Springer Series in Statistics

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.

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 — *Juha M. Alho*
Evanston, Illinois, USA — *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 — *Juha M. Alho*
Evanston, Illinois, USA — *Bruce D. Spencer*

Contents

List of Examples

List of Figures

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 19^{th} 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 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 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 $F(.)$ is continuous this probability is zero. If $F(.)$ is differentiable, then $F'(.) = f(.)$ 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 $f(y) = (2\pi)^{-1/2}\sigma^{-1} \exp(-(y - \mu)^2/(2\sigma^2))$. $\diamondsuit$

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 $x \in \{0, 1\}$. $\diamondsuit$

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 $p(.)$ is differentiable, then $f(x) = -p'(x)$.

The joint probability of events $A_1, \ldots, A_n$ is $P(A_1 \cap \ldots \cap A_n)$, but we sometimes write $P(A_1, \ldots, A_n)$ 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 $X_1, \ldots, X_n$ are random variables, their *joint distribution function* is $F(x_1, x_2, .., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, .., X_n \leq x_n)$. Writing column vectors $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{X} = (X_1, \ldots, X_n)^T$, 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 $f(.)$, or if X takes discrete values $x_1, x_2, \ldots$, then

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

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, $E[\mathbf{X}] = (E[X_1], \ldots, E[X_n])^T$. 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 $\mathbf{X} = (X_1, \ldots, X_n)^T$ 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 $\text{Cov}(\mathbf{a} + \mathbf{BX}) = \mathbf{B}\text{Cov}(\mathbf{X})\mathbf{B}^T$.

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 $f_i(x_i; \boldsymbol{\theta})$, $i = 1, \ldots, n$, where $\boldsymbol{\theta}$ is, say, a $k \times 1$ vector of parameters belonging to some set $\boldsymbol{\Theta} \subset \mathbb{R}^k$. 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 $i = 1, \ldots, n$. 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 $\text{Cov}(\mathbf{S}_i(\boldsymbol{\theta})) = \mathcal{I}_i(\boldsymbol{\theta})$, $i = 1, \ldots, n$, 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 $i = 1, \ldots, n$. I.e., this is a matrix whose (r, s) element is $\partial^2/\partial\boldsymbol{\theta}_r\partial\boldsymbol{\theta}_s \log(f_i(X_i; \boldsymbol{\theta}))$. Their sum $\mathbf{H}(\boldsymbol{\theta}) = \mathbf{H}_1(\boldsymbol{\theta}) + \cdots + \mathbf{H}_n(\boldsymbol{\theta})$ 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, $E[S_i(\boldsymbol{\theta})\mathbf{S}_i(\boldsymbol{\theta})^T] = \mathrm{Cov}(\mathbf{S}_i(\boldsymbol{\theta})) = \mathcal{I}_i(\boldsymbol{\theta})$, 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

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

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 $\mathbf{S}(\boldsymbol{\theta})/n \to \mathbf{0}$,[1] as $n \to \infty$, 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 $\mathrm{Cov}(\hat{\boldsymbol{\theta}}) = \mathcal{I}(\boldsymbol{\theta})^{-1}$. 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

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

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 a 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 $o(.)$ and $O(.)$ 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 $x \to L$ if $a(x_n)$ is $o(b(x_n))$ or $O(b(x_n))$ for any sequence $\{x_n\}_{n=1}^{\infty}$ with $x_n \to L$. For example, $6x^4$ is $O(x^4)$ and $o(x^5)$ as $x \to \infty$, and $6x^4$ is $O(x^4)$ and $o(x^3)$ as $x \to 0$.

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.