

---

**PROBABILITY, STATISTICS AND  
MODELLING IN PUBLIC HEALTH**

---

# **PROBABILITY, STATISTICS AND MODELLING IN PUBLIC HEALTH**

Edited by

**MIKHAIL NIKULIN**

Université Victor Segalin Bordeaux 2, France

V. Steklov Mathematical Institute, Saint Petersburg, Russia

**DANIEL COMMENGES**

Université Victor Segalin Bordeaux 2, France

**CATHERINE HUBER**

Université René Descartes, Paris, France



**Springer**

Library of Congress Control Number: 2005052019

ISBN-10: 0-387-26022-6      e-ISBN: 0-387-26023-4

ISBN-13: 978-0387-26022-8

Printed on acid-free paper.

© 2006 Springer Science+Business Media, Inc.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, Inc., 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 in the United States of America.

9 8 7 6 5 4 3 2 1

[springeronline.com](http://springeronline.com)

**Dedicated to Marvin ZELÉN**

---

## Preface

On September 23, 2003 Marvin Zelen was awarded the title of Docteur Honoris Causa de l'Université Victor Segalen Bordeaux 2, Bordeaux, France. Professor Zelen was the third biostatistician to receive this title after David Cox (1999) and Norman Breslow (2001). To mark the occasion and the importance of the contribution of Professor Zelen in development of biostatistics in public health and especially in the *War on Cancer*, a special symposium, *Probabilités, Statistics and Modelling in Public Health*, was organized in Marvin's honor by Daniel Commenges and Mikhail Nikulin. This workshop took place on September 22-23, 2003, in Bordeaux. Several well known biostatisticians from Europe and America were invited. A special issue of *Lifetime Data Analysis* was published (Volume 10, No 4), gathering some of the works discussed at this symposium. This volume gathers a larger number of papers, some of them being extended versions of papers published in the *Lifetime Data Analysis* issue, others being new. We present below several details of the biography of Professor Zelen.

Marvin Zelen is Professor of Statistics at the Harvard School of Public Health in Boston. He is one of the major researchers in the field of statistical methods in public health.

Since 1960, Professor Zelen constantly worked in several fields of applied statistics, specifically in biology and epidemiology of cancer. He is very well known for his work on clinical trials in oncology, on survival analysis, reliability and planning of experiments and prevention. His papers have now become classics among epidemiologists and biostatisticians who work in the field of cancer.

Since 1967, Professor Zelen was involved in different scientific groups such as the Eastern Cooperative Oncology Group, the Veteran's Administration Lung Cancer Group, the Gastrointestinal Tumor Study Group, and the Radiation Therapy Oncology Group to do statistical research in cancer clinical trials in the USA. Professor Zelen made also significant contributions to reliability theory and random processes, mainly Markov and semi-Markov pro-

cesses, in biostatistics and epidemiology. Professor Zelen is famous all over the world for the development of the Biostatistics Department in the Harvard School of Public Health. He received several awards for his contributions to statistical methodology in the biomedical field. Among them, in 1967, the Annual Award, Washington Academy of Science, for Distinguished Work in Mathematics, in 1992, the *Statistician of the Year* award of Boston Chapter of the American Statistical Association, and, in 1996, the *Morse Award for Cancer Research*.

We thank all participants of the workshop in Bordeaux and all colleagues and friends of Marvin for supporting us in the organization of the meeting in Bordeaux and for their contributions in preparation of this volume. Especially we thank Thelma Zelen, Mei-Ling Ting Lee, Stephen Lagakos, Dave Harrington, Bernard Begaud, Roger Salamon, Valia Nikouline, Elizabeth Cure and the participants of the European Seminar *Mathematical Methods for Reliability, Survival Analysis and Quality of Life* for their help in organization of the meeting and preparation of the proceedings. We thank also l'IFR-99 "Santé Publique" for financial support of our project.

We sincerely hope that this volume will serve as a valuable reference for statisticians.

*Mikhail Nikulin, Daniel Commenges and Catherine Huber*, editors  
March, 2005, Bordeaux

---

# Contents

## **Forward and Backward Recurrence Times and Length Biased Sampling: Age Specific Models**

<i>Marvin Zelen</i> .....	1
1 Introduction .....	1
2 Motivating Problems and Preliminary Results .....	2
2.1 Chronic Disease Modeling .....	2
2.2 Early Detection Modeling .....	3
2.3 Preliminary Results .....	3
3 Development of the Chronic Disease Model .....	4
3.1 Forward Recurrence Time Distribution .....	5
3.2 Backward Recurrence Time Distribution .....	6
3.3 Length Biased Sampling and the Survival of Prevalent Cases .....	6
3.4 Chronological Time Modeling .....	8
4 Early Detection Disease Model .....	9
5 Discussion .....	10
References .....	11

## **Difference between Male and Female Cancer Incidence Rates: How Can It Be Explained?**

<i>Konstantin G. Arbeev, Svetlana V. Ukraintseva, Lyubov S. Arbeeve, Anatoli I. Yashin</i> .....	12
1 Introduction .....	12
2 Data .....	14
3 Three Components of the Individual Aging Process .....	15
4 The Incorporated Ontogenetic Model of Cancer .....	16
5 Application of the Ontogenetic Model to Data on Cancer Incidence Rate by Sex .....	17
6 Conclusion .....	20
References .....	21

**Non-parametric estimation in degradation-renewal-failure models**

*V. Bagdonavičius, A. Bikelis, V. Kazakevičius, M. Nikulin* . . . . . 23

1 Introduction . . . . . 23

2 Model. . . . . 24

3 Decomposition of a counting process associated with  $Z(T)$  . . . . . 25

4 Estimation . . . . . 27

    4.1 The data . . . . . 27

    4.2 Estimation of  $\Lambda$  . . . . . 28

    4.3 Large sample properties of  $\hat{\Lambda}$  . . . . . 30

    4.4 Estimation of the probability  $\mathbf{p}_j(\mathbf{z})$  . . . . . 35

References . . . . . 36

**The Impact of Dementia and Sex on the Disablement in the Elderly**

*P.Barberger-Gateau, V.Bagdonavičius, M.Nikulin, O.Zdorova-Cheminade,* . . . . . 37

1 Introduction . . . . . 37

    1.1 Data. . . . . 38

2 Degradation model . . . . . 40

3 Estimation of the mean degradation . . . . . 41

4 Application to the PAQUID data . . . . . 43

    4.1 The estimated mean of the disablement process in men and women . . . . . 43

    4.2 The estimated mean of the disablement process in demented and non-demented subjects . . . . . 43

    4.3 The estimated mean of the disablement process in demented and non-demented men . . . . . 44

    4.4 The estimated mean of the disablement process in demented and non-demented women . . . . . 45

    4.5 The estimated mean of the disablement process in demented men and women. . . . . 46

    4.6 The estimated mean of the disablement process in non-demented men and women. . . . . 47

    4.7 The estimated mean of the disablement process in high and low educated subjects . . . . . 48

5 Joint model for degradation-failure time data . . . . . 49

References . . . . . 50

**Nonparametric Estimation for Failure Rate Functions of Discrete Time semi-Markov Processes**

*Vlad Barbu, Nikolaos Limnios.* . . . . 53

1 Introduction . . . . . 53

2 Preliminaries . . . . . 54

    2.1 The Discrete Time semi-Markov Model . . . . . 54



2.2	Basic Results on semi-Markov Chains Estimation . . . . .	58
3	Failure Rates Estimation . . . . .	59
	Asymptotic Confidence Intervals for Failure Rates . . . . .	63
4	Proofs . . . . .	63
5	Numerical Example . . . . .	68
	References . . . . .	70

**Some recent results on joint degradation and failure time modeling**

	<i>Vincent Couallier</i> . . . . .	73
1	Introduction . . . . .	73
2	Joint models for degradation and failure time modeling . . . . .	74
2.1	Failure time as hitting times of stochastic processes . . . . .	75
	stochastic degradation defined as diffusion . . . . .	75
	A gamma process as degradation process . . . . .	75
	A marked point process as degradation . . . . .	76
	A mixed regression as degradation process : the general path model . . . . .	77
2.2	Failure times with degradation-dependent hazard rate . . . . .	78
2.3	The joint model : a mixed regression model with traumatic censoring . . . . .	79
3	Some recent results in semiparametric estimation in the general path model . . . . .	80
3.1	Linear estimation . . . . .	80
3.2	Nonlinear estimation . . . . .	82
3.3	Estimation of the reliability functions . . . . .	84
	References . . . . .	87

**Estimation in a Markov chain regression model with missing covariates**

	<i>Dorota M. Dabrowska, Robert M. Elashoff, Donald L. Morton</i> . . . . .	90
1	Introduction . . . . .	90
2	The model and estimation . . . . .	92
2.1	The model . . . . .	92
2.2	Example . . . . .	96
2.3	Estimation . . . . .	99
2.4	Random censoring . . . . .	104
3	A data example . . . . .	107
	References . . . . .	117

**Tests of Fit based on Products of Spacings**

	<i>Paul Deheuvels, Gérard Derzko</i> . . . . .	119
1	Introduction and Main Results. . . . .	119
1.1	Introduction. . . . .	119
1.2	Some Relations with the Kullback-Leibler Information . . . . .	121
2	Proofs. . . . .	124

2.1	A useful Theorem.....	124
2.2	Appendix.....	133
	References .....	135

**A Survival Model With Change-Point in Both Hazard and Regression Parameters**

	<i>Dupuy Jean-François</i> .....	136
1	Introduction .....	136
2	Notations and construction of the estimators .....	137
	2.1 Preliminaries .....	137
	2.2 The estimators .....	138
3	Convergence of the estimators.....	139
	References .....	143

**Mortality in Varying Environment**

	<i>M.S. Finkelstein</i> .....	145
1	Introduction .....	145
2	Damage accumulation and plasticity .....	146
	2.1 Proportional hazards .....	146
	2.2 Accelerated life model .....	148
	2.3 Other models .....	151
	2.4 Damage accumulation and plasticity. Period Setting .....	154
3	Concluding remarks .....	157
	References .....	157

**Goodness of Fit of a joint model for event time and nonignorable missing Longitudinal Quality of Life data**

	<i>Sneh Gulati, Mounir Mesbah</i> .....	159
1	Introduction and Preliminaries .....	159
2	The Dropout Process.....	161
3	The Model of Dupuy and Mesbah (2002) .....	162
4	The Test of Goodness of Fit .....	164
5	Conclusion .....	166
6	References .....	166

**Three approaches for estimating prevalence of cancer with reversibility. Application to colorectal cancer**

	<i>C.Gras, J.P.Daurès and B.Tretarre</i> .....	169
1	Introduction .....	169
2	Definitions.....	170
3	Three approaches for estimating prevalences .....	171
	3.1 Transition Rate Method.....	171
	Method .....	171
	Model specifications .....	174
	Mortality rates .....	174
	Incidence rates .....	174

Transition rates from the disease . . . . . 174  
 Age-specific non recovery prevalence estimates . . . . . 175  
 3.2 A parametric model [CD97]. . . . . 175  
     Method . . . . . 176  
     Model specifications . . . . . 177  
 3.3 Counting Method estimates . . . . . 177  
 4 Results . . . . . 178  
 5 Discussion . . . . . 180  
 References . . . . . 184

**On statistics of inverse gamma process as a model of wear**

*B.P. Harlamov* . . . . . 187  
 1 Introduction . . . . . 187  
 2 Inverse process with independent positive increments . . . . . 188  
     Initial definitions . . . . . 188  
     Moments of the first exit time distributions . . . . . 189  
     Inverse gamma process . . . . . 190  
     One-dimensional distribution . . . . . 191  
     Example 1 . . . . . 193  
     Example 2 . . . . . 193  
     Multi-dimensional distribution . . . . . 196  
 3 Estimation of parameters . . . . . 197  
     The direct way of data gathering . . . . . 197  
     Approximate maximum likelihood estimates . . . . . 198  
     Inverse way of data gathering . . . . . 199  
     Inverse way of data gathering when dealing with a  
         continuous wear curve . . . . . 199  
     Soft ware . . . . . 201  
 References . . . . . 201

**Operating Characteristics of Partial Least Squares in  
 Right-Censored Data Analysis and Its Application in  
 Predicting the Change of HIV-I RNA**

*Jie Huang, David Harrington* . . . . . 202  
 1 Introduction . . . . . 203  
 2 Analysis Methods . . . . . 204  
 3 Simulation studies . . . . . 209  
 4 A Description of the Data . . . . . 213  
 5 The Data Analysis . . . . . 215  
 6 Summary and Discussion . . . . . 224  
 References . . . . . 227

**Inference for a general semi-Markov model and a sub-model for independent competing risks**

*Catherine Huber-Carol, Odile Pons, Natacha Heutte* . . . . . 231

1 Introduction . . . . . 231

2 Framework . . . . . 232

3 Independent Competing Risks Model . . . . . 234

4 General Model . . . . . 235

5 Case of a bounded number of transitions . . . . . 238

6 A Test of the Hypothesis of Independent Competing Risks. . . . . 239

7 Proofs . . . . . 241

References . . . . . 244

**Estimation Of Density For Arbitrarily Censored And Truncated Data**

*Catherine Huber, Valentin Solev, Filia Vonta* . . . . . 246

1 Introduction. . . . . 246

2 Partitioning the total observation time . . . . . 247

    2.1 Random covering. . . . . 247

    2.2 Short-cut covering. . . . . 248

    2.3 The mechanism of truncation and censoring . . . . . 249

3 The distribution associated with random covering. . . . . 250

4 The distribution of random vector  $(L(x), R(x), L(z), R(z))$ . . . . . 253

5 The distribution of random vector  $(L(X), R(X), L(Z), R(Z))$ . . . . . 255

6 Maximum likelihood estimators. . . . . 256

    6.1 The bracketing Hellinger  $\varepsilon$ -entropy . . . . . 257

    6.2 Hellinger and Kullback-Leibler distances. . . . . 259

    6.3 Estimation in the presence of a nuisance parameter . . . . . 262

References . . . . . 265

**Statistical Analysis of Some Parametric Degradation Models**

*Waltraud Kahle, Heide Wendt* . . . . . 266

1 Introduction . . . . . 266

2 A Degradation Model . . . . . 267

    2.1 The distribution of  $(T_n)$ . . . . . 268

    2.2 Marking the sequence  $(T_n)$  . . . . . 270

3 Maximum Likelihood Estimates . . . . . 271

4 Moment Estimates . . . . . 274

5 Comparison of Maximum Likelihood and Moment Estimates . . . . . 276

6 Conclusion . . . . . 277

References . . . . . 278

**Use of statistical modelling methods in clinical practice**

*Klyuzhev V.M., Ardashev V.N., Mamchich N.G., Barsov M.I., Glukhova S.I.* . . . . . 280

1 Introduction . . . . . 280

2 Methods of statistical modelling . . . . . 280

3 Results ..... 281  
 References ..... 284

**Degradation-Threshold-Shock Models**

*Axel Lehmann* ..... 286  
 1 Introduction ..... 286  
 2 Degradation-Threshold-Shock-Models ..... 288  
     2.1 Degradation-Threshold-Models ..... 292  
     2.2 Degradation-Shock-Models ..... 293  
 3 Maximum Likelihood Estimation ..... 294  
 4 Concluding remarks ..... 296  
 References ..... 297

**Comparisons of Test Statistics Arising from Marginal Analyses of Multivariate Survival Data**

*Qian H. Li, Stephen W. Lagakos* ..... 299  
 1 Introduction ..... 299  
 2 The WLW Method and Definitions of Test Statistics ..... 301  
 3 Asymptotic Properties of the Test Statistics under Contiguous Alternatives ..... 303  
 4 Comparisons of Test Statistics ..... 304  
     4.1 Equal  $\mu_1, \mu_2, \dots, \mu_K$  ..... 304  
     4.2 Unequal  $\mu_1, \mu_2, \dots, \mu_K$  ..... 306  
     4.3 Special Correlation Structures ..... 306  
 5 Determining Sample Size and  $K$  ..... 307  
 6 Example: Recurring Opportunistic Infections in HIV/AIDS ..... 310  
 7 Discussion ..... 311  
 References ..... 314

**Nonparametric Estimation and Testing in Survival Models**

*Henning Läuter, Hannelore Liero* ..... 319  
 1 Stating the Problem ..... 319  
 2 Nonparametric Estimators ..... 322  
     2.1 Model with censoring ..... 322  
     2.2 The Nelson-Aalen estimator for the cumulative hazard function ..... 323  
     2.3 A kernel estimator for the hazard function ..... 324  
 3 Testing the Hazard Rate ..... 325  
     3.1 An asymptotic  $\alpha$ -test ..... 326  
     3.2 Application to the example ..... 327  
         Conclusions ..... 327  
 4 Some further remarks ..... 328  
 5 About the Extension to the Model with Covariates ..... 329  
 References ..... 331

**Selecting a semi-parametric estimator by the expected log-likelihood**

*Benoit Liqueur, Daniel Commenges* ..... 332

1 Introduction ..... 332

2 The expected log-likelihood as theoretical criterion ..... 334

    2.1 Definitions and notations ..... 334

    2.2 The expected log-likelihood ..... 334

    2.3 Case of right-censored data ..... 335

    2.4 Case of explanatory variable ..... 336

3 Estimation of ELL ..... 336

    3.1 Likelihood cross-validation : LCV ..... 336

    3.2 Direct bootstrap method for estimating ELL ( $ELL_{boot}$  and  $ELL_{iboot}$ ) ..... 337

    3.3 Bias corrected bootstrap estimators ..... 338

4 Simulation ..... 338

    4.1 Kernel estimator ..... 339

    4.2 Penalized likelihood estimator ..... 342

5 Choosing between stratified and unstratified survival models ..... 343

    5.1 Method ..... 343

    5.2 Example ..... 345

6 Conclusion ..... 346

References ..... 347

**Imputing responses that are not missing**

*Ursula U. Müller, Anton Schick, Wolfgang Wefelmeyer* ..... 350

1 Introduction ..... 350

2 Efficient influence functions ..... 352

3 Efficient estimators ..... 357

    Achnowledgment ..... 362

References ..... 362

**Bivariate Decision Processes**

*Martin Newby* ..... 364

1 Introduction ..... 364

2 The Structure of the Model ..... 366

3 Inspection Policies ..... 366

4 The Inspection Cycle ..... 367

    4.1 System Renewal ..... 367

    4.2 Arbitrary Restoration ..... 368

5 Optimal Policies ..... 368

    5.1 Average Cost Criterion ..... 369

    5.2 Total Cost Criterion ..... 369

    5.3 Obtaining Solutions ..... 370

6 Lévy Processes as Degradation Models ..... 370

7 Examples ..... 371

7.1	Maximum Process . . . . .	371
7.2	The Integrated Process . . . . .	372
7.3	The Absolute Value . . . . .	372
7.4	Bessel Processes . . . . .	373
7.5	Models for Imperfect Inspection . . . . .	374
8	Summary . . . . .	375
	References . . . . .	375

**Weighted Logrank Tests With Multiple Events**

	<i>C. Pinçon, O. Pons</i> . . . . .	378
1	Introduction and notations . . . . .	378
2	Asymptotic distribution of $(LR_1, LR_2)'$ under $H_0$ in a copula model . . . . .	380
2.1	Preliminary results for the martingales under $H_0$ . . . . .	381
2.2	Asymptotic distribution of $(LR_1, LR_2)'$ under $H_0$ . . . . .	384
2.3	What if the joint censoring distributions or the joint survival functions differ in groups $A$ and $B$ under $H_0$ ? . . . . .	386
3	Simulations study . . . . .	388
4	Application . . . . .	389
5	Discussion . . . . .	390
	References . . . . .	391

**Explained Variation and Predictive Accuracy in General Parametric Statistical Models: The Role of Model Misspecification**

	<i>Susanne Rosthøj, Niels Keiding</i> . . . . .	392
1	Introduction . . . . .	392
2	Measures of explained variation . . . . .	393
2.1	Definition of the explained variation . . . . .	394
2.2	Estimation of the explained variation . . . . .	395
3	Misspecification and definition of the predictive accuracy . . . . .	397
4	The failure time model . . . . .	399
5	Which estimation method to choose - model based or not? . . . . .	401
6	Acknowledgement . . . . .	402
7	Appendix . . . . .	402
	References . . . . .	403

**Optimization of Breast Cancer Screening Modalities**

	<i>Yu Shen, Giovanni Parmigiani</i> . . . . .	405
1	Introduction . . . . .	405
2	Model . . . . .	407
2.1	Natural History of Breast Cancer . . . . .	407
2.2	Survival Distributions and Mortality . . . . .	410
2.3	Sensitivities of Mammography and Clinical Breast Examinations . . . . .	411
2.4	Costs of Screening Programs . . . . .	412

3 Optimization of Screening Strategies and Sensitivity Analyses . . . 413  
 4 Discussion . . . . . 415  
 References . . . . . 416

**Sequential Analysis of Quality of Life Rasch Measurements**

*Veronique Sebille, Mounir Mesbah* . . . . . 421  
 1 Introduction . . . . . 421  
 2 Methods . . . . . 423  
   2.1 IRT models . . . . . 423  
   2.2 The Rasch Model . . . . . 424  
   2.3 Estimation of the parameters . . . . . 424  
   2.4 Sequential Analysis . . . . . 425  
     Traditional Sequential Analysis . . . . . 425  
     Sequential Analysis based on Rasch measurements . . . . . 426  
     Estimation of parameters . . . . . 427  
     Z and V statistics . . . . . 427  
   2.5 The Sequential Probability Ratio Test and the  
     Triangular Test . . . . . 428  
   2.6 Study framework . . . . . 428  
 3 Results . . . . . 430  
 4 Discussion . . . . . 434  
 5 Conclusion . . . . . 435  
 6 References . . . . . 436  
 7 Appendix 1 . . . . . 438  
   7.1 1. MLE of  $\sigma$  under  $H_0(\mu = \mu_0 = 0)$  . . . . . 438  
   7.2 2. Efficient score:  $Z(X)$  statistic under  $H_0(\mu = \mu_0 = 0)$  . . 438  
   7.3 3. Fisher’s information:  $V(X)$  statistic under  
      $H_0(\mu = \mu_0 = 0)$  . . . . . 438  
 8 Appendix 2 . . . . . 439  
   8.1 Stopping boundaries for the one-sided SPRT and TT . . . 439

**Three Types of Hazard Functions Curves Described**

*Sidorovich G.I., Shamansky S.V., Pop V.P., Rukavicin O.A.* . . . . . 440  
 1 Patients and method . . . . . 440  
 2 Results . . . . . 441  
 References . . . . . 445

**On the Analysis of Fuzzy Life Times and Quality of Life Data**

*Reinhard Viertl* . . . . . 446  
 1 Introduction . . . . . 446  
 2 Fuzzy data . . . . . 447  
 3 Empirical reliability functions for fuzzy life times . . . . . 448  
 4 Generalized classical statistical inference for fuzzy data . . . . . 449  
 5 Generalized Bayesian inference in case of fuzzy information . . . . 450  
 6 Conclusion . . . . . 451  
 References . . . . . 451



**Statistical Inference for Two-Sample and Regression Models with Heterogeneity Effect: A Collected-Sample Perspective**

*Hong-Dar Isaac Wu* ..... 452

1 Introduction ..... 452

2 Two-Sample Models ..... 453

    2.1 Two-sample location-scale model ..... 454

    2.2 Two-sample transformation model ..... 455

3 Hazards Regression ..... 456

4 Non-proportional Hazards Model ..... 460

5 Extensions and Brief Discussion ..... 462

References ..... 463

**Failure Distributions Associated With General Compound Renewal Damage Processes**

*S. Zacks* ..... 466

1 Introduction ..... 466

2 The General Compound Renewal Damage Process, and The Associated Failure Distribution ..... 467

3 Compound Poisson With Exponential Damage ..... 469

4 Compound Poisson With Erlang Damage ..... 473

References ..... 474

**Index** ..... 477

---

## List of Contributors

**K. G. Arbeev** Center for Demographic Studies, Duke University, 2117 Campus Drive, Box 90408, Durham, NC 27708-0408, USA  
arbeev@cds.duke.edu

**L. S. Arbeeva** Ulyanovsk State University, Leo Tolstoy St. 42, 432700 Ulyanovsk, Russia  
arbeev@mail.ru

**V.N. Ardashev** Burdenko Main Military Clinical Hospital, Moscow, Russia

**V. Bagdonavičius**  
Department of Mathematical Statistics, Vilnius University, Lithuania  
vilius@sm.u-bordeaux2.fr

**P. Barberger-Gateau** IFR 99 Santé Publique, Université Victor Segalen Bordeaux 2, France  
nikou@sm.u-bordeaux2.fr

**Vlad Barbu**  
Université de Technologie de Compiègne, Laboratoire

de Mathématiques Appliquées de Compiègne, BP 20529, 60205 Compiègne, France  
barbu@dma.utc.fr

**Barsov M.I.**  
Burdenko Main Military Clinical Hospital, Moscow, Russia

**A. Bikelis**  
Vilnius University, Naugarduko 24, Vilnius, Lithuania  
marius@post.omnitel.net

**D. Commenges** INSERM E0338, Université Victor Segalen Bordeaux 2, 146 rue Léo Saignat, 33076 Bordeaux Cedex, FRANCE.  
daniel.commenges@isped.u-bordeaux2.fr

**V. Couallier**  
Equipe Statistique Mathématique et ses Applications  
U.F.R. Sciences et Modélisation, Université Victor Segalen Bordeaux 2  
146 rue Leo Saignat  
33076 Bordeaux cedex FRANCE  
couallier@sm.u-bordeaux2.fr

**D. M. Dabrowska**

Department of Biostatistics,  
University of California, Los  
Angeles, CA 90095-1772  
dmdabrowska@yahoo.com

**J.P.Daurès**

Laboratoire de Biostatistique,  
Institut Universitaire de  
Recherche Clinique, 641 avenue de  
Doyen Gaston Giraud, 34093  
Montpellier, France.

**P. Deheuvels**

L.S.T.A., Université Paris VI, 7  
avenue du Château, F  
92340 Bourg-la-Reine, France  
pd@ccr.jussieu.fr

**G. Derzko**

Sanofi-Synthélabo Recherche, 371  
rue du Professeur Joseph Blayac,  
34184 Montpellier Cedex 04, France  
Gerard.Derzko@sanofi-aventis.com

**J-F. Dupuy**

Laboratoire de Statistique et  
Probabilités, Université Paul  
Sabatier,  
118, route de Narbonne, 31062  
Toulouse cedex 4, France  
dupuy@math.ups-tlse.fr

**R. M. Elashoff**

Department of Biostatistics,  
University of California, Los  
Angeles, CA 90095-1772

**M.S. Finkelstein**

Department of Mathematical  
Statistics  
University of the Free State  
PO Box 339, 9300 Bloemfontein,  
Republic of South Africa  
and Max Planck Institute for  
Demographic Research  
Konrad-Zuse-Strasse 1  
18057 Rostock, Germany  
FinkelM.SCI@mail.uovs.ac.za

**S.I. Glukhova**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**C. Gras**

Laboratoire de Biostatistique,  
Institut Universitaire de  
Recherche Clinique, 641 avenue de  
Doyen Gaston Giraud, 34093  
Montpellier, France.  
claudine.gras@iurc.montp.inserm.fr

**S. Gulati**

Department of Statistics, The Hon-  
ors College, Florida International  
University,  
Miami, FL 33199,USA  
gulati@fiu.edu

**B.P. Harlamov**

Institute of Problems of Mechanical  
Engineering,  
Russian Academy of Sciences,  
Saint-Petersburg,  
harlamov@random.ipme.ru

**D. Harrington**

Department of Biostatistics,  
Harvard School of Public Health,  
and Department of Biostatistical  
Science, Dana-Farber Cancer  
Institute, 44 Binney Street, Boston,  
Massachusetts 02115, U.S.A.  
dph@jimmy.harvard.edu

**N. Heutte**

IUT de Caen, Antenne de  
Lisieux, Statistique et Traitement  
Informatique des Données. 11,  
boulevard Jules Ferry 14100 Lisieux,  
France  
N.Heutte@lisieux.iutcaen.  
unicaen.fr

**J. Huang**

Department of Preventive Medicine,  
Feinberg School of  
Medicine, Northwestern University,  
680 N. Lake Shore Drive Suite  
1102, Chicago, Illinois 60611, U.S.A.  
jjhuang@northwestern.edu

**C. Huber-Carol**

University Paris 5, 45 rue des  
Saints-Pères, 75270  
Paris Cedex 06, France and U 472  
INSERM, 16bis avenue P-V  
Couturier, 94 800, Villejuif, France  
catherine.huber@univ-paris5.fr

**W. Kahle**

Otto-von-Guericke-University,  
Faculty of Mathematics,  
D-39016 Magdeburg, Germany  
waltraud.kahle@mathematik.  
uni-magdeburg.de

**V. Kazakevičius**

Vilnius University, Naugarduko 24,  
Vilnius, Lithuania  
Vytautas.kazakevicius.maf.vu.lt

**N. Keiding**

Department of Biostatistics,  
University of Copenhagen,  
Blegdamsvej 3, DK-2200 Copen-  
hagen N, Denmark  
nk@biostat.ku.dk

**V.M. Klyuzhev**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**S. W. Lagakos**

Department of Biostatistics, Harvard  
School of Public Health  
655 Huntington Avenue,  
Boston MA 02115  
lagakos@hsph.harvard.edu

**H. Läuter**

Institute of Mathematics, University  
of Potsdam  
laeuter@rz.uni-potsdam.de

**A. Lehmann**

Otto-von-Guericke-University  
Magdeburg  
Institute of Mathematical Stochastics  
PF 4120, D-39016 Magdeburg,  
Germany  
axel.lehmann@mathematik.  
uni-magdeburg.de

**Q. H. Li**

Food and Drug Administration  
Center  
for Drug and Evaluation Research,  
HFD-705  
7500 Standish Place, Metro Park  
North  
(MPN) II, Rockville,  
MD 20855  
liq@cder.fda.gov

**H. Liero**

Institute of Mathematics,  
University of Potsdam  
liero@rz.uni-potsdam.de

**N. Limnios**

Université de Technologie de  
Compiègne, Laboratoire  
de Mathématiques Appliquées de  
Compiègne  
Nikolaos.Limnios@utc.fr

**B. Lique**

Laboratoire de Statistique et Analyse  
des Données,  
BHSM, 1251 avenue centrale BP 47  
38040 Grenoble Cedex 09, FRANCE

**N.G. Mamchich**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**M. Mesbah**

Laboratoire de Statistique Théorique  
et Appliquée (LSTA),  
Université Pierre et Marie Curie -  
Paris VI, Boîte 158, - Bureau  
8A25 - Plateau A. 175 rue du  
Chevaleret,  
75013 Paris, France  
mesbah@ccr.jussieu.fr

**D. L. Morton**

John Wayne Cancer Institute,  
Santa Monica, CA 90404

**U. U. Müller**

Fachbereich 3, Universität Bremen,  
Postfach 330 440, 28334 Bremen,  
Germany  
uschi@math.uni-bremen.de

**M. Newby**

Centre for Risk Management,  
Reliability and Maintenance  
City University  
LONDON EC1V 0HB

**M. Nikulin**

99 Santé Publique, Université Victor  
Segalen Bordeaux 2,  
France  
nikou@sm.u-bordeaux2.fr

**G. Parmigiani**

Departments of Oncology,  
Biostatistics and Pathology  
Johns Hopkins University,  
Baltimore, MD 21205  
gp@jhu.edu

**C. Pinçon**

EA 3614 - Laboratoire de Biomathé-  
matiques  
3, rue du Professeur Laguesse - 59006  
Lille cédex - France.  
cpincon@pharma.univ-lille2.fr

**O. Pons**

Département MIA - INRA  
Domaine de Vilvert - 78352  
Jouy-en-Josas cédex - France.  
Odile.Pons@jouy.inra.fr

**V.P. Pop**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**S. Rosthøj**

Department of Biostatistics,  
University of Copenhagen,  
Blegdamsvej 3, DK-2200 Copen-  
hagen N, Denmark  
S.Rosthoej@biostat.ku.dk

**O.A. Rukavicin**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**A. Schick**

Department of Mathematical  
Sciences, Binghamton University,  
Binghamton, NY 13902-6000, USA  
anton@math.binghamton.edu

**V. Sebille**

Laboratoire de Biostatistiques,  
Faculté de Pharmacie, Université de  
Nantes, 1  
rue Gaston Veil, BP 53508, 44035  
Nantes Cedex 1, France.  
veronique.sebille@univ-nantes.fr

**S.V. Shamansky**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**Yu Shen**

Department of Biostatistics and  
Applied Mathematics  
M. D. Anderson Cancer Center,  
University of Texas  
Houston, TX 77030  
yshen@mdanderson.org

**G.I. Sidorovich**

Burdenko Main Military Clinical  
Hospital, Moscow,  
Russia

**V. Solev**

Steklov Institute of Mathematics at  
St. Petersburg, nab.  
Fontanki, 27 St.Petersburg 191023  
Russia,  
solev@pdmi.ras.ru

**B.Tretarre**

Registre des Tumeurs de l'Hérault,  
bâtiment  
recherche, rue des Apothicaires  
B.P. 4111,  
34091 Montpellier Cedex 5.

**S. V. Ukraintseva**

Center for Demographic Studies,  
Duke University, 2117 Campus  
Drive, Box 90408, Durham, NC  
27708-0408, USA  
ukraintseva@cds.duke.edu

**R. Viertl**

Vienna University of Technology,  
1040 Wien, Austria  
R.Viertl@tuwien.ac.at

**F. Vonta**

Department of  
Mathematics and Statistics, Univer-  
sity of Cyprus P.O. Box 20537,  
CY-1678, Nicosia, Cyprus,  
vonta@ucy.ac.cy

**W. Wefelmeyer**

Mathematisches Institut, Universität  
zu Köln,  
Weyertal 86-90, 50931 Köln,  
Germany  
wefelmeyer@math.uni-koeln.de

**H. Wendt**

Otto-von-Guericke-University,  
Faculty of Mathematics,  
D-39016 Magdeburg, Germany

**H.-D. I. Wu**

School of Public Health, China  
Medical University,  
91 Hsueh-Shih Rd., Taichung 404,  
TAIWAN.  
honda@mail.cmu.edu.tw

**A. I. Yashin**

Center for Demographic Studies,  
Duke University, 2117 Campus  
Drive, Box 90408, Durham, NC  
27708-0408, USA  
yashin@cds.duke.edu

**S. Zacks**

Department of Mathematical  
Sciences  
Binghamton University  
shelly@math.binghamton.edu

**M. Zelen**

Harvard School of Public Health and  
the Dana-Farber Cancer Institute  
Boston, MA 02115, U.S.A.

**O. Zdorova-Cheminade**

Université Victor Segalen  
Bordeaux 2, France

---

# Forward and Backward Recurrence Times and Length Biased Sampling: Age Specific Models

Marvin Zelen<sup>1</sup>

Harvard School of Public Health and the Dana-Farber Cancer Institute  
Boston, MA 02115, U.S.A. `name@email.address`

**Summary.** Consider a chronic disease process which is beginning to be observed at a point in chronological time. The backward recurrence and forward recurrence times are defined for prevalent cases as the time with disease and the time to leave the disease state respectively, where the reference point is the point in time at which the disease process is being observed. In this setting the incidence of disease affects the recurrence time distributions. In addition, the survival of prevalent cases will tend to be greater than the population with disease due to length biased sampling. A similar problem arises in models for the early detection of disease. In this case the backward recurrence time is how long an individual has had disease before detection and the forward recurrence time is the time gained by early diagnosis; i.e. until the disease becomes clinical by exhibiting signs or symptoms. In these examples the incidence of disease may be age related resulting in a non-stationary process. The resulting recurrence time distributions are derived as well as some generalization of length-biased sampling.

## 1 Introduction

Consider a sequence of events occurring over time in which the probability distribution between events is stationary. Consider a randomly chosen interval having endpoints which are events and select at random a time point in the interval. The forward recurrence time is defined as the time from the random time point to the next event; the backward recurrence time is the time from the time point to the previous event; cf. Cox and Miller [CM65].

An example illustrating these recurrence times is the so-called “waiting time paradox”; cf. Feller [FEL71]. Suppose the events are defined as bus arrivals at a particular location. A person arriving at the bus stop has a waiting time until the next bus arrives. The waiting time is the forward recurrence time. The backward recurrence time is how long the person missed the previous bus.

Backward and forward recurrence times play an important role in several biomedical applications. However in many instances the distribution of events

may have a distribution which changes with time. Furthermore time may be chronological or age. In some applications it may be necessary to consider two time scales incorporating both chronological time and age.

In addition, a closely related topic is length biased sampling . Referring to the bus waiting problem, when the individual arrives at the bus stop, she is intersecting a time interval having endpoints consisting of the previous bus arrival and the next arrival. Implicitly these intervals are chosen so that the larger the interval, the greater the probability of selecting it. The selection phenomena is called length bias sampling.

We will consider two motivating examples for generalizing the recurrence time distributions and length biased sampling. One example deals with a model of the natural history of a chronic disease . The other example refers to modeling the early detection of disease . The mathematics of the examples are the same. However, they are both important in applications and we use both to motivate our investigation. This paper is organized as follows. Section 2 describes the two motivating examples and summarizes results for stationary processes. Section 3 develops the model for the chronic disease example; section 4 indicates the necessary changes for the early detection example. The paper concludes with a discussion in section 5.

## 2 Motivating Problems and Preliminary Results

### 2.1 Chronic Disease Modeling

Consider a population and a chronic disease such that at any point in time a person may be disease free ( $S_0$ ), alive with disease ( $S_a$ ) or may have died of the specific disease ( $S_d$ ). The natural history of the disease will be  $S_0 \rightarrow S_a \rightarrow S_d$ . The transitions  $S_0 \rightarrow S_a$  corresponds to the (point) incidence of the disease and  $S_a \rightarrow S_d$  describes the (point) mortality.

Of course an individual may die of other causes or may be cured by treatment. Our interest is in disease specific mortality. Hence an individual who dies of other causes while in  $S_a$  is regarded as being censored for the particular disease. An individual who is cured of a disease will still be regarded as being in  $S_a$  and eventual death due to other causes will be viewed as a censored observation. This model is a progressive disease model and is especially applicable for many chronic diseases — especially some cancers, cardiovascular disease and diabetes.

Consider a study where at some point in time, say,  $t_0$  this population will be studied. At this point in time some individuals will be disease free ( $S_0$ ) while others will be alive with disease ( $S_a$ ). Those in  $S_a$  are prevalent cases. The backward recurrence time is how long a prevalent case has had disease up to the time  $t_0$ . The forward recurrence time refers to the eventual time of death of the prevalent cases using  $t_0$  as the origin. The sum of the backward and forward recurrence times is the total survival of prevalent cases.



## 2.2 Early Detection Modeling

Consider a population in which at any point in time a person may be in one of three states: disease free ( $S_0$ ), pre-clinical ( $S_p$ ), or clinical ( $S_c$ ). The pre-clinical state refers to individuals who have disease, but there are no signs or symptoms. The individual is unaware of having disease. The clinical state refers to the clinical diagnosis of the disease when the disease interferes with the functioning of an organ system or causes pain resulting in the individual seeking medical help leading to the clinical diagnosis of the disease. The natural history of the disease is assumed to be  $S_0 \rightarrow S_p \rightarrow S_c$ . Note that the transition from  $S_0 \rightarrow S_p$  is never observed. The transition  $S_p \rightarrow S_c$  describes the disease incidence. The aim of an early detection program is to diagnose individuals in the pre-clinical state using a special examination. If indeed, the early detection special examination does diagnose disease in the pre-clinical state, the disease will be treated and the natural history of the disease will be interrupted. As a result, the transition  $S_p \rightarrow S_c$  will never be observed. The time gained by earlier diagnosis is the forward recurrence time and the time a person has been in the pre-clinical state before early diagnosis is the backward recurrence time. If  $t_0$  is the time (either age or chronological time) in which the disease is detected, we then have an almost identical model as the chronic disease model simply by renaming the states.

## 2.3 Preliminary Results

Consider a non-negative random variable  $T$  having the probability density function  $q(t)$ . A length biased sampling process chooses units with a probability proportional to  $t$  ( $t < T \leq t + dt$ ). Samples of  $T$  are drawn from a length biased process. Suppose the random variable is randomly split into two parts ( $U, V$ ) so that  $T = U + V$ . The random variable  $U$  and  $V$  are the backward and forward recurrence times. The model assumes that for fixed  $T = t$  ( $t < T \leq t + dt$ ) a point  $u$  is chosen according to a uniform distribution over the interval  $(0, t)$ . Then if  $q_f(v)$  and  $q_b(u)$  are the probability density functions of the forward and backward recurrence times it is well known that with length biased sampling for selecting  $T$ ; cf. Cox and Miller [CM65].

$$q_f(t) = q_b(t) = Q(t)/m, \quad t > 0 \quad (1)$$

where  $Q(t) = \int_t^\infty q(x)dx$  and  $m = \int_0^\infty Q(x)dx$ .

Also the p.d.f. of  $T$  is

$$f(t) = tq(t)/m. \quad (2)$$

Note that the first moments of these distributions are:

$$\int_0^\infty \frac{tQ(t)}{m} dt = \frac{m}{2}(1 + C^2),$$

$$\int_0^\infty \frac{t^2q(t)}{m} dt = m(1 + C^2) \quad (3)$$

where  $C = \sigma/m$  is the coefficient of variation associated with  $q(t)$ . If  $q(t)$  is the exponential distribution with mean  $m$ , the forward and backward recurrence times have the same exponential distribution as  $q(t)$  and  $C = 1$ .

A reviewer suggested that a simpler way to discuss these results is to initially assume that the joint distribution of  $(U, V)$  is  $f(u, v) = q(u+v)I(u \geq 0, v \geq 0)/m..$  Then all the results above are readily derived. Implication in this assumption is  $f(u/T) = 1/t$  and length biased sampling.

### 3 Development of the Chronic Disease Model

In this section we will investigate generalizations of the distribution of the backward and forward recurrence times using the chronic disease model as a motivating example. We remark that for the chronic disease model, the process may have been going on for a long time before being observed at time  $t_0$ .

Suppose at chronological time  $t_0$  the disease process is being observed. The prevalent cases at time  $t_0$  will have an age distribution denoted by  $b(z|t_0)$ . We will initially consider the prevalent cases who have age  $z$ . Later by weighting by the age distribution for the whole population we will derive properties of the prevalent cases for the population. The prevalent cases could be regarded as conditional on the time  $t_0$  when observations began. Another model is that the prevalent cases could be assumed to have arisen by sampling the population at a random point in time which is  $t_0$ . We shall consider both situations.

Define

$$a(z|t_0) = \begin{cases} 1 & \text{if individual of age } z \text{ is in } S_a \text{ at time } t_0. \\ 0 & \text{if individual of age } z \text{ is not in } S_a \text{ at time } t_0, \\ & \text{but was incident with disease before age } z. \end{cases}$$

$$a(t_0) = \begin{cases} 1 & \text{if individual is in } S_a \text{ at time } t_0. \\ 0 & \text{if individual is not in } S_a \text{ at time } t_0, \\ & \text{but was incident with disease before time } t_0. \end{cases}$$

$$P(z|t_0) = P\{a(z|t_0) = 1\}, \quad P_0 = P\{a(t_0) = 1\} = \int_0^{t_0} P(z|t_0)b(z|t_0)dz(4)$$

Note that someone with disease at time  $t_0$  having age  $z$  was born in the year  $v = t_0 - z$ . Hence the probability distribution of ages at time  $t_0$  is equivalent to the distribution of birth cohorts at time  $t_0$ .

### 3.1 Forward Recurrence Time Distribution

Define

$$\begin{aligned} T_f &= \text{Forward recurrence time random variable} \\ q_f(t|z)dt &= P\{t < T_f \leq t + dt \mid a(z|t_0) = 1\} \\ Q_f(t|z) &= P\{T_f > t \mid a(z|t_0) = 1\} \\ I(\tau)d\tau &= P\{S_0 \rightarrow S_a \text{ during } \tau, \tau + d\tau\} \end{aligned}$$

where  $\tau$  refers to the age of incidence. Consider the probability of being in  $S_a$  at time  $t_0$  and having age  $z$ . If an individual becomes incident at age  $\tau$ , then  $P\{a(z|t_0) = 1|\tau\} = P\{T > z - \tau\} = Q(z - \tau)$ . Multiplying by  $I(\tau)d\tau$  and integrating over the possible values of  $\tau$  ( $0 < \tau \leq z$ ) results in

$$P\{a(z|t_0) = 1\} = \int_0^z I(\tau)Q(z - \tau)d\tau \quad (5)$$

This probability applies to the birth cohort year  $v = t_0 - z$ ; i.e. an individual born in year  $v$  who is prevalent at time  $t_0$  having age  $z$ .

Consider the joint distribution of an individual having age  $z$  at time  $t_0$  and staying in  $S_a$  for at least an additional  $t$  time units. If  $\tau$  is the age of entering  $S_a$ , then

$$P(z|t_0, \tau)Q_f(t|z, \tau) = P\{T > z - \tau + t\} = Q(z - \tau + t)$$

and multiplying by  $I(\tau)d\tau$  and integrating over  $(0, z)$  gives

$$P(z|t_0)Q_f(t|z) = \int_0^z I(\tau)Q(z - \tau + t)d\tau \quad (6)$$

In the above it is assumed that the time entering  $S_a$  ( $\tau$ ) is not known, requiring integration over possible values of ( $\tau$ ). Consequently the p.d.f. of the forward recurrence time is

$$q_f(t|z) = -\frac{d}{dt}Q_f(t|z) = \int_0^z I(\tau)q(z - \tau + t)d\tau/P(z|t_0) \quad (7)$$

Suppose the incidence is constant,  $I(\tau) = I$  then

$$q_f(t|z) = [Q(t) - Q(t + z)] / \int_0^z Q(y)dy. \quad (8)$$

If  $Q(z)$  is negligible, then

$$q_f(t|z) \sim Q(t)/m$$

which is the usual forward recurrence time distribution for a stationary process.

Define  $q_f(t|t_0)$  as the forward recurrence time averaged over the population. By definition we can write

$$P(a(t_0) = 1)q_f(t|t_0) = \int_0^{t_0} P(z|t_0)q_f(b|z)b(z|t_0)dz \quad (9)$$

When the age distribution is uniform so that  $b(z|t_0) = b$  then it can be shown, cf. Zelen and Feinleib [ZF69]

$$\int_0^\infty q_f(t|t_0)P(a|t_0) = 1)dt_0 / \int_0^\infty P(a(t_0) = 1)dt_0 = Q(t)/m.$$

Thus if the sampling point is regarded as a random point in time, the forward recurrence time distribution as  $t_0 \rightarrow \infty$  is the same as the stationary forward recurrence time distribution.

### 3.2 Backward Recurrence Time Distribution

The backward recurrence time refers to the time in  $S_a$  up to time  $t_0$  (or age  $z$ ). Let  $T_b$  be the backward recurrence time random variable and  $q_b(t|z)$  be the conditional p.d.f. with  $Q_b(t|z) = \int_t^z q_b(y|z)dy$ . Note that  $0 < t \leq z$ . Then using the same reasoning as in deriving the forward recurrence time distribution we have

$$P\{T_b > t, a(z|t_0) = 1\} = P(z|t_0)Q_b(t|z) = \int_0^{z-t} I(\tau)Q(z-\tau)d\tau \quad (10)$$

which allows the calculation of  $q_b(t|z)$ ; i.e.,

$$q_b(t|z) = I(z-t)Q(t)/P(z|t_0), \quad 0 < t \leq z \quad (11)$$

When  $I(\tau) = I$ ,  $q_b(t|z) = Q(t)/\int_0^z Q(y)dy$ .

Finally the average backward recurrence time distribution is

$$q_b(t|t_0) = Q(t) \int_t^{t_0} I(z-t)b(z|t_0)dz/P_0 \quad (12)$$

Note the distinction between  $q_b(t|z)$  and  $q_b(t|t_0)$ . The former refers to individuals having age  $z$  at time  $t_0$  whereas the latter refers to the weighted average over age for prevalent cases at time  $t_0$ . When  $b(z|t_0) = b$ , we can integrate over  $t_0$  and show that the backward recurrence time averaged over  $t_0$  is  $Q(t)/m$ .

### 3.3 Length Biased Sampling and the Survival of Prevalent Cases

As pointed out earlier, the prevalence cases are not a random sample of cases, but represent a length biased sample. In this section, we investigate the

consequences of length biased sampling when disease incidence is age-related. We also derive the survival of prevalent cases.

Define  $T = T_b + T_f$  which is the time in which prevalent cases are in  $S_a$ . This is the survival of prevalent cases from the time when they become incident with disease. We will derive  $f(t|z)$ , the *pdf* of the time in  $S_a$  for prevalent cases who have age  $z$  at chronological time  $t_0$ . Since the age  $z$  is fixed at time  $t_0$ , it is necessary to consider  $t > z$  and  $t \leq z$  separately. If  $t$  is fixed and  $t > z$ , then  $P\{a(z|t_0) = 1 \mid t > z\} = \int_0^z I(\tau)d\tau$ . Similarly, if  $t$  is fixed and  $t < z$ , in order to be prevalent at time  $t_0$  and be of age  $z$ , it is necessary that  $z - t < \tau < z$ . Thus, we have for fixed  $t$  ( $t < T \leq t + dt$ )

$$P\{a(z|t_0) = 1 \mid t < T \leq t + dt\} = \begin{cases} \int_0^z I(\tau)d\tau, & \text{if } t > z \\ \int_{z-t}^z I(\tau)d\tau, & \text{if } t \leq z \end{cases} \quad (13)$$

Note that  $\int_{z-t}^t I(\tau)d\tau$  is an increasing function of  $t$ . Consequently, individuals with long sojourn times in  $S_a$  have a greater probability of being in  $S_a$  at time  $t_0$ . Our development is a generalization of the usual considerations of length biased sampling as we have shown how length biased sampling is affected by the transition into  $S_a$ . The usual specification of length biased sampling is to assume  $P\{a(z) = 1 \mid t < T \leq t + dt\} \propto t$ , which in our case would be true if  $I(\tau) = I$  and  $t \leq z$ . We also remark that  $P\{a(z|t_0) = 0 \mid t < T \leq t + dt\} = \int_0^{z-t} I(\tau)d\tau$  refers to individuals, conditional on having survival  $t < T \leq t + dt$ , who entered  $S_a$  and died before time  $t_0$ , but would have been age  $z$  at time  $t_0$  if they had lived. Another interpretation of this probability is that a birth cohort born in  $v = z - t$  was incident with disease but died before reaching age  $z$ . Using (13) the joint distribution of  $a(z|t_0)$  and  $T$  is

$$P\{a(z|t_0) = 1, t < T \leq t + dt\} = \begin{cases} q(t)dt \int_0^z I(\tau)d\tau, & \text{if } t > z \\ q(t)dt \int_{z-t}^z I(\tau)d\tau, & \text{if } t \leq z. \end{cases} \quad (14)$$

Therefore, the time in  $S_a$  for cases prevalent at  $t_0$  and having age  $z$  is

$$f(t|z)dt = \frac{P\{a(z|t_0) = 1, t < T \leq t + dt\}}{P(z)}. \quad (15)$$

Some simplifications occur if  $I(\tau) = I$ . Then

$$f(t|z) = \begin{cases} zq(t)/\int_0^z Q(x)dx & \text{if } t > z \\ tq(t)/\int_0^z Q(x)dx & \text{if } t \leq z \end{cases} \quad (16)$$

If  $q(t)$  is negligible in the neighborhood of  $z$ , and  $t \leq z$ , then  $f(t|z) \simeq tq(t)/m$  which is the usual distribution for the sum of the forward and backward recurrence time random variables.

Using the same development, we can calculate  $f(t|a(z|t_0) = 0)$  which refers to the survival of individuals who died before  $t_0$ , but would have been age  $z$  at time  $t_0$ . Since

$$P\{a(z|t_0) = 0, t < T \leq t + dt\} = \left[ \int_0^{z-t} I(\tau) d\tau \right] q(t) dt, t \leq z$$

and

$$P(a(z|t_0) = 0) = \int_0^z \left[ \int_0^{z-t} I(\tau) d\tau \right] q(t) dt$$

we have

$$f(t|a(z|t_0) = 0) = \frac{\left[ \int_0^{z-t} I(\tau) d\tau \right] q(t)}{P(a(z|t_0) = 0)} \text{ if } t \leq z \quad (17)$$

which is the distribution of those who died before time  $t_0$ , but would have been age  $z$  at  $t_0$  if they had lived. If  $I(z) = I$ , the distribution is

$$f(t|a(z|t_0) = 0) = \frac{(1 - \frac{t}{z})q(t)}{\int_0^z (1 - \frac{t}{z})q(t) dt} \text{ for } t \leq z. \quad (18)$$

Note that if  $z \rightarrow \infty$ , then

$$f(t|a(z|t_0) = 0) = q(t)$$

which is the population survival pdf.

### 3.4 Chronological Time Modeling

Suppose that the incidence is a function of chronological time rather than age. Also, in some cases,  $t_0$  may be regarded as far removed from the origin as the disease process has been going on a long time. Then the equations for the forward and backward times may be modified by replacing  $z$  by  $t_0$ . Therefore, we have

$$\begin{aligned} q_f(t|t_0) &= \int_0^{t_0} I(\tau) q(t_0 - \tau + t) dt / P(t_0) \\ q_b(t|t_0) &= I(t_0 - t) Q(t) / P(t_0) \\ f(t|t_0) &= \begin{cases} q(t) \int_0^{t_0} I(\tau) d\tau & \text{if } t > t_0 \\ q(t) \int_{t_0-t}^{t_0} I(\tau) d\tau & \text{if } t \leq t_0 \end{cases} \\ f(t|a(t_0) = 0) &= q(t) \int_0^{t_0-t} I(\tau) d\tau / P\{a(t_0) = 0\} \text{ for } t \leq t_0 \end{aligned} \quad (19)$$

with  $P(t_0) = P\{a(t_0) = 1\} = \int_0^{t_0} I(\tau) Q(t_0 - \tau) d\tau$ .

If  $I(\tau) = I$  then