Statistics for Industry and Technology

Series Editor

N. Balakrishnan
McMaster University
Department of Mathematics and Statistics
1280 Main Street West
Hamilton, Ontario L8S 4K1
Canada

Editorial Advisory Board

Max Engelhardt
EG&G Idaho, Inc.
Idaho Falls, ID 83415

Harry F. Martz
Group A-1 MS F600
Los Alamos National Laboratory
Los Alamos, NM 87545

Gary C. McDonald
NAO Research & Development Center
30500 Mound Road
Box 9055
Warren, MI 48090-9055

Kazuyuki Suzuki
Communication & Systems Engineering Department
University of Electro Communications
1-5-1 Chofugaoka
Chofu-shi
Tokyo 182
Japan

Scan Statistics

Methods and Applications

Joseph Glaz
Vladimir Pozdnyakov
Sylvan Wallenstein
Editors

Birkhäuser
Boston • Basel • Berlin

Editors
Joseph Glaz
Department of Statistics, U-4120
University of Connecticut
215 Glenbrook Rd.
Storrs, CT 06269-4120, USA
joseph.glaz@uconn.edu

Vladimir Pozdnyakov
Department of Statistics, U-4120
University of Connecticut
215 Glenbrook Rd.
Storrs, CT 06269-4120, USA
vladimir.pozdnyakov@uconn.edu

Sylvan Wallenstein
Department of Community
& Preventive Medicine
Box 1057
Mount Sinai School of Medicine
1 Gustave Levy Place
New York, NY 10029, USA
sylvan.wallenstein@mssm.edu

ISBN 978-0-8176-4748-3 e-ISBN 978-0-8176-4749-0
DOI 10.1007/978-0-8176-4749-0

Library of Congress Control Number: 2009926299

Mathematics Subject Classification (2000): 60C05, 60D05, 60G30, 60G35, 60G55, 60G63, 60G70, 60J22, 60M02, 62P10, 62P12, 62P25, 62P30, 62M30, 62N05

Printed on acid-free paper

Birkhäuser Boston is part of Springer Science+Business Media (www.birkhauser.com)

In honor of Joseph I. Naus

Contents

Preface

In the last ten years the area of scan statistics has risen to prominence in the field of applied probability and statistics. A recent search with Google Scholar lists 1780 references to scan statistics, 988 of which are from the last five years. It is quite impressive that about 200 articles on scan statistics are published each year. About 60 percent of the articles focus on spatial scan statistics and their applications. In addition to challenging theoretical problems, the area of scan statistics has exciting applications in many areas of science and technology, including: archaeology, astronomy, bioinformatics, biosurveillance, computer science, electrical engineering, epidemiology, food sciences, genetics, geography, material sciences, molecular biology, physics, reconnaissance, reliability and quality control, and telecommunication.

This volume has been edited in honor of Joseph Naus's seventieth birthday. The leading chapter, "Joseph Naus: Father of the Scan Statistic," by Sylvan Wallenstein, provides a comprehensive and interesting historical account of the early stages of research in the area of scan statistics, initiated by Joseph Naus almost half a century ago. The rest of the chapters have been arranged in alphabetical order of surnames of their leading authors.

In this volume, we have gathered a group of experts in the field of probability and statistics that have made significant contributions to the area of scan statistics, to review major developments in this area over the last ten years and to present recent or new results as well as point out new directions for future research. The contents of this volume illustrate the depth and the diversity of the methods and applications of the area. We hope that this volume will provide a comprehensive survey of the major recent developments in this area of research and will serve as a valuable reference and source for researchers in applied probability and statistics and in many other areas of science and technology. Graduate students interested in this area of research will find this volume to be of great value, as it points out many interesting and challenging research directions that they could pursue. This volume is suitable for use in teaching a graduate-level seminar course in applied probability and statistics.

Our sincere thanks go to all the authors, who showed great enthusiasm and support for this project. We appreciate their cooperation throughout the course of the project in submitting their articles on time and their help in reviewing the manuscripts. Additional thanks go to Mrs. Debbie Iscoe for her

Joseph Naus

support with issues related to typesetting this volume. Our special thanks go to N. Balakrishnan, Series Editor of *Statistics for Industry and Technology*, Regina Gorenshteyn, Associate Editor, and Tom Grasso, Editor, Computational Sciences and Engineering, Birkhäuser Boston (Springer) for their continual support and encouragement throughout the preparation of this volume.

Joseph Glaz thanks his wife, Sarah, and his son, Ron, for their continual loving support and encouragement. Vladimir Pozdnyakov thanks his mother, Valentina, and his late father, Ivan Ivanovich, as many called him, for nurturing Vladimir's interest in mathematics. Sylvan Wallenstein thanks his wife, Helene, for her love and encouragement, as well as for proofreading.

Storrs, CT, USA **J. Glaz**

Storrs, CT, USA **V. Pozdnyakov**

New York, NY, USA **S. Wallenstein**

Contributors

Balakrishnan, N. McMaster University, Hamilton, Ontario, Canada
bala@univmail.cis.mcmaster.ca

Boutsikas, M. University of Piraeus, Piraeus, Greece
mbouts@unipi.gr

Chan, H.P. National University of Singapore, Singapore, Republic of Singapore
stachp@nus.edu.sg

Chen, J. University of Massachusetts, Boston, MA, USA
jie.chen@umb.edu

Cooper, G.F. University of Pittsburgh, Pittsburgh, PA, USA
gfc@pitt.edu

Costa, M.A. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
macosta@ufmg.br

Das, K. Carnegie Mellon University, Pittsburgh, PA, USA
kaustav@cs.cmu.edu

Duarte, A.R. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
andersonrd@ufmg.br

Duczmal, L. Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
duczmal@ufmg.br

Glaz, J. University of Connecticut, Storrs, CT, USA
joseph.glaz@uconn.edu

Fu, J.C. University of Manitoba, Winnipeg, Manitoba, Canada
fu@cc.umanitoba.ca

Haiman, G. UFR de Mathématiques, Université de Lille 1, Lille, France
haiman@ccr.jussieu.fr

Hoh, J. Yale University, New Haven, CT, USA
josephine.hoh@yale.edu

Jiang, X. University of Pittsburgh, Pittsburgh, PA, USA
xij6@pitt.edu

Joshi, S.W. Slippery Rock University of Pennsylvania, Slippery Rock, PA, USA
sharadchandra.joshi@sru.edu

Koli, R.E. Watershed Surveillance and Research Institute, Jalgaon, India
rek.jalasri@gmail.com

Koutras, M. University of Piraeus, Piraeus, Greece
mkoutras@unipi.gr

Kulldorff, M. Harvard University, Boston, MA, USA
martin_kulldorff@hms.harvard.edu

Lou, W.Y.W. University of Toronto, Toronto, Ontario, Canada
wendy.lou@utoronto.ca

Milienos, F. University of Piraeus, Piraeus, Greece
fmilien@unipi.gr

Myers, W.L. Pennsylvania State University, University Park, PA, USA
wlm@psu.edu

Neill, D.B. Carnegie Mellon University, Pittsburgh, PA, USA
neill@cs.cmu.edu

Ng, H.K.T. Southern Methodist University and Baylor Research Institute, Dallas, TX, USA
ngh@mail.smu.edu

Ott. J. Beijing Institute of Genomics, Beijing, China
ottjurg@yahoo.com

Patil, G.P. Pennsylvania State University, University Park, PA, USA
gpp@stat.psu.edu

Perone-Pacifico, M. Sapienza University of Rome, Rome, Italy
marco.peronepacifico@uniroma1.it

Pozdnyakov, V. University of Connecticut, Storrs, CT, USA
vladimir.pozdnyakov@uconn.edu

Preda, C. Faculté de Médecine, Université de Lille 2, Lille, France
cpreda@univ-lille2.fr

Robin, S. AgroParisTech/INRA, Paris, France
stephane.robin@agroparistech.fr

Schbath, S. INRA, Jouy-en-Josas, France
schbath@jouy.inra.fr

Schneider, J. Carnegie Mellon University, Pittsburgh, PA, USA
schneide@cs.cmu.edu

Stefanov, V.T. University of Western Australia, Crawley, Australia
stefanov@maths.uwa.edu.au

Steele, J.M. University of Pennsylvania, Philadelphia, PA, USA
steele@wharton.upenn.edu

Tango, T. National Institute of Public Health, Wako-shi, Japan
tango@niph.go.jp

Tavares, R. Universidade Federal de Ouro Preto, Ouro Preto, Brazil
tavares@iceb.ufop.br

Tu, I-P. Academia Sinica, Taipei, Taiwan
iping@stat.sinica.edu.tw

Verdinelli, I. Sapienza University of Rome, Rome, Italy, and
Carnegie Mellon University, Pittsburgh, PA, USA
isabella@stat.cmu.edu

Wallenstein, S. Mount Sinai School of Medicine, New York, NY, USA
sylvan.wallenstein@mssm.edu

Zhang, N.R. Stanford University, Stanford, CA, USA
nzhang@stanford.edu

List of Tables

List of Figures

1

Joseph Naus: Father of the Scan Statistic

Sylvan Wallenstein

Department of Community and Preventive Medicine, Mount Sinai School of Medicine, New York, NY, USA

Abstract: Currently, the literature on the scan statistic is vast, growing exponentially in diverse directions, with contributions by many researchers and groups. As time goes on, the early history of the problem bears telling. Joseph Naus, the father of the scan statistic, originated the modern work on the topic. The process took almost twenty years to reach maturity; I have chosen Naus (1982) as the definition of this maturity. The very name "scan statistic" does not appear to have become attached to the problem for fifteen years, and the interconnections to what is now one problem, in both statement of the problem and common methods of solution, was far from obvious originally. This chapter will not attempt a full review of all of Naus's statistical contributions, or even a full review of his contributions as they concern the scan statistic. Instead, it will focus on a few themes that had already originated in Naus's first twenty years of written research (1962–1982), and briefly continue with those threads to the present. Since these early themes include such general issues as applications of the scan statistic, mentoring graduate students, and specific methodological issues, the review will encompass a significant portion of Dr. Naus's research, without making claim to being exhaustive regarding either his research or the much broader topic of research he influenced on the scan statistic.

This chapter is divided into five parts:

1. Naus (1963), Naus's Ph.D. thesis, and the state of the art prior to 1965.
2. Naus's six singly authored first papers, covering all aspects of the problem, and focusing on exact solutions.
3. The first jointly published papers with Naus and his first five Ph.D. students working on the scan, focusing on exact values.
4. Two key publications in 1979–1982 that brought various strands together.
5. A shorter description of "later" work focusing on themes previously introduced.

Keywords and phrases: Scan statistic

J. Glaz et al. (eds.), *Scan Statistics: Methods and Applications*, Statistics for Industry and Technology, DOI 10.1007/978-0-8176-4749-0_1,

1.1 Naus (1963): Ph.D. Thesis

Joseph Naus graduated from the City College of New York in 1959 with a BBA in Economics. He began graduate study in Economics at Harvard the following year, where his advisor was Robert Dorfman. He was advised, as preparation for his graduate studies in economics, to broaden his knowledge in several areas — one of them being statistics. One of his first courses was taught by Arthur Dempster. The field intrigued him and seemed (and perhaps was, at the time) appreciably more manageable than the seemingly broader field of economics. At some point within his first year, he switched to the Statistics Department.

In his third year of graduate study (his second in statistics), Naus was spending an appreciable portion of the time in the Applied Science Division of the Operations Evaluation Group (OEG), which was funded through a contract awarded to MIT from the Navy. Naus notes in the preface to his thesis that on April 10, 1961, Jacinto Steinhardt outlined "Two Probability Problem Areas of Immediate Concern" to the OEG group. The first problem is stated as arising from "naval needs," which was probably motivated with the Navy wanting to know something about future buildup of naval forces in one region of the ocean. Nevertheless, from what Joe remembers, the problem was stated in general terms, though apparently with some emphasis on the two-dimensional problem. Naus, as a member of the OEG, began work on the problem in the fall of 1961 and on June 29, 1962 wrote up his results, Naus (1962), in ASD (Applied Science Division) Paper 8. This technical report, written before the thesis, is referenced in a footnote in Ederer, Myers, and Mantel (1964), which is apparently the first citation of Naus's work.

This line of research continued in a later contract with the Navy and culminated in a thesis approved in October 1963, under the direction of Frederick Mosteller in the Department of Statistics at Harvard, titled "Clustering of Random Points in the Line and Plane." The thesis acknowledged appreciation to Jerome Klotz, who had an appointment in the Business School.

The one-dimensional aspect of the problem, as stated in the thesis, concerns N points independently drawn from a random variable X on $[0, 1)$, with cumulative distribution $F(x)$. $P(k; N, w|F(x))$ is the probability that as some subinterval of length w scans the interval $[0, 1)$, it contains at least k of the N points on that larger interval. (Naus (1963) used the notation n instead of k, and sometimes referred to the problem as the "big N/little n" problem, but in keeping with later literature, this paper will use k for the size of the cluster.) When no argument is given for $F(x)$, X is assumed to follow a uniform distribution, so that $P(k; N, w)$ is the probability that given N points uniformly distributed on $[0, 1)$, there exists a subinterval of width w containing k or more points. As will

be pointed out below, this problem was but one of four parts of the "general" (one-dimensional) problem that would eventually emerge. The four subdivisions of the problem are formed by (i) conditioning, or not, on the total number of points in the interval, and (ii) considering discrete or continuous events. Various aspects of the problem would be studied for two decades, with some of the problems addressed in their own papers giving exact solutions. It was not until Naus (1982) that all four problems were put in the same framework and a single generic approximation was given for all four cases. Perhaps twenty years seems like a long time, but it should be noted that in addition to Naus's students and readers, giants of the field such as Mosteller and Karlin who dealt with various aspects of the four-fold problem also "missed" the global connection. In addition, it took considerable time and effort to lay the foundation to find exact values for the probabilities.

Some special cases of the problem had been previously considered. For $k = N$, the problem was one of finding the distribution of the range with the solution given by Burnside (1928, p. 22); for $k = 2$, the problem relates to the smallest distance between N points with the solution given by Parzen (1960, p. 304). Naus (1963) cites Feller (1958), who had noted the problem but stated that it involves complicated sample spaces, and thus implicitly did not have a simple solution.

The other papers most directly related to Naus's thesis project were Silberstein (1945), Berg (1945), and Mack (1948, 1950). These investigators were apparently the first to address the clustering problem beyond the special cases. They focused on the expected number of clusters, a topic that was apparently not to be addressed again for over thirty years when Glaz and Naus (1983) addressed the issue.

Mosteller, Naus's advisor, had worked on the discrete problem, which a decade or so later would be linked to the yet unnamed scan statistic, but for the first decade the link would remain unexplored. In the early 1960s, the two natural extensions to previous work were to $k = 3$ and to $k = N-1$, with $k = 3$ being the more promising. Naus recalls that another student of Mosteller, Tom Lehrer, who would later achieve fame as a well-known musical satirist, worked on the problem for $k = 3$. Apparently unbeknownst to Naus, and to this author ten years later, was a paper by Elteren and Gerrits (1961) that "nibbled" on the $k = 3$ problem by using a direct integration approach for $N = 6, 7, 8$.

One approach of Silberstein and Mack that Naus apparently used was the polynomial approach. Silberstein (1945) had noted that $P(k, N, w)$ is a polynomial in w of order N. Mack (1948) notes that the polynomial expression may change in different regions of $[0, 1)$. Naus exploits this observation in his thesis, as a lemma that helps move from a derivation for a particular value of w (typically $w = 1/L$, L an integer) to all w.

Naus's ground-breaking approach, which perhaps appears obvious in retrospect (but is not really, for one must know its limitations), was to phrase the problem in terms of paths, particularly what he termed 2-paths and L-paths, and then use combinatoric techniques, particularly the reflection principle, which allowed an exact solution to be computed. Whether an event of interest occurred, depended on whether the move of a path down preceded, or followed, a move up. This involved an analogy between points dropping in and out of an interval, and a cluster of points. Two different parts of the distribution were thus tackled: $w = 1/2$, and $k > N/2$. But it is perhaps even harder to realize the situation in which the argument fell apart, and why the condition $k > N/2$ is so critical. This is summarized as a footnote in both Naus (1963) and Naus (1965a).

Chapter 2 of the thesis found $P(k; N, w|F(x))$ for $k > N/2$, and Chapter 3 found limiting distributions. Chapter 4 explored two-dimensional generalizations, while the last chapter gave applications. It is probably the topic in the second chapter that sparked the greatest progress in subsequent papers and in research in the field up to about 1990. Already in the thesis, Naus showed an interest in a wide range of applications, for example, relating his work to work of Daniel Bernoulli concerning the "mutual inclinations of the planets."

In the thesis and in a later paper, Naus contrasts this "scan" approach with that based on a "fixed grid." The contrast can best be illustrated when $w = 1/L$, L an integer, in which case the fixed grid approach is based on the maximum number of events in any of the L intervals, while the "scan approach" is based on the maximum number of cases as the interval of length w scans the $[0, 1)$ interval. Naus seems to have used "scan" in this restricted context, more than in an attempt to label the statistic.

The beginning of Joe's work on the scan coincided with the start of his married life. During this period, he married Sarah Rosen who was originally from New Jersey. They had met after Joe's first year of graduate study. Their first daughter, Alisa, was born in 1962 while Joe was at Harvard, and their second daughter, Laura, was born in 1965 while Joe was at Rutgers. At Harvard, Joe remembers living in a small apartment with minimal extras, and commuting to Harvard by bicycle.

Joseph Irwin Naus's thesis was approved in September 1963, and the Ph.D. degree was awarded officially in January of the following year. In the 1963–1964 academic year, he continued his work full time as an operations research analyst at the Institute of Naval Studies.

1.2 The Early Papers Touching All Aspects of the Problem: 1965–1968

This section covers Naus's first five papers. In addition to the references previously cited, Naus was by this time aware of the asymptotic distribution for a scan-like statistic in Menon (1964). He found that this asymptotic approximation was not adequate. This problem would continue to plague approximations of the scan based on asymptotic theory and continue to provide justification for the search for exact values. Later, approximations, as opposed to asymptotic values, would be used with some measure of success.

Naus's first position after Harvard was as an assistant professor of statistics at Rutgers, joining the department in 1964 and having an appointment there in 1964–1966. As will be noted in this section, the professional collaborations and informal discussions between Naus and fellow faculty members were to be productive.

1.2.1 Maximum cluster of points on a line, Naus (1965a)

Interestingly, this first paper of Naus cites only Berg (1945), Mack (1948), Silberstein (1945), and Naus (1963). Since, as noted above, the contribution of the cited articles involved at most integration methods, the ideas in the paper were generated entirely by Naus, with possible help from his mentors at Harvard, and possibly later, Rutgers.

To understand the context of Naus's work, we introduce only a little notation, almost all in this paragraph. As noted above, derivations are simplified by considering the case $w = 1/L$, L an integer, so that the $[0, 1)$ interval can be viewed as divided into L parts. The event A denotes that one of these L intervals has k or more points, i.e. that at least one of the L cell occupancy numbers is at least k. The event B_i denotes the event that (i) A^c, all the L subintervals contain fewer than k points, and that (ii) there exists an interval of length w that overlaps the ith and $i + 1$st disjoint intervals that contains k or more points. Setting $B = \cup B_i$, $P(k, N, w) = P(A) + P(B)$. As Naus implicitly realizes, calculation of $P(B)$ becomes more complicated to the extent that more intersections of the B_i's have to be considered. The probabilities of intersections become rapidly more complicated, as the number of events increase, particularly for consecutive i's. By keeping k large relative to N, (i) the number of possible intersections of B_i is limited, and (ii) simpler methods can be used to calculate the probabilities needed. Specifically, for $k > N/2$, the only events possible are B_i and $B_i \cap B_{i+1}$. As would not be noted until much later, under this restriction, the probability for the latter event is the sum of two, rather than six, terms.