

Springer Series in Statistics

Advisors:

P. Bickel, P. Diggle, S. Fienberg,

U. Gather, I. Olkin, S. Zeger

Juha M. Alho and Bruce D. Spencer

Statistical Demography and Forecasting

With 33 Illustrations

 Springer

Juha Alho
Department of Statistics
University of Joensuu
Joensuu, Finland

Bruce Spencer
Department of Statistics
Northwestern University
Evanston, IL 60208
USA

Library of Congress Control Number: 2005926699 (hard cover)
Library of Congress Control Number: 2005927649 (soft cover)

ISBN 10: 0-387-23530-2 (hard cover)
ISBN 13: 978-0387-23530-1 (hard cover)
ISBN 10: 0-387-22538-2 (soft cover)
ISBN 13: 978-0387-22538-8 (soft cover)

Printed on acid-free paper.

© 2005 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. (TB/MVY)

9 8 7 6 5 4 3 2 1 SPIN 11011019 (hard cover) SPIN 11013662 (soft cover)

springeronline.com

To Irja and Donna

Preface

Statistics and demography share important common roots, yet as academic disciplines they have grown apart. Even a casual survey of leading journals shows that cross-references are rare. This is unfortunate, because many social problems call for a multi-disciplinary approach. Both statistics and demography are necessary ingredients in any serious analysis of the sustainability of pension or health care systems in the aging societies, in the assessment of potential inequities of formula-based allocations to local governments, in the estimation of the size of elusive populations such as drug users, in the investigation of the consequences of social ills such as unemployment, and so forth. This book was written to bring together much of the basic statistical theory and methodology for estimating and forecasting population growth and its components of births, deaths, and migration. Although relatively simple mathematical methods have traditionally been used to assess demographic trends and their role in the society, use of modern statistical methods offers significant advantages for more accurately measuring population and vital rates, for forecasting the future, and for assessing the uncertainty of the demographic estimates and forecasts.

For statisticians the book provides a unique introduction to demographic problems in a familiar language. For demographers, actuaries, epidemiologists, and professionals in related fields the book presents a unified statistical outlook on both classical methods of demography and recent developments. The book provides a self-contained introduction to the statistical theory of demographic rates (births, deaths, migration) in a multi-state setting. The book has a dual character. On the one hand, it is a monograph that can be consumed by a lone reader. There are many results that have appeared in journals or working papers only. Some appear here for the first time. The book is also useful as a classroom text, and includes exercises and complements to explore special topics in detail without interrupting the flow of the text. More than half of the book is readily accessible to undergraduates, but to fully benefit from the complete text may require more maturity.

Joensuu, Finland
Evanston, Illinois, USA

Juha M. Alho
Bruce D. Spencer

Acknowledgments

This book was some 15 years in the making. We are grateful to many colleagues and students for advice, encouragement and helpful comments, both specific and general. We thank Bill Bell, Katie Bench, Henry Bienen, Petra Can, Tom Espenshade, Steve Fienberg, Marty Frankel, Olavi Haimi, Joan Hill, Jan Hoem, Jeff Jenkins, Jay Kadane, Anne Kearney, Nico Keilman, Nathan Keyfitz, Donna Kostanich, Bill Kruskal, Esa Läärä, Jukka Lassila, Ron Lee, Risto Lehtonen, Chijien Lin, Lincoln Moses, Fred Mosteller, Tom Mule, Jukka Nyblom, Erkki Pahkinen, Päivi Partanen, Rita Petroni, Jiahe Qian, Dave Raglin, Chris Rhoads, Gregg Robinson, Mikko A. Salo, the late I. Richard Savage, Eric Schindler, Tom Severini, Eric Song, Richard Suzman, Shripad Tuljapurkar, Tarmo Valkonen, Jim Vaupel, Nic van de Walle, Larry Wu, Sandy Zabell. Shelby Haberman and Mary Mulry went above and beyond the call in close reading and advice. Responsibility for remaining errors, of course, remains with the authors.

During preparation of the book we received financial support from U.S. National Institute on Aging grant R01 AG10156-01A1 to Northwestern University; The Searle Fund grant on Limits of Empirical Social Science for Policy Analysis, to Northwestern University; U.S. Census Bureau contract 50-YABC-7-66020 with Abt, Associates; Academy of Finland Grants 8684, 41495, and 201408, Statistics Finland Grant 5012, and European Commission Grant HPSE-CT-2001-00095 to University of Joensuu; and European Commission Grant QLRT-2001-02500 to the Research Institute of the Finnish Economy.

Joensuu, Finland
Evanston, Illinois, USA

Juha M. Alho
Bruce D. Spencer

Contents

Preface	vii
Acknowledgments	ix
List of Examples	xix
List of Figures	xxv
Chapter 1. Introduction	1
1. Role of Statistical Demography	1
2. Guide for the Reader	4
3. Statistical Notation and Preliminaries	4
Chapter 2. Sources of Demographic Data	9
1. Populations: Open and Closed	9
2. <i>De Facto</i> and <i>De Jure</i> Populations	11
3. Censuses and Population Registers	15
4. Lexis Diagram and Classification of Events	16
5. Register Data and Epidemiologic Studies	19
5.1. Event Histories from Registers	19
5.2. Cohort and Case-Control Studies	19
5.3. Advantages and Disadvantages	20
5.4. Confounding	22
6. Sampling in Censuses and Dual System Estimation	24
Exercises and Complements	27
Chapter 3. Sampling Designs and Inference	31
1. Simple Random Sampling	32
2. Subgroups and Ratios	35
3. Stratified Sampling	36
3.1. Introduction	36
3.2. Stratified Simple Random Sampling	37
3.3. Design Effect for Stratified Simple Random Sampling	38
3.4. Poststratification	39
4. Sampling Weights	40
4.1. Why Weight?	40

4.2. Forming Weights	41
4.3. Non-Response Adjustments	43
4.4. Effect of Weighting on Precision	45
5. Cluster Sampling	46
5.1. Introduction	46
5.2. Single Stage Sampling with Replacement	47
5.3. Single Stage Sampling without Replacement	47
5.4. Multi-Stage Sampling	49
5.5. Stratified Samples	50
6. Systematic Sampling	52
7. Distribution Theory for Sampling	53
7.1. Central Limit Theorems	53
7.2. The Delta Method	55
7.3. Estimating Equations	56
8. Replication Estimates of Variance	61
8.1. Jackknife Estimates	61
8.2. Bootstrap Estimates	62
8.3. Replication Weights	63
Exercises and Complements	64
Chapter 4. Waiting Times and Their Statistical Estimation	71
1. Exponential Distribution	71
2. General Waiting Time	76
2.1. Hazards and Survival Probabilities	76
2.2. Life Expectancies and Stable Populations	79
2.2.1. Life Expectancy	79
2.2.2. Life Table Populations and Stable Populations	81
2.2.3. Changing Mortality	82
2.2.4. Basics of Pension Funding	84
2.2.5. Effect of Heterogeneity	85
2.3. Kaplan-Meier and Nelson-Aalen Estimators	85
2.4. Estimation Based on Occurrence-Exposure Rates	88
3. Estimating Survival Proportions	91
4. Childbearing as a Repeatable Event	93
4.1. Poisson Process Model of Childbearing	93
4.2. Summary Measures of Fertility and Reproduction	96
4.3. Period and Cohort Fertility	101
4.3.1. Cohort Fertility is Smoother	101
4.3.2. Adjusting for Timing	103
4.3.3. Effect of Parity on Pure Period Measures	104
4.4. Multiple Births and Effect of Pregnancy on Exposure Time	106
5. Poisson Character of Demographic Events	107
6. Simulation of Waiting Times and Counts	109
Exercises and Complements	110

Chapter 5. Regression Models for Counts and Survival	117
1. Generalized Linear Models	118
1.1. Exponential Family	118
1.2. Use of Explanatory Variables	119
1.3. Maximum Likelihood Estimation	119
1.4. Numerical Solution	120
1.5. Inferences	121
1.6. Diagnostic Checks	122
2. Binary Regression	123
2.1. Interpretation of Parameters and Goodness of Fit	123
2.2. Examples of Logistic Regression	124
2.3. Applicability in Case-Control Studies	129
3. Poisson Regression	130
3.1. Interpretation of Parameters	130
3.2. Examples of Poisson Regression	131
3.3. Standardization	133
3.4. Loglinear Models for Capture-Recapture Data	136
4. Overdispersion and Random Effects	138
4.1. Direct Estimation of Overdispersion	139
4.2. Marginal Models for Overdispersion	139
4.3. Random Effect Models	140
5. Observable Heterogeneity in Capture-Recapture Studies	143
6. Bilinear Models	146
7. Proportional Hazards Models for Survival	150
8. Heterogeneity and Selection by Survival	154
9. Estimation of Population Density	156
10. Simulation of the Regression Models	158
Exercises and Complements	159
Chapter 6. Multistate Models and Cohort-Component Book-Keeping	166
1. Multistate Life-Tables	167
1.1. Numerical Solution Using Runge-Kutta Algorithm	167
1.2. Extension to Multistate Case	168
1.3. Duration-Dependent Life-Tables	172
1.3.1. Heterogeneity Attributable to Duration	172
1.3.2. Forms of Duration-Dependence	173
1.3.3. Aspects of Computer Implementation	174
1.3.4. Policy Significance of Duration-Dependence	175
1.4. Nonparametric Intensity Estimation	175
1.5. Analysis of Nuptiality	177
1.6. A Model for Disability Insurance	179
2. Linear Growth Model	180
2.1. Matrix Formulation	180

2.2. Stable Populations	183
2.3. Weak Ergodicity	185
3. Open Populations and Parametrization of Migration	186
3.1. Open Population Systems	186
3.2. Parametric Models	186
3.2.1. Migrant Pool Model	187
3.2.2. Bilinear Models	187
4. Demographic Functionals	189
5. Elementwise Aspects of the Matrix Formulation	191
6. Markov Chain Models	191
Exercises and Complements	193
 Chapter 7. Approaches to Forecasting Demographic Rates	 198
1. Trends, Random Walks, and Volatility	198
2. Linear Stationary Processes	201
2.1. Properties and Modeling	202
2.1.1. Definition and Basic Properties	202
2.1.2. ARIMA Models	203
2.1.3. Practical Modeling	206
2.2. Characterization of Predictions and Prediction Errors	210
2.2.1. Stationary Processes	210
2.2.2. Integrated Processes	211
2.2.3. Cross-Correlations	216
3. Handling of Nonconstant Mean	216
3.1. Differencing	216
3.2. Regression	218
3.3. Structural Models	219
4. Heteroscedastic Innovations	220
4.1. Deterministic Models of Volatility	221
4.2. Stochastic Volatility	222
Exercises and Complements	223
 Chapter 8. Uncertainty in Demographic Forecasts: Concepts, Issues, and Evidence	 226
1. Historical Aspects of Cohort-Component Forecasting	228
1.1. Adoption of the Cohort-Component Approach	228
1.2. Whelpton's Legacy	228
1.3. Do We Know Better Now?	231
2. Dimensionality Reduction for Mortality	234
2.1. Age-Specific Mortality	234
2.2. Cause-Specific Mortality	236
3. Conceptual Aspects of Error Analysis	238
3.1. Expected Error and Empirical Error	238
3.2. Decomposing Errors	238
3.2.1. Error Classifications	238
3.2.2. Alternative Decompositions	240

3.3. Acknowledging Model Error	240
3.3.1. Classes of Parametric Models	240
3.3.2. Data Period Bias	241
3.4. Feedback Effects of Forecasts	242
3.5. Interpretation of Prediction Intervals	244
3.5.1. Uncertainty in Terms of Subjective Probabilities	244
3.5.2. Frequency Properties of Prediction Intervals	248
3.6. Role of Judgment	249
3.6.1. Expert Arguments	249
3.6.2. Scenarios	250
3.6.3. Conditional Forecasts	251
4. Practical Error Assessment	251
4.1. Error Measures	252
4.2. Baseline Forecasts	253
4.3. Modeling Errors in World Forecasts	256
4.3.1. An Error Model for Growth Rates	256
4.3.2. Second Moments	257
4.3.3. Predictive Distributions for Countries and the World	259
4.4. Random Jump-off Values	261
4.4.1. Jump-off Population	262
4.4.2. Mortality	263
5. Measuring Correlatedness	264
Exercises and Complements	267
Chapter 9. Statistical Propagation of Error in Forecasting	269
1. Törnqvist's Contribution	269
2. Predictive Distributions	271
2.1. Regression with a Known Covariance Structure	271
2.2. Random Walks	274
2.3. ARIMA(1,1,0) Models	276
3. Forecast as a Database and Its Uses	277
4. Parametrizations of Covariance Structure	278
4.1. Effect of Correlations on the Variance of a Sum	279
4.2. Scaled Model for Error	280
4.3. Structure of Error in Migration Forecasts	283
5. Analytical Propagation of Error	284
5.1. Births	284
5.2. General Linear Growth	285
6. Simulation Approach and Computer Implementation	287
7. Post Processing	289
7.1. Altering a Distributional Form	289
7.2. Creating Correlated Populations	292
7.2.1. Use of Seeds	292
7.2.2. Sorting Techniques	293
Exercises and Complements	294

Chapter 10. Errors in Census Numbers	296
1. Introduction	296
2. Effects of Errors on Estimates and Forecasts	297
2.1. Effects on Mortality Rates	297
2.2. Effects on Forecasts	298
2.3. Effects on Evaluation of Past Population Forecasts	298
3. Use of Demographic Analysis to Assess Error in U.S. Censuses	299
4. Assessment of Dual System Estimates of Population Size	300
5. Decomposition of Error in the Dual System Estimator	303
5.1. A Probability Model for the Census	303
5.2. Poststratification	304
5.3. Overview of Error Components	305
5.4. Data Error Bias	308
5.5. Decomposition of Model Bias	309
5.5.1. Synthetic Estimation Bias and Correlation Bias	309
5.5.2. Poststratified Estimator	310
5.6. Estimation of Correlation Bias in a Poststratified Dual System Estimator	312
5.7. Estimation of Synthetic Estimation Bias in a Poststratified Dual System Estimator	314
6. Assessment of Error in Functions of Dual System Estimators and Functions of Census Counts	316
6.1. Overview	316
6.2. Computation	317
Exercises and Complements	319
Chapter 11. Financial Applications	327
1. Predictive Distribution of Adjustment for Life Expectancy Change	327
1.1. Adjustment Factor for Mortality Change	327
1.2. Sampling Variation in Pension Adjustment Factors	329
1.3. The Predictive Distribution of the Pension Adjustment Factor	330
2. Fertility Dependent Pension Benefits	332
3. Measuring Sustainability	335
4. State Aid to Municipalities	337
5. Public Liabilities	339
5.1. Economic Series	340
5.2. Wealth in Terms of Random Returns and Discounting	340
5.3. Random Public Liability	341
Exercises and Complements	342
Chapter 12. Decision Analysis and Small Area Estimates	344
1. Introduction	344
2. Small Area Analysis	345

3.	Formula-Based Allocations	346
3.1.	Theoretical Construction	346
3.1.1.	Apportionment of the U.S. House of Representatives	347
3.1.2.	Rationale Behind Allocation Formulas	348
3.2.	Effect of Inaccurate Demographic Statistics	349
3.3.	Beyond Accuracy	350
4.	Decision Theory and Loss Functions	351
4.1.	Introduction	351
4.2.	Decision Theory for Statistical Agencies	353
4.3.	Loss Functions for Small Area Estimates	357
4.4.	Loss Functions for Apportionment and Redistricting	359
4.1.1.	Apportionment	359
4.1.2.	Redistricting	360
4.5.	Loss Functions and Allocation of Funds	361
4.5.1.	Effects of Over- and Under-Allocation	361
4.5.2.	Formula Nonoptimality	362
4.5.3.	Optimal Data Quality with Multiple Statistics and Uses	363
5.	Comparing Risks of Adjusted and Unadjusted Census Estimates	363
5.1.	Accounting for Variances of Bias Estimates	364
5.2.	Effect of Unmeasured Biases on Comparisons of Accuracy	365
6.	Decision Analysis of Adjustment for Census Undercount	365
7.	Cost-Benefit Analysis of Demographic Data Exercises and Complements	367
		368
	References	371
	Author Index	397
	Subject Index	405

List of Examples

Chapter 2. Sources of Demographic Data	9
1.1. Who Counts in the U.S. Census?	10
1.2. Who Belongs to the Sami Population?	10
2.1. Accident Rates in Nordic Countries.	12
2.2. Undercount in U.S. Censuses.	12
2.3. What Is a Household?	14
2.4. Corporate Demography.	14
3.1. Nigerian Censuses.	15
5.1. British Doctors' Study.	20
5.2. Doll and Hill Study.	21
6.1. Underreporting of Occupational Diseases.	26
6.2. Numbers of Drug Users.	26
Chapter 3. Sampling Designs and Inference	31
1.1. The 1970 Draft Lottery in the U.S.	33
1.2. Child Stunting.	34
3.1. NELS:88 Base-Year School Sample.	37
3.2. Design Effects for NELS:88.	39
3.3. Poststratification in the 1990 U.S. Post Enumeration Survey (PES).	40
4.1. NELS:88 First Followup Schools.	41
4.2. Extreme Weights in the 1990 U.S. PES.	42
4.3. Nonparticipation in a Survey in an STD Clinic.	43
4.4. The Dual System Estimator as a Propensity-Weighted Census.	44
4.5. Extreme Weights in the Survey of Consumer Finance.	46
5.1. Survey of the Homeless in Chicago.	48
5.2. NELS:88 Sample of Students.	51
5.3. The U.S. Current Population Survey.	51
6.1. Systematic Sampling of Private Schools in the National Assessment of Educational Progress.	53
7.1. Model-Based Variance of the Dual System Estimator (DSE).	56
7.2. Design-Based Variance of the Dual System Estimator (DSE).	58

7.3.	Parameter Interpretation Under An Erroneous Model.	59
7.4.	Fieller Intervals for a Ratio Estimator.	60
Chapter 4. Waiting Times and Their Statistical Estimation		71
1.1.	Memorylessness of Exponential Waiting Time.	72
1.2.	Independent Causes of Death.	72
1.3.	Cross-Sectional Heterogeneity of Constant Hazard Rates.	75
1.4.	Gamma Distribution for Frailty.	75
2.1.	Weibull Distribution.	77
2.2.	Linear Survival Functions.	77
2.3.	Balducci Model for Survival Function.	77
2.4.	Competing Risks.	77
2.5.	Mortality and Marital Status in Finland.	78
2.6.	Effect of Changes in Hazards on Life Expectancy.	83
2.7.	Life Expectancy Calculation from Kaplan-Meier Estimates.	86
2.8.	Survival Probabilities for Habsburgs.	86
2.9.	Actuarial Estimator.	89
2.10.	Distribution of Death During First Year.	90
2.11.	Proportion of Deaths During First Days.	90
4.1.	Age-Specific Fertility Rates for Italy and the U.S.	95
4.2.	Finnish Fertility, 1776–1999.	97
4.3.	Time Trends in Sex Ratios in Finland.	98
4.4.	Alternative Measures of Mean Age at Childbearing, Finland 2000.	100
4.5.	Parity Progression Ratios.	105
6.1.	Simulation of Weibull Random Variates.	109
Chapter 5. Regression Models for Counts and Survival		117
1.1.	Exponential Distribution.	118
1.2.	Bernoulli Distribution.	118
1.3.	Leverage in Simple Generalized Linear Model.	122
2.1.	Sex Ratios of the Habsburgs.	124
2.2.	Child Mortality among the Habsburgs.	125
2.3.	Testing Effects of Exposure on Illness.	125
2.4.	Detecting Confounding.	127
2.5.	Choosing the Sword.	127
3.1.	Poisson Models for Births.	131
3.2.	Mortality of Young Widows.	132
3.3.	Age-Period-Cohort Problem.	132
3.4.	Number of the Habsburg Offspring.	132
3.5.	Regression Models for Rates of Small Areas.	132
3.6.	Relative Risk of Mortality for Unemployed.	135
3.7.	Triple Systems Estimates of Numbers of Drug Users.	138
4.1.	Overdispersion in Habsburg Cohort Sizes.	142
5.1.	Heterogeneity in Reporting of Occupational Disease.	145
5.2.	Heterogeneity in Census Enumeration Probabilities.	145

6.1. Lee-Carter Model for Mortality.	147
6.2. Mortality among Elderly.	149
7.1. A Simple Example of Cox Regression.	150
7.2. A Simple Example of Cox Regression with Censoring.	151
7.3. Changes in Mortality of the Habsburgs.	153
7.4. Time-Varying Covariates.	154
7.5. Likelihood for Matched Studies.	154
Chapter 6. Multistate Models and Cohort-Component Book-Keeping	166
1.1. Runge-Kutta Illustration.	167
1.2. A Three-State Labor Force Model.	169
1.3. Hazards Producing a Linear Solution.	170
1.4. Remarriage Probability Varies with Time Spent Non-married.	172
2.1. Two-Sex Problem.	183
4.1. Marriage Prevalence as a Functional.	190
4.2. Life Expectancy as a Functional.	190
4.3. Age Dependency Ratio.	190
4.4. A Relation between Prevalence and Incidence.	190
6.1. Metapopulation of Butterflies.	192
Chapter 7. Approaches to Forecasting Demographic Rates	198
1.1. Cohort Fertility Is Smoother.	199
1.2. Cholesky Decomposition.	201
2.1. MA(q) Processes.	203
2.2. AR(1) Processes.	203
2.3. EWMA Processes.	205
2.4. Vital Processes Appear Nonstationary.	207
2.5. Standard Error Under AR(1) Residuals.	211
2.6. Correlations of Forecast Errors For AR(1) Processes.	211
2.7. Correlations of Forecast Errors for Integrated AR(1) Processes.	212
2.8. Standard Error and Random Error.	212
3.1. Forecasting a Random Walk with a Drift.	217
3.2. Trend in Finnish Fertility up to 1930.	217
3.3. Alternative Time Series Forecasts of the U.S. Growth Rate.	219
3.4. Stochastic Local Level Process.	220
3.5. Stochastic Linear Trend Process.	220
4.1. A Heteroscedastic Process with Time Invariant Autocorrelations.	222
Chapter 8. Uncertainty in Demographic Forecasts: Concepts, Issues, and Evidence	226
1.1. Cohort Approach to Fertility Forecasting.	231
1.2. Effect of Marriage Duration on Fertility.	232
1.3. Was the Baby-Boom a Unique Phenomenon?	232

1.4.	Trend Extrapolation Versus Judgment.	232
1.5.	Counterintuitive Data on Economic Shocks and Demographics.	233
2.1.	Rates of Mortality Decline in Europe.	235
2.2.	Emerging Cause of Death.	237
3.1.	Sensitivity to Assumptions.	239
3.2.	Planning Optimism.	244
3.3.	Achieving Approximate Consensus on Probabilities.	246
3.4.	Elicitation of Probabilities via Betting.	247
3.5.	Assessing Prediction Intervals for ARIMA Forecasts.	248
3.6.	Mortality Differences Across Countries.	250
3.7.	Fertility in the Mediterranean Countries.	250
3.8.	Migration to Germany.	250
4.1.	Error Estimates for Fertility Forecasts in Europe.	254
4.2.	Error Estimates for Mortality Forecasts in Europe.	255
5.1.	Constant Correlations Across Ages.	265
5.2.	Constant Correlations Across Causes of Death.	265
5.3.	Uncorrelated Errors for Different Vital Rates.	265
5.4.	Constant Correlations Across Countries within a Region.	266
Chapter 9. Statistical Propagation of Error in Forecasting		269
2.1.	Posterior of an AR(1) Process With Known Autocorrelations.	274
2.2.	Conditional Likelihood of an AR(1) Process.	274
2.3.	Predictive Distribution of a Random Walk.	275
2.4.	Predictive Distribution of a Random Walk With a Drift.	275
4.1.	Independence, AR(1), and Perfect Dependence.	279
4.2.	Error in a Cohort Survival Setting.	279
4.3.	Autoregressive Model for Correlations Across Age.	281
4.4.	Specifying a Linear Process to Match Judgment.	282
5.1.	Representation of a Closed Female Population.	285
6.1.	Storage Space Required by the Database.	288
7.1.	Stochastic Forecast Database for Finland.	290
Chapter 10. Errors in Census Numbers		296
4.1.	Post Enumeration Surveys in the 1990 and 2000 U.S. Censuses.	300
4.2.	Post Enumeration Survey in the U.K. in 2001.	302
5.1.	Artificial Example of Probability Model for a Census.	304
5.2.	Error Components in the 1990 U.S. PES.	307
5.3.	Error Components in the 2000 U.S. A.C.E.	307
5.4.	Estimates of Correlation Bias Based on DA Totals.	312
5.5.	Estimates of Correlation Bias Based on DA Sex Ratios.	313
5.6.	Surrogate Variables for Undercount and Overcount in the 2000 U.S. Census.	315

Chapter 12. Decision Analysis and Small Area Estimates	344
4.1. Asymmetric Consequences of Forecast Error.	351
4.2. Posterior Risk Under Linear Loss.	353
4.3. When Policy Makers Prefer Error to Accuracy.	354
4.4. Non-Adjustment of Undercount Estimates for Correlation Bias.	356
4.5. Adjustment for Correlation Bias for Hispanics in the 2000 U.S. Census.	356
4.6. Alternative Estimates of Population.	357
4.7. Value Judgements in Sample Allocation.	358
4.8. Expected Loss of Adjusted and Unadjusted 2000 U.S. Census for Redistricting.	360
6.1. Expected Loss of Adjusted and Unadjusted 1990 U.S. Census.	365
6.2. Expected Loss of Adjusted and Unadjusted 2000 U.S. Census, A.C.E. Revision II.	366
7.1. Decennial Census.	367
7.2. Mid-Decade Census.	368

List of Figures

Chapter 2. Sources of Demographic Data	
1. Lexis Diagram.	17
2. Example of Confounding.	24
Chapter 4. Waiting Times and Their Statistical Estimation	
1. Log of Mortality Hazard for the Married, Widowed, and Single and Divorced Women in Finland, in 1998.	78
2. Log of the Hazard Increment of Mortality in Finland in 1881–1890 and 1986–1990, for Females and Males.	82
3. Survival Probabilities for Females and Males among the Members of the Main Line of the Family of Habsburgs.	87
4. The Distribution of Life Times of Those Born in 1994, Who Died in Age Zero, in Finland.	90
5. Total Fertility Rate in Finland in 1776–1999 and in the United States in 1920–1999.	97
6. Sex Ratio at Birth (Actual and Smoothed) in Finland in 1751–2000.	98
7. Approximate Completed Fertility for Birth Cohorts Born in Finland in 1905–1965.	102
Chapter 6. Multistate Models and Cohort-Component Book-Keeping	
1. Average Relative Risk of Remarriage Among Widowed and Divorced as a Function of the Duration of Widowhood and Divorce, Respectively.	173
2. Possible State Transitions in Nuptiality Processes.	177
3. Relative Risk of Death Among Married as a Function of the Duration of Marriage: Average, in Age 30, in Age 40, and in Age 50.	178
4. Distribution of Time Spent in the Divorced State, if Ever Divorced, for a Single at Age 17.	179

5. Average Density of Male Migration in Finland, Across Three Regions, During 1987–1997.	188
6. Two Most Important Patterns of Deviation from Average Age Distribution of Migration Intensity.	188
7. Coefficients of Deviations from the Mean for the Six Flows, During 1987–1997.	189
Chapter 7. Approaches to Forecasting Demographic Rates	
1. Hypothetical Cohort and Period Fertility Under a Pure Period Random Walk Model.	199
2. Hypothetical Mortality Rates and a Moving Average Estimate of their Level.	204
3. The Growth Rate of the U.S. Population in 1900–1999, and Three Forecasts: AR(1) and ARIMA(2,1,0) with and without a Constant Term.	208
4. Total Fertility Rate of Finland in 1920–1996, and its Forecast for 1997–2021 with 50% Prediction Intervals.	214
5. (A) Lag-Plot of the First Differences $Y(t)$ at Lag 1. (B) Lag-Plot of the First Differences $Y(t)$ at Lag 2.	215
6. Absolute First Differences of the U.S. Growth Rate in 1900–1999, and an Exponentially Smoothed Trend Estimate.	221
Chapter 8. Uncertainty in Demographic Forecasts: Concepts, Issues, and Evidence	
1. Smoothed Rate of Decline in Age-Specific Mortality for Females and Males and its Median Across 11 European Countries, for Females, and for Males.	235
2. Distribution of Absolute Errors of Decline in Growth Rate.	243
3. Change in the Expected Value for the Probability of Heads in a Sequence of Coin Tossing Experiments for an Individual with a Prior Expectation of 0.9 and an Individual with a Prior Expectation of 0.1.	247
4. Median Relative Error of Fertility Forecast as a Function of Lead Time for Six Countries with Long Data Series, their Average, and a Random Walk Approximation.	254
5. Median Relative Error of Mortality Forecast as a Function of Lead Time for Nine Countries with Long Data Series, their Average, and a Random Walk Approximation.	256
Chapter 9. Statistical Propagation of Error in Forecasting	
1. Predictive Distribution of a Fertility Measure and its Modified Distribution.	291
Chapter 11. Financial Applications	
1. Predictive Distribution of the Adjustment Factor in 2010–2060: Median, First and Third Quartiles, and First and Ninth Deciles.	332

2. Predictive Distribution of Old-Age Dependency Ratio (Ages 60+/Ages 20–59) in Finland in 2010, 2030, and 2050. 334
3. Pension Contributions, as % of the Total Wages of the Covered Employees, in Finland in 1995–2070, Under Current Rules and Under a Fertility Dependent Rule, if the Population Follows the High Old-Age Dependency Ratio Variant. 335
4. Replacement Rate and Contribution Rate Under Full Wages Indexation and Full Wage-Bill Indexation, and an Example of Potential Viable Region $\{(c, r) | c \leq 0.38, r \geq 0.28\}$. 336
5. Relative Burden of Social and Health Care Allocations in 1940–1997 in Finland, and the Median, Quartiles, and First and Ninth Deciles of its Predictive Distribution in 1998–2050. 339

1

Introduction

1. Role of Statistical Demography

The world population exceeded six billion (6,000,000,000) in 1999. According to current United Nations projections, in 2050 the population is expected to be 9.3 billion, although under plausible scenarios it might be as low as 7.7 billion or as high as 10.9 billion. In all cases, the increase will intensify competition for arable land, clean water, and raw materials. Soil erosion and deforestation will continue in many parts of the world. The increased production of food, housing, and consumer goods will increase the production of greenhouse gases and, thus, contribute to climate change.

Underneath the global trends there is a great diversity. In the middle of the 19th century, European women gave birth to five children or more, on average. A newborn was expected to live 40 years or less. In a matter of a century the average number of children dropped to two and life expectancy rose to over 60 years. Many developing countries (notably China) have later followed a similar path, but a key factor in the uncertainty regarding global trends is whether all developing countries will go through a similar transition, and if so, at what pace.

Even within the industrialized world a great diversity persists. The average number of children per woman (as measured by the total fertility rate) varies from 1.2 children per woman in Italy and Spain, to 2.0 in the United States. The U.S. value is over 50% higher than that of the primarily catholic Mediterranean countries that have had a history of relatively high fertility! Yet, all values are below the level (approximately 2.1) that is needed for population replacement. Although births currently exceed deaths, this is a temporary phenomenon caused by an age-distribution that still has relatively many people in the child-bearing ages. In the near future the situation will change, and the age-distributions of the industrialized countries will be older than in any national population ever before on earth. This will put stress on the health care and retirement systems, a stress whose magnitude is not fully appreciated by decision makers, yet.

The “graying” of the industrialized populations will be accentuated by two factors. First, the large baby-boom cohorts born after World War II will be retiring in 2010–2020. This may prove to be a one time phenomenon, but no-one can say

for certain that fertility fluctuations would have come to an end. The second factor is the continuing increase in longevity. Forecasters have repeatedly assumed that the decline in mortality cannot continue for more than a decade or two, only to have been proved wrong by the subsequent development.

Interestingly, populations can be quite heterogeneous with respect to life expectancy, as well. Women live longer than men, the rich and the well-educated live longer than the poor and the less-educated, and those in marriage live longer than those divorced, for example. The elderly are in many ways disadvantaged in the current industrialized societies. A happier future may lay ahead, if only by selection: it is possible that we will see a well-educated, healthy and wealthy retired population that is capable of exercising political power for its own benefit.

Since the rate of population growth in the developing countries far exceeds that of the industrialized countries, the geographic distribution of the world population will change. For example, the combined population of Europe and North America is currently 17% of the world population, but since the combined population is not expected to change by 2050, its share is expected to drop to 11%. A key social policy issue is to what extent the declining trend is counterbalanced by immigration from the less developed regions. An influx of immigrants would probably be advantageous to the elderly, since the immigrants could keep the economies growing and the “pay-as-you-go” retirement systems solvent. However, those in working age may reasonably see immigrants as competing in the same labor market, so racism and xenophobia may also gain ground.

Apart from global issues, demographics has an important role in the day-to-day decision making of national and local governments. Ever since the biblical times demographic data have served as a basis of taxation, military conscription, apportionment of political representation, and allocation of funds. Systematic biases in data may cause inequities across ethnic domains or geographic regions. When small areas are considered, random variations may cause inequalities in treatment. Lack of timeliness is always a potential source of systematic bias, but the remedy of frequent adjustments adds an element of unpredictability in the planning by local units.

Relatively simple mathematical methods have traditionally been used to assess demographic trends and their role in the society. The methods have typically been based on the measurement of demographic rates by age and sex. Summary measures, such as total fertility rate and life expectancy can then be calculated. A substantive line of research tries to explain variation in the rates across social groups, regions, or time, in terms of sociological or economic concepts. Another, less ambitious line of research tries to elucidate the long-term implications of the current rates. Classical methods from matrix algebra and differential and integral equations are used in the latter.

Simple methods have served and, undoubtedly, will continue to serve demography well. However, there are three reasons for expanding a demographer’s toolkit into a statistical direction. First, as noted above, there is considerable interest in exploring variations in demographic rates in ever finer subpopulations. For example, if we find that young widows have an elevated risk of death but numbers

are small, how can we know that this is not due to chance? Or, if the duration of unemployment is associated with mortality, how can this be evaluated? Cross tabulations are a classical, but clumsy, way to study such issues. In epidemiology, cross tabulations have largely been replaced by statistical relative risk regression techniques. We believe the same will happen in demography. Apart from simply adding new techniques to a demographer's toolkit, a methodological consequence is that principles of statistical inference, in particular the assessment of estimation error, should become a standard part of demographic analysis.

Second, many of the issues mentioned above involve forecasting in one way or another. In econometrics, the standard way to handle forecasting problems is to use statistical time-series techniques. We believe demographers can also benefit from the time-series toolkit provided that it is judiciously applied, in a manner that respects the demographic context. Demographic forecasts can then be made using data driven techniques, in addition to the judgmental methods that are currently favored. A methodological consequence of the adaptation of such techniques is that forecast uncertainty can be handled probabilistically. For example, instead of merely saying that it is plausible that world population is between 7.7 and 10.9 billion in 2050, we may say that it is within such an interval with a specific probability. Empirical analyses based on the accuracy of earlier U.N. forecasts suggest that in this case the probability is roughly 95%.

Third, even though the quality of basic demographic data on population size is likely to continue to improve, more elusive populations have become of concern. For example, we need information on the spread of drug use to assess its cost to the society and to determine the success anti-drug policies. Direct enumeration is, clearly, out of the question. Or, we need estimates of populations by health status to anticipate future demands on institutional care and housing that are accessible to those physically impaired. Such populations present us with complex definitional challenges, and information concerning them must be derived via statistical techniques that may suffer both from biases and sampling error.

After these remarks we are reminded of two characterizations of the demographic profession. Jim Vaupel has defined a demographer as "someone who knows Lexis". Earlier Joel Cohen defined a demographer as "someone who forecasts population wrong", and a mathematical demographer as "someone who uses mathematics to forecast population wrong". Perhaps we could define a statistical demographer as "someone who knows Lexis, forecasts population wrong, but can at least quantify the uncertainty".

We have written this book with two types of readers in mind. First, we have thought of a mathematically oriented demographer, who is interested in learning the statistical outlook on the familiar problems. We have tried to define all relevant concepts in the book. However, the exposition is necessarily brief, so previous familiarity with basic mathematical statistics, regression analysis, and time-series analysis is probably necessary for a full understanding of many of the arguments. Second, we have thought of a statistician, who is interested in working with demographic problems. We have tried to present the central demographic concepts in the context of statistical models, and indicate conditions under which the classical

demographic procedures are optimal. Empirical examples are provided to give a flavor of what makes demography interesting. In addition to demographers and statisticians, we have thought of, for example, economists interested in pension and health care problems, epidemiologists interested in risk assessment, and actuaries and public health people interested in gerontology as potential readers of the book.

The application of statistical models in demography is not always straight forward, however. Along the way we try to indicate how a blind application of statistics can lead to unacceptable results. In fact, a central virtue of demographic teaching is a kind of “source criticism”, in which one examines, much like a historian does, the mechanisms that have produced the data being analyzed. The most fashionable statistical analysis is not worth much if it is applied to data that are not what they seem. The book points out such issues, so it may be of a more general methodological interest to statistical readers.

2. Guide for the Reader

The book was originally conceived as a monograph intended for a lone reader. There are many results that have appeared in journals or working papers only. Some appear here for the first time. Yet, we have included exercises and complements to permit the use of the book in classroom. Some of the technical material is useful for reference (e.g., formulas for estimators and variances), and may be skipped on a first reading. Guidance is provided throughout the book. Parts of the earlier versions of the book have been used at the Universities of Joensuu and Jyväskylä, Finland; Örebro University, Sweden; Max Planck Institute at Rostock, Germany; and Northwestern University, U.S.A., to teach advanced undergraduate and graduate students in statistics and demography. For a statistical audience, additional discussion of the demographic issues has often proved useful. For a demographic audience, we have spent more time on the basics of statistics.

At least three threads of thought can be distinguished within the book:

- * Chapters 2 and 4–6 provide an introduction to Statistical Demography; a shorter course that might be called Biometrics is obtained from Chapters 2 and 4;
- * Chapters 2–4, 10 and 12 provide an introduction to the Demographic Data Sources and their Quality;
- * Chapters 4, 6–9 and 11 provide an introduction to Demographic Forecasting; a shorter course concentrating on Demographics of Pensions and Public Finances is obtained from sections of Chapters 4, 8–9, and 11.

In each case, other chapters provide supporting material.

3. Statistical Notation and Preliminaries

The remainder of this chapter introduces some notation for random variables and their distributions emphasizing vector and matrix formulations. We also give a heuristic review of basic results from maximum likelihood estimation that we

assume as known in the sequel. Additional reminders/results will appear interspersed in the text, where needed. Some references for this material, at the same general mathematical level of the text, include Rice (1995), DeGroot (1987), Lindsey (1996), Azzalini (1996) and, at a more advanced mathematical level, Rao (1973), Severini (2000), Bickel and Doksum (2001), and Williams (2001).

The probability of an event A will be denoted by $P(A)$. If X is a *random variable* (i.e., a function whose value is determined by a random experiment), its *distribution function* or *cumulative distribution function (c.d.f.)* is $F(x) = P(X \leq x)$. The probability that X exactly equals x is $P(X = x) = F(x) - \lim_{h \searrow 0} F(x - h)$. Note that whenever $F(\cdot)$ is continuous this probability is zero. If $F(\cdot)$ is differentiable, then $F'(\cdot) = f(\cdot)$ is the *density function* of X .

Example 3.1. Normal (Gaussian) Distributions. The *standard normal distribution* $N(0, 1)$ has the expectation 0 and variance 1. Its density is $f(x) = (2\pi)^{-1/2} \exp(-x^2/2)$. Suppose X has this distribution, or $X \sim N(0, 1)$, then $Y = \mu + \sigma X$ has the normal (Gaussian) distribution $N(\mu, \sigma^2)$ with mean μ and variance σ^2 . The density of Y is $f(y) = (2\pi)^{-1/2} \sigma^{-1} \exp(-(y - \mu)^2/(2\sigma^2))$. \diamond

Example 3.2. Bernoulli Distribution. If X takes the value 1 with probability p and 0 with probability $1 - p$, then X has a Bernoulli distribution with parameter p , or $X \sim \text{Ber}(p)$. In this case $P(X = x) = p^x(1 - p)^{1-x}$, where $0 \leq p \leq 1$ and $x \in \{0, 1\}$. \diamond

In mathematical demography one typically considers $X \geq 0$ and it is often more convenient to work with *survival probabilities* $p(x) = P(X > x)$ than with c.d.f.'s. If $p(\cdot)$ is differentiable, then $f(x) = -p'(x)$.

The joint probability of events A_1, \dots, A_n is $P(A_1 \cap \dots \cap A_n)$, but we sometimes write $P(A_1, \dots, A_n)$ for short. The *conditional probability* of one event given another is defined as $P(A_1|A_2) = P(A_1 \cap A_2)/P(A_2)$, when $P(A_2) > 0$. If X_1, \dots, X_n are random variables, their *joint distribution function* is $F(x_1, x_2, \dots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n)$. Writing column vectors $\mathbf{x} = (x_1, \dots, x_n)^T$ and $\mathbf{X} = (X_1, \dots, X_n)^T$, with T denoting transpose, we may also write $F(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x})$ where the inequality holds for each component.

The *expectation* of X is denoted by $E[X]$. If X has density $f(\cdot)$, or if X takes discrete values x_1, x_2, \dots , then

$$E[X] = \int_{-\infty}^{\infty} xf(x) dx \quad \text{or} \quad E[X] = \sum_i x_i P(X_i = x_i), \quad (3.1)$$

respectively. If X and Y are random variables and a and b are scalars, then we have the linearity property $E[aX + bY] = aE[X] + bE[Y]$. The *variance* of X is defined as $\text{Var}(X) = E[(X - E[X])^2]$. It has the property $\text{Var}(a + bX) = b^2 \text{Var}(X)$.

The expectation of a random vector \mathbf{X} is defined componentwise, $E[\mathbf{X}] = (E[X_1], \dots, E[X_n])^T$. If \mathbf{a} is a vector and \mathbf{B} is a matrix such that $\mathbf{a} + \mathbf{B}\mathbf{X}$ is well-defined, then $E[\mathbf{a} + \mathbf{B}\mathbf{X}] = \mathbf{a} + \mathbf{B}E[\mathbf{X}]$. The *covariance* between X_1 and

X_2 is defined as $\text{Cov}(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])]$. The covariance matrix of $\mathbf{X} = (X_1, \dots, X_n)^T$ is an $n \times n$ matrix $\text{Cov}(\mathbf{X})$ whose (i, j) element is $\text{Cov}(X_i, X_j)$. Using vector notation we may write $\text{Cov}(\mathbf{X}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T]$. It has the property $\text{Cov}(\mathbf{a} + \mathbf{B}\mathbf{X}) = \mathbf{B}\text{Cov}(\mathbf{X})\mathbf{B}^T$.

The *conditional expectation* of X_1 given X_2 is denoted by $E[X_1|X_2]$. It has the linearity property of the usual expectation. It may be shown that, when the moments exist, $E[X_1] = E[E[X_1|X_2]]$. The *conditional variance* is $\text{Var}(X_1|X_2) = E[X_1^2|X_2] - E[X_1|X_2]^2$. It has the property, $\text{Var}(X_1) = E[\text{Var}(X_1|X_2)] + \text{Var}(E[X_1|X_2])$. Similarly, the *conditional covariance* is defined as $\text{Cov}(X_1, X_2|X_3) = E[X_1X_2|X_3] - E[X_1|X_3]E[X_2|X_3]$ and has the property $\text{Cov}(X_1, X_2) = E[\text{Cov}(X_1, X_2|X_3)] + \text{Cov}(E[X_1|X_3], E[X_2|X_3])$.

Example 3.3. Multivariate Normal Distribution. Suppose a $k \times 1$ vector \mathbf{X} has $E[\mathbf{X}] = \boldsymbol{\mu}$ and $\text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}$. It has a multivariate normal distribution, $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if $\mathbf{a}^T \mathbf{X} \sim N(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$ for any $k \times 1$ vector \mathbf{a} . If $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$, the identity matrix, then $\mathbf{X}^T \mathbf{X} \sim \chi^2$ distribution with $k \geq 1$ degrees of freedom. \diamond

The multivariate normal distribution is an example of a parametric family of distributions. Consider n independent observations X_i coming from densities $f_i(x_i; \boldsymbol{\theta})$, $i = 1, \dots, n$, where $\boldsymbol{\theta}$ is, say, a $k \times 1$ vector of parameters belonging to some set $\Theta \subset \mathbb{R}^k$. We do not assume here that the observations are necessarily identically distributed, because in regression applications of interest they typically are not. For example, in normal theory regression, if X_i would be the dependent variable and \mathbf{z}_i would be a vector of explanatory variables, we would have the density $f_i(x_i; \boldsymbol{\theta}) = (2\pi)^{-1/2} \sigma^{-1} \exp(-(x_i - \mathbf{z}_i^T \boldsymbol{\beta})^2 / (2\sigma^2))$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2)^T$.

When viewed as a function of $\boldsymbol{\theta}$ the probability of the observed data is called the *likelihood function*, $L(\boldsymbol{\theta}) = f_1(x_1; \boldsymbol{\theta}) \cdots f_n(x_n; \boldsymbol{\theta})$. The natural logarithm of the likelihood function is the *loglikelihood function* $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$. The principle of maximum likelihood means that we try to determine a value of $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$, or equivalently $\ell(\boldsymbol{\theta})$. The maximizing value (if one exists) is called a *maximum likelihood estimator (MLE)*. Define a $k \times 1$ vector of partial derivatives $\mathbf{S}_i(\boldsymbol{\theta}) = \partial/\partial \boldsymbol{\theta} \log(f_i(x_i; \boldsymbol{\theta}))$ for each $i = 1, \dots, n$. Their sum $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{S}_1(\boldsymbol{\theta}) + \cdots + \mathbf{S}_n(\boldsymbol{\theta})$ is called the *score* (e.g., Rao 1973, 367), and the MLE solves the system of k equations $\mathbf{S}(\boldsymbol{\theta}) = \mathbf{0}$.

Before the observations $X_i = x_i$ have been made, the score is a random variable, because its components are random: $\mathbf{S}_i(\boldsymbol{\theta}) = \partial/\partial \boldsymbol{\theta} \log(f_i(X_i; \boldsymbol{\theta}))$. Assuming that the order of differentiation and integration can be changed, we have that $E[\mathbf{S}_i(\boldsymbol{\theta})] = \partial/\partial \boldsymbol{\theta} \int f_i(x_i; \boldsymbol{\theta}) dx_i = \mathbf{0}$. The latter equality holds because the integral equals 1 for all $\boldsymbol{\theta}$. Therefore, the expectation of the score is $E[\mathbf{S}(\boldsymbol{\theta})] = \mathbf{0}$. Write $\text{Cov}(\mathbf{S}_i(\boldsymbol{\theta})) = \mathcal{I}_i(\boldsymbol{\theta})$, $i = 1, \dots, n$, and define $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}_1(\boldsymbol{\theta}) + \cdots + \mathcal{I}_n(\boldsymbol{\theta})$. It follows that $\text{Cov}(\mathbf{S}(\boldsymbol{\theta})) = \mathcal{I}(\boldsymbol{\theta})$, because the observations are independent. This is one form of the so-called *Fisher information* of the sample. Subject to regularity conditions on densities $f_i(x_i; \boldsymbol{\theta})$ (that may involve conditions on both the range of values of possible explanatory variables and on the tails of the density), none of components of the score $\mathbf{S}_i(\boldsymbol{\theta})$ take too large a share of the variance of the score,

so one can appeal to the central limit theorem to assert the asymptotic normality of the score. Therefore, we have that $\mathbf{S}(\boldsymbol{\theta}) \sim N(\mathbf{0}, \mathcal{I}(\boldsymbol{\theta}))$ asymptotically.

Example 3.4. Score tests. Consider a hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Under the null hypothesis, $\mathbf{a}^T \mathbf{S}(\boldsymbol{\theta}_0) \sim N(0, \mathbf{a}^T \mathcal{I}(\boldsymbol{\theta}_0) \mathbf{a})$ for any $k \times 1$ vector \mathbf{a} , so depending on the alternative hypothesis, a large number of the so-called *score tests* can be constructed. \diamond

Define a $k \times k$ matrix $\mathbf{H}_i(\boldsymbol{\theta}) = \partial^2 / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \log(f_i(X_i; \boldsymbol{\theta}))$, for each $i = 1, \dots, n$. I.e., this is a matrix whose (r, s) element is $\partial^2 / \partial \theta_r \partial \theta_s \log(f_i(X_i; \boldsymbol{\theta}))$. Their sum $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{H}_1(\boldsymbol{\theta}) + \dots + \mathbf{H}_n(\boldsymbol{\theta})$ is called the *Hessian*. By a direct calculation one can show that $E[\mathbf{H}_i(\boldsymbol{\theta})] = \partial^2 / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T \int f_i(x_i; \boldsymbol{\theta}) dx_i - E[\mathbf{S}_i(\boldsymbol{\theta}) \mathbf{S}_i(\boldsymbol{\theta})^T]$. As in the case of the score, the first term on the right hand side is zero. Using the result, $E[\mathbf{S}_i(\boldsymbol{\theta}) \mathbf{S}_i(\boldsymbol{\theta})^T] = \text{Cov}(\mathbf{S}_i(\boldsymbol{\theta})) = \mathcal{I}_i(\boldsymbol{\theta})$, we find an alternative expression for Fisher information, $-E[\mathbf{H}(\boldsymbol{\theta})] = \mathcal{I}(\boldsymbol{\theta})$.

Example 3.5. Fisher Information for Normal Distribution. Consider the normal distribution $N(\mu, \sigma^2)$. Let $\boldsymbol{\theta} = (\mu, \sigma^2)^T$. The Fisher information $\mathcal{I}(\boldsymbol{\theta})$ is given by the matrix

$$\begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/(2\sigma^4) \end{bmatrix}. \quad (3.2)$$

If instead we take $\boldsymbol{\theta} = (\mu, \sigma)^T$ then the lower diagonal entry of $\mathcal{I}(\boldsymbol{\theta})$ changes to $2/\sigma^2$. \diamond

Suppose $\hat{\boldsymbol{\theta}}$ is the MLE. By Taylor's theorem there is vector $\boldsymbol{\theta}'$ between the MLE and the true value $\boldsymbol{\theta}$ such that $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{S}(\boldsymbol{\theta}) + \mathbf{H}(\boldsymbol{\theta}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. We get from this that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = -\mathbf{H}(\boldsymbol{\theta}')^{-1} \mathbf{S}(\boldsymbol{\theta})$ provided that the inverse exists. Subject to regularity conditions $\mathbf{S}(\boldsymbol{\theta})/n \rightarrow \mathbf{0}$,¹ as $n \rightarrow \infty$, and $\mathbf{H}(\boldsymbol{\theta})/n$ has a limit $\mathbf{H}^*(\boldsymbol{\theta})$ that is a continuous function of $\boldsymbol{\theta}$ at least in the neighborhood of the true parameter value. In this case the MLE also converges to $\boldsymbol{\theta}$, so it is *consistent*. Being essentially a linear function of the score, the MLE inherits the multivariate normal distribution from the score and asymptotically $\text{Cov}(\hat{\boldsymbol{\theta}}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$. For practical inferential purposes we may assume, for large n , that $\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, -\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1})$. This leads to the so-called *Wald tests*.

There is yet a third type of test that naturally arises from the above theory. Consider a hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. Using a second order Taylor series development for $\ell(\boldsymbol{\theta})$ around $\hat{\boldsymbol{\theta}}$ and noting that $\mathbf{S}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$, we get that

$$2\ell(\hat{\boldsymbol{\theta}} - \ell(\boldsymbol{\theta}_0)) = -(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \mathbf{H}(\boldsymbol{\theta}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \quad (3.3)$$

where $\boldsymbol{\theta}'$ is a point between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. The asymptotic result given for the Wald tests shows that the right hand side has an approximate χ^2 distribution with k degrees of freedom. This is one form of the so-called *likelihood ratio test*. The three tests are

¹ This can mean either convergence in probability or almost sure convergence (Rice 1995, 164).

asymptotically equivalent, but their small sample characteristics may differ (Rao 1973, 415–418).

We conclude with definition of $o(\cdot)$ and $O(\cdot)$ notation. Let $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$ be two sequences of numbers. We say that a_n is $o(b_n)$ if $\lim_n |a_n/b_n| = 0$, and $a_n = O(b_n)$ if $|a_n/b_n|$ is bounded when n is large. To allow continuous arguments we say that $a(x)$ is $o(b(x))$ or $O(b(x))$ as $x \rightarrow L$ if $a(x_n)$ is $o(b(x_n))$ or $O(b(x_n))$ for any sequence $\{x_n\}_{n=1}^{\infty}$ with $x_n \rightarrow L$. For example, $6x^4$ is $O(x^4)$ and $o(x^5)$ as $x \rightarrow \infty$, and $6x^4$ is $O(x^4)$ and $o(x^3)$ as $x \rightarrow 0$.

2

Sources of Demographic Data

1. Populations: Open and Closed

We can think of a population size as a *process*. At any given time t a set of individuals satisfy the membership criterion of the population. In the case of a geographic area, for example, the criterion is “being in the area”. The population can increase via births and in-migration. It can decrease via deaths and out-migration.¹ Thus, births, deaths, and migration form the relevant *vital processes*.

Traditionally, the term *vital event* is used for births, deaths, marriages and divorces but not for migration (cf., Shryock and Siegel 1976, 20). Although this usage has an origin in civil registration, the distinction is not useful in statistical demography and we consider vital processes to include migration. Changes of marital status can be vital processes, if the population of interest has been defined in terms of marital status, but so can be such processes as getting a job or becoming unemployed, if the population is defined in terms of employment status.

In a limiting case we define a population as *closed* if it has no vital processes. A closed population is simply a set of individuals. (In demography it is common to call a population closed even if it experiences births and deaths. We take here a broader view.) In most demographic applications a population is open in some respects. For example, in a follow-up study of a fixed set of individuals, the population is closed with respect to births and in-migration, but it is open with respect to deaths. Annoyingly from the researcher’s point of view, such a population may, in practice, be open to out-migration and other forms of attrition or loss from follow-up, as well.

As discussed below, the distinction between closed and open populations is important in the design of the data collection for demographic studies. However, in most parts of this book we have the prototype of national population in mind. National populations are open to births, deaths, migration etc.

¹ A population can also change when its definition changes, e.g., when a country, state, or city annexes or de-annexes an area. Such changes do not involve vital processes, and analysis of past data on population change should make allowance for any significant boundary changes that occurred.