.

PROOF. From the same reason as in the overlapping case, we see that

$$\psi^{(x_1,\dots,x_m)} = t\psi_l \quad \text{for each initial state } (x_1,x_2,\dots,x_m).$$

Since all the p.g.f.'s of the conditional distributions are the same, we denote it by ψ. Of course, ψ is the (unconditional) p.g.f. of the number of non-overlapping occurrences of "1"-runs of length l until τ regardless of the initial distribution.

Then, we see that the following systems of equations hold:

$$\begin{cases} \psi_l &= p_{11\ldots1}\psi_{l+1} + q_{11\ldots1}\psi \\ \psi_{l+1} &= p_{11\ldots1}\psi_{l+2} + q_{11\ldots1}\psi \\ & \cdots \\ \psi_{2l-1} &= p_{11\ldots1}t\psi_{2l} + q_{11\ldots1}\psi \end{cases} \tag{23.2}$$

$$\begin{cases} \psi_{2l} &= p_{11\ldots1}\psi_{2l+1} + q_{11\ldots1}\psi \\ \psi_{2l+1} &= p_{11\ldots1}\psi_{2l+2} + q_{11\ldots1}\psi \\ & \cdots \\ \psi_{3l-1} &= p_{11\ldots1}t\psi_{3l} + q_{11\ldots1}\psi \end{cases} \tag{23.3}$$

$$\cdots$$

$$\begin{cases} \psi_{(\nu-1)l} &= p_{11\ldots1}\psi_{(\nu-1)l+1} + q_{11\ldots1}\psi \\ \psi_{(\nu-1)l+1} &= p_{11\ldots1}\psi_{(\nu-1)l+2} + q_{11\ldots1}\psi \\ & \cdots \\ \psi_{\nu l-1} &= p_{11\ldots1}t\psi_{\nu l} + q_{11\ldots1}\psi \end{cases} \tag{23.4}$$

$$\begin{cases} \psi_{\nu l} &= p_{11\ldots1}\psi_{\nu l+1} + q_{11\ldots1}\psi \\ \psi_{\nu l+1} &= p_{11\ldots1}\psi_{\nu l+2} + q_{11\ldots1}\psi \\ & \cdots \\ \psi_{\nu l+\mu-1} &= p_{11\ldots1}1 + q_{11\ldots1}\psi\ . \end{cases} \tag{23.5}$$

By solving (23.2), (23.3), (23.4) and (23.5), we have

$$\begin{aligned} \psi_l &= \psi(1-p_{11\ldots1}{}^l)\,\frac{1-(p_{11\ldots1}{}^l t)^{\nu-1}}{1-p_{11\ldots1}{}^l t} + \psi(p_{11\ldots1}{}^l t)^{\nu-1}(1-p_{11\ldots1}{}^\mu) \\ &\quad + p_{11\ldots1}{}^{l(\nu-1)+\mu}t^{\nu-1}. \end{aligned}$$

Then, $t\psi_l = \psi$ implies (23.1), which completes the proof. ■

Remark. It is easy to see that the formula (23.1) agrees with

$$\frac{p^k t^{[k/l]}}{1-(1-p)\sum_{i=0}^{k-1} p^i t^{[i/l]}}\ ,$$

which was derived as p.g.f. of the distribution of the number of non-overlapping occurrences of "1"-runs of length l until τ in independent trials with success probability p [*cf.* Corollary 3.2 of Aki and Hirano (1995)].

Here, we give a recurrence formula of probabilities of this distribution.

Proposition 23.2.1 *We denote by $P(n)$ the probability that the number of non-overlapping occurrences of "1"-runs of length l until the first occurrence of "1"-run of length k is n. Then, the following recurrence formula holds:*

$$\begin{cases} P(n) & = 0 \quad \textit{for } n < \nu \\ P(\nu) & = p^{k-l} \\ P(\nu+1) & = -p^k + p^{k-l} - \alpha P(1) \\ P(n) & = P(n-1) - \alpha P(n-\nu) - \beta P(n-\nu-1) \quad \textit{for } n \geq \nu+2, \end{cases}$$

where $\alpha = (p^\mu - p^l)p^{l(\nu-1)}$ and $\beta = p^{\nu l} - p^k$.

Proof. It is easily seen that (23.1) can be rewritten as

$$t^\nu \left\{ \frac{p^{k-l} - p^k t}{1 - t + \alpha t^\nu + \beta t^{\nu+1}} \right\}.$$

We set

$$\frac{p^{k-l} - p^k t}{1 - t + \alpha t^\nu + \beta t^{\nu+1}} = \sum_{n=0}^{\infty} f(n) t^n,$$

and then multiply $1 - t + \alpha t^\nu + \beta t^{\nu+1}$ on both sides. Then, by comparing the coefficient of t^n we see that

$$\begin{cases} f(0) = p^{k-l} \\ f(1) - f(0) + \alpha f(1-\nu) + \beta f(-\nu) = -p^k \\ f(n) - f(n-1) + \alpha f(n-\nu) + \beta f(n-\nu-1) = 0 \quad (n \geq 2). \end{cases}$$

This completes the proof. ∎

Next, we study the distribution of the number of occurrences of "1"-runs of exact length l until τ. Let ξ_l ($l = m, m+1, \ldots, k-1$) be the number of occurrences of "1"-runs of exact length l until τ. We denote by $\varphi^{(x_1,\ldots,x_m)}(t)$ the p.g.f. of the conditional distribution of ξ_l given that $X_{-m+1} = x_1, \ldots, X_0 = x_m$. Let $\varphi_i(t)$ be the p.g.f. of the conditional distribution of ξ_ℓ given that we start with a "1"-run of length i, where $i \geq m$. As in the previous cases, we observe

$$\varphi^{(x_1,\ldots,x_m)} = \varphi_l. \tag{23.6}$$

Here, we note that $\varphi^{(x_1,\ldots,x_m)}$ does not depend on initial condition X_{-m+1}, $\ldots, X_0$. So, we set $\varphi^{(x_1,\ldots,x_m)} = \varphi$. Furthermore, we have the following system of equations:

$$\begin{cases} \varphi_l & = p_{11\ldots1}\varphi_{l+1} + q_{11\ldots1} t\varphi \\ \varphi_{l+1} & = p_{11\ldots1}\varphi_{l+2} + q_{11\ldots1}\varphi \\ & \cdots \\ \varphi_{k-1} & = p_{11\ldots1} + q_{11\ldots1}\varphi \; . \end{cases} \tag{23.7}$$

From (23.6) and (23.7), we obtain

$$\varphi = \frac{p_{11\ldots1}{}^{k-l}}{p_{11\ldots1}{}^{k-l} + q_{11\ldots1} - q_{11\ldots1}t} \; .$$

Hence, we have the following result.

Theorem 23.2.3 *For $l = m, m+1, \ldots, k-1$, let ξ_l be the number of occurrences of "1"-runs of exact length l until the first occurrence of "1"-run of length k. Then the distribution of ξ_l is the geometric distribution, i.e., $\xi_l \sim G(p_{11\ldots1}{}^{k-l}/(p_{11\ldots1}{}^{k-l} + q_{11\ldots1}))$.*

Finally, we examine the distribution of the number of occurrences of "1"-runs of length l or more until τ. Let ν_l $(l = m, m+1, \ldots, k-1)$ be the number of occurrences of "1"-runs of length l or more until τ. We denote by $\rho^{(x_1,\ldots,x_m)}(t)$ the p.g.f. of the conditional distribution of ν_l given that $X_{-m+1} = x_1, \ldots, X_0 = x_m$. Let $\rho_i(t)$ be the p.g.f. of the conditional distribution of ν_l given that we start with a "1"-run of length i, where $i \geq m$. Then, we observe

$$\rho^{(x_1,\ldots,x_m)} = \rho_l. \tag{23.8}$$

Here, we note that $\rho^{(x_1,\ldots,x_m)}$ does not depend on the initial condition $X_{-m+1}, \ldots, X_0$. So, we set $\rho^{(x_1,\ldots,x_m)} = \rho$. Furthermore, we have the following system of equations:

$$\begin{cases} \rho_l &= p_{11\ldots1}\rho_{l+1} + q_{11\ldots1}t\rho \\ \rho_{l+1} &= p_{11\ldots1}\rho_{l+2} + q_{11\ldots1}t\rho \\ & \cdots \\ \rho_{k-1} &= p_{11\ldots1}t + q_{11\ldots1}t\rho \; . \end{cases} \tag{23.9}$$

From (23.8) and (23.9), we obtain

$$\rho = \frac{p_{11\ldots1}{}^{k-l}t}{1 - t + p_{11\ldots1}{}^{k-l}t} \; .$$

We can state this result as follows.

Theorem 23.2.4 *For $l = m, m+1, \ldots, k-1$, let ν_l be the number of occurrences of "1"-runs of length l or more until the first occurrence of "1"-run of length k. Then the distribution of ν_l is the geometric distribution of order 1, i.e., $\nu_l \sim G_1(p_{11\ldots1}{}^{k-l})$.*

23.3 Case $l < m$

In the last section we investigated the distributions of the numbers of overlapping and non-overlapping occurrences of success-runs of length l and of the numbers of success-runs of exact length l and of length l or more until τ in the m-th order Markov dependent trials when $l \geq m$. We are interested in the case $l < m$. However, in general, when $l < m$, the distributions depend on the initial condition of the m-th order Markov chain and are not necessarily so simple as the case $l \geq m$.

In this section we give a method for deriving the p.g.f. of the conditional distribution of the number of overlapping occurrences of success-runs of length l under each initial condition until τ in the m-th order Markov dependent trials when $l < m$.

A sequence which follows the m-th order Markov chain depends on the past occurrences of length m. A set of $\{0, 1\}$-sequence of length m consists of 2^m elements, and can be uniquely regarded as a binary number. Further we translate it into a decimal number. For example, when $m = 3$, $p_{101} = p_5$. Let $N_m = \{0, 1, 2, \ldots, 2^m - 1\}$ and let f_i $(i = 0, 1)$ be the mapping from N_m to N_m such that

$$f_i(x) = 2x + i \pmod{2^m}, \text{ for } i = 0, 1$$

and define g_l by

$$g_l(j) = \begin{cases} j - l + 1 & \text{if } j \geq l \\ 0 & \text{if } j < l. \end{cases}$$

We denote by $\phi^{(x)}$ (for each $x \in N_m$) the p.g.f. of the conditional distribution of the number of overlapping occurrences of success-runs of length l until τ. Then we have the following system of 2^m equations of conditional p.g.f.'s

$$\begin{aligned} \phi^{(x)} &= q_x \phi^{(f_0(x))} + p_x q_{f_1(x)} \phi^{(f_0 \circ f_1(x))} t^{g_l(1)} \\ &\quad + p_x p_{f_1(x)} q_{f_1^2(x)} \phi^{(f_0 \circ f_1^2(x))} t^{g_l(2)} \\ &\quad + \ldots + p_x p_{f_1(x)} \cdots p_{f_1^{k-2}(x)} q_{f_1^{k-1}(x)} \phi^{(f_0 \circ f_1^{k-1}(x))} t^{g_l(k-1)} \\ &\quad + p_x p_{f_1(x)} \cdots p_{f_1^{k-1}(x)} t^{g_l(k)}, \quad \text{for each } x \in N_m \end{aligned}$$

by considering all possibilities of the first occurrence of 0. In this case, the system is linear with respect to the conditional p.g.f's, and so we can solve it by using computer algebraic systems. An illustrative example is given in Hirano, Aki and Uchida (1996).

References

1. Aki, S. (1985). Discrete distributions of order k on a binary sequence, *Annals of the Institute of Statistical Mathematics*, **37**, 205–224.

2. Aki, S., Balakrishnan, N. and Mohanty, S. G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **48**, 773–787.

3. Aki, S. and Hirano, K. (1993). Discrete distributions related to succession events in a two-state Markov chain, In *Statistical Science & Data Analysis* (Eds., K. Matusita, M. L. Puri and T. Hayakawa), pp. 467–474, Utrecht: VSP Publishers.

4. Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes, *Annals of the Institute of Statistical Mathematics*, **46**, 193–202.

5. Aki, S. and Hirano, K. (1995). Joint distributions of numbers of success-runs and failures until the first consecutive k successes, *Annals of the Institute of Statistical Mathematics*, **47**, 225–235.

6. Balakrishnan, N., Balasubramanian, K. and Viveros, R. (1995). Start-up demonstration tests under correlation and corrective action, *Naval Research Logistics*, **42**, 1271–1276.

7. Balasubramanian, K., Viveros, R. and Balakrishnan, N. (1993). Sooner and later waiting time problems for Markovian Bernoulli trials, *Statistics & Probability Letters*, **18**, 153–161.

8. Chao, M. T., Fu, J. C. and Koutras, M. V. (1995). Survey of reliability studies of consecutive-k-out-of-n:F & related systems, *IEEE Transactions on Reliability*, **40**, 120–127.

9. Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Volume 1, Third edition, New York: John Wiley & Sons.

10. Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050–1058.

11. Goldstein, L. (1990). Poisson approximation and DNA sequence matching, *Communications in Statistics—Theory and Methods*, **19**, 4167–4179.

12. Hirano, K. (1994). Consecutive-k-out-of-n:F systems, *Proceedings of the Institute of Statistical Mathematics*, **42**, 45–61 (in Japanese).

13. Hirano, K. and Aki, S. (1993). On number of occurrences of success runs of specified length in a two-state Markov chain, *Statistica Sinica*, **3**, 313–320.

14. Hirano, K., Aki, S. and Uchida, M. (1996). Distributions of numbers of success-runs until the first consecutive k successes in higher order Markov dependent trials, *Research Memorandum No. 603*, The Institute of Statistical Mathematics, Tokyo, Japan.

15. Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

16. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second edition, New York: John Wiley & Sons.

17. Koutras, M. V. and Alexandrou, V. A. (1995). Runs, scans and urn model distributions: a unified Markov chain approach, *Annals of the Institute of Statistical Mathematics*, **47**, 743–766.

18. Ling, K. D. (1988). On binomial distributions of order k, *Statistics & Probability Letters*, **6**, 247–250.

19. Mohanty, S. G. (1994). Success runs of length k in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **46**, 777–796.

20. Mood, A. M. (1940). The distribution theory of runs, *Annals of Mathematical Statistics*, **11**, 367–392.

21. Philippou, A. N., Georghiou, C. and Philippou, G. N. (1983). A generalized geometric distribution and some of its properties, *Statistics & Probability Letters*, **1**, 171–175.

22. Rajarshi, M. B. (1974). Success runs in a two-state Markov chain, *Journal of Applied Probability*, **11**, 190–192.

23. Schwager, S. J. (1983). Run probability in sequences of Markov-dependent trials, *Journal of the American Statistical Association*, **78**, 168–175.

24. Uchida, M. and Aki, S. (1995). Sooner and later waiting time problems in a two-state Markov chain, *Annals of the Institute of Statistical Mathematics*, **47**, 415–433.

25. Viveros, R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data, *Journal of Quality Technology*, **25**, 119–130.

24

On Multivariate Distributions of Various Orders Obtained by Waiting for the r-th Success Run of Length k in Trials With Multiple Outcomes

Demetrios L. Antzoulakos and Andreas N. Philippou

University of Crete, Rethymno, Greece
University of Patras, Patras, Greece

Abstract: A sequence of independent trials with $m + 1$ mutually exclusive outcomes S, F_1, F_2, ..., F_m is considered until the occurrence of the r-th non-overlapping success run of length k, and the distributions of related random vectors are derived. First a new genesis scheme is established for the multivariate negative binomial distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988). It is shown that it is the distribution of the sum of two random vectors: the i-th component of the first one is the number of occurrences of F_i and the i-th component of the second one is the total number of S's which precede directly the occurrences of F_i but do not belong to any success run of length k $(1 \leq i \leq m)$. Furthermore, we obtain exact distributions of random vectors whose components are numbers of failures, non-overlapping runs of failures, successes, overlapping success runs of length l and success runs of length at least l. The majority of the above problems are also treated in the case of the generalized sequence of order k and corresponding results are established regarding the multivariate extended negative binomial distribution of order k of Philippou and Antzoulakos (1990). The present paper generalizes several results of Aki and Hirano (1994, 1995).

Keywords and phrases: Multivariate distributions of order k, type I, extended, negative binomial, geometric, genesis scheme, success, failure of type i, run

24.1 Introduction

The exact probability mass function of the number of trials until the occurrence of the first success run of length k in Bernoulli trials was obtained by Philippou, Georghiou and Philippou (1983), who called it the *geometric distribution of order* k and derived from its study the *negative binomial distribution of order* k and the *Poisson distribution of order* k [also see Philippou (1983, 1984)]. Since then, a substantial number of papers have appeared on univariate distributions of order k; we mention the papers by Aki (1985), Aki, Balakrishnan and Mohanty (1996), Aki, Kuboki and Hirano (1984), Balakrishnan, Mohanty and Aki (1997), Balasubramanian, Viveros and Balakrishnan (1993), Charalambides (1986), Ebneshahrashoob and Sobel (1990), Fu and Koutras (1994), Godbole (1990), Hirano (1986), Ling (1988), Mohanty (1994), Panaretos and Xekalaki (1986), Philippou (1988), Philippou and Makri (1986) and references therein. The distribution theory of the multivariate distributions of order k was initiated by Philippou, Antzoulakos and Tripsiannis (1988), who obtained appropriate multivariate analogs for several univariate distributions of order k. Further results and/or new multivariate distributions of order k were obtained by Antzoulakos and Philippou (1994), Ling and Tai (1990), Philippou and Antzoulakos (1990), Philippou and Tripsiannis (1991), Philippou, Antzoulakos and Tripsiannis (1990) and Tripsiannis (1993). A comprehensive review of these developments can be found in the recent book by Johnson, Kotz and Balakrishnan (1997).

In the present paper, we establish new genesis schemes for two specific multivariate distributions of order k. The first one is the multivariate negative binomial distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988) [also see Antzoulakos and Philippou (1991)]. It was denoted there by $\overline{MNB}_{k,I}(r;Q_1,Q_2,\ldots,Q_m)$ but it will be denoted here by $MNB_k(r;q_1,q_2,\ldots,q_m)$. This distribution is the multivariate analog of the (suitably shifted) negative binomial distribution of the same order of Philippou, Georghiou and Philippou (1983) which is denoted here by $NB_k(r;p)$. The second one is the multivariate extended negative binomial distribution of order k of Philippou and Antzoulakos (1990). It was denoted there by $\overline{MENB}_k(r;q_{11},\ldots,q_{mk})$ but it will be denoted here by $MENB_k(r;q_{11},\ldots,q_{mk})$. This distribution is the multivariate analog of the (suitably shifted) extended negative binomial distribution of the same order of Aki (1985) which is denoted here by $ENB_k(r;p_1,p_2,\ldots,p_k)$. We note that the support of the above mentioned univariate distributions begins with 0. For k=1, the above mentioned multivariate distributions reduce of course to the usual multivariate negative binomial distribution [see, for example, Patil *et al.* (1984, p. 107)] which is denoted here by $MNB_1(r;q_1,q_2,\ldots,q_m)$.

In Section 24.2, we consider a sequence of independent trials with $m+1$ possible outcomes S, F_1, F_2, $\ldots$, F_m until the occurrence of the r-th non-overlapping S-run of length k. In Theorem 24.2.1, we obtain a new genesis scheme of $MNB_k(r; q_1, q_2, \ldots, q_m)$. It is shown that it is the distribution of the sum of two random vectors $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$ and $\mathbf{Z} = (Z_1, Z_2, \ldots Z_m)$: Y_i $(1 \leq i \leq m)$ denotes the number of occurrences of F_i, and Z_i $(1 \leq i \leq m)$ denotes the total number of S's which precede directly the occurrences of F_i but do not belong to any S-run of length k. The distribution of each one of the above mentioned random vectors is also obtained (see Propositions 24.2.1 and 24.2.2). We also obtain distributions of random vectors whose components are numbers of overlapping S-runs of length l, and S-runs of length at least l which precede directly the occurrences of F_i but do not belong to any S-run of length k (see Propositions 24.2.3 and 24.2.4). Furthermore, the joint distribution of the numbers of non-overlapping F_i-runs of length k_i $(1 \leq i \leq m)$ is obtained (see Proposition 24.2.5). In Section 24.3, we treat the majority of the above problems in the case of the generalized sequence of order k and we establish corresponding results regarding the $MENB_k(r; q_{11}, \ldots, q_{mk})$ (see Theorem 24.3.1 and Proposition 24.3.1).

The results of Sections 24.2 and 24.3, for $r = 1$, may be specialized to new results regarding the multivariate geometric distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988) and the multivariate extended geometric distribution of order k of Philippou and Antzoulakos (1990), respectively. We shall not discuss this specialization here, however, due to lack of space. We mention that our present paper generalizes several results of Aki and Hirano (1994, 1995), who employed other methods for their derivations. Also see Antzoulakos and Philippou (1995).

In order to avoid unnecessary repetitions, we mention here that in this paper x_{ij} $(1 \leq i \leq m$ and $1 \leq j \leq k)$ are non-negative integers as specified, and m, k, r and l are fixed positive integers, and $l \leq k-1$. In addition, whenever sums and products are taken over i and j, ranging from 1 to m and from 1 to k, respectively, we shall omit these limits for notational simplicity.

24.2 Independent Trials

Consider a sequence of independent trials each of which has $m+1$ mutually exclusive outcomes S, F_1, F_2, $\ldots$, F_m, where $\Pr[S] = p$ and $\Pr[F_i] = q_i$, $1 \leq i \leq m$ $(0 < q_i < 1,\ 0 < \sum_i q_i < 1$ and $p = 1 - \sum_i q_i)$. We call success and failure of type i the outcomes S and F_i, respectively. In the following theorem, we obtain a new genesis scheme of $MNB_k(r;\ q_1,\ q_2,\ \ldots,\ q_m)$.

Theorem 24.2.1 *Consider the above sequence of independent trials until the occurrence of the r-th non-overlapping success run of length k. Let Y_i and*

Z_i $(1 \leq i \leq m)$ be random variables denoting, respectively, the number of failures of type i and the total number of successes which precede directly the occurrences of failures of type i but do not belong to any success run of length k. Set $X_i = Y_i + Z_i$ $(1 \leq i \leq m)$. Then, the random vector $\mathbf{X}$ is distributed as $MNB_k(r;\, q_1,\, q_2,\, \ldots,\, q_m)$, i.e., for $x_i = 0, 1, 2, \ldots,$ $1 \leq i \leq m$,

$$\Pr[\mathbf{X} = \mathbf{x}] = p^{kr} \sum_{\Sigma_j j x_{ij} = x_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r-1} \Pi_i p^{x_i} \left(\frac{q_i}{p}\right)^{\sum_j x_{ij}}$$

PROOF. For any fixed non-negative integers $x_1, x_2, \ldots, x_m$, a typical element of the event $(\mathbf{X} = \mathbf{x})$ is an arrangement,

$$\alpha_1 \alpha_2 \ldots \alpha_{r-1+\sum_i \sum_j x_{ij}} \underbrace{SS \ldots S}_{k}, \tag{24.1}$$

of the outcomes S, F_1, F_2, $\ldots$, F_m such that $r-1$ of the α's are of the form $e_k = \underbrace{SS \ldots S}_{k}$, x_{ij} of the α's are of the form $e_{ij} = \underbrace{SS \ldots S}_{j-1} F_i$ $(1 \leq i \leq m$ and $1 \leq j \leq k)$, and

$$x_i = y_i + z_i = \Sigma_j x_{ij} + \Sigma_j (j-1)\, x_{ij} = \Sigma_j j x_{ij}, \quad i = 1, 2, \ldots, m. \tag{24.2}$$

Fix x_{ij}, $1 \leq i \leq m$ and $1 \leq j \leq k$ (r is fixed). Then, the number of the above arrangements is

$$\binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r-1}, \tag{24.3}$$

and each one of them has probability

$$p^{kr} \Pi_i \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} = \Pi_i p^{x_i} \left(\frac{q_i}{p}\right)^{\sum_j x_{ij}}, \tag{24.4}$$

by the independence of the trials, the definition of e_{ij} and e_k, $\Pr[F_i] = q_i$ $(1 \leq i \leq m)$ and $\Pr[S] = p$. The theorem then follows directly from (24.2)–(24.4). ∎

Remark 24.2.1 For $m = 1$, Theorem 24.2.1 reduces to a variant of Theorem 3.1(a) of Philippou (1984) regarding the genesis of the (suitably shifted) negative binomial distribution of order k, and for $r = 1$ it provides a new type of genesis of the multivariate geometric distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988).

In the following proposition, we obtain the distribution of the random vector $\mathbf{Y}$.

Proposition 24.2.1 *Consider the sequence of independent trials and the random variables Y_i $(1 \leq i \leq m)$ as in Theorem 24.2.1. Then, the random vector $\mathbf{Y}$ is distributed as $MNB_1(r; \hat{q}_1,\, \hat{q}_2,\, \ldots,\, \hat{q}_m)$, where $\hat{q}_i = q_i \left(1 - p^k\right)(1-p)^{-1}$, $1 \leq i \leq m$.*

PROOF. For any fixed non-negative integers $y_1, y_2, \ldots, y_m$, a typical element of the event $(\mathbf{Y} = \mathbf{y})$ is an arrangement as (24.1) with the conditions (24.2) on the non-negative integers x_{ij} substituted by $\sum_j x_{ij} = y_i$, $1 \le i \le m$. Therefore, proceeding along the same lines as those in the proof of Theorem 24.2.1, we have

$$\begin{aligned}
\Pr[\mathbf{Y} = \mathbf{y}] &= p^{kr} \sum_{\sum_j x_{ij} = y_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r - 1} \Pi_i \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} \\
&= p^{kr} \frac{(\sum_i y_i + r - 1)!}{\Pi_i y_i! (r-1)!} \sum_{\sum_j x_{ij} = y_i} \Pi_i \frac{(\Sigma_j x_{ij})!}{\Pi_j x_{ij}!} \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} \\
&= p^{kr} \frac{(\sum_i y_i + r - 1)!}{\Pi_i y_i! (r-1)!} \Pi_i \sum_{\sum_j x_{ij} = y_i} \frac{(\Sigma_j x_{ij})!}{\Pi_j x_{ij}!} \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} \\
&= p^{kr} \frac{(\sum_i y_i + r - 1)!}{\Pi_i y_i! (r-1)!} \Pi_i \left(\Sigma_j q_i p^{j-1}\right)^{y_i},
\end{aligned}$$

by the multinomial theorem. The last relation establishes the proposition. ■

In the following proposition, we obtain the distribution of the random vector $\mathbf{Z}$.

Proposition 24.2.2 *Consider the sequence of independent trials and the random variables Z_i $(1 \le i \le m)$ as in Theorem 24.2.1. Then, the random vector $\mathbf{Z}$ is distributed as $MNB_{k-1}(r; q_1, q_2, \ldots, q_m)$.*

PROOF. For any fixed non-negative integers $z_1, z_2, \ldots, z_m$, a typical element of the event $(\mathbf{Z} = \mathbf{z})$ is an arrangement as (24.1) with the conditions (24.2) on the non-negative integers x_{ij} substituted by $\sum_{j=1}^{k-1} j x_{i,j+1} = z_i$, $1 \le i \le m$. Therefore, proceeding along the same lines as those in the proof of Theorem 24.2.1, we have

$$\begin{aligned}
\Pr[\mathbf{Z} = \mathbf{z}] &= p^{kr} \sum_{\mathbf{1}} \sum_{\sum_{j=1}^{k-1} j x_{i,j+1} = z_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r - 1} \Pi_i \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} \\
&= p^{kr} \sum_{\sum_{j=1}^{k-1} j x_{i,j+1} = z_i} \binom{\sum_i \sum_{j=1}^{k-1} x_{i,j+1} + r - 1}{x_{12}, \ldots, x_{1k}, \ldots, x_{m2}, \ldots, x_{mk}, r - 1} \\
&\quad \times \Pi_i \Pi_{j=1}^{k-1} \left(q_i p^j\right)^{x_{i,j+1}} \sum_{\mathbf{1}} \Pi_i \binom{\gamma_1 + \sum_{n=1}^{i} x_{n1}}{x_{i1}} q_i^{x_{i1}},
\end{aligned} \tag{24.5}$$

where $\gamma_1 = r - 1 + \sum_i \sum_{j=1}^{k-1} x_{i,j+1}$ and $\sum_{\mathbf{1}}$ means sum over all non-negative integers x_{i1}, $1 \le i \le m$. Now, using successively the identity

$$\sum_{j=0}^{\infty} \binom{\gamma + j}{j} x^j = (1 - x)^{-(\gamma+1)}, \quad |\, x \,| < 1, \tag{24.6}$$

[see, for example, Abramowitz and Stegum (1965, p. 822)], we have that

$$\sum_{\mathbf{1}} \Pi_i \binom{\gamma_1 + \sum_{n=1}^{i} x_{n1}}{x_{i1}} q_i^{x_{i1}} = (1 - \sum_i q_i)^{-(\gamma_1+1)} = p^{-\left(r + \sum_i \sum_{j=1}^{k-1} x_{i,j+1}\right)}. \tag{24.7}$$

Therefore, from (24.5) and (24.7), we get

$$\begin{aligned}\Pr[\mathbf{Z} = \mathbf{z}] &= p^{(k-1)r} \sum_{\sum_{j=1}^{k-1} j x_{i,j+1} = z_i} \binom{\sum_i \sum_{j=1}^{k-1} x_{i,j+1} + r - 1}{x_{12}, \ldots, x_{1k}, \ldots, x_{m2}, \ldots, x_{mk}, r-1} \\ &\quad \times \Pi_i \Pi_{j=1}^{k-1} \left(q_i p^{j-1}\right)^{x_{i,j+1}},\end{aligned}$$

and this establishes the proposition. ■

In the following proposition, we are dealing with overlapping success runs of length l ($l \leq k-1$). This counting scheme of runs was proposed by Ling (1988) [also see Fu and Koutras (1994), Godbole (1992) and Hirano *et al.* (1991)]. Consider the sequence of independent trials as in Theorem 24.2.1 and let U_i ($1 \leq i \leq m$) be random variables denoting the total number of overlapping success runs of length l which precede directly the occurrences of failures of type i but do not belong to any success run of length k. As an example, for the case $k = 5$, $r = 2$, $m = 2$ and $l = 2$, we give the sequence $SSSSSF_1SSSF_2SF_1SSSF_1SSSSF_2SSSSS$, where $U_1 = 2$ and $U_2 = 5$.

Proposition 24.2.3 *The random vector* $\mathbf{U}$ *is distributed as* $MNB_{k-l}(r;\, q_1,\, q_2,\, \ldots,\, q_m)$.

PROOF. For any fixed non-negative integers $u_1, u_2, \ldots, u_m$, a typical element of the event ($\mathbf{U} = \mathbf{u}$) is an arrangement as (24.1) with the conditions (24.2) on the non-negative integers x_{ij} substituted by $\sum_{j=1}^{k-l} j x_{i,j+l} = u_i$, $1 \leq i \leq m$. Therefore, proceeding along the same lines as those in the proof of Theorem 24.2.1, we have

$$\begin{aligned}\Pr[\mathbf{U} = \mathbf{u}] &= p^{kr} \sum_{\mathbf{2}} \sum_{\sum_{j=1}^{k-l} j x_{i,j+l} = u_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r-1} \Pi_i \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}} \\ &= p^{kr} \sum_{\sum_{j=1}^{k-l} j x_{i,j+l} = u_i} \binom{\sum_i \sum_{j=1}^{k-l} x_{i,j+l} + r - 1}{x_{1,l+1}, \ldots, x_{1k}, \ldots, x_{m,l+1}, \ldots, x_{mk}, r-1} \\ &\quad \times \Pi_i \Pi_{j=1}^{k-l} \left(q_i p^{j+l-1}\right)^{x_{i,j+l}} \\ &\quad \times \sum_{\mathbf{2}} \Pi_i \Pi_{j=1}^{l} \binom{\gamma_2 + \delta_i + \sum_{n=1}^{j} x_{in}}{x_{ij}} \left(q_i p^{j-1}\right)^{x_{ij}},\end{aligned} \tag{24.8}$$

where $\gamma_2 = r - 1 + \sum_i \sum_{j=1}^{k-l} x_{i,j+l}$, $\delta_1 = 0$, $\delta_i = \sum_{n=1}^{i-1} \sum_{j=1}^{l} x_{nj}$ for $i \geq 2$, and $\sum_{\mathbf{2}}$ means sum over all non-negative integers x_{ij}, $1 \leq i \leq m$ and $1 \leq j \leq l$. Using successively (24.6), we get

$$\begin{aligned}\sum_{\mathbf{2}} \Pi_i \Pi_{j=1}^{l} \binom{\gamma_2 + \delta_i + \sum_{n=1}^{j} x_{in}}{x_{ij}} \left(q_i p^{j-1}\right)^{x_{ij}} &= \left(1 - \sum_i \sum_{j=1}^{l} q_i p^{j-1}\right)^{-(\gamma_2+1)} \\ &= p^{-l\left(r + \sum_i \sum_{j=1}^{k-l} x_{i,j+l}\right)}. \end{aligned} \tag{24.9}$$

Therefore, from (24.8) and (24.9), we get

$$\begin{aligned} &\Pr[\mathbf{U} = \mathbf{u}] \\ &= p^{(k-l)r} \sum_{\sum_{j=1}^{k-l} j x_{i,j+l} = u_i} \binom{\sum_i \sum_{j=1}^{k-l} x_{i,j+l} + r - 1}{x_{1,l+1}, \ldots, x_{1k}, \ldots, x_{m,l+1}, \ldots, x_{mk}, r-1} \\ &\quad \times \Pi_i \Pi_{j=1}^{k-l} \left(q_i p^{j-1}\right)^{x_{i,j+l}}, \end{aligned}$$

and this establishes the proposition. ■

Remark 24.2.2 Proposition 24.2.3, for $l = 1$, reduces to Proposition 24.2.2.

In the following proposition, we are dealing with success runs of length at least l ($l \leq k-1$). This counting scheme of runs is of great statistical importance [see, for example, Gibbons (1971) and Goldstein (1990)]. Consider the sequence of independent trials as in Theorem 24.2.1 and let W_i ($1 \leq i \leq m$) be random variables denoting the total number of success runs of length at least l which precede directly the occurrences of failures of type i but do not belong to any success run of length k. As an example, for the case $k = 5$, $r = 2$, $m = 2$ and $l = 2$, we give the sequence $SSSSSF_1SSSSF_2SF_1SSF_1SSSF_2SSSSS$, where $W_1 = 1$ and $W_2 = 2$.

Proposition 24.2.4 *The random vector* $\mathbf{W}$ *is distributed as* $MNB_1(r;\, \bar{q}_1, \bar{q}_2, \ldots, \bar{q}_m)$, *where* $\bar{q}_i = q_i \left(1 - p^{k-l}\right)(1-p)^{-1}$, $1 \leq i \leq m$.

PROOF. For any fixed non-negative integers $w_1, w_2, \ldots, w_m$, a typical element of the event $(\mathbf{W} = \mathbf{w})$ is an arrangement as (24.1) with the conditions (24.2) on the non-negative integers x_{ij} substituted by $\sum_{j=l+1}^{k} x_{ij} = w_i$, $1 \leq i \leq m$. Therefore, proceeding along the same lines as those in the proof of Theorem 24.2.1, we have

$$\Pr[\mathbf{W} = \mathbf{w}] = p^{kr} \sum_{\mathbf{3}} \sum_{\sum_{j=l+1}^{k} x_{ij} = w_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r-1} \Pi_i \Pi_j \left(q_i p^{j-1}\right)^{x_{ij}},$$

where $\sum_3$ means sum over all non-negative integers x_{ij}, $1 \leq i \leq m$ and $1 \leq j \leq l$. Following the proof of Proposition 24.2.3, we can easily show that

$$\begin{aligned}
&\Pr[\mathbf{W} = \mathbf{w}] \\
&= p^{(k-l)r} \sum_{\sum_{j=l+1}^{k} x_{ij} = w_i} \binom{\sum_i \sum_{j=l+1}^{k} x_{ij} + r - 1}{x_{1,l+1}, \ldots, x_{1k}, \ldots, x_{m,l+1}, \ldots, x_{mk}, r-1} \\
&\quad \times \Pi_i \Pi_{j=l+1}^{k} \left(q_i p^{j-l-1}\right)^{x_{ij}}.
\end{aligned}$$

Next, using the proof of Proposition 24.2.1 and the multinomial theorem, we get

$$\begin{aligned}
\Pr[\mathbf{W} = \mathbf{w}] &= p^{(k-l)r} \frac{(\sum_i w_i + r - 1)!}{\Pi_i w_i!(r-1)!} \Pi_i \sum_{\sum_{j=l+1}^{k} x_{ij} = w_i} \frac{\left(\sum_{j=l+1}^{k} x_{ij}\right)!}{\Pi_j x_{ij}!} \\
&\quad \times \Pi_{j=l+1}^{k} \left(q_i p^{j-l-1}\right)^{x_{ij}} \\
&= p^{(k-l)r} \frac{(\sum_i w_i + r - 1)!}{\Pi_i w_i!(r-1)!} \Pi_i \left(\sum_{j=l+1}^{k} q_i p^{j-l-1}\right)^{w_i}.
\end{aligned}$$

The last relation establishes the proposition. ∎

Corollary 24.2.1 *Consider the sequence of independent trials as in Theorem 24.2.1 and let Y, Z, U and W be random variables denoting, respectively, the number of failures of any type, the number of successes, the number of overlapping success runs of length l, and the number of success runs of length at least l, which do not belong to any success run of length k. Then,*

(a) Y is distributed as $NB_1(r; p^k)$;

(b) Z is distributed as $NB_{k-1}(r; p)$;

(c) U is distributed as $NB_{k-l}(r; p)$;

(d) W is distributed as $NB_1(r; p^{k-l})$.

PROOF. It is an immediate consequence of Propositions 24.2.1–24.2.4 by noting that (a) $Y = \sum_i Y_i$, $Z = \sum_i Z_i$, $U = \sum_i U_i$, and $W = \sum_i W_i$ and (b) if $\mathbf{X}$ is distributed as $MNB_k(r; q_1, q_2, \ldots, q_m)$, then $\sum_i X_i$ is distributed as $NB_k(r; p)$, where $p = 1 - \sum_i q_i$ [see, Philippou, Antzoulakos and Tripsiannis (1990)]. ∎

Next, we obtain the joint distribution of the numbers of non-overlapping runs of failures of type i of length k_i ($1 \leq i \leq m$ and $k_i \geq 1$).

Proposition 24.2.5 *Consider the sequence of independent trials as in Theorem 24.2.1 and let V_i ($1 \leq i \leq m$) be random variables denoting the number* *of non-overlapping runs of failures of type i of length k_i ($k_i \geq 1$). Then, the random vector* $\mathbf{V}$ *is distributed as* $MNB_1(r; \tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_m)$, *where*

$$\tilde{q}_i = \frac{q_i^{k_i}\,(1-q_i)\left(1-q_i^{k_i}\right)^{-1}}{1-\left[(p-p^k)\,(1-p^k)^{-1}\right]-\sum_i\left(q_i-q_i^{k_i}\right)\left(1-q_i^{k_i}\right)^{-1}}, \quad i=1,2,\ldots,m.$$

PROOF. We denote by E_0 and E_i ($1 \leq i \leq m$) a non-overlapping success run of length k and a non-overlapping run of failures of type i of length k_i, respectively. Let W_r be a random variable denoting the number of trials until the r-th occurrence of E_0. Let $p_n(h)$, $0 \leq n \leq m$, be the probability of the event that at the h-th trial the sooner event between $E_0, E_1, \ldots, E_m$ occurs and the sooner event is E_n. Then,

$$\left\{\begin{array}{l}
\Pr[\mathbf{V}=\mathbf{v}, W_r=w]=0 \text{ if some } v_i<0\ (1\leq i\leq m), \text{ or } w<kr,\\
\Pr[\mathbf{V}=\mathbf{v}, W_r=kr]=0 \text{ if some } v_i\neq 0\ (1\leq i\leq m),\\
\Pr[\mathbf{V}=\mathbf{0}, W_r=kr]=p^{kr},\\
\Pr[\mathbf{V}=\mathbf{v}, W_r=w]=\displaystyle\sum_{h=k}^{w-k(r-1)} p_0(h)\Pr[\mathbf{V}=\mathbf{v}, W_{r-1}=w-h]\\
\qquad\qquad\qquad\qquad\quad +\ \displaystyle\sum_i\sum_{h=k_i}^{w-kr} p_i(h)\Pr[\mathbf{V}=\mathbf{v}-\mathbf{1}_i, W_r=w-h],
\end{array}\right\} \tag{24.10}$$

where $\mathbf{0}=(0,0,\ldots,0)$, and $\mathbf{1}_i$ ($1\leq i\leq m$) denotes the m-dimensional vector with a "1" at the i-th entry and "0" elsewhere. Set

$$g_0(t)=\sum_{h=k}^{\infty} p_0(h)t^h \text{ and } g_i(t)=\sum_{h=k_i}^{\infty} p_i(h)t^h, \quad i=1,2,\ldots,m, \tag{24.11}$$

and

$$G_r(\mathbf{s},t)=\sum_{v_1=0}^{\infty}\cdots\sum_{v_m=0}^{\infty}\sum_{w=kr}^{\infty}\Pr[\mathbf{V}=\mathbf{v}, W_r=w]t^w\Pi_i s_i^{v_i}. \tag{24.12}$$

Using (24.10)–(24.12), we get

$$\begin{aligned}
G_r(\mathbf{s},t) &= \sum_{v_1=0}^{\infty}\cdots\sum_{v_m=0}^{\infty}\sum_{w=kr}^{\infty}\sum_{h=k}^{w-k(r-1)} p_0(h)\Pr[\mathbf{V}=\mathbf{v}, W_{r-1}=w-h]t^w\Pi_i s_i^{v_i}\\
&\quad+\sum_i\sum_{v_1=0}^{\infty}\cdots\sum_{v_m=0}^{\infty}\sum_{w=kr}^{\infty}\sum_{h=k_i}^{w-kr} p_i(h)\Pr[\mathbf{V}=\mathbf{v}-\mathbf{1}_i, W_r=w-h]\\
&\quad\times t^w\Pi_i s_i^{v_i}\\
&= \sum_{h=k}^{\infty} p_0(h)t^h\sum_{v_1=0}^{\infty}\cdots\sum_{v_m=0}^{\infty}\sum_{w=k(r-1)+h}^{\infty}\Pr[\mathbf{V}=\mathbf{v}, W_{r-1}=w-h]
\end{aligned}$$

$$\times\ t^{w-h}\Pi_i s_i^{v_i}$$
$$+ \sum_i \sum_{h=k_i}^{\infty} p_i(h)t^h \sum_{v_1=0}^{\infty} \cdots \sum_{v_m=0}^{\infty} \sum_{w=kr+h}^{\infty} \Pr[\mathbf{V} = \mathbf{v} - \mathbf{1}_i,$$
$$W_r = w - h]t^{w-h}\Pi_i s_i^{v_i}$$
$$= g_0(t)G_{r-1}(\mathbf{s}, t) + G_r(\mathbf{s}, t) \sum_i s_i g_i(t), \tag{24.13}$$

and

$$G_1(\mathbf{s}, t) = g_0(t) \left(1 - \sum_i s_i g_i(t)\right)^{-1}. \tag{24.14}$$

From (24.13) and (24.14), we get

$$G_r(\mathbf{s}, t) = \left(\frac{g_0(t)}{1 - \sum_i s_i g_i(t)}\right)^r. \tag{24.15}$$

Eq. (24.15) implies that $\mathbf{V}$ is distributed as $MNB_1(r; \tilde{q}_1, \tilde{q}_2, \ldots, \tilde{q}_m)$, where $\tilde{q}_i = g_i(1)$, $1 \le i \le m$. By setting $t = 1$, $x_i = 1$ and $x_j = 0$ $(0 \le j \ne i < \infty)$ in the expression of $\phi(t)$ in Theorem 2.1 of Aki (1992), we obtain

$$g_i(1) = \frac{q_i^{k_i}(1 - q_i)\left(1 - q_i^{k_i}\right)^{-1}}{1 - \left[(p - p^k)(1 - p^k)^{-1}\right] - \sum_i \left(q_i - q_i^{k_i}\right)\left(1 - q_i^{k_i}\right)^{-1}}, \quad i = 1, 2, \ldots, m,$$

and this establishes the proposition. ■

Remark 24.2.3 Propositions 24.2.1–24.2.5, for $r = 1$, yield new results regarding the multivariate geometric distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988). Also, Propositions 24.2.1–24.2.5 and Corollary 24.2.1, for $m = 1$ and $r = 1$, yield respective results of Aki and Hirano (1994), and Corollary 24.2.1 for $r = 1$ reduces to respective results of Aki and Hirano (1995).

24.3 Generalized Sequence of Order k

In this section, the majority of the problems of Section 24.2 are treated in the case of the generalized sequence of order k of Philippou and Antzoulakos (1990). This sequence is an extension of independent trials with multiple outcomes and may be suitable for considering succession events in practical situations where independence of trials cannot be assumed. We recall first the definition of this sequence from Philippou and Antzoulakos (1990).

Definition 24.3.1 An infinite sequence $\{T_n\}_{n=0}^{\infty}$ of $\{0,1,2,\ldots,m\}$-valued random variables is said to be the generalized sequence of order k with parameters q_{ij}, $0 < q_{ij} < 1$ for $1 \le i \le m$ and $1 \le j \le k$, and $\sum_i q_{ij} < 1$, if

(a) $T_0 \neq 0$ almost surely and

(b) $\Pr[T_n = i \mid T_{n-1} = t_{n-1}, \ldots, T_1 = t_1, T_0 = t_0] = q_{ij},$
$$i = 1,2,\ldots,m,\; n \ge 1,$$

where $j = c - k[(c-1)/k]$, c is the smallest integer which satisfies $t_{n-c} \neq 0$, and $[x]$ denotes the greatest integer in x.

We call success and failure of type i the outcomes $0,1,2,\ldots,m$, respectively. A direct consequence of the above definition is that

$$\Pr[T_n = 0 \mid T_{n-1} = t_{n-1}, \ldots, T_1 = t_1, T_0 = t_0] = p_j = 1 - \textstyle\sum_i q_{ij},$$

where j is as above, and that the conditional distribution of $T_{n+1}, T_{n+2}, \ldots$ given that $T_n \neq 0$ is equal to the distribution of $T_1, T_2, \ldots$. Also, the generalized sequence of order k, for $m = 1$, reduces to the binary sequence of order k of Aki (1985) [also see Dhar and Jiang (1995)].

In the following theorem, a new genesis scheme of $MENB_k(r; q_{11}, \ldots, q_{mk})$ is obtained.

Theorem 24.3.1 *Consider the above generalized sequence of order k until the occurrence of the r-th non-overlapping success run of length k, and let X_i $(1 \le i \le m)$ be as in Theorem 24.2.1. Then, the random vector* $\mathbf{X}$ *is distributed as $MENB_k(r;\, q_{11}, \ldots, q_{mk})$, i.e., for $x_i = 0,1,2,\ldots$, $1 \le i \le m$,*

$$\begin{aligned}\Pr[\mathbf{X} = \mathbf{x}] &= (p_1 p_2 \cdots p_k)^r \sum_{\sum_j j x_{ij} = x_i} \binom{\sum_i \sum_j x_{ij} + r - 1}{x_{11}, \ldots, x_{mk}, r-1} \\ &\quad \times \Pi_i \Pi_j \left(p_0 p_1 p_2 \cdots p_{j-1} q_{ij}\right)^{x_{ij}},\end{aligned}$$

with the convention that $p_0 = 1$.

PROOF. For any fixed non-negative integers $x_1, x_2, \ldots, x_m$, a typical element of the event $(\mathbf{X} = \mathbf{x})$ is an arrangement

$$\alpha_1 \alpha_2 \ldots \alpha_{r-1+\sum_i \sum_j x_{ij}} \underbrace{00\ldots0}_{k},$$

of the outcomes $0,1,2,\ldots,m$ such that $r-1$ of the α's are of the form $e_k = \underbrace{00\ldots0}_{k}$, x_{ij} of the α's are of the form $e_{ij} = \underbrace{00\ldots0}_{j-1} i$, and $\sum_j j x_{ij} = x_i$, $(1 \le i \le m$ and $1 \le j \le k)$.

Fix x_{ij}, $1 \le i \le m$ and $1 \le j \le k$ (r is fixed). Definition 24.3.1 implies that (i) the probability of the pattern e_{ij} is $p_0 p_1 p_2 \cdots p_{j-1} q_{ij}$ since the previously first occurring failure of any type is $1 + sk$ $(0 \le s \le r-1)$ places before it, and (ii) the probability of the pattern e_k is always $p_1 p_2 \cdots p_k$. Therefore the probability of each one of the above arrangements is

$$(p_1 p_2 \cdots p_k)^r \, \Pi_i \Pi_j \, (p_0 p_1 p_2 \cdots p_{j-1} q_{ij})^{x_{ij}} .$$

The theorem then can be easily established by means of the proof of Theorem 24.2.1. ■

Remark 24.3.1 Theorem 24.3.1, for $r = 1$, gives a new proof of Proposition 3.1 of Philippou and Antzoulakos (1990) regarding the genesis of the multivariate extended geometric distribution of order k. For $m = 1$, it gives a new genesis scheme of the (suitably shifted) extended negative binomial distribution of order k of Aki (1985).

In the sequel, we obtain results for the generalized sequence of order k which correspond to those of Propositions 24.2.1–24.2.4 of Section 24.2.

Proposition 24.3.1 *Consider the generalized sequence of order k as in Theorem 24.3.1 and the random variables Y_i, Z_i, U_i and W_i $(1 \le i \le m)$ as in Propositions 24.2.1–24.2.4, respectively. Then,*

(a) $\mathbf{Y}$ *is distributed as* $MNB_1(r; \hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m)$;

(b) $\mathbf{Z}$ *is distributed as* $MENB_{k-1}(r; q_{12}, \ldots, q_{1k}, \ldots, q_{m2}, \ldots, q_{mk})$;

(c) $\mathbf{U}$ *is distributed as* $MENB_{k-l}(r; q_{1,l+1}, \ldots, q_{1k}, \ldots, q_{m,l+1}, \ldots, q_{mk})$;

(d) $\mathbf{W}$ *is distributed as* $MNB_1(r; \bar{Q}_1, \bar{Q}_2, \ldots, \bar{Q}_m)$;

where $\hat{Q}_i = \sum_j p_0 p_1 p_2 \cdots p_{j-1} q_{ij}$, *and* $\bar{Q}_i = \sum_{j=l+1}^{k} p_{l+1} p_{l+2} \cdots p_{j-1} q_{ij}$, $1 \le i \le m$, *with the convention that* $p_0 = 1$ *and for* $j = l+1$, $p_{l+1} p_{l+2} \cdots p_{j-1} = 1$.

PROOF. The proof follows along the same lines as those in the proofs of Propositions 24.2.1–24.2.4. ■

Corollary 24.3.1 *Consider the generalized sequence of order k as in Theorem 24.3.1 and let Y, Z, U and W be random variables denoting, respectively, the number of failures of any type, the number of successes, the number of overlapping success runs of length l, and the number of success runs of length at least l, which do not belong to any success run of length k. Then,*

(a) Y *is distributed as* $NB_1(r; p_1 p_2 \cdots p_k)$;

(b) Z *is distributed as* $ENB_{k-1}(r; p_2, p_3, \ldots, p_k)$;

(c) U *is distributed as* $ENB_{k-l}(r; p_{l+1}, p_{l+2}, \ldots, p_k)$;

(d) W *is distributed as* $NB_1(r; p_{l+1} p_{l+2} \cdots p_k)$.

Remark 24.3.2 The generalized sequence of order k reduces to the case of independent trials of Section 24.2 if we set $q_{ij} = q_i$, $1 \le i \le m$ and $1 \le j \le k$, which implies $p_1 = p_2 = \ldots = p_k = p = 1 - \sum_i q_i$. So, the corresponding results of Section 24.2 can be regarded as corollaries of those in this section. Also, Proposition 24.3.1, for $r = 1$, gives new results regarding the multivariate extended geometric distribution of order k of Philippou and Antzoulakos (1990), and Corollary 24.3.1, for $r = 1$, gives respective results of Aki and Hirano (1995).

References

1. Abramowitz, M. and Stegun, I. A. (1965). *Handbook of Mathematical Functions*, New York: Dover.

2. Aki, S. (1985). Discrete distributions of order k on a binary sequence, *Annals of the Institute of Statistical Mathematics*, **37**, 205–224.

3. Aki, S. (1992). Waiting time problems for a sequence of discrete random variables, *Annals of the Institute of Statistical Mathematics*, **44**, 363–378.

4. Aki, S., Balakrishnan, N. and Mohanty, S. G. (1996). Sooner and later waiting time problems for success and failure runs in higher order Makov dependent trials, *Annals of the Institute of Statistical Mathematics*, **48**, 773–787.

5. Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes, *Annals of the Institute of Statistical Mathematics*, **46**, 193–202.

6. Aki, S. and Hirano, K. (1995). Joint distributions of number of success-runs and failures until the first consecutive k successes, *Annals of the Institute of Statistical Mathematics*, **47**, 225–235.

7. Aki, S., Kuboki, H., and Hirano, K. (1984). On discrete distributions of order k, *Annals of the Institute of Statistical Mathematics*, **36**, 431–440.

8. Antzoulakos, D. L. and Philippou, A. N. (1991). A note on the multivariate negative binomial distributions of order k, *Communications in Statistics—Theory and Methods*, **20**, 1389–1399.

9. Antzoulakos, D. L. and Philippou, A. N. (1994). Expressions in terms of binomial coefficients for some multivariate distributions of order k, In *Runs and Patterns in Probability: Selected Papers* (Eds., A. P. Godbole and S. G. Papastavridis), pp. 1–14, Dordrecht: Kluwer Academic Publishers.

10. Antzoulakos, D. L. and Philippou, A. N. (1995). Distributions of the numbers of successes and failures until the r-th overlapping and non-overlapping success run of length k, *Submitted for publication.*

11. Balakrishnan, N., Mohanty, S. G. and Aki, S. (1997). Start-up demonstration tests under Markov dependence model with corrective action, *Annals of the Institute of Statistical Mathematics* (to appear).

12. Balasubramanian, K., Viveros, R. and Balakrishnan, N. (1993). Sooner and later waiting time problems for Markovian Bernoulli trials, *Statistics & Probability Letters*, **18**, 153–161.

13. Charalambides, C. A. (1986). On discrete distributions of order k, *Annals of the Institute of Statistical Mathematics*, **38**, 557–568.

14. Dhar, K. S. and Jiang, X. (1995). Probability bounds on the finite sum of the binary sequence of order k, *Journal of Applied Probability*, **32**, 1014–1027.

15. Ebneshahrashoob, M. and Sobel, M. (1990). Sooner and later waiting time problems for Bernoulli trials: frequency and run quotas, *Statistics & Probability Letters*, **9**, 5–11.

16. Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **89**, 1050–1058.

17. Gibbons, J. D. (1971). *Nonparametric Statistical Inference*, New York: McGraw-Hill.

18. Godbole, A. P. (1990). Specific formulae for some success run distributions, *Statistics & Probability Letters*, **10**, 119–124.

19. Godbole, A. P. (1992). The exact and asymptotic distribution of overlapping success runs, *Communications in Statistics—Theory and Methods*, **21**, 953–967.

20. Goldstein, L. (1990). Poisson approximations and DNA sequence matching, *Communications in Statistics—Theory and Methods*, **19**, 4167–4179.

21. Hirano, K. (1986). Some properties of the distributions of order k, In *Fibonacci Numbers and Their Applications* (Eds., A. N. Philippou, G. E. Bergum and A. F. Horadam), pp. 43–53, Dordrecht: D. Reidel.

22. Hirano, K., Aki, S., Kashiwagi, N. and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order k, *Statistics & Probability Letters*, **11**, 503–509.

23. Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

24. Ling, K. D. (1988). On binomial distributions of order k, *Statistics & Probability Letters*, **6**, 247–250.

25. Ling, K. D. and Tai, T. H. (1990). On bivariate binomial distributions of order k, *Soochow Journal of Mathematics*, **16**, 211–220.

26. Mohanty, S. G. (1994). Success runs of length k in Markov dependent trials, *Annals of the Institute of Statistical Mathematics*, **46**, 777–796.

27. Panaretos, J. and Xekalaki, E. (1986). On some distributions arising from certain generalized sampling schemes, *Communications in Statistics—Theory and Methods*, **15**, 873–891.

28. Patil, G. P., Boswell, M. T., Joshi, S. W. and Ratnaparkhi, M. V. (1984). *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work—Volume 1*, Fairland: International Co-operative Publishing House.

29. Philippou, A. N. (1983). Poisson and compound Poisson distributions of order k and some of their properties, *Zapiski Nauchnykh Seminarov Leningradskogo Otdeleniya Matematicheskogo Instituta im. V. A. Steklova AN SSSR*, **130**, 175–180 (in Russian, English summary).

30. Philippou, A. N. (1984). The negative binomial distribution of order k and some of its properties, *Biometrical Journal*, **26**, 789–794.

31. Philippou, A. N. (1988). On multiparameter distributions of order k, *Annals of the Institute of Statistical Mathematics*, **40**, 467–475.

32. Philippou, A. N. and Antzoulakos, D. L. (1990). Multivariate distributions of order k on a generalized sequence, *Statistics & Probability Letters*, **9**, 453–463.

33. Philippou, A. N., Antzoulakos, D. L. and Tripsiannis, G. A. (1988). Multivariate distributions of order k, *Statistics & Probability Letters*, **7**, 207–216.

34. Philippou, A. N., Antzoulakos, D. L. and Tripsiannis, G. A. (1990). Multivariate distributions of order k, part II, *Statistics & Probability Letters*, **10**, 29–35.

35. Philippou, A. N., Georghiou, C. and Philippou, G. N. (1983). A generalized geometric distribution and some of its properties, *Statistics & Probability Letters*, **1**, 171–175.

36. Philippou, A. N. and Makri, F. S. (1986). Successes, runs and longest runs, *Statistics & Probability Letters*, **4**, 211–215.

37. Philippou, A. N. and Tripsiannis, G. A. (1991). Multivariate Pólya and inverse Pólya distributions of order k, *Biometrical Journal*, **33**, 225–236.

38. Tripsiannis, G. A. (1993). Modified multivariate Pólya and inverse Pólya distributions of order k, *Journal of the Indian Society of Statistics and Operations Research*, **14**, 1–14.

25

A Multivariate Negative Binomial Distribution of Order k Arising When Success Runs are Allowed to Overlap

Gregory A. Tripsiannis and Andreas N. Philippou

Democritus University of Thrace, Alexandroupolis, Greece
University of Patras, Patras, Greece

Abstract: Ling (1989) introduced and studied a negative binomial distribution of order k, type III, which he denoted by $NB_{k,III}(r,p)$, as the probability distribution of the number of Bernoulli trials $M_r^{(k)}$ until the occurrence of r possibly overlapping success runs of length k [see also Hirano *et al.* (1991)]. In the present paper, independent trials are considered with $m+1$ possible outcomes and the multivariate negative binomial distribution of order k, type III, say $\overline{MNB}_{k,III}(r;q_1,\ldots,q_m)$, is introduced as the distribution of a random vector $\boldsymbol{Y}$ which is a multivariate analogue of $Y = M_r^{(k)} - (k+r-1)$. The probability generating function, mean and variance-covariance, and several distributional properties of $\boldsymbol{Y}$ are established. The present paper generalizes to the multivariate case shifted versions of results of Ling (1989) and Hirano *et al.* (1991) on $NB_{k,III}(r,p)$. Three new results on $NB_{k,III}(r,p)$ or/and its shifted version are derived first; another one arises as a corollary of a proposition on $\overline{MNB}_{k,III}(r;q_1,\ldots,q_m)$.

Keywords and phrases: Multivariate distributions of order k, type III negative binomial, probability generating function, mean, variance-covariance, distributional properties, limiting case

25.1 Introduction

The study of multivariate distributions of order k was initiated by Philippou, Antzoulakos and Tripsiannis (1988, 1990), who introduced and studied the multivariate negative binomial, Poisson, k-point, logarithmic series and modified logarithmic series distributions of order k. These two papers generalized several results of Sibuya, Yoshimura and Shimizu (1964), Patil and Bildikar (1967), Philippou (1986, 1988), Philippou, Georghiou and Philippou (1983), Aki, Kuboki and Hirano (1984), Aki (1985) and Hirano and Aki (1987) on multivariate distributions and distributions of order k.

Since then, a number of papers have appeared dealing with new (or the above) multivariate distributions of order k. Ling and Tai (1990) derived bivariate binomial distributions of order k, while Philippou and Antzoulakos (1990), Philippou and Tripsiannis (1991) and Antzoulakos and Philippou (1994) generalized to the multivariate case the respective work of Aki (1985), Philippou, Tripsiannis and Antzoulakos (1989) and Godbole (1990). One may refer to the recent book of Johnson, Kotz and Balakrishnan (1997) for a comprehensive review of developments on multivariate distributions of order k.

All those of the above-mentioned papers which discuss success runs consider them to be non-overlapping. Ling (1988, 1989) allowed success runs to overlap. In the first paper, he derived and studied the binomial distribution of order k, type II. In the second one, he introduced and studied the negative binomial distribution of order k, type III, denoted by $NB_{k,III}(r,p)$, as the distribution of the number of Bernoulli trials $M_r^{(k)}$ until the occurrence of the r-th possibly overlapping success run of length k. Among other results, he obtained its probability generating function, mean and variance.

In the present paper, we also allow success runs to overlap. In Section 25.2, we reconsider $M_r^{(k)}$ and we establish a new formula for its probability distribution function (see Theorem 25.2.1). Next, we introduce a related random variable Y which lends itself to a multivariate generalization, and we note that it is distributed as the (shifted) negative binomial distribution of order k, type III, say $\overline{NB}_{k,III}(r,p)$ (see Theorem 25.2.2). Finally, we derive the multivariate negative binomial distribution of order k, type III, say $\overline{MNB}_{k,III}(r;q_1,\ldots,q_m)$, as the distribution of the multivariate analogue of $Y = M_r^{(k)} - (k+r-1)$ (see Theorem 25.2.3). In Section 25.3, we first derive the probability generating function, and hence the mean and variance-covariance of $\overline{MNB}_{k,III}(r;q_1,\ldots,q_m)$ (see Theorem 25.3.1), and then we show that $\boldsymbol{Y}$ can be represented as a sum of independent random vectors (see Theorem 25.3.2). We also establish a new genesis scheme of $\overline{NB}_{k,III}(r,p)$ as a corollary to Proposition 25.3.1, and we find the limit of $\overline{MNB}_{k,III}(r;q_1,\ldots,q_m)$ when $q_i \to 0$ and $rq_i \to \lambda_i$ as $r \to \infty$ (see Theorem 25.3.3).

The present paper extends several results of Ling (1989) and Hirano *et al.* (1991) on the shifted versions of type III negative binomial distribution of order k to the multivariate case. It also generalizes several results on the multivariate negative binomial distribution [see, for example, Patil *et al.* (1984) and Johnson, Kotz and Balakrishnan (1997)] to the corresponding results on order k distributions.

In order to avoid unnecessary repetitions, we mention here that in this paper whenever sums and products are taken over i and j ranging from 1 to m and 1 to $k+r-1$, respectively, we shall omit these limits. Also, vectors and random vectors are always m-dimensional vectors.

25.2 Multivariate Negative Binomial Distribution of Order k, Type III

We first derive an exact formula for the probability distribution function of the negative binomial distribution of order k, type III, which is an alternative to that given by Ling (1989).

Theorem 25.2.1 *Let $M_r^{(k)}$ be a random variable denoting the number of Bernoulli trials until the occurrence of the r-th possibly overlapping success run of length k. Then, for $n \geq k+r-1$*

$$\Pr[M_k^{(r)} = n] = \sum_{l=k}^{k+r-1} \sum \binom{n_1 + \cdots + n_{k+r-1}}{n_1, \ldots, n_{k+r-1}} p^n \left(\frac{q}{p}\right)^{n_1+\cdots+n_{k+r-1}},$$

where the inner summation is taken over all non-negative integers $n_1, \ldots, n_{k+r-1}$ such that

$$\sum_{j=1}^{k+r-1} jn_j + l = n \text{ and } \sum_{j=k+1}^{k+r-1} (j-k)n_j + l = k+r-1, \; k \leq l \leq k+r-1. \tag{25.1}$$

Proof. For any fixed non-negative integer n, a typical element of the event $(M_r^{(k)} = n)$ is an arrangement

$$\alpha_1\alpha_2 \ldots \alpha_{n_1+\cdots+n_{k+r-1}} \underbrace{SS\ldots S}_{l}$$

of the letters F (Failure) and S (Success), such that n_j of the α's are $e_j = \underbrace{SS\ldots S}_{j-1}F$ $(1 \leq j \leq k+r-1)$ and they satisfy (25.1).

Fix n_j $(1 \leq j \leq k+r-1)$ (r is fixed). Then the number of such arrangements is

$$\binom{n_1 + \cdots + n_{k+r-1}}{n_1, \ldots, n_{k+r-1}}$$

and each one of them has probability $p^n(q/p)^{n_1+\cdots+n_{k+r-1}}$.

The theorem then readily follows, since the non-negative integers n_j $(1 \leq j \leq k+r-1)$ may vary subject to (25.1). ■

We have the following simple corollary from Theorem 25.2.1.

Corollary 25.2.1 *Let $M_r^{(k)}$ be the random variable as in Theorem 25.2.1, and let $Y = M_r^{(k)} - (k+r-1)$. Then, Y is distributed as the shifted negative binomial distribution of order k, type III, $\overline{NB}_{k,III}(r,p)$, i.e. for $y \geq 0$,*

$$\Pr[Y = y] = \sum_{l=k}^{k+r-1} \sum \binom{y_1 + \cdots + y_{k+r-1}}{y_1, \ldots, y_{k+r-1}} p^{y+k+r-1} \left(\frac{q}{p}\right)^{y_1+\cdots+y_{k+r-1}},$$

where the inner summation is taken over all non-negative integers $y_1, \ldots, y_{k+r-1}$ such that

$$\sum_{j=1}^{k} j y_j + k \sum_{j=k+1}^{k+r-1} y_j = y \text{ and } \sum_{j=k+1}^{k+r-1} (j-k) y_j + l = k + r - 1, \; k \leq l \leq k + r - 1.$$

We now present the following theorem which may be generalized to the multivariate case (as will be seen later).

Theorem 25.2.2 *In a sequence of independent Bernoulli trials with success probability p $(0 < p < 1)$ until the occurrence of the r-th $(r \geq 1)$ possibly overlapping S-run of length k $(k \geq 1)$, consider random variables X and L_j $(1 \leq j \leq X)$ denoting the number of outcomes F and the length of the S-run preceding directly the j-th occurrence of F, respectively. Furthermore, let $\widetilde{L}_j = \min\{L_j, k-1\}$, $L = \widetilde{L}_1 + \cdots + \widetilde{L}_X$, and $Y = X + L$. Then, the random variable Y is distributed as $\overline{NB}_{k,III}(r,p)$.*

PROOF. It may be established along the same lines as those in the proof of Theorem 25.2.1, or by noting that $Y = M_r^{(k)} - (k + r - 1)$. ■

The next result generalizes Theorem 25.2.2 to the multivariate case and provides a genesis scheme for a new multivariate distribution of order k.

Theorem 25.2.3 *In a sequence of independent trials with $m + 1$ possible outcomes $F_1, \ldots, F_m$ and S with probabilities $q_i = \Pr[F_i]$ and $p = \Pr[S]$ $(0 < q_i < 1$ for $1 \leq i \leq m$, $\Sigma_i q_i < 1$ and $p = 1 - \Sigma_i q_i)$ until the occurrence of the r-th $(r \geq 1)$ possibly overlapping S-run of length k $(k \geq 1)$, consider random variables X_i $(1 \leq i \leq m)$ and L_{ij} $(1 \leq i \leq m$ and $1 \leq j \leq X_i)$ denoting the number of outcomes F_i and the length of the S-run preceding directly the j-th occurrence of F_i, respectively. Furthermore, let $\widetilde{L}_{ij} = \min\{L_{ij}, k-1\}$, $L_i = \widetilde{L}_{i1} + \cdots + \widetilde{L}_{iX_i}$, and $Y_i = X_i + L_i$. Then, for $y_i = 0, 1, \ldots$ $(1 \leq i \leq m)$, we have*

$$\Pr[Y_1 = y_1, \ldots, Y_m = y_m] = \sum_{l=k}^{k+r-1} \sum \frac{(\Sigma_i \Sigma_j y_{ij})!}{\Pi_i \Pi_j y_{ij}!} p^{\Sigma_i y_i + k + r - 1} \prod_i \left(\frac{q_i}{p}\right)^{\Sigma_j y_{ij}},$$

where the inner summation is taken over all non-negative integers y_{ij} $(1 \leq i \leq m$ and $1 \leq j \leq k + r - 1)$ such that

$$\sum_{j=1}^{k} j y_{ij} + k \sum_{j=k+1}^{k+r-1} y_{ij} = y_i \qquad (1 \leq i \leq m)$$

and

$$\sum_i \sum_{j=k+1}^{k+r-1} (j-k)y_{ij} + l = k+r-1, \qquad k \le l \le k+r-1. \tag{25.2}$$

PROOF. For any fixed non-negative integers $y_1, \ldots, y_m$, a typical element of the event $(Y_1 = y_1, \ldots, Y_m = y_m)$ is an arrangement

$$\alpha_1\alpha_2 \ldots \alpha_{\Sigma_i\Sigma_j y_{ij}} \underbrace{SS\ldots S}_{l}$$

of the letters $F_1, \ldots, F_m$ and S, such that y_{ij} of the α's are $e_{ij} = \underbrace{SS...S}_{j-1}F_i$ $(1 \le i \le m$ and $1 \le j \le k+r-1)$, and satisfy (25.2).

The proof of the theorem then follows along the same lines as those in the proof of Theorem 25.2.1. ■

For $k = 1$, Theorem 25.2.3 simply reduces to a sampling derivation of the multivariate negative binomial distribution [see, for example, Patil *et al.* (1984, p. 107) and Johnson, Kotz and Balakrishnan (1997)]; further, for $m = 1$, it reduces to the derivation of the (shifted) negative binomial distribution of order k, type III. We therefore introduce the following definition.

Definition 25.2.1 A random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ is said to have the multivariate negative binomial distribution of order k, type III, with parameters $r, q_1, \ldots, q_m$ $(r \ge 1,\ 0 < q_i < 1$ for $1 \le i \le m$, $\Sigma_i q_i < 1$ and $p = 1 - \Sigma_i q_i)$, to be denoted by $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$ if $\Pr[\boldsymbol{Y} = \boldsymbol{y}]$ is given by (25.2) for $y_i = 0, 1, \ldots$ $(1 \le i \le m)$.

We observe that

$$\overline{MNB}_{k,III}(1; q_1, \ldots, q_m) = \overline{MG}_{k,I}(q_1, \ldots, q_m), \tag{25.3}$$

where the latter denotes the multivariate geometric distribution of order k, type I, of Philippou, Antzoulakos and Tripsiannis (1988); also, Theorem 25.2.3 provides a new genesis scheme for this distribution for the case $r = 1$.

25.3 Characteristics and Distributional Properties of $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$

In this section we derive the probability generating function (pgf), the mean and variance-covariance, as well as several distributional properties of the multivariate negative binomial distribution of order k, type III. We first establish the following lemma.

Lemma 25.3.1 *Let $\boldsymbol{Y}^{(r)} = (Y_1^{(r)}, \ldots, Y_m^{(r)})'$, $r \geq 1$, be a random vector distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$, and let $\boldsymbol{W} = (W_1, \ldots, W_m)'$ be a random vector distributed as $\overline{MG}_{k,I}(q_1, \ldots, q_m)$. Then, for $r \geq 2$,*

$$\Pr[\boldsymbol{Y}^{(r)} = \boldsymbol{Y}^{(r-1)}] = p \text{ and } \Pr[\boldsymbol{Y}^{(r)} = \boldsymbol{Y}^{(r-1)} + \boldsymbol{k}_i + \boldsymbol{W}] = q_i \quad (1 \leq i \leq m),$$

where $\boldsymbol{k}_i$ is an m-dimensional vector with its i-th component equal to k and all of its other components equal to 0, and $Y_i^{(r-1)}$ and W_i $(1 \leq i \leq m)$ are independent random variables.

PROOF. Suppose it is given that $\boldsymbol{Y}^{(r-1)} = \boldsymbol{y}^{(r-1)}$. Either the next outcome is S, or it is F_i $(1 \leq i \leq m)$. If the next outcome is S, which occurs with probability p, then $\boldsymbol{Y}^{(r)} = \boldsymbol{y}^{(r-1)}$ or equivalently $\boldsymbol{Y}^{(r)} = \boldsymbol{Y}^{(r-1)}$. Therefore, $\Pr[\boldsymbol{Y}^{(r)} = \boldsymbol{Y}^{(r-1)}] = p$. If the outcome is F_i $(1 \leq i \leq m)$, which occurs with probability q_i, then for $1 \leq s \neq i \leq m$ we have

$$Y_i^{(r)} = y_i^{(r-1)} + k + w_i \text{ and } Y_s^{(r)} = y_s^{(r-1)} + w_s, \quad 1 \leq s \neq i \leq m,$$

or equivalently

$$Y_i^{(r)} = Y_i^{(r-1)} + k + W_i \text{ and } Y_s^{(r)} = Y_s^{(r-1)} + W_s, \quad 1 \leq s \neq i \leq m. \tag{25.4}$$

Therefore, $\Pr[\boldsymbol{Y}^{(r)} = \boldsymbol{Y}^{(r-1)} + \boldsymbol{k}_i + \boldsymbol{W}] = q_i$ $(1 \leq i \leq m)$. We also note that $Y_i^{(r-1)}$ (W_i) depends on the independent trials before (after) the occurrence of the outcome F_i. Therefore, $Y_i^{(r-1)}$ and W_i $(1 \leq i \leq m)$ are independent, and this completes the proof of the lemma. ■

We note that Lemma 25.3.1 generalizes a shifted version of Lemma 3.1 of Ling (1989).

The following result provides a recurrence relation on r for the pgf of the random vector $\boldsymbol{Y}^{(r)}$, $g_{\boldsymbol{Y}^{(r)}}(.)$, an exact formula for $g_{\boldsymbol{Y}^{(r)}}(.)$, and the mean and variance-covariance of $\boldsymbol{Y}^{(r)}$.

Theorem 25.3.1 *Let $\boldsymbol{Y}^{(r)}$, $r \geq 1$, be a random vector distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$ with pgf $g_{\boldsymbol{Y}^{(r)}}(\boldsymbol{t})$. Then, for $r \geq 2$, we have*

(i) $g_{\boldsymbol{Y}^{(r)}}(\boldsymbol{t}) = g_{\boldsymbol{Y}^{(r-1)}}(\boldsymbol{t}) \left(p + g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t}) \Sigma_i q_i t_i^k\right)$, $|t_i| \leq 1$, $1 \leq i \leq m$;

(ii) $$g_{\boldsymbol{Y}^{(r)}}(\boldsymbol{t}) = g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t}) \left(p + g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t}) \Sigma_i q_i t_i^k\right)^{r-1} = \frac{p^k}{1 - \Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \left(p + \frac{p^k}{1 - \Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \Sigma_i q_i t_i^k\right)^{r-1},$$ $|t_i| \leq 1$, $1 \leq i \leq m$;

(iii) $E[Y_i^{(r)}] = \frac{q_i}{p^k} \Sigma_{j=1}^k j p^{j-1} + (r-1) \frac{q_i}{p^k} \left(k p^k + (1-p) \Sigma_{j=1}^k j p^{j-1}\right)$, $1 \leq i \leq m$;

$$\begin{aligned}
\textit{(iv)}\ \ Var(Y_i^{(r)}) &= \tfrac{q_i}{p^k}\ \Sigma_{j=1}^k j(j-1)p^{j-1} + \tfrac{q_i^2}{p^{2k}}\left(\Sigma_{j=1}^k jp^{j-1}\right)^2 + \tfrac{q_i}{p^k}\ \Sigma_{j=1}^k jp^{j-1}\\
&+ (r-1)q_ik^2 + (r-1)(1-p)\tfrac{q_i}{p^k}\left(\Sigma_{j=1}^k jp^{j-1} + \Sigma_{j=1}^k j(j-1)p^{j-1}\right)\\
&+ (r-1)\tfrac{q_i^2}{p^{2k}}\left(kp^k + (1-p)\Sigma_{j=1}^k jp^{j-1}\right)\left((1+p)\Sigma_{j=1}^k jp^{j-1} - kp^k\right),\\
&\qquad 1 \le i \le m;
\end{aligned}$$

$$\begin{aligned}
\textit{(v)}\ \ Cov(Y_i^{(r)}, Y_s^{(r)}) &= 3(r-1)k\ \tfrac{q_iq_s}{p^k}\ \Sigma_{j=1}^k jp^{j-1} - (r-1)\tfrac{q_iq_s}{p^{2k}}\\
&\times\left(kp^k + (1-p)\Sigma_{j=1}^k jp^{j-1}\right)\\
&- (r-1)\tfrac{q_iq_s}{p^{2k}}\left(kp^k + (1-p)\Sigma_{j=1}^k jp^{j-1}\right)\left(\Sigma_{j=1}^k jp^{j-1}\right)\\
&+ [3\{r(1-p)+p\} - 2]\ \tfrac{q_iq_s}{p^{2k}}\left(\Sigma_{j=1}^k jp^{j-1}\right)^2, \quad 1 \le i \ne s \le m.
\end{aligned}$$

PROOF. For $|t_i| \le 1$ $(1 \le i \le m)$ we have, by means of Lemma 25.3.1 and relation (25.4)

$$\begin{aligned}
g_{\boldsymbol{Y}^{(r)}}(\boldsymbol{t}) &= E\left[t_1^{Y_1^{(r)}} \cdots t_m^{Y_m^{(r)}}\right]\\
&= pE\left[t_1^{Y_1^{(r-1)}} \cdots t_m^{Y_m^{(r-1)}}\right]\\
&\quad + \sum_i q_iE\left[t_1^{Y_1^{(r-1)}+W_1} \cdots t_i^{Y_i^{(r-1)}+k+W_i} \cdots t_m^{Y_m^{(r-1)}+W_m}\right]\\
&= pE\left[t_1^{Y_1^{(r-1)}} \cdots t_m^{Y_m^{(r-1)}}\right]\\
&\quad + \sum_i q_it_i^kE\left[t_1^{Y_1^{(r-1)}} \cdots t_m^{Y_m^{(r-1)}}\right]E\left[t_1^{W_1} \cdots t_m^{W_m}\right]\\
&= g_{\boldsymbol{Y}^{(r-1)}}(\boldsymbol{t})\left(p + g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t})\sum_i q_it_i^k\right),
\end{aligned}$$

which establishes Part (i) of the theorem.

Next, using successively Part (i), it can be seen that

$$g_{\boldsymbol{Y}^{(r)}}(\boldsymbol{t}) = g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t})\left(p + g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t})\Sigma_iq_it_i^k\right)^{r-1}.$$

But, $\boldsymbol{Y}^{(1)}$ is distributed as $\overline{MG}_{k,I}(q_1, \ldots, q_m)$, because of (25.3). Therefore,

$$g_{\boldsymbol{Y}^{(1)}}(\boldsymbol{t}) = \frac{p^k}{1 - \Sigma_iq_i\Sigma_{j=1}^k p^{j-1}t_i^j}, \qquad |t_i| \le 1,\ 1 \le i \le m$$

[see Philippou, Antzoulakos and Tripsiannis (1988)]. The last two relations establish Part (ii) of the theorem.

Parts (iii)–(v) follow very simply from the expression of $g_{\boldsymbol{Y}}(\boldsymbol{t})$ in Part (ii). ■

For $m = 1$, Theorem 25.3.1 gives the pgf, mean and variance of the shifted negative binomial distribution of order k, type III, which are consistent with the corresponding results of Ling (1989) [see also Theorem 4.1 of Hirano *et al.* (1991)]. In addition, for $k = 1$, it provides the pgf, mean and variance-covariance of the usual multivariate negative binomial distribution [see, for example, Patil *et al.* (1984, p. 108) and Johnson, Kotz and Balakrishnan (1997, Chapter 36)].

Next, we represent the random vector $\boldsymbol{Y}$, which is distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$, as a sum of r independent random vectors which are related to $\overline{MG}_{k,I}(q_1, \ldots, q_m)$.

Theorem 25.3.2 *Let* $\boldsymbol{Y}$ *be a random vector distributed as* $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$, $r \geq 2$. *Also, let* $\boldsymbol{W}^{(j)} = (W_1^{(j)}, \ldots, W_m^{(j)})'$, $1 \leq j \leq r$, *be independent random vectors distributed as* $\overline{MG}_{k,I}(q_1, \ldots, q_m)$. *Now, consider random variables* Z_j $(1 \leq j \leq r-1)$ *such that*

$$\Pr[Z_j = i] = q_i \quad \text{and} \quad \Pr[Z_j = 0] = p \quad (1 \leq i \leq m \text{ and } 1 \leq j \leq r-1).$$

Suppose that $W_i^{(j)}$ *and* Z_j $(1 \leq i \leq m$ *and* $1 \leq j \leq r)$ *are mutually independent. Next, for each* $1 \leq i \leq m$ *and* $1 \leq j \leq r-1$, *define random vectors* $\boldsymbol{X}^{(j)}$ *as follows:*

$$\boldsymbol{X}^{(j)} = \begin{cases} \boldsymbol{k}_i + \boldsymbol{W}^{(j)} & \text{if } Z_j = i, \\ \boldsymbol{0} & \text{if } Z_j = 0, \end{cases}$$

where $\boldsymbol{k}_i$ *is a vector with its* i*-th component equal to* k *and all of its other components equal to 0, and* $\boldsymbol{0}$ *is a vector whose components are all 0. Then,* $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(r-1)}, \boldsymbol{W}^{(r)}$ *are independent random vectors and*

$$\boldsymbol{Y} = \boldsymbol{X}^{(1)} + \cdots + \boldsymbol{X}^{(r-1)} + \boldsymbol{W}^{(r)}.$$

PROOF. For $|t_i| \leq 1$, $1 \leq i \leq m$ and $1 \leq j \leq r-1$, we have

$$\begin{aligned}
g_{\boldsymbol{X}^{(j)}}(\boldsymbol{t}) &= E\left[t_1^{X_1^{(j)}} \cdots t_m^{X_m^{(j)}}\right] \\
&= pE\left[t_1^0 \cdots t_m^0\right] + q_1 E\left[t_1^{k+W_1^{(j)}} t_2^{W_2^{(j)}} \cdots t_m^{W_m^{(j)}}\right] \\
&\qquad + \cdots + q_m E\left[t_1^{W_1^{(j)}} t_2^{W_2^{(j)}} \cdots t_m^{k+W_m^{(j)}}\right] \\
&= p + E\left[t_1^{W_1^{(j)}} t_2^{W_2^{(j)}} \cdots t_m^{W_m^{(j)}}\right] \Sigma_i q_i t_i^k \\
&= p + \frac{p^k}{1 - \Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \, \Sigma_i q_i t_i^k,
\end{aligned}$$

since $\boldsymbol{W}^{(j)}$ $(1 \leq j \leq r)$ is distributed $\overline{MG}_{k,I}(q_1, \ldots, q_m)$. The independence of $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(r-1)}, \boldsymbol{W}^{(r)}$, which follows from their definition, implies that

$$g_{\boldsymbol{X}^{(1)}+\cdots+\boldsymbol{X}^{(r-1)}+\boldsymbol{W}^{(r)}}(\boldsymbol{t}) = \left(p + \frac{p^k}{1 - \Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j}\, \Sigma_i q_i t_i^k\right)^{r-1} \times \frac{p^k}{1 - \Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j},$$

which is the pgf of $\boldsymbol{Y}$ by Part (ii) of Theorem 25.3.1. Therefore,

$$\boldsymbol{Y} = \boldsymbol{X}^{(1)} + \cdots + \boldsymbol{X}^{(r-1)} + \boldsymbol{W}^{(r)},$$

which completes the proof of the theorem. ■

For $m = 1$, Theorem 25.3.2 reduces to a result for $\overline{NB}_{k,III}(r,p)$, which corresponds to Theorem 4.2 of Hirano *et al.* (1991) regarding $NB_{k,III}(r,p)$.

The following proposition, which can be easily derived from Part (ii) of Theorem 25.3.1, provides a new genesis scheme for the multivariate negative binomial distribution of order k, type III, with proper parameters.

Proposition 25.3.1 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ be a random vector distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$. Further, let $r_0, r_1, \ldots, r_n$ $(n \geq 1)$ be non-negative integers such that $0 = r_0 < r_1 < \cdots < r_n = m$, and let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ be a random vector such that $Z_s = \sum_{i=r_{s-1}+1}^{r_s} Y_i$ $(1 \leq s \leq n)$. Then, $\boldsymbol{Z}$ is distributed as $\overline{MNB}_{k,III}(r; Q_1, \ldots, Q_n)$, where $Q_s = \sum_{i=r_{s-1}+1}^{r_s} q_i$ for $1 \leq s \leq n$.*

For $n = 1$, Proposition 25.3.1 yields the following corollary, which provides a new genesis scheme for $\overline{NB}_{k,III}(r,p)$.

Corollary 25.3.1 *Let $\boldsymbol{Y} = (Y_1, \ldots, Y_m)'$ be a random vector distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$. Then, $Y = \sum_i Y_i$ is distributed as $\overline{NB}_{k,III}(r,p)$, where $p = 1 - q_1 - \cdots - q_m$.*

We finally give a limit theorem for $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$, which generalizes the shifted version of a corresponding result of Hirano *et al.* (1991) on $NB_{k,III}(r,p)$ to the multivariate case.

Theorem 25.3.3 *Let $\boldsymbol{Y}^{(r)} = (Y_1^{(r)}, \ldots, Y_m^{(r)})'$ be a random vector distributed as $\overline{MNB}_{k,III}(r; q_1, \ldots, q_m)$, and let $Z_1, \ldots, Z_m$ be independent random variables such that*

$$\Pr[Z_i = kz_i] = e^{-\lambda_i} \frac{\lambda_i^{z_i}}{z_i!}, \qquad z_i = 0, 1, \ldots \ (\lambda_i > 0,\ 1 \leq i \leq m).$$

Set $\boldsymbol{Z} = (Z_1, \ldots, Z_m)'$ and assume that $q_i \to 0$ and $rq_i \to \lambda_i$ as $r \to \infty$. Then,

$$\Pr[\boldsymbol{Y}^{(r)} = \boldsymbol{y}] \to \Pr[\boldsymbol{Z} = k\boldsymbol{y}], \qquad y_i = 0, 1, \ldots \ (1 \leq i \leq m).$$

PROOF. For $|t_i| \leq 1$, $1 \leq i \leq m$, we have from Part (ii) of Theorem 25.3.1

$$\begin{aligned}
&g_{\mathbf{Y}^{(r)}}(\boldsymbol{t}) \\
&= \frac{p^k}{1-\Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \left(p + \frac{p^k}{1-\Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \, \Sigma_i q_i t_i^k \right)^{r-1} \\
&= \frac{p^k}{1-\Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \left(1 - \frac{\Sigma_i r q_i \left(1 - \frac{p^k}{1-\Sigma_i q_i \Sigma_{j=1}^k p^{j-1} t_i^j} \, t_i^k \right)}{r} \right)^{r-1} \\
&\to \exp\left\{ -\Sigma_i \lambda_i \left(1 - t_i^k \right) \right\} \qquad \text{as } r \to \infty,
\end{aligned}$$

which is the pgf of $\boldsymbol{Z}$. Hence, the theorem. ∎

For $k = 1$, Theorem 25.3.3 reduces to a known result on the multivariate negative binomial distribution [see, for example, Patil *et al.* (1984, p. 109) and Johnson, Kotz and Balakrishnan (1997, Chapter 36)].

Acknowledgement. The authors wish to thank Dr. Demetrios Antzoulakos for a number of helpful comments.

References

1. Aki, S. (1985). Discrete distributions of order k on a binary sequence, *Annals of the Institute of Statistical Mathematics*, **37**, 205–224.

2. Aki, S., Kuboki, H. and Hirano, K. (1984). On discrete distributions of order k, *Annals of the Institute of Statistical Mathematics*, **36**, 431–440.

3. Antzoulakos, D. L. and Philippou, A. N. (1994). Expressions in terms of binomial coefficients for some multivariate distributions of order k, In *Runs and Patterns: Selected Papers* (Eds., A. P. Godbole and S. G. Papastavridis), pp. 1–14, Dordrecht: Kluwer Academic Publishers.

4. Godbole, A. P. (1990). Specific formulae for some success run distributions, *Statistics & Probability Letters*, **10**, 119–124.

5. Hirano, K. and Aki, S. (1987). Properties of the extended distributions of order k, *Statistics & Probability Letters*, **6**, 67–69.

6. Hirano, K., Aki, S., Kashiwagi, N. and Kuboki, H. (1991). On Ling's binomial and negative binomial distributions of order k, *Statistics & Probability Letters*, **11**, 503–509.

7. Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*, New York: John Wiley & Sons.

8. Ling, K. D. (1988). On binomial distributions of order k, *Statistics & Probability Letters*, **6**, 247–250.

9. Ling, K. D. (1989). A new class of negative binomial distributions of order k, *Statistics & Probability Letters*, **7**, 371–376.

10. Ling, K. D. and Tai, T. H. (1990). On bivariate binomial distributions of order k, *Soochow Journal of Mathematics*, **16**, 211–220.

11. Patil, G. P. and Bildikar, S. (1967). Multivariate logarithmic series distribution as a probability model in population and community ecology and some of its statistical properties, *Journal of the American Statistical Association*, **62**, 655–674.

12. Patil, G. P., Boswell, M. T., Joshi, S. W. and Ratnaparkhi, M. V. (1984). *Dictionary and Classified Bibliography of Statistical Distributions in Scientific Work, Volume 1*, Fairland: International Cooperative Publishing House.

13. Philippou, A. N. (1986). Distributions and Fibonacci polynomials of order k, longest runs, and reliability of consecutive-k-out-of-n:F systems, In *Fibonacci Numbers and Their Applications* (Eds., A. N. Philippou, G. E. Bergum and A. F. Horadam), pp. 203–227, Dordrecht: D. Reidel.

14. Philippou, A. N. (1988). On multiparameter distributions of order k, *Annals of the Institute of Statistical Mathematics*, **40**, 467–475.

15. Philippou, A. N. and Antzoulakos, D. L. (1990). Multivariate distributions of order k on a generalized sequence, *Statistics & Probability Letters*, **9**, 453–463.

16. Philippou, A. N. and Tripsiannis, G. A. (1991). Multivariate Pólya and inverse Pólya distributions of order k, *Biometrical Journal*, **33**, 225–236.

17. Philippou, A. N., Antzoulakos, D. L. and Tripsiannis, G. A. (1988). Multivariate distributions of order k, *Statistics & Probability Letters*, **7**, 207–216.

18. Philippou, A. N., Antzoulakos, D. L. and Tripsiannis, G. A. (1990). Multivariate distributions of order k, Part II, *Statistics & Probability Letters*, **10**, 29–35.

19. Philippou, A. N., Georghiou, C. and Philippou, G. N. (1983). A generalized geometric distribution of order k, *Statistics & Probability Letters*, **1**, 171–175.

20. Philippou, A. N., Tripsiannis, G. A. and Antzoulakos, D. L. (1989). New Pólya and inverse Pólya distributions of order k, *Communications in Statistics—Theory and Methods*, **18**, 2125–2137.

21. Sibuya, M., Yoshimura, I. and Shimizu, R. (1964). Negative multinomial distribution, *Annals of the Institute of Statistical Mathematics*, **16**, 409–426.

PART VI

Applications to Distribution Theory

26

The Joint Energy Distributions of the Bose-Einstein and of the Fermi-Dirac Particles

I. Vincze and R. Tőrös

Mathematical Institute of the Hungarian Academy of Sciences, Budapest
Research Institute of Technical Chemistry of the Hungarian Academy of Sciences, Budapest

Abstract: Vincze (1959, 1961) derived the Planck-Bose-Einstein (PBE) probability density function for the energy distribution of black body radiation. This derivation was based on an expression for the information measure (negentropy) belonging to a continuous random variable and on the Bose-Einstein statistics; the quantum hypothesis of Planck was not needed. The present note considers the joint distribution of several variables. In earlier works, the method for solving the extreme value problem happened with a method due to Kullback and Leibler (1951) and the result did not agree with the formula given by Planck (1900), unless certain element was neglected. In this paper, the authors consider again the extreme value problem but through Lagrange method of the theory of variation which yields the PBE formula exactly. For the Fermi-Dirac case, the procedure goes on the same line. As a consequence, one has the property "indistinguishability of the particles" as a tool for arriving at the results, but *the dependence of the random variables* is essential. This suggests that our procedure is appropriate even for gases which are neither bosons nor fermions, but the particles influence each other and their distributions are not independent [see Wilczek (1991)].

Keywords and phrases: Joint probability distribution, Bose-Einstein statistic, statistical physics, information, entropy

26.1 Introduction

The original derivation of the energy distribution in the case of the black body radiation was given by Planck (1900); he used the maximum entropy principle introduced by Boltzmann (1896). Planck, for avoiding combinatorial difficulties, assumed the famous quantum hypothesis; he dropped essentially the mechanical character of the mechanism based on the phase space. Several further derivations of Planck's formula followed over the years, but first of them was given by Einstein; recognizing the genious idea of Bose, he has given a procedure for obtaining Planck's result adding also an interpretation (indistinguishable particles) referring to the combinatorial tool used in the method. This justifies the name Planck-Bose-Einstein (PBE) distribution. Vincze (1959, 1961) extended the Boltzmann (Shannon) discrete entropy formula for continuous random variables based on the Kolmogorov measure theoretical foundation of the probability theory using the fact: while the entropy term when refining the partition of space of elementary events tends to infinity in fairly regular cases, the complementary notion of the information (negentropy) may tend to a finite value. In formulas:

$$E_n = \sum_{i=1}^{n} p_i \log p_i^{-1} = \log n + \sum_{i=1}^{n} p_i \log \frac{p_i}{\frac{1}{n}}$$

or

$$E_n + I_n = \log n, \qquad \text{with } I_n = \sum_{i=1}^{n} p_i \log \frac{p_i}{\frac{1}{n}} \,. \tag{26.1}$$

26.2 Derivation of the Joint Distribution and of the Joint Entropy

26.2.1 On the method

Let us have a sequence of *independent*, identically distributed random variables

$$X_1, X_2, \ldots, X_N, \ldots \,. \tag{26.2}$$

From now on, X_i will be the energy of a particle of a gas system $\mathcal{E}$. Let $\Pr[X_i < x \mid T_0] = \phi(x)$, the common distribution of X_i's under an initial state T_0 of $\mathcal{E}$. Due to an external effect, the common distribution of the X_i's became $\Pr[X_i < x \mid T] = F(x)$, i.e., $\mathcal{E}$ turns to be in the state T. We now assume the following:

(a) The basic distribution of the random variables in (26.2) will be $\phi(x)$, but after an external effect their actual distribution will be $F(x)$, which is the consequence of the constraint. The situation is completely analogous to the mechanics: having a field, the path of a particle will be determined by the law of the field in spite of the fact that a constraint is present, which will influence the actual path in a noticeable fashion.

(b) The set of possible values of the random variables in (26.2) will remain unchanged under a durable effect of the external factor. This means the existence of the Radon-Nikodym derivative $\psi(x) = \frac{dF(x)}{d\phi(x)}$.

Constructing the empirical distribution function $\phi_N(x)$ from the first N terms of (26.2) which, according to the theorem of Glivenko, tends to $\phi(x)$ uniformly on the whole real line. But due to the external effect, $\phi_N(x)$ will be close to a distribution function $F(x)$ different from $\phi(x)$. The probability of this event is small, and it tends to zero when $N \to \infty$, but its N-th root may tend to a finite limit. This is expressed in the large deviation theorem of Sanov (1961) which has the form

$$\begin{aligned} \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{1}{N} \log \Pr \left[\sup_x |\phi_N(x) - F(x)| < \varepsilon \right] \\ = - \int_{-\infty}^{\infty} \log \frac{dF(x)}{d\phi(x)} \, dF(x), \end{aligned} \tag{26.3}$$

i.e., asymptotically,

$$\Pr \left[\sup_x |\phi_N(x) - F(x)| < \varepsilon \right] \approx e^{-N \int_{-\infty}^{\infty} \log \frac{dF(x)}{d\phi(x)} dF(x)}. \tag{26.4}$$

This is the mathematical form of our assumption (a), when the variables are independent and satisfy simple requirements.

The exponent is $-N$ times the information, which corresponds in the discrete case to I_n in (26.1).

Eq. (26.3) is valid both for discrete and for continuous random sequences.

Now the maximum entropy principle (the maximum probability principle) corresponds to the minimum information principle; i.e., they are equivalent:

$$\int_{-\infty}^{\infty} \log \frac{dF(x)}{d\phi(x)} \, dF(x) = \text{minimum},$$

under the conditions:

$$\int_{-\infty}^{\infty} dF(x) = 1, \tag{26.5}$$

$$\int_{-\infty}^{\infty} x dF(x) = m \neq \int_{-\infty}^{\infty} x d\phi(x) = m_0. \tag{26.6}$$

The solution to this problem leads to the Boltzmann distribution in the discrete case, and to the Gibbs-distribution in the continuous case [Kullback-Leibler].

Now adopting the heuristic approach given by Bose and Einstein, we turn to the *dependent, multivariate* case. We *accept as information the left hand side in (26.3) or (26.4)* and we shall determine this "distance" in the same way as was done by Sanov: dividing the real axis into n parts, we calculate the probability of the random variable ν_i which is the number of particles having energies falling into the i-th interval among the N sample elements. Then, we let n and N tend to infinity.

26.2.2 Joint distribution of the number of particles in energy intervals

Considering a large system of particles, the characteristic of the particles was the energy as a random variable to be investigated. We assume they are identically distributed and the particles are indistinguishable. We now take a large number (say N), of samples of size s.

Denoting the i-th characteristic in the k-th sample by $X_{k,i}$, $i = 1, 2, \ldots, s$, we have the vectors

$$\boldsymbol{X}_k = (X_{k,1}, X_{k,2}, \ldots, X_{k,s}), \qquad k = 1, 2, \ldots, N. \tag{26.7}$$

Our aim is to determine the joint distribution of these s random variables under the assumptions mentioned above that they are identically distributed and the particles indistinguishable. Let us denote the joint distribution by

$$\begin{aligned} &\Pr[X_{k,1} < x_1,\ X_{k,2} < x_2, \ldots, X_{ks} < x_s] \\ &= \ F(x_1, x_2, \ldots, x_s) = F(x_{i_1}, x_{i_2}, \ldots, x_{i_s}), \end{aligned} \tag{26.8}$$

for any permutation $(i_1, i_2, \ldots, i_n)$ of $(1, 2, \ldots, n)$ which means that the indistinguishability implies the exchangeability of the random variables.

The determination of the joint distribution will happen – as usual in the physics – in such a way that we determine the empirical distribution based on the sample of size N. Then we let N to infinity and in the limiting case we consider the "distance" of the probability under the actual distribution (state T) from the distribution under the original state (T_0). The latter will be denoted by

$$\phi(x_1, x_2, \ldots, x_s) = \phi(x_{i_1}, x_{i_2}, \ldots, x_{i_s}). \tag{26.9}$$

For the determination of $F(x_1, x_2, \ldots, x_s)$, we need the knowledge of $\phi(x_1, x_2, \ldots, x_n)$, but as in number of cases this is unknown. So some assumption (hypothesis) will be taken on it whose validity can be justified by a comparison of the experimental data with the theoretical values computed under $\phi(x_1, x_2, \ldots, x_s)$. This is the method we explained in our earlier section and we now turn to the calculation.

Our sample vectors are elements of the positive quadrant of the Euclidean space E_s^+. We divide the s real axis $(0, \infty)$ into n parts by means of the system of points

$$0 \le y_{l0} < y_{l1} < y_{l2} < \cdots < y_{ln} = \infty, \qquad l = 1, 2, \ldots, s;\ n = 1, 2, \ldots$$

This system divides the positive quadrant E_s^+ into n^s s-dimensional cubes. Such a cube will be called a *cell* and the question is how the N observations will be distributed into the n^s cells. Following Bose and Einstein, we shall divide each cell into z parts (which corresponds to a refinement of the energy levels).

We assume that each – original – cell has the same probability $\frac{1}{n^s}$; this means that in the case of $s = 2$, the following relation holds (here we have two axis only, which will be denoted by x_1 and x_2):

$$\begin{aligned}\Delta\phi(x_{1,i}, x_{2,j}) &= \phi(x_{1,i+1}, x_{2,j+1}) - \phi(x_{1,i}, x_{2,j+1}) \\ &\quad - \phi(x_{1,i+1}, x_{2,j}) + \phi(x_{1,i}, x_{2,j}),\end{aligned}$$

where

$$i, j = 0, 1, 2, \ldots, n-1.$$

We shall use the notation $\nu_{i_1,i_2,\ldots,i_s}$ for the random variable, which is the number of vectors $\boldsymbol{X}_K$ falling into the corresponding cell of E_s^+ ($i_j = 1, 2, \ldots, n$; $j = 1, 2, \ldots, s$, $k = 1, 2, \ldots, N$). According to the Bose-Einstein principle, the probability that

$$\nu_{i_1,i_2,\ldots,i_s} = N_{i_1,i_2,\ldots,i_s}$$

with

$$\sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \cdots \sum_{i_s=1}^{n} N_{i_1,i_2,\ldots,i_s} = N,$$

has the value

$$P_{n^s,N} = \frac{\prod_{i_1=1}^{n} \prod_{i_2=1}^{n} \cdots \prod_{i_s=1}^{n} \binom{N_{i_1,i_2,\ldots,i_s}+z-1}{N_{i_1,i_2,\ldots,i_s}}}{\binom{N+n^s z-1}{N}}. \tag{26.10}$$

For the sake of simplicity the choice $N = n^s z$ will be taken, where n and z will tend to infinity. Making use of Stirling's formula in the form $N! \approx \left(\frac{N}{e}\right)^N \sqrt{2\pi N}$, we obtain

$$\begin{aligned}P_{n^s,N} &\sim \frac{N^N (n^s z - 1)^{n^s z - 1}}{(N + n^s z - 1)^{N + n^s z - 1}} \prod_{i_1=1}^{n} \prod_{i_2=1}^{n} \cdots \\ &\quad \prod_{i_s=1}^{n} \frac{\left(\frac{N_{i_1,i_2,\ldots,i_s}}{N} + \frac{1}{n^s}\right)^{N_{i_1,i_2,\ldots,i_s}+z}}{\left(\frac{N_{i_1,i_2,\ldots,i_s}}{N}\right)^{N_{i_1,i_2,\ldots,i_s}} \left(\frac{1}{n^s}\right)^z}.\end{aligned}$$

For the N-th root of the term before the product sign, we have the asymptotic value $\frac{1}{4}$.

This probability then has the form

$$\begin{aligned}P_{n^s,N} &\sim \frac{1}{4^N} \prod_{i_1=1}^{n} \prod_{i_2=1}^{n} \cdots \\ &\quad \prod_{i_s=1}^{n} \left(1 + \frac{\frac{1}{n^s}}{\frac{N_{i_1,i_2,\ldots,i_s}}{N}}\right)^{N_{i_1,i_2,\ldots,i_s}+z} \left(\frac{\frac{N_{i_1,i_2,\ldots,i_s}}{N}}{\frac{1}{n^s}}\right)^z.\end{aligned}$$

As

$$\begin{aligned} \frac{N_{i_1,i_2,\ldots,i_s}}{N} &= \Delta F(x_{i_1}, x_{i_2}, \ldots, x_{i_s}) \equiv \Delta F, \\ \frac{1}{n^s} &= \Delta\phi(x_{i_1}, x_{i_2}, \ldots, x_{i_s} \equiv \Delta\phi, \end{aligned}$$

we obtain the following relation

$$\frac{1}{N}\log P_{n^s,N} \sim \log\frac{1}{4} + \sum_{i=1}^{n}\sum_{i_2=1}^{n}\cdots\sum_{i_s=1}^{n}\left[(\Delta F + \Delta\phi)\log\left(1 + \frac{\Delta\phi}{\Delta F}\right) + \Delta\phi\log\frac{\Delta F}{\Delta\phi}\right].$$

From now on we assume that when $N \to \infty$, the two joint distributions induce measures on E_s^+, which are absolutely continuous with respect to each other, i.e., the Radon-Nikodym derivative

$$\psi(x_{1,2}, \ldots, x_s) = \frac{dF(x_1, x_2, \ldots, x_s)}{d\phi(x_1, x_2, \ldots, x_s)} \tag{26.11}$$

exists.

With N and z tending to infinity, we obtain for the logarithm of the information according to Vincze (1959)

$$H_{B-E} = -k\int_{E_s^+} dF(\boldsymbol{x})\log\left[\frac{\psi(\boldsymbol{x})}{(1+\psi(\boldsymbol{x}))^{1+\frac{1}{\psi(\boldsymbol{x})}}}\right], \tag{26.12}$$

where

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_s) \in E_s^+.$$

Through a similar procedure, we come to the thermodynamic probability of the Fermi-Dirac statistic

$$H_{F-D} = -k\int_{E_s^+} dF(\boldsymbol{x})\log\frac{\psi(\boldsymbol{x})}{(1-\frac{1}{\alpha}\psi(\boldsymbol{x}))^{\frac{\alpha}{\psi(\boldsymbol{x})}}}, \tag{26.13}$$

where α constant is larger than the maximum of $\psi(\boldsymbol{x})$.

26.3 Determination of the Limit Distributions

The solution of the extremum problem may happen in two ways.

(a) The Kullback-Leibler method

Using the Lagrange multipliers $\boldsymbol{\lambda}$, μ for $m_i = \int x_i dF(\boldsymbol{x})$, $\int dF = 1$, we obtain the unconditional problem

$$I_{B-E} + \mu \int_{E_s^+} dF(\boldsymbol{x} \mid T) + \int_{E_s^+} (\boldsymbol{\lambda}, \boldsymbol{x}) dF(\boldsymbol{x} \mid T) = \text{minimum} \tag{26.14}$$

which can be written in the form

$$\int_{E_s^+} dF(\boldsymbol{x} \mid T) \log \frac{dF(\boldsymbol{x} \mid T)}{e^{-\mu-(\boldsymbol{\lambda},\boldsymbol{x})}(1+\psi(\boldsymbol{x}))^{1+\frac{1}{\psi(\boldsymbol{x})}} dF(\boldsymbol{x} \mid T_0)} = \min, \tag{26.15}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_s)$ abbreviation was used and $(\boldsymbol{\lambda}, \boldsymbol{x})$ denotes the scale product.

(26.14) is the Bose-Einstein case, while

$$\int_{E_s^+} dF(\boldsymbol{x} \mid T) \log \frac{dF(\boldsymbol{x} \mid T)}{e^{-\mu-(\boldsymbol{\lambda},\boldsymbol{x})}(1-\frac{1}{\alpha}\psi(\boldsymbol{x}))^{1-\frac{\alpha}{\psi(\boldsymbol{x})}} dF(\boldsymbol{x} \mid T_0)} = \min \tag{26.16}$$

in the Fermi-Dirac case.

Choosing μ in both cases so that the denominator is a density function, we can again conclude that the minimums are zero:

$$dF(\boldsymbol{x} \mid T) = e^{-\mu(\boldsymbol{\lambda},\boldsymbol{x})}(1+\psi(\boldsymbol{x}))^{1+\frac{1}{\psi(\boldsymbol{x})}} dF(\boldsymbol{x} \mid T_0) \tag{26.17}$$

in the Bose-Einstein, and

$$dF(\boldsymbol{x} \mid T) = e^{-\mu-(\boldsymbol{\lambda},\boldsymbol{x})} \left(1 - \frac{1}{\alpha}\phi(\boldsymbol{x})\right)^{1-\frac{\alpha}{\psi(\boldsymbol{x})}} dF(\boldsymbol{x} \mid T_0) \tag{26.18}$$

in the Fermi-Dirac case. Unfortunately, these expressions are not linear in the unknown $dF(\boldsymbol{x} \mid T)$, and so the solution can be obtained using numerical calculations.

It can be seen easily that neglecting the exponents $\frac{1}{\psi}$ and $\frac{\alpha}{\psi}$, respectively, in (26.17) and (26.18), we obtain exactly the Bose-Einstein and the Fermi-Dirac formula.

(b) The Euler-Lagrange method

We have to solve the minimization problem in (26.14) under the specified conditions by using the standard method of the theory of variation. For this, we suppose the existence of the density functions assuming the necessary continuity properties. Then we have

$$I_{B-E} = \int_{E_s^+} f(\boldsymbol{x} \mid T) \log \frac{f(\boldsymbol{x} \mid T)}{f(\boldsymbol{x} \mid T_0)(1+\psi(\boldsymbol{x}))^{1+\frac{1}{\psi(\boldsymbol{x})}}}\, d\boldsymbol{x}. \tag{26.19}$$

As the term $\frac{df(\boldsymbol{x}|T)}{dx}$ does not occur, we have the equation

$$\log f(\boldsymbol{x} \mid T) - \log(f(\boldsymbol{x} \mid T) + f(\boldsymbol{x} \mid T_0)) + (\mu + (\boldsymbol{\lambda}, \boldsymbol{x})) = 0 \tag{26.20}$$

and hence

$$f(\boldsymbol{x} \mid T) = \frac{f(\boldsymbol{x} \mid T_0)}{e^{\mu+(\boldsymbol{\lambda},\boldsymbol{x})} - 1} \tag{26.21}$$

which has the form of Planck's formula

$$f(\boldsymbol{x}, T) = \frac{f(\boldsymbol{x}, T_0)}{e^{\mu+(\boldsymbol{\lambda},\boldsymbol{x})} + \frac{1}{\alpha}}\,. \tag{26.22}$$

26.4 Discussion

The final results of this paper are the entropy formulae H_{B-E} in (26.12), H_{F-D} in (26.13) and the joint distributions in (26.17) and (26.18) [or (26.21) and (26.22)]. Both formulae contain the *a priori* distribution $F(\boldsymbol{x}, T_0)$ or the density function $f(\boldsymbol{x}, T_0)$. The comparison of results with experiments (or the numerical results) require the knowledge of *a priori* distribution. Here, it was assumed as a constant $\frac{1}{n^s}$. The calculations of physics, however, use for *a priori* densities $f(x, T_0) \sim \sqrt{x}$ in case of massive particles and $f(x, T_0) \sim x^2$ in case of Planck normal spectrum. Now the generalization of *a priori* densities in case of joint distribution of energies $x_1, x_2, \dots, x_s$ requires further efforts. The present results are that taking $f(x, T_0) \sim x_1^2 x_2^2$ for $s = 2$ does not yield the Planck distribution of normal spectrum (black hole radiation). In our opinion, the knowledge of *a priori* density means the knowledge of a deeper natural law. The next step is therefore to find that function so that the deduced Planck distribution would agree with the well known distribution. This is—among others—our future work.

Acknowledgement. The first author's research was supported by the Hungarian National Foundation for Scientific Research, Grant no. T-016384.

References

1. Boltzmann, L. (1896). *Volesungen über Gastheorie*, Berlin.
2. Bose, S. N. (1925). Planck's Gesetz und Lichtquanten Hypothese, *Zeitschrift Physics*, **26**, 178–181.
3. Kullback, S. (1959). *Information and Statistics*, New York: John Wiley & Sons.
4. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79–86.
5. Planck, M. (1900). Zur Theorie des Gesetzes der Energieverteilung in Normalspektrum, *Verband der Deutschen Physikalischen Gesellschaft.*
6. Sanov, I. N. (1961). On the probability of large deviations of random variables, *IMS and AMS Selected Translations in Mathematical Statistics and Probability I*, pp. 213–244.
7. Vincze, I. (1959). On an interpretation of a concept of information theory, *Matematikai Lapok*, **10**, 255–266 (in Hungarian). The English version in *Transactions of the Second Prague Conference*, Academia Prague, 1960.
8. Vincze, I. (1968). On certain distribution theorems of the statistical mechanics, *Publications of the Mathematics and Physics Class of the Hungarian Academy of Science*, **19**, 117–130 (in Hungarian).
9. Vincze, I. (1974). On the maximum probability principle in statistical physics, In *Progress in Statistics* (Eds., J. Gani, K. Sarkadi and I. Vincze), *Colloquia Mathematica Societatis János Bolyai,* **9**; *European Meeting of Statisticians*, Budapest, 1972.
10. Vincze, I. and Tőrös (1994). On the Bose-Einstein distribution, *Third International Conference on Lattice Path Combinatorics and Applications*, Faculty of Mathematical Sciences, University Delhi, Abstract, p. 16.
11. Wilczek, F. (1991). Anyons, *Scientific American*, **11**, 24.

27

On Modified q-Bessel Functions and Some Statistical Applications

A. W. Kemp

University of St Andrews, St Andrews, Scotland

Abstract: The paper defines two modified q-Bessel functions and uses their properties in the study of the distribution of the differences of (i) two Euler variables, (ii) two Heine variables (these are q-analogues of Poisson variables); three-term recurrence relations for the probabilities are obtained and the logconcavity of the distributions is established. A particular case of (ii) is the bilateral discrete normal distribution—this involves Jacobi's triple product. The Euler and the Heine distributions are both special cases of the generalized Euler distribution. The difference of two generalized Euler variables is discussed briefly; other special cases lead to Ramanujan's ${}_1\psi_1$ sum and to its finite form.

Keywords and phrases: q-Poisson distribution, Euler distribution, Heine distribution, generalized Euler distribution, bilateral discrete normal distribution, q-Bessel functions, Jacobi's triple product, Ramanujan's ${}_1\psi_1$ sum

27.1 Introduction

The distribution of the difference of two independent Poisson random variables X_1 and X_2, with parameters θ_1 and θ_2, was first investigated by Irwin (1937) for the special case $\theta_1 = \theta_2$. The more general case $\theta_1 \neq \theta_2$ has been studied by Skellam (1946), de Castro (1952), Prekopa (1952), and others, who showed that

$$\Pr[X_1 - X_2 = n] = e^{-\theta_1-\theta_2}(\theta_1/\theta_2)^{n/2} I_n(2\sqrt{\theta_1\theta_2}),$$

where $I_n(\cdot)$ is the modified Bessel function of the first kind. This distribution arises when a physical effect is measured by the difference of two independent counts which are modelled by Poisson random variables.

Strackee and van der Gon (1962) have commented, "In a steady state the number of light quanta, emitted or absorbed in a definite time, is distributed according to a Poisson distribution. In view thereof, the physical limit of perceptible contrast in vision can be studied in terms of the difference between two independent variates each following a Poisson distribution." Since 1962 there has been a great deal of research concerning the behaviour of subatomic particles; this has led to the development in the physics literature of a q-Poisson distribution with probability mass function (pmf)

$$\Pr[X = x] = \frac{\theta^x \Pr[X = 0]}{(1-q)(1-q^2)\cdots(1-q^x)}, \qquad x = 0, 1, \ldots, \tag{27.1}$$

where $\Pr[X = 0] = \prod_{j\geq 0}(1-\theta q^j)$, $0 < \theta < 1$, and $0 < q < 1$; for example, see Jing (1994). It is the same distribution as that introduced into the statistical literature under the name *Euler distribution* by Benkherouf and Bather (1988) as a prior distribution for a stopping time strategy when sequentially drilling for oil. Benkherouf and Bather also put forward the *Heine distribution* with pmf

$$\Pr[Y = y] = \frac{\eta^y q^{y(y-1)/2} \Pr[Y = 0]}{(1-q)(1-q^2)\cdots(1-q^y)}, \qquad y = 0, 1, \ldots, \tag{27.2}$$

where $\Pr[Y = 0] = \prod_{j\geq 0}(1+\eta q^j)^{-1}$, $0 < \eta$, $0 < q < 1$ as a feasible prior. For properties and other modes of genesis of the two distributions, see Kemp (1992a,b); both can be regarded as q-analogues of the Poisson distribution. More recently, Benkherouf and Alzaid (1993) have developed a third type of prior distribution which they called a *generalized Euler distribution.* This has the pmf

$$\Pr[W = w] = \frac{\theta^w (1-a)(1-aq)\cdots(1-aq^{w-1}) \Pr[W = 0]}{(1-q)(1-q^2)\cdots(1-q^w)}, \qquad w = 0, 1, \ldots, \tag{27.3}$$

where $\Pr[W = 0] = \prod_{j\geq 0}(1-\theta q^j)/(1-a\theta q^j)$, $0 < a < 1$, $0 < \theta < 1$, $0 < q < 1$; Benkherouf and Alzaid remarked that the Euler and the geometric distributions are the special cases $a = 0$ and $a = q$, respectively. Its existence and modes of genesis under less restrictive parameter conditions are studied in Kemp (1996a); the Heine distribution arises when $a \to \infty$, $\theta \to 0^-$, and when $a \to -\infty$, $\theta \to 0^+$ such that $a\theta = -\eta$, $\eta > 0$.

The aim of the present paper is to study the differences of (i) two Euler random variables, (ii) two Heine random variables, using q-analogues of the ordinary modified Bessel function of the first kind

$$I_\nu(z) = e^{-i\nu\pi/2} J_\nu(ze^{i\pi/2}) = \frac{(z/2)^\nu}{\Gamma(\nu+1)}\, {}_0F_1(-; \nu+1; z^2/4), \quad -\pi < \arg z \leq \pi/2,$$

and to comment briefly on the distribution of the difference of two generalized Euler random variables.

Section 27.2 sets out the notation used in the paper, defines the modified q-Bessel functions and gives their properties. Section 27.3 deals with the distribution of the difference of two Euler random variables; Section 27.4 does the same for the distribution of the difference of two Heine random variables and notes that Jacobi's triple product occurs in a particular case. Special cases of the difference of two generalized Euler random variables are mentioned in Section 27.5; these involve Ramanujan's ${}_1\psi_1$ sum and a finite form of it.

27.2 Notation

The Gasper-Rahman definition of a basic hypergeometric function (q-series) [Gasper and Rahman (1990)] will be used throughout:

$$\begin{aligned} &{}_A\phi_B\left(\begin{matrix} a_1,\dots,a_A \\ b_1,\dots,b_B \end{matrix};\ q,z\right) \\ &= {}_A\phi_B(a_1,\dots,a_A;b_1,\dots,b_B;q,z) \\ &= \sum_{j=0}^{\infty}\frac{(a_1;q)_j\dots(a_A;q)_j z^j}{(b_1;q)_j\dots(b_B;q)_j(q;q)_j}\left[(-1)^j q^{\binom{j}{2}}\right]^{B-A+1}, \end{aligned}$$

where $(a;q)_0 = (a)_{q,0} = 1$, $(a;q)_j = (a)_{q,j} = (1-a)(1-aq)\cdots(1-aq^{j-1})$. When $A = B+1$, this agrees with the Bailey-Slater definition [Bailey (1935) and Slater (1966)] that was adopted in Erdélyi *et al.* (1953), Johnson, Kotz and Kemp (1992), and Kemp (1992a,b):

$$\begin{aligned} {}_A\Phi_B\left[\begin{matrix} a_1,\dots,a_A \\ b_1,\dots,b_B \end{matrix};\ q,z\right] &= {}_A\Phi_B[a_1,\dots,a_A;b_1,\dots,b_B;q,z] \\ &= \sum_{j=0}^{\infty}\frac{(a_1;q)_j\dots(a_A;q)_j\ z^j}{(b_1;q)_j\dots(b_B;q)_j(q;q)_j}. \end{aligned}$$

Where $B < A-1$ ($B > A-1$) in the Bailey-Slater notation, Gasper and Rahman add $A-B-1$ extra denominator ($B-A+1$ extra numerator) parameters equal to zero. Thus, all expressions in the Bailey-Slater notation can be stated in the Gasper-Rahman notation but not all expressions in the Gasper-Rahman notation can be stated in the Bailey-Slater notation.

Similarly, Gasper and Rahman (1990) define the bilateral basic hypergeometric function (q-series) as

$$\begin{aligned} &{}_r\psi_s\left(\begin{matrix} a_1,\dots,a_r \\ b_1,\dots,b_s \end{matrix};\ q,z\right) \\ &= \sum_{j=-\infty}^{\infty}\frac{(a_1;q)_j\dots(a_r;q)_j\ z^j}{(b_1;q)_j\dots(b_s;q)_j}(-1)^{(s-r)j}q^{(s-r)j(j-1)/2}. \end{aligned}$$

Work on the q-Bessel functions $J_\nu^{(1)}(z;q)$ and $J_\nu^{(2)}(z;q)$ has been carried out in various notations from Jackson (1904) and Hahn (1949) onward; here we use Ismail's (1981) notation.

We define

$$I_\nu^{(k)}(z;q) = e^{-i\nu\pi/2} J_\nu^{(k)}(ze^{i\pi/2};q), \qquad k = 1, 2, \tag{27.4}$$

where

$$\begin{aligned} J_\nu^{(1)}(z;q) &= \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty} (z/2)^\nu {}_2\phi_1(0,0;q^{\nu+1};q,-z^2/4), \\ J_\nu^{(2)}(z;q) &= \frac{(q^{\nu+1};q)_\infty}{(q;q)_\infty} (z/2)^\nu {}_0\phi_1(-;q^{\nu+1};q,-z^2q^{\nu+1}/4), \end{aligned}$$

$|z| < 2$, $0 < q < 1$, giving

$$I_n^{(1)}(z;q) = \frac{(q^{n+1};q)_\infty}{(q;q)_\infty} (z/2)^n {}_2\phi_1(0,0;q^{n+1};q,z^2/4), \tag{27.5}$$

$$I_n^{(2)}(z;q) = \frac{(q^{n+1};q)_\infty}{(q;q)_\infty} (z/2)^n {}_0\phi_1(-;q^{n+1};q,z^2q^{n+1}/4). \tag{27.6}$$

The relationship between the two functions is

$$I_n^{(2)}(z;q) = (z^2/4;q)_\infty I_n^{(1)}(z;q); \tag{27.7}$$

this follows as in Hahn (1949). Also

$$I_{-n}^{(k)}(z;q) = I_n^{(k)}(z;q), \qquad |z| < 2,\ k = 1, 2. \tag{27.8}$$

The following recurrence and logconcavity properties will be required.

Property 27.2.1

$$-zq^n I_{n+1}^{(k)}(z;q) + zI_{n-1}^{(k)}(z;q) = 2(1-q^n) I_n^{(k)}(z;q), \qquad |z| < 2,\ k = 1, 2. \tag{27.9}$$

This is a direct result of the q-contiguous relationships between the q-series and can be easily proved by comparison of the coefficients of z^j.

Property 27.2.2

$$f_n(z) \equiv \frac{I_{n-1}^{(k)}(z;q) I_{n+1}^{(k)}(z;q)}{I_n^{(k)}(z;q) I_n^{(k)}(z;q)} < 1, \qquad |z| < 2,\ 0 < q < 1,\ k = 1, 2. \tag{27.10}$$

When $n > 0$, $x = z/2$, then from (27.7)

$$\begin{aligned} f_n(z) &= \left(\frac{1-q^n}{1-q^{n+1}}\right) \frac{{}_2\phi_1(0,0;q^{n+2};q,x^2)\ {}_0\phi_1(-;q^n;q,x^2q^n)}{{}_2\phi_1(0,0;q^{n+1};q,x^2)\ {}_0\phi_1(-;q^{n+1};q,x^2q^{n+1})} \\ &= \frac{(1-q^n)\left[1+\sum_{j\geq 1}(q^{2n+1+j};q)_j x^{2j}\{(q;q)_j(q^n;q)_j(q^{n+2};q)_j\}^{-1}\right]}{(1-q^{n+1})\left[1+\sum_{j\geq 1}(q^{2n+1+j};q)_j x^{2j}\{(q;q)_j(q^{n+1};q)_j(q^{n+1};q)_j\}^{-1}\right]}. \end{aligned} \tag{27.11}$$

Subtracting the numerator from the denominator of (27.11) gives

$$(1-q)q^n\left[1+\sum_{j\geq 1}\frac{(q^{2n+1+j};q)_j(x^2q)^j}{(q;q)_j(q^{n+1};q)_j(q^{n+2};q)_j}\right] > 0,$$

and so (27.10) is true for $n > 0$. From (27.8), it is also true for $n < 0$. Finally, when $n = 0$, from (27.8) and (27.7)

$$\begin{aligned} f_0(z) &= \left(\frac{I_1^{(k)}(z;q)}{I_0^{(k)}(z;q)}\right)^2 \\ &= \frac{x^2\ {}_2\phi_1(0,0;q^2;q,x^2)\ {}_0\phi_1(-;q^2;q,x^2q^2)}{(1-q)^2\ {}_2\phi_1(0,0;q;q,x^2)\ {}_0\phi_1(-;q;q,x^2q)}. \end{aligned} \tag{27.12}$$

Subtracting the numerator from the denominator now gives

$$(1-q)^2 + x^2q\sum_{j\geq 0}\frac{(q^{j+3};q)_j(x^2q)^j}{\{(q^2;q)_j\}^3} > 0.$$

This establishes the logconcavity and hence the unimodality of $I_n^{(k)}(z;q)$, $|z|<2$, $0<q<1$, $k = 1,2$.

Property 27.2.3

$$I_n^{(2)}(z;q) = \frac{(z/2)^n}{(q;q)_\infty}\ {}_1\phi_1(z^2/4;0;q,q^{n+1}). \tag{27.13}$$

From the limiting form $\epsilon \longrightarrow 0$ of Heine's transform

$$\begin{aligned} {}_2\phi_1(a,0;c;q,y) &= \lim_{\epsilon\longrightarrow 0} {}_2\phi_1(a,\epsilon;c;q,y) \\ &= \lim_{\epsilon\longrightarrow 0}\frac{(\epsilon;q)_\infty(ay;q)_\infty}{(c;q)_\infty(y;q)_\infty}\ {}_2\phi_1(c/\epsilon,y;ay;q,\epsilon) \\ &= \frac{(ay;q)_\infty}{(c;q)_\infty(y;q)_\infty}\ {}_1\phi_1(y;ay;q,c), \end{aligned} \tag{27.14}$$

it follows, with $a = 0$, that $I_n^{(1)}(z;q)$ can be expressed in terms of a confluent basic hypergeometric function

$$I_n^{(1)}(z;q) = \frac{(z/2)^n}{(q;q)_\infty (z^2/4;q)_\infty} \, {}_1\phi_1(z^2/4;0;q,q^{n+1}); \tag{27.15}$$

remembering (27.7), this gives (27.13).

Remark 27.2.1 When $z^2/4 = q$, then

$$I_n^{(2)}(2\sqrt{q};q) = (q;q)_\infty \, I_n^{(1)}(2\sqrt{q};q) = \frac{q^{n/2}}{(q;q)_\infty} \sum_{j=0}^{\infty} (-1)^j q^{j(2n+j+1)/2}. \tag{27.16}$$

Lemma 27.2.1

$$(g;q)_\infty \, {}_1\phi_1(0;g;q,h) = (h;q)_\infty \, {}_1\phi_1(0;h;q,g), \qquad |g| < 1, \ |h| < 1. \tag{27.17}$$

Using (27.14) and Heine's transform again,

$$\begin{aligned} {}_1\phi_1(0;g;q,h) &= \lim_{c\to\infty} {}_2\phi_1(c,0;g;q,h/c) \\ &= \lim_{c\to\infty} \frac{(h;q)_\infty}{(g;q)_\infty} \, {}_1\phi_1(h/c;h;q,g). \end{aligned}$$

27.3 The Distribution of the Difference of Two Euler Random Variables

From (27.1), the probability generating function (pgf) of the Euler distribution is

$${}_1\phi_0(0;-;q,\theta s)/{}_1\phi_0(0;-;q,\theta) = \frac{(\theta;q)_\infty}{(\theta s;q)_\infty}, \qquad 0 < \theta < 1, 0 < q < 1,$$

by Heine's theorem. The pgf of the difference of two independent Euler random variables with parameters (λ, q) and (μ, q) is therefore

$$G_E(s) = (\lambda;q)_\infty \, (\mu;q)_\infty \, {}_1\phi_0(0;-;q,\lambda s) \, {}_1\phi_0(0;-;q,\mu/s). \tag{27.18}$$

The moment properties can be obtained via the factorial cumulant generating function

$$\begin{aligned} &K_E(t) \\ &= \ln G_E(1+t) \\ &= \sum_{j\geq 0} \left[\ln(1-\lambda q^j) + \ln(1-\mu q^j) - \ln\{1-\lambda q^j(1+t)\} - \ln\{1-\mu q^j/(1+t)\}\right] \\ &= -\sum_{j\geq 0} \left[\ln\left(1 - \frac{\lambda q^j t}{1-\lambda q^j}\right) + \ln\left(1 + \frac{t}{1-\mu q^j}\right) - \ln(1+t)\right]. \end{aligned}$$

Hence or otherwise,

$$\begin{aligned}\mu_1' &= K_{[1]} = \sum_{j\geq 0} \frac{(\lambda-\mu)q^j}{(1-\lambda q^j)(1-\mu q^j)}, \\ \mu_2 &= K_{[2]} + \mu_1' = \sum_{j\geq 0}\left[\frac{\lambda q^j}{(1-\lambda q^j)^2} + \frac{\mu q^j}{(1-\mu q^j)^2}\right], \text{ etc.}\end{aligned}$$

Multiplying together the two q-series in (27.18) and collecting together the terms in s gives

$$\begin{aligned}G_E(s) &= \sum_{n=-\infty}^{\infty} p_n s^n \\ &= (\lambda;q)_\infty\,(\mu;q)_\infty \\ &\quad\times\left[{}_2\phi_1(0,0;q;q,\lambda\mu) + \sum_{n=1}^{\infty}\frac{(\lambda^n s^n + \mu^n s^{-n})}{(q;q)_n}\,{}_2\phi_1(0,0;q^{n+1};q,\lambda\mu)\right]\end{aligned} \tag{27.19}$$

and hence

$$p_n = (\lambda;q)_\infty\,(\mu;q)_\infty \frac{\lambda^n}{(q;q)_n}\,{}_2\phi_1(0,0;q^{n+1};q,\lambda\mu),\ n = 0,1,2,\ldots, \tag{27.20}$$

$$= (\lambda;q)_\infty\,(\mu;q)_\infty \left(\frac{\lambda}{\mu}\right)^{n/2} I_n^{(1)}(2\sqrt{\lambda\mu};q) \tag{27.21}$$

[(27.21) holds for n negative as well as positive.]

From the recurrence relation (27.9), it follows that

$$\mu q^n p_{n+1} = \lambda p_{n-1} - (1-q^n)p_n. \tag{27.22}$$

The logconcavity property (27.10) shows that

$$\frac{p_n p_{n+2}}{p_{n+1}^2} = f_{n+1}(2\sqrt{\lambda\mu}) < 1. \tag{27.23}$$

The distribution of the difference of two independent Euler variables is therefore logconcave. This implies that it is unimodal and that it has an increasing hazard (failure) rate.

There is an alternative way of expressing p_n as a series. From (27.21) and the confluent property (27.15),

$$p_n = \frac{(\lambda;q)_\infty\,(\mu;q)_\infty\,\lambda^n}{(q;q)_\infty\,(\lambda\mu;q)_\infty}\,{}_1\phi_1(\lambda\mu;0;q,q^{n+1}), \quad n = 0,1,2,\ldots; \tag{27.24}$$

this will generally converge more rapidly than (27.20). When n is negative, the roles of λ and μ are interchanged. For $\lambda\mu = q$, (27.24) gives

$$p_n = \frac{(\lambda;q)_\infty\,(q/\lambda;q)_\infty\,\lambda^n}{\{(q;q)_\infty\}^2}\,[1 - q^{n+1} + q^{2n+3} - q^{3n+6} + \cdots], \tag{27.25}$$

cf. (27.16). We note the relationship to Andrews' (1986) expansion

$$\prod_{n=1}^{\infty} \frac{(1-q^n)^2}{(1-zq^n)(1-z^{-1}q^{n-1})} = \sum_{N,r=-\infty}^{\infty} (-1)^{r+N} z^N q^{(r^2-N^2)/2+(r+N)/2}, \tag{27.26}$$

where the summation over N and r is constrained by $r \geq |N|$ and where $1 < |z| < |q|^{-1}$. Andrews obtained this expansion by carrying out a partial fraction decomposition of the left hand side and applying geometric series expansions to the denominators in the partial fraction decomposition.

27.4 The Distribution of the Difference of Two Heine Random Variables

The Heine distribution has the pgf

$$ {}_0\phi_0(-;-;q,-\eta s)/{}_0\phi_0(-;-;q,-\eta) = \frac{(-\eta s;q)_\infty}{(-\eta;q)_\infty}, \qquad 0 < \eta,\ 0 < q < 1,$$

cf. (27.2), by Heine's theorem. The difference of two independent Heine random variables with parameters (λ, q) and (μ, q) therefore has the pgf

$$G_H(s) = {}_0\phi_0(-;-;q,-\lambda s)\ {}_0\phi_0(-;-;q,-\mu/s)\,/\,\{(-\lambda;q)_\infty\,(-\mu;q)_\infty\} \tag{27.27}$$

and the factorial cumulant generating function

$$\begin{aligned} K_H(t) &\\ &= \ln G_H(1+t) \\ &= \sum_{j\geq 0}\Big[\ln\{1+\lambda q^j(1+t)\} + \ln\{1+\mu q^j/(1+t)\} - \ln(1+\lambda q^j) - \ln(1+\mu q^j)\Big] \\ &= \sum_{j\geq 0}\left\{\ln\left(1+\frac{\lambda q^j t}{1+\lambda q^j}\right) + \ln\left(1+\frac{t}{1+\mu q^j}\right) - \ln(1+t)\right\}. \end{aligned}$$

The mean and variance are

$$\begin{aligned} \mu_1' &= \sum_{j\geq 0} \frac{(\lambda-\mu)q^j}{(1+\lambda q^j)(1+\mu q^j)}, \\ \mu_2 &= \sum_{j\geq 0}\left[\frac{\lambda q^j}{(1+\lambda q^j)^2} + \frac{\mu q^j}{(1+\mu q^j)^2}\right]. \end{aligned}$$

Multiplication of the two q-series in (27.27) and collection of the terms in s gives

$$\begin{aligned} G_H(s) &= \{(-\lambda;q)_\infty\,(-\mu;q)_\infty\}^{-1} \times \Bigg[\,{}_0\phi_1(-;q;q,\lambda\mu) \\ &\qquad + \sum_{n=1}^{\infty} \frac{(\lambda^n s^n + \mu^n s^{-n})q^{n(n-1)/2}}{(q;q)_n}\,{}_0\phi_1(-;q^{n+1};q,\lambda\mu q^n)\Bigg]. \end{aligned} \tag{27.28}$$

Therefore,

$$\begin{aligned} p_n &= \{(-\lambda;q)_\infty\,(-\mu;q)_\infty\}^{-1}\,\frac{\lambda^n q^{n(n-1)/2}}{(q;q)_n}\,{}_0\phi_1(-;q^{n+1};q,\lambda\mu q^n) \\ &\qquad n = 0,1,2,\ldots, \qquad (27.29) \\ &= \{(-\lambda;q)_\infty\,(-\mu;q)_\infty\}^{-1}\left(\frac{\lambda}{\mu}\right)^{n/2} q^{n^2/2} I_n^{(2)}(2\sqrt{\lambda\mu/q};q) \qquad (27.30) \end{aligned}$$

[(27.30) holds for n negative as well as n positive.] The recurrence relation between these probabilities is [from (27.9)], *cf.* (27.22),

$$\mu p_{n+1} = \lambda q^n p_{n-1} - (1-q^n)q\,p_n. \tag{27.31}$$

We now have

$$\begin{aligned} \frac{p_n p_{n+2}}{p_{n+1}^2} &= \frac{q I_n^{(2)}(2\sqrt{\lambda\mu/q};q) I_{n+2}^{(2)}(2\sqrt{\lambda\mu/q};q)}{I_{n+1}^{(2)}(2\sqrt{\lambda\mu/q};q) I_{n+1}^{(2)}(2\sqrt{\lambda\mu/q};q)} \qquad (27.32) \\ &= q f_{n+1}(2\sqrt{\lambda\mu/q}) < 1 \qquad (27.33) \end{aligned}$$

by (27.10). The difference of two independent Heine variables therefore also has a logconcave, unimodal distribution with an increasing hazard (failure) rate.

From (27.30) and (27.13)

$$p_n = \frac{\lambda^n q^{n(n-1)/2}}{(-\lambda;q)_\infty\,(-\mu;q)_\infty\,(q;q)_\infty}\,{}_1\phi_1(\lambda\mu/q;0;q,q^{n+1}),\ n = 0,1,2,\ldots, \tag{27.34}$$

and similarly with λ and μ interchanged for n negative. When $\lambda\mu = q$, this becomes

$$p_n = \frac{\lambda^n q^{n(n-1)/2}}{(-\lambda;q)_\infty\,(-q/\lambda;q)_\infty\,(q;q)_\infty}. \tag{27.35}$$

This is the bilateral discrete normal distribution of Kemp (1996b). Summing over n gives Jacobi's triple product

$$(-\lambda;q)_\infty\,(-q/\lambda;q)_\infty\,(q;q)_\infty = \sum_{n=-\infty}^{\infty} \lambda^n q^{n(n-1)/2}. \tag{27.36}$$

27.5 Comments on the Distribution of the Difference of Two Generalized Euler Random Variables

The Euler distribution is the special case $a = 0$ of the generalized Euler distribution with pgf

$$\frac{{}_1\phi_0(a; -; q, \theta s)}{{}_1\phi_0(a; -; q, \theta)} = \frac{(a\theta s; q)_\infty\ (\theta; q)_\infty}{(\theta s; q)_\infty\ (a\theta; q)_\infty},$$

cf. (27.3). Kemp (1996a) has found that the parameter constraints of Benkherouf and Alzaid (1993) can be relaxed to (i) $a < 1$, $0 < \theta < 1$, $0 < q < 1$, (ii) $a = q^{-m}$, $m \in \mathbb{Z}^+$, $\theta < 0$, $0 < q < 1$.

The difference of two independent generalized Euler random variables with parameters (A, λ, q) and (B, μ, q) therefore has the pgf

$$\begin{aligned} G_{GE}(s) &= \frac{(\lambda; q)_\infty\ (\mu; q)_\infty}{(A\lambda; q)_\infty\ (B\mu; q)_\infty}\ {}_1\phi_0(A; -; q, \lambda s)\ {}_1\phi_0(B; -; q, \mu/s) \\ &= \frac{(\lambda; q)_\infty\ (\mu; q)_\infty}{(A\lambda; q)_\infty\ (B\mu; q)_\infty} \left[\sum_{n\geq 0} \frac{(A; q)_n\ \lambda^n s^n}{(q; q)_n}\ {}_2\phi_1(Aq^n, B; q^{n+1}; q, \lambda\mu) \right. \\ &\qquad \left. + \sum_{n\geq 1} \frac{(B; q)_n\ \mu^n s^{-n}}{(q; q)_n}\ {}_2\phi_1(A, Bq^n; q^{n+1}; q, \lambda\mu) \right]. \end{aligned} \tag{27.37}$$

The factorial cumulant generating function follows as before by taking the logarithm of the pgf and substituting $1 + t$ for s.

From (27.37), the probabilities are

$$p_n = \frac{(\lambda; q)_\infty\ (\mu; q)_\infty}{(A\lambda; q)_\infty\ (B\mu; q)_\infty}\ \frac{(A; q)_n\ \lambda^n}{(q; q)_n}\ {}_2\phi_1(Aq^n, B; q^{n+1}; q, \lambda\mu), \tag{27.38}$$

for $n = 0, 1, 2, \ldots$ and similarly, with interchange of parameters, for n negative; they are not directly related to the modified q-Bessel functions. Certain special cases are examined for their inherent interest.

As noted earlier, the Euler and Heine distributions are particular cases of the generalized Euler distribution. The pgf of the difference formed by subtracting a Heine random variable from an Euler random variable is therefore the special case of (27.37) with $A = 0$, $B \to -\infty$, $\mu \to 0^+$, $B\mu = -\beta$. From (27.38)

$$p_n = \frac{(\lambda; q)_\infty}{(-\beta; q)_\infty}\ \frac{\lambda^n}{(q; q)_n}\ {}_1\phi_1(0; q^{n+1}; q, -\lambda\beta),\ n = 0, 1, 2, \ldots, \tag{27.39}$$

and hence from Lemma 27.2.1

$$p_n = \frac{(\lambda;q)_\infty\,(-\lambda\beta;q)_\infty}{(-\beta;q)_\infty\,(q;q)_\infty}\,\lambda^n\,{}_1\phi_1(0;-\lambda\beta;q,q^{n+1}),\ n=0,1,2,\ldots; \tag{27.40}$$

similarly, with interchange of parameters, for n negative. The series in (27.40) has alternating signs; also it will generally converge more rapidly than (27.39).

For a second special case, let $AB\lambda\mu = q$. Then the ${}_2\phi_1(\cdot)$ series in (27.37) can be summed by the q-Gauss sum,

$${}_2\phi_1(a,b;c;q,ca^{-1}b^{-1}) = \frac{(c/a;q)_\infty\,(c/b;q)_\infty}{(c;q)_\infty\,(ca^{-1}b^{-1};q)_\infty},$$

and (27.37) becomes

$$\begin{aligned} G_2(s) &= \frac{(\lambda;q)_\infty\,(qA^{-1}B^{-1}\lambda^{-1};q)_\infty}{(A\lambda;q)_\infty\,(qA^{-1}\lambda^{-1};q)_\infty} \\ &\quad\times\sum_{n=-\infty}^{\infty}\frac{(A;q)_n\,\lambda^n s^n (qA^{-1};q)_\infty\,(q^{n+1}B^{-1};q)_\infty}{(q;q)_\infty\,(qA^{-1}B^{-1};q)_\infty} \\ &= \frac{(\lambda;q)_\infty\,(qA^{-1}B^{-1}\lambda^{-1};q)_\infty\,(qA^{-1};q)_\infty\,(qB^{-1};q)_\infty}{(A\lambda;q)_\infty\,(qA^{-1}\lambda^{-1};q)_\infty\,(q;q)_\infty\,(qA^{-1}B^{-1};q)_\infty} \\ &\quad\times{}_1\psi_1(A;qB^{-1};q,\lambda s). \end{aligned} \tag{27.41}$$

Setting $s=1$ in (27.41) gives Ramanujan's ${}_1\psi_1$ sum.

If $B\to-\infty$, $\mu\to 0^+$, such that $B\mu = qA^{-1}\lambda^{-1}$, then we have a generalized Euler random variable minus a Heine random variable and the corresponding pgf is

$$G_3(s) = \frac{(\lambda;q)_\infty\,(qA^{-1};q)_\infty}{(A\lambda;q)_\infty\,(qA^{-1}\lambda^{-1};q)_\infty\,(q;q)_\infty}\,{}_1\psi_1(A;0;q,\lambda s). \tag{27.42}$$

A Heine random variable with parameters (α,q) minus a generalized Euler random variable with parameters (B,μ,q), where $-\alpha = qB^{-1}\mu^{-1}$, has the pgf

$$\begin{aligned} G_4(s) &= \frac{(-q\alpha^{-1}B^{-1};q)_\infty\,(qB^{-1};q)_\infty}{(-\alpha;q)_\infty\,(-q\alpha^{-1};q)_\infty\,(q;q)_\infty}\sum_{n=-\infty}^{\infty}\frac{\alpha^n s^n q^{n(n-1)/2}}{(qB^{-1};q)_n} \\ &= \frac{(-q\alpha^{-1}B^{-1};q)_\infty\,(qB^{-1};q)_\infty}{(-\alpha;q)_\infty\,(-q\alpha^{-1};q)_\infty\,(q;q)_\infty}\,{}_0\psi_1(-;qB^{-1};q,-\alpha s); \end{aligned} \tag{27.43}$$

this is, of course, the previous distribution with (A,λ) and (B,μ) interchanged and the support reversed. Putting $s=1$ provides a summation formula for a ${}_0\psi_1(\cdot)$ series.

Finally, consider the second set of parameter constraints for the generalized Euler distribution. Suppose that $A=q^{-m}$, $B=q^{-\ell}$, $\lambda<0$, $\mu<0$, where

$m, \ell \in \mathbb{Z}^+$. Suppose also that $\lambda\mu = q$. Both generalized Euler distributions are now finite convolutions of Bernoulli random variables and their difference has finite support. The pgf of their difference is

$$\begin{aligned} G_5(s) &= \frac{{}_1\phi_0(q^{-m}; -; q, \lambda s)\, {}_1\phi_0(q^{-\ell}; -; q, q\lambda^{-1}s^{-1})}{{}_1\phi_0(q^{-m}; -; q, \lambda)\, {}_1\phi_0(q^{-\ell}; -; q, q\lambda^{-1})} \\ &= \frac{1}{(q^{-m}\lambda; q)_m\, (q^{1-\ell}/\lambda; q)_\ell} \sum_{n=-\ell}^{m} \frac{(q^{-m}; q)_n\, \lambda^n s^n}{(q; q)_n} \\ &\qquad \times {}_2\phi_1(q^{n-m}, q^{-\ell}; q^{n+1}; q, q) \\ &= \frac{(q; q)_{m+\ell}}{(q^{-m}\lambda; q)_m\, (q^{1-\ell}/\lambda; q)_\ell} \sum_{n=-\ell}^{m} \frac{(-\lambda)^n s^n q^{-(m-n)(\ell+n)-n(n+1)/2}}{(q; q)_{m-n}\, (q; q)_{\ell+n}} \end{aligned} \tag{27.44}$$

by the q-Vandermonde summation formula. Setting $s = 1$ gives a doubly finite form of Ramanujan's ${}_1\psi_1(a; b; q, z)$ sum without the usual restrictions $|b/a| < |z| < 1$.

References

1. Andrews, G. E. (1986). *q-Series: Their Development and Application in Analysis, Number Theory, Combinatorics, Physics, and Computer Algebra*, Providence, RI: American Mathematical Society.

2. Bailey, W. N. (1935). *Generalized Hypergeometric Series*, London: Cambridge University Press.

3. Benkherouf, L. and Alzaid, A. A. (1993). On the generalized Euler distribution, *Statistics & Probability Letters*, **18**, 323–326.

4. Benkherouf, L. and Bather, J. A. (1988). Oil exploration: sequential decisions in the face of uncertainty, *Journal of Applied Probability*, **25**, 529–543.

5. de Castro, G. (1952). Note on differences of Bernoulli and Poisson variables, *Portugaliae Mathematica*, **11**, 173–175.

6. Erdélyi, A. *et al.* (1953). *Higher Transcendental Functions*, Vols. I, II, New York: McGraw-Hill.

7. Gasper, G. and Rahman, M. (1990). *Basic Hypergeometric Series*, Cambridge: Cambridge University Press.

8. Hahn, W. (1949). Beiträge zur Theorie der Heineschen Reihen, *Mathematische Nachrichten*, **2**, 340–379.

9. Irwin, J. O. (1937). The frequency distribution of the difference between two independent variates following the same Poisson distribution, *Journal of the Royal Statistical Society, Series A*, **100**, 415–416.

10. Ismail, M. E. H. (1981). The basic Bessel functions and polynomials, *SIAM Journal of Mathematical Analysis*, **12**, 454–468.

11. Jackson, F. H. (1904). On generalized functions of Legendre and Bessel, *Transactions of the Royal Society of Edinburgh*, **41**, 1–28.

12. Jing, S. (1994). The q-deformed binomial distribution and its asymptotic behaviour, *Journal of Physics A: Mathematical and General*, **27**, 493–499.

13. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second edition, New York: John Wiley & Sons.

14. Kemp, A. W. (1992a). Heine-Euler extensions of the Poisson distribution, *Communications in Statistics—Theory and Methods*, **21**, 571–588.

15. Kemp, A. W. (1992b). Steady state Markov chain models for the Heine and Euler distributions, *Journal of Applied Probability*, **29**, 869–876.

16. Kemp, A. W. (1996a). Further models for the generalized Euler distribution, (submitted for publication).

17. Kemp, A. W. (1996b). Characterizations of a discrete normal distribution, *Journal of Statistical Planning and Inference* (to appear).

18. Prekopa, A. (1952). On composed Poisson distributions, IV, *Acta Mathematica Academiae Scientiarum Hungaricae*, **3**, 317–325.

19. Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations, *Journal of the Royal Statistical Society, Series A*, **109**, 296.

20. Slater, L. J. (1966). *Generalized Hypergeometric Functions*, London: Cambridge University Press.

21. Strackee, J. and van der Gon, J. J. D. (1962). The frequency distribution of the difference between two Poisson variates, *Statistica Neerlandica*, **16**, 17–23.

28

A q-Logarithmic Distribution

C. David Kemp

University of St Andrews, St Andrews, Scotland

Abstract: The distribution considered here is obtained by replacing the ordinary Gaussian hypergeometric function by an appropriate q-hypergeometric (basic hypergeometric) function in the Kemp (1968) formulation of the probability generating function for the logarithmic distribution. Properties of the resulting q-logarithmic distribution are examined; a birth-and-death process model for group sizes is proposed whose equilibrium distribution is the q-logarithmic.

Keywords and phrases: Basic hypergeometric function, q-analogues of distributions, q-logarithmic distribution, group size model

28.1 Introduction

The ${}_AF_B$ hypergeometric function is defined as

$$
\begin{aligned}
{}_AF_B\left[\begin{matrix} a_1,\ldots,a_A \\ b_1,\ldots,b_B \end{matrix};z\right] &= {}_AF_B(a_1,\ldots,a_A;b_1,\ldots,b_B;z) \\
&= \sum_{n=0}^{\infty}\frac{(a_1)_n\ldots(a_A)_n}{(b_1)_n\ldots(b_B)_n}\frac{z^n}{n!}\,,
\end{aligned}
\tag{28.1}
$$

where

$$
(a)_n = \begin{cases} 1 & \text{if } n=0 \\ a(a+1)\ldots(a+n-1) & \text{if } n=1,2,\ldots. \end{cases}
\tag{28.2}
$$

Gasper and Rahman (1990) defined an ${}_A\phi_B$ q-hypergeometric function as

$$
\begin{aligned}
{}_A\phi_B\left[\begin{matrix} a_1,\dots,a_A \\ b_1,\dots,b_B \end{matrix}; q, z\right] &= {}_A\phi_B(a_1,\dots,a_A;b_1,\dots,b_B;q,z) \\
&= \sum_{n=0}^{\infty} \frac{(a_1;q)_n \dots (a_A;q)_n z^n}{(b_1;q)_n \dots (b_B;q)_n (q;q)_n} \left[(-1)^n q^{\binom{n}{2}}\right]^{B-A+1},
\end{aligned}
\tag{28.3}
$$

where

$$
(a;q)_n = \begin{cases} 1 & \text{if } n = 0 \\ (1-a)(1-aq)\dots(1-aq^{n-1}) & \text{if } n = 1,2,\dots \end{cases}
\tag{28.4}
$$

The Gasper-Rahman definition is a generalization of the definition which had been in use for many years and which did not include the expression in square brackets. In the important cases where $A = B+1$, the previous definition and (28.3) are identical. In particular,

$$
{}_2\phi_1(a,b;c;q,z) = 1 + \frac{(1-a)(1-b)}{(1-c)(1-q)}z + \frac{(1-a)(1-aq)(1-b)(1-bq)}{(1-c)(1-cq)(1-q)(1-q^2)}z^2 + \dots .
\tag{28.5}
$$

Note that

$$
\lim_{q\to 1} \frac{(q_a;q)_n}{(1-q)^n} = a(a+1)\dots(a+n-1) = (a)_n ,
\tag{28.6}
$$

and hence

$$
\lim_{q\to 1} {}_2\phi_1(q^a, q^b; q^c; q, z) = {}_2F_1(a,b;c;z)
\tag{28.7}
$$

(the usual Gaussian hypergeometric function).

The theory of basic hypergeometric functions arose from the study of partitions in the 18th century, but it is only recently that they have begun to be seriously applied in discrete statistical distribution theory.

Nearly thirty years ago, A. W. Kemp (1968) showed how the study of a very wide family of discrete distributions could be unified by expressing their probability generating functions (pgf's) in terms of generalized hypergeometric functions. The simple limiting relationship between q-hypergeometric and ordinary hypergeometric functions exemplified by (28.7) above suggests the study of q-analogues of the Kemp family of distributions. In the sequel, we examine an obvious q-analogue of the logarithmic distribution and indicate an application to group-size distributions.

28.2 A q-Logarithmic Distribution

The logarithmic distribution [see, for example, Johnson, Kotz and Kemp (1992, Chapter 7)] is an important discrete distribution, which has many practical applications. It is *inter alia* a member of the Kemp family of distributions. Its pgf is

$$\begin{aligned} G(s) &= \frac{\log(1-\theta s)}{\log(1-\theta)} = s\frac{{}_2F_1(1,1;2;\theta s)}{{}_2F_1(1,1;2;\theta)} \\ &= a\left\{\theta s + \frac{\theta^2 s^2}{2} + \frac{\theta^3 s^3}{3} + \dots\right\}, \end{aligned} \tag{28.8}$$

where $0 < \theta < 1$ and $a = -[\log(1-\theta)]^{-1}$.

Consider the distribution with pgf

$$\begin{aligned} H(s) &= \frac{\theta s\, {}_2\phi_1(q,q;q^2;q,\theta s)}{\theta\, {}_2\phi_1(q,q;q^2;q,\theta)} = \frac{s\, {}_2\phi_1(q,q;q^2;q,\theta s)}{{}_2\phi_1(q,q;q^2;q,\theta)} \\ &= C\left\{\theta s \sum_{x=0}^{\infty} \frac{(q;q)_x (q;q)_x}{(q^2;q)_x (q;q)_x} (\theta s)^x\right\} \\ &= C\left\{\sum_{x=0}^{\infty} \frac{(1-q)}{(1-q^{x+1})} (\theta s)^{x+1}\right\} \\ &= C\left\{\theta s + \frac{(\theta s)^2}{1+q} + \frac{(\theta s)^3}{1+q+q^2} + \frac{(\theta s)^4}{1+q+q^2+q^3} + \dots\right\}, \end{aligned} \tag{28.9}$$

where $C = [\theta\, {}_2\phi_1(q,q;q^2;q,\theta)]^{-1}$.

It is immediately apparent from (28.7) that

$$\lim_{q\to 1} H(s) = G(s), \tag{28.10}$$

hence we regard the distribution with pgf $H(s)$ as a q-analogue of the logarithmic distribution and refer to it as the q-log distribution. It is this distribution that is the main concern of this paper.

The probability mass function (pmf) of the q-log distribution is

$$p_x = \begin{cases} C\,\theta^x \Big/ \sum_{j=0}^{x-1} q^j & x = 1,2,3,\dots \\ 0 & \text{elsewhere} \end{cases}, \quad 0 < q < 1,\ 0 < \theta < 1. \tag{28.11}$$

Like those of the logarithmic distribution, the q-log probabilities $\{p_x\}$ form a strictly monotonic decreasing sequence and the distribution is reverse J-shaped.

Suppose that we shift the q-log distribution one unit to the left, so that the first non-zero probability is p_0 rather than p_1. Then, for $x > 0$,

$$\begin{aligned} \frac{p_{x-1}p_{x+1}}{{p_x}^2} &= \frac{(1-q^x)^2}{(1-q^{x-1})(1-q^{x+1})} \\ &> 1, \end{aligned} \tag{28.12}$$

since

$$(1-q^x)^2 - (1-q^{x-1})(1-q^{x+1}) = q^{x-1}(1-q)^2 > 0; \tag{28.13}$$

hence, the distribution is log-convex.

It is well-known that if the logarithmic distribution is similarly shifted then the resultant distribution is infinitely divisible. Log-convexity is a sufficient condition for infinite divisibility of any distribution for which $p_x > 0$, $x \geq 0$. Hence the shifted q-log distribution must also be infinitely divisible. Log-convexity also implies that the distribution has a decreasing hazard (failure) rate.

While the q-log distribution approaches the logarithmic distribution as $q \to 1$, when $q \to 0$ it approaches the geometric distribution with pgf

$$K(s) = s\frac{(1-\theta)}{(1-\theta s)} \tag{28.14}$$

and pmf

$$p_x = \begin{cases} (1-\theta)\theta^{x-1} & x = 1,2,3,\ldots \\ 0 & \text{elsewhere} \end{cases} ; \quad 0 < \theta < 1 \tag{28.15}$$

[for properties of the geometric distribution, see Johnson, Kotz and Kemp (1992, Chapter 5)].

If we apply a standard Heine transformation [see, for example, Gasper and Rahman (1990, p. 241)] to ${}_2\phi_1(q,q;q^2;q,\theta s)$, we can show that

$$\begin{aligned} H(s) &= C(1-q)\left\{\frac{\theta}{(1-\theta)}\frac{s(1-\theta)}{(1-\theta s)} + \frac{q\theta}{(1-q\theta)}\frac{s(1-q\theta)}{(1-q\theta s)}\right. \\ &\qquad \left. + \frac{q^2\theta}{(1-q^2\theta)}\frac{s(1-q^2\theta)}{(1-q^2\theta s)} + \frac{q^3\theta}{(1-q^3\theta)}\frac{s(1-q^3\theta)}{(1-q^3\theta s)} + \ldots\right\}. \end{aligned} \tag{28.16}$$

Hence, the q-log distribution can be interpreted as a mixture of geometric distributions where the weights are the expected values of the respective distributions.

Although there is no simple summation formula for ${}_2\phi_1(q,q;q^2;q,\theta)$, numerical computation of the $\{p_x\}$ of the q-log distribution is straightforward for most (q,θ) combinations. It is very simple to compute, store, and sum the terms $\theta^x / \sum_{j=0}^{x-1} q^j$ sequentially until individual terms become negligibly small (say, $< 10^{-8}$). Individual terms are then multiplied by the reciprocal of the sum to give the $\{p_x\}$.

While it is not possible to obtain closed expressions for the moments of the distribution, we can obtain the following expression for the mean from (28.16) by remembering that $H'(1) = E[X]$ and $H(1) = 1$:

$$E[X] = \left\{ \frac{\frac{1}{(1-\theta)^2} + \frac{q}{(1-q\theta)^2} + \frac{q^2}{(1-q^2\theta)^2} + \cdots}{\frac{1}{(1-\theta)} + \frac{q}{(1-q\theta)} + \frac{q^2}{(1-q^2\theta)} + \cdots} \right\}. \tag{28.17}$$

In the same way, we use $H''(1)$ to obtain

$$E[X(X-1)] = \left\{ \frac{\frac{2\theta}{(1-\theta)^3} + \frac{2q^2\theta}{(1-q\theta)^3} + \frac{2q^4\theta}{(1-q^2\theta)^3} + \cdots}{\frac{1}{(1-\theta)} + \frac{q}{(1-q\theta)} + \frac{q^2}{(1-q^2\theta)} + \cdots} \right\}. \tag{28.18}$$

We can proceed similarly for higher order factorial moments.

Alternatively, for any (q, θ) it is easy to compute values for the moments directly as the $\{p_x\}$ are being computed.

For constant θ, as q is *increased* from 0 to 1 the mean *decreases* from $1/(1-\theta)$ to $a\theta/(1-\theta)$ and both the variance and the third central moment also *decrease*. For a constant mean, however, as q *increases* both θ and the variance also *increase*. For example, if θ is held constant at 0.743, the mean and variance of the limiting geometric distribution are 3.89 and 11.23, while those of the limiting logarithmic are 2.13 and 3.75. But if the *mean* is held constant at 3.89 (the limiting geometric value above), then the value of θ for the limiting logarithmic distribution is 0.900 and the variance also *increases* to 23.50. For a constant mean, the distribution becomes shorter-tailed as q decreases.

Given empirical data which are assumed to have been generated by a q-log model, the ease with which the probabilities and hence the likelihood can be computed suggests the use of direct search methods for the maximum likelihood estimation of the parameters.

28.3 A Group Size Model for the Distribution

Group size models have received considerable attention in the statistical literature since Yule (1924); for information and references see, for example, Kemp and Kemp (1992) and Morgan (1976, 1993). A typical statistical model regards the group size distribution as the equilibrium distribution arising from a birth-death process. We adopt this approach. We assume that the minimum size of a group is unity, so the process starts with a group of size $x \geq 1$, and that the following birth and death rates λ_i, μ_i apply:

$$\begin{array}{lllllll} \lambda_1 & = & \lambda & \mu_1 & = & 0 & \\ \lambda_x & = & \lambda\left(\frac{1-q^x}{1-q}\right) & \mu_x & = & \mu\left(\frac{1-q^x}{1-q}\right) & x > 1, \end{array} \quad (28.19)$$

i.e., for fixed q, both the birth and death rates increase as group size increases but reducing q reduces all the rates except the first.

Standard results for a birth-death process enable us to obtain the probabilities of the equilibrium distribution of group size as

$$p_x = p_1 \frac{\lambda_1\lambda_2\cdots\lambda_{x-1}}{\mu_2\mu_3\cdots\mu_x} = p_1 \frac{\lambda^{x-1}}{\mu^{x-1}}\left(\frac{1-q}{1-q^x}\right). \quad (28.20)$$

Hence, the pgf must be

$$H(s) = sp_1 \sum_{x=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^{x-1}\left(\frac{1-q}{1-q^x}\right)s^{x-1} = sp_1\,{}_2\phi_1\left(q, q; q^2; q, \frac{\lambda}{\mu}s\right), \quad (28.21)$$

i.e., the equilibrium distribution of group size is q-logarithmic.

This model can also be reformulated in terms of a random walk whose analysis results in the same distribution.

References

1. Gasper, G. and Rahman, M. (1990). *Basic Hypergeometric Series*, Cambridge: Cambridge University Press.

2. Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate Discrete Distributions*, Second edition, New York: John Wiley & Sons.

3. Kemp, A. W. (1968). A wide class of discrete distributions and the associated differential equations, *Sankhyā, Series A*, **30**, 401–410.

4. Kemp, A. W. and Kemp, C. D. (1992). A group-dynamic model and the lost-games distribution, *Communications in Statistics—Theory and Methods*, **21**, 791–798.

5. Morgan, B. J. T. (1976). Stochastic models of grouping changes, *Advances in Applied Probability*, **8**, 30–57.

6. Morgan, B. J. T. (1993). Expected size distributions in models of group dynamics, *Journal of Applied Probability*, **30**, 1–16.

7. Yule, G. U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S., *Philosophical Transactions of the Royal Society of London*, **B213**, 21–87.

29

Bernoulli Learning Models: Uppuluri Numbers

K. G. Janardan

Eastern Michigan University, Ypsilanti, MI

Abstract: When an operator works on a machine, it is reasonable to assume that in a learning situation the probability of making an error changes from trial to trial, thus leading to a general Bernoulli process. In this paper, some results concerning two random variables - S_n, the number of errors made in a fixed number n of trials and T_r, the waiting time for the r-th success in this general Bernoulli process are studied. The probability distributions of S_n and T_r satisfy certain recursive relations and lead to interesting connections with the results found in combinatorial theory. The details of some interesting results which lead to what we call Uppuluri numbers, and in some special cases, relate to signless Stirling numbers of the first kind and to Pascal's triangle are also presented.

Keywords and phrases: Learning process, probability models, Stirling numbers, Pascal triangle

29.1 Introduction

Mathematical models for natural or social phenomena facilitate the development of any science when a sufficient body of quantitative information has been collected. This collection can be used to point out the direction in which models should be constructed and to test the adequacy of such models. These models in turn, are often used in organizing and interpreting data and in proposing new direction for experimental research. Among the areas of social sciences, psychology is one area where numerous attempts have been made to construct quantitative models for learning phenomena when learning is viewed as a product of experience.

Parvin and Grammas (1980) defined the technical progress as a learning process and examined this view in the context of the theory of learning in psychology. They developed interesting models for the production of a service or goods where man and machine are input factors. In particular, they offered a mathematical model in which they assumed the probability that an operator makes an error changes from trial to trial. In fact they assumed that the probability that the operator makes an error at the n-th trial is $1/(n+1)$ and studied the distribution of the random variable S, the number of errors made in a fixed number of trials.

In this paper, the probability distributions of the random variable S_n, which counts the number of errors made when the probability p_i changes from trial to trial, and that of the random variable T_r, the waiting time for the r-th success, are obtained in a general Bernoulli process. Parvin and Grammas results form special cases of our model. The probability distributions of S_n and T_r satisfy certain recursive relations in terms of what we call Uppuluri numbers, and lead to interesting connections with results found in combinatorial theory.

29.2 The General Model

Consider a sequence of random variables $\{X_n; n = 1, 2, \ldots\}$ that are independent and assume the values only 0 or 1. In general, the event $\{X_n = 1\}$ implies that the success has occurred on the n-th trial. In the learning model situation of the type discussed in Parvin and Grammas, the event $\{X_n = 1\}$ may be interpreted as the event that the performance of the task (job) resulted in no error on the n-th execution of the task. The event $\{X_n = 0\}$ then implies that an error has occurred on the n-th operation. Let p_n be the probability no error is made on the n-th operation (trial):

$$p_n = \Pr[X_n = 1] \tag{29.1}$$

Let q_n denote the probability that an error is made on the n-th operation:

$$q_n = \Pr[X_n = 0]. \tag{29.2}$$

Clearly $p_n + q_n = 1$ for each n. Each X_n is one of the simplest kinds of random variables, a Bernoulli random variable.

Now, for each n, define the random variable S_n as follows: Let S_n be the number of error free executions of the task (number of successes) in n operations (trials). It is of interest to find the probability distribution of S_n and other statistical properties. In particular, what can be said of $\Pr[S_n = k]$ for $k = 0, 1, \ldots, n$ (the probability of exactly k successes in n trials) and the mean $E(S_n)$ and the variance $V(S_n)$ of S_n?

It follows from the well known properties, for $0 \leq S_n \leq n$,

$$S_n \;=\; X_1 + X_n + \cdots + X_n \tag{29.3}$$

$$E(S_n) = \sum_{j=1}^{n} p_j \tag{29.4}$$

$$V(S_n) = \sum_{j=1}^{n} p_j q_j \tag{29.5}$$

A closed expression for the probability distribution of S_n does not exist in the literature. Our objective is to obtain one. However, in the case of constant success probability, $p_n = p$ for all n, the general model reduces to well known binomial distribution:

$$\Pr[S_n = k] = \binom{n}{k} p^k q^{n-k} \tag{29.6}$$

To find an expression for the general case, we first construct a recurrence relation. The event $\{S_{n+1} = k\}$ can occur in exactly one of two mutually exclusive ways. Either $\{S_n = k\}$ already occurred and a failure occurs on the $(n+1)$ trial or $\{S_n = k-1\}$ occurred and a success occurs on the n-th trial. Then using standard laws of probability, we get

$$\Pr[S_{n+1} = k] = q_{n+1} \Pr[S_n = k] + p_{n+1} \Pr[S_n = k-1] \tag{29.7}$$

In applying this, it is important to note that the probability of any S_n taking a negative value or a value greater than n is zero.

The recursion in (29.7) can be used to develop the probability distribution of S_n for small n [see Uppuluri and Piziak (1984)]. From (29.7), it can be shown that

$$\Pr[S_n = 0] = \prod_{j=1}^{n} q_j \tag{29.8}$$

and

$$\Pr[S_n = n] = \prod_{j=1}^{n} p_j \tag{29.9}$$

Now define

$$R(n,k) = \frac{\Pr[S_n = k]}{\Pr[S_n = 0]} . \tag{29.10}$$

Thus,

$$\Pr[S_n = k] = R(n,k) \prod_{j=1}^{n} q_j . \tag{29.11}$$

Some easy algebra converts the recursion in (29.7) to

$$R(n+1,k) = R(n,k) + \beta_{n+1} R(n,k-1) \tag{29.12}$$

where β_n is the odds ratio: $\beta_n = \frac{p_n}{q_n}$.

It follows from the recursion relation in (29.12)

$$\begin{cases} R(n,0) &= 1 \\ R(n,n) &= \Pi_{j=1}^{n}\beta_j \\ R(n,1) &= \sum_{j=1}^{n}\beta_j \,. \end{cases} \tag{29.13}$$

The solution of the recursion relation can be shown to be

$$U(n,k;\boldsymbol{\beta}) = \sum_{i_1<i_2<\cdots<i_k} \beta_{i_1}\beta_{i_2}\ldots\beta_{i_k} \tag{29.14}$$

with $U(n,0;\boldsymbol{\beta}) = R(n,0) = 1$. The sum in (29.14) denotes the sum of all products for every subset $\{i_1,i_2,\ldots,i_k\}$ of values from the set $\{1,2,\ldots,n\}$ such that $i_1 < i_2 < \cdots < i_k$, for $1 \leq k \leq n$. We call the expression in (29.14) Uppuluri numbers. It may be noted that if we take $\beta_j = j$ (i.e., $p_j = jq_j$), $U(n,k;\boldsymbol{\beta})$ reduces to the signless or absolute Stirling numbers of the first kind: $s(n,k) = \sum i_1 i_2 \ldots i_k$ where summation is extended over all combinations $\{i_1,i_2,\ldots,i_k\}$ of positive integers $\{1,2,\ldots,n\}$.

Even more interesting is that, given β's and the relation (29.14), a general probabilistic learning model can be constructed as

$$\Pr[S_n = k] = \varphi_n\, U(n,k;\boldsymbol{\beta}) \qquad \text{for } k = 0,1,\ldots,n \tag{29.15}$$

where $\varphi_n = \Pi_{j=1}^{n} q_j$ and $U(n,0;\boldsymbol{\beta}) = 1$.

If we define γ_i as reciprocal of β_i, i.e., $\gamma_i = \frac{q_i}{p_i}$ for all i, then we can also write the probability function of S_n as

$$\Pr[S_n = k] = \theta_n\, U_1(n,k;\boldsymbol{\gamma}) \tag{29.16}$$

where $\theta_n = \prod_{i=1}^{n} p_i$ and $U_1 = \sum_{i_1<i_2<\cdots<i_{n-k}} \gamma_{i_1}\gamma_{i_2}\cdots\gamma_{i_{n-k}}$.

It may be noted here that the Uppuluri numbers $U_1(n,k;\boldsymbol{\gamma})$ are calculated from the elements which are reciprocals of the ones used in the definition of $U(n,k;\boldsymbol{\beta})$.

29.2.1 Special cases of the general probabilistic model

The Parvin-Grammas Model

If we take $p_i = \frac{i}{i+1}$ and $q_i = \frac{1}{i+1}$ for $i = 1,2,\ldots n$, then the mean and variance of S_n, are given by

$$\begin{cases} \mu_n &= \sum_{j=1}^{n} p_j &= \sum_{j=1}^{n} \frac{j}{j+1} \\ \sigma_n^2 &= \sum_{j=1}^{n} p_j q_j &= \sum_{j=1}^{n} \frac{j}{(j+1)^2} \,. \end{cases} \tag{29.17}$$

These expressions are the same as those obtained in Parvin and Grammas (1980).

In this case, the recursion in (29.12) reduces to

$$s(n+1,k) = s(n,k) + (n+1)s(n,k-1) \tag{29.18}$$

where $s(n,k)$ is the signless Stirling numbers of the first kind, and the general learning model (29.15) reduces to

$$\Pr[S_n = k] = \frac{1}{(n+1)!}\sum i_1 i_2 \ldots i_k. = \frac{s(n,k)}{(n+1)!} \tag{29.19}$$

This shows that the model considered by Parvin and Grammas leads to signless Stirling numbers of the first kind.

The Binomial Model

If we assume no learning occurs from trial to trial, the probabilities p_i are constant, i.e., $p_i = p$, for all i, then the recursion relation in (29.12) reduces to

$$U(n+1,k;\beta) = U(n,k;\beta) + \beta U(n,k-1;\beta) \tag{29.20}$$

where $\beta = \frac{p}{q}$ and $U(n,k;\beta)$ in (29.14) reduces to

$$U(n,k;\beta) = \sum_{j=1}^{\binom{n}{k}} \left(\frac{p}{q}\right)^k \tag{29.21}$$

and the general learning model reduces to the binomial model as in (29.6). If β is taken as equal to 1 in the recurrence relation (29.20), (i.e., when $p = q = 1/2$), (29.20) reduces to the familiar Pascal's triangle identity:

$$\binom{n+1}{k} = \binom{n}{k} + \binom{n}{k-1}. \tag{29.22}$$

The First Generation Rumor Model

If we take $p_i = \theta/(\theta + i)$ and hence $\gamma_i = i/\theta$ in (29.16), then the probability function reduces to

$$\Pr[S_n = k] = s(n,k)\,\frac{\theta^k}{\theta^{[n]}}\,, \qquad \text{for } k = 1,2,\ldots,n \tag{29.23}$$

where $\theta^{[n]} = \theta(\theta+1)(\theta+2)\ldots(\theta+n-1)$. Let n be the total number of persons hearing a rumor, with

$$\theta = \frac{\text{Intensity of source transmission}}{\text{Intensity of between individual transmission}}$$

then the number of those who have heard the rumor from the source will have the probability model given by (29.23) [see Bartholomew (1982)].

The Random Record Model

Let $Y_0, Y_1, \ldots$ be independent and identically distributed random variables with common distribution function F. Define the upper record X_j by letting

$$X_j = \begin{cases} 1 & \text{if } Y_i > Y_k \text{ for } k = 0, 1, 2, \ldots, j-1 \\ 0 & \text{otherwise} \end{cases}$$

Then X_j's are independent and $\Pr[X_j = 1] = \frac{1}{j+1}$. Let S_n be the number of records up to n. Since $S_n = \sum_{j=1}^{n} X_j$, the probability distribution of S_n is [see Westcott (1977)] same as (29.19).

29.3 Waiting Time Learning Models

Let us now consider the general case of the waiting times learning models with arbitrary choice of probabilities from operation to operation (trial). Specifically, assume

$$\Pr[\text{an error is made on the } n\text{-th operation}] = q_n, \qquad 0 < q_n \leq 1,$$

and

$$\Pr[\text{no error is made on the } n\text{-th operation}] = p_n, \qquad q_n + p_n = 1.$$

Let the random variable T_r denote the waiting time (i.e., number of machine operations required before) the r-th error-free execution of a task. From (29.3) we know that S_n denotes the number of error-free executions in n trials. Thus,

$$\Pr[T_{r+1} = n+1] = \Pr[S_n = r] p_{n+1} \tag{29.24}$$

or

$$\frac{\Pr[T_{r+1} = n+1]}{p_{n+1}} = \Pr[S_n = r] \tag{29.25}$$

where $\Pr[S_n = r]$ satisfies the recurrence relation:

$$\Pr[S_n = r] = q_n \Pr[s_{n-1} = r] + p_n \Pr[S_{n-1} = r-1]. \tag{29.26}$$

Applying (29.25) to (29.26) we can write (29.26) as

$$\frac{\Pr[T_{r+1} = n+1]}{p_{n+1}} = \frac{q_n}{p_n} \Pr[T_{r+1} = n] + \Pr[T_r = n] \tag{29.27}$$

and

$$\Pr[T_r = r] = p_1 p_2 \ldots p_r. \tag{29.28}$$

By taking $n = r + 1$ in (29.27), we can write (29.27) as

$$\frac{\Pr[T_{r+1} = r + 2]}{p_{n+2}} = p_1 p_2 \ldots p_{r+1} q_{r+1} + p_{r+2} \Pr[T_r = r + 1] \qquad (29.29)$$

for $r = 1, 2, 3, \ldots$

Successive evaluation of probabilities in (29.29) in terms of p's and q's allows reduction of (29.29), eventually, to

$$\begin{aligned} \Pr[T_r = r + 1] &= p_1 p_2 \ldots p_{r+1} \left(\frac{q_1}{p_1} + \frac{q_2}{p_2} + \cdots + \frac{q_r}{p_r} \right) \\ &= \theta_{r+1}(\gamma_1 + \gamma_2 + \cdots + \gamma_r) \\ &= \theta_{r+1} \sum_{j=1}^{r} \gamma_j \end{aligned} \qquad (29.30)$$

where $\theta_{r+1} = \Pi_{j=1}^{r+1} p_j$ and $\gamma_i = \frac{q_i}{p_i}$, for $i = 1, 2, \ldots, r$.

Similarly, we can show that

$$\Pr[T_r = r + 2] = \theta_{r+2} \sum_{i_1 < i_2}^{\binom{r+1}{2}} \gamma_{i_1} \gamma_{i_2}. \qquad (29.31)$$

Next we begin developing interesting connections with results in combinatorial theory. A look at the expressions in (29.30) and (29.31) suggests two ideas. First, there is an algorithm for constructing the probability distribution, $\Pr[T_r = r + m]$ for any m; and second, the number of summands in each case, are $\binom{r+m-1}{m}$. For $m = 1$, we get the number of summands in (29.30) as r and for $m = 2$, the number of summands in (29.31) as $\binom{r+1}{2}$. Thus,

$$\Pr[T_r = r + m] = \theta_{r+m} \sum_{i_1 < i_2 < \cdots < i_m} \gamma_{i_1} \gamma_{i_2} \ldots \gamma_{i_m} \qquad (29.32)$$

where the number of terms in the sum is $\binom{r+m-1}{m}$.

Hence, to pursue the interesting connection between the waiting time learning models and certain classes of numbers in combinatorial theory, we restate the Uppuluri numbers as defined in (29.14):

$$U(n, m; \gamma) = \sum_{i_1 < i_2 < \cdots < i_m} (\gamma_{i_1} \gamma_{i_2} \ldots \gamma_{i_m}) \qquad (29.33)$$

where the sum denotes the sum of all products $\gamma_{i_1} \gamma_{i_2} \ldots \gamma_{i_m}$ for every subset $\{i_1, i_2, \ldots, i_m\}$ of values from the set $\{1, 2, \ldots, n\}$ such that $i_1 < i_2 < \cdots < i_m$, $1 \leq m \leq n$.

Therefore, the probability distribution of T_r can be defined as

$$\Pr[T_r = k] = \theta_k U(k - 1, k - r; \gamma) \qquad (29.34)$$

where $\theta_k = \Pi_{j=1}^{k} p_j$, $U(k - 1, k - r; \gamma)$ is as defined in (29.33) , and γ_j's are the odds ratios, $\gamma_j = \frac{q_j}{p_j}$ for all j.

29.3.1 Special cases of waiting time learning models

The Parvin and Grammas Model

If we make Parvin and Grammas assumption that $p_k = \frac{k}{k+1}$ and $q_k = \frac{1}{k+1}$ for $k = 1, 2, \ldots$, then $\theta_k = \frac{k!}{(k+1)!}$ and $\gamma_k = \frac{1}{k}$. Hence, the probability distribution in (29.34) reduces to

$$\Pr[T_r = k] = \frac{1}{(k+1)} \sum_{i_1 < i_2 < \cdots < i_{k-r}} \frac{1}{i_1 i_2 \ldots i_{k-r}} . \tag{29.35}$$

From the properties of Stirling numbers of the first kind, [see Charalambides and Singh (1988)] we have

$$(k-1)! \sum_{i_1 < i_2 < \cdots < i_{k-r}} \frac{1}{i_1 i_2 \ldots i_{k-r}} = \sum_{i_1 < i_2 \ldots i_{r-1}} i_1 i_2 \ldots i_{r-1} = s(k-1, r-1) \tag{29.36}$$

Substituting (29.36) into (29.35) we obtain

$$\Pr[T_r = k] = \frac{1}{(k-1)!(k+1)} \, s(k-1, r-1) \tag{29.37}$$

This result shows that the distribution of T_r, under Parvin and Grammas assumption, is again related to the absolute Stirling numbers of the first kind. The first few terms of $s(n, m)$ are [see Parvin and Grammas (1980)]:

$$\begin{cases} s(n,1) &= \frac{n(n+1)}{2} \\ s(n,2) &= \frac{n(n+1)}{2} \frac{(n-1)(3n+2)}{12} \\ s(n,3) &= \left[\frac{n(n+1)}{2}\right]^2 \frac{(n-1)(n-2)}{12} \end{cases} \tag{29.38}$$

Using (29.38) we can obtain the probability distribution of T_1 the waiting time for the first success as

$$\Pr[T_1 = k] = \frac{1}{(k-1)!(k+1)} = \frac{k}{(k+1)!} \, , \text{ for } k = 1, 2, \ldots \tag{29.39}$$

The probability distribution of T_2, the waiting time for 2 successes can be shown to be

$$\Pr[T_2 = k] = \frac{k^2(k-1)}{2(k-2)!(k+1)} \, , \text{ for } k = 2, 3, \ldots \tag{29.40}$$

These results agree with those obtained by Uppuluri and Janardan (1985). The expected values of T_1 and T_2 can be derived as $E(T_1) = (e-1)$ and $E(T_2) = \frac{3}{2}\, e - 1$. The variances of T_1 and T_2 can be shown to be $V(T_1) = 3e - e^2$ and $V(T_2) = e(26 - 9e)/4$.

Similarly, the probability distribution of T_3, the waiting time for the third success can be derived as

$$\Pr[T_3 = k] = \frac{k^2(k-1)(k-2)(3k-1)}{24(k+1)!}, \quad \text{for } k = 3, 4, \ldots \tag{29.41}$$

and $E(T_3) = 4.3233$ (using Maple software).

The Random Record Waiting Time Model

In the record model considered in Section 29.2.1, let T_r denote the waiting time before the r-th upper record is obtained. Then the probability distribution of T_r is as given in (29.37).

The Negative Binomial Model

If we assume that no learning occurs from trial to trial, the probabilities p_i are constant, i.e., $p_i = p$, for all i. Then, in (29.33) $\gamma = \frac{q}{p}$, $\theta_k = p^k$ and

$$U(k-1, r-1; \gamma) = \sum_{j=1}^{\binom{k-1}{r-1}} (q/p)^{k-r} = \binom{k-1}{r-1} q^{k-r} p^{-k+r}. \tag{29.42}$$

Thus, the distribution of the random variable T_r reduces to the negative binomial distribution:

$$\Pr[T_r = k] = \binom{k-1}{r-1} p^r q^{k-r} \qquad \text{for } k = r, r+1, \ldots. \tag{29.43}$$

The Waiting Time Model for the First Generation Rumor

In the first generation rumor model considered in Section 29.2.1, let T_r denote the number of people required to be sampled before the r-th person who has heard the rumor from the source is found, then the probability distribution of T_r is given by

$$\Pr[T_r = k] = \frac{k\theta^{k-r}}{(\theta+1)^{[k]}} s(k-1, r-1), \quad \text{for } k = r, r+1, \ldots \tag{29.44}$$

The special cases discussed above are only a few among several of the learning models of interest. Based on the general results in (29.15) and (29.34), one can develop several other learning models. These are left to the reader.

References

1. Bartholomew, D. J. (1982). *Stochastic Models for Social Processes*, Third edition, New York: John Wiley & Sons.

2. Charalambides, C. A. and Singh, J. (1988). A review of the Stirling numbers, their generalizations and statistical applications, *Communications in Statistics–Theory and Methods*, **17**, 2533–2595.

3. Parvin, M. and Grammas, G. M. (1980). On a probabilistic learning model, *Policy Analysis and Information Systems*, **4**, 47–51.

4. Riordan, J. (1958). *An Introduction to Combinatorial Analysis*, New York: John Wiley & Sons.

5. Uppuluri, V. R. R. and Janardan, K. G. (1985). Man and machine: Waiting times in learning models, *Journal of Organizational Behaviour and Statistics*, **2**, 163–176.

6. Uppuluri, V. R. R. and Piziak, R. (1984). General probabilistic learning models, *Policy and Information*, **8**, 71–84.

7. Westcott, M. (1977). The random record model, *Proceedings of the Royal Society of London, Series A*, **356**, 529–547.

PART VII

Applications to Nonparametric Statistics

30

Linear Nonparametric Tests Against Restricted Alternatives: The Simple-Tree Order and The Simple Order

S. Chakraborti and W. Schaafsma

University of Alabama, Tuscaloosa, AL
University of Groningen, Groningen, The Netherlands

Abstract: The problem of testing homogeneity of distributions against some ordered alternatives is considered without making specific parametric model assumptions. The alternatives of interest include the simple-tree order and the simple order. The approach is based on a reformulation of the testing problems, regarding them as being composed of a finite number of two-sample sub-testing problems. The Mann-Whitney-Wilcoxon test is used, as an example, for the component problems, the main issue being how to combine dependent test statistics into a single efficient overall test. Keeping the practitioner's needs in mind, a linear function of the individual test statistics is considered and the Abelson-Tukey-Schaafsma-Smid principle (of minimizing "the maximum shortcoming") is applied to derive the optimal (minimax) coefficients in the linear combination. The advantages of using optimal weights are illustrated by making ARE comparisons with the Jonckheere-Terpstra-Tryon-Hettmansperger test. The restriction to the class of linear combinations is attractive from the practitioner's viewpoint but the theoretical statistician will argue that such linear combinations leading to "somewhere" most powerful tests makes them questionable from an overall point of view. Hence, alternative approaches are discussed, at least to some extent. The main purpose of the paper, however, is to assist the practitioner who has decided to use a linear combination but worries about the weights to be chosen. It is argued that the restriction by linearity is less questionable for the simple order than for the simple-tree order.

Keywords and phrases: Nonparametric, one-way data, patterned alternatives, many-to-one problem, upward trend problem, most stringent somewhere most powerful test, combination of dependent tests; positive orthant

30.1 Introduction

Let $X_{01}, X_{02}, ..., X_{0n_0}, ..., X_{k1}, X_{k2}, ..., X_{kn_k}$ be $(k+1)$ independent random samples, where $X_{hj}, j = 1, 2, ..., n_h$, are from a population with a continuous cumulative distribution function F_h, $h = 0, 1, ..., k$. An observation X_{hj} may represent, for example, the response measured on the j-th unit receiving the h-th treatment. In some applications, especially in the development of new drugs, there is often a control or a standard treatment with some special significance. In such situations, we shall use the index 0 to denote the control treatment.

In this paper, we consider testing problems associated with the hypothesis that the treatments are homogeneous against some directional alternatives. Thus, the null hypothesis

$$H_0 : F_0 = F_1 = ... = F_k$$

is tested against some "one-sided" alternative hypothesis. Although our approach is more generally applicable, we focus, for clarity, on the two one-sided alternatives to homogeneity that have received a great deal of attention in the literature. The first of these concerns the situation where one of the treatments is a control. The remaining k treatments are compared with the control so as to ascertain whether any of these is better than the control, when it is known, a-priori, that none of these can be any worse. This is commonly known as the "comparison of treatments with a control" or the "many-to-one" problem. We assume that "better" means "higher" and formulate the testing problems as follows.

Testing Problem 1: Test H_0 against

$$H_1 : F_0 \geq F_1, F_0 \geq F_2, ..., F_0 \geq F_k,$$

with at least one strict inequality. This means that under the alternative there exists at least one treatment that is better than the control, in other words, there exists at least one $i = 1, 2, ..., k$ such that $F_0(x) > F_i(x)$, for at least one real x.

The second problem we consider may be relevant in situations where it is postulated, a-priori, that the distributions are in an increasing order, say from 0 to k. In the literature, this is often referred to as the "upward trend" problem.

Testing Problem 2: Test H_0 against

$$H_2 : F_0 \geq F_1 \geq F_2 \geq ... \geq F_k,$$

with at least one strict inequality. Under the upward trend alternative, X_{hi} is stochastically smaller than $X_{h+1,j}$ for $i = 1, 2, ..., n_h, j = 1, 2, ..., n_{h+1}$ and $h = 0, 1, ..., k-1$.

30.2 Background

The testing of H_0 is usually considered with the "unrestricted" alternative H_3 in mind (explicitly or implicitly) where $H_0 \cup H_3$ implies that

$$F_{\pi(0)} \geq F_{\pi(1)} \geq ... \geq F_{\pi(k)}$$

holds for some permutation π of $\{0, 1, ..., k\}$. The Kruskal-Wallis test is one of the possible tests in this case. The problem (H_0, H_2) can then be regarded as a modification of (H_0, H_3), requiring a trend analogue of the Kruskal-Wallis test (see Remark 30.6.3). The restricted alternatives H_1 and H_2 are related in the sense that H_1 is much wider than H_2 or, in logical terms, H_2 implies H_1. These restricted alternatives should conform with the experimenter's prior knowledge which, of course, may be more or less vague. It is a common belief, however, that this vagueness does not exclude the possibility that violations of certain almost logical inequalities are impossible. These inequalities are incorporated in the inference mechanism by using them to define the parameter space. Since the parameter space under a specific set of inequalities is smaller than that induced by the global alternative, one should be able to improve upon the classical solution in the sense that the power of the test is increased, at least for the major part of the alternative hypothesis.

During the fifties and the sixties, there was considerable interest in such testing problems and many ad hoc procedures were proposed. Bartholomew (1961) gave a survey emphasizing the use of the likelihood-ratio principle, which, unfortunately, does not lead to easy null distributions. Many ad hoc proposals have been based on test statistics of a linear type. These tests were identified as "somewhere most powerful" in Schaafsma and Smid (1966). A minimax principle for determining the optimum weights in such a linear test statistic was explored in Abelson and Tukey (1963). In terms of the Neyman-Pearson theory, their "maximin-r" principle becomes that of minimizing the maximum shortcoming. Hence, maximin linear statistics correspond to most stringent somewhere most powerful tests. One might prefer to construct the most stringent test in the class of *all* level-α tests. Such constructions were made by Van Zwet and Oosterhoff (1967) and Schaafsma (1968, 1971). If the alternative hypothesis is "not too wide", then being somewhere most powerful is not too harmful and the best linear test may be preferable over the most stringent, as well as over the likelihood-ratio test—its maximum shortcoming will be larger than that of its competitors. But in the greater part of the interior of the alternative, especially in the neighborhood of the parameter values where the most stringent linear test is most powerful, the power properties of the best linear test may be preferable. Much depends on the wideness of the alternative. We shall argue that H_1 may be too wide while H_2 is sufficiently narrow to wholeheartedly recommend the use of such best linear tests.

There is a considerable literature about testing problems with inequality constraints on the parameter space. The reader might consult the monographs of Schaafsma (1966), Oosterhoff (1969), Barlow *et al.* (1972), Snijders (1979), Robertson, Wright and Dykstra (1988) and Akkerboom (1990). The last mentioned reference is about a compromise between various principles: the attention is restricted to a class of manageable "circular likelihood-ratio" tests and from this class the most stringent one is chosen. Circular likelihood-ratio tests have been advanced by Pincus (1987). The problem of comparing treatments with a control has been studied in some of these references as an application of the "simple-tree" order whereas the problem of upward trend has been studied under the heading of "simple" order.

In the present, paper a less ambitious perspective is chosen. We assume that the practitioner has decided to use a linear test because he/she wants to have a complete understanding of the null distribution "without having to study too much literature." Concerning the choice of weights in the linear test statistic, he/she might make a subjective choice based on the knowledge of the problem or he/she might consult the present paper where a choice is recommended on the basis of the special mathematical form of the alternative.

30.3 Objectives

Our objective is to consider a relatively new approach to the Testing Problems 1 and 2. The procedures to be developed for Problem 1 are somewhat questionable whereas those for Problem 2 can be firmly recommended. The approach is based on a reformulation of the testing problems, regarding them as being composed of a finite number (k) of sub-testing problems, each of which is a two-sample problem. This is appealing since suitable two-sample test statistics are available for the sub-testing problems. However, these statistics are dependent and thus the main question becomes how to combine the k dependent component test statistics to construct one efficient overall test for testing H_0 against H_1 (or H_2). The approach leads to relatively easy elaborations and a deeper understanding of some theory already available.

30.4 Exploration and Reformulation

It is useful to note that for the simple-tree problem, the null hypothesis H_0 and the model $H_0 \cup H_1$ can be regarded as the intersections of the sub-null hypotheses, $H_{01}, H_{02}, ..., H_{0k}$ and the sub-models $H_{01} \cup H_{11}, ..., H_{0k} \cup H_{1k}$, respectively, where

$$H_{0h} : F_0 = F_h \text{ and } H_{1h} : F_0 \geq F_h, \quad \text{with } \; F_0(x) > F_h(x) \text{ for some } x.$$

For the upward trend problem, the null hypothesis H_0 and the alternative hypothesis H_2 can similarly be regarded as the intersections of the sub-null and the sub-alternative hypotheses, $H_{01}, H_{02}, ..., H_{0k}$ and $H_{21}, H_{22}, ..., H_{2k}$, respectively, where

$$H_{0h} : F_{h-1} = F_h \text{ and } H_{2h} : F_{h-1} \geq F_h, \text{with } \; F_{h-1}(x) > F_h(x) \text{ for some } x.$$

It has been suggested that the choice of these sub-problems (H_{0h}, H_{1h}) and (H_{0h}, H_{2h}) is somewhat intuitive and that alternative choices might lead to different theory and solutions. We agree to some extent with this criticism. The alternative choices to be discussed, however, are not very natural. The purely formal mathematical-statistical approaches discussed in Schaafsma (1968) or Snijders (1979) are either inexact (the first reference) or mathematically involved and complicated (the second reference). We do not claim that the present paper is the first that considers the optimal choice of weights in a nonparametric setting. This problem has already been discussed in some papers including Johnson and Mehrotra (1971) and Tryon and Hettmansperger (1973). The two testing problems are treated separately as follows.

30.5 Test for the Simple-Tree Problem

The problem of testing H_{0h} against H_{1h} is a two-sample problem with a one-sided alternative, between samples 0 and h, with n_0 and n_h observations, respectively. Intuitively, a test for this problem should be based on the two samples $(X_{01}, ..., X_{0n_0})$ and $(X_{h1}, ..., X_{hn_h})$. In this paper, we assume that the experimenters do not wish to make any parametric model assumptions about the underlying distributions and therefore a nonparametric test is the preferred option, probably one based on the ranks. The most popular two-sample rank test is the Wilcoxon rank-sum test which can be based on the Mann-Whitney statistic:

$$T_h = \sum_{r=1}^{n_h} \sum_{s=1}^{n_0} I(X_{hr} \geq X_{0s}), \quad h = 1, 2, ..., k.$$

The Mann-Whitney statistics will be used to exemplify the proposed theory. Other two-sample statistics (such as Chernoff-Savage statistics) can be chosen as well, but they will require some adaptation. Our approach leads to an overall test statistic which is a linear combination of $T_1, T_2, ..., T_k$, say $\Sigma_{h=1}^{k} c_h T_h$. The choice of the weights $c_1, c_2, ..., c_k$ is the main issue. Note, however, that such linear tests may have poor power characteristics for some sub-alternatives at or near the "edges" of the space defined by the alternative. As this alternative is very wide, linear tests for this problem are subject to considerable criticism. This is why considerable attention has been paid in the literature to some non-linear types of tests. Nevertheless, as far as applications are concerned, practitioners usually prefer the linear tests due to their ease of implementation and interpretation.

The moments of T_h, under the null hypothesis, can be obtained using standard methods and are

$$E[T_h] = n_h n_0/2, \qquad \mathrm{var}(T_h) = n_h n_0(n_h + n_0 + 1)/12,$$

and

$$\mathrm{cov}(T_g, T_h) = n_g n_h n_0/12.$$

It will be convenient to consider the standardized two-sample statistics

$$S_h = \frac{T_h - (n_h n_0/2)}{\sqrt{n_h n_0(n_h + n_0 + 1)/12}},$$

$h = 1, 2, ..., k$. It is easy to see that

$$\mathrm{cov}(S_g, S_h|H_0) = \mathrm{corr}(S_g, S_h|H_0) = \gamma_{gh} = \begin{cases} 1 & \text{if } g = h \\ a_g a_h & \text{if } g \neq h, \end{cases} \tag{30.1}$$

where

$$a_h = n_h^{1/2}(n_0 + n_h + 1)^{-1/2}.$$

With respect to the alternative hypothesis, for any choice of (F_0, F_h), the expectation of S_h is given by $E[S_h] = \sqrt{12 n_0} a_h \pi_h = \mu_h$, say, where

$$\pi_h = P(X_{hr} \geq X_{0s}) - 1/2 = \int F_0 dF_h - 1/2 = \int (1 - F_h) dF_0 - 1/2.$$

The distribution of $\boldsymbol{T} = (T_1, T_2, ..., T_k)'$ (hence of $\boldsymbol{S} = (S_1, S_2, ..., S_k)'$) can be approximated, in large samples, using results in the literature of nonparametric statistics [see for example, Puri (1965)]. It follows that for any fixed alternative $F_0, F_1, ..., F_k$, the distribution of $\boldsymbol{S}$ can be approximated by a k-variate normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)'$ and some covariance matrix. Under H_0, the covariance matrix is $\Gamma = ((\gamma_{gh}))$, given in (30.1). In terms of the standardized statistics, the testing problem H_0 against H_1 can be restated (at least globally) as testing

$$H_0^* : \boldsymbol{\mu} = \mathbf{0} \quad \text{against} \quad H_1^* : \boldsymbol{\mu} \geq \mathbf{0}, \quad \text{with} \quad \boldsymbol{\mu}'\boldsymbol{\mu} \neq 0, \tag{30.2}$$

which resembles the well-known combination of tests problem, the difference being that the test statistics are now dependent.

A useful approximation is obtained by making the assumption that $\boldsymbol{S}$ has a $N_k(\boldsymbol{\mu}, \Gamma)$ distribution, where Γ is exactly given by (30.1) and H_0^* is tested against H_1^*. This is an important step and a justification of this approach, for large samples, can be given via contiguity theory [the reader is referred to Snijders (1979) for details]. Intuitively, when sample sizes are large, $\boldsymbol{S}$ will be approximately normally distributed with covariance matrix Γ under a sequence of contiguous alternatives. Alternatives $(F_0, F_1, ..., F_k)$ are contiguous (to the hypothesis H_0) if the difference between F_h and F_0 is of the order of $(n_h + n_0)^{-1/2}, h = 1, 2, ..., k$.

The reader will have noticed that intuitive arguments are involved here. It is indeed true that H_0^* and H_1^* are "practically" but not logically equivalent to H_0 and H_1, the practical equivalence being generated by the restriction to the statistics $\boldsymbol{T}$ or $\boldsymbol{S}$.

A canonical form

For motivational purposes, the testing problem H_0^* against H_1^* in terms of $\boldsymbol{S} \sim N_k(\boldsymbol{\mu}, \Gamma)$ is transformed such that a "canonical form" appears with $\Gamma = I_k$, the $k \times k$ identity matrix. Since Γ is nonsingular, there exists a matrix $B = (b_1, b_2, ..., b_k)$ with columns b_h such that $B\Gamma B' = I_k$ and hence $B'B = \Gamma^{-1}$. Defining $\boldsymbol{V} = B\boldsymbol{S}$, we arrive at the canonical formulation where $\boldsymbol{V} \sim N_k(\boldsymbol{\xi}, I_k)$ and $H : \boldsymbol{\xi} = \boldsymbol{0}$ has to be tested against $A : \boldsymbol{\xi} \in K$, where K denotes the convex polyhedral cone

$$\{\Sigma_{h=1}^k \xi_h b_h; \xi_1, \xi_2, ..., \xi_k \geq 0\},$$

spanned by the columns of B. An advantage of the treatment in this canonical formulation is that we can discuss the problem using usual (Euclidean) geometrical arguments. Disadvantages are that a more or less arbitrary matrix B is involved and that the solutions need to be reformulated in terms of $\boldsymbol{S} = B^{-1}\boldsymbol{V}$.

Remark 30.5.1 Recall that the null covariance matrix of $\boldsymbol{S}$ is Γ, where $\Gamma_{gg} = 1$ and $\Gamma_{gh} = a_g a_h$, as mentioned earlier. The matrix Γ can be expressed as $\Delta + \boldsymbol{a}\boldsymbol{a}'$ where Δ is the diagonal matrix with $\Delta_{hh} = 1 - a_h^2$ and $\boldsymbol{a} = (a_1, a_2, ..., a_k)'$. It can be shown that the inverse of the matrix Γ is

$$\Gamma^{-1} = \Delta^{-1} - \frac{\Delta^{-1}\boldsymbol{a}\boldsymbol{a}'\Delta^{-1}}{1 + \boldsymbol{a}'\Delta^{-1}\boldsymbol{a}} \quad . \tag{30.3}$$

Some useful simplifications result by noting the following. First,

$$\Delta^{hh} = (1 - a_h^2)^{-1}. \tag{30.4}$$

Secondly, $\boldsymbol{a}'\Delta^{-1}$ is the row vector with h-th element $a_h(1 - a_h^2)^{-1}$ and hence $\Delta^{-1}\boldsymbol{a}\boldsymbol{a}'\Delta^{-1}$ is a $k \times k$ matrix with diagonal element $a_h^2(1 - a_h^2)^{-2}$, $h = 1, 2, ..., k$.

Finally,

$$a'\Delta^{-1}a = \Sigma_{h=1}^{k} a_h^2(1-a_h^2)^{-1}. \tag{30.5}$$

Noting that $a_h = n_h^{1/2}(n_0+n_h+1)^{-1/2}$, we then have $\Delta^{hh} = (n_0+1)^{-1}(n_h+n_0+1)$, $a_h(1-a_h^2)^{-1} = (n_0+1)^{-1}n_h^{1/2}(n_h+n_0+1)^{1/2}$, $a_h^2(1-a_h^2)^{-1} = (n_0+1)^{-1}n_h$, so that $a'\Delta^{-1}a = (n_0+1)^{-1}\Sigma_{h=1}^{k} n_h$. Therefore, from (30.3), the (h,h)-th element of the inverse of Γ is

$$\Gamma^{hh} = (n_0+1)^{-1}(N+1)^{-1}(n_0+n_h+1)(N-n_h+1), \quad h = 1,2,...k, \tag{30.6}$$

where $N = \Sigma_{h=0}^{k} n_h$ is the total sample size. These expressions will be useful in computations later on.

30.5.1 Some particular cases

Case 1

For illustrative purposes, first consider the problem of testing $H_0^* : \boldsymbol{\mu} = \mathbf{0}$ against some fixed alternative $A_1^* : \boldsymbol{\mu} = \boldsymbol{\mu}_{(1)}$, on the basis of $\boldsymbol{S}$, where $\boldsymbol{\mu}_{(1)} \geq \mathbf{0}$. In the canonical form, this corresponds to testing $\boldsymbol{\xi} = B\boldsymbol{\mu} = \mathbf{0}$ against $\boldsymbol{\xi} = \boldsymbol{\xi}_{(1)} = B\boldsymbol{\mu}_{(1)} \in K$. In this case, applying the Neyman-Pearson Fundamental Lemma, it is easy to show that the most powerful size α test rejects H_0 in favor of A_1 if

$$\left(\frac{\boldsymbol{\xi}_{(1)}}{\|\boldsymbol{\xi}_{(1)}\|}, \boldsymbol{V}\right) \geq z_\alpha, \tag{30.7}$$

where z_α is the upper α-th quantile of the standard normal distribution and (x,y) and $\|.\|$ denote the Euclidean inner product and norm, respectively. The above test can be expressed as

$$\frac{\boldsymbol{\mu}'_{(1)}B'\boldsymbol{V}}{\|B\boldsymbol{\mu}_{(1)}\|} = \frac{\boldsymbol{\mu}'_{(1)}\Gamma^{-1}\boldsymbol{S}}{(\boldsymbol{\mu}'_{(1)}\Gamma^{-1}\boldsymbol{\mu}_{(1)})^{1/2}} \geq z_\alpha. \tag{30.8}$$

Thus, the optimal (most powerful) size α test is based on a linear function of the standardized two-sample statistics.

Case 2

Next consider the case where the experimenter has a less specific alternative in mind than in Case 1. If, for example, $F_1 = F_2 = ... = F_k$ is considered relevant, because the treatments are only slight variations of a given one whereas the control corresponds to a placebo, then one may be interested in testing H_0 against $F_0 \geq F_1 = F_2 = ... = F_k$. In this situation, one may decide to pool the k treatment samples together and calculate the usual "two-sample" Mann-Whitney statistic between the pooled sample and the control sample. It can be seen that this statistic is given by $\Sigma_{h=1}^{k} T_h$. Intuitively, any reasonable approach via $\boldsymbol{S}$ and $\boldsymbol{BS}$ should result (at least asymptotically) in the same solution. That this is the case with our approach is now illustrated.

If $F_1 = F_2 = ... = F_k$, then $\mu_{(1)} = \tau \boldsymbol{a}$, where $\tau = (12n_0)^{1/2}\{\int F_0 dF_h - 1/2\}$. Hence, $\boldsymbol{\xi}_{(1)} = \tau \boldsymbol{Ba}$ and the linear test statistic in (30.8) simplifies to

$$\boldsymbol{a}'\Gamma^{-1}\boldsymbol{S}/(\boldsymbol{a}'\Gamma^{-1}\boldsymbol{a})^{1/2}.$$

It follows from (3) that $\boldsymbol{a}'\Gamma^{-1} = \boldsymbol{a}'\Delta^{-1}/(1+\boldsymbol{a}'\Delta^{-1}\boldsymbol{a})$ and hence $\boldsymbol{a}'\Gamma^{-1}\boldsymbol{a} = \boldsymbol{a}'\Delta^{-1}\boldsymbol{a}/(1+\boldsymbol{a}'\Delta^{-1}\boldsymbol{a})$. Thus, we obtain

$$\boldsymbol{a}'\Delta^{-1}\boldsymbol{S}/\{\boldsymbol{a}'\Delta^{-1}\boldsymbol{a}(1+\boldsymbol{a}'\Delta^{-1}\boldsymbol{a})\}^{1/2}, \tag{30.9}$$

which is a linear function of the standardized two-sample statistics S_h.

Using the fact that for $a_h = n_h^{1/2}(n_h + n_0 + 1)^{-1/2}$, $\boldsymbol{a}'\Delta^{-1}\boldsymbol{S} = \Sigma_{h=1}^k a_h(1-a_h^2)^{-1}S_h = \Sigma_{h=1}^k (n_0+1)^{-1} n_h^{1/2}(n_h+n_0+1)^{1/2}S_h$ and that $\boldsymbol{a}'\Delta^{-1}\boldsymbol{a} = \Sigma_{h=1}^k\{a_h^2(1-a_h^2)^{-1}\} = (n_0+1)^{-1}\Sigma_{h=1}^k n_h$, the statistic in (30.9) further simplifies to

$$\frac{\Sigma_{h=1}^k n_h^{1/2}(n_h+n_0+1)^{1/2}S_h}{\sqrt{(\Sigma_{h=1}^k n_h)(\Sigma_{h=1}^k n_h + n_0 + 1)}} = \frac{\Sigma_{h=1}^k T_h - n_0(\Sigma_{h=1}^k n_h)/2}{\sqrt{n_0(\Sigma_{h=1}^k n_h)(\Sigma_{h=1}^k n_h + n_0 + 1)/12}},$$

which is just the standardized two-sample Mann-Whitney test statistic computed between the control sample and the k treatment samples combined (see the formulas for the expectation and the variance of T_h at the beginning of this section). This verifies our earlier statement. Indeed, such a test statistic has been proposed by Fligner and Wolfe (1982).

Case 3

Finally, consider the situation where there is no "reason" or "information" available, a-priori, to focus on some specific sub-alternative of A. This, for example, is the case where k essentially different treatments are compared with a control (the placebo). The situation is similar to what Hettmansperger and Norton (1987) called "specifying a vague pattern." It is obvious that no uniformly optimal test procedure will exist in this case because the optimal test statistic $(\boldsymbol{\xi}_{(1)}/\|\boldsymbol{\xi}_{(1)}\|, \boldsymbol{V})$ depends on the choice of the "direction" $\boldsymbol{\xi}_{(1)}/\|\boldsymbol{\xi}_{(1)}\|$ in K or, equivalently, on the "half-line" $l = \{\rho\boldsymbol{\xi}_{(1)}; \rho > 0\}$ spanned by it. If we still wish to use a linear combination $\Sigma_{h=1}^k c_h V_h$ (or $\Sigma_{h=1}^k c_h' S_h$ or $\Sigma_{h=1}^k c_h'' T_h$), then we need to discuss which choice of $\xi_{(1)}$ or half-line l is most appropriate. From a geometrical point of view, various proposals can be made. Note that, in a sense, the convex cone $K = \{\Sigma_{h=1}^k \mu_h b_h; \mu_1, \mu_2, ..., \mu_k \geq 0\}$ has to be"represented" by some half-line l. A geometrically attractive proposal is to consider the centroid line

$$l_1 = \{\rho\Sigma_{h=1}^k(b_h/\|b_h\|); \rho > 0\}.$$

The axis of the inscribed cone and the axis of the circumscribed cone (if it is inside K) are two other possibilities.

From a statistical point of view, it seems reasonable to focus on the half-line $l_0 \subset K$ which minimizes the maximum angle $\Psi(l, m)$, as a function of l, the maximum being taken over any other half-line $m \subset K$. It can be shown that the optimal half-line l_0 coincides with the axis of the circumscribed cone if $l_0 \subset K$. Abelson and Tukey (1963) motivated this approach by arguing that a minimum correlation is maximized by such a choice. Schaafsma (1966) approached the problem from a Neyman-Pearson point of view and showed that the test which is most powerful level α against l_0 has the smallest maximum shortcoming among all somewhere most powerful level α tests, the shortcoming of a test being defined in the usual way as the difference between the envelope power and the power actually achieved. Such a test is called the most stringent somewhere most powerful level α (MSSMPα) test. The main arguments behind obtaining the MSSMP test are as follows.

Note that for some fixed $\boldsymbol{\nu}$, the most powerful size α test of $\boldsymbol{\nu} = \mathbf{0}$ against a fixed simple alternative $\boldsymbol{\nu} \in K$, is

$$\phi_l = \begin{cases} 1 & \text{if } \boldsymbol{\nu}'\boldsymbol{V} \geq z_\alpha \|\boldsymbol{\nu}\| \\ 0 & \text{otherwise.} \end{cases} \tag{30.10}$$

Let $l = \{\rho\boldsymbol{\nu}; \rho > 0\}$ denote the half-line spanned by $\boldsymbol{\nu}$. Similarly, let $m = \{\rho\boldsymbol{\eta}; \rho > 0\}$ denote the half-line spanned by some other fixed $\boldsymbol{\eta} \in K$. It is easy to see [originally observed by Smid and explicitly shown in Schaafsma (1966)] that the maximum of the shortcoming of the test ϕ_l, on the half-line m, is an increasing function of the angle

$$\Psi(l, m) = \cos^{-1}\left\{\frac{(\boldsymbol{\nu}, \boldsymbol{\eta})}{\|\boldsymbol{\nu}\|\,\|\boldsymbol{\eta}\|}\right\}$$

between l and m. Hence, the maximum of the shortcoming of ϕ_l as a function of $\boldsymbol{\eta} \in K$ is attained on the half-line m which maximizes $\Psi(l, m)$. In the Schaafsma-Smid theory, the optimal half-line $l = l_0$ is chosen as that half-line $l \subset K$ such that $\max_m \Psi(l, m)$ is minimized. In principle, l_0 can be obtained by applying the equiangular-or-closer principle described by Abelson and Tukey (1963). The edges $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_k$ of K are given by $\boldsymbol{e}_h = \{\rho\boldsymbol{b}_h; \rho > 0\}$ and the equiangular half-line l_0' is obtained by equating the angles $\Psi(l_0', \boldsymbol{e}_h), h = 1, 2, \ldots, k$. If the solution of these equations satisfies $l_0' \subset K$, then $l_0' = l_0$ is the required minimax half-line.

As we noted before, a drawback of this presentation is that it is based on a canonical formulation: a matrix B has to be chosen from the collection of all B such that $B\Gamma B' = I_k$. The solutions obtained will have to be translated back to $\boldsymbol{S} = B^{-1}\boldsymbol{V}$. These somewhat cumbersome steps can be avoided by using a more direct approach for testing $H_0^* : \boldsymbol{\mu} = \mathbf{0}$ against $H_1^* : \boldsymbol{\mu} \geq \mathbf{0}$ on the basis of $\boldsymbol{S} \sim N_k(\boldsymbol{\mu}, \Gamma)$, if R^k is endowed with the inner product $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}'\Gamma^{-1}\boldsymbol{y}$. Using this representation, the equiangular half-line l_0 is now obtained. The k equations to be solved are

$$\begin{aligned}\cos\Psi(l_0, \boldsymbol{e}_h) &= \frac{(\boldsymbol{e}_h, \boldsymbol{\mu})}{\sqrt{(\boldsymbol{e}_h, \boldsymbol{e}_h)(\boldsymbol{\mu}, \boldsymbol{\mu})}} \\ &= \frac{\boldsymbol{e}_h'\Gamma^{-1}\boldsymbol{\mu}}{\sqrt{(\boldsymbol{e}_h'\Gamma^{-1}\boldsymbol{e}_h)(\boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{\mu})}} \\ &= c, \quad h = 1, 2, ..., k,\end{aligned}$$

where c is some constant to be determined later. For our case of the positive orthant, the edge $\boldsymbol{e}_h$ is given by $\{\rho(0, ..., 0, 1, 0, ..., 0)'; \rho > 0\}$, where the one is in the h-th position. The constant c has an important geometrical and statistical interpretation. It is the cosine of the minimax angle (if the solution to the above system of equations lies in the k dimensional positive orthant) and it determines the maximum shortcoming of the most stringent somewhere most powerful test. In fact, this maximum shortcoming can be shown [Schaafsma (1966, p. 38)] to be approximately equal to $(1 - \alpha)(1 - c)$. Thus, if the minimax angle is small, then c is large (close to 1) and the maximum shortcoming of the MSSMP test is small. On the other hand, if the minimax angle is large, then c is small and the maximum shortcoming of the MSSMP test is large. In the present situation, but not in Section 30.6, the minimax angle turns out to be large and hence the efficacy of any linear test is questionable; the relevance of the restriction to the class of somewhere most powerful (linear) tests is doubtful. See Section 30.7 for some alternative, from a theoretical viewpoint, less questionable, approaches.

However, as an applied statistician will be reluctant to go beyond the class of linear test statistics, it may make sense to continue with the derivation of the best linear test, even if this is known to be unsatisfactory from a theoretical point of view, as is the case in the simple-tree problem.

30.5.2 Derivation of the MSSMP test

Recall that to find the MSSMP weight vector $\boldsymbol{\nu}$, we need to solve the equations

$$\cos\Psi(l_0, \boldsymbol{e}_h) = c, \quad h = 1, 2, ..., k,$$

where $\boldsymbol{e}_h = \rho(0, ..., 0, 1, 0, ..., 0)'$. This leads to the equations

$$\boldsymbol{e}_h'\Gamma^{-1}\boldsymbol{\mu} = \sqrt{\boldsymbol{e}_h'\Gamma^{-1}\boldsymbol{e}_h}, \qquad h = 1, 2, ..., k,$$

if $\boldsymbol{\mu}$ or $(\boldsymbol{\nu})$ is normalized such that c is $(\boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{\mu})^{-1/2}$. The above system of equations can be expressed as

$$I_k\Gamma^{-1}\boldsymbol{\mu} = \boldsymbol{\delta}, \tag{30.11}$$

where

$$\boldsymbol{\delta} = (\sqrt{\Gamma^{11}}, \sqrt{\Gamma^{22}}, ..., \sqrt{\Gamma^{kk}})'.$$

The solution to (30.11) is given by

$$\boldsymbol{\mu} = \Gamma\boldsymbol{\delta}, \tag{30.12}$$

where we need to verify that $\boldsymbol{\mu} = \Gamma\boldsymbol{\delta} \geq \mathbf{0}$. For the simple-tree problem, this verification is a matter of straightforward computation (since in this case both Γ and $\boldsymbol{\delta}$ have nonnegative components) and can be made for every design $n_0, n_1, ..., n_k$. Rewriting the test given by the critical region

$$\frac{\boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{S}}{\sqrt{\boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{\mu}}} \geq z_\alpha,$$

in terms of $\boldsymbol{\delta}$ by using (30.12), we obtain the following.

Result 30.5.1 *The approximate MSSMPα test for the simple-tree problem (H_0, H_1) (based on Mann-Whitney-Wilcoxon statistics) is given by*

$$\frac{\boldsymbol{\delta}'\boldsymbol{S}}{\sqrt{\boldsymbol{\delta}'\Gamma\boldsymbol{\delta}}} \geq z_\alpha,$$

where $\boldsymbol{S} = (S_1, S_2, ..., S_k)'$, $S_h = \frac{T_h - n_0 n_h/2}{\sqrt{n_h n_0 (n_h + n_0 + 1)/12}}$, and T_h is the Mann-Whitney statistic between sample 0 and sample h. The minimax weights are given by

$$\delta_h = \sqrt{\Gamma^{hh}} = \sqrt{(1 + \frac{n_h}{n_0 + 1})(1 - \frac{n_h}{N+1})}, \tag{30.13}$$

and $\boldsymbol{\delta}'\Gamma\boldsymbol{\delta} = \boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{\mu} = c^{-2}$ is equal to

$$\frac{(n_0 + 1)\Sigma_{h=1}^{k}(N - n_h + 1) + \{\Sigma_{h=1}^{k}\sqrt{n_h(N - n_h + 1)}\}^2}{(n_0 + 1)(N + 1)}. \tag{30.14}$$

Remark 30.5.2 When $n_1 = n_2 = ... = n_k, \delta_h$ is a constant and the approximate MSSMP test is based on the sum of the S_h or equivalently on the sum of the Mann-Whitney statistics T_h.

Remark 30.5.3 The maximum shortcoming of the approximate MSSMP test is attained on each of the k edges of the positive orthant. The cosine of the minimax angle is given by

$$c = (\boldsymbol{\mu}'\Gamma^{-1}\boldsymbol{\mu})^{-1/2} = (\boldsymbol{\delta}'\Gamma\boldsymbol{\delta})^{-1/2},$$

where $\boldsymbol{\delta}'\Gamma\boldsymbol{\delta}$ is given in (30.14). The quantity c determines the maximum shortcoming of the MSSMPα test. In general, the minimax angle depends on k and the sample sizes. When $n_1 = n_2 = ... = n_k = an_0, 0 < a < 1$, from (30.14) we find

$$c^{-2} = k\left(1 - \frac{an_0}{N+1}\right)\left(1 + k\frac{an_0}{n_0+1}\right). \tag{30.15}$$

Hence for large n_0 and N, the cosine of the minimax angle, c, can be approximated by

$$\{k(1+ak-a)\}^{-1/2}. \tag{30.16}$$

Remark 30.5.4 It may be noted that $\delta_h^2 = \Gamma^{hh}$, so that the weights for the approximate MSSMPα test are determined by the diagonal elements of the inverse of the (asymptotic) null correlation matrix of $\boldsymbol{S}$. This leads to the following result.

Proposition 30.5.1 *Suppose $\boldsymbol{S}$ is (asymptotically) normally distributed with mean $\boldsymbol{\mu}$ and a correlation matrix Γ^* and that $H_0^* : \boldsymbol{\mu} = \mathbf{0}$ is to be tested against $H_1^* : \boldsymbol{\mu} \geq \mathbf{0}$ on the basis of $\boldsymbol{S}$. Let Γ be the (asymptotic) correlation matrix of $\boldsymbol{S}$ under H_0^*. The approximate MSSMPα test for testing H_0^* against H_1^* is given by*

$$\frac{\Sigma_{h=1}^k \delta_h S_h}{\sqrt{\boldsymbol{\delta}'\Gamma\boldsymbol{\delta}}} \geq z_\alpha,$$

where

$$\delta_h = \sqrt{\Gamma^{hh}} = \sqrt{\frac{\textit{co-factor of } (h,h)\textit{-th element of } \Gamma}{\textit{determinant of } \Gamma}},$$

provided $\Gamma\boldsymbol{\delta} \geq \mathbf{0}$. The cosine of the minimax angle c is given by $(\boldsymbol{\delta}'\Gamma\boldsymbol{\delta})^{-1/2}$.

Remark 30.5.5 From Proposition 30.5.1, in general, the approximate most stringent (or minimax) weights δ_h are proportional to

$$\sqrt{\text{co-factor of } (h,h)\text{-th element of } \Gamma}.$$

Thus, as far as the minimax weights are concerned, attention can be restricted to the co-factors of the diagonal elements of the null correlation matrix of the standardized statistics. This allows easier computation of the minimax weights in a number of applications, including the next one with the simple-order. Of course, the condition that $\Gamma\boldsymbol{\delta} \geq \mathbf{0}$ still needs to be verified.

30.6 Test for the Simple Order Problem

We now consider Testing Problem 2 or the problem of testing homogeneity against the simple order alternative (upward trend). In this case, the problem of testing H_{0h} against H_{2h} is a two-sample problem with a one-sided alternative between samples $h-1$ and h, with n_{h-1} and n_h observations, respectively. The test statistic for this problem will be based on the adjacent samples, $(X_{h-1,1}, ..., X_{h-1,n_{h-1}})$ and $(X_{h1}, ..., X_{hn_h})$. As in the case of the simple-tree alternative, we use the Wilcoxon rank-sum test based on the Mann-Whitney statistic

$$T_{h-1,h} = \Sigma_{r=1}^{n_h}\Sigma_{s=1}^{n_{h-1}} I(X_{hr} \geq X_{h-1,s}),$$

between the $(h-1)^{th}$ and the h^{th} sample, $h = 1, 2, ..., k$. The reader might observe that some intuition is involved in these choices. We agree that alternative choices are possible and that our choice is something "unique" or "most appropriate." The standardized statistic is

$$U_h = \frac{T_{h-1,h} - n_{h-1}n_h/2}{\sqrt{n_{h-1}n_h(n_{h-1}+n_h+1)/12}} \,. \tag{30.17}$$

Observe that for any (F_{h-1}, F_h),

$$E[U_h] = \sqrt{12}a_h\pi_h = \mu_h,$$

where

$$a_h = n_{h-1}^{1/2}n_h^{1/2}(n_{h-1}+n_h+1)^{-1/2}$$

and

$$\begin{aligned}\pi_h &= \Pr[X_{h,r} \geq X_{h-1,s}] - 1/2 \\ &= \int F_{h-1}dF_h - 1/2 = \int (1-F_h)dF_{h-1} - 1/2.\end{aligned}$$

It is known that [see, for example, Tryon and Hettmansperger (1973)]

$$\text{cov}(T_{h-1,h}, T_{h,h+1}|H_0) = -n_hn_{h-1}n_{h+1}/12, \qquad h = 1, 2, ..., k-1.$$

Therefore, $\text{cov}(U_h, U_{h+1}|H_0)$ equals

$$\gamma_{h,h+1} = \gamma_{h+1,h} = -\sqrt{\frac{n_{h-1}n_{h+1}}{(n_{h-1}+n_h+1)(n_h+n_{h+1}+1)}},$$

$h = 1, 2, ..., k-1$, and all other covariances (or correlations) among the U_h are 0, while $\text{var}(U_h) = 1$. For any fixed alternative $F_0, F_1, ..., F_k$, the asymptotic distribution of the random vector of standardized statistics $\boldsymbol{U} = (U_1, U_2, ..., U_k)'$

can be obtained from the results of Puri (1965). The asymptotic distribution is a k-variate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, ..., \mu_k)'$ and some correlation matrix. Under H_0, the correlation matrix is $\Gamma = ((\gamma_{gh}))$, where γ_{gh} is specified above.

Thus, as in the simple-tree case, the problem (H_0, H_2) of testing against the simple order can be recast in terms of the standardized statistics as testing $H_0^* : \boldsymbol{\mu} = \mathbf{0}$ against $H_1^* : \boldsymbol{\mu} \geq \mathbf{0},\ \boldsymbol{\mu}'\boldsymbol{\mu} \neq 0$. In other words, the testing problem at hand is approximately equivalent to testing against the positive orthant on the basis of dependent test statistics. Again, we may proceed as in Cases 1 and 2 to derive the approximately most powerful size α tests for the two alternatives: (i) where $\boldsymbol{\mu}$ is specified, and (ii) where the F_i's are equi-distant (in some sense); the details are omitted. We, however, concentrate on the "vague pattern," namely the situation where no a-priori information is available to focus on any sub-alternative of A. In this case, we can obtain the approximately MSSMPα test by applying Proposition 30.5.1. This is illustrated below.

30.6.1 Derivation of the (A)MSSMP test

From Proposition 30.5.1, the minimax weights (apart from the condition $\Gamma\boldsymbol{\delta} > \mathbf{0}$, which has to be verified) are given by $\delta_h = \sqrt{\Gamma^{hh}}$ and therefore the problem reduces to finding the diagonal elements of the inverse of the null correlation matrix Γ. To this end, by Remark 30.5.5, it is sufficient to find the co-factors of the diagonal elements of Γ. Now, observe that for large sample sizes, $\gamma_{h,h+1}$ is approximately equal to

$$-\sqrt{\frac{n_{h-1}n_{h+1}}{(n_{h-1} + n_h)(n_h + n_{h+1})}} = d_h, \tag{30.18}$$

say. Thus, the null correlation matrix Γ is approximately equal to

$$\gamma_{gh} = \begin{cases} 1 & \text{if} \quad g = h \\ d_g & \text{if} \quad h = g + 1, g = 1, 2, ..., k - 1 \\ 0 & \text{if} \quad |g - h| \geq 2. \end{cases} \tag{30.19}$$

The approximation to $\gamma_{h,h+1}$ is useful because it simplifies calculation of the inverse of the matrix Γ. To this end, it may be noted that Γ is a $k \times k$ band symmetric matrix of bandwidth 3 with elements $d_1, ..., d_{k-1}$ on the super and sub diagonals (above and below the main diagonal consisting of 1's), d_h being given by (30.18), and zero's elsewhere. Let Δ_h denote the determinant of a $h \times h$ band symmetric matrix with elements $d_1, ..., d_{h-1}$ on the super (and the sub) diagonal, with $\Delta_0 = \Delta_1 = 1$. The required co-factors can be obtained from the following result.

Lemma 30.6.1 *The co-factor of the (h, h)-th element of Γ is given by*

$$C_{hh} = \Delta_{h-1}(d_1, d_2, ..., d_{h-2})\Delta_{k-h}(d_{h+1}, d_{h+2}, ..., d_{k-1}).$$

Using the special form of d_h, it can be established that

$$C_{hh} = \frac{\prod_{i=1}^{k-1} n_i}{\prod_{i=0}^{k-1}(n_i + n_{i+1})}\, s_{h-1}(N - s_{h-1})\left(\frac{1}{n_h} + \frac{1}{n_{h-1}}\right), \tag{30.20}$$

$h = 1, 2, ..., k$, where $s_h = \Sigma_{g=0}^{h} n_g$ and $s_k = N$.

A proof of this Lemma is given in the Appendix.

Proposition 30.5.1 can be applied provided we can verify that $\Gamma\boldsymbol{\delta} \geq \mathbf{0}$. After some algebraic manipulations, it can be seen that this condition is equivalent to the condition that the MSSMP weights for the simple order problem are nondecreasing, as observed in Schaafsma (1966). A proof of the latter (for every design $n_0, ..., n_k$) can be found in the same reference on page 74. However, as the algebraic manipulations are somewhat involved and the cited reference may not be easily available, we give an indication of the proof of this fact in the second part of the appendix.

Result 30.6.1 *The approximate MSSMPα test for the simple order problem (H_0, H_2) (based on Mann-Whitney-Wilcoxon statistics) is*

$$\frac{\Sigma_{h=1}^{k}\delta_h U_h}{\sqrt{\boldsymbol{\delta}'\Gamma\boldsymbol{\delta}}} \geq z_\alpha,$$

where U_h is given by (30.17) and the most stringent (minimax or maximin) weights are proportional to

$$\sqrt{s_{h-1}(N - s_{h-1})}\sqrt{\frac{1}{n_h} + \frac{1}{n_{h-1}}}, \qquad h = 1, 2, ..., k. \tag{30.21}$$

Remark 30.6.1 When $n_0 = n_1 = ... = n_k$, we have $d_h = -1/2$, and the minimax weights are proportional to

$$\sqrt{h(k + 1 - h)}, \quad h = 1, 2, ..., k, \tag{30.22}$$

so that the approximate MSSMP test for the simple order problem, with equal sample sizes, is based on

$$\mathrm{S} = \Sigma_{h=1}^{k}\sqrt{h(k + 1 - h)}T_{h-1,h}, \tag{30.23}$$

where we recall that $T_{h-1,h}$ is the Mann-Whitney statistic between the adjacent samples $h - 1$ and h, $h = 1, 2, ..., k$.

Remark 30.6.2 When $n_0 = n_1 = ... = n_k$, the cosine of the minimax angle, c, can be shown to be equal to

$$\frac{(k+1)^{1/2}}{[\sum_{i=1}^{k+1}\{\sqrt{(i-1)(k+2-i)} - \sqrt{i(k+1-i)}\}^2]^{1/2}}\,. \tag{30.24}$$

Remark 30.6.3 For the simple order problem with equal sample sizes, an alternative test may be based on

$$T = \sum_{h=1}^{k} \sqrt{h(k+1-h)}(\bar{R}_h - \bar{R}_{h-1}), \tag{30.25}$$

where $\bar{R}_h$ is the average of the ranks assigned to sample h in the combined ranking of the $k+1$ samples. It can be shown that, for fixed k and large sample sizes, $ARE(S,T) = 1$, so that the tests based on S and T are equally efficient.

For the simple order problem, linear combinations of pairwise Mann-Whitney (more generally, Chernoff-Savage) statistics have been considered by Tryon and Hettmansperger (1973). They show that, when $F_i(x) = F(x - \theta_i)$ and sample sizes are equal, corresponding to every linear combination of pairwise Mann-Whitney statistics $\Sigma_{j=0}^{k-1}\Sigma_{h=j+1}^{k} g_{jh}T_{jh}$, there is an "equivalent" statistic based on adjacent Mann-Whitney statistics $\Sigma_{h=1}^{k} a_h T_{h-1,h}$, where

$$a_h = \Sigma_{i=0}^{h-1}\Sigma_{j=h}^{k} g_{ij}, \quad h = 1, 2, ..., k.$$

Further, they show that when the spacings $\theta_i - \theta_{i-1}$ are equal for $i = 1, 2, ...k$, the maximum Pitman efficacy is achieved when $g_{jh} = 1$ for all j and h (equivalently when $a_h = h(k-h+1)$). Thus, the popular nonparametric test for the simple order (upward trend) problem proposed by Terpstra (1952) [also by Jonckheere (1954) and more generally by Puri (1965)] which is based on

$$\mathrm{J} = \Sigma_{j=0}^{k-1}\Sigma_{h=j+1}^{k} T_{jh},$$

has the maximum Pitman efficacy among the class of linear combinations of Mann-Whitney statistics, when the spacings and the sample sizes are equal. It follows that when sample sizes are equal, an asymptotically equivalent form of the J statistic is

$$\mathrm{H} = \Sigma_{h=1}^{k} h(k+1-h)T_{h-1,h}, \tag{30.26}$$

which is a linear combination of the adjacent Mann-Whitney statistics. It is clear that the test based on H is an asymptotically somewhere most powerful nonparametric test and is different from the MSSMP test based on S. Thus, a power comparison between the two tests is of interest.

30.6.2 Power comparisons

For large sample sizes, a comparison between S and H can be made in terms of the Pitman asymptotic relative efficiency (ARE). Using asymptotic means and null variances and proceeding, for example, as in Chakraborti and Desu (1991), it can be shown that for local alternatives

$$\theta_h = \theta_{h-1} + \Omega_h N^{-1/2}, \ \Omega_h \geq 0, h = 1, 2, ..., k,$$

with at least one of the Ω's being positive, the ARE is given by

$$\text{ARE(S,H)} = \frac{(\Sigma_{h=1}^{k} a_h \lambda_{h-1} \lambda_h \Omega_h)^2}{(\Sigma_{h=1}^{k} b_h \lambda_{h-1} \lambda_h \Omega_h)^2} \frac{\text{Var}(N^{-3/2}\text{H}|H_0)}{\text{Var}(N^{-3/2}\text{S}|H_0)}, \tag{30.27}$$

where $a_h = \sqrt{h(k+1-h)}, b_h = a_h^2$, and $\lambda_h = \lim_{N\to\infty}(n_h/N)$. We evaluate the ARE when the sample sizes are equal for two configurations of the Ω's: (i) where all the Ω_h's are zero except the first one, and (ii) where the Ω_h's are all equal (the equal spacings case). The ARE values are given in Table 30.1.

Table 30.1: ARE(S,H) at $\alpha = .05$

k	ARE1	ARE2
3	1.08514	.97195
4	1.18990	.94231
5	1.30104	.91535
6	1.41432	.89158
7	1.52808	.87069
8	1.64157	.85225
9	1.75443	.83586
10	1.86654	.82119

It is seen that in the first case (see ARE1) the MSSMP test is considerably more efficient (especially for $k \geq 5$) than the Jonckheere-Terpstra-Tryon-Hettmansperger test. It should be noted, however, that these alternatives with a jump between populations 0 and 1 and no jump elsewhere are somewhat extreme. In the second case (see ARE 2), the MSSMP test is at an obvious loss of power and the statistic H is more efficient, especially for large k. In general, it is difficult to choose between such somewhere most powerful tests on the basis of ARE alone and some further research on the small-sample power comparison, may provide more insight. Nevertheless, the idea of minimizing the maximum shortcoming is found to have some useful appeal.

It may be noted that Johnson and Mehrotra (1971) also considered some nonparametric tests for the problem of simple order. They proposed a class of linear tests (hereafter referred to as M tests) based on the Abelson-Tukey-Schaafsma-Smid weights and the sum of scores associated with the observations from the i-th sample. For equal sample sizes, Johnson and Mehrotra compared (see their Table 3.2) the asymptotic power of their normal scores test with the asymptotic powers of Bartholomew's LR ($\bar{\chi}^2$) test, Puri's test and the global (χ^2) test for $k = 3, 4, 8$ and 12. The conclusions may be summarized as follows. For small to moderate k, say for $k \leq 5$, the M test and Puri's test are more powerful than the LR test when the parameters are equally spaced, with Puri's test being slightly more powerful than the M test. The situation is reversed when there is only a single change in the parameters, in which case the LR test is more powerful with the JN test running a close second, especially when $k \geq 5$. As expected, all of the three tests are more powerful than the global test in all cases.

30.7 Extending the Class of SMP Tests

As noted earlier, the class of SMP (linear) tests has the advantage that probability distributions, power properties, etc., are easily obtained. However, the property of linear tests being somewhere most powerful is somewhat peculiar. On the one hand, it is nice to know that a test is optimal at least somewhere (for the MSSMP test along l_0) in the parameter space, but on the other hand it is obvious that reducing the shortcoming to 0 along the half-line $l_0 \subset K$ will have the effect that considerable shortcoming may be expected elsewhere in the parameter space, in particular along the edges. Put differently, the MSSMP test is uniformly most powerful against the alternative $\boldsymbol{\xi} \in l_0$ and hence it is a useful test provided that l_0 is a "good" representation of the entire convex cone K. It follows that the MSSMP test will be questionable if the convex cone K is very wide or, equivalently, if the minimax angle is very large.

For the simple-tree problem, the cone K is indeed very wide so that the MSSMP test has large shortcomings on or near the edges. This can be expressed more succinctly by computing c, say, in the case where $n_0 = n_1 = \ldots = n_k$. For equal sample sizes we have, from (30.16), $c = k^{-1}$. If $k = 3$, for example, then $c = 0.33$ and the minimax angle $\Psi_0 = \cos^{-1}(c)$ equals 70.73 degrees. Hence, the maximum shortcoming of the MSSMP.05 test is approximately 0.63. This is unsatisfactorily large; it is even larger than the maximum shortcoming of the (global) χ^2 test which is 0.28 [see Akkerboom (1990, Table 4.3.2, p. 179)].

For the simple order problem, however, the situation is quite different. To illustrate, again assume $k = 3$ and equal sample sizes. In this 4-sample case, from (30.24), we find $c = 0.81$, so that the minimax angle Ψ_0 equals 36.21 degrees. Hence the maximum shortcoming of the MSSMP.05 test is, approximately, 0.18 and thus the MSSMP test seems acceptable for the simple order problem though, of course, the maximum shortcoming is still considerable.

In view of the above findings, some statisticians have concluded that in situations like the ones considered, linear tests are not satisfactory (unless there are good reasons to focus on the half-lines where they are most powerful, like in the case of the H test). Giving up linearity, however, makes things extremely complicated. In Van Zwet and Oosterhoff (1967) and Schaafsma (1968), the most stringent (MS) level α tests were constructed without making any other restrictions than the level α one. Bartholomew (1959, 1961a,b) and other authors have explored the likelihood-ratio (LR) principle. For the simple order problem, Chacko (1963) proposed a rank analogue of Bartholomew's LR test which can be used when the sample sizes are equal. Shorack (1967) extended Chacko's procedure to the case of unequal sample sizes. However, although it is relatively easy, in some cases, to compute the outcome of the LR test statistic, it is often far less easy to study its null distribution. It may be noted that for equal samples sizes, the critical values for the LR test can be found in Robertson, Wright and Dykstra (1988) for the simple order and the simple-tree order, when $3 \leq k \leq 24$ and $2 \leq k \leq 23$, respectively. For unequal sample sizes Robertson, Wright and Dykstra (1988) tabulated the critical values for the simple order when k equals 3 and 4. Critical values for the simple-tree order can be obtained from these tables. In general, for unequal sample sizes, one may use the FORTRAN programs by Bohrer and Chow (1978) and Sun (1988) to compute the P-values for the LR tests. These, however, underscore some of the difficulties a practitioner might face when implementing the LR tests. As a result, many ad hoc proposals and approximations have been considered in the literature. The reader is again referred to the book by Robertson, Wright and Dykstra (1988) for a review of these developments.

Akkerboom (1990) contains an elaboration of what might be called a "middle of the road" approach, incorporating some features of the extensiveness of the cone K without sacrificing too much of the simplicity of the (linear) MSSMP tests. Instead of focusing in on the (pointed) cone K (and the half-line l_0), Akkerboom considered a circular cone

$$C = C(l_0, \omega) = \{\boldsymbol{\xi} \in R^k : \boldsymbol{b}'\boldsymbol{\xi} \geq \|\boldsymbol{\xi}\| \cos(\omega)\},$$

where $\boldsymbol{b}$ is the unit vector along $l_0 = \{\lambda \boldsymbol{b} : \lambda \geq 0\}$, the axis of C, and 2ω is the opening angle of $C(\boldsymbol{b} \in R^k; 0 \leq \omega \leq \pi/2)$. The semi-angle ω can be chosen such that the likelihood ratio test for $H : \boldsymbol{\xi} = \mathbf{0}$ versus $A : \boldsymbol{\xi} \in C$, which had been studied earlier by Pincus (1987), has minimum maximum shortcoming among all tests of this type, on the alternative K we are really interested in.

We will not discuss the details of the Akkerboom approach because these are well documented. We simply conclude this section by remarking that the Akkerboom-Pincus approach tends to provide a considerable reduction of the maximum shortcoming. For instance, for the simple-tree problem when $k = 3$ and $n_0 = n_1 = n_2 = n_3 = 1$ as before, $c = 0.33, \Psi_0 = 70.73$ degrees and Akkerboom's procedure suggests taking ω equal to 70 degrees, whereby the maximum shortcoming is reduced to about 0.27 [see Akkerboom (1990), Table 4.3.2, p. 179)]. The resulting test is what he called an approximately most stringent circular likelihood ratio (MSCLR) test. Calculations such as the one above suggest that some improvements can be achieved by going beyond the linear tests. In practice, however, researchers are likely to be attracted by the simplicity of the linear tests because this allows easy computation of P-values, and by the related idea that the true parameter value $\boldsymbol{\xi}$, though not exactly on some half-line l, will not be "too far from l." In the present paper, our attention therefore has been focused on linear tests, the most stringent somewhere most powerful one in particular.

Acknowledgements. Work on this research started while the first author was on sabbatical leave at the University of Groningen; support received during the visit is gratefully acknowledged. Partial support for the work was provided by a summer research award from the College of Commerce and Business Administration, University of Alabama.

Appendix

Part 1

First, we prove Lemma 30.6.1 about the co-factor of the (h, h)-th element of Γ for the upward trend problem. Note that for particular values of $k = 2, 3, 4$, etc., the formula for C_{hh} can be verified by direct algebraic manipulations. For a general proof, observe that the first expression for C_{hh} can be obtained by inspection, however, the second expression [Eq. (30.20)] needs verification. It is clear that the proof will be complete if we can show the following:

$$\Delta_{h-1}(d_1, d_2, ..., d_{h-2}) = \frac{s_{h-1}}{n_{h-1}} \frac{\prod_{g=1}^{h-1} n_g}{\prod_{g=0}^{h-2}(n_g + n_{g+1})} \tag{30.28}$$

and

$$\Delta_{k-h}(d_{h+1}, d_{h+2}, ..., d_{k-1}) = \frac{N - s_{h-1}}{n_h} \frac{\prod_{g=h}^{k-1} n_g}{\prod_{g=h}^{k-1}(n_g + n_{g+1})} . \tag{30.29}$$

We prove Eq. (30.28), Eq. (30.29) can be proved in a similar manner. It is clear that proving (30.28) is equivalent to proving

$$\Delta_{h-1}(d_1, d_2, ..., d_{h-2}) = \frac{\Sigma_{g=0}^{h-1} n_g \prod_{g=1}^{h-2} n_g}{\prod_{g=0}^{h-2}(n_g + n_{g+1})} . \tag{30.30}$$

We apply complete induction to prove (30.30). First, we can directly show that the formula holds for $k = 2, 3, 4$ and $h = 1, 2, ..., k$. Now assume that for any given k, the formula holds for some h. We then need to show that it also holds for $h + 1$. Thus, we need to show that

$$\Delta_{h+1}(d_1, d_2, ..., d_h) = \frac{\Sigma_{g=0}^{h+1} n_g \prod_{g=1}^{h} n_g}{\prod_{g=0}^{h}(n_g + n_{g+1})} . \tag{30.31}$$

Because of the structure of Γ, it can be seen that

$$\Delta_{h+1}(d_1, d_2, ..., d_h) = \Delta_h(d_1, d_2, ..., d_{h-1}) - d_h^2 \Delta_{h-1}(d_1, d_2, ..., d_{h-2}). \quad (30.32)$$

Substituting for Δ_{h-1} and Δ_{h-2} and noting that

$$d_h = -\sqrt{\frac{n_{h-1}n_{h+1}}{(n_{h-1}+n_h)(n_h+n_{h+1})}},$$

we see that (30.30) holds iff

$$\frac{\Sigma_{g=0}^{h+1} n_g \prod_{g=1}^{h} n_g}{\prod_{g=0}^{h}(n_g+n_{g+1})} - \frac{\Sigma_{g=0}^{h} n_g \prod_{g=1}^{h-1} n_g}{\prod_{g=0}^{h-1}(n_g+n_{g+1})} + \frac{n_{h-1}n_{h+1}}{(n_{h-1}+n_h)(n_h+n_{h+1})} \frac{\Sigma_{g=0}^{h-1} n_g \prod_{g=1}^{h-2} n_g}{\prod_{g=0}^{h-2}(n_g+n_{g+1})} = 0,$$

which can be shown to be true after some routine algebraic manipulations.

Part 2

Here, we prove that $\Gamma\boldsymbol{\delta} \geq \mathbf{0}$ for the simple order problem. To verify this, or equivalently that, $(\Gamma\boldsymbol{\delta})_h \geq 0$, $h = 1, 2, ..., k$, we focus on $h = 2, 3, ..., k-1$. The values $h = 1$ and $h = k$ require a separate treatment which is not too difficult. The crux of the matter is that

$$\begin{aligned}(\Gamma\boldsymbol{\delta})_h &= \gamma_{h,h-1}\delta_{h-1} + \delta_h + \gamma_{h,h+1}\delta_{h+1} \\ &= d_{h-1}\delta_{h-1} + \delta_h + d_h\delta_{h+1},\end{aligned}$$

where d_h is given by (30.18) and δ_h is given by (30.21). To establish the positivity of this expression, δ_h has to be split up into two parts. Substitution of (30.18) and (30.21) and elementary computations provide that the expression is equal to

$$\sqrt{\frac{n_{h-1}n_h}{n_{h-1}+n_h}}\left[\frac{\sqrt{s_{h-1}(N-s_{h-1})} - \sqrt{s_{h-2}(N-s_{h-2})}}{n_{h-1}} - \frac{\sqrt{s_h(N-s_h)} - \sqrt{s_{h-1}(N-s_{h-1})}}{n_h}\right]$$

which is positive since the function $x \to \sqrt{x(N-x)}$ is concave and hence "difference quotients are decreasing". (Note that a difference of difference quotients is involved.) This is the proof originally conceived by Smid and briefly indicated in Schaafsma (1966, p. 74).

References

1. Abelson, R. P. and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order, *Annals of Mathematical Statistics*, **34**, 1347–1369.

2. Akkerboom, C. J. (1990). *Testing Problems with Linear or Angular Inequality Constraints,* New York: Springer-Verlag.

3. Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions,* New York: John Wiley & Sons.

4. Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives, *Biometrika*, **46**, 36–48.

5. Bartholomew, D. J. (1961a). A test of homogeneity of means under restricted alternatives, *Journal of the Royal Statistical Society, Series B,* **23**, 239–281.

6. Bartholomew, D. J. (1961b). Ordered tests in the analysis of variance, *Biometrika*, **48**, 325–332.

7. Bohrer, R. and Chow, W. (1978). Weights for one-sided multivariate inference, *Applied Statistics*, **27**, 100–103.

8. Chacko, V. J. (1963). Testing homogeneity against ordered alternatives, *Annals of Mathematical Statistics*, **34**, 945–956.

9. Chakraborti, S. and Desu, M. M. (1991). Linear rank tests for comparing treatments with a control when data are subject to unequal patterns of censorship, *Statistica Neerlandica,* **45**, 227–254.

10. Fligner, M. A. and Wolfe, D. A. (1982). Distribution-free tests for comparing several treatments with a control, *Statistica Neerlandica,* **36**, 119–127.

11. Hettmansperger, T. P. and Norton, R. M. (1987). Tests for patterned alternatives in k-sample problems, *Journal of the American Statistical Association,* **82**, 292–299.

12. Johnson, R. A. and Mehrotra, K. G. (1971). Some c-sample nonparametric tests for ordered alternative, *Journal of the Indian Statistical Association,* **9,** 8–23.

13. Jonckheere, A. R. (1954). A distribution-free k-sample test against ordered alternatives, *Biometrika,* **41,** 133–145.

14. Oosterhoff, J. (1969). *Combination of One-sided Statistical Tests,* Mathematical Centre Tracts 28, Mathematisch Centrum, Amsterdam.

15. Pincus, R. (1987). Testing linear hypotheses under restricted alternatives, *Mathematische Operationsforschung und Statistik, Series Statistics*, **6**, 733–751.

16. Puri, M. L. (1965). Some distribution free K-sample rank tests of homogeneity against ordered alternatives, *Communications in Pure and Applied Mathematics,* **18,** 51–63.

17. Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference,* New York: John Wiley & Sons.

18. Schaafsma, W. (1966). *Hypothesis Testing Problems with the Alternative Restricted by a Number of Linear Inequalities,* Groningen, The Netherlands: Noordhoff.

19. Schaafsma, W. (1968). A comparison of the most stringent and the most stringent somewhere most powerful test for certain problems with restricted alternatives, *Annals of Mathematical Statistics,* **39,** 531–546.

20. Schaafsma, W. (1971). The Neyman-Pearson theory for testing statistical hypotheses, *Statistica Neerlandica,* **25,** 1–27.

21. Schaafsma, W. and Smid, L. J. (1966). Most stringent somewhere most powerful tests against alternatives restricted by a number of linear inequalities, *Annals of Mathematical Statistics,* **37,** 1161–1172.

22. Shorack, G. R. (1967). Testing against ordered alternatives in model I analysis of variance; normal theory and nonparametric, *Annals of Mathematical Statistics*, **38**, 1740–1753.

23. Snijders, T. (1979). *Asymptotic Optimality Theory for Testing Problems with Restricted Alternatives,* Mathematical Centre Tracts 113, Mathematisch Centrum, Amsterdam.

24. Sun, H-J. (1988). A FORTRAN subroutine for computing normal orthant probabilities of dimensions up to nine, *Communications in Statistics–Simulation and Computation*, **17**, 1097–1111.

25. Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend when ties are present in one ranking, *Indagationes Mathematicae,* **14,** 324–333.

26. Tryon, P. and Hettmansperger, T. P. (1973). A class of non-parametric tests for homogeneity against ordered alternatives, *Annals of Statistics,* **1,** 1061–1070.

27. Van Zwet, W. R. and Oosterhoff, J. (1967). On the combination of independent test statistics, *Annals of Mathematical Statistics*, **38,** 659–680.

31

Nonparametric Estimation of the Ratio of Variance Components

M. Mahibbur Rahman and Z. Govindarajulu

University of Swaziland, Swaziland
University of Kentucky, Lexington, KY

Abstract: Consider the one-way random effects model $Z_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$, where α_i and ϵ_{ij} are mutually independent and distributed with mean zero and variance σ_α^2 and σ_ϵ^2, respectively. An estimation procedure is developed for estimating $\rho^2 = \sigma_\alpha^2/\sigma_\epsilon^2$, the ratio of variance components, using a method similar to that proposed by Shorack (1969) for estimating the ratio of two scale parameters. An extensive Monte Carlo study is carried out in order to investigate the performance of the proposed method when compared with the other two methods provided by Groggel, Wackerly and Rao (1987) and Bhattacharyya (1977). The study suggests that the proposed method is superior to those two alternative methods for certain ranges of the model parameters. Finally, we propose a small-sample adjustment which improves the performance of the estimation method.

Keywords and phrases: Random effects model, nonparametric method, Hodges-Lehmann type of estimator, variance component ratio, Moses rank like test

31.1 Introduction

The statistical technique known as *analysis of variance* was initially developed by R. A. Fisher for facilitating the analysis and interpretation of data obtained from field trials and laboratory experiments in agricultural and biological research. Now this research tool is widely used in almost every discipline. The principal interest of analysis of variance lies in estimating (or testing hypothesis about) linear functions of the effects in the assumed model. If the assumed model is a random effects model, the prime interest concerning the effects becomes estimating or testing hypothesis about the variances of the random effects, known as *variance components.*

For example, a laboratory experiment is designed to study the maternal ability of mice. Suppose the experiment uses litter weights of ten-day old litters as the measure of maternal ability as in Young, Legates and Farthing (1965) and that six litters from each of four dams, all of one breed, constitute the data. A suitable model for analyzing the data is the one-way balanced random effects model which can be represented as

$$Z_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad i = 1, 2, \ldots, m, \; j = 1, 2, \ldots, n, \tag{31.1}$$

where Z_{ij} is the weight of the i-th litter from the j-th dam, μ an unknown constant, α_i the effect due to i-th litter, and the error terms, ϵ_{ij}, are mutually independent random variables having the same distribution. The general null hypothesis we usually want to test concerning variance components is

$$H_0: \; \mathrm{Var}(\alpha_i) \leq C \; \mathrm{Var}(\epsilon_{ij}) \qquad \text{for all } i, j$$

where $C > 0$ is specified. Furthermore, the standard analysis of variance techniques require the assumption that both α_i and ϵ_{ij} are normally distributed with zero means. The rejection or acceptance of the above null hypothesis concludes only that $\rho^2 = \sigma_\alpha^2/\sigma_\epsilon^2$ either greater or smaller than the specified value of C, where σ_α^2 and σ_ϵ^2 are the variances of α_i and ϵ_{ij}, respectively. On the other hand, an estimate of ρ^2 provides the actual relative magnitudes of the variance components which is more informative than saying it is greater or smaller than a certain value.

There are certainly many situations in which normality assumptions regarding random effects are fully justified, and in such cases the analysis of variance is the preferred technique. However, certainly there are situations in which the normality assumption is not at all justified and in those cases, use of the analysis of variance technique can provide misleading results. Hence, it is important to derive distribution-free methods for estimating (or for testing hypothesis about) ρ^2 and in this paper we provide one such method of estimation. The proposed estimation procedure does not require normality assumption but does require the assumption of independence between the two random effects. Also, this procedure ensures the non-negativity of the estimates which is another advantage over the classical analysis of variance estimates.

Hodges and Lehmann (1963) considered an estimation procedure for the two-sample location parameters in the case where $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are independent random samples from populations having continuous distribution functions $F_X(t) = \Psi((t - \nu)/\sigma_1)$ and $F_Y(t) = \Psi((t - \eta)/\sigma_2)$, respectively, with ν and η as the respective location parameters (medians) and σ_1 and σ_2 as the respective scale parameters. They derived a test statistic $h(X, Y)$ which has a symmetric distribution under the null hypothesis. If $h(X, Y)$ is the number of pairs (i, j) such that $X_i < Y_j$ $(1 \leq i \leq m;\ 1 \leq j \leq n)$, then the test based on $h(X, Y)$ becomes Wilcoxon's rank-sum test for the two-sample location problem. The estimate of the difference between two location parameters is then the median of the set of mn differences $Y_j - X_i$. Several researchers tried to improvise this estimating procedure of the difference between two location parameters for the two-sample scale problem.

One of those attempts was initiated by Bhattacharyya (1977), who introduced a version of the Hodges-Lehmann estimator for estimating the ratio of the two scale parameters $(\rho = \sigma_2/\sigma_1)$ by applying it to the Ansari-Bradley test statistic. The resulting estimator of ρ is the median of all the ratios (x/y) where (x, y) are all 'relevant pairs' for the observed outcome. A relevant pair is a pair (x, y) where x and y are both positive. Ansari and Bradley (1960) introduced a distribution-free rank test for the two-sample dispersion problem where it is assumed that the location parameters ν and η or $\nu - \eta$ are known. In order to make the combined sample symmetric about zero, we need to know the value of the location parameters.

Moses (1963) introduced a rank-like distribution-free test of the ratio of two scale parameters, which does not require any assumption regarding the location parameters. Shorack (1969) derived an estimator based on Moses' rank-like test by adopting techniques provided by Hodges and Lehmann (1963). This estimation procedure divides both samples into subgroups of equal size. The estimator then becomes the median of all possible ratios of the group sample variances.

A recent attempt to improvise the Hodges-Lehmann estimator to the scale problem is due to Groggel, Wackerly and Rao (1987). The authors used the model (31.1) where α_i and ϵ_{ij} are independent random variables with continuous distributions which are symmetric about zero and have variances σ_α^2 and σ_ϵ^2, respectively, and they also assumed that α_i/σ_α and $\epsilon_{ij}/\sigma_\epsilon$ have the same distribution. This implies that α_i and $\rho\epsilon_{ij}$ (or $\ln|\alpha_i|$ and $\ln|\rho\epsilon_{ij}|$) have the same distributions. Hence, $\ln|\alpha_i|$ and $\ln|\epsilon_{ij}|$ have distributions which differ only in the location parameter, namely, $\ln \rho$. Thus, the authors reduced the two-sample scale problem to a two-sample location problem where $\ln \hat{\rho}^2$ is the median of the sets of $\ln|\rho\epsilon_{ij}| - |\ln \alpha_i|$.

Section 31.2 explains the computational procedure of the proposed method. In Section 31.3, we present a summary of a Monte Carlo study which compares the estimates obtained from the proposed method with those of Groggel, Wackerly and Rao (1987) and Bhattacharyya (1977). The last section provides an adjustment for the proposed estimation procedure so as to reduce the bias of the estimates.

31.2 Proposed Estimation Procedure

Both Bhattacharyya (1977) and Shorack (1969) introduced a Hodges-Lehmann type of estimator based on two independent random samples from two populations. In addition to that, Bhattacharyya's method requires information about location parameters, but Shorack's method does not require this information. The objective of this study is to develop an estimation procedure for $\rho^2 = \sigma_\alpha^2/\alpha_\epsilon^2$, the ratio of the variance components in one-way balanced random effect model in (31.1). Let us assume that the treatment effects (α_i) and the error terms (ϵ_{ij}) are independently distributed with continuous distributions $F_\alpha(t) = \Psi((t-\nu)/\sigma_\alpha)$ and $F_\epsilon(t) = \Psi((t-\nu)/\sigma_\epsilon)$. Further, let us assume that the location parameters which are means are equal to zero. Note that, under the above assumptions, the distributions differ only in scale parameters.

The values of α_i and ϵ_{ij} in the model (31.1) are not observable since we observe only the values of Z_{ij} which come from a single population. So, we cannot apply Shorack's method directly to this model since it requires that the samples must come from two populations which are symmetric. In order to apply Shorack's method for estimating ρ^2, we assume that α_i and ϵ_{ij} are independent random variables with distribution Ψ that is symmetric about zero. Also, we need to have two independent samples, one consisting of α_i's and another consisting of ϵ_{ij}'s, but the individual α_i and ϵ_{ij} are not observable. To construct two independent samples of observations α_i and ϵ_{ij}, we introduce the "pseudosamples" as defined by Groggel, Wackerly and Rao (1987) which are transformed Z_{ij}'s, such that for large m and n, the transformed Z_{ij}'s behave essentially like independent samples of α_i of size m and that of ϵ_{ij} of size $n = mn$. Although there are different ways in which these pseudosamples can be constructed, we will use the method called as 'means method' as described by Groggel, Wackerly and Rao (1987).

Define

$$\begin{aligned} \bar{Z}_{i.} &= n^{-1}\sum_{j=1}^{n} Z_{ij}, & \bar{Z}_{..} &= (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}, \\ \bar{\epsilon}_{i.} &= n^{-1}\sum_{j=1}^{n} \epsilon_{ij}, & \bar{\epsilon}_{..} &= (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n} \epsilon_{ij}, \end{aligned}$$

and

$$\bar{\alpha} = m^{-1}\sum_{i=1}^{m} \alpha_i.$$

Then, the pseudosamples based on means are formed as

$$X_{ij} = Z_{ij} - \bar{Z}_{i.} = \epsilon_{ij} - \bar{\epsilon}_{i.} \qquad (31.2)$$

and

$$Y_i = \bar{Z}_{i.} - \bar{Z}_{..} = \alpha_i + \bar{\epsilon}_{i.} - \bar{\alpha} - \bar{\epsilon}_{..}\,, \quad i = 1, 2, \ldots, m, \; j = 1, 2, \ldots, n. \quad (31.3)$$

Since the first moment of the underlying distribution is assumed to be zero, it is clear that, for large m and n, $\bar{\epsilon}_{i.}$, $\bar{\epsilon}_{..}$ and $\bar{\alpha}$ all converge in probability to zero. Hence, for large m and n, ϵ_{ij} $(= X_{ij})$ and α_i $(= Y_i)$ become two independent random samples of size $N = mn$ and m, respectively, generated by the transformed Z_{ij}'s. Now, we are ready to apply Shorack's procedure for estimating the ratio of the variance components.

The proposed estimation procedure, for the ratio of the variance components, consists of the following steps:

Step 1. We select a positive integer $2 \leq k \leq m$ and randomly divide the pseudosamples X_{ij} and Y_i into N' and m' subgroups of size k, respectively. We will discard any extra observations. Note that $N' = [N/k]$ and $m' = [m/k]$, where $[a]$ indicates the largest integer contained in a.

Step 2. Let $X_{p1}, X_{p2}, \ldots, X_{pk}$ denote the p-th subgroup of X-pseudosamples for $p = 1, 2, \ldots, N'$ and let $Y_{q1}, Y_{q2}, \ldots, Y_{qk}$ denote the q-th subgroup of Y-pseudosamples for $q = 1, 2, \ldots, m'$.

Step 3. Define $C_1, C_2, \ldots, C_{N'}$ by

$$C_p = \sum_{s=1}^{k} (X_{ps} - \bar{X}_p)^2, \quad p = 1, 2, \ldots, N', \text{ where } \bar{X}_p = \sum_{s=1}^{k} (X_{ps}/k).$$

Step 4. Define $D_1, D_2, \ldots, D_{m'}$ by

$$D_q = \sum_{t=1}^{k} (Y_{qt} - \bar{Y}_q)^2, \quad q = 1, 2, \ldots, m', \text{ where } \bar{Y}_q = \sum_{t=1}^{k} (Y_{qt}/k).$$

Step 5. Form the $m'N'$ ratios D_q/C_p for $q = 1, 2, \ldots, m'$ and $p = 1, 2, \ldots, N'$.

Step 6. Let $U^{(1)} \leq U^{(2)} \leq \cdots \leq U^{(m'N')}$ denote the ordered values of D_q/C_p and let $\hat{\rho}^2$ be the Hodges-Lehmann type estimator of $\rho^2 = \sigma_\alpha^2/\sigma_\epsilon^2$. Then, if $m'N'$ is odd, say $m'N' = 2r + 1$, we have $\hat{\rho}^2 = U^{(r+1)}$ and if $m'N'$ is even, say $m'N' = 2r$, then $\hat{\rho}^2 = [U^{(r)}U^{(r+1)}]^{1/2}$.

31.3 Monte Carlo Comparison

An extensive Monte Carlo study was carried out in order to compare the various methods of point estimates of $\rho^2 = \sigma_\alpha^2/\sigma_\epsilon^2$. Using SAS/IML (Statistical Analysis System/Interactive Matrix Language) procedure, random numbers were generated from four distributions that are symmetric about zero. The four distributions used were normal, uniform, logistic, and double-exponential. In each case, the resulting random numbers were used to form responses in the balanced one-way random effect model (31.1). Without loss of generality, we assumed $\mu = 0$. Hence, the model (31.1) becomes

$$Z_{ij} = \alpha_i + \epsilon_{ij}, \qquad i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, n.$$

The mn responses in each model were formed by generating $m + mn$ random numbers from one of the selected distributions. Throughout this simulation study, we assumed that $\sigma_\epsilon^2 = 1$ for simplicity so that $\rho^2 = \sigma_\alpha^2$. Multiplying the first m of these random numbers by a constant (which is, in fact, $\rho^2 = \sigma_\alpha^2$), we obtained the simulated values of α_i. The multiplier for the remaining mn numbers is one (because we fixed $\sigma_\epsilon^2 = 1$). These mn numbers provided the simulated values of ϵ_{ij}. The simulated responses were then obtained by adding the α_i and ϵ_{ij} values. The multipliers used were 0.10, 0.25, 0.50 and 1.00 in order to obtain the effects for different values of ρ^2.

For each of the four distributions, various size models were generated. The size of the model was determined by the combination of m (the number of treatments) and n (the number of observations per treatment). Three different values used for both m and n were 6, 12 and 18. The different subgroup sizes, denoted by k, that were used are 2, 3, 4, 6, 12 and 18, depending on the treatment size m.

For every combination of distribution, model size (m, n) and ρ^2-value, 3000 sets of responses were generated and the point estimates of ρ^2 were computed using the proposed estimation method with different k (subgroup size), and the methods described in Groggel, Wackerly and Rao (1987) and Bhattacharyya (1977). Note that Bhattacharyya estimated the ratio of two scale parameters. To apply this procedure for estimating the ratio of the two variances, we assumed the location parameters to be means instead of medians and σ_ϵ^2 and σ_α^2 to be variances of the underlying distributions.

A summary of the Monte Carlo study is presented in Tables 31.1 - 31.4. The tables present the mean, standard deviation and mean square error (m.s.e.) for the estimate of ρ^2 for each selected distribution, for different model size and ρ^2. The average estimates obtained from the methods of Groggel, Wackerly and Rao (1987) and Bhattacharyya (1977) were found to be consistently very close to each other with similar standard deviation for all of the four selected distributions. Hence, for the rest of the discussion, we will call these two methods as the G-B methods. As expected, all procedures yielded average estimates which approach the desired values as model size (m, n) and ρ^2 increase with decreasing standard deviations. For each distribution, higher subgroup size k produced a better estimate for the ratio of variance components among all k in the proposed method. The subgroup size $k \geq 4$ yielded highly biased estimates with high standard deviations for all m. In all distributions, for all ρ^2 and m, the G-B methods failed to perform well in most cases when compared to the proposed method with respect to the standard deviation of the estimates for $k \geq 6$.

For each model size (m, n) and each ρ^2, the proposed method yielded slightly higher average estimates as compared to the G-B methods, but, the standard deviation of the estimates remained consistently smaller for $k \geq 6$. Notice here that the proposed as well as the G-B methods produced average estimates which are positively biased. But the amount of bias decreased when both model size and ρ^2 increased so that the mean square error (m.s.e.) of the estimates reduced significantly when compared to those of the G-B methods. Except for a few smaller model sizes, the mean square errors of the estimates remained consistently smaller for the proposed method for $k \geq 6$ when compared with those of the G-B methods.

The findings for the logistic distribution (see Table 31.2) are quite similar to those in the case of the normal distribution (see Table 31.1) for all procedures except the corresponding average estimates and standard deviations of the estimates are a little higher than those for the normal distribution case. The results obtained for the uniform distribution (see Table 31.3) produced small bias, but significantly smaller standard deviation when compared with those obtained for the normal distribution. The findings in the case of double-exponential responses were similar to those of the uniform distribution for all procedures except that the corresponding average estimates are slightly smaller than those for uniform responses.

In summary, the proposed method with largest subgroup size which equals m produces best results among all k within the proposed method. When compared with the G-B methods, the case when $k \geq 6$ performs better for each distribution. The G-B methods consistently performed poorly with respect to the standard deviation of the estimates for all selected distributions and for every m and ρ^2. Finally, all of the findings suggest choosing the largest possible subgroup size k which equals the treatment size m in order to obtain the best estimates; however, when $m \geq 12$, the proposed estimates were found to be better than those of the G-B methods for subgroup size k as small as 6.

31.4 Adjustment for Bias

From the pseudosamples (31.2) and (31.3), it is clear that the expression on the right hand side of (31.4) and (31.5) approach σ_ϵ^2 and σ_α^2, respectively, when both m and n become large:

$$\begin{aligned} \mathrm{Var}(X_{ij}) &= n^{-1}(n-1)\sigma_\epsilon^2 && (31.4)\\ \mathrm{Var}(Y_i) &= m^{-1}(m-1)\sigma_\alpha^2 + (mn)^{-1}(m-1)\sigma_\epsilon^2. && (31.5) \end{aligned}$$

Hence, for large m and n,

$$\rho^2 = \sigma_\alpha^2/\sigma_\epsilon^2 \doteq \mathrm{Var}(Y_i)/\mathrm{Var}(X_{ij}).$$

However, for small or moderate model sizes, the ratio

$$\begin{aligned} \mathrm{Var}(Y_i)/\mathrm{Var}(X_{ij}) &= \left[\frac{(m-1)\sigma_\alpha^2}{m} + \frac{(m-1)\sigma_\epsilon^2}{mn}\right]\left[\frac{(n-1)\sigma_\epsilon^2}{n}\right]^{-1} \\ &= \frac{n(m-1)}{m(n-1)}\left[\rho^2 + \frac{1}{n}\right] = \tilde{\rho}^2 \quad \text{(say)}, && (31.6) \end{aligned}$$

where $\tilde{\rho}^2$ is the actual value of ρ^2 when we use pseudosamples. Hence, for small or moderate model sizes, we may be estimating values which are quite different from those desired, i.e., we are estimating $\tilde{\rho}^2$, instead of ρ^2.

Using (31.6), we propose the following small-sample adjustment in estimating the ratio of two variance components:

$$Adj\ \hat{\rho}^2 = \frac{m(n-1)}{n(m-1)}\,\hat{\rho}^2 - \frac{1}{n}\,, \tag{31.7}$$

where $\hat{\rho}^2$ is the estimate of ρ^2 obtained from the sample observations. Notice that $Adj\ \tilde{\rho}^2$ depends on model size (m, n). If m and n are very small, then for small ρ^2, the adjustment (31.7) may produce negative estimates. In order to avoid negative estimates, we impose the restriction

$$\hat{\rho}^2 \geq \frac{m-1}{m(n-1)}\,. \tag{31.8}$$

Hence, if the estimate obtained from sample data satisfies (31.8), we perform the small-sample adjustment in (31.7); otherwise, we leave the estimate unadjusted. The adjustment in (31.7) will improve the performance of the proposed method by reducing the bias since $\hat{\rho}^2$ overestimates the desired value.

Concluding Remarks. The inference may depend on the particular grouping into subsamples. However, for m large (*say* ≥ 25), this should not influence the inference. Miller (1968) has given a jackknife estimate of log of the ratio of variances. Using this estimate as a test criterion for testing the null hypothesis of equality of the variances, Miller (1968, pp. 574-575) surmised that his test may be slightly more efficient than that of Shorack (1969), especially when the underlying populations are light-tailed (the reverse may be true for heavy-tailed distributions such as contaminated normal distributions).

Also, in defining the pseudosamples given by (31.2) and (31.3), they could have been based on deviations from sample medians.

References

1. Ansari, A. R. and Bradley, R. A. (1960). Rank sum test for dispersion, *Annals of Mathematical Statistics*, **31**, 1174–1189.

2. Bhattacharyya, H. T. (1977). Nonparametric estimation of ratio of scale parameters, *Journal of the American Statistical Association*, **72**, 459–463.

3. Groggel, D. J., Wackerly, D. D. and Rao, P. V. (1987). Nonparametric estimation in one-way random effect models, *ASA Proceedings of the Statistical Computing Section*, pp. 102–107.

4. Hodges, J. L. and Lehmann, E. L. (1963). Estimate of location based on rank tests, *Annals of Mathematical Statistics*, **34**, 598–611.

5. Miller, R. G. (1968). Jackknife variances, *Annals of Mathematical Statistics*, **39**, 567–582.

6. Moses, L. E. (1963). Rank tests for dispersion, *Annals of Mathematical Statistics*, **34**, 973–983.

7. Shorack, G. R. (1969). Testing and estimating ratios of scale parameters, *Journal of the American Statistical Association*, **64**, 999–1013.

8. Young, C. W., Legates, J. E. and Farthing, B. R. (1965). Pre- and postnatal influences on growth prolificacy and maternal performance in mice, *Genetics*, **52**, 553–561.

Table 31.1: Mean, standard deviation and mean square error of 3000 estimates for ρ^2 when population is normal

			m = 6					m = 12						m = 18					
			Proposed est. when					Proposed estimate when						Proposed estimate when					
ρ^2	n		k=2	k=3	k=6	Gr.*	Bh.'	k=3	k=4	k=6	k=12	Gr.*	Bh.'	k=3	k=6	k=9	k=18	Gr.*	Bh.'
0.10	6	mean	.5063	.4165	.3869	.3478	.3407	.3704	.3589	.3518	.3489	.3376	.3332	.3481	.3378	.3346	.3341	.3277	.3247
		std	.6665	.3865	.2931	.3030	.3183	.2324	.2074	.1849	.1670	.1980	.2033	.1807	.1455	.1389	.1310	.1552	.1574
		mse	.6093	.2495	.1683	.1532	.1593	.1271	.1101	.0916	.0899	.0957	.0957	.0942	.0777	.0701	.0720	.0759	.0753
	12	mean	.2944	.2535	.2372	.2125	.2067	.2268	.2234	.2194	.2166	.2069	.2044	.2159	.2107	.2082	.2075	.2013	.1996
		std	.2959	.2070	.1644	.1644	.1669	.1359	.1204	.1079	.0986	.1130	.1149	.1060	.0843	.0795	.0757	.0924	.0926
		mse	.1253	.0664	.0458	.0397	.0392	.0345	.0297	.0259	.0233	.0242	.0241	.0247	.0193	.0177	.0173	.0188	.0185
	18	mean	.2311	.2048	.1911	.1689	.1674	.1853	.1819	.1785	.1773	.1703	.1685	.1767	.1744	.1733	.1721	.1677	.1666
		std	.2358	.1638	.1280	.1297	.1317	.1092	.0956	.0864	.0801	.0922	.0928	.0824	.0683	.0640	.0608	.0729	.0729
		mse	.0728	.0378	.0247	.0216	.0219	.0192	.0158	.0141	.0124	.0134	.0133	.0127	.0102	.0094	.0088	.0099	.0097
0.25	6	mean	.7960	.6447	.6014	.5396	.5365	.5767	.5576	.5491	.5450	.5255	.5218	.5443	.5274	.5223	.5219	.5122	.5104
		std	1.028	.6025	.4581	.4698	.4920	.3646	.3244	.2872	.2598	.3064	.3140	.2818	.2259	.2130	.2024	.2413	.2457
		mse	1.356	.5188	.3334	.3046	.3242	.2396	.1998	.1576	.1545	.1698	.1725	.1660	.1280	.1095	.1149	.1270	.1282
	12	mean	.5406	.4618	.4332	.3863	.3825	.4128	.4064	.3985	.3940	.3760	.3730	.3928	.3831	.3789	.3776	.3655	.3636
		std	.5585	.3757	.3043	.3021	.3110	.2483	.2195	.1942	.1783	.2069	.2096	.1919	.1524	.1448	.1379	.1642	.1649
		mse	.3963	.1860	.1262	.1099	.1143	.0882	.0727	.0598	.0525	.0587	.0590	.0572	.0410	.0364	.0353	.0403	.0410
	18	mean	.4473	.4010	.3751	.3336	.3313	.3632	.3586	.3509	.3480	.3341	.3323	.3480	.3428	.3401	.3381	.3289	.3277
		std	.4598	.3182	.2532	.2608	.2645	.2106	.1861	.1680	.1557	.1796	.1809	.1626	.1341	.1250	.1190	.1415	.1417
		mse	.2503	.1241	.0798	.0750	.0766	.0572	.0464	.0401	.0339	.0393	.0395	.0360	.0266	.0238	.0219	.0262	.0261
0.50	6	mean	1.267	1.028	.9595	.8634	.8721	.9179	.8889	.8761	.8716	.8376	.8369	.8736	.8429	.8357	.8356	.8193	.8184
		std	1.613	.9694	.7367	.7558	.7982	.5858	.5161	.4562	.4158	.4892	.5015	.4523	.3605	.3378	.3222	.3861	.3923
		mse	3.193	1.218	.7538	.7034	.7756	.5178	.4175	.3135	.3110	.3532	.3650	.3441	.2476	.2015	.2164	.2510	.2553
	12	mean	.9476	.8100	.7603	.6759	.6744	.7212	.7115	.6978	.6895	.6573	.6555	.6901	.6706	.6633	.6611	.6388	.6374
		std	.9970	.6595	.5390	.5396	.5507	.4535	.3855	.3386	.3114	.3613	.3660	.3333	.2664	.2529	.2414	.2834	.2842
		mse	1.194	.5311	.3582	.3221	.3337	.2369	.1933	.1538	.1329	.1553	.1582	.1472	.1001	.0869	.0842	.0996	.0997
	18	mean	.8113	.7315	.6817	.6061	.6049	.6594	.6519	.6385	.6325	.6079	.6072	.6343	.6228	.6183	.6145	.5987	.5976
		std	.8334	.5781	.4610	.4784	.4863	.3793	.3369	.3032	.2820	.3256	.3270	.2970	.2438	.2274	.2165	.2587	.2594
		mse	.7915	.3877	.2455	.2401	.2475	.1692	.1366	.1168	.0971	.1177	.1184	.1062	.0745	.0657	.0600	.0767	.0768
1.00	6	mean	2.209	1.799	1.676	1.519	1.549	1.597	1.550	1.532	1.524	1.461	1.472	1.534	1.473	1.463	1.463	1.434	1.439
		std	2.769	1.708	1.298	1.342	1.426	1.023	.8960	.7965	.7297	.8600	.8854	.7926	.6293	.5892	.5633	.6744	.6849
		mse	9.134	3.556	2.142	2.067	2.335	1.404	1.105	.8082	.8078	.9525	1.006	.9134	.6199	.4848	.5324	.6433	.6623
	12	mean	1.755	1.505	1.415	1.253	1.260	1.339	1.320	1.295	1.280	1.219	1.221	1.284	1.244	1.233	1.228	1.185	1.188
		std	1.870	1.234	1.009	1.015	1.037	.8030	.7148	.6272	.5782	.6722	.6814	.6152	.4925	.4700	.4479	.5210	.5239
		mse	4.068	1.778	1.190	1.095	1.143	.7597	.6137	.4806	.4129	.5001	.5133	.4597	.3022	.2625	.2526	.3060	.3100
	18	mean	1.536	1.391	1.294	1.152	1.157	1.251	1.241	1.213	1.201	1.156	1.157	1.206	1.182	1.173	1.167	1.138	1.139
		std	1.586	1.098	.8763	.9123	.9282	.7194	.6395	.5744	.5346	.6175	.6221	.5677	.4647	.4321	.4120	.4955	.4979
		mse	2.803	1.360	.8548	.8556	.8862	.5810	.4672	.3959	.3264	.4057	.4119	.3650	.2492	.2170	.1976	.2646	.2673

* Estimates obtained by the method of Groggel et al (1987).

' Estimates obtained by the method of Bhattacharyya (1977).

Table 31.2: Mean, standard deviation and mean square error of 3000 estimates for ρ^2 when population is logistic

			m = 6					m = 12						m = 18					
			Proposed est. when					Proposed estimate when						Proposed estimate when					
ρ^2	n		k=2	k=3	k=6	Gr.*	Bh.'	k=3	k=4	k=6	k=12	Gr.*	Bh.'	k=3	k=6	k=9	k=18	Gr.*	Bh.'
0.10	6	mean	.5421	.4421	.4026	.3851	.3798	.3813	.3728	.3619	.3598	.3618	.3575	.3660	.3531	.3472	.3458	.3574	.3534
		std	.6927	.4324	.3030	.3383	.3594	.2465	.2178	.1938	.1832	.2167	.2205	.1907	.1511	.1403	.1367	.1674	.1692
		mse	.6753	.3040	.1833	.1957	.2074	.1399	.1218	.0996	.1011	.1155	.1149	.1071	.0869	.0764	.0791	.0943	.0928
	12	mean	.3129	.2713	.2523	.2366	.2335	.2429	.2344	.2283	.2258	.2266	.2239	.2288	.2212	.2179	.2153	.2254	.2231
		std	.3591	.2332	.1873	.1953	.2030	.1456	.1330	.1173	.1104	.1275	.1285	.1146	.0962	.0894	.0862	.1055	.1055
		mse	.1742	.0838	.0583	.0568	.0590	.0416	.0358	.0302	.0280	.0323	.0319	.0297	.0239	.0214	.0207	.0268	.0263
	18	mean	.2538	.2230	.2091	.1986	.1956	.1989	.1951	.1887	.1877	.1889	.1873	.1852	.1797	.1764	.1751	.1812	.1804
		std	.2774	.1938	.1567	.1615	.1626	.1205	.1103	.0999	.0957	.1103	.1103	.0911	.0762	.0717	.0695	.0816	.0821
		mse	.1006	.0527	.0364	.0358	.0356	.0243	.0212	.0185	.0169	.0201	.0198	.0156	.0121	.0110	.0105	.0133	.0132
0.25	6	mean	.8502	.6934	.6321	.5938	.5957	.5928	.5792	.5656	.5643	.5576	.5558	.5681	.5487	.5429	.5422	.5461	.5420
		std	1.108	.7138	.5041	.5381	.5758	.3963	.3490	.3169	.3052	.3444	.3548	.3028	.2462	.2313	.2286	.2625	.2653
		mse	1.589	.7062	.4001	.4077	.4510	.2746	.2302	.1826	.1919	.2132	.2194	.1929	.1499	.1274	.1376	.1566	.1556
	12	mean	.5642	.4919	.4626	.4271	.4245	.4380	.4229	.4131	.4113	.4026	.3994	.4092	.3983	.3942	.3914	.3983	.3960
		std	.6567	.4377	.3693	.3605	.3775	.2721	.2469	.2204	.2142	.2335	.2366	.2060	.1791	.1684	.1652	.1908	.1914
		mse	.5300	.2502	.1816	.1613	.1729	.1094	.0908	.0752	.0719	.0778	.0783	.0678	.0541	.0475	.0473	.0584	.0579
	18	mean	.4963	.4359	.4121	.3839	.3811	.3829	.3774	.3687	.3684	.3608	.3585	.3575	.3483	.3438	.3431	.3436	.3423
		std	.5673	.3943	.3260	.3230	.3267	.2361	.2187	.2035	.1971	.2136	.2144	.1779	.1539	.1451	.1431	.1589	.1601
		mse	.3825	.1900	.1326	.1222	.1239	.0734	.0641	.0580	.0529	.0579	.0577	.0432	.0333	.0298	.0291	.0340	.0341
0.50	6	mean	1.352	1.107	1.016	.9437	.9595	.9413	.9229	.9045	.9049	.8793	.8819	.9009	.8741	.8659	.8691	.8525	.8520
		std	1.782	1.187	.8526	.8934	.9728	.6493	.5725	.5268	.5131	.5536	.5721	.4904	.4061	.3858	.3857	.4190	.4271
		mse	3.903	1.779	.9936	.9951	1.157	.6164	.5066	.3929	.4272	.4503	.4731	.4012	.3049	.2498	.2850	.2997	.3063
	12	mean	.9865	.8597	.8138	.7400	.7435	.7610	.7363	.7215	.7203	.6937	.6916	.7080	.6921	.6871	.6848	.6812	.6799
		std	1.146	.7798	.6765	.6312	.6628	.4838	.4364	.3957	.3890	.4094	.4148	.3610	.3181	.3004	.2979	.3300	.3323
		mse	1.550	.7375	.5561	.4561	.4986	.3022	.2463	.2056	.1998	.2052	.2088	.1736	.1381	.1200	.1229	.1417	.1428
	18	mean	.8979	.7872	.7504	.6902	.6897	.6896	.6795	.6678	.6688	.6445	.6430	.6429	.6294	.6231	.6232	.6125	.6114
		std	1.035	.7226	.6090	.5944	.6027	.4291	.3981	.3777	.3660	.3821	.3852	.3234	.2836	.2684	.2665	.2869	.2891
		mse	1.229	.6047	.4336	.3895	.3993	.2201	.1907	.1796	.1625	.1669	.1688	.1250	.0972	.0872	.0862	.0949	.0960
1.00	6	mean	2.367	1.939	1.787	1.638	1.702	1.639	1.603	1.581	1.586	1.589	1.534	1.561	1.525	1.513	1.522	1.460	1.467
		std	3.146	2.126	1.560	1.608	1.709	1.149	1.016	.9462	.9317	.9729	1.006	.8697	.7302	.6989	.7033	.7349	.7543
		mse	11.77	5.403	3.053	2.993	3.700	1.729	1.396	1.078	1.211	1.216	1.299	1.071	.8091	.6435	.7675	.7521	.7877
	12	mean	1.824	1.598	1.517	1.362	1.384	1.407	1.360	1.337	1.338	1.275	1.278	1.304	1.278	1.272	1.271	1.244	1.246
		std	2.101	1.467	1.293	1.175	1.233	.9081	.8126	.7487	.7392	.7614	.7755	.6722	.5967	.5643	.5637	.6035	.6096
		mse	5.096	2.512	1.940	1.512	1.669	.9904	.7903	.6744	.6608	.6554	.6791	.5444	.4338	.3740	.3914	.4238	.4324
	18	mean	1.699	1.491	1.427	1.299	1.307	1.299	1.280	1.265	1.269	1.211	1.213	1.212	1.190	1.182	1.183	1.151	1.151
		std	1.975	1.378	1.174	1.133	1.151	.8141	.7565	.7242	.7031	.7174	.7253	.6169	.5436	.5149	.5138	.5442	.5478
		mse	4.392	2.142	1.562	1.374	1.420	.7522	.6512	.6274	.5668	.5593	.5715	.4255	.3319	.2985	.2977	.3190	.3229

* Estimates obtained by the method of Groggel et al (1987).

' Estimates obtained by the method of Bhattacharyya (1977).

Table 31.3: Mean, standard deviation and mean square error of 3000 estimates for ρ^2 when population is uniform

			m = 6					m = 12						m = 18					
			Proposed est. when					Proposed estimate when						Proposed estimate when					
ρ^2	n		k=2	k=3	k=6	Gr.*	Bh.'	k=3	k=4	k=6	k=12	Gr.*	Bh.'	k=3	k=6	k=9	k=18	Gr.*	Bh.'
0.10	6	mean	.4894	.3918	.3647	.2958	.2892	.3345	.3286	.3289	.3313	.2733	.2685	.3271	.3212	.3248	.3285	.2752	.2712
		std	.6300	.3612	.2750	.2803	.2916	.2116	.1878	.1725	.1577	.1662	.1683	.1673	.1356	.1280	.1229	.1338	.1347
		mse	.5485	.2156	.1457	.1169	.1208	.0998	.0875	.0770	.0784	.0577	.0567	.0796	.0673	.0633	.0673	.0486	.0475
	12	mean	.2682	.2250	.2119	.1623	.1585	.2121	.2058	.2058	.2059	.1659	.1633	.2010	.2009	.2040	.2052	.1665	.1650
		std	.2846	.1704	.1341	.1227	.1234	.1201	.1053	.0943	.0882	.0886	.0891	.0942	.0769	.0704	.0677	.0718	.0722
		mse	.1093	.0447	.0305	.0189	.0186	.0270	.0223	.0201	.0190	.0122	.0119	.0191	.0161	.0155	.0156	.0095	.0094
	18	mean	.2215	.1881	.1777	.1364	.1337	.1705	.1665	.1677	.1688	.1361	.1647	.1659	.1637	.1659	.1666	.1369	.1361
		std	.2190	.1366	.1060	.0969	.0972	.0930	.0807	.0712	.0670	.0687	.0689	.0733	.0596	.0543	.0519	.0553	.0553
		mse	.0627	.0264	.0173	.0107	.0106	.0136	.0109	.0099	.0092	.0060	.0059	.0097	.0076	.0072	.0071	.0044	.0043
0.25	6	mean	.7588	.6006	.5534	.4499	.4426	.5280	.5172	.5153	.5164	.4427	.4375	.5170	.5055	.5084	.5126	.4476	.4443
		std	.9792	.5473	.3908	.3845	.3980	.3198	.2788	.2550	.2300	.2530	.2557	.2531	.2024	.1881	.1783	.2043	.2071
		mse	1.217	.4224	.2448	.1878	.1955	.1795	.1491	.1241	.1239	.1011	.1005	.1353	.1062	.0943	.1007	.0808	.0807
	12	mean	.5127	.4152	.3902	.3067	.3019	.3927	.3796	.3761	.3740	.3204	.3173	.3744	.3686	.3719	.3723	.3238	.3224
		std	.5460	.3054	.2375	.2212	.2234	.2117	.1800	.1580	.1455	.1584	.1594	.1679	.1295	.1178	.1122	.1289	.1295
		mse	.3671	.1206	.0761	.0522	.0526	.0652	.0492	.0409	.0366	.0300	.0299	.0437	.0308	.0279	.0276	.0221	.0220
	18	mean	.4393	.3717	.3476	.2747	.2716	.3410	.3323	.3309	.3311	.2836	.2817	.3332	.3247	.3266	.3273	.2869	.2857
		std	.4295	.2534	.1899	.1824	.1842	.1759	.1493	.1286	.1184	.1319	.1325	.1404	.1079	.0976	.0918	.1063	.1070
		mse	.2203	.0790	.0456	.0339	.0344	.0392	.0291	.0241	.0206	.0185	.0186	.0266	.0172	.0154	.0144	.0127	.0127
0.50	6	mean	1.203	.9551	.8760	.7324	.7312	.8587	.8366	.8272	.8252	.7405	.7396	.8319	.8148	.8152	.8189	.7534	.7521
		std	1.513	.8324	.5794	.5968	.6225	.5048	.4292	.3862	.3427	.4006	.4070	.3956	.3081	.2837	.2646	.3220	.3267
		mse	2.784	.9000	.4771	.4102	.4410	.3835	.2975	.2303	.2232	.2183	.2230	.2715	.1940	.1621	.1717	.1679	.1703
	12	mean	.9053	.7299	.6796	.5479	.5438	.6952	.6720	.6594	.6543	.5830	.5806	.6646	.6488	.6512	.6505	.5932	.5924
		std	.9370	.5140	.3852	.3755	.3817	.3649	.3027	.2623	.2363	.2729	.2752	.2874	.2131	.1934	.1824	.2220	.2235
		mse	1.042	.3170	.1807	.1433	.1476	.1713	.1212	.0942	.0797	.0813	.0823	.1097	.0676	.0581	.0559	.0580	.0585
	18	mean	.8035	.6782	.6306	.5077	.5049	.6266	.6086	.6025	.6014	.5329	.5311	.6127	.5944	.5948	.5950	.5415	.5406
		std	.7796	.4477	.3262	.3244	.3289	.3141	.2619	.2229	.2016	.2358	.2376	.2518	.1863	.1677	.1563	.1908	.1917
		mse	.6999	.2322	.1235	.1053	.1082	.1147	.0804	.0633	.0509	.0567	.0574	.0761	.0436	.0371	.0335	.0381	.0384
1.00	6	mean	2.093	1.671	1.522	1.309	1.335	1.523	1.481	1.452	1.443	1.349	1.357	1.488	1.434	1.429	1.431	1.380	1.386
		std	2.579	1.399	.9429	1.015	1.088	.8687	.7329	.6400	.5595	.6979	.7111	.6785	.5150	.4683	.4301	.5593	.5682
		mse	7.850	2.409	1.161	1.126	1.297	1.028	.7689	.5434	.5096	.6092	.6338	.6988	.4541	.3556	.3710	.4578	.4720
	12	mean	1.683	1.358	1.257	1.309	1.335	1.300	1.256	1.227	1.214	1.111	1.113	1.245	1.209	1.210	1.206	1.133	1.134
		std	1.701	.9275	.6750	.6804	.6992	.6709	.5444	.4696	.4145	.4994	.5075	.5248	.3768	.3414	.3192	.4071	.4103
		mse	3.363	.9888	.5221	.4638	.4901	.5405	.3621	.2722	.2179	.2617	.2705	.3356	.1860	.1541	.1444	.1836	.1865
	18	mean	1.537	1.292	1.196	.9751	.9768	1.197	1.161	1.145	1.141	1.032	1.032	1.172	1.135	1.131	1.130	1.052	1.053
		std	1.482	.8375	.5965	.6090	.6184	.5895	.4866	.4107	.3665	.4441	.4475	.4734	.3423	.3071	.2838	.3596	.3619
		mse	2.485	.7867	.3945	.3715	.3830	.3865	.2629	.2002	.1544	.1982	.2014	.2538	.1354	.1116	.0976	.1321	.1338

* Estimates obtained by the method of Groggel et al (1987).

' Estimates obtained by the method of Bhattacharyya (1977).

Table 31.4: Mean, standard deviation and mean square error of 3000 estimates for ρ^2 when population is double-exponential

			m = 6					m = 12						m = 18					
			Proposed est. when					Proposed estimate when						Proposed estimate when					
ρ^2	n		k=2	k=3	k=6	Gr.*	Bh.'	k=3	k=4	k=6	k=12	Gr.*	Bh.'	k=3	k=6	k=9	k=18	Gr.*	Bh.'
0.10	6	mean	.5037	.3905	.3634	.2839	.2753	.3306	.3244	.3261	.3292	.2601	.2550	.3234	.3189	.3239	.3279	.2623	.2584
		std	.7007	.3638	.2716	.2663	.2662	.2094	.1854	.1699	.1553	.1612	.1632	.1658	.1340	.1273	.1215	.1295	.1299
		mse	.6540	.2167	.1431	.1048	.1016	.0970	.0848	.0750	.0767	.0516	.0506	.0774	.0658	.0627	.0667	.0431	.0420
	12	mean	.2726	.2229	.2089	.1513	.1471	.2082	.2021	.2041	.2048	.1557	.1537	.1984	.1993	.2032	.2046	.1570	.1559
		std	.2902	.1768	.1339	.1159	.1158	.1160	.1022	.0924	.0863	.0828	.0833	.0920	.0755	.0691	.0664	.0674	.0676
		mse	.1140	.0464	.0298	.0161	.0156	.0252	.0209	.0194	.0184	.0099	.0098	.0182	.0156	.0151	.0153	.0077	.0076
	18	mean	.2236	.1828	.1711	.1232	.1208	.1675	.1634	.1667	.1677	.1278	.1266	.1632	.1620	.1649	.1659	.1289	.1282
		std	.2271	.1348	.1025	.0897	.0897	.0901	.0779	.0693	.0652	.0638	.0639	.0710	.0579	.0527	.0504	.0514	.0516
		mse	.0669	.0250	.0156	.0085	.0084	.0127	.0101	.0095	.0088	.0048	.0047	.0090	.0071	.0069	.0068	.0034	.0034
0.25	6	mean	.7671	.5938	.5487	.4310	.4236	.5221	.5115	.5113	.5129	.4242	.4203	.5125	.5018	.5066	.5114	.4304	.4279
		std	1.002	.5412	.3853	.3738	.3864	.3146	.2734	.2481	.2237	.2422	.2464	.2487	.1963	.1846	.1741	.1945	.1977
		mse	1.273	.4110	.2377	.1725	.1795	.1730	.1431	.1192	.1191	.0890	.0897	.1307	.1020	.0924	.0986	.0704	.0707
	12	mean	.5153	.4112	.3847	.2892	.2860	.3861	.3736	.3726	.3719	.3049	.3018	.3705	.3661	.3703	.3712	.3109	.3095
		std	.5470	.2969	.2235	.2058	.2093	.2027	.1724	.1524	.1396	.1461	.1472	.1619	.1248	.1130	.1077	.1189	.1194
		mse	.3696	.1141	.0681	.0439	.0451	.0596	.0450	.0383	.0343	.0244	.0243	.0407	.0291	.0265	.0263	.0178	.0178
	18	mean	.4422	.3620	.3412	.2576	.2544	.3355	.3268	.3285	.3289	.2712	.2692	.3285	.3222	.3248	.3259	.2760	.2748
		std	.4330	.2398	.1787	.1659	.1670	.1690	.1417	.1228	.1124	.1204	.1213	.1343	.1031	.0926	.0871	.0975	.0983
		mse	.2244	.0701	.0402	.0276	.0279	.0359	.0260	.0222	.0189	.0149	.0151	.0242	.0158	.0142	.0133	.0102	.0103
0.50	6	mean	1.214	.9461	.8677	.0737	.7032	.8483	.8289	.8209	.8193	.7167	.7176	.8328	.8093	.8122	.8168	.7349	.7350
		std	1.523	.8203	.5636	.5715	.5915	.4921	.4167	.3713	.3283	.3785	.3863	.3849	.2953	.2734	.2542	.3028	.3085
		mse	2.830	.8718	.4529	.3682	.3911	.3635	.2818	.2169	.2097	.1902	.1965	.2589	.1829	.1557	.1650	.1469	.1504
	12	mean	.8943	.7169	.6673	.5170	.5164	.6835	.6611	.6539	.6503	.5619	.5594	.6585	.6446	.6486	.6484	.5771	.5764
		std	.9414	.4960	.3590	.3465	.3564	.3462	.2847	.2486	.2225	.2485	.2518	.2757	.2018	.1826	.1716	.2019	.2040
		mse	1.041	.2930	.1569	.1203	.1273	.1535	.1070	.0855	.0721	.0656	.0669	.1012	.0617	.0535	.0515	.0467	.0474
	18	mean	.8054	.6601	.6188	.4784	.4755	.6156	.5989	.5977	.5974	.5145	.5128	.6043	.5900	.5917	.5926	.5271	.5260
		std	.7852	.4184	.3018	.2909	.2956	.2982	.2457	.2093	.1875	.2128	.2146	.2396	.1747	.1564	.1453	.1724	.1740
		mse	.7097	.2007	.1052	.0851	.0880	.1023	.0702	.0561	.0446	.0455	.0462	.0683	.0386	.0329	.0297	.0304	.0310
1.00	6	mean	2.110	1.652	1.506	1.267	1.294	1.507	1.468	1.441	1.432	1.322	1.334	1.481	1.426	1.424	1.427	1.365	1.372
		std	2.643	1.366	.9030	.9598	1.018	.8454	.7029	.6063	.5270	.6525	.6727	.6600	.4855	.4445	.4057	.5213	.5317
		mse	8.220	2.293	1.071	.9929	1.123	.9721	.7137	.4991	.4651	.5296	.5643	.6669	.4174	.3343	.3471	.4053	.4216
	12	mean	1.693	1.334	1.236	.9812	.9848	1.279	1.236	1.215	1.207	1.079	1.082	1.234	1.202	1.204	1.202	1.113	1.116
		std	1.734	.8818	.6226	.6186	.6351	.6333	.5059	.4385	.3833	.4494	.4564	.5020	.3524	.3174	.2951	.3660	.3702
		mse	3.491	.8897	.4435	.3830	.4036	.4792	.3118	.2389	.1899	.2083	.2150	.3072	.1652	.1367	.1280	.1469	.1505
	18	mean	1.532	1.256	1.173	.9233	.9260	1.177	1.142	1.135	1.134	1.002	1.004	1.157	1.126	1.124	1.125	1.031	1.032
		std	1.493	.7780	.5445	.5425	.5540	.5589	.4527	.3808	.3354	.3964	.4010	.4490	.3164	.2824	.2595	.3205	.3240
		mse	2.513	.6708	.3267	.3002	.3124	.3440	.2253	.1722	.1305	.1571	.1608	.2265	.1161	.0954	.0832	.1037	.1060

* Estimates obtained by the method of Groggel et al (1987).

' Estimates obtained by the method of Bhattacharyya (1977).

32

Limit Theorems for M-Processes Via Rank Statistics Processes

M. Hušková

Charles University, Praha, Czech Republic

Abstract: The purpose of this paper is to study limit behavior of processes related to M-estimators using certain properties of related rank statistics processes proved in Hušková (1996b). The results are then employed to obtain the limit distribution of M-statistics for change point problem, particularly, to get the limit behavior of the test statistics for detection of a change and limit behavior of estimators of the change.

Keywords and phrases: M-statistics, rank statistics, weighted maxima, location model

32.1 Introduction

Let $X_1, \ldots, X_n$ be independent random variables with distribution functions $F(\cdot;\theta_1), \ldots, F(\cdot;\theta_n)$, where $\theta_1, \ldots, \theta_n$ are parameters belonging to an open interval $\Theta \subseteq R^1$.

Consider the M-statistics

$$S_k(\psi) = \sum_{i=1}^{k} \psi(X_i; \theta_n(\psi)), \qquad k = 1, \ldots, n, \tag{32.1}$$

where ψ is a score function fulfilling certain conditions and $\theta_n(\psi)$ is the M-estimator generated by the score function ψ defined as any solution of the equation

$$\sum_{i=1}^{n} \psi(X_i; \theta) = 0 \tag{32.2}$$

or some asymptotically equivalent estimator.

We show that, under quite mild conditions, the limit behavior of various weighted supremum and L_p- functionals of $S_k(\psi)$, $k = 1, \dots, n$, is the same as the respective functionals of partial sums of independent variables.

The proofs are based on the fact that if $\theta_1 = \cdots = \theta_n$ (i.e., if $X_1, \dots, X_n$ are i.i.d. random variables) then the random vector $\{\sum_{i=1}^k \psi(X_i; \theta_n^*);\ k = 1, \dots, n\}$ has the same distribution as $\{\sum_{i=1}^k \psi(X_{R_i}; \theta_n^*);\ k = 1, \dots, n\}$, where $(R_1, \dots, R_n)$ is a random permutation of $(1, \dots, n)$ independent of $(X_1, \dots, X_n)$ and θ_n^* is a symmetric statistic of $X_1, \dots, X_n$. Then given $(X_1, \dots, X_n)$, the random vector $\{\sum_{i=1}^k \psi(X_{R_i}; \theta_n^*);\ k = 1, \dots, n\}$ is the vector of two-sample rank statistics, because

$$\Pr[R_1 = r_1, ..., R_n = r_n] = \frac{1}{n!}$$

for any permutation $(r_1, \dots, r_n)$ of $(1, \dots, n)$, asymptotic results on two-sample rank statistics processes [Hušková (1996b)] under the hypothesis of randomness can be employed.

As a consequence, the results on the tests and estimators for the change point problem based on M-statistics (32.1) proved in Antoch and Hušková (1994) and Hušková (1996a) are shown to hold true under weaker assumptions. The present proofs are also much simpler. Some new results are also derived.

Results on other procedures for the problem formulated in (32.15) and (32.16) below are quoted in references.

In the following section we study separately the case $\theta_1 = \cdots = \theta_n$ – the null hypothesis (Section 32.2) and the case $\theta_1 = \cdots = \theta_m \neq \theta_{m+1} = \cdots = \theta_n$ – the alternative hypothesis (Section 32.3).

32.2 Case $\theta_1 = \cdots = \theta_n$

In this section, we suppose that $X_1, \dots, X_n$ are i.i.d. random variables with common distribution $F(x; \theta_0)$, i.e., $\theta_1 = \cdots = \theta_n$ $(= \theta_0)$, $\theta_0 \in \Theta$.

Define the process

$$V_n(t) = S_{[nt]}(\psi)/(\sqrt{n}\sigma_n(\psi)), \qquad t \in (0, 1), \tag{32.3}$$

where

$$\sigma_n^2(\psi) = \frac{1}{n} \sum_{i=1}^n \psi^2(X_i; \theta_n(\psi)). \tag{32.4}$$

At first, we modify Theorem 1 and Theorem 2 in Hušková (1996b) in a way suitable for our purposes.

$\psi(x;\theta)$ be a score function such that, as $n \to \infty$,

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n}\psi^2(X_i;\theta_n(\psi)) \geq c\right] \to 1 \tag{32.5}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|\psi(X_i;\theta_n(\psi))|^{2+\Delta} = O_p(1) \tag{32.6}$$

for some $c > 0$ and for some $\Delta > 0$.

(i) Then, as $n \to \infty$

$$\{V_n(t);\, t \in (0,1)\} \xrightarrow{d} \{B(t);\, t \in (0,1)\},$$

where $\{B(t);\, t \in (0,1)\}$ is a Brownian bridge.

(ii) Let q be a positive function, defined on $(0,1)$, increasing in a neighbourhood of zero and decreasing in a neighbourhood of one. Then,

$$I_{0,1}(q,c) = \int_0^1 \frac{1}{t(1-t)}\exp\left\{-\frac{cq^2(t)}{t(1-t)}\right\}dt < \infty \tag{32.7}$$

for some $c > 0$ if and only if, as $n \to \infty$,

$$\sup_{0\leq t\leq 1}\{|V_n(t)|/q(t)\} \xrightarrow{D} \sup_{0\leq t\leq 1}\{|B(t)|/q(t)\}, \tag{32.8}$$

where $\{B(t); t \in [0,1]\}$ is a Brownian bridge.

(iii) Assertion (32.8) with $q(t) = 1$, $t \in (0,1)$, remains true even if in (32.6) $\Delta = 0$ and

$$n^{-1/2}\max_{1\leq i\leq n}\{|\psi(X_i;\theta_n(\psi))|\} = o_p(1) \tag{32.9}$$

are fulfilled.

(iv) Let w be a positive function on $(0,1)$ and $0 < \alpha < \infty$. Then, as $n \to \infty$,

$$\int_0^1 |V_n(t)|^\alpha/w(t)dt \xrightarrow{D} \int_0^1 (|B(t)|)^\alpha/w(t)\,dt \tag{32.10}$$

if and only if

$$\int_0^1 \frac{(t(1-t))^{\alpha/2}}{w(t)}\,dt < \infty.$$

Theorem 32.2.2 *Let the assumptions of Theorem 32.2.1 be satisfied.*

(i) If $0 < t_1(n) < t_2(n) < n$ *and as* $n \to \infty$

$$u(n) = \frac{(n - t_1(n))t_2(n)}{t_1(n)(n - t_2(n))} \to \infty,$$

then for any $y \in R^1$*, as* $n \to \infty$,

$$\begin{aligned}&\Pr\Big[\sqrt{2\log\log u(n)} \max_{t_1(n)\le k<t_2(n)} \Big\{\sqrt{\frac{n}{k(n-k)}}\,\frac{1}{\sigma_n(\psi)}\,|S_k(\psi)|\Big\}\\ &\quad \le y + 2\log\log u(n) + \frac{1}{2}\log\log\log u(n) - \frac{1}{2}\log(\pi)\Big]\\ &\quad \to \exp\{-2\exp(-y)\}.\end{aligned} \tag{32.11}$$

(ii) If moreover, as $n \to \infty$,

$$G \to \infty, \qquad G^{-1}n^{2/(2+\Delta)}\log n \to 0 \tag{32.12}$$

then, as $n \to \infty$*, for any* $y \in R^1$

$$\begin{aligned}&\Pr\Big[\sqrt{2\log\frac{n}{G}} \max_{G<k\le n} \Big\{\frac{1}{\sqrt{G}}\,\frac{1}{\sigma_n(\psi)}\,|S_k(\psi) - S_{k-G}(\psi)|\Big\}\\ &\quad \le y + 2\log\frac{n}{G} + \frac{1}{2}\log\log\frac{n}{G} - \frac{1}{2}\log\pi\Big] \to \exp\{-2\exp(-y)\}\end{aligned} \tag{32.13}$$

and

$$\begin{aligned}&\Pr\Big[\sqrt{2\log\frac{n}{G}} \max_{G<k<n-G} \Big\{\frac{1}{\sqrt{2G}}\,\frac{1}{\sigma_n(\psi)}\,|S_{k+G}(\psi) - 2S_k(\psi) + S_{k-G}(\psi)|\Big\}\\ &\quad \le y + 2\log\frac{n}{G} + \frac{1}{2}\log\log\frac{n}{G} - \frac{1}{2}\log(4\pi/9)\Big] \to \exp\{-2\exp(-y)\},\end{aligned} \tag{32.14}$$

where $\sigma_n^2(\mathbf{a})$ *is defined in (32.4).*

PROOF OF THEOREMS 32.2.1 AND 32.2.2. Since $X_1, \ldots, X_n$ are i.i.d. random variables, the distribution of $\{\sum_{i=1}^k \psi(X_i; \theta_n(\psi));\ k = 1, \ldots, n\}$ is the same as that of $\{\sum_{i=1}^k \psi(X_{R_i}; \theta_n(\psi));\ k = 1, \ldots, n\}$, where $(R_1, \ldots, R_n)$ is a random permutation of $(1, \ldots, n)$ independent of $(X_1, \ldots, X_n)$. Then given $(X_1, \ldots, X_n)$ the random vector $\{\sum_{i=1}^k \psi(X_{R_i}; \theta_n(\psi));\ k = 1, \ldots, n\}$ forms a vector of rank statistics

$$\sum_{i=1}^k a(R_i), \qquad k = 1, \ldots, n,$$

where $a(i) = \psi(X_i; \theta_n(\psi))$, $i = 1, \ldots, n$. Notice that, due to the definition of $\theta_n(\psi)$, $\overline{a}_n = \frac{1}{n}\sum_{i=1}^n a(i) = 0$. In view of the assumptions (32.5), (32.6) and (32.9), these scores fulfill assumptions of Theorem 1 and Theorem 2 in Hušková (1996b) and hence Theorem 32.2.1 and Theorem 32.2.2 both hold true. ■

Remark 32.2.1 The assertions of Theorem 32.2.1 and Theorem 32.2.2 remain true if $\psi(X_i;\, \theta_n(\psi))$ is replaced by $\psi(X_i; \theta_n^*) - \overline{\psi}_n$, $i = 1, \ldots, n$, where θ_n^* is any symmetric function of $X_1, .., X_n$ and $\overline{\psi}_n = \frac{1}{n}\sum_{i=1}^n \psi(X_i; \theta_n^*)$.

The assumptions (32.5) and (32.6) are quite mild, and they are fulfilled by various sets of assumptions. Typically we need slightly more than these for the asymptotic normality of the estimator $\theta_n(\psi)$. One of the possible sets is formulated in the following theorem in terms of the assumptions on the score function $\psi(x;\theta)$ and the distribution functions $F(x;\theta)$.

Theorem 32.2.3 *Let $X_1, \ldots, X_n$ be i.i.d. random variables with common distribution function $F(x;\theta_0), \theta_0 \in \Theta$, where Θ is an open interval, and let $\psi(x;\theta)$ be a score function fulfilling*

A.1 the score function $\psi(x;\theta)$ is nondecreasing in the second argument;

A.2 the function $\lambda(t) = \int \psi(x;\theta_o + t)dF(x;\theta_o)$, $t \in R^1$ is continuously differentiable at $t = 0$, $\lambda(0) = 0$ and the derivative $\lambda'(0) > 0$;

A.3 $\int \psi^2(x;\theta_0+t)dF(x;\theta_0)$ is bounded away from 0 in a neighborhood of $t = 0$;

A.4 $\int |\psi(x;\theta_0+t)|^{2+\Delta}dF(x;\theta_0)$ is bounded in a neighborhood of $t = 0$ for some $\Delta > 0$.

Then (32.5) and (32.6) hold true.

PROOF. The proof is quite simple. The assumptions of the theorem guarantee the asymptotic normality and the $\sqrt{n}$–consistency of the estimator $\theta_n(\psi)$ [see Theorem 7.2.2A in Serfling (1981)] which together with the assumptions implies (32.5) and (32.6). ■

The above theorems provide limit distributions under the null hypothesis (and also approximations for the critical values) for the test statistics used for the testing problem

$$H_0 : \theta_1 = \cdots = \theta_n = (\theta_0) \tag{32.15}$$

against

$$\begin{aligned} H_1 : \quad & \text{there exist } 1 \leq m < n \text{ such that} \\ & (\theta_0) = \theta_1 = \cdots = \theta_m \neq \theta_{m+1} = \cdots = \theta_n (= \theta_o + \delta_n), \end{aligned} \tag{32.16}$$

where m, θ_0 and δ_n are unknown parameters. This problem is known as the *change point problem* and m is called the *change point*.

At the end, we recall the test statistics based on $S_k(\psi)$ for this testing problem:

$$\begin{aligned}
T_1(q) &= \sup_{0\le t\le 1} \{|V_n(t)|/q(t)\} \\
T_2(w) &= \int_0^1 |V_n(t)|^\alpha / w(t) dt \\
T_3(G) &= \max_{G<k\le n} \left\{ \frac{1}{\sqrt{G}} \frac{1}{\sigma_n(\psi)} |S_k(\psi) - S_{k-G}(\psi)| \right\} \\
T_4(G) &= \max_{G<k<n-G} \left\{ \frac{1}{\sqrt{2G}} \frac{1}{\sigma_n(\psi)} |S_{k+G}(\psi) - 2S_k(\psi) + S_{k-G}(\psi)| \right\}.
\end{aligned}$$

These statistics as the test statistics for the testing problem (32.15) and (32.16) are discussed in Antoch and Hušková (1994) and Hušková (1996a).

32.3 Change Point Alternatives

Applying similar arguments as in the previous section, we derive the limit behavior of $T_1(q_v)$, $T_3(G)$ and $T_4(G)$ and some related statistics under alternative (32.16), where $q_v(t) = \{t(1-t)\}^v$, $0 \le v \le 1/2$.

We consider the following assumptions.

B.1 $X_1, \dots, X_n$ are independent random variables such that $X_1, \dots, X_m$ have common distribution function $F(x; \theta_0)$ and $X_{m+1}, \dots, X_n$ have common distribution function $F(x; \theta_0 + \delta_n)$, where m, δ_n and $\theta_0 \in \Theta$, are unknown parameters, Θ is an open interval;

B.2 there exists $\eta \in (0,1)$ such that $m = [n\eta]$, where $[a]$ denotes the integer part of a;

B.3 as $n \to \infty$,

$$\delta_n \to 0, \qquad |\delta_n| \sqrt{\frac{n}{\log\log n}} \to \infty;$$

B.4 as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \psi^2(X_i; \theta_n(\psi)) \xrightarrow{P} c^2 \tag{32.17}$$

$$\frac{1}{m} \sum_{i=1}^{m} \psi^2(X_i; \theta_n(\psi)) \xrightarrow{P} c^2 \tag{32.18}$$

for some $c^2 > 0$;

B.5 as $n \to \infty$,

$$\frac{n}{m(n-m)\delta_n} S_m(\psi) \xrightarrow{P} b, \qquad b \neq 0.$$

The assumptions (B.1) - (B.5) correspond to the alternative (32.16) when the magnitude of the change δ_n is small ($\delta_n \to 0$).

We shall study here also the following estimators of the change point m that are related to the statistics $T_1(q_v)$ and $T_4(G)$:

$$\widehat{m}_1(v) = \min\left\{k;\ \max_{1 \le i \le n}\left\{\frac{|S_i(\psi)|}{q_v(i/n)}\right\} = \frac{|S_k(\psi)|}{q_v(k/n)}\right\},\ 0 \le v \le 1/2 \tag{32.19}$$

$$\widehat{m}_4(G) = \min\Big\{k;\ \min_{G < i \le n-G}\{|S_{i+G}(\psi) - 2S_i(\psi) + S_{n,i-G}(\psi)|\} = |S_{k+G}(\psi) - 2S_k(\psi) + S_{k-G}(\psi)|\Big\}. \tag{32.20}$$

The estimators corresponding to the other test statistics can be introduced and studied accordingly [see Antoch and Hušková (1994) and Hušková (1996a)]. They have anticipated properties.

Theorem 32.3.1 *Let assumptions (A.1), (B.1) - (B.5) and (32.6) be satisfied. Then the limit distribution of $T_1(q_v)$ is the same as that of*

$$\frac{1}{\sqrt{n} c\, q_v(m/n)} \sum_{i=1}^{m} \psi(X_i; \theta_n(\psi))$$

for $0 \le v \le 1/2$ and

$$\frac{b^2\delta_n^2}{c^2}(\widehat{m}_1(q_v) - m) \xrightarrow{d} \min\Big\{z \in R^1;\ \max\{W(t) - |t| g_v(t),\ t \in R^1\} = W(z) - |z| g_v(z)\Big\}, \tag{32.21}$$

where $0 \le v \le 1/2$,

$$g_v(t) = \begin{cases} (1-v)(1-\eta) + v\eta, & t < 0, \\ (1-v)\eta + v(1-\eta), & t > 0, \end{cases}$$

and

$$W(t) = \begin{cases} W_1(-t), & t < 0, \\ W_2(t), & t > 0, \end{cases}$$

with $\{W_1(t),\ t > 0\}$ and $\{W_2(t),\ t > 0\}$ being independent Wiener processes.

PROOF. Since $X_1, \ldots, X_m$ are i.i.d. random variables and $\theta_n(\psi)$ is a symmetric function of $X_1, \ldots, X_n$, the vector $\{\sum_{i=1}^{k} \psi(X_i; \theta_n(\psi));\ k = 1, \ldots, m\}$ has the same distribution as the random vector $\{\sum_{i=1}^{k}\psi(X_{R_i^-};\ \theta_n(\psi));\ k=1,\dots,m\}$, where $(R_1^-,\dots,R_n^-)$ is a random permutation of $(1,\dots,m)$.

Similarly, since $X_{1+m},..,X_n$ are i.i.d. random variables, $\theta_n(\psi)$ is a symmetric function of $X_1,\dots,X_n$ and since

$$\sum_{i=1}^{k}\psi(X_i;\theta_n(\psi)) = -\sum_{i=1+k}^{n}\psi(X_i;\theta_n(\psi)),\ k=1,\dots,n,$$

the vector $\{\sum_{i=1}^{k}\psi(X_i;\theta_n(\psi));\ k=1+m,\dots,n\}$ has the same distribution as the random vector $\{-\sum_{i=1+k}^{n}\psi(X_{R_i^+};\theta_n(\psi));\ k=1+m,\dots,n\}$, where $(R_{1+m}^+,\dots,R_n^+)$ is a random permutation of $(m+1,\dots,n)$.

Denote

$$\begin{aligned} Z^+(k) &= \textstyle\sum_{i=1+k}^{n}(a(R_i^+)-\overline{a}_m^+), & k&=m+1,\dots,n\\ Z^-(k) &= \textstyle\sum_{i=1}^{k}(a(R_i^-)-\overline{a}_m^-), & k&=1,..,m, \end{aligned}$$

where

$$\begin{aligned} a(i) &= \psi(X_i;\theta_n(\psi)), & i&=1,2,\dots,n,\\ \overline{a}_m^+ &= \tfrac{1}{n-m}\textstyle\sum_{i=m+1}^{n}a(i), & \overline{a}_m^- &= \tfrac{1}{m}\textstyle\sum_{i=1}^{m}a(i). \end{aligned}$$

Notice that

$$\frac{n-m}{n}\,\overline{a}_m^+ + \frac{m}{n}\,\overline{a}_m^- = 0.$$

Then given $X_1,\dots,X_n$, the random vectors $\{Z^+(k);\ k=1+m,\dots,n\}$ and $\{Z^-(k);\ k=1,\dots,m\}$ are independent vectors of two-sample-like rank statistics with

$$\begin{aligned} \Pr[R_1^-=r_1,\dots,R_m^-=r_m] &= \frac{1}{m!}\,,\\ \Pr[R_{m+1}^+=r_{m+1},\dots,R_n^+=r_n] &= \frac{1}{(n-m)!}\,, \end{aligned}$$

where $(r_1,\dots,r_m)$ and $(r_{m+1},\dots,r_n)$ are any permutations of $(1,\dots,m)$ and $(m+1,\dots,n)$, respectively.

Thus the assumptions of this theorem ensure those of Theorem 32.2.1 and Theorem 32.2.2 (i) hold true for $\{Z^+(k);\ k=1+m,\dots,n\}$ and $\{Z^-(k);\ k=1,\dots,m\}$.

Now, we are ready to prove (32.21). By Theorem 32.2.1 applied to $Z^+(k)$, $k=m+1,\dots,n$, and $Z^-(k)$, $k=1,\dots,m$, conditionally given $(X_1,\dots,X_n)$, we have as $n\to\infty$

$$\max_{1\le k\le m-R_n\delta_n^{-2}}\left\{\frac{(Z^-(k))^2}{k(m-k)\delta_n^2}\right\} = o_p(1), \tag{32.22}$$

$$\max_{m-R_n\delta_n^{-2}\le k\le m}\left\{\frac{(Z^-(k))^2}{k}\right\} = o_p(1), \tag{32.23}$$

$$\max_{1\le k\le m-R_n\delta_n^{-2}}\Big\{\frac{|Z^-(k)|}{(m-k)|\delta_n|}\Big\} = o_p(1), \tag{32.24}$$

$$\max_{m+R_n\delta_n^{-2}\le k<n}\Big\{\frac{(Z^+(k))^2}{(n-k)(k-m)\delta_n^2}\Big\} = o_p(1), \tag{32.25}$$

$$\max_{m<k\le m+R_n\delta_n^{-2}}\Big\{\frac{(Z^+(k))^2}{n-k}\Big\} = o_p(1), \tag{32.26}$$

$$\max_{m+R_n\delta_n^{-2}\le k<n}\Big\{\frac{|Z^+(k)|}{(k-m)|\delta_n|}\Big\} = o_p(1) \tag{32.27}$$

for any $R_n \to \infty$, $R_n(\delta_n^2 n)^{-1}$.

Applying slightly modified Theorem 32.2.1 (i) to $\{Z^+(k);\ k=m+1,\dots,n\}$ and $\{Z^-(k);\ k=1,\dots,m\}$, we observe given $(X_1,\dots,X_n$ that as $n\to\infty$

$$\{\delta_n Z^-(m-[t\delta_n^{-2}])/c;\ t\in[0,A]\} \xrightarrow{d} \{W_1(t);\ t\in[0,A]\} \tag{32.28}$$

and

$$\{\delta_n Z^+(m+[t\delta_n^{-2}])/c;\ t\in[0,A]\} \xrightarrow{d} \{W_2(t);\ t\in[0,A]\}, \tag{32.29}$$

where $\{W_2(t);\ t\in[0,A]\}$ and $\{W_2(t);\ t\in[0,A]\}$, are independent Wiener processes, $A>0$. The relations (32.22) – (32.29) holds true also unconditionally.

Then using these relations we proceed along the line of proof of Theorem 1 in Gombay and Hušková (1996); namely, (32.22) – (32.27) imply the consistency

$$\delta_n^2(\widehat{m}_1(q_v)-m) = O_p(1) \tag{32.30}$$

and the limit distribution of $\widehat{m}_1(q_v)$ can be concluded from (32.23), (32.26), (32.28) and (32.29). The assertion (32.21) is proved.

Concerning the limit distribution of $T_1(q_v)$, we notice that by (32.30) as $n\to\infty$

$$P\Big(T_1(q_v) = \max_{|k-m|\le A_n\delta_n^{-2}}\Big\{\frac{|S_k(\psi)|}{\sqrt{n}\sigma_n(\psi)q_v(k/n)}\Big\}\Big) \to 1$$

for any $A_n\to\infty$. Now, by (32.23), (32.26) and by assumption (B.5), we have

$$\begin{aligned}
&\max_{|k-m|\le A_n\delta_n^{-2}}\Big\{|S_k(\psi)-S_m(\psi)|\Big\}\\
&= O_p\Big(\max_{|k-m|\le A_n\delta_n^{-2}}\{|S_k(\psi)-\frac{k}{m}S_m(\psi)|\}\Big) + O_p\Big(|S_m(\psi)|\frac{A_n}{\delta_n^2 n}\Big)\\
&= O_p\Big(\frac{A_n}{\delta_n\sqrt{n}}\Big).
\end{aligned}$$

Then choosing A_n such that $A_n\to\infty$ and $\frac{A_n}{\delta_n\sqrt{n}}\to 0$ and realizing that $q_v(t)$ has a finite and bounded away from zero derivative in a neighborhood of $t=\eta$, we conclude that $T_1(q_v)$ has the desired limit distribution. ■

Theorem 32.3.2 *Let assumptions (A.1), (B.1) - (B.5), (32.6) and (32.12) be satisfied and let as $n \to \infty$*

$$\frac{\delta_n \sqrt{G}}{\sqrt{\log\log n}} \to \infty. \tag{32.31}$$

Then the limit distribution of $T_4(G)$ is the same as that of

$$\frac{1}{c\sqrt{2G}}\Big(\sum_{i=m+1}^{m+G} \psi(X_i;\theta_n(\psi)) - \sum_{i=m-G+1}^{m} \psi(X_i;\theta_n(\psi))\Big),$$

and

$$\frac{b^2\delta_n^2}{c^2}\left(\widehat{m}_4(G) - m\right) \xrightarrow{d} \min\Big\{z \in R^1;\, \max\{W(t) - |t|/\sqrt{6},\ t \in R^1\} = W(z) - |z|/\sqrt{6}\Big\}, \tag{32.32}$$

where $\{W(t),\ t \in R^1\}$ is the process defined in Theorem 32.3.1.

PROOF. It is very similar to that of Theorem 32.3.1 and therefore is omitted. ■

The assumptions in Theorems 32.3.1 and 32.3.2 are quite mild. As in Section 32.2, there exist several sets of assumptions ensuring the validity of (B.4) and (B.5). One of the possible sets is formulated in the following theorems in terms of the score function $\psi(x;\theta)$ and the distribution function $F(x;\theta)$.

For this purpose we introduce the functions:

$$\begin{array}{lcll}
\lambda_i(t,v) & = & \int \psi^i(x;\theta_0+t)dF(x;\theta_0+v), & (t,v) \in R^2,\ i=1,2, \\
\lambda_{2+\Delta}(t,v) & = & \int |\psi(x;\theta_0+t)|^{2+\Delta}dF(x;\theta_0+v), & (t,v) \in R^2, \\
\lambda_1^{(1)}(t,v) & = & \frac{\partial}{\partial t}\,\lambda_1(t,v), & (t,v) \in R^2,
\end{array}$$

for some $\Delta > 0$.

Theorem 32.3.3 *Let assumptions (A.1) and (B.1) - (B.3) be satisfied and let*

C.1 the first partial derivatives of $\lambda_1(t,v)$ are continuous in a neighborhood of $(t,v) = (0,0)$, $\lambda_1(0,0) = 0$, $\lambda_1^{(1)}(t,0) > 0$ in a neighborhood of $t = 0$ and

$$\frac{\partial}{\partial v}\,\lambda_1(t,v)|_{(t,v)=(0,0)} = -\lambda_1^{(1)}(t,0);$$

C.2 $\lambda_2(t,v)$ is positive and continuous at $(t,v) = (0,0)$;

C.3 $\lambda_{2+\Delta}(t,v)$ is bounded in a neighborhood of $(t,v) = (0,0)$

also be satisfied.

Then the assertion (32.21) remains true with $b = \lambda_1'(0,0)$ *and* $c = \lambda_2(0,0)$ *and if, moreover, as* $n \to \infty$

$$|\lambda_1(t,v) - \lambda_1^{(1)}(0,0)(t-v)| \leq D_1|t-v|^{1+\beta}, \qquad |t| + |v| \leq D_2 \tag{32.33}$$

and

$$\sqrt{n}|\delta_n|^{1+\beta} \to 0 \tag{32.34}$$

for some $\beta > 0$, $D_1 > 0$, $D_2 > 0$, *then as* $n \to \infty$

$$(\eta(1-\eta))^{v-1/2}\left(T_1(q_v) - \delta_n\sqrt{n}\frac{\lambda_1^{(1)}(0,0)}{\sqrt{\lambda_2(0,0)}}\frac{\eta(1-\eta)}{q_v(\eta)}\right) \xrightarrow{d} N(0,1). \tag{32.35}$$

PROOF. First, we show that the assumptions (B.4) and (B.5) are satisfied and that $b = \lambda_1^{(1)}(0,0)$, $c = \lambda_2(0,0)$.

Using standard methods one can derive that as $n \to \infty$

$$\theta_n(\psi) = \theta_0 + \delta_n\frac{n-m}{n} + o_p(\delta_n) \tag{32.36}$$

which together with the assumptions (A.1), (C.1) – (C.3) implies as $n \to \infty$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{m}(\psi(X_i;\theta_n(\psi)) - \lambda_1(\theta_n(\psi),0)) = O_p(\frac{1}{\sqrt{n}\delta_n}) = o_p(1), \tag{32.37}$$

$$\frac{1}{m}\sum_{i=1}^{m}\Big(\psi^2(X_i;\theta_n(\psi)) - \lambda_2(\theta_n(\psi),0)\Big) = o_p(1), \tag{32.38}$$

$$\begin{aligned}\frac{1}{n}\Big(&\sum_{i=1}^{m}(|\psi^2(X_i;\theta_n(\psi))|^{2+\Delta} - \lambda_{2+\Delta}(\theta_n(\psi),0)\\ &+ \sum_{i=m+1}^{n}(|\psi^2(X_i;\theta_n(\psi))|^{2+\Delta} - \lambda_{2+\Delta}(\theta_n(\psi),\delta_n)\Big) = o_p(1).\end{aligned} \tag{32.39}$$

Applying again the representation (32.36) and assumption (C.1), we obtain as $n \to \infty$

$$\frac{n}{(n-m)\delta_n}\;\lambda_1(\theta_n(\psi),0)) = \lambda_1^{(1)}(0,0) + o_p(1),$$

$$\begin{aligned}\lambda_s(\theta_n(\psi),0) &= \lambda_s(0,0)) + o_p(1), \qquad s = 2, 2+\Delta,\\ \lambda_s(\theta_n(\psi),\delta_n) &= \lambda_s(0,0) + o_p(1), \qquad s = 2, 2+\Delta.\end{aligned}$$

These relations together with (32.37) – (32.39) imply the assertions.

If additionally (32.33) and (32.34) are satisfied, we have the following representation

$$\begin{aligned}\theta_n(\psi) &= \theta_0 + \delta_n \frac{n-m}{n} - \frac{1}{\lambda_1^{(1)}(0,0)n} \sum_{i=1}^{n} \Big\{ \psi(X_i;\theta_0) \\ &\quad - E\psi(X_i;\theta_0) \Big\} + o_p(n^{-1/2}).\end{aligned}$$

and then

$$\lambda_1(\theta_n(\psi),0) = \delta_n \frac{n-m}{n} \lambda_1^{(1)}(0,0) - \frac{1}{n}\sum_{i=1}^{n} \psi(X_i;\theta_0) + o_p(n^{1/2}).$$

This together with Theorem 32.3.1 and (32.37) ensures that (32.35) holds true. ■

Theorem 32.3.4 *Let assumptions (A.1), (B.1) - (B.3), (C.1) - (C.3), (32.12) and (32.31) be satisfied, then (32.31) remains true with* $b = \lambda_1^{(1)}(0,0)$ *and* $c = \lambda_2(0,0)$. *If, moreover, (32.33) and* $\sqrt{n}|\delta_n|^{1+\beta} \to 0$ *hold true, then as* $n \to \infty$

$$T_4(G) - \sqrt{G}\delta_n \frac{\lambda_1^{(1)}(0,0)}{\sqrt{2\lambda_2(0,0)}} \xrightarrow{d} N(0,1).$$

PROOF. It is very similar to that of Theorem 32.3.3 and hence is omitted. ■

Remark 32.3.1 If the parameter θ is a shift in location, i.e., $F(x;\theta) = F(x-\theta)$, and $\psi(x;\theta) = \psi(x-\theta)$, then the assumptions can be further simplified.

Remark 32.3.2 Concerning the choice of ψ, we take either one of the ψ functions considered in robust statistics or we put

$$\psi(x;\theta) = \frac{f'(x,\theta)}{f(x,\theta)},$$

where $f'(x;\theta) = \frac{\partial}{\partial\theta} f(x;\theta)$. In the later case, the test procedure based on $T_1(q_{1/2})$ is asymptotically equivalent to the maximum likelihood test for the problem formulated at the end of Section 32.2 and the estimator $m_1(1/2)$ is asymptotically equivalent to the maximum likelihood estimator of m.

Acknowledgements. This research was partially supported by grant GAČR-201/94/0472.

References

1. Antoch, J. and Hušková, M. (1989). Some M-tests for detection of a change in linear models, In *Proceedings of the 4th Prague Symposium on Asymptotic Statistics* (Eds., M. Hušková and P. Mandl), pp. 123–136, Praha: Charles University Press.

2. Antoch, J. and Hušková, M. (1994). Procedures for detection of multiple changes in series of independent observations, In *Proceedings of the 5th Prague Symposium on Asymptotic Statistics* (Eds., M. Hušková and P. Mandl), pp. 3–20, Berlin: Physica-Verlag.

3. Antoch, J., Hušková, M. and Veraverbeke, N. (1995). Change point problem and bootstrap, *Journal of Nonparametric Statistics*, **5**, 123–144.

4. Bhattacharrya, P. K. and Brockwell, P. J. (1976). The minimum of an additive process with applications to signal estimation in storage theory, *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **37**, 51–75.

5. Brodsky, B. E. and Darkhovsky, B. S. (1993). *Nonparametric Methods in Change-Point Problems*, Kluwer: Dordrecht.

6. Carlstein, E. (1988). Nonparametric change-point estimation, *Annals of Statistics*, **16**, 188–197.

7. Csörgö, M. and Horváth, L. (1987). Nonparametric tests for the change point problem, *Journal of Statistical Planning and Inference*, **17**, 1–9.

8. Csörgö, M. and Horváth, L. (1988). Invariance principles for changepoint problems, *Journal of Multivariate Analysis*, **27**, 151–168.

9. Csörgö, M. and Horváth, L. (1988). Nonparametric methods for the change point problem, In *Handbook of Statistics—Volume 7* (Eds., P. R. Krishnaiah and C. R. Rao), pp. 403–425, Amsterdam: North-Holland.

10. Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change point estimators, *Annals of Statistics*, **19**, 1471–1475.

11. Gombay, E. and Hušková, M. (1966). Rank based estimators of the change-point, *Submitted for publication.*

12. Huber, P. J. (1981). *Robust Statistics*, New York: John Wiley & Sons.

13. Hušková, M. (1996a). Estimation of a change in linear models, *Statistics & Probability Letters*, **26**, 13–24.

14. Hušková, M. (1996b). Limit theorems for rank statistics, *Statistics & Probability Letters*, (to appear).

15. Sen, P. K. (1981). *Sequential Nonparametrics*, New York: John Wiley & Sons.

16. Serfling, B. (1980). *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons.

Author Index

Subject Index