Parameterization and Curve Fitting

An early attempt to describe demographic observations in a mathematical formula was contributed by Abraham DeMoivre, whose hypothesis (1725, p. 4) "consists in supposing that the number of lives existing at any age is proportional to the number of years intercepted between the age given and the extremity of old age," i.e.

$$l_x = l_0\left(1 - \frac{x}{\omega}\right),$$

from which we have for the force of mortality

$$\mu_x = \frac{1}{\omega - x}.$$

DeMoivre chose the upper limit $\omega = 86$ for its quite good fit to Halley's (1693) life table, and qualified his analysis by setting a lower age limit of 12. His assumption that l_x is a linear function of age was used for short age intervals by Joshua Milne (1815) to develop his q_x formula and it remains in use in that context. As an approximation to the full life table it has long been obsolete. DeMoivre (1733) was also first to develop the normal distribution, discussed below in connection with fertility.

We begin this chapter with an article by Benjamin Gompertz (1825) who used the better mortality statistics of his time to suggest the expressions

$$\mu_x = Bc^x,$$
$$l_x = l_0 g^{(c^x - 1)},$$

where B, c, and g are constants; in effect a law that mortality increases exponentially with age. Gompertz was innovative in that he reasoned from a law of mortality in finding l_x, where a century earlier DeMoivre had not attempted to justify his implied mortality function. Gompertz is as much remembered for having reasoned well: his Law is a better description of the upper ages than DeMoivre's hypothesis that populations collapse towards a specific age limit. From Gompertz are derived Farr's (1864) and indirectly Reed and Merrell's (1939) excellent approximations to ${}_nq_x$.

An improvement on Gompertz' Law was contributed by William Makeham (1860, 1867), whose 1867 article is included here as paper 31. In a careful examination of causes of death (as they were then listed), Makeham found that overall mortality levels could be better represented if a constant term were added to μ_x to account for causes of mortality not dependent on age, a possibility Gompertz also had noted. This gives

$$\mu_x = A + Bc^x,$$
$$l_x = l_0 s^x g^{(c^x - 1)}.$$

For ages beyond childhood and youth Makeham's Law combines intuitive plausibility and a close fit to observed mortality.

Other equations have been suggested that give a slightly better fit to mortality rates overall but have spurious inflection points or are otherwise not intuitive. For these the reader may consult Hugh M. Wolfenden (1942, pp. 79—85; 1954, pp. 164—167). A further contribution to the interpretation of Gompertz' and Makeham's Laws and their generalization will be found in Brillinger (1961).

The lack of a satisfactory analytic expression for mortality that includes both infancy and adulthood has driven workers in the field to the alternative approach provided by model life tables, which average observed mortality schedules by regression, usually on life expectancy at birth or at age ten where the strong impact of infant mortality is not felt. An early attempt was published by the United Nations (1955); but by far the best known is that of Ansley Coale and Paul Demeny (1966). From their examination of historical mortality patterns in countries having reliable data, Coale and Demeny were able to isolate four general patterns and to construct regressions of the forms

$${}_nq_x = a + b\mathring{e}_{10},$$
$$\ln {}_nq_x = a + b\mathring{e}_{10}$$

for the separate age groups within each of the four families of tables. Their blending of the regressions and general methodology are given as paper 32. How the model tables are applied to countries with poor demographic data is discussed in paper 33, from *United Nations Manual IV* (1967) which Coale and Demeny prepared. An application of factor analysis to the problem of sorting out existing life tables and making the models will be found in Ledermann and Breas (1959).

William Brass (Brass and Coale 1968) has suggested an ingenious combination of analytic curves and model tables that provides greater flexibility in table use, which we include here as paper 34. Brass takes a selected model table as the standard, say $l_x^{(s)}$, and adjusts it to an observed set of rates by a regression of the form

$$\ln[(1 - l_x)/l_x] = \alpha + \beta \ln[(1 - l_x^{(s)})/l_x^{(s)}].$$

The equation preserves the extreme values $l_0=1$, $l_\omega=0$, and allows model tables to be used in cases where observed rates are incomplete and not quite of the (usually European) pattern the models reflect. The Brass technique also has much potential value for forecasting, since in many instances the constants α and β show fairly clear trends over time (Brass 1974, pp. 546—551).

The early attention and degree of success achieved in mapping and analyzing mortality patterns contrasts sharply with the very meager efforts that were made to understand marriage and fertility. Most of the 19th century elapsed between Nicander's publication of fertility rates and their use by Milne (1815), and Richard Böckh's (1886, p. 30) introduction of the Net Reproduction Rate, defined as the integral

$$R_0 = \int_\alpha^\beta p(a)m(a)da\,,$$

where $p(a)$ is the probability that an individual survives from birth to age a and $m(a)$ is the expected number of offspring of the same sex born to him (her) in the interval a to $a+da$. The integration is across all fertile ages and hence gives expected progeny to an individual just born. [Richard R. Kuczynski, at one time a student of Böckh's and an assistant in the Berlin statistical office, in several of his works drew attention to the implications of the rate for population survival and growth; e.g. (Kuczynski 1928, pp. 41—42): "The pertinent question is not: is there an excess of births over deaths? but rather: are natality and mortality such that a generation which would be permanently subject to them would during its lifetime, that is until it has died out, produce sufficient children to replace that generation? If, for instance, 1,000 newly born produce in the course of their lives exactly 1,000 children, the population after the death of the older 1,000 will remain unaltered ... and if natality and mortality remain permanently the same, the population will always exactly hold its own. If more than 1,000 children are produced by a generation of 1,000 newly born, the population will increase; if less than 1,000 are produced, the population will decrease and finally die out."] Neither the net reproduction rate nor the simpler age-specific fertility rates were in use in England when Cannan carried out his population projections in 1895.

With Lotka's contributions to stable population theory good approximations to the net maternity function and related measures became immediately valuable, and several attempts to fit equations to them were made. The earliest employ Pearson distributions (Elderton and Johnson 1969, pp. 35—109; Kendall and Stuart 1969, Vol. 1, pp. 148—154), which are solutions to the differential equation

$$\frac{d\ln f(x)}{dx} = \frac{(x-a)}{b_0+b_1x+b_2x^2},$$

most but not all having a single mode at $x=a$ and making smooth contact with the x-axis at the extremities. They include the normal distribution, investigated by Dublin and Lotka (1925); the Pearson Type I by S.J. Pretorius (1930) and Lotka (1933); and Pearson Types I and III by S.D. Wicksell (1931). (For the normal, $b_1=b_2=0$; for the Type III, $b_2=0$.)

The normal, Type III and an exponential introduced by Hadwiger (1940) are analyzed in detail in Keyfitz (1968, pp. 141—169). We include here as paper 35 Wicksell's remarks about the normal and Pearson curves.

None of the equations we mention has been defined to express either marriage patterns or birth interval and family size distributions, and the most recent work in the field has been directed toward providing this essential base. A large measure of success has been achieved by Ansley Coale (Coale 1971; Coale and McNeil 1972; Coale and Trussell 1974), in work directed toward the development of model nuptiality and fertility tables. The work combines four elements. Drawing on evidence from a number of populations, Coale (1971) found that the age-specific risk of marriage among those who ever marry could be closely approximated by a double exponential (Gompertz) distribution, differing between populations in origin (age at which marriages essentially begin) and the relative intensity of the marriage process once underway. For fertility by duration of marriage, Coale suggested taking as a standard that of a natural fertility population (i.e., one in which birth limitation is not intentionally practiced), reducible by an exponential decay function to represent fertility patterns imposed by family planning practice.

In later work extending Griffith Feeney's (1972) suggestion that the age-specific probability of marriage could be represented as the convolution of a normal distribution to represent age at eligibility with an exponential distribution representing delays between eligibility and marriage, Coale and McNeil found that as further exponentials are introduced the distribution takes as its limiting value a Makeham curve of the form (Coale and McNeil 1972, p. 744)

$$\bar{g}(x) = \frac{\lambda}{\Gamma(\alpha/\lambda)} \exp\{-\alpha(x-\mu) - \exp[-\lambda(x-\mu)]\},$$

where Γ is the gamma function, $\mu = a + \dfrac{\Gamma'(\alpha/\lambda)}{\lambda\Gamma(\alpha/\lambda)}$, and a is the mean of $\bar{g}(x)$. The distribution has the important properties that it closely matches the probability distribution $g(x)$ generated by the Coale risk function and can be approximated to a high degree of accuracy using 1 to 3 exponentials. The distribution is thus both empirical and intuitive. The process of fitting the curve is greatly simplified by taking advantage of the similarities between different populations to establish a standard distribution. This distribution, fitted by two constants, and the distribution for fertility by duration of marriage, fitted by one constant, form the base for the Coale and Trussell (1974) model fertility tables. Paper 36 is from this article. [For another model expressing age at first marriage, also with an intuitive base, see Hernes (1972). The various models about equally reflect observed marriage patterns.]

The future growth of populations was first treated mathematically by Pierre-François Verhulst (1838), in response to Malthus' argument that populations would tend to grow exponentially until checked by resource limits. By implication, long range projections might be formed without detailed reference to current age structure, fertility or mortality, an assumption compatible with the lack of fertility information in Verhulst's time. His suggestion, paper 37, was to represent population growth by the logistic

$$r_t = b\left(1 - \frac{P_t}{k}\right),$$

$$P_t = \frac{k}{1+e^{a-bt}},$$

which makes the intrinsic growth rate r_t a linearly decreasing function of population size P_t, k being the asymptotic population. The equation was independently applied to U.S. population growth by Raymond Pearl and Lowell Reed (1920), whose article here follows Verhulst's. Pearl and Reed's satisfaction with the logistic was not warranted when they wrote; in particular they failed to notice that the good fit of their equation reflected extraordinary contributions of technology, immigration and territorial expansion to U.S. population growth, making the correspondence of r_t and P_t rather more fortuitous than imperative. (As it turned out, the curve was 0.3% under the 1930 census but over by 3.5% at the next count. The higher figure translates as a 57% overestimate of intercensal growth. It may be compared with Whelpton's fine componant projection for 1940, 0.2% above the census, and his 6.0% and 49% underestimates for 1950 population and 1940—1950 intercensal growth respectively. In this case the better method was not a better guarantee of success.) Pearl and Reed toward the end of the paper take care in suggesting limitations of the logistic, and the reader will find there some of the reasons why they should not have used it; others were provided by Lancelot Hogben (1931, pp. 176—184) and William Feller (1940).

In another work that deserves mention, Thomas Edmonds (1852) applied Gompertz' Law to population growth by defining the rate of population increase as:

$$\frac{dP}{dt} = Bc^t,$$

with the distinction that $c<1$ in order that the rate of growth would decline rather than increase over time. (In Verhulst, this corresponds to the damping factor $\phi(p)$.) Edmonds' equations are those later reapplied to mortality by Farr (1864), and he is almost certainly Farr's immediate source. The equations have not been remembered in population projections.

The chapter concludes with an examination by Otis Dudley Duncan (1958) of human spatial measurement and efforts that have been made to fit curves to city size distributions, using either Felix Auerbach's (1913) rank-size rule or Mario Saibante's (1928) power function. These analyses confront measurement problems more complex than we have treated elsewhere in the book.

Readers should be aware that the equations presented here are fitted to real data in different ways: DeMoivre's and those given by Duncan are solved by regression; the early fertility functions by moments; and Gompertz and Verhulst used simultaneous equations with selected data points. [Here, note that we might not follow the same approach. Curves in which constants enter non-linearly are accessible to least squares fitting by computer, using iterative methods; neither moment nor selected-point fittings are of equivalent quality.]

The method of least squares was first suggested by Carl Friedrich Gauss in personal communications and by Adrien Marie Legendre (1805) in print; fitting by moments is due principally to Pearson (1893, 1948).