Demographic Analysis: Stochastic Approach

Krishnan Namboodiri
Professor Emeritus
Ohio State University, Columbus, Ohio, USA

November 4, 2015

1 Introduction

Demographers study population dynamics: changes in population size and structure resulting from fertility (reproduction), mortality (deaths), and spatial and social mobility. The focus may be the world population or a part of it, such as the residents of a country or the patients of a hospital. Giving birth, dying, shifting usual place of residence, and trait changes (e.g., getting married) are called events. Each event involves transition from one "state" to another (e.g., from never-married state to married state). A person is said to be "at risk" or "exposed to the risk" of experiencing an event, if for that person the probability of that experience is greater than zero. The traits influencing the probability of experiencing an event are called the risk factors of that event (e.g., high blood pressure, in the case of ischemic heart disease). Demographic data are based on censuses, sample surveys, and information reported to offices set up for continuously recording demographic events. Some observational studies can be viewed as random experiments. For an individual selected at random from a population at time t, the value of the variable $y_{t+\theta}$, denoting whether that individual will be alive as of a subsequent moment $t + \theta$ is unpredictable. This unpredictability of the value of $y_{t+\theta}$ qualifies the observational study involving observations at times t and $t + \theta$ to be considered as a random experiment and $y_{t+\theta}$, as a random variable, defined by a set of possible values it may take (e.g., 1 if alive at time $t + \theta$, and 0, otherwise), with a probability function associated therewith (Kendall and Buckland 1971). The interval between a fixed date and a subsequent event is a random variable, in the above-mentioned sense.

The term rate is used in demography for the number of events (e.g., deaths) expressed per unit of some other quantity, such as person-years at risk (often expressed per 1000). For example, the *crude death rate* (annual number of deaths expressed per 1000 mid-year population) in Japan in 2007 was 9. The mid-year population in such calculations is an approximation to the sum of the person-years lived by the members of the population involved during the specified year. Death rates calculated for sub- populations, homogeneous, to some degree, with respect to one or more relevant risk factors are called specific

death rates. Examples are age-specific and age-sex specific death rates.

A *life-table* shows the life-and-death history of a group of persons, called a *cohort*, born at the same time (e.g., a year), as the cohort members survive to successive ages or die in the intervals, subject to the mortality conditions portrayed in a schedule of age-specific death rates. An account of the origin, nature, and uses of life tables is available in P. R. Cox (1975). Life tables have become powerful tools for the analysis of non-renewable (non-repeatable) processes. If a repeatable process, such as giving births, can be split into its non-renewable components (e.g., births by birth order) then each component can be studied, using the life-table method. The term: *survival analysis* is applied to the study of non-renewable processes, in general. Associated with the survival rate is the hazard rate, representing the instantaneous rate of failure (to survive). Hazard rate corresponds to the instantaneous death rate or force of mortality, as used in connection with life tables.

2 Macro-Level Focus

A great deal of demographic research is linked directly or indirectly to model construction and validation, viewing observations as outcomes of random experiments. Birth-and-death process (see Kendall, 1948; Bhat, 1984) is a continuous time, integer valued, counting process, in which population size at time t, remains constant, increases by one unit (a birth), or decreases by one unit (a death), over the period: t to $t + \Delta t$. Time-trend in population size is studied using branching processes, in a simple version of which, each member of each generation produces offspring, in accordance with a fixed probability law common to all members (see, e.g., Grimmett and Stirzaker, 1992 for a discussion of simple as well as complex models of branching processes). The logistic process for population growth of the "birth-and-death" type views the instantaneous rates of birth and death per individual alive at a given moment as linear functions of population size (see Brillinger, 1981; Goel and Dyn, 1979; Mollison, 1995). For compositional analysis, one may apply an appropriate log-ratio transformation to the composition of interest, and treat the resulting values as a random vector from a multivmultivariate normal distribution (see Aitchison, 1986; Namboodiri, 1991).

Using the component model (see Keyfitz, 1971) of population projection, one obtains internally consistent estimates of the size and age-sex composition of populations as of future years by combining hypothesized patterns of change in fertility, mortality, and migration. On the basis of such projections, issues such as the following can be examined: (1) Reduction in population growth rate resulting from the elimination of deaths due to a specific cause, e.g. heart disease; (2) Relative impact on the age-composition, in the long-run, of different combinations of population- change components (e.g., fertility and mortality); and (3) tendency of populations to "forget" the past features (e.g., age composition) if the components of population dynamics were to continue to operate without change over a sufficiently long time.

To estimate and communicate the uncertainty of population projections, the practitioners have been combining "high," "medium," and "low" scenarios for the components of population change in various ways (e.g., "high" fertility combined with "low" mortality to produce "high" population projection) to show different possibilities regarding future population size and composition. Since such demonstrations of uncertainties have no probabilistic interpretations, Lee and Tuljapurkar, among others, have pioneered efforts to develop and popularize the use of stochastic population projections (see Lee, 2004). Lee and Tuljapurkar (1994) demonstrated, for example, how to forecast births and deaths, from time-series analyses of fertility and mortality data for the United States, and then combine the results with deterministically estimated migration to forecast population size and composition. They used in the demonstration, products of stochastic matrices.

Comparison of the simple non-stochastic trend model: $y_t = \beta_0 + \beta_t(t) + e_t$, with the stochastic (random-walk with a drift) model: $y_t = \alpha_0 + y_{t-1} + e_t$, where e_t's are $NID(0, \sigma_e)$ for all t, shows that even when the error terms have equal variance (σ_e^2) in the two models, the prediction intervals for the latter are wider than those of the former: For a forecast horizon H, the variance of the forecast error (the departure of the forecast from the actual) in the case of $y_t = \beta_0 + \beta_1(t) + e_t$ is σ_e^2, while the corresponding quantity is $H\sigma_e^2$, in the case of $y_t = \alpha_0 + y_{t-1} + e_t$.

3 Micro-Level Processes

At the micro level, one focuses on events (such as giving birth to the first child, dying, recovering from illness, and so on) experienced by individuals. In event histories, points of time at which transitions occur (e.g., from not in labor force to employed) are represented by a sequence of non-negative random variables: $(T_1, T_2, \ldots)$, and the differences: $V_k = T_k - T_{k-1}$, $k = 2, 3, \ldots$, are commonly referred to as waiting times. Comprehensive discussions of waiting times are available, for example, in: Cleves et al. (2004); Collett (2003); Elandt-Johnson and Johnson (1980/1999); and Lawless (1982/2003).

D. R. Cox (1972) introduced, what has come to be known as, the proportional hazards model: $\lambda(t) = \lambda_0(t)\psi(z_1, z_2, \ldots, z_k)$, where "$t$" represents time, and the multiplier, $\psi(z_1, z_2, \ldots, z_k)$, is positive and time-independent. A special form of the model is: $\lambda(t) = \lambda_0(t) \exp(\Sigma \beta_j z_j)$, in which $\{\beta_j\}$ are unknown regression coefficients.

An important feature of waiting time is heterogeneity (variation among individuals) in the hazard rate (see Sheps and Menken, 1973; Vaupel et al., 1979; Heckman and Singer, 1982). Heterogeneity is incorporated often as a multiplier in the Cox proportional hazards model. For example, the hazard function for the ith individual may be specified as: $\lambda(t) = \lambda_0(t)\nu_i \exp(\Sigma \beta_j Z_{ij})$, representing an individual-specific, unobserved heterogeneity factor by ν_i. Vaupel et al. (1979) called such models: "frailty" models.

Heckman and Singer (1982) suggested the specification of the unobserved

heterogeneity factor in $\lambda(t) = \lambda_0(t)\nu_i \exp(\Sigma\beta_j Z_{ij})$, as a K-category discrete random variable. Thus the ith individual is presumed to belong to one of K groups. The value of K is determined empirically so as to maximize the likelihood of the sample on hand, under a specified (e.g., the exponential or Weibull) form for $\lambda_0(t)$. In the presence of heterogeneity, inference becomes sensitive to the form assumed for the hazard function (see, e.g., Trussell and Richards, 1985).

As Sheps and Perin (1963) and Menken (1975), among others, have pointed out, simplified models, unrealistic though they may be, have proved useful in gaining insights such as that a highly effective contraceptive used by a rather small proportion of a population reduces birth rates more than does a less effective contraceptive used by a large proportion of the population.

Some fertility researchers have been modeling parts rather than the whole of the reproductive process. The components of birth intervals have been examined, with emphasis on the physiological and behavioral determinants of fertility (see Leridon, 1977). Another focus has been abortions, induced and spontaneous (see: Abramson, 1973; Potter et al., 1975; Michels and Willett, 1996). Fecundability investigations have been yet another focus (see Menken, 1975; Wood et al., 1994). Menken (1975) alerts researchers to the impossibility of reliably estimating fecundability from survey data. The North Carolina Fertility Study referred to in Dunson and Zhou (2000) is of interest in this connection: In that study couples were followed up from the time they discontinued birth control in order to attempt pregnancy. The enrolled couples provided base-line data and then information regarding ovulation in each menstrual cycle, day-by-day reports on intercourse, first morning urine samples, and the like. Dunson and Zhou present a Bayesian Model and Wood et al. (1994) present a multistate model for the analysis of fecundability and sterility.

To deal with problems too complex to be addressed using analytic models, researchers have frequently been adopting the simulation strategy, involving computer-based sampling and analysis at the disaggregated (e.g., individual) level. See, for example, the study of (1) kinship-resources for the elderly (Murphy, 2004; Wachter, 1997); (2) female family-headship (Moffit and Rendall, 1995); (3) AIDs and the elderly (Wachter et al., 2002); and (4) the impact of heterogeneity on the dynamics of mortality (Vaupel and Yashin, 1985; Vaupel et al., 1979). Questions such as the following arise: Is it possible to reproduce by simulation the world-population dynamics, detailing the changes in the demographic-economic-spatial-social DESS) complex, over the period, say: 1900-2000? Obviously, in order to accomplish such a feat, one has to have a detailed causal model of the observed changes to be simulated. As of now no satisfactory model of that kind is available. Thinking along such lines, demographers might begin to view micro-simulation as a challenge and an opportunity to delve into the details of population dynamics.

Based on an article from Lovric, Miodrag (2011), *International Encyclopedia of Statistical Science*. Heidelberg: Springer Science+Business Media, LLC.

Dr. Krishnan Namboodiri was Robert Lazarus Professor of Population Studies at The Ohio State University, Columbus, Ohio, USA, (1984-2000) and has

been Professor Emeritus at the same institution since 2000. Before joining The Ohio State University, he was Assistant Professor, Associate Professor, Professor, and Chairman, Department of Sociology, University of North Carolina at Chapel Hill, USA, (1966-1984); Reader in Demography, University of Kerala, India, (1963-1966). Dr. Namboodiri was Editor of Demography (1976-1979), and Associate Editor of a number of professional journals such as Mathematical Population Studies (1985-1989). He has authored or co-authored over 80 publications including 12 books. He is a Fellow of the American Statistical Association, and is a recipient of honors such as Lifetime Achievement Award from Kerala University, and has been consultant from time to time to Ford Foundation, World Bank, United Nations, and other organizations.

References and Further Readings

Abramson, F.D. (1973) High foetal mortality and birth intervals. *Popul. Stud.* 27:235-242.

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data.* Chapman and Hall, London.

Bhat, U.N. (1984) *Elements of Applied Stochastic Processes, 2nd edn.* Wiley, New York.

Brillinger, D.R. (1981) Some aspects of modern population mathematics. *Can. J. Stat.* 9:173-194.

Cleves, M., Gould, W.G., and Gutierrez, R.G. (2004) *An Introduction to Survival Analysis Using Stata, Revised edn.* Stata Press, College Station.

Collett, D. (2003) *Modeling Survival Data in Medical Research, 2nd edn.* Chapman and Hall/CRC, Boca Raton.

Cox, D.R. (1972) Regression models and life tables (with discussion). *J. Roy. Stat. Soc., Series B* 34:187-202.

Cox, P.R. (1975) *Population Trends, Vols I.II.* Her Majesty's Stationary Office, London.

Dunson, D.B. and Zhou, H. (2000) A Bayesian model for fecundability and sterility. *J. Am. Stat. Assoc.* 95(452):1054-1062.

Elandt-Johnson, R. and Johnson, N. (1980/1999) *Survival Models and Data Analysis.* Wiley, New York.

Goel, N.S. and Dyn, N.R. (1979) *Stochastic Models in Biology.* Academic Press, New York.

Grimmett, G.R. and Stirzaker, D.R. (1992) *Probability and Random Processes, 2nd edn.* Clarendon Press, Oxford.

Heckman, J.J. and Singer, B. (1982) Population heterogeneity in demographic models. In: *Multidimensional Mathematical Demography*, Land, K.C., Rogers, A. (eds). Academic Press, New York, pp. 567-599.

Kendall, D.G. (1948) A generalized birth and death process. *Ann. Math. Stat.* 19:1-15.

Kendall, M.G. and Buckland, W.R. (1971) *Dictionary of Statistical Terms, 3rd edn.* Hafner, New York.

Keytz, N. (1971) Models. *Demography* 8:329-352.

Lawless, J.F. (1982/2003) *Statistical Models and Methods for Lifetime Data, 2nd edn.* Wiley, New York.

Lee, R.D. (2004) Quantifying our ignorance: Stochastic forecasts of population and public budgets. In: *Aging, Health, and Public Policy: Demographic and Economic Perspectives*, Waite LJ (ed). *Population and Development Review (A Special Supplement)* 30:153-175.

Lee, R.D. and Tuljapurkar, S. (1994) Stochastic population projections for the United States beyond high, medium, and low. *J. Am. Stat. Assoc.* 89(438):1175-1189.

Leridon, H. (1977) *Human Fertility: The Basic Components* (trans: Helzner, J.F.). University of Chicago Press, Chicago.

Menken, J. (1975) Biometric models of fertility. *Soc. Forces* 54:52-65.

Michels, K.B. and Willett, W.C. (1996) Does induced or spontaneous abortion affect the risk of cancer? *Epidemiology* 7:521-528.

Moffit, R.A. and Rendall, M.S. (1995) Cohort trends in the lifetime distribution of family headship in the United States, 1968-1985. *Demography* 32:407-424.

Mollison, D. (ed) (1995) *Epidemic Models: Their Structure and Relation to Data.* Cambridge University Press, London.

Murphy, M. (2004) Tracing very long-term kinship networks using SOCSIM. *Demogr. Res.* 10:171-196.

Namboodiri, K. (1991) *Demographic Analysis: A Stochastic Approach.* Academic Press, San Diego/New York.

Potter, R.G., Ford, K., and Boots, B. (1975) Competition between spontaneous and induced abortions. *Demography* 12:129-141.

Shepsm, M.C. and Menken, J. (1973) *Mathematical Models of Conception and Birth.* University of Chicago Press, Chicago.

Sheps, M.C. and Perin, E.B. (1963) Changes in birth rates as a function of contraceptive effectiveness: Some applications of a stochastic model. *Am. J. Public Health* 53:1031-1046.

Trussell, T.J. and Richards, T. (1985) Correcting for unobserved heterogeneity in hazard models: An application of the Heckman-Singer model for demographic data. In: *Sociological Methodology*, Tuma, N.B. (ed). Jossey-Bass, San Francisco, pp. 242-276.

Vaupel, J.W., Manton, K.G., and Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography* 16:439-454.

Vaupel, J.W. and Yashin, A.J. (1985) Heterogeneity ruses: Some surprising effects of selection in population dynamics. *Am. Stat.* 39:176-185.

Wachter, K.W. (1997) Kinship resources for the elderly. *Philos. T. Roy. Soc. B* 352:1811-1817.

Wachter, K.W., Knodel, J.E., and Vanlandingham, M. (2002) AIDs and the elderly of Thailand: Projecting familial impacts. *Demography* 39:25-41.

Wood, J.W., Holman, D.J., Yashin, A.I., Peterson, R.J., Weinstein, M., and Chang, M.C. (1994) A multistate model of fecundability and sterility.

Demography 31:403-426.