



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

© Birkhäuser Boston, a part of Springer Science+Business Media, LLC 2009

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Birkhäuser Boston, c/o Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

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	xv
Contributors	xvii
List of Tables	xxi
List of Figures	xxv
1 Joseph Naus: Father of the Scan Statistic	1
<i>S. Wallenstein</i>	
1.1 Naus (1963): Ph.D. Thesis	2
1.2 The Early Papers Touching All Aspects of the Problem: 1965–1968	5
1.2.1 Maximum cluster of points on a line, Naus (1965a) . . .	5
1.2.2 Clustering in two dimensions, Naus (1965b)	6
1.2.3 Power comparisons, Naus (1966a)	7
1.2.4 Application of Karlin–McGregor (1959) theorem, Naus (1966b)	7
1.2.5 Birthday problem #1, Naus (1968)	9
1.3 Joseph Naus’s Students in 1967–1978, Exploitation of Ballot Problem Results, Broadening of Problem	9
1.4 Two Key Publications, 1979–1982	14
1.4.1 Indexed bibliography	14
1.4.2 Approximations	15
1.5 Later Work, Briefly Noted	16
References	20
2 Precedence-Type Tests for the Comparison of Treatments with a Control	27
<i>N. Balakrishnan and H.K.T. Ng</i>	
2.1 Introduction	27
2.2 Review of Precedence-Type Tests	29
2.2.1 Precedence test	30

2.2.2	Weighted maximal precedence test	31
2.2.3	Minimal Wilcoxon rank-sum precedence test	31
2.3	Test Statistics for Comparing $k - 1$ Treatments with Control . .	33
2.3.1	Tests based on precedence statistic	33
2.3.2	Tests based on weighted maximal precedence statistic . .	35
2.3.3	Tests based on minimal Wilcoxon rank-sum precedence statistic	39
2.4	Exact Power Under Lehmann Alternative	41
2.5	Discussion	42
2.6	Illustrative Example	44
Appendix A: Probability Mass Function of (M_2, \dots, M_k) Under the Null Hypothesis		45
Appendix B: Probability Mass Function of (M_2, \dots, M_k) Under the Lehmann Alternative		48
References		53
3	Extreme Value Results for Scan Statistics	55
<i>M.V. Boutsikas, M.V. Koutras, and F.S. Milienos</i>		
3.1	Introduction	55
3.2	Definitions and Notation	57
3.3	The Binary Scan Statistic	60
3.3.1	Bounds and approximations	61
3.3.2	Asymptotic results	67
3.3.3	Extreme value results	70
3.4	Scan Statistic Exceedances	71
3.4.1	Compound Poisson approximation for $W_{n,k,r(u)}$	71
3.4.2	Convergence of threshold-based scan statistics under maximum domain of attraction assumptions . . .	74
3.4.3	Examples	79
References		84
4	Boundary Crossing Probability Computations in the Analysis of Scan Statistics	87
<i>H.P. Chan, I.-P. Tu, and N.R. Zhang</i>		
4.1	Introduction	87
4.2	Theoretical Developments	88
4.3	Applications in Spatial Scan Statistics	92
4.3.1	Searching for a source of muon particles in the sky . . .	93
4.3.2	Case-control epidemiological studies	96

4.4	Recent Applications in Genomics	97
4.4.1	Biomolecular sequence analysis	98
4.4.2	Detecting changes in DNA copy number	100
4.5	Concluding Remarks	103
	References	104
5	Approximations for Two-Dimensional Variable Window Scan Statistics	109
	<i>J. Chen and J. Glaz</i>	
5.1	Introduction	109
5.2	Two-Dimensional Discrete Scan Statistics	110
5.3	Variable Window Discrete-Type Scan Statistics	117
5.3.1	Unconditional case	117
5.3.2	Conditional case	119
5.4	Numerical Results	121
5.4.1	Unconditional case	121
5.4.2	Conditional case	121
5.5	Summary	125
	References	126
6	Applications of Spatial Scan Statistics: A Review	129
	<i>M.A. Costa and M. Kulldorff</i>	
6.1	Introduction	129
6.2	Brief Methodological Overview	130
6.3	Applications in Medical Imaging	132
6.4	Applications in Cancer Epidemiology	132
6.5	Applications in Infectious Disease Epidemiology	134
6.6	Applications in Parasitology	136
6.7	Other Medical Applications	137
6.8	Applications in Veterinary Medicine	138
6.9	Applications in Forestry	138
6.10	Applications in Geology	139
6.11	Applications in Astronomy	139
6.12	Applications in Psychology	140
6.13	Applications to Accidents	140
6.14	Applications in Criminology and Warfare	140
6.15	Applications in Demography	141
6.16	Applications in the Humanities	141
6.17	Scan Statistic Software	141
6.18	Discussion	142
	References	142

7	Extensions of the Scan Statistic for the Detection and Inference of Spatial Clusters	153
	<i>L. Duczmal, A.R. Duarte, and R. Tavares</i>	
7.1	Introduction	153
7.2	Irregularly Shaped Spatial Clusters	154
7.3	Data-Driven Spatial Cluster Detection Models	163
7.4	Applications	167
	References	167
8	1-Dependent Stationary Sequences and Applications to Scan Statistics	179
	<i>G. Haiman and C. Preda</i>	
8.1	Introduction	179
8.2	Application of the Approximations (8.6) and (8.7) to One-Dimensional Scan Statistics	184
8.2.1	Application to one-dimensional continuous scan statistics	184
8.2.2	Application to one-dimensional discrete scan statistics	186
8.3	Application of the Method to Two-Dimensional Scan Statistics	188
8.3.1	Application to continuous scan statistics	189
8.3.2	Application to discrete scan statistics	190
	References	191
9	Scan Statistics in Genome-Wide Scan for Complex Trait Loci	195
	<i>J. Hoh and J. Ott</i>	
9.1	Introduction	195
9.2	Methods	196
9.3	Applications	197
9.3.1	Autism data	197
9.3.2	Schizophrenia data	198
9.3.3	Parkinson's disease data	198
9.3.4	Age-related macular degeneration (AMD) data	199
9.4	Discussion	199
	References	200
10	On Probabilities for Complex Switching Rules in Sampling Inspection	203
	<i>W.Y.W. Lou and J.C. Fu</i>	
10.1	Introduction	203
10.2	Notation and Finite Markov Chain Imbedding	205

10.3 Main Results 206

10.4 Numerical Examples of Switching Rules 210

 10.4.1 Example 1: Tightened to normal inspection 210

 10.4.2 Example 2: Normal to tightened inspection 210

 10.4.3 Example 3: Discontinuation of inspection 211

 10.4.4 Example 4: Three-level modeling 214

10.5 Summary and Discussion 216

References 218

11 Bayesian Network Scan Statistics for Multivariate Pattern Detection 221

D.B. Neill, G.F. Cooper, K. Das, X. Jiang, and J. Schneider

11.1 Introduction 221

 11.1.1 Event surveillance 222

 11.1.2 The spatial scan statistic 223

 11.1.3 The univariate Bayesian spatial scan statistic 225

 11.1.4 Bayesian networks 226

11.2 The Multivariate Bayesian Scan Statistic 228

 11.2.1 Methods 229

 11.2.2 Evaluation 231

 11.2.3 Discussion 232

11.3 The Agent-Based Bayesian Scan Statistic 235

 11.3.1 Methods 236

 11.3.2 Evaluation 238

 11.3.3 Discussion 238

11.4 The Anomalous Group Detection Method 240

 11.4.1 Methods 241

 11.4.2 Evaluation 244

 11.4.3 Discussion 245

References 246

12 ULS Scan Statistic for Hotspot Detection with Continuous Gamma Response 251

G.P. Patil, S.W. Joshi, W.L. Myers, and R.E. Koli

12.1 Introduction 252

12.2 Basic Ideas 253

12.3 ULS Scan Statistic 254

12.4 Computational Aspects 256

12.5 Testing Significance of the Scan Statistic 258

12.6 Gamma Response Model 258

 12.6.1 Monte Carlo simulation 260

12.7	Details of Software Implementation	260
12.8	Construction of the ULS Scan Tree	263
12.9	A Case Study	265
12.9.1	Description of Pennsylvania hexagonal biodiversity data	265
12.9.2	Pennsylvania elevation hotspot and illustrative data items and format	266
12.10	Conclusions	267
	References	268
13	False Discovery Control for Scan Clustering	271
	<i>M. Perone-Pacifico and I. Verdinelli</i>	
13.1	Introduction	271
13.2	The Basics of Multiple Testing	272
13.3	The Method	274
13.3.1	False discovery control for uncountably many tests . . .	275
13.3.2	The test statistic	277
13.4	Clusters Shaving for Bias Correction	278
13.5	Power Increase Through Multiple Bandwidths	280
13.6	Examples	281
13.6.1	Mixture of uniforms	281
13.6.2	Smooth density with diagonal contours	282
13.6.3	Cosmological data	285
	References	286
14	Martingale Methods for Patterns and Scan Statistics	289
	<i>V. Pozdnyakov and J.M. Steele</i>	
14.1	Introduction	289
14.2	Patterns in an Independent Sequence	290
14.2.1	A gambling approach to the expected value	290
14.2.2	Gambling on a generating function	292
14.2.3	Second and higher moments	293
14.3	Compound Patterns and Gambling Teams	294
14.3.1	Expected time	295
14.3.2	The generating function and the second moment	296
14.4	Patterns in Markov Dependent Trials	299
14.4.1	Two-state Markov chains and a single pattern	299
14.4.2	Two-state Markov chains and compound patterns	302
14.4.3	Finite state Markov chains	304
14.5	Applications to Scans	308
14.5.1	Second moments and distribution approximations	309
14.5.2	Scan for clusters of a certain word	312

14.6	Concluding Remarks	316
	References	316
15	How Can Pattern Statistics Be Useful for DNA Motif Discovery?	319
	<i>S. Schbath and S. Robin</i>	
15.1	Introduction	319
15.2	Words with Exceptional Frequency	320
15.2.1	Sequence models	322
15.2.2	Mean and variance for the count	325
15.2.3	Word count distribution	327
15.2.4	p -values and scores of exceptionality	333
15.2.5	Example of DNA motif discovery	335
15.3	Words with Exceptional Distribution	339
15.3.1	Compound Poisson process	339
15.3.2	Words significantly unbalanced between two sequences	339
15.3.3	Detecting regions significantly enriched with or devoid of a word	341
15.4	More Sophisticated Patterns	342
15.4.1	Family of words	342
15.4.2	Structured motifs	345
15.5	Ongoing Research and Open Problems	346
	References	347
16	Occurrence of Patterns and Motifs in Random Strings	351
	<i>V.T. Stefanov</i>	
16.1	Introduction	351
16.2	Patterns: Discrete-Time Models	353
16.3	Patterns: General Discrete-Time and Continuous-Time Models	356
16.3.1	Waiting times	356
16.3.2	Joint generating functions associated with waiting times	358
16.4	Compound Patterns	359
16.4.1	Compound patterns containing a small number of single patterns	359
16.4.2	Weighted counts of compound patterns	361
16.4.3	Structured motifs	362
	References	364

17 Detection of Disease Clustering	369
<i>T. Tango</i>	
17.1 Introduction	369
17.2 Temporal Clustering	370
17.2.1 Disjoint tests	370
17.2.2 Scan statistics for individual time points data	370
17.2.3 Clustering index	371
17.2.4 Other methods	372
17.2.5 Illustration with congenital oesophageal atresia data	373
17.2.6 Illustration with trisomy data	375
17.3 Spatial Clustering	377
17.3.1 Tests based on adjacencies	378
17.3.2 Tests based on scanning regions	378
17.3.3 Spatial scan statistics	380
17.3.4 Clustering index	381
17.3.5 Other methods	383
17.3.6 Illustration with gallbladder cancer mortality data	384
17.4 Discussion	386
References	388
Index	393

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-Pacífico, 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

Table 2.1	Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = n_2 = n_3 = n = 10, 15$ and 20	36
Table 2.2	Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 4$, $n_1 = n_2 = n_3 = n_4 = n = 10, 15$ and 20	37
Table 2.3	Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = 10$, $n_2 = n_3 = 15$ and $n_4 = 15, n_2 = n_3 = 20$	38
Table 2.4	Near 5% critical values and exact levels of significance (l.o.s.) for P_1, P_2, T_1, T_2, W_1 and W_2 with $k = 3$, $n_1 = 10$, $n_2 = n_3 = n_4 = 15$ and $n_1 = 15, n_2 = n_3 = n_4 = 20$	38
Table 2.5	Power values under Lehmann alternative for $k = 3$, $n_1 = n_2 = n_3 = 10$, $r = 4(1)10$ and $\gamma_2 = \gamma_3 = \gamma = 0.2(0.2)1.0$	42
Table 2.6	Power values under Lehmann alternative for $k = 4$, $n_1 = \dots = n_4 = 10$, $r = 4(1)10$ and $\gamma_2 = \gamma_3 = \gamma_4 = \gamma = 0.2(0.2)1.0$	43
Table 2.7	Appliance cord life data from Nelson (1982, p. 510) (* denotes censored observations).	45
Table 2.8	Values of (m_{1i}, \dots, m_{8i}) and the statistics $P_{(8)i}$, $M_{(8)i}$ and $W_{(8)i}$ for $i = 2, 3$	45
Table 5.1	Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .001$	121
Table 5.2	Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .0025$	122
Table 5.3	Comparison of power for i.i.d. Bernoulli distribution with $p_0 = .005$	122
Table 5.4	Comparison of power for i.i.d. binomial distribution with $L = 5$ and $p_0 = .001$	122
Table 5.5	Comparison of power for i.i.d. Poisson distribution with $\mu_0 = .001$	123

Table 5.6 Comparison of power for $a = 10$ for i.i.d. Bernoulli model. . **123**

Table 5.7 Comparison of power for $a = 25$ for i.i.d. Bernoulli model. . **124**

Table 5.8 Comparison of power for $a = 50$ for i.i.d. Bernoulli model. . **124**

Table 5.9 Comparison of power for $L = 5$ and $a = 50$ for i.i.d. binomial model. **124**

Table 5.10 Comparison of power for $a = 100$ for i.i.d. Poisson model. . **125**

Table 5.11 Comparison of power for $a = 300$ for i.i.d. Poisson model. . **125**

Table 8.1 Approximations for $\mathbf{P}(S \leq x)$ by approximations (8.25) and (8.7). $T = 1001$ **185**

Table 8.2 Approximations for $\mathbf{P}(S \leq x)$ by Haiman (2007) and Naus (1982), $X_i \sim \mathcal{B}(1, p)$, $p = 0.1$, $m = 30$ **188**

Table 8.3 Approximation for $\mathbf{P}(S \leq n)$. $L = 500$, $K = 500$, $\lambda = 0.01$. **190**

Table 8.4 Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \text{Poisson}(0.25)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$ **191**

Table 8.5 Approximation for $\mathbf{P}(S \leq x) : X_{i,j} \sim \mathcal{B}(5, 0.05)$, $m_1 = m_2 = 5$, $L = 5$, $K = 5$, $M = 10^9$ **191**

Table 10.1 Distribution of $W(\Lambda)$ for some selected ρ, p and k in Example 1. **210**

Table 10.2 Samples of the waiting time distribution of $W(\Lambda)$ in Example 2 with $\Lambda = \bigcup_{i=1}^4 \Lambda_i$ and $k = 5$ **212**

Table 12.1 Computational time for selected datasets. **257**

Table 12.2 Biodiversity data for Pennsylvania hexagonal tessellates. . . **265**

Table 14.1 Fixed window scans: at least 3 failures out of 10 consecutive trials, $\mathbf{P}(Z_n = 1) = .01$, $\mu = 30822$, $\sigma = 30815$ **310**

Table 14.2 Fixed window scans: at least 4 failures out of 20 consecutive trials, $\mathbf{P}(Z_n = 1) = .05$, $\mu = 481.59$, $\sigma = 469.35$ **310**

Table 14.3 Variable window: at least 2 failures out of 10 trials or at least 3 failures out of 50 trials, $\mathbf{P}(Z_n = 1) = .01$, $\mu = 795.33$, $\sigma = 785.85$ **311**

Table 14.4 Double scans: at least 2 type II failures out of 10 trials or at least 3 failures of any kind out of 10 trials, $\mathbf{P}(Z_n = 1) = .04$, $\mathbf{P}(Z_n = 2) = .01$, $\mu = 324.09$, $\sigma = 318.34$ **312**

Table 15.1 Expected counts of `aagtgcggt` and `accgcactt` in random sequences having on average the same composition as the *H. influenzae* complete genome. **321**

Table 15.2	Statistics of gctggtgg in the complete genome (left) and in the backbone genome (right) of <i>E. coli</i> K12 under various models <i>Mm</i> . The rank is obtained while sorting the 65,536 scores by decreasing order.	336
Table 15.3	The 10 most exceptionally frequent 7-letter words under model M5 in the <i>S. aureus</i> complete genome. Columns correspond respectively to the word, its observed count, its estimated expected count, its normalizing factor, its score of over-representation under model M5, its observed skew and its skew score under model M0.	338
Table 17.1	$n = 35$ cases of oesophageal atresia and tracheo-oesophageal fistula over 2191 days from 1950 to 1955. Day 1 was set as <i>1 January 1950</i> . (Data from Knox, 1959)	374
Table 17.2	Frequency of trisomy among karyotyped spontaneous abortions of pregnancies, by calendar month of the last menstrual period, July 1975 to June 1977, in three New York hospitals. (Data from Wallenstein, 1980; Tango, 1984) . . .	376

List of Figures

Figure 2.1	Schematic representation of a precedence life-test.	30
Figure 3.1	Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the Pareto distribution $F(x) = 1 - x^{-2}$, $x \geq 1$.	81
Figure 3.2	Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the uniform distribution $F(x) = x$, $0 < x < 1$.	82
Figure 3.3	Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the exponential distribution $F(x) = 1 - e^{-x}$, $x \geq 0$.	83
Figure 3.4	Exact (simulated) and approximate distribution for $Y_{m:r:k}$ for the (standard) normal distribution $\Phi(x)$, $x \in \mathfrak{R}$.	83
Figure 4.1	The x coordinate represents the locations of three well-known virus genomes. The y coordinate represents either half the length of the palindromic patterns (top plots), $u^{-1}N_u(t - u/2)$ for the unweighted case (middle plots) or $u^{-1}S_u(t - u/2)$ for the weighted case (bottom plots). The dotted lines are threshold levels corresponding to p-values of 0.001. The inverted triangles are experimentally validated origins of replication.	100
Figure 10.1	The distributions of the waiting time for discontinuation of inspection, $P[W(D) = n]$ versus n for the three inspection levels of MIL STD 105E in Example 3. For Inspection Level I, $p_n = 0.953$, $p_t = 0.809$, and $EW(D) = 1271$; for Inspection Level II, $p_n = 0.984$, $p_t = 0.858$, and $EW(D) = 24421$; and for Inspection Level III, $p_n = 0.985$, $p_t = 0.901$, and $EW(D) = 109381$.	214
Figure 10.2	Distribution of $W(\Lambda)$ for Example 4 with $\Lambda = \bigcup_{i=1}^7 \Lambda_i$ and $p_1 = 0.7, p_2 = 0.2, p_3 = 0.1$.	216

Figure 11.1	Demonstration of the spatial scan statistic.	224
Figure 11.2	Bayesian network representation of the MBSS method. Solid ovals represent observed quantities, and dashed ovals represent hidden quantities that are modeled. The counts $c_{i,m}^t$ are directly observed, while the baselines $b_{i,m}^t$ and the parameter priors for each stream (α_m, β_m) are estimated from historical data.	229
Figure 11.3	Example of a probability map computed by MBSS. Darker shading indicates a higher probability that the given zip code has been affected.	233
Figure 11.4	General Bayesian network representation of stream-based scan approaches. Relative risks $q_{i,m}^t$ are conditioned on the event type E_k and region S , and may be correlated. Counts $c_{i,m}^t$ are conditionally independent given the relative risks $q_{i,m}^t$ and baselines $b_{i,m}^t$	233
Figure 11.5	Bayesian network representation of the ABSS method. Solid oval represents observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is directly observed.	236
Figure 11.6	General Bayesian network representation of agent-based scan approaches. Solid oval represents observed quantities, and dashed ovals represent hidden quantities that are modeled. Each agent's value of C_r is conditioned on the event type E_k and region S , and these values may be correlated by additional hidden nodes.	239
Figure 11.7	Plot of detection precision vs. recall for (left) ED dataset and (right) PIERS dataset, from Das <i>et al.</i> (2008).	244
Figure 12.1	Illustrative data.	255
Figure 12.2	Cells topologically sorted.	256
Figure 12.3	The ULS tree.	256
Figure 12.4	Overall data structure.	261
Figure 12.5	Abstract response model class.	262
Figure 12.6	Input data file for elevation hotspot. The size is 1 here since all cells have the same area.	266
Figure 12.7	Elevation hotspot is in gray.	267
Figure 12.8	Topographical map of Pennsylvania.	267

Figure 13.1	Bias in kernel density estimation: The solid line is the true density f . The dashed line is the expected kernel density estimator f_H , for small (A) and large (B) bandwidths.	279
Figure 13.2	Bias correction: Vertical lines delimit the true clusters. Horizontal lines show not bias-adjusted (A) and bias-adjusted (B) rejection regions for different bandwidths.	280
Figure 13.3	Contour plot of density in (13.11).	282
Figure 13.4	Unshaved (left panels) and shaved (right panels) rejection regions for small, intermediate, and large bandwidths.	283
Figure 13.5	False discovery proportion (panel A) and power (panel B) for unshaved (dashed) and shaved (solid) rejection regions as functions of bandwidth.	284
Figure 13.6	Clusters detected combining different bandwidths.	284
Figure 13.7	True density (A) and detected clusters (B). In plot B, the solid line represents the conservative null hypothesis in (13.3), the dashed line the null in (13.2).	285
Figure 13.8	Observed data points (A) and detected clusters (B).	286
Figure 15.1	Four occurrences of <i>aataa</i> in sequence S leading to two clumps of <i>aataa</i> , the first one of size 1 and the second one of size 3.	326
Figure 15.2	Exceptionality scores for the 65,536 8-letter words in the <i>E. coli</i> backbone. Left: Boxplots of the scores under models M0 to m6. Right: Scores under models M1 (x -axis) and M6 (y -axis).	337
Figure 15.3	Over-representation scores under M5 and skew scores under M0 for the most over-represented 7-letter words (over-representation scores greater than 5) in the complete genome of <i>S. aureus</i> . The four best candidates (motifs A to D) are indicated. Motif C (<i>gaagcgg</i>) is the functional Chi site of <i>S. aureus</i>	338
Figure 15.4	Significance of the intensity peaks for the occurrences of the Chi site of <i>H. influenzae</i>	342
Figure 17.1	The SMRs of gallbladder cancer (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan (1996–2000).	384
Figure 17.2	The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by SaTScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan.	385

- Figure 17.3 The most likely cluster (shaded area) and the secondary cluster (a lighter shaded area) detected by FleXScan for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan. . . . **385**
- Figure 17.4 Two centers of clustering areas (shaded area) detected by Tango's spatial clustering index for gallbladder cancer mortality data (male) in three prefectures, Niigata, Fukushima, and Yamagata, in Japan. **386**

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

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 i th 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.