Rates and Proportions

POP 502 / ECO 572/ SOC 532 • SPRING 2017

Proportions

Proportions are used to describe situations with two possible outcomes. For example newborn babies are classified shortly after birth as males or females. There are approximately 105 male births for every 100 female births, a sex ratio at birth of 105. The *proportion* male is therefore 105/205 = 0.512 or 51.2%.

Proportions are often interpreted as estimates of probabilities. We could say that the probability that a newborn baby is male is 0.512.

> To calculate a proportion you divide the number of units who possess the attribute of interest by the total number of units under observation. Proportions (and probabilities) are always between zero and one.

Rates

Rates are used to describe the frequency of occurrence of events over time. For example your heart rate is the number of times your heart beats per minute. Women aged 15-44 in the U.S. have births at a rate of 62.5 births per 1000 woman-years of exposure. Note the key role played by time in both examples.

> To calculate a rate you divide the number of events observed in a period of time by a measure of exposure based on the number of units under observations and the time each one was at risk. Rates cannot be negative but they can exceed one.

For annual rates exposure is often estimated as the mid-year population. Some events terminate exposure but others don't. (Think of births and deaths.)

Usage

Rates and probabilities can be confusing because usage is not consistent. Some authors use the two terms interchangeably and many use "rate" to describe proportions. For example the contraceptive prevalence "rate" is just the proportion of women currently using contraception. There is no time interval involved. At least in this case there is no risk of confusion. The most egregious example is the "infant mortality rate", which is an estimate of the probability that a child will die before reaching its first birthday. This causes confusion because, as we shall see, we can also calculate a rate. We will return to this issue when we compute life tables.

Direct and Indirect Standardization

POP 502 / ECO 572/ SOC 532 • SPRING 2017

Rates as Weighted Averages

Any general rate can be written as a weighted average of group-specific rates, with weights given by the proportion of exposure time in each group. The basic result (see page 23 in the textbook) is

$$R = \sum c_i m_i$$

where c_i reflects the composition and m_i the rate for group i. A general rate can change over time (or differ across groups) because of changes in composition, even if the group-specific rates are constant. (Frequently the groups are defined by age.)

Example: The CDR in Kazakhstan is lower than in Sweden: 7.42 compared to 10.55 deaths per 1000 population. We suspect this is due to its younger age structure. Note: All calculations are shown in the website using R and Stata. The key results are summarized in the following table combining different rates and compositions:

Composition	Rates		
	Kazakhstan	Sweden	Average
Kazakhstan	7.42	4.20	5.81
Sweden	16.34	10.55	13.44
Average	11.88	7.37	

Direct Standardization

The purpose of standardization is to facilitate comparison of rates over time (or across groups) by removing the effect of composition. The direct standardized rate for a given population combines the population's group-specific rates m_i with the composition of a standard population c_i^S:

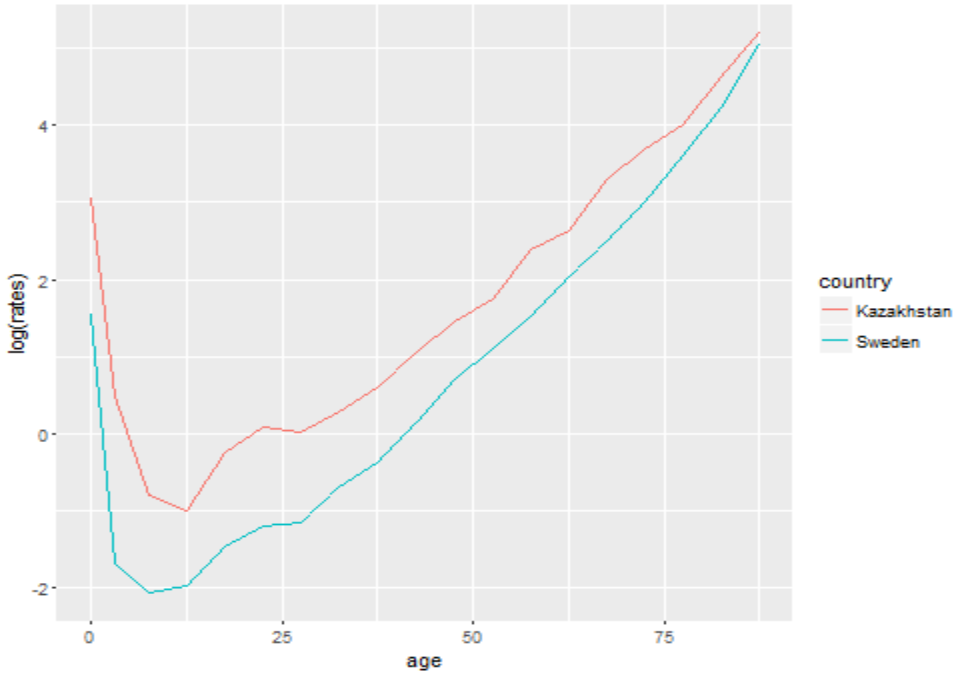
$$DSR = \sum c_i^S m_i$$

We interpret this rate as a *counterfactual*: the rate that would be observed if the population had the standard composition but its own group-specific rates. Using a uniform distribution as the standard composition yields the average rate. (A closely related measure is the total fertility rate, which has a synthetic cohort interpretation.) A good choice of standard to compare two populations is the average composition.

Example: If Kazakhstan had Sweden's age structure the CDR would be 16.34, which is 77% higher than the CDR in Sweden. Standardizing both countries using the average age structure gives rates of 11.88

and 7.37, or 61% higher mortality in Kazakhstan. The Swedish standard gives more weight to older ages where the relative difference is smaller, see the graph.

Figure 1. Age-Specific Mortality Rates



Decomposition of a Difference between Two Rates

The difference between two general rates can be due to differences in group-specific rates and differences in composition:

- The effect of the rates can be ascertained by comparing standardized rates that use a standard composition and the actual rates, say $\sum c^S m_i^A$ for population A and $\sum c^S m_i^B$ for population B, where c_i^S denotes the standard composition.
- The effect of the composition can be ascertaining by comparing summary measures that use the actual compositions and a set of standard rates, say $\sum c_i^A m_i^S$ and $\sum c_i^B m_i^S$, where m_i^S denotes the standard rates.

One difficulty with this approach is that these two components do not always add up to the original difference, which has led some authors to propose adding "interaction" terms. A much simpler approach is to use as standard the average of the two populations, taking

$$c_i^S = \frac{1}{2}(c_i^A + C_i^B) \text{ and } m_i^S = \frac{1}{2}(m_i^A + m_i^B).$$

The decomposition is then exact, and no additional terms are needed (see page 28 in the textbook). This works for proportions too!

Example: The CDR is 3.12 points *lower* in Kazakhstan than in Sweden. If both countries had the average age composition the CDR would be 4.51 *higher* in Kazakhstan. If both countries had the same rates the CDR would be 7.63 *lower* in Kazakhstan. In other words the difference of -3.12 between Kazakhstan and Sweden results from higher rates (+4.51) compensated by a younger age structure (-7.63) in Kazakhstan.

Standardized Mortality Ratio

Computation of a direct standardized rate requires knowledge of the group-specific rates. Can we adjust for compositional effects if these are not known? Yes, provided we know the composition. We can compute the rate that would be observed if the population had its own composition but a standard set of rates, and compare this to the observed rate (which results from the same composition but the actual rates).

The standardized mortality ratio is the ratio of observed to expected deaths (or death rates) under the standard:

$$SMR = \frac{R}{\sum c_i m_i^S}$$

We interpret this ratio as a *counterfactual*: proportionally how much higher (or lower) mortality would be if the group-specific rates were the same as in the standard. The same idea applies to other events. (The Princeton fertility index I_f, for example, has the same construction.)

Example: The SMR for Kazakhstan using Sweden as the standard is 7.42/4.20 = 1.77. Thus, the CDR in Kazakhstan is 77% higher than it would be if it had Sweden's age-specific mortality rates (but its own age structure).

Indirect Standardization

A closely-related approach is to approximate the direct standardized rate using a two-step procedure: we figure out the effect of composition using the observed and standard compositions with the standard rates, and then apply this as a correction to the observed rate. In symbols:

$$ISR = R\,\frac{R^S}{\sum c_i m_i^S}$$

where R^S is the general rate in the standard population.

Example: The indirect standardized mortality rate for Kazakhstan using Sweden as the standard is 7.42 x (10.55/4.20) = 18.64. (This is an approximation to the direct standardized rate of 16.34 obtained earlier.)

Period Life Tables

POP 502 / ECO 572/ SOC 532 • SPRING 2017

Rates to Probabilities

To convert a rate to a probability all we need is ${}_nm_x$, as we can then use

$$ {}_nq_x = \frac{n \quad {}_nm_x}{1 + (n - \quad {}_na_x) \quad {}_nm_x} $$

This is the one and only formula you need to build a life table, everything else is pretty intuitive.

Coale and Demeny suggest values of ${}_na_x$ for young ages, see Table 3.3 on page 48.

For the last open-ended age group we use

$$ {}_\infty a_x = \frac{1}{{}_\infty m_x} $$

or equivalently set ${}_\infty q_x = 1$.

Stationary Population

Interpreting a life table as a stationary population with l_0 births per year (see page 53):

- l_x is the number age x at last birthday
- ${}_nL_x$ is the number between ages x and $x + n$
- T_x is the population age x and above
- T_0 is the total population size
- ${}_nd_x$ is the number of deaths between ages x and $x + n$
- e_x is the mean age at death of people dying in a given year

What's the crude death rate? And the crude birth rate? The growth rate?

Mortality as a Continuous Process

- $l(x)$ is the number surviving to exact age x out of $l(0)$
- The death density $d(x)$ is

$$d(x) = \lim_{n\to 0} \frac{{}_n d_x}{n} = \lim_{n\to 0} \frac{l(x) - l(x+n)}{n} = -l'(x)$$

- The force of mortality $\mu(x)$ is

$$\mu(x) = \lim_{n\to 0} \; {}_n m_x = \lim_{n\to 0} \frac{{}_n d_x}{{}_n L_x} = \lim_{n\to 0} \frac{{}_n d_x}{n l(x)} = \frac{d(x)}{l(x)} = -\frac{d}{dx} \log l(x)$$

- Integrating both sides and starting from $l(0)$ we get

$$l(x) = l(0) e^{-\int_0^x \mu(a) da}$$

The similarity to "the most important formula in demography" should not go unnoticed.

More on page 69

Probabilistic Interpretation

Let X denote a random variable representing age at death in a mortality regime given by $l(x)$

- Probability of surviving to age x is

$$\Pr(X > x) = \frac{l(x)}{l(0)}$$

- The density of age at death is

$$\lim_{dx\to 0} \frac{\Pr(X \in (x, x+dx))}{dx} = \frac{d(x)}{l(0)}$$

- The *conditional* density of age at death given survival to x is the force of mortality.

$$\lim_{dx\to 0} \frac{\Pr(X \in (x, x+dx) \mid X > x)}{dx} = \frac{d(x)}{l(x)} = \mu(x)$$

Rates to Probabilities Revisited

The probability of dying between ages x and $x+n$ conditional on having survived to x can be written as

$${}_n q_x = 1 - \frac{l(x+n)}{l(x)} = 1 - e^{-\int_x^{x+n} \mu(a) da}$$

If we assume that the force of mortality is *constant* between ages x and $x+n$ and estimate it using ${}_n m_x$ we get

$${}_n q_x = 1 - e^{-n \, {}_n m_x}$$

A simple formula for converting rates to probabilities. Note that ${}_n a_x$ is not needed but is implicit. Solving for ${}_n a_x$ in the usual formula for converting rates to probabilities we get

$$ {}_n a_x = n + \frac{1}{m} - \frac{n}{q} $$

and substituting the simple formula for ${}_n q_x$ we get the result on page 46.

Population and Life Table Rates

We now have some tools to have a closer look at these rates (see pages 61-62).

The life table mortality rates are

$$ {}_n m_x = \frac{\int_x^{x+n} l\,(a)\mu(a)da}{\int_x^{x+n} l\,(a)da} $$

a weighted average of the force of mortality in the age group with weights proportional to the stationary population.

The mortality rates observed in a population subject to the same force of mortality are

$$ {}_n M_x = \frac{\int_x^{x+n} N\,(a)\mu(a)da}{\int_x^{x+n} N\,(a)da} $$

where $N(a)$ is the population density at age a, so the rate is a weighted average of the force of mortality with weights proportional to the observed population.

There are only two cases in which equating these is right

1. When the population is stationary, so $n(a) \propto l(a)$, and
2. When the rates are constant within each age group.

This is one reason why I like the assumption of piecewise constant rates. In practice one hopes any differences within a single year of age will be too small to matter.

Stationary Populations

POP 502/ ECO 572/ SOC 532 • SPRING 2017

We discussed the interpretation of the life table as a stationary population. The following example illustrates some of the ideas.

A Social Planning Application

The population of a certain country is stationary, following the abridged life table below, which for simplicity we assume applies to both males and females. There are 10,000 births per year and no migration.

Age x	l_x	${}_nd_x$	${}_nq_x$	${}_nL_x$	T_x	e_x
0	100000	1147	0.01147	99197	7123096	71.23
1	98853	264	0.00267	394798	7023899	71.05
5	98589	164	0.00166	492511	6629101	67.24
10	98425	160	0.00163	491756	6136590	62.35
15	98265	604	0.00615	490137	5644834	57.45
20	97661	789	0.00808	486262	5154697	52.78
25	96872	665	0.00686	482658	4668436	48.19
30	96207	609	0.00633	479525	4185778	43.51
35	95598	778	0.00814	476170	3706254	38.77
40	94820	1269	0.01338	471198	3230084	34.07
45	93551	2154	0.02302	462814	2758886	29.49
50	91397	3528	0.03860	448818	2296072	25.12
55	87869	5434	0.06184	426642	1847254	21.02
60	82435	8030	0.09741	393303	1420612	17.23
65	74405	11342	0.15244	345084	1027309	13.81
70	63063	14656	0.23240	279824	682225	10.82
75	48407	16383	0.33844	201217	402401	8.31
80+	32024	32024	1.00000	201185	201185	6.28

School education is mandatory between ages 5 and 15. A presidential candidate proposes a social solidarity plan that includes an annual allowance of $2,000 for each child under the minimum school leaving age, a housing grant of $5,000 in cash to each couple on their first marriage, a cash grant of $400 for deaths under age 10 and $800 for deaths at age 10 and above, and an old age pension of $6,000 per year to each person over age 65.

(a) What's the total population of the country?
(b) If all citizens marry for the first time at age 25, what's the annual number of marriages?
(c) What's the total number of children under minimum school living age?

(d) What's the annual number of deaths under age 10? At ages 10 and over?
(e) What's the number of old age pensioners?
(f) How much would the candidate's plan cost per year?
(g) Which benefit is the most expensive?
(h) If three-quarters of the population between ages 15 and 25 and half of the population between ages 25 and 65 are employed in the labor force, how much tax would each have to pay to support the social solidarity plan entirely out of direct taxation?

There are only 10,000 births per year, so we need to adjust the radix of the life table to match. For simplicity I assume below that this is the case, applying a "reduction factor" of 0.1 to all columns representing counts (l_x, ${}_nd_x$, ${}_nL_x$, T_x)

a) The total population is $T_0 = 712{,}310$. (Not 7 million.)
b) The number of people reaching age 25 each year is l_{25} but the number of marriages is half that because it takes two to Tango: $\frac{1}{2}l_{25} = 4{,}844$
c) The number of children under age 15 is ${}_{15}L_0 = T_0 - T_{15}$=147,826.
d) The annual number of deaths under age 10 is ${}_{10}d_0 = l_0 - l_{10} = 157.5$. The number of deaths at ages 10 or older is ${}_{\infty}d_{10} = l_{10} = 9{,}842.5$. These add up to 10,000. (You may round if you wish.)
e) The number of old age pensioners or people at ages 65 and above is $T_{65} = 102{,}731$.
f) Here is a quick annual budget to go with the campaign promises:

Item	Calculation	Cost ($1,000's)
Child allowance	\$2,000 × 147,826	295,652
Housing grant	\$5,000 × 4,844	24,218
Death grant	\$400 × 157.5 + \$800× 9,842.5	7,937
Old age pension	\$6,000 × 102,731	616,385
Total		944,193

The total cost is just shy of a billion dollars. (Total w/o rounding is \$944,192,800.)
g) The most expensive benefit is the old age pension, by far.
h) The population at ages 15 to 25 is $T_{15} - T_{25} = 97{,}640$ and at ages 25 to 65 is $T_{25} - T_{65} = 364{,}113$. Given the proportions in the labor force (75% and 50%) we have 255,286 taxpayers (0.75 × 97640 + 0.50 ×364,113 = 255,286). The assessment need to balance the books is \$3,699 per person employed (\$944,192,800/255,286.2=\$3,698.57).

This problem is adapted from A. H. Pollard, F. Yusuf and G. N. Pollard (1990) *Demographic Techniques.* Third Edition. Sydney: Pergamon Press.

Life Expectancy Decompositions

Section 3.10 in the textbook considers the decomposition of a change in life expectancy in terms of the contribution of each age group. At issue are questions such as "How much of the gain in life expectancy in a recent period can be attributed to reductions in infant and child mortality?" We can answer the question in continuous or discrete time. The textbook focuses on the discrete method and gives a formula and example. We'll briefly review both, focusing on expectation of life at birth.

Pollard

Pollard (1982) proposed a continuous-time decomposition. Recall that expectation of life at age a is the ratio of time lived after age a to the number of survivors to that age. Specifically, at birth we have

$$e(0) = \frac{1}{l(0)}\int_0^\infty l\,(x)dx = \int_0^\infty e^{-M(x)}\,dx$$

where $M(x) = \int_0^x \mu\,(a)da$ is the cumulative force of mortality up to age x. The difference in life expectancy between two time periods (or two countries) may be written as

$$e_2(0) - e_1(0) = \int_0^\infty (\,e^{-M_2(x)} - e^{-M_1(x)})dx$$

Let us factor out the initial survival $e^{-M_1(x)} = l_1(x)/l_1(0)$ to obtain

$$e_2(0) - e_1(0) = \frac{1}{l_1(0)}\int_0^\infty (\,e^{M_1(x)-M_2(x)} - 1)l_1(x)dx$$

At this point we can integrate by parts using the fact that $l_1(x)$ is the derivative of $-\int_x^\infty l_1\,(a)da = -l_1(x)e_1(x)$, to obtain the main result

$$e_2(0) - e_1(0) = \frac{1}{l_1(0)}\int_0^\infty (\,\mu_1(x) - \mu_2(x))e^{M_1(x)-M_2(x)}l_1(x)e_1(x)dx$$

Pollard notes that expanding the exponential in a Taylor series and taking just the leading term (which is 1) leads to the "well-known" approximation $\Delta e(0) \approx \Delta\mu(x)e_1(x)l_1(x)/l_1(0)$. In words, a small change in death rates at age x changes life expectancy at birth by the expectation of life remaining at x times the probability of surviving to that age.

Noting that $e^{-M_1(x)} = l_1(x)/l_1(0)$ we can do some cancellation to obtain the simpler formula

$$e_2(0) - e_1(0) = \int_o^\infty (\mu_1(x) - \mu_2(x)) \frac{l_2(x)}{l_2(0)} e_1(x) dx$$

Which combines survival probabilities in the new regime ($l^2(x)$) with life expectancy in the old ($e_1(x)$. Reversing the labels leads to an equivalent expression in terms of $l_1(x)$ and $e_2(x)$. Both exact.

This method is largely of theoretical interest because to apply it we need to evaluate the integrals, which requires approximations. (See Pollard's paper if you are interested.)

Arriaga

Arriaga (1988) proposed a discrete-time decomposition that is much easier to apply to conventional abridged life tables. We consider the contribution of a change in mortality rates at ages x to $x + n$ on life expectancy at age $a < x$. We focus here on life expectancy at birth, so $a = 0$. For consistency with the textbook I'll use superscripts for the two time periods or countries.

I find that it helps follow the argument to consider the average person-years lived at ages x to $x + n$, which Arriaga calls a "temporary" life expectancy and denotes $= {}_ne_xL_x/l_x$. Changing mortality at ages x to $x + n$ has an affect at those ages and as we'll see, also an effect at later ages.

The first component, sometimes called the *direct* effect, reflects the fact that in the new regime people spend on average ${}_ne_x^2$ years at those ages instead of ${}_ne_x^1$, provided of course they make it to age x, so this first component is

$$\frac{l_x^1}{l_0^1}({}_ne_x^2 - {}_ne_x^1) = \frac{l_x^1}{l_0^1}(\frac{{}_nL_x^2}{l_x^2} - \frac{{}_nL_x^1}{l_x^1})$$

This is the first part of equation (3.11) in the textbook.

The second component reflects the fact that we now have more people coming out of the age group x to $x + n$. In the first regime we had l_{x+n}^1 exiting, but we now have $l_x^1 l_{x+n}^2/l_x^2$ exiting. (It may help to think of the last ratio as the conditional probability of surviving from x to $x + n$ in the second regime.) The additional survivors represent more person-years at later ages even if the rates at those ages don't change and still average ${}_\infty e_{x+n}^1$ years. But of course the rates themselves have changed, and they will average ${}_\infty e_{x+n}^2$ years instead. The second component is then

$$\frac{1}{l_0^1}(l_x^1 \frac{l_{x+n}^2}{l_x^2} - l_{x+n}^1){}_\infty e_{x+n}^2 = \frac{T_{x+n}^2}{l_0^1}(\frac{l_x^1}{l_x^2} - \frac{l_{x+n}^1}{l_{x+n}^2})$$

This is the second part of (3.11) in the textbook.

Arriaga further splits this term into an *indirect* effect attributable to the additional survivors at old rates, and an *interaction* effect due to the fact that those survivors face new rates. We will not distinguish these, but if you are interested the indirect effect is easily

computed using ${}_\infty e^1_{x+n}$ instead of ${}_\infty e^2_{x+n}$ in the above formula and the interaction is the difference between the two.

For the last open-ended age group there is only a direct effect, computed as

$$\frac{l^1_x}{l^1_0}\left(\frac{T^2_x}{l^2_x} - \frac{T^1_x}{l^1_x}\right)$$

This is expression (3.12) in the textbook, but it is really just a special case of the first formula above because ${}_\infty L_x = T_x$ for the open-ended age group.

If you apply these formulas to the example in the textbook make sure you use ${}_n L^2_x$ and T_x as printed, because accumulating person-years by age gives slightly difference results, probably because of rounding. The data needed are available in file `box34.dat` in my website.

Pollard (1988) shows that his continuous-time formulation and Arriaga's discrete-time analysis are exactly equivalent in the limit when one uses finer and finer age intervals, with Pollard's equation corresponding to the sum of Arriaga's direct, indirect and interaction effects.

Keyfitz

Keyfits considered the effects of absolute and relative changes in mortality at every age, and you'll find a nice writeup in Keyfitz and Caswell (2005, Section 4.3).

They show that if the rates change from $\mu(x)$ to $\mu(x) + \delta$, then the derivative of life expectancy w.r.t. δ evaluated at zero is $-\overline{x}e_0$ where $\overline{x}$ is the mean age in the stationary population. For example if life expectancy is 70 and mean age is 35, reducing all rates by 0.001 would increase life expectancy by 2.45 years.

An alternative scenario posits a proportionate change in age-specific rates, where the force of mortality goes from $\mu(x)$ to $\mu(x)(1 + \delta)$. In this case the survival probability is raised to the power $(1 + \delta)$, and the resulting integral is hard to evaluate except in special cases. They show, however, that the derivative of the log of life expectancy w.r.t. δ can be written approximately as the product $-H\delta$, where

$$H = -\frac{\int \log\left(\frac{l(x)}{l(0)}\right) l(x)dx}{\int l(x)dx}$$

is a measure knows as *entropy*, defined as the negative weighted average of $\log l(x)/l(0)$ using $l(x)$ as the weights. If everyone lived to age ω and then died $l(x)/l(0$ would be one and its log zero, so $H = 0$. At the other extreme, if the force of mortality is constant we have an exponential distribution with $l(x)/l(0) = e^{-\mu x}$, for which $H = 1$.

The product $-H\delta$ estimates the relative change in life expectancy after a proportionate change in age-specific mortality. For U.S. females in 2013 entropy was around 0.134, so a ten percent decline in mortality at all ages would increase life expectancy by about 1.34%

or 1.09 years. A quick calculation using a single year life table and reducing rates by 10% yields an actual increase in life expectancy of 1.41 years.

These results are useful because they provide some insight into the relationship between death rates and life expectancy, but they are not terribly realistic because they apply to small absolute or relative changes at all ages. However, the decomposition procedures discussed above have very wide applicability and happen to be exact.

Revised 2/27/2017

Model Life Tables

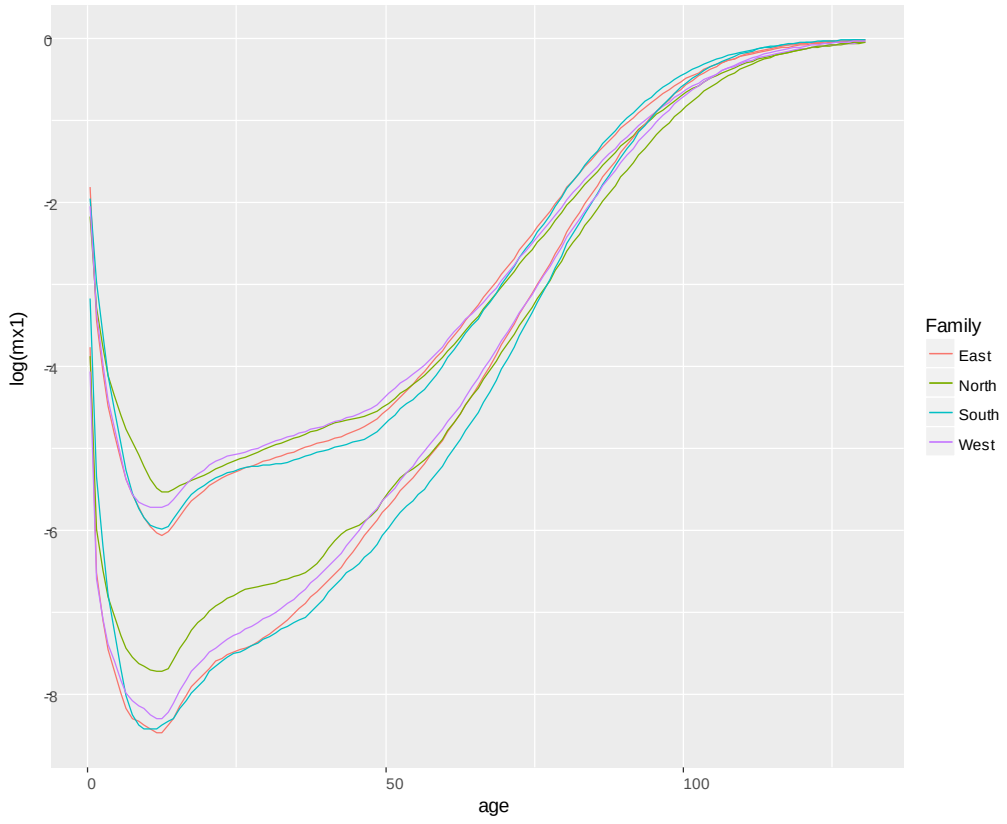
Pop 502/ Eco 572/ Soc 532 • Spring 2017

No single mathematical formula fits the wide range of observed age patterns of mortality, so several authors have developed empirical standards. We review the Coale-Demeny-Vaughan model life tables (a.k.a. the Princeton Regional Model Life Tables), the Brass Relational Logit model and a modification, and the log-quadratic model of Wilmoth and collaborators. The Coale-Demeny and original Brass models are discussed in Section 9.1 of the textbook. Later we will dicuss the Lee-Carter model used in forecasting mortality.

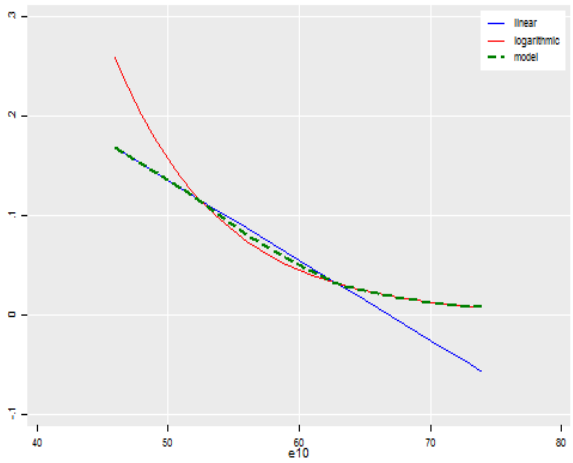
Coale-Demeny

Coale and Demeny examined a large number of life tables from countries with reliable data, mostly in Europe, and used regression methods to build four families of model life tables, labeled "West", "North", "East" and "South" because they corresponded roughly to regions of Europe, with variants for males and females. Each family is indexed by a single parameter representing the level of mortality.

The figure below shows the West female model life tables when life expectancy is 50 and 75 years, to give you an idea of the shape of the models.



The method of construction is not without interest. For each age group 0, 1-4, 5-9, ..., 75-79 in standard abridged life tables they regressed ${}_nq_x$ on e_{10}. They used a linear regression and a logarithmic one with $\log_{10}10000_nqx$ as the outcome. Table XI in their book has the coefficients. Neither model was fully adequate, but the two regressions always crossed twice within the range of e_{10}. They decided to use the linear model to the left of the first crossing, the logarithmic to the right of the second, and the average of the two in between, as shown on the right



To proceed beyond age 80 they used a Gompertz extension. Briefly they use ${}_5q_{75}$ to estimate $\mu(77.5)$ and a linear regression to predict $\mu(105)$, use these two values to estimate the Gompertz slope, and then use that to estimate death and survival probabilities for ages 80-84 to 90-94 and 95+. Time lived in 0-1 and 1-5 was computed using the regression equations we already encountered, see Table 3.3 on page 48. Time lived after age 80 was computed by numerical integration of the Gompertz survival function.

The published set of tables has 25 levels, designed to give life expectancy at birth from 20 to 80 in steps of 2.5 years. An oddity of the tables is that the value of e_{10} is not the same as the input value used to generate the death probabilities, which is not surprising in a non-linear system; it is best to think of that as just a seed. These tables have been widely used in demographic applications for many years.

The UN has published a set of "Coale-Demeny" tables which differ from the original published set because they were extended beyond age 90 (all the way to 130) and because they smoothed some of the rates. The UN also has its own set of model life tables, with families called "General", "Far Eastern", "South Asian", "Latin American" and "Chilean", but they haven't quite overcome the popularity of the Coale-Demeny system. The whole set of nine model schedules with life expectancy from 20 to 100 is available from the UN in a spreadsheet.

Brass Relational Logits

Brass proposed a model where he first transforms the survival function by taking logits, which he defines as (assuming a radix of one)

$$Y(l_x) = \frac{1}{2}\log\frac{1-l(x)}{l(x)}$$

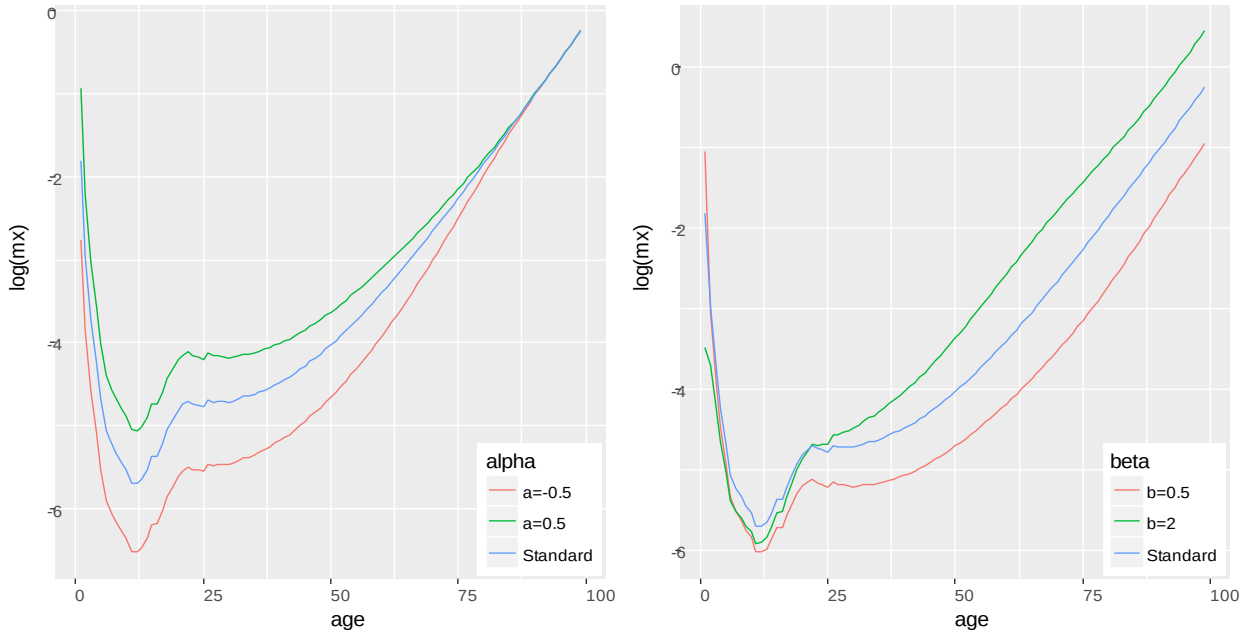
Note the arcane use of $1/2$, which was common among British statisticians, and the fact that he works with $(1-p)/p$, which changes the sign.

He then writes the transformed schedule as a linear function of a standard logit schedule $Y(l_x^s)$, so that

$$Y(l_x) = \alpha + \beta Y(l_x^s)$$

The essential idea is that a simple transformation of the l_x function allows relating any survival curve to another, including the standard. It is very easy to check visually the fit of the model because a plot of the observed logits versus the model logits should yield a straight line.

The plots below show how the parameter α reflects the level of mortality and β the shape of the schedule, or balance between child and adult mortality. (I find the plots of the force of mortality more informative than the plots of the survival functionsdepende shown in the textbook.)



The relational logit model found many applications in countries with limited data, particularly in Africa. The website shows an application of the model to the mortality of Seychelles males in 1971-75, with an excellent fit. The website also has the original single and five-year standards. (The latter is just for convenience, as obviously the survival at five-year age intervals is a subset of the single-year standard.)

Tools for Demographic Estimation (TDE) relies heavily on relational models. Instead of using the Brass standard, however, they propose taking as the standard one of the model life tables in the Coale-Demeny system or in the UN family, usually the Princeton West or the UN general family, with life expectancy 60. Their online materials have extensive notes on the choice of a standard and they provide a spreadsheet with the nine standards.

The Modified Logit System

Murray and collaborators noted in 2003 that the original Brass system could be made to fit much better if instead of the logit function they used a different transformation, while maintaining the idea that after transformation the schedule should be a linear function of a standard, so

$$Z(l_x) = \alpha + \beta Z(l_x^s)$$

After examining a large number of life tables they decided to use a modified logit transformation which incorporates corrections based on mortality at ages 5 and 60. Specifically,

$$Z(l_x) = Y(l_x) + \gamma_x \left(1 - \frac{Y(l_5)}{Y(l_5^s)}\right) + \theta_x \left(1 - \frac{Y(l_{60})}{Y(l_{60}^s)}\right)$$

where γ_x and θ_x are constants and l_x^s is a standard survival function, defined for ages 1,5(5)80, with different values for males and females, all chosen to improve the fit of the model to a large training set of life tables.

The γ_x and θ_x coefficients are zero at ages 5 and 60, which makes this transformation identical to the Brass logit at those ages. At other ages the transformation differs from the logit, with age-specific corrections depending on departures from the standard at ages 5 and 60. Note that the transformation of the standard *is* the Brass logit, as the correction terms vanish. Putting everything together the model is

$$Y(l_x) + \gamma_x \left(1 - \frac{Y(l_5)}{Y(l_5^s)}\right) + \theta_x \left(1 - \frac{Y(l_{60})}{Y(l_{60}^s)}\right) = \alpha + \beta Y(l_x^s)$$

Because the model is linear on the parameters it can be fitted by OLS, regressing $Z(l_x)$ on $Y(l_x^s)$. This yields fitted values $\hat{Z}(l_x)$, which can be converted to survival probabilities solving for $Y(\hat{l}_x)$ in the above equation.

(My presentation follows closely the paper. TDE write the model with the Brass logit $Y(l_x)$ on the left-hand-side and everything else on the right, which I think is confusing because the outcome also appears on the right in the guise of $Y(l_5)$ and $Y(l_{60})$. Because they do this, they note that the signs of their coefficients are reversed relative to the paper. In the end the results are, of course, exactly the same.)

The Log-Quadratic System

In 2012 Wilmoth and collaborators published a new 2-parameter system of model life tables. The model is based on a very large set of reliable life tables from the Human Mortality database, which Wilmoth had started in his Berkeley days.

The basic idea is that the log of the mortality rate in the usual abridged life table is a quadratic function of the log of ${}_5q_0$ called h for short, and a second parameter k which controls the shape of the schedule.

$$\log {}_n m_x = a_x + b_x h + c_x h^2 + v_k k$$

where a_x, b_x, c_x and v_k are constants, which were estimated from a large training set of reliable life tables.

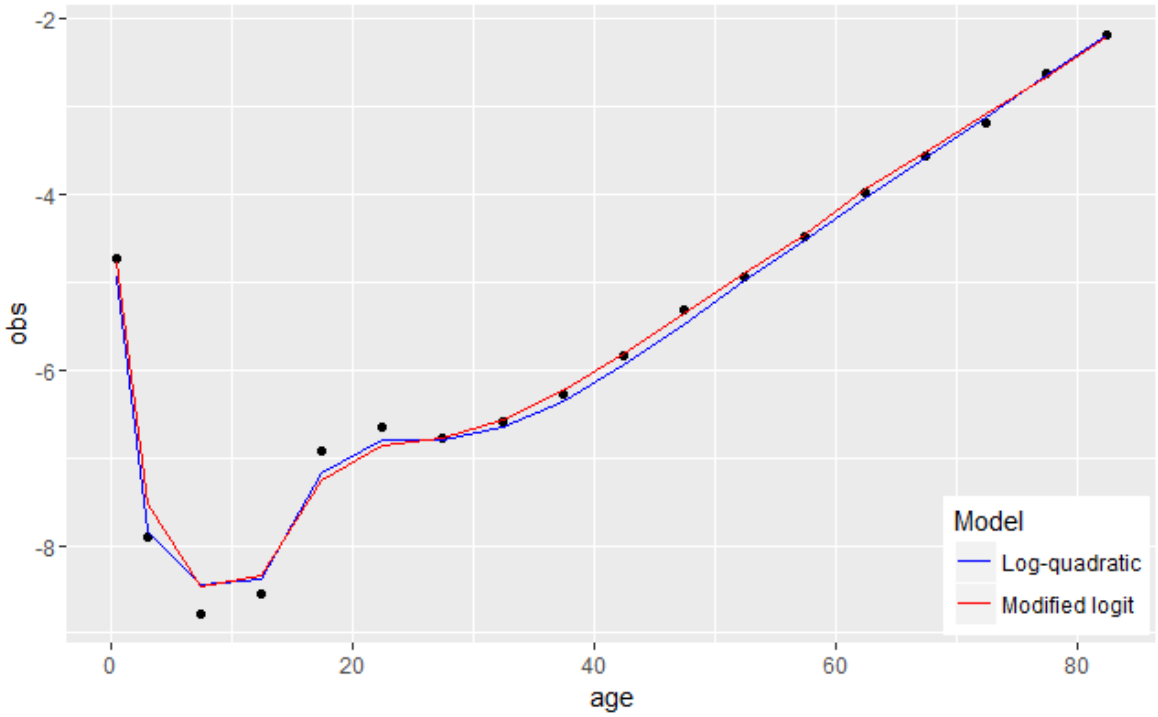
An interesting feature of the model is that all the coefficients for the age group 1-4 are zero, so the equation does not produce an estimate of $\log {}_4 m_1$. The reason is that one of the parameters is the log of ${}_5 q_0$ and the authors obviously care about consistency. So they use the fitted value of ${}_1 m_0$ to produce ${}_1 q_0$, then estimate ${}_4 q_1 = 1 - (1 - {}_5 q_0)/(1 - {}_1 q_0)$ and finally convert this to a rate. These steps use the Coale-Demeny ${}_n a_x$ values that we have encountered already. Because of this extra work, the fitted life table reproduces exacly the ${}_5 q_0$ value used as an input parameter.

I think this model represents the current state of the art and appears to fit well a wide variety of mortality schedules, but the modified logit system is a strong competitor.

All of these models can be used to estimate mortality from limited data. TDE has an application estimating a complete life table for Kenya using only measures of child and adult mortality, namely ${}_5 q_0$ and ${}_{45} q_{15}$.

Austrian Males in 1992

The figure below shows the log-mortality rates in the life table we estimated for Austrian males in 1992, which is Box 3.1 in the textbook, and fitted values based on the modified logit and the log-quadratic models.



As you can see, the models have some difficulty following the rates for teenagers and young adults, but do pretty well at the very young ages, and even better after age 25. In my opinion, it would be very hard to get a better fit without adding a third parameter.

The code used to produce this graph is available in the course website, including functions to evaluate and fit the models.

Kaplan-Meier and Cox

We consider non-parametric estimation of the survival function using cohort data. Specifically, we assume we have observations $t_1, \ldots, t_n$ of survival times as well as indicators $d_1, \ldots, d_n$ that take the value 1 if the observation ended with the event of interest and 0 otherwise.

One-Sample: Kaplan-Meier

If there was no censoring the obvious estimate of the probability of surviving to t would be the empirical survival function

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} I\,(t_i > t)$$

or proportion alive at t.

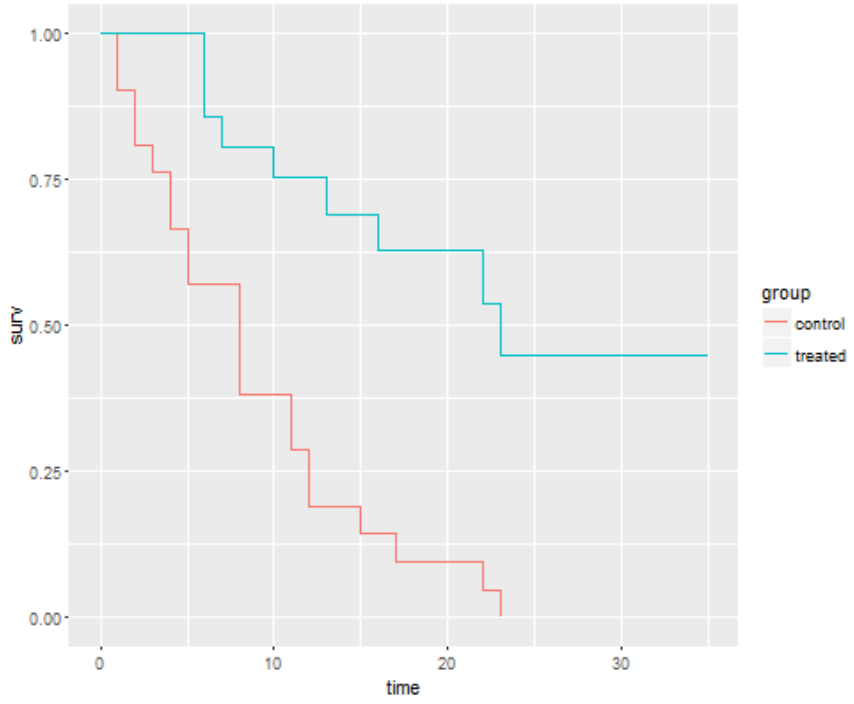
Kaplan-Meier extended the estimator to censored data. They focus on the *distinct ordered* event times (not counting censoring times), which I'll denote $t_{(i)}$. Let d_i denote the number of events at $t_{(i)}$ and n_i be the number alive, and hence at risk, *just before* $t(i)$. The Kaplan-Meier or *product limit* estimate is then

$$\hat{S}(t) = \prod_{i:t_{(i)} \le t} (\,1 - \frac{d_i}{n_i})$$

Note that d_i/n_i estimates the probability of an event at $t_{(i)}$ given the number at risk (like ${}_nq_x$) and one minus that or $1 - d_i/n_i$ is the probability of surviving that failure time conditional on survival up to that point, so the product is an unconditional survival probability up to t (like l_x.)

The estimate is a step function with discontinuities at the observed failure times. If there is no censoring the estimator coincides with the empirical survival function, so it is a generalization for censored data.

In the website we compute Kaplan-Meier estimators for time in remission of leukemia patients in two groups, treated and controls. The figure below shows the estimated survival curves. One group has no censoring and the estimate is just the proportion surviving to each duration; in the end all relapse. In the treated group we note that after 35 weeks almost half the patients remain in remission.



We can compute the standard error of the estimator using the delta method. Briefly, the method approximates the variance of a function of a random variable using a first-order Taylor series expansion, which gives

$$\text{var}(f(X)) \approx [f'(X)]^2 \text{var}(X)$$

In our case $\hat{S}(t)$ is a product, so we first take logs and assume independence of the conditional survival probabilities, so

$$\text{var}(\log \hat{S}(t)) = \sum_{i:t_{(i)} \leq t} \text{var}\,(\log p_i)$$

We estimate the variance of p_i using the binomial formula, so $\text{var}(p_i) = p_i q_i / n_i$ and then use the delta method to obtain

$$\text{var}(\log p_i) = \frac{1}{p_i^2} \text{var}(p_i) = \frac{q_i}{p_i n_i}$$

and then apply the delta method again, this time to go from $\log \hat{S}(t)$ to $S(t)$:

$$\text{var}(\hat{S}(t)) = [\hat{S}(t)]^2 \sum_{i:t_{(i)} \leq t} \frac{q_i}{p_i n_i}$$

This is known as Greenwood's formula and predates the Kaplan-Meier estimator by 32 years! It was first proposed in the context of actuarial life tables for cancer survival in 1926.

Regression: Cox Proportional Hazards

Cox proposed a general solution to the problem of doing regression analysis with survival data without having to make strong assumptions about the shape of the hazard or force of mortality. I will use the standard statistical notation to emphasize the fact that this model has a wide range of applications beyond mortality.

The basic proportional hazards model assumes that

$$\lambda(t,x) = \lambda_0(t)e^{x'\beta}$$

where $\lambda(t,x)$ is the hazard at time t for a subject with covariate values x, $\lambda_0(t)$ is a *baseline* hazard that applies to everyone at time t and $e^{x'\beta}$ is a *relative risk* for a subject with covariates values x compared to a subject with $x = 0$.

A simple example may help fix ideas. Suppose there are only two groups and x takes the value 1 for one group (say, treated) and 0 for the other (say, the control group). Then the model says

$$\lambda(t,x) = \begin{cases} \lambda_0(t), & \text{if } x = 0 \\ \lambda_0(t)e^{\beta}, & \text{if } x = 1 \end{cases}$$

In this case $\lambda_0(t)$ denotes the risk at time t in the control group, and e^{β} denotes the relative risk in the treated group at any given time, compared to the control group at the same time.

There are extensions of the model where the covariates may change over time, of their effects may be non-proportional, or both, but here we will focus on the simpler case.

Cox did not just contribute a model but also a way to estimate it without making any assumptions about the shape of the underlying hazard. Like previous workers, he focuses on the distinct ordered failure times $t_{(i)}$.

Suppose first that there are no ties in the observation times, so one and only one person fails at $t_{(i)}$. Let's call this person $j(i)$. Let R_i denote the risk set, or indices of all subjects alive just before $t_{(i)}$. The probability that the person who failed at $t_{(i)}$ would be $j(i)$ conditional on the risk set is

$$L_i = \frac{\lambda(t_i, x_{j(i)})}{\sum_{j \in R_i} \lambda\,(t_i, x_j)}$$

If we write the risk as the product of the baseline risk times the relative risk, we find that the baseline hazard cancels out and the probability in question becomes

$$L_i = \frac{e^{x_{j(i)}'\beta}}{\sum_{j \in R_i} e^{x_j'\beta}}$$

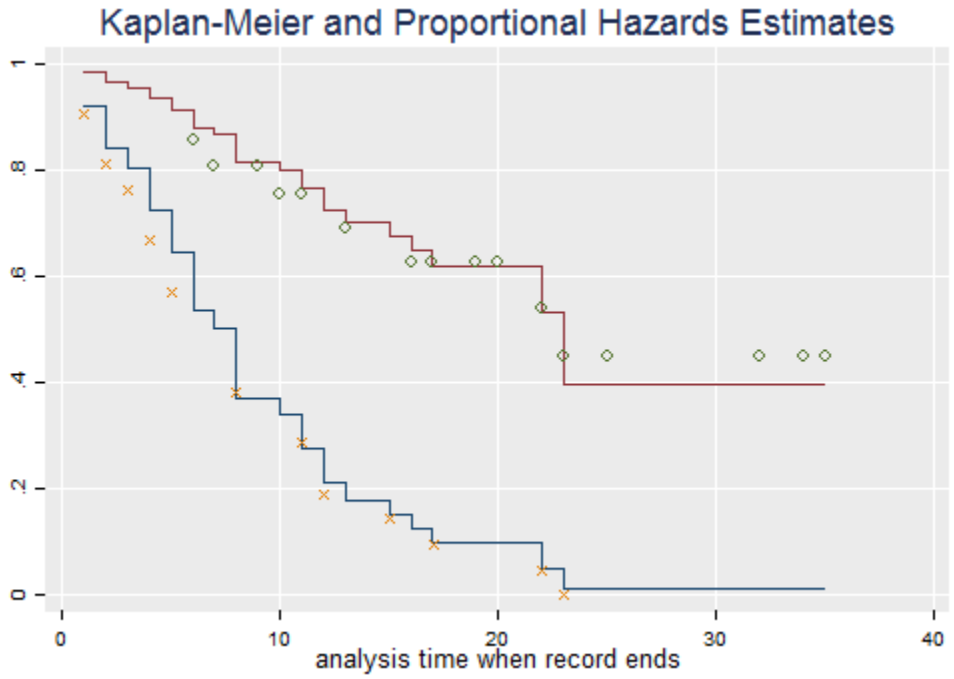
an expression that depends only on β. Cox proposed treating the product of these conditional probabilities over all distinct failure times as if it were a likelihood function, maximizing it to obtain an estimate of β.

The product is known as Cox's *partial likelihood* and the resulting estimator shares many of the optimal properties of maximum likelihood, with a small loss of efficiency compared to making full (and correct!) parametric assumptions.

Calculation of the estimate is more complicated if there are tied failure times. The numerator is simply the product of the relative risks over the d_i who fail, but in principle the denominator requires considering all possible ways of selecting d_i failures out of n_i in the risk set, which may not be feasible. Not surprisingly, there are several approximations. The simplest one is Breslow's, which takes as denominator the sum of relative risks raised to the power d_i. Efron proposed a better approximation that requires only modest computational effort, and can be motivated by breaking the ties.

The website shows how to fit Cox's model to the leukemia remission data. We find a maximum partial likelihood estimate of -1.572 using Efron's method. Exponentiating this estimate we conclude that the risk of relapse is 79% lower in the treated group than in the controls at any duration of remission [exp(-1.572)= 0.208].

It is possible to obtain estimates of the baseline survival function by adapting the Kaplan-Meier logic *after* fitting a Cox model to obtain an estimate of β. The logic involves using the relative risks as weights. The figure below overlays Cox proportional-hazard estimates on the Kaplan-Meier estimates we obtained earlier, showing a good fit.



This is essentially Figure 1 in Cox's original paper. An alternative diagnostic plot in the log-log scale is shown on the website

Unobserved Heterogeneity

POP 502 / ECO 572/ SOC 532 • SPRING 2017

The life table summarizes the experience of a population. This is not representative of the experience of individuals unless these are homogenous. This unit deals with the consequences of unobserved heterogeneity, with an application to the mortality cross-over in Black and White mortality in the U.S. A classic reference is Vaupel, Manton and Stallard (1979).

The Multiplicative Frailty Model

A popular approach to modeling unobserved heterogeneity assumes that the hazard $\mu_i(x)$ for individual i at age x is the product of two terms, a baseline hazard $\mu_0(x)$ and a multiplicative term θ_i representing the individual's frailty, so

$$\mu_i(x) = \mu_0(x)\theta_i$$

A person with $\theta = 1$ represents the baseline risk. A person with $\theta = 1.5$ has 50% *higher* risk than our reference individual *at every age.* A person with $\theta = 0.5$ has 50% *lower* risk than the reference individual. The formulation is just like a proportional hazards model, except that we don't observe a person's frailty.

Let $p_i(x)$ denote the probability that individual i will survive to age x,

$$p_i(x) = \Pr\{X > x\} = \frac{l_i(x)}{l_i(0)}$$

This is just $l_i(x)$ if the radix is 1. From our results relating survival probabilities to hazards we have

$$p_i(x) = p_0(x)^{\theta_i}$$

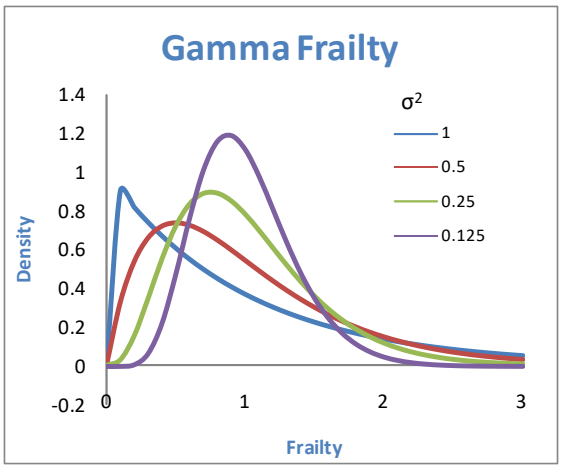
where $p_0(x)$ is the baseline survival probability. This follows from writing the survival probability as $p_i(x) = \exp\{-\int_0^x \mu_i(a)da\}$ and then substituting the model for the individual hazard. So if the reference individual has an 80% change of living to age 60, one with $\theta = 1.5$ has only a 72% chance, whereas one with $\theta = 0.5$ has an 89% chance.

Gamma Frailty

The next step is to assume that frailty has a distribution in the population. A common assumption is to postulate a gamma distribution, which has density

$$g(\theta) = \frac{\theta^{\alpha-1}e^{-\beta\theta}\beta^{\alpha}}{\Gamma(\alpha)}$$

with parameters α and β . The mean is α/β and the variance is α/β^2, so we often set $\alpha = \beta = 1/\sigma^2$ to get a mean of one and a desired variance. The figure on the side shows gamma densities with mean one and variances of one, a half, a quarter and 1/8.



One could use other distributions for frailty, and results have been obtained for discrete mixtures and for cases where frailty has an inverse Gaussian or a compound Poisson distribution, but gamma is by far the most popular choice.

Population Average Survival

The survival function we estimate with a life table is an average for individuals with different frailties. Suppose the entire population consisted of just the three individuals in the initial example. Then the average probability of living to age 60 would be 80.3%, the average of 80, 72 and 89.

Of course a population will have more than three individuals, so we average using the distribution of frailty. The average survival probability in the population is then

$$p(x) = \int_0^\infty p_0(a)^\theta g(\theta) d\theta$$

In general this is not the same as the baseline. We call $p(x)$ the *population-average* survival.

If frailty has a gamma distribution with mean one and variance σ^2, then with a bit of algebra one can show that the population survival is given by

$$p(x) = \frac{1}{[1 + \sigma^2 H_0(x)]^{\frac{1}{\sigma^2}}}$$

where $H_0(x) = \int_0^x \mu_0(a) da$ is the integrated baseline hazard. This is a Pareto distribution of the second kind.

Survival functions are useful but not terribly informative, so I turn attention to the hazard.

Population Average Hazard

To compute the population hazard we proceed from first principles, taking the negative log of the survival probability to obtain a cumulative (or integrated) hazard and then differentiating to obtain the hazard. If we follow that approach it can be shown that

$$\mu(x) = \mu_0(x) E(\theta | X > x)$$

where $E(\theta | X > x)$ is the expected value of frailty among survivors to age x.

A more specific result can be obtained if we assume that frailty at birth has a gamma distribution with mean one and variance σ^2. In that case the mean frailty of survivors to age x is

$$E(\theta|X > x) = \frac{1}{1 + \sigma^2 H_0(x)}$$

where $H_0(x)$ is the integrated baseline hazard, as before. Thus, under gamma frailty the population average hazard is

$$\mu(x) = \frac{\mu_0(x)}{1 + \sigma^2 H_0(x)}$$

At birth mean frailty is one and the population average hazard is the same as the baseline individual hazard. As time goes by, however, the mean frailty of survivors declines, becoming less than one, and as a result the population average hazard becomes lower than the baseline individual hazard. This can be seen from the fact that the integrated hazard, which increases with age, is in the denominator of the formulas for the mean frailty and the population average hazard.

Our interpretation is this result is that the frail tend to die first, so over time the population becomes increasingly selected, consisting of individuals who are more robust. Note also that frailty declines faster (so selection operates more quickly) when the population is more heterogeneous to start with (larger σ^2) or the risk is higher (larger baseline hazard $\mu_0(x)$ and hence larger $H_0(x)$).

Example: To fix ideas consider a situation where the hazard is constant over time for each individual but there is heterogeneity of frailty. Specifically suppose the individual hazard is $\mu_0\theta$, where μ_0 is the baseline hazard and θ denotes frailty. If frailty has a gamma distribution then the population hazard is

$$\mu(x) = \frac{\mu_0}{1 + \sigma^2 \mu_0 x}$$

and declines from μ_0 at birth to zero as $x \to \infty$. It will decline faster for larger μ_0 or larger σ^2.

A constant hazard model doesn't work well for mortality but it approximates other situations, such as time to conception among fecund women trying to conceive a child. Assume that each woman's fecundability is constant over time, at least for a few months, but women differ in their fecundability. According to these results the population hazard would decline over time even though it's constant for each woman. This occurs because more fecund women tend to conceive first, and the survivors become increasingly selected for lower fecundability.

When frailty at birth has a gamma distribution one can show that the distribution of frailty among survivors to age x is also gamma with the mean given above and variance

$$var(\theta|X > x) = \frac{\sigma^2}{\left(1 + \sigma^2 H_0(x)\right)^2}$$

Note that the variance at birth is σ^2 but over time the variance get smaller and smaller, so the population becomes more homogeneous. Frailty, of course, also declines. An interesting feature of gamma frailty is that the coefficient of variation (standard deviation over mean) remains constant.

The Inversion Formula

So far we have gone from individual to population hazards. Can we go the other way? The answer is yes, and leads to interesting applications.

If frailty has a gamma distribution then one can show that the baseline hazard satisfies

$$\mu_0(t) = \mu(x)e^{\sigma^2 H(x)}$$

where $H(x) = \int_0^x \mu(a)da$ is the cumulative (integrated) population hazard. (The negative log of the population survival function with radix 1.)

Example. We considered earlier how a gamma mixture of exponentials leads to a declining population hazard. I now show that a constant population hazard can be viewed as a mixture of something else. If the population hazard is constant then $\mu(x) = \mu$ and the cumulative hazard is $H(x) = \mu x$. Plugging these functions into the inversion formula we find that the baseline individual hazard is

$$\mu_0(x) = \mu e^{\sigma^2 \mu x}$$

an exponential function of x which we recognize as a Gompertz hazard. Thus, we have the remarkable result that a population that shows a constant hazard may result from individuals with gamma distributed heterogeneity who face hazards that increase exponentially with time.

The Identification Problem

You may begin to suspect that we have a bit of an identification problem here, because a flat population hazard could also result from a homogeneous population where each individual's hazard is flat. All we can estimate is hazards for populations, or groups. It pays to be aware, however, that the hazards for the individuals may be different. In particular, we can't distinguish heterogeneity from negative duration dependence.

The Mortality Cross-Over

The online computing logs illustrate these ideas with an application to U.S. mortality. We start with population average survival and hazard curves for blacks and whites and note the well-documented mortality cross-over. We then use the inversion formula to find subject-specific hazards for blacks and whites that do not cross, yet under heterogeneity lead to population-average hazards that do cross. The underlying explanation is that blacks face higher mortality at younger ages and hence become more highly selected at older ages. The alternative explanation is age misreporting, which may be particularly prominent among older blacks because of the lack of birth certificates.

Competing Risks

POP 502 / ECO 572/ SOC 532 • SPRING 2017

The textbook has a good discussion of multiple decrement life tables and we reproduce online the calculations in Boxes 4.1 and 4.2. Here is just a quick summary of some of the main ideas.

Continuous Time Formulation

We assume that there are J causes of failure and that every failure can be attributed to one and only one of these causes. We define a cause-specific hazard or force of mortality

$$\mu_j(x) = \lim_{dx \to 0} \frac{1}{dx} \Pr\{X \in (x, x+dx) | X > x \text{ and } J = j\}$$

As the limit of the probability of death due to cause i in a small interval after x given survival to x.

Using the law of total probability, the total force of mortality is a simple sum

$$\mu(x) = \sum_j \mu_j(x)$$

From the hazard we can calculate overall survival the usual way. We can also compute the probability of dying by age x due to cause j, also known as the *cumulative incidence* function

$$I_j(x) = \Pr\{X \leq x \text{ and } J = j\} = \frac{1}{l(0)} \int_0^x l(a)\mu_j(a)da$$

The integrand reflects the probability of surviving *all* causes up to age a times the conditional probability of then dying due to cause j. All of these functions reflect what happens when all causes are competing or acting at the same time.

One may also calculate a "survival function" based on cause j alone, or based on all causes other than j, but these are counterfactuals predicated on the assumption of independence of the underlying risks. Unfortunately this assumption cannot be verified empirically.

The Multiple Decrement Life Table

The additional data are simply counts of deaths by age due to cause j, say ${}_nD_x^j$, which add up to all deaths at that age. Let ${}_nR_x^j = {}_nD_x^j / {}_nD_x$ denote the proportion of deaths at ages x to $x+n$ due to cause j. The exposure remains the same, as everybody is exposed to all causes at any given time. We then calculate a life table using *all* causes of death as usual, but then add a few columns:

$${}_nq_x^j = {}_nq_x \ {}_nR_x^j$$

is the conditional probability of death at ages $x, x+n$ due to cause j, and

$$ {}_nd_x^j = l_x \, {}_nq_x^j = {}_nd_x \; {}_nR_x^j $$

Is the number of deaths at ages $x, x+n$ due to cause j. Accumulating counts of deaths and dividing by l_0 estimates the incidence function, or cumulative probability of death due to cause j. The textbook also defines (in my opinion in a slight abuse of notation) l_x^j as the number of people age x who will eventually leave the life table due to cause j, obtained by accumulating ${}_nd_x^j$ from age x onwards.

The example in Box 4.1 distinguishes female deaths due to neoplasms and all other causes and estimates that 21.2% of female births will die of neoplasm if 1991 rates were to prevail.

Associated Single-Decrement Life Tables

Associated with cause j is a cause-specific force of mortality $\mu_j(x)$. The associated single-decrement life table attempts to estimate what survival would look like if this was the only cause operating. The honest answer is that we don't know, the question is a *counterfactual* that requires the strong assumption that eliminating the other causes would leave $\mu_j(x)$ unchanged. This is equivalent to assuming independence of the underlying risks.

Then there is a technical problem concerning how to construct the life table. The textbook discusses three approaches.

1. We can calculate a cause-specific rate ${}_nM_x^j$ dividing deaths dues to cause j by total exposure, equate that to the life table rate ${}_nm_x^j$ and proceed "as usual", making assumptions about ${}_na_x$, which may of course be different from the assumptions made for the overall table.

2. Assume that cause-specific risks are constant within an age group and convert the rates to probabilities accordingly. This is my preferred approach. The textbook notes that is "logically consistent and easy to apply" and preferable when the assumption is tenable, which of course is only true for small age intervals.

3. Follow Chiang in assuming that the cause-specific force of mortality is proportional to the overall force of mortality in the age interval x to $x+n$, so $\mu_j(a) = R_j \, \mu(a)$ when $a \in (x, x+n)$. The proportionality factor is estimated using ${}_nR_x^j$, the proportion of all deaths in the age interval that are due to cause j. This is a proportional hazards assumption and leads to estimating the probability of surviving the interval when only cause j is operating as

$$ {}_np_x^j = {}_np_x^{\,{}_nR_x^j} $$

From these survival probabilities we can get the survival function as well as the conditional probabilities of death. Calculating time lived, which we need to get expectation of life, requires assumptions about ${}_na_x$. A common solution is to keep the values in the full life table, but the textbook notes that if only one cause is operating the age distribution in the interval will be younger, and recommends a graduation approach, using formulas 4.6 and 4.8. The former is based on fitting a quadratic to age at death over three adjacent intervals. The latter interpolates between two extremes based on the cases when 0 or 100% of the deaths in the interval are due to cause j.

IMHO the differences between these approaches are relatively minor and pale in comparison to the real problem, which is not knowing how the force of mortality for cause j would change if all other causes were eliminated.

Cause-Deleted Life Tables

This is formally exactly the same problem, but instead of assuming $\mu_j(x)$ is the only force operating, we assume that this cause has been eliminated, but all *other* causes continue to operate as before. In other words the overall force of mortality is now $\mu(x) - \mu_j(x)$. A cause-deleted life table is thus exactly the same as the associated single-decrement life table for all other causes.

Exactly the same issues noted above apply. The key assumption is that eliminating a cause of death would leave the other forces of mortality unchanged. The life table would change, of course, but only because exposure would change.

The website reproduces the calculations in Box 4.2 using Chiang's method and confirms that if neoplasm were eliminated but the rates for all other causes remained unchanged, life expectancy would increase from 78.92 to 82.46 years, a gain of 3.54 years.

I also show that using instead the simpler assumption that all forces of mortality are constant within each age interval leads to almost identical estimates.

It is worth noting that if there is heterogeneity and people at high risk of cancer are also at high risk of dying of other causes, the increase in life expectancy might be less than calculated. On the other hand if those survivors benefit from lower future death rates for other causes the gain might be larger than estimated. In short, the calculation is just a counterfactual based on the assumption that eliminating some causes of death doesn't change others and all age-specific death rates remain the same.

Current Status Life Tables

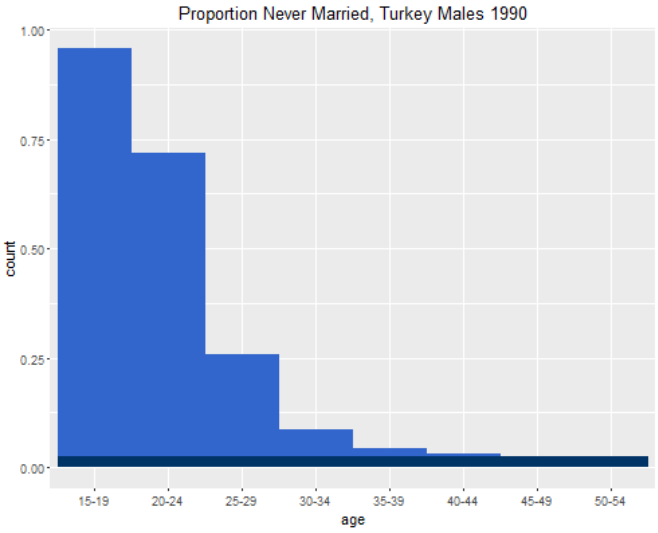
POP 502 / ECO 572/ SOC 532 • SPRING 2017

Singulate Mean Age at Marriage

Hajnal (1953) proposed a method for estimating the distribution of age at marriage from proportions single by age. The idea is that in a closed population with no mortality the proportion of women who remain single at age x is a direct estimate of $l(x)/l(0)$. Summing proportions single over age we obtain an estimate of time lived in the single state. From that we can obtain mean age at marriage by analogy with expectation of life. The only slight complication is that not everyone marries. The solution is quite simple, we rescale the survival function so it applies only to those who will marry, using

$$l^*(x) = \frac{l(x) - l(\infty)}{1 - l(\infty)}$$

In the website I go through the exercise on page 90 of the textbook, working with proportions single among Turkish men in 1990. I start at age 15 because no-one marries before then. The time lived single between ages 15 and 50 is just 5 times the sum of proportions single: 10.59 years. The proportion that remains single by age 50 is estimated at 2.33% by averaging the last two age groups. The mean age at marriage of those who marry by age 50 is then estimated as



$$15 + \frac{10.59 - 35{\times}0.0233}{1 - 0.0233} = 25.004$$

We start with 10.59 and subtract the time spent single by those who don't marry by age 50, namely 35×0.0233 = 0.816 years, scale up dividing by the proportion who marry by age 50, and add the 15 years everyone spends single from birth to age 15. In this example the adjustment makes little difference because almost every one marries by age 50.

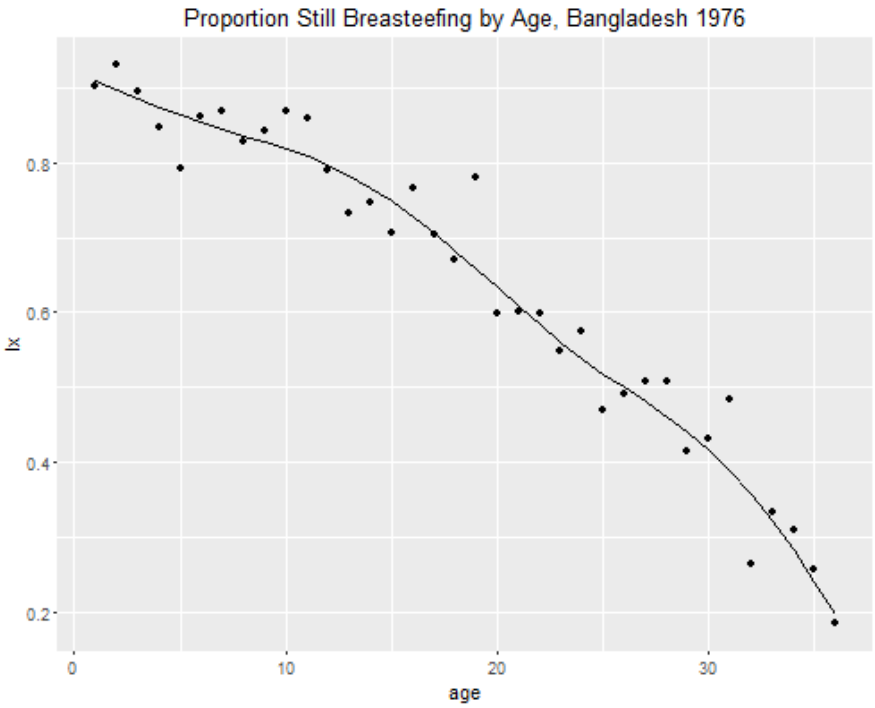
The resulting estimate is called the *singulate mean age at marriage* or SMAM. It does not represent the conditional mean age at marriage of a real cohort unless nuptiality has remained constant over the last 35 years of so. Even without that interpretation, however, it is a useful summary of period nuptiality, as the unscaled version represents the total time lived in the single state in a given period. For example in 1990 Turkish men spent 10.59 of the 35 years from ages 15 to 50 single. If people delay or forego marriage in a calendar period this indicator would be expected to increase.

Duration of Breastfeeding

Current status life tables are particularly useful in the study of duration of breastfeeding. One could, of course, build a cohort life table using retrospective reports of breastfeeding duration, but these are notoriously unreliable. We illustrate these ideas using data from the World Fertility Survey (WFS) of Bangladesh, conducted in 1976.

The WFS asked questions on breastfeeding for the last and next to last births only, and coded the answers in terms of the last closed and the open birth intervals. A life table based on retrospective reports of breastfeeding duration would combine both pieces of information, restricting the analysis to births in the last three years or so to ensure a representative sample of births. As we have seen earlier, the raw data show very substantial heaping on multiples of 12, and to a lesser extent some multiples of six. This could represent a real tendency for women to wean their children after achieving a milestone such as age two, but it could also represent bad data with substantial rounding to whole years.

An alternative approach is to tabulate whether the child is still being breastfed or not by current age. This is a direct estimate of $l(x)/l(0)$, with the estimate at very young ages representing the proportion who are ever breastfed. The figure below shows a current status life table by single months of age, based on all births in the 36 months before the survey. If the birth happened to be the last one I got breastfeeding status from the open interval data, otherwise I assumed it was weaned.



The first thing to note is that there is no precipitous drop in the proportion of children being breastfed around age two, as you would expect if the increased risk of weaning at this milestone was real. The other problem is that with single-month data we have small sample sizes and the proportions still breastfeeding are erratic, and fail to decline monotonically as any decent survival function should.

There is an algorithm called *pool adjacent violators* that obtains a monotone estimate by averaging successive entries that violate the monotonicity constraint. An alternative solution is to smooth the estimates, as I have done in the figure by using a regression spline with internal knots at 12 and 24 months. Note that the proportion still breastfeeding at age x completed months is attributed to exact age $x + 1/2$.

We can also compute mean duration of breastfeeding for all births by simply summing the proportions still breastfed by age, which is equivalent to computing time lived in the breastfeeding state. In our example the area under the curve is 23.0 months using the raw data and 22.8 months using the spline. If one wanted to compute the mean only for children who are breastfed one could adjust the survival curve dividing by the proportion ever breastfed. Assuming 95% breastfeed the mean would be 24.0 months.

Incidence-Prevalence

If all one wanted is the mean there is an even simpler estimate called the incidence-prevalence estimator. This is in common use in epidemiology, where one can approximate the duration of a disease in months dividing the number of new cases per month (incidence) by the number of existing cases at a given time (prevalence). Treating breastfeeding as a "disease" we would divide the number of births per month by the number of children being breastfed at the time of the survey. If the number of births is relatively constant over time this is exactly the same as summing the proportions ever breastfed by age. We show online that the incidence-prevalence estimate, computed as the overall proportion sill breastfeeding over 0.95/36, an estimate of "new cases" per month, is 23.8 months.

Right and Left Censoring

Note in closing that with current status data *all* observations are censored at their current age:

- children still breastfeeding are *right* censored; all we know is that they will breastfeed longer than their current age, and
- children weaned are *left* censored; all we know is that they were breastfeed less than their current age.

Yet we can still estimate a survival curve!

Age at Marriage

Pop 502/ ECO 572/ SOC 532 • SPRING 2017

We consider models for age at first marriage proposed by Coale (1971) and by Hernes (1972), with an application to predict the "future" of first unions in Colombia. The textbook discusses the first of these models in Section 9.2. We start by introducing some common notation treating age as continuous.

Notation

Let $F(a)$ denote the proportion ever married by exact age a, so $F(\infty)$ is the proportion who ever marry. The derivative $f(a) = F'(a)$ can be described as the marriage density (or frequency of first marriages) at exact age a. The complement $1 - F(a)$ is the proportion who remain never married by exact age a, and is analogous to the survival function $l(x)/l(0)$. Dividing first marriage frequencies by the proportion single we obtain the hazard of first marriage $\mu(a) = f(a)/(1 - F(a))$.

Coale-McNeil

Coale (1971) examined first marriage frequencies in a number of countries and discovered that if he adjusted them for the proportion who eventually marry and for the location and scale of age at marriage they all had almost exactly the same shape (see his Figures 3 and 4). He used reliable data from Sweden to represent this common shape, leading to the model

Here c is the proportion who ever marry and $G_s()$ is the Swedish schedule of proportions ever married by age a among women who eventually marry. To map age in the population of interest to age in the Swedish standard you subtract a_0 and divide by k. The parameter a_0 is described as the age by which a "consequential" number of marriages first occur, and k represents the "pace" of marriage relatively to the Swedish standard. For example if $k = 1$ people marry just as fast as in Sweden, but if $k = 2$ they marry more slowly, taking two years to achieve the same proportions married that Sweden achieves in just one year.

$$F(a) = c\, G_s\left(\frac{a - a_0}{k}\right)$$

In later work Coale and McNeil (1972) found an analytic expression that fits the Swedish standard very well and turned out to be related to the gamma distribution. They also showed that to a very close approximation the model could be described as consisting of the sum of a normally distributed random variable and three exponential waiting times. The first component could represent the age at which one enters the marriage market and the delays could represent the time needed to find a suitable partner, the length of the courtship, and the length of the engagement period. Data from a 1959 survey in France provided some partial support for this behavioral interpretation.

In work I did with Trussell developing methods to fit this model to survey data by maximum likelihood we used the mean μ and standard deviation σ instead of a_0 and k, so that

$$F(a) = c\, G_0\left(\frac{a - \mu}{\sigma}\right)$$

where $G_0(z)$ is a standardized schedule with mean zero and variance one. This yields parameters that are easier to interpret. If $z = (a - \mu)/\sigma$ denotes standardized age the schedule $G_0(z)$ can be computed in terms of the c.d.f. $\Gamma()$ of the gamma distribution as $1 - \Gamma(\exp(-1.896\ z + 0.805)\ , 0.604)$ in R, Stata or Excel.

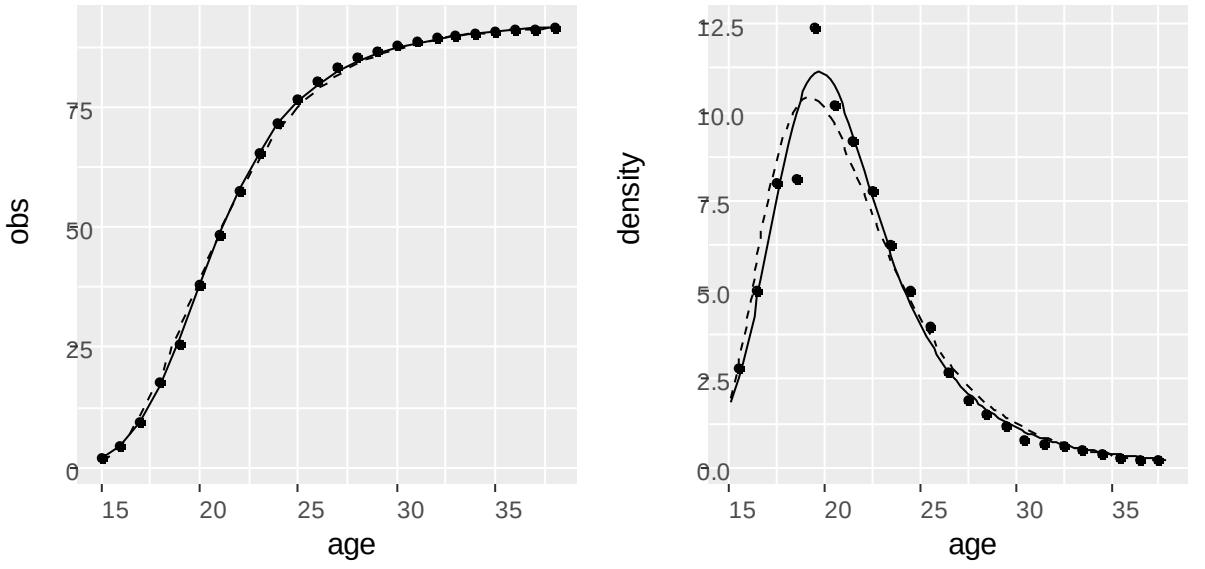


Figure 1 Coale-McNeil and Hernes Nuptiality Models Fit to U.S. Women born in 1920-24

Figure 1 shows the results of fitting the Coale-McNeil model as well as Hernes's model (discussed below) to the 1920-24 cohort of U.S. white women, the data used in Hernes's original paper. The fits to the cumulative schedule are extremely close, but the fits to the first marriage frequencies differ around age 20. We estimate that 92.2% of the cohort ever marry, and that age at marriage has a mean of 21.64 and a standard deviation of 4.56 among those who marry.

Hernes

Hernes (1972) proposed a model that has an interesting behavioral basis in terms of a diffusion process. Specifically, he postulates that people marry as a result of (1) social pressure, which increases as more and more people in a cohort have married, and (2) a person's own attractiveness, which regrettably declines exponentially with age. He develops the model in terms of a differential equation, but I find it easier to think in terms of the hazard of marrying, which is simply the product of the two influences:

$$\mu(a) = A\, e^{-ra} F(a)$$

Actually he writes b^a where I write e^{-ra}, but this is the same thing with $b = e^{-r}$ and $r = -\log(b)$. My notation is intended to remind you of the population growth equation and the exponential survival curve. Basically the model assumes that we lose attractiveness at a constant rate r per year. We often measure age from 15 so A represents attractiveness at that age.

A minor drawback of the model is that you must start the process with someone already married, otherwise there's no social pressure to marry and nothing ever happens. To see this point note that if

$F(a)$ is zero then the hazard is zero and nobody marries. This is not a serious issue in practice; the model usually predicts a small but non-zero probability of being married at very young ages.

The hazard is the ratio of the density $f(a) = F'(a)$ to the survival $1 - F(a)$, so we can write the model as (his equation 7)

$$F'(a)\, Ae^{-ra}F(a)(1 - F(a))$$

Solving this differential equation requires a boundary condition that can best be expressed in terms of the proportion who eventually marry, $F(\infty)$, which for consistency with the previous model I will denote c. The resulting solution looks rather complicated (see his equation 10 or the somewhat simpler forms 12 and 13), but can be simplified drastically by taking logits to obtain:

$$\text{logit}\big(F(a)\big) = \text{logit}(c) - \frac{A}{r}e^{-ra}$$

Hernes illustrates his model by fitting it to data from the U.S. for white women born in 1920-24, with the results shown in Figure 1. He gets an attractiveness parameter of 1.046 at age 15 with a decay rate of 15.7% per year (so his $b = 0.855$) and 92.99% eventually marrying. (I get a slightly better OLS fit with $A = 0.980$, $r = 0.148$ and $c = 0.936$, and a maximum likelihood fit with $A = 1.01$, $r = 0.152$ and $c = 0.931$.)

Age at Marriage in Colombia

The next application comes from a paper I wrote with Trussell years ago, using data from the Colombian World Fertility Survey conducted in 1976. We were particularly interested in checking how well models would predict future marriages. We focused on the cohort aged 35-39 at interview, which had already gone through the prime marrying ages and fitted the Coale-McNeil model. We then repeated the fit using the data we would have had if we had interviewed those cohorts 15 years earlier, when they were only 20-24. Later I fitted the Hernes model to the same data. Figure 2 depicts the fits

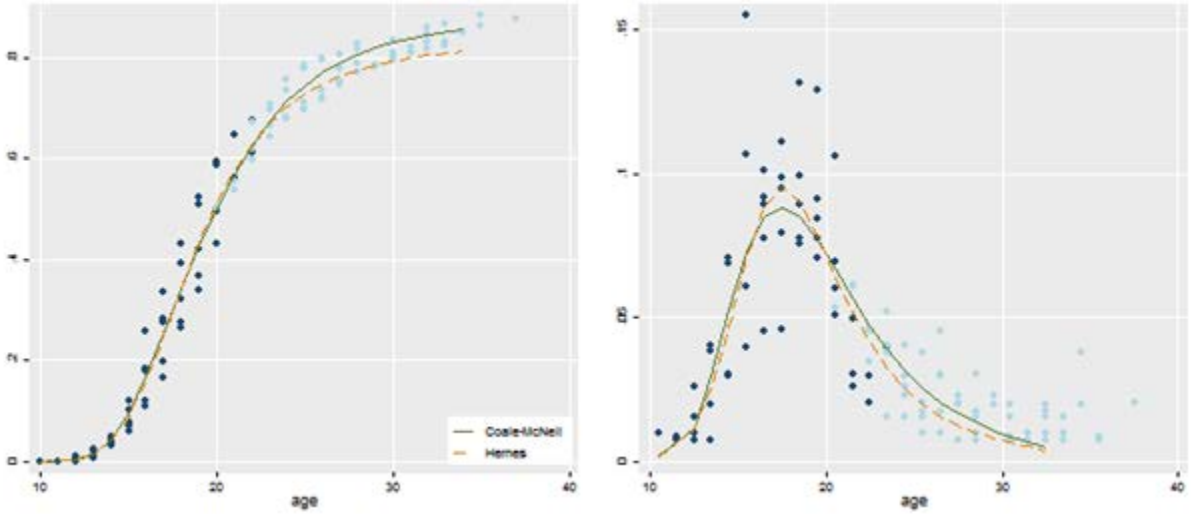


Figure 2. Proportions Ever Married and Marriage Frequencies in Colombia Fitted at Ages 20-24

The bottom line is that the Coale-McNeil did a pretty good job predicting the course of marriage for these cohorts. The Hernes model adapted better to the younger ages, but didn't predict the future

quite as well.

The table that follows shows the parameter estimates used in the figure:

Model	Parameter	Cohort at 35-39	15 Years Earlier
Coale-McNeil	Mean	20.44	20.15
	St. Dev	5.38	5.10
	Pem	0.885	0.874
Hernes	Attractiveness	0.652	0.732
	Decay	0.151	0.188
	Pem	0.887	0.830

Please refer to the computing logs in the course website for code showing how to reproduce the graphs. The functions used to compute the models are available in Stata and R in a package called nuptfer that includes a few classic nuptiality and fertility models.

In Stata type `net from http://data.princeton.edu/eco572/stata` and follow the instructions to install the nuptfer package. The documentation is at http://data.princeton.edu/eco572/nuptfer.html.

In R you need to install Hadley Wickman's devtools package using `install.packages("devtools")`, and then install nuptfer from GitHub `using install_github("grodri/nuptfer")`. The package includes documentation.

Fertility and Reproduction

POP 502/ ECO 572/ SOC 532 • SPRING 2017

This topic is covered in some detail in Chapter 5 of the textbook. Here is a summary of the main ideas. (Separate notes deal with birth intervals and the proximate determinants of fertility.)

Period Fertility

Make sure you are familiar with the standard period measures, the Crude Birth Rate (CBR) or births per 1000 population (or person-years), the General Fertility Rate (GFR) or births per 1000 woman-years in the reproductive ages (usually 15-44), and the Total Fertility Rate (TFR) equal to the sum of single-year age-specific fertility rates (or five times the sum of 5-year age-specific fertility rates) divided by 1,000.

The TFR is interpreted as the average number of children a (synthetic) cohort of women would have if it went through life having children at the current age-specific fertility rates. It can also be interpreted as a standardized GFR, where the standard age distribution is uniform (has the same number of women at every age). The actual GFR uses the age distribution as weights.

The Total Marital Fertility Rate (TMFR) is constructed in a similar fashion using age-specific marital fertility rates, or marital births per 1000 married woman-years in each age group. It represents the expected number of children a woman would have if she married at 15 and stayed married through the reproductive years having children at current marital fertility rates. A much better measure would use duration of marriage rather than age as the clock, but the requisite data are rarely available.

The Princeton Fertility Indices

In his study of fertility in historical European population Coale wanted to compare fertility rates adjusted for age, but the required age-specific fertility rates were not always available. He did, however, have the age distributions. So he adopted an indirect approach. He used as standard a set of rates from the Hutterites, a group with very high fertility, and computed an index of fertility defined as the ratio of observed to expected births if all women had children at Hutterite rates:

$$I_f = \frac{B}{\sum H_i W_i} = \frac{\sum F_i W_i}{\sum H_i W_i}$$

where B is the number of births, F_i stands for age-specific fertility rates (with H_i for the Hutterite rates) and W_i is the number of women in the age group. The analogy with the SMR should not go unnoticed.

Assuming that births occur only within marriage, Coale was able to decompose I_f into an index of marriage I_m and an index of marital fertility I_g, such that

$$I_f = I_m I_g$$

The index of marital fertility is defined in a form analogous to the index of general fertility as

$$I_g = \frac{\sum F_i^L W_i^L}{\sum H_i W_i^L} \text{ or } \frac{B}{\sum H_i W_i^L}$$

where F_i^L is the age-specific marital fertility rate and W_i^L is the number of married women in that age group (the L stands for legitimate). If there is no extra-marital fertility $B = B^L = \sum F_i^L W_i^L$, the numerator of both I_g and I_f. The index of marriage is defined as

$$I_m = \frac{\sum H_i W_i^L}{\sum H_i W_i}$$

and is essentially a weighted average of proportions married by age using the Hutterite fertility rates as weights. (Notation might be clearer if we used M_i instead of W_i^L for married women in age group i.)

Box 5.2 in the textbook illustrates the calculations for the French village of Tourouvre-au-Perche in 1801, for which $I_f = 0.364, I_g = 0.70$ and $I_m = 0.52$. Fertility was 36.4% of what it would be if all women had births at Hutterite rates. This was due in part to lower marital fertility, as births were only 70% what they would be if married women had children at Hutterite rates. The main explanation, however, lies in the proportions married by age, which amount to only 52% when weighted by Hutterite rates. (The fertility decline in Europe turned out to be driven largely by marriage delays.)

Reproduction

The term reproduction is used to denote single-sex fertility rates, usually female births to women. The corresponding age-specific rates are called *maternity* rates, and are computed with only female births in the numerator. The textbook uses ${}_nF_x^F$ to denote the maternity rate for ages x to $x + n$, and $m(a)$ to denote the maternity function in continuous time.

The Gross Reproduction Rate (GRR) is the female equivalent of the TFR, essentially a sum of age-specific *maternity* rates, which is interpreted as the average number of daughters a woman would have if she went through the reproductive span having daughters at current maternity rates. In continuous time, if the reproductive years go from ages α to β the GRR is

$$GRR = \int_\alpha^\beta m(a)da$$

A better measure of reproduction is the Net Reproduction Rate (NRR), which multiplies each maternity rate ${}_nF_x^F$ by the probability of surviving to that age, using ${}_nL_x^F/nl_0$ from an appropriate female life table. We interpret NRR as the average number of daughters a newborn woman would have if she was subject to the observed survival probabilities and maternity rates. In continuous time

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

where $p(a) = l(a)/l(0)$ is the probability of surviving from birth to exact age a.

Obviously the NRR must be one for the female population to replace itself, in which case we say that fertility is at *replacement level*. (This doesn't say anything about the age pattern of fertility, as many schedules can lead to replacement level.) We'll now see what this level implies in terms of more common measures, namely the GRR and the TFR.

If every woman survived from birth to the end of the reproductive years the GRR and NRR would be equal. Coale has shown that to a good approximation

$$NRR = p(A_M)GRR$$

where A_M is the mean age of the maternity schedule, a weighted average using the maternity function as the weights:

$$A_M = \frac{\int a\, m(a)da}{\int m(a)da}$$

In practice A_M is computed as a weighted sum, working with the midpoints of the age groups $x + \frac{n}{2}$ and with weights given by the maternity rates ${}_nF_x^F$.

Using this approximation, replacement level fertility corresponds to $GRR = 1/p(A_M)$. If the sex ratio at birth does not depend on the age of the mother and there are 2.05 total births for each female birth, then $TFR = 2.05\ GFR$ and the replacement level TFR is approximately $2.05/p(A_M)$. In most developed countries $p(A_M)$ is pretty close to one and the replacement level TFR is about 2.1. In high mortality countries many women die before reaching the mean age of childbearing and the replacement level TFR can be much higher. Espenshade, Guzmán and Westoff noted surprising global variation in replacement fertility, ranging from less than 2.1 to nearly 3.5.

Box 5.5 in the textbook shows that in the U.S. in 1991 the GRR was 1.013 daughters per woman and the NRR was 0.995, so fertility was below replacement level. The fact that the two indices are so similar indicates a fairly high probability of surviving to the mean age of childbearing, in fact about 98.2%. The mean age of the maternity schedule is not shown in the textbook, but can easily be verified to be 26.5.

A final note of caution: an $NRR > 1$ indicates that the population is growing, as each cohort of women leaves behind a larger cohort of daughters, but it doesn't tell us how fast it is growing. As we'll see, this also depends on when they have their daughters, or more precisely on the mean length of a generation.

Revised 3/29/2017

Fertility Models

POP 502 / ECO 572/ SOC 532 • SPRING 2017

We now discuss age patterns of fertility, with a brief review of the Henry, Coale, Coale-Trussell, Page, Brass and Schmertmann models. The textbook deals with this subject in Section 9.3.

Natural Fertility and Control

Henry discovered that many natural fertility populations (where there is no conscious attempt to control the number of children) differed in the level of fertility but had a similar age profile. We call this the natural fertility pattern and embody it in a schedule $n(a)$. The actual age-specific rates in a natural fertility population are proportional to this schedule. The shape of the schedule is represented by the top line in Figure 1.

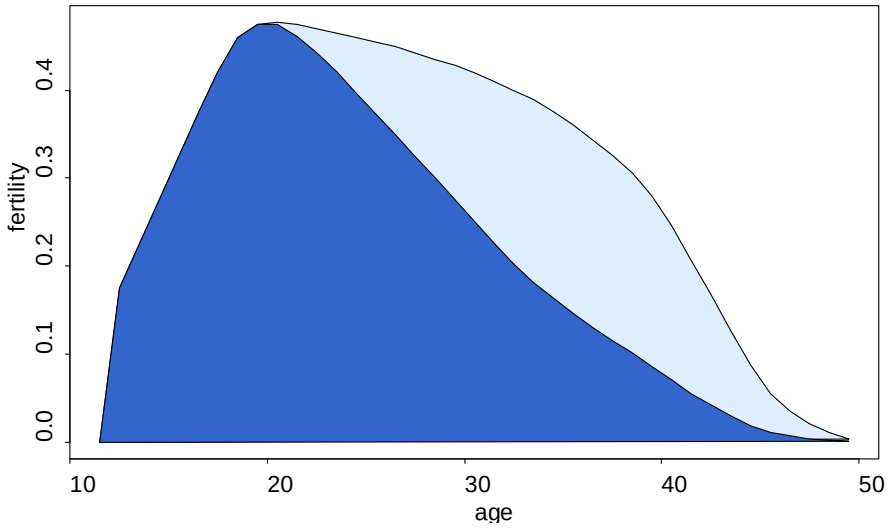


Figure 1. Natural and Marital Fertility

Coale discovered that populations that controlled their fertility (consciously limiting the number of children) exhibit a typical pattern of deviation from natural fertility, with departures in the log scale increasing with age according to a typical schedule $v(a)$. This led him to propose a model where marital fertility at age a is

$$m(a) = M\, n(a) e^{-mv(a)}$$

Here M is a parameter representing the level of marital fertility, and m is a parameter representing the degree of control. Note that dividing by the natural fertility schedule $n(a)$ and taking logs we obtain

$$\log \frac{m(a)}{n(a)} = \alpha + \beta v(a)$$

with intercept $\alpha = \log(M)$ and slope $\beta = -m$. This makes it very easy to check visually if a given fertility schedule follows Coale's model. The model is also easy to fit by simple linear regression. An even better approach is to use Poisson regression, as suggested by Bröstrom and Trussell.

The lower line in Figure 1 shows a marital fertility schedule where $M = 1$ and $m = 1$. The lighter area shows the extent to which fertility falls below natural fertility as a result of control. The textbook shows a simple application to data from Mali in 1955-6 with $M = 0.76$ indicating a level of natural fertility 24% below Henry's level, and $m = 0.189$ showing little evidence of control. The fit is not particularly good.

General Fertility by Age

Coale and Trussell combined Coale's model of marital fertility with the Coale-McNeil model of marriage, writing the general fertility rate at age a as the product of the proportion married at that age by the age-specific marital fertility rate, assuming no fertility outside marriage. (This assumption is a lot more reasonable if we define "marriage" to include both legal and consensual unions.) Thus

$$f(a) = cG_0\left(\frac{a-\mu}{\sigma}\right) Mn(a)e^{-m\,v(a)}$$

It is easy to see that the parameters c and M are not separately identified, so we can't distinguish the level of natural fertility from the proportion who eventually marry from age-specific fertility rates alone. (Unless, of course, we have data on proportions married and marital fertility and fit the two components of the model separately.)

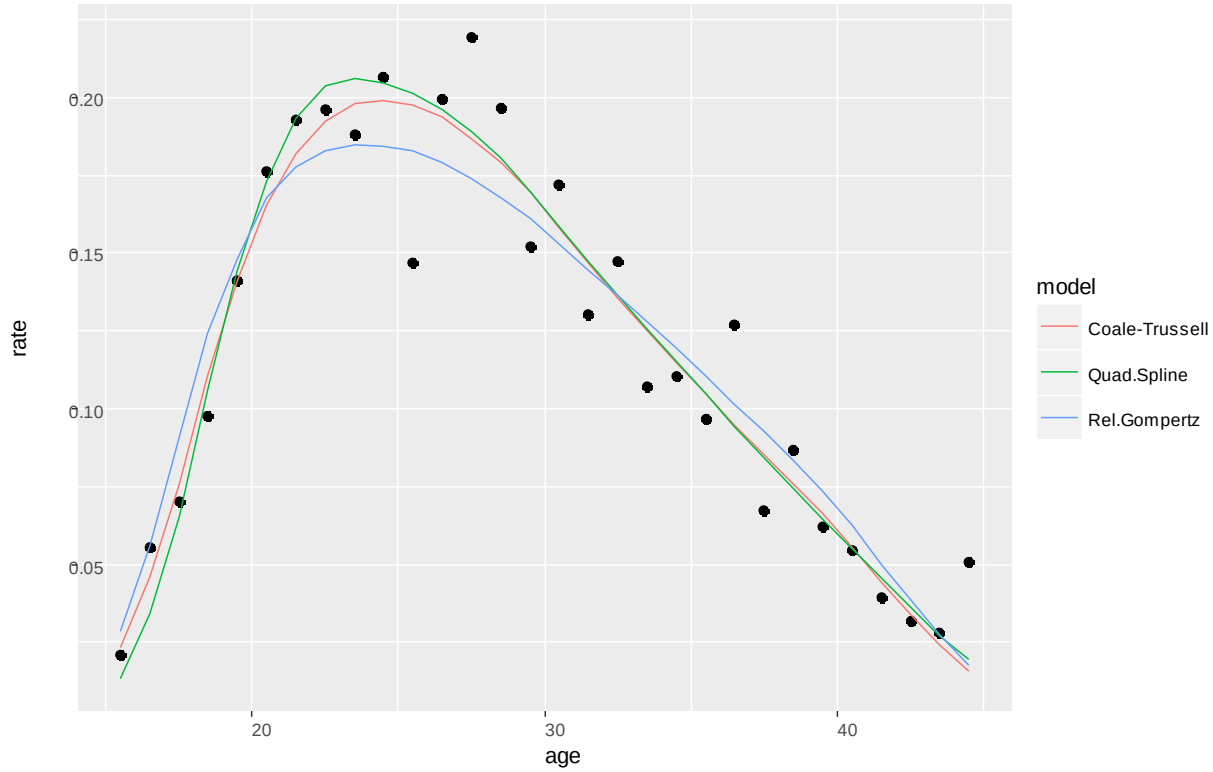


Figure 2. Age-specific Fertility Rates and Model Fits, Brazil DHS 1986

In a paper with Philipov we show how to fit this model using maximum quasi-likelihood procedures. A simple alternative is to use non-linear least squares. Figure 2 shows our fit to data from the Brazil Demographic and health Survey of 1986 using births in the four years before the survey. The observed rates are noisy but the model does an excellent job smoothing the data. (The figure also shows two other models discussed below.)

The estimated parameters (μ, σ, M, m) are (20.52, 5.10, 0.591, 0.771), indicating that average age of entry into union (exposure) is 20.52 with a standard deviation of 5.10, and that there is substantial fertility control among those in union (exposed). Because the Brazilian DHS includes information about unions we were able to fit the two components of the model separately. We got (22.12, 5.39, 0.954) from the union component and (0.738, 0.993) from the marital fertility component. The separate models imply later entry into risk and more control, but do not fit the observed rates as well as the combined model (details not shown).

Marital Fertility by Age and Duration

Page proposed a model of marital fertility where the level of natural fertility depends on a woman's age as in the previous models, but the degree of control depends on the duration of marriage instead of age. She further discovered that the pattern of departure from natural fertility viewed as a function of union duration rather than age was linear in the log scale, so no special schedule of control was needed. This leads to the following model of marital fertility by age and duration

$$m(a,d) = Mn(a)e^{\beta d}$$

Or, dividing by the natural fertility schedule and taking logs.

$$\log\frac{m(a,d)}{n(a)} = \alpha + \beta d$$

where $\alpha = \log(M)$. In work with Cleland we fitted this model to data from the WFS and showed that the degree of control parameter β is strongly correlated with contraceptive use, particularly use for limiting, and the level of natural fertility parameter α is correlated with breastfeeding duration as well as contraceptive use for spacing. We therefore called α and β the limiting and spacing parameters. In subsequent work we used this model to study social determinants of fertility in terms of their effects on limiting and spacing behavior by letting each of the parameters depend on covariates.

The Relational Gompertz Model

Brass proposed a relational Gompertz model of fertility. The basic idea is to transform the proportion of cumulative fertility achieved by age a using a log-log transformation

$$Y(a) = -\log(-\log\left(\frac{F(a)}{F}\right))$$

where F is short for $F(\infty)$, representing the total fertility rate, and then assume that the transformed schedule is a linear function of a standard schedule

$$Y(a) = \alpha + \beta Y_s(a)$$

which was derived by Booth and is available in the *Tools for Demographic Estimation* website. (Interestingly the standard was derived by examining a large collection of Coale-Trussell schedules.)

The transformation in the first equation above is called a *gompit* and is related to the Gompertz growth function, just like a *logit is* related to the logistic growth function. The Gompertz growth function in turn, is closely related to the Gompertz survival function.

The model is very easy to fit by OLS if we accumulate the observed rates and then divide by the TFR to obtain proportion of total fertility achieved at each age, and has applications in indirect estimation. Unfortunately it doesn't fit the Brazilian data very well, as you can see in Figure 2.

The Quadratic Spline Model

Schmertmann (2003) proposed a quadratic spline model with four "graphically intuitive" parameters, a, P, H and R. The first three represent the ages at which fertility rises above zero, reaches its peak, and falls to half its peak level, and the last one is the peak fertility.

Figure 2 shows that the model fits the Brazilian data reasonably well. (The residual sums of squares are 0.0087 for Coale-Trussell, 0.0090 for Schmertmann and 0.0110 for Brass, but note that the relational Gompertz model has three parameters and the other two models have four each.) The quadratic spline index ages are 13.85, 23.18 and 35.66, and the peak level is 0.206.

The model is a quadratic spline, so it can be written as

$$f(a) = R \sum_{k=0}^{4} \theta_k (a - \xi_k)_+^2$$

where ξ_k are the knots and θ_k the coefficients for $k = 0, \ldots, 4$. These parameters can be obtained from the index ages by solving a system of linear equations which imposes a number of constraints to reduce the number of shape parameters from 10 to three.

A nice feature of the model is that it can be integrated analytically to obtain cumulative fertility as a cubic spline

$$F(a) = \frac{R}{3} \sum_{k=0}^{4} \theta_k (a - \xi_k)_+^3$$

When working with grouped data I recommend computing the rate for ages $(x, x + n)$ as the difference $F(x + n) - F(a)$, which is more accurate than the mid-point rate $f(x + \frac{n}{2})$.

Birth Intervals

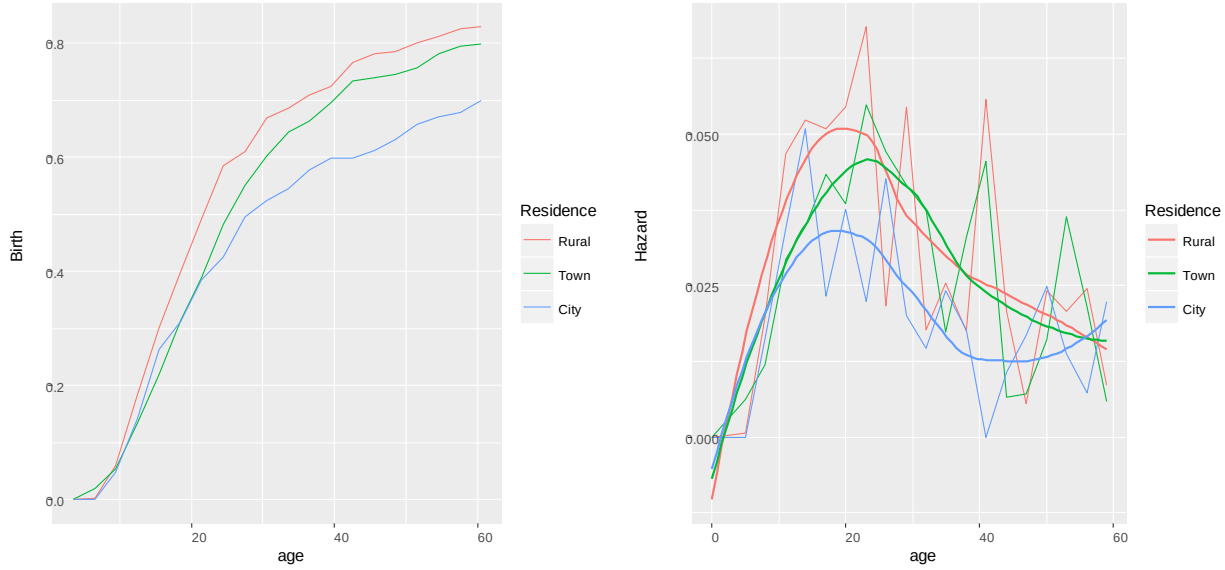
POP 502 / ECO 572/ SOC 532 • SPRING 2017

Birth interval analysis provides a more detailed view of the family building process than conventional fertility rates. We discuss briefly life table analysis of birth intervals and models of conception and birth, topics discussed in Section 5.4 of the textbook.

Life Table Analysis of Birth Intervals

Life table techniques can be used to study the progression from one parity to the next. Of interest here is the force of fertility, a hazard function that reflects the risk of moving to the next parity by duration since the last birth, and the *birth function* $B(d)$, or proportion who have moved to the next parity by duration d since the previous birth. (This function is analogous to our old friends $F(a)$, the proportion married by age a, and $1 - l(x)/l(0)$, the complement of the survival function.) The proportion who eventually move on is called the *parity progression ratio.* The average time it takes to move is the *length of the birth interval*.

Hobcraft and I did an illustrative analysis of birth intervals using data from the Colombian World Fertility Survey. The figures below show the birth and hazard functions for the transition from second to third birth by childhood type of place of residence, for women who had a second birth in the ten years preceding the survey. (The original paper used all births, so results are slightly different.)



The birth function shows that women who grew up in cities are less likely to make the transition to a third birth than those who grew up in towns or rural areas. The hazards functions are noisier, as you might expect, but smoothing shows a typical pattern where the hazard rises quickly to reach a maximum after one or two year and then declines. The rise is due to women coming out of the post-partum non-susceptible period, and the decline can be attributed to fertility control and/or selectivity.

Most women who make the transition to the next parity do so within five years, so one can summarize the *quantum* of fertility using the birth function at five years, and the *tempo* using the mean birth interval for those who make the transition within five years. As shown in the summary table on the right, women who grew up in rural areas have much higher parity progression ratios than those who grew up in towns or cities. They also have the shortest birth intervals, but the relationship between interval length and childhood residence is not monotonic, as women who grew up in towns have the longest intervals.

Residence	Quantum	Tempo
Rural	82.3	19.64
Town	79.3	22.06
City	67.7	20.24

The first birth interval is different from the others because it doesn't start with a birth. Traditionally demographers have studied the transition from first marriage to first birth, but this is fraught with difficulties because of premarital births. Ideally one would want a better marker of the start of exposure, but the necessary data are rarely available. A better strategy is to think directly in terms of entry into motherhood. The Coale-McNeil model of age at first marriage has been used successfully to model age at first birth. Birth intervals, like fertility rates, can be based on period or cohort data, with a synthetic cohort interpretation in the latter case. Cohorts can be defined by year of birth, year of entry into motherhood, or the year in which a specific parity is reached.

A side note: in life table analysis of birth intervals we compute rates dividing the number of births of a given order (say second births) by women in the previous parity (those with one birth), who are of course the only ones at risk of making that particular transition (to a second birth). We also index the process by duration since previous birth. Sometimes analysts compute order-specific fertility rates by age, dividing births of a given order to women in an age group by all women in the age group, regardless of parity. These rates are then summed to obtain order-specific TFRs, which are then interpreted as synthetic parity progression ratios. The age-order specific rates have the nice property that they add up to the overall ASFR, but they are not true event-exposure rates. (Clearly women with three children are not exposed to have a second child.) This is the same distinction we encountered before in terms of marriage frequencies and marriage rates, or between death densities and hazards.

Models of Conception and Birth

There is a long tradition of work on mathematical models of conception and birth by Sheps, Menken, Potter, Bongaarts, and others; and some of this work is reviewed in Section 5.4 in the textbook. Here we set aside issues of unobserved heterogeneity to focus on a few key ideas that will be useful later.

A typical birth interval has three components, a non-susceptible period that ends with the resumption of ovulation, a waiting time that ends with conception, and a gestation period that ends with the next birth. The first interval is different in that it doesn't start with a non-susceptible period. The figure below shows these components starting with the waiting time to conception.

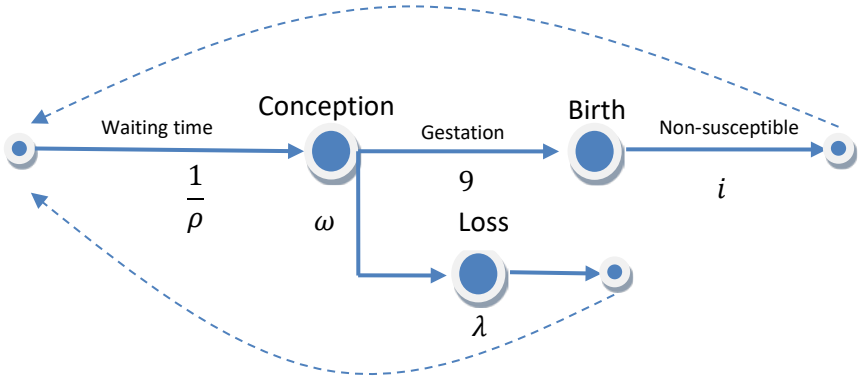


Figure 1. A Simple Model of Conception and Birth

In a homogeneous population with constant fecundability ρ the waiting time to conception is $1/\rho$. (This is a standard result for Bernoulli trials.) If $\rho = 0.2$ it would take an average of 5 months to conceive, and adding a typical gestation period of 9 months would lead to a first birth interval of 14 months. The length of the non-susceptible period depends on such practices as post-partum abstinence and the length of breastfeeding. If this phase lasts i months, then subsequent birth intervals would last on average $1/\rho + 9 + i$ months. If $i = 7.5$ months we get an average birth interval length of 21.5 months.

So far we have ignored pregnancy losses, such as miscarriages and still births. If the probability that a pregnancy will not end in a live birth is ω it takes on average $\omega/(1-\omega)$ losses before a successful live-birth conception. (This is the same standard result as above, but counting the delay, or failures before the first success.) Let λ denote the length of gestation plus the non-susceptible period following a pregnancy loss. (One could separate these two segments, but I follow the textbook in considering them together.) Each loss adds $1/\rho + \lambda$ months to the waiting time to a live-birth conception, the $1/\rho$ months it took to conceive plus gestation and infecundity following the loss. The total length of the birth interval is then

$$\left(\frac{1}{\rho} + \lambda\right)\frac{\omega}{1-\omega} + \frac{1}{\rho} + 9 + i$$

(Combining the two terms on $1/\rho$ yields equation 5.13 in the textbook, which uses s_b and s_ω for the no-susceptible periods following a birth and a loss; the latter includes gestation but the former doesn't, so I prefer using different symbols.) If 20% of pregnancies are wasted, so $\omega = 0.2$, one would average 0.25 losses before a live birth conception. Assuming that gestation plus infecundity takes up $\lambda = 5$ months, each loss would add 10 months (5 to conceive and 5 for gestation and non-susceptibility). The average 0.25 losses would then add a total of 2.5 months to the birth interval. Under these conditions, women would have a first birth after 16.5 months, with another birth following every 24 months.

Birth intervals can be translated into expected number of children by calculating how many intervals fit into the reproductive period. A woman who married at age 15 and had children following our simple model until age 45 (so her reproductive span is 360 months) would have one birth after 16.5 months and then (360-16.5)/24 = 14.3 more, for a total of 15.3 births. (We will encounter this number again

later.) If she married at 25 instead she would have 'only' 10.3 children, and we can easily see the effect of delaying marriage.

This model can also be used to estimate the effect of contraception on fertility. We say that a method has *effectiveness e* if the monthly probability of conception ρ is reduced to $\rho(1-e)$; so a 90% effective method reduces fecundability to 10% of what it would be otherwise. It is easy to see that the waiting time to conception is then $1/\rho(1-e)$ instead of $1/\rho$, so the average wait would be 1/0.02 = 50 months instead of 5 with 90% effective contraception. The first birth interval would then be 72.75 months, subsequent birth intervals would be 80.25 months, and our mythical woman would have 4.58 births over 30 years. If she also waited to marry at 25 instead of 15 she would have 3.08 children.

An interesting application of this simple model is to compute births averted by abortion. You'd think an abortion averts exactly one birth, but this ignores two facts: (1) a woman having an abortion would become susceptible much sooner than if she had carried the pregnancy to term, particularly if there is prolonged lactational infecundity, and (2) some of the pregnancies that are aborted would have resulted in a loss anyway, with the fraction depending on the timing of abortion. Under the assumptions used so far and with no contraception, an abortion adds 10 months to the birth interval (5 to conceive and 5 in gestation and infecundity), so it prevents 10/24=0.435 births. Using 90% effective contraception an abortion increases the birth interval from 80.25 to 135.25 months, thus preventing 55/80.25 = 0.685 births.

The textbook notes that an abortion effectively prevents one birth when contraception is very effective and the waiting time to conception dominates the birth interval, but this is only true if there are no other pregnancy losses. Equation 5.15 effectively assumes that abortions occur very early, so a fraction ω are redundant and the best you can do is avert $1-\omega$ births. A more realistic model would distinguish early and late losses and allow for the timing of abortion. A simple solution is to add a delay for recognized losses before the decision to abort of the form $(1/\rho + \lambda_e)\omega_e/(1-\omega_e)$ where λ_e is the gestation and infecundity associated with an early loss (say 4 months) and ω_e is the probability of an early loss (say 0.12 instead of 0.20). This increases the length of an interval with 90% effective contraception from 80.25 to 142.61 and averts 0.777 births. (With 99% effective contraception the original model gives 0.785 and adding an allowance for losses 0.893.)

These calculations are extremely simplistic because they ignore age effects and heterogeneity across women, but they have the advantage that they can be carried out 'on the back on an envelope'. More realistic models require simulation. Potter has looked at births averted when abortion is added to contraception using data from Taiwan and a simulation model called ACCOFERT, which has the added advantage of letting some of the parameters vary with age. He concludes that "if abortions are being performed in the third month of pregnancy upon 30-year-old women who are regularly practicing 98 percent effective contraception, the mean number of births averted per operation is 0.85; but if the same women are not practicing contraception at all, births averted per abortion average only 0.45." Another simulation model you may find interesting is SOCSIM, developed by Hammer and Watcher.

Proximate Determinants

POP 502 / ECO 572/ SOC 532 • SPRING 2017

The textbook has a brief discussion of Bongaarts's proximate determinants framework in Section 5.2. This is also a summary with a bit of historical context, a couple of examples, and a brief overview of recent developments.

Davis-Blake: The Intermediate Variables

In 1956 Davis and Blake published a very influential paper noting that any social factor affecting fertility had to act through one of eleven *intermediate variables,* which they grouped into three main categories:

I. Factors Affecting Exposure to Intercourse ("Intercourse Variables")

A. Those governing the formation and dissolution of unions in the reproductive period

1. Age at entry into sexual unions
2. Permanent celibacy
3. Time spent after or between unions (because of divorce, separation, or death of husband)

B. Those governing the exposure to intercourse within unions

4. Voluntary abstinence
5. Involuntary abstinence (from impotence, illness, temporary separations)
6. Coital frequency (excluding periods of abstinence)

II. Factors Affecting Exposure to Conception ("Conception Variables")

7. Fecundity or infecundity as affected by involuntary causes
8. Use or non-use of contraception (mechanical, chemical or other)
9. Fecundity of infecundity as affected by voluntary causes (sterilization)

III. Factors Affection Gestation and Successful Parturition ("Gestation Variables")

10. Fetal mortality from involuntary causes
11. Fetal mortality from voluntary causes

This list served as a conceptual framework for many years. It would take some twenty years, however, before a quantitative model would emerge.

Bongaarts: The Proximate Determinants

In 1978 Bongaarts recast the list in terms of eight variables that he called the *proximate determinants* of fertility, distinguishing three types of factors representing exposure, deliberate fertility control, and natural fertility. The complete list follows. One of the most innovative aspects of the list is the explicit recognition of the important role of lactional infecundity as a determinant of fertility.

I. Exposure factors

1. Proportions married

II. Deliberate marital fertility control factors

2. Contraception
3. Abortion

III. Natural marital fertility factors

4. Lactational infecundity
5. Fequency of intercourse
6. Sterility
7. Spontaneous intrauterine mortality
8. Duration of the fertility period

Bongaarts also argued that only the first four proximate determinants varied enough across populations to play a significant role in explaining fertility levels and differentials, and went on to propose a quantitative framework that explains the observed level of fertility in terms of the four proximate determinants using a simple multiplicative model:

$$\mathrm{TFR} = C_m \, C_c \, C_a \, C_i \, \mathrm{TF}$$

Here TFR is the total fertility rate, and TF is a maximum potential natural total fertility rate, often taken to be 15.3, a number we encountered before in the context of models of conception and birth.

The four indices C_m, C_c, C_a and C_i represent the *fertility reducing* effects of marriage, contraception, abortion and post-partum infecundity. If all these indices were one then women would have on average 15.3 children. In reality, of course, the indices are usually considerably less than one.

The indices can all be defined in terms of ratios of fertility rates. In particular, the **index of marriage** is the ratio of the TFR to the TMFR

$$C_m = \frac{\mathrm{TFR}}{\mathrm{TMFR}}$$

If we assume no extra marital births this index is a weighted average of proportions married by age with the marital fertility rates as weights. (The similarity to I_m should not go unnoticed.)

The **index of contraception** can be defined as a ratio of the TMFR to the average number of births a woman married throughout the reproductive ages would have in the absence of contraception. The index depends on contraceptive prevalence among married women and on the effectiveness of the methods used, and is often estimated using the following equation

$$C_c = 1 - 1.18 \, u \, e$$

where u is the average proportion of married women currently using contraception in each age group and e is the average effectiveness of the methods used. When age-specific prevalence rates are not available u is estimated as the overall contraceptive prevalence rate, or proportion of married women using contraception. In the absence of effectiveness data values are usually borrowed from another population, often the rates given for the U.S. in 1970 in Table 1 of Bongaarts's paper. Alternatively one could use recent estimates from *Contraceptive Technology*. The constant 1.18 is related to the proportion of married women that is non-sterile and represents an approximation. If nobody uses contraception C_c is one. When all non-sterile women are protected by perfect methods C_c is zero.

The **index of abortion** is defined as the ratio of the TFR to what the TFR would be if women had no abortions. We know by now that an abortion averts on average less than one birth, with the exact number depending on the practice of contraception following the abortion. Bongaarts estimates that an abortion averts on average $b = 0.4(1 + u)$ births, or between 0.4 when no contraception is practiced and 0.8 when all women who have abortions use contraception. Ideally u should measure contraceptive use among women who have an abortion, but it is often estimated using the proportion of married women using contraception. The incidence of abortion is estimated using the total abortion rate, TA. This leads to the index

$$C_a = \frac{\text{TFR}}{\text{TFR} + 0.4(1 + u)\,\text{TA}}$$

Unfortunately data on abortions are scarce and notoriously unreliable, so this factor is often estimated as a residual. Westoff has developed regression equations to estimate the total abortion rate from other data, including contraceptive use and the TFR.

The **index of lactational infecundity** measures the fertility reducing effect of breastfeeding, and is based on the simple model of conception and birth discussed earlier. The post-partum non-susceptible period lasts between 1.5 months and two years, depending on the duration of breastfeeding. This segment is followed by the waiting time to conception, which is typically 7.5 months. Spontaneous pregnancy losses add an average of 2 months to the waiting time. This is followed by 9 months of gestation leading to a live birth. Thus, the typical birth interval is 1.5 + 7.5 + 2 + 9 = 20 months without lactation and i + 7.5 + 2+9 = $i + 18.5$ more generally, so the fertility reducing effect of breastfeeding can be estimated as the ratio

$$C_i = \frac{20}{18.5 + i}$$

where i is the average duration of the infecundable period from birth to the first postpartum ovulation, often equated to the duration of breastfeeding.

When data are available to estimate all four indices the model yields an estimate of TN, which is obtained dividing the TFR by the product of the indices. This value shouldn't be too far from 15.3, depending on the effects of the other four proximate determinants and the model fit.

Korea 1960-70 and U.S. 1965-73

Bongaarts applies his model to explain the declines in fertility in Korea between 1960 and 1970 and in the U.S. between 1965 and 1973. The table and figures below summarize his results.

TABLE 1. Proximate Determinants in Korea and the U.S.

Proximate	Korea		U.S.	
Determinant	1960	1970	1965	1973
Infecundity	0.56	0.66	0.93	0.93
Abortion	0.97	0.84	1.00	0.95
Contraception	0.97	0.76	0.31	0.22
Marriage	0.72	0.58	0.61	0.57
TFR	6.13	4.05	2.72	1.67

Looking first at Korea, the TFR declined from 6.13 to 4.06 as a result of increases in contraception and abortion and, to a lesser extent, a decline in marriage, which together compensated for a reduced effect of lactational infecundity, as the duration of breastfeeding was reduced (from 17.4 to 11.9 months).

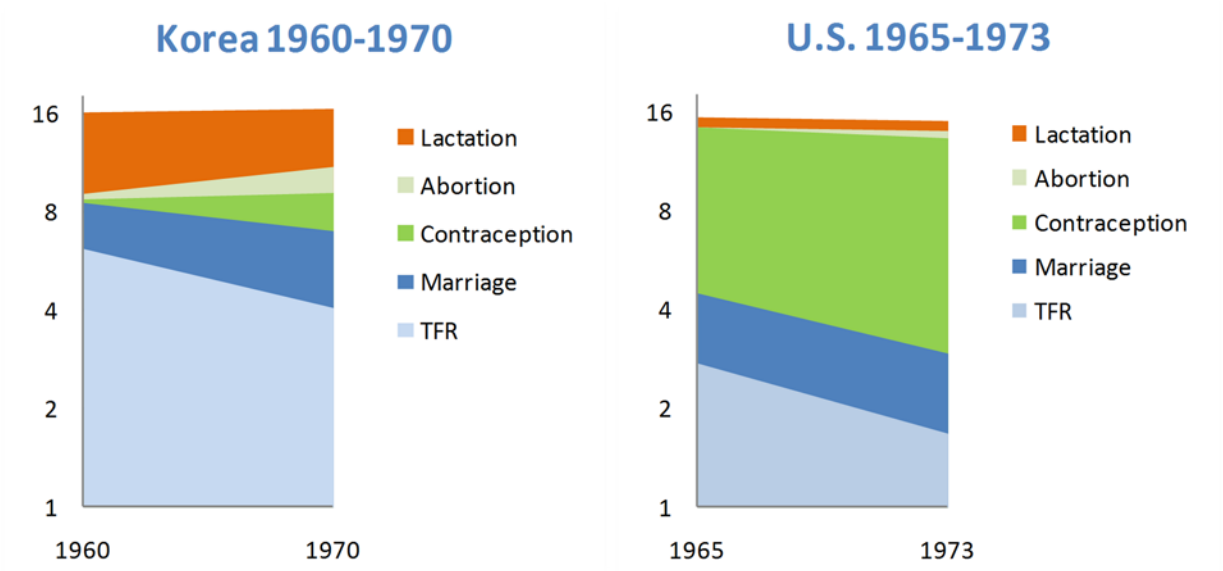


FIGURE 1. Proximate Determinants in Korea and the U.S.

In contrast, in the U.S. practically all the decline from 2.72 to 1.67 children per woman can be attributed to increased contraceptive use, with small effects of marriage and abortion and no change at all in breastfeeding. Comparing across countries we are struck by the much larger impact of contraception in the U.S. and the important and persistent role of lactational infecundity in Korea, where it was reducing fertility by 34% in 1970.

A technical note on the figures: in a multiplicative model the order of the factors does not alter the product, but if the results are shown as proportionate reductions from a maximum such as 15.3, as is customary, the visual impression depends very much on the order in which the factors are introduced. Bongaarts (Figure 4) starts with lactational infecundity, which translates its effect on births as if all women were married and not using contraception or abortion. If one was to follow the Davis-Blake order, going from exposure to intercourse to conception and birth, marriage would come first and its effect would look larger. A simple solution is to use a log scale when plotting the fertility rates after applying each index, so the effects are additive and the order becomes immaterial. This is the approach that Hobcraft and I used when we applied this framework to the analysis of repeat fertility surveys in the Dominican Republic in 1975 and 1980, and is the approach used in the above figures.

Bongaarts: Time for a Tune-Up

Changes in reproductive behavior over the last 30 years have required some adjustments of the model. Stover (1998) proposed a number of revisions and more recently Bongaarts (2015) noted that it was "time for a tune-up" and proposed some updates to make the model more accurate in contemporary populations. We review briefly the latest revision.

A key feature of the revised model an emphasis on age-specific indices, which are now easily obtained from DHS surveys, particularly for marriage and contraception. (The original model also had an age-specific version, but it was rarely used because the required data were not available.) The aggregate index is a weighted average of the age-specific indices as explained below.

Another change concerns timing: because the proximate determinants can only affect fertility 9 months later, and DHS surveys calculate fertility rates for the three years before the survey, the revision uses interpolation between two surveys to estimate the indices 27 months (or 2.25 years) before the survey. Following a discussion of the revised indices we go through an example using data from Colombia.

Sexual Exposure

Because extra-marital sex and pregnancy are becoming more prevalent, Bongaarts proposed modifying the index of marriage by counting women who are in unions plus unmarried women who are pregnant, report sex in the last month, use contraception, or are in the post-partum insusceptible period, in a renamed index of sexual exposure. The new index is

$$c_m^*(a) = m(a) + ex(a)$$

where $m(a)$ is the proportion married/in union, and $ex(a)$ is a measure of extra-marital exposure based on the criteria listed above.

The aggregate index is a weighted average of these proportions with weights equal to the age-specific fertility rates among exposed women, which are in turn estimated as $f(a)/c_m^*(a)$. This makes the index a ratio of the TFR to a total *exposed* (rather than marital) fertility rate.

Contraception

The revision addresses three issues:

(1) the original model ignored overlap between contraceptive use and post-partum infecundity, but this overlap has become significant in societies with long durations of breastfeeding or abstinence; the solution is simply to exclude women in post-partum amenorrhea or abstinence when calculating the index of contraception;

(2) the overall index was based on contraceptive use among all married women aged 15-49 and therefore was affected by the age composition of women in unions; the revised index uses age-specific prevalence rates among exposed women; and

(3) the model allowed effectiveness to depend on the method mix (in the aggregate model) or age (in the age-specific model) but not both; the revision considers both age and method mix.

The new index is then

$$c_c^*(a) = 1 - r^*(a)\big(u^*(a) - o(a)\big)e^*(a)$$

where $u^*(a)$ is age-specific prevalence among exposed women, $o(a)$ is the overlap with post-partum infecundity, $e^*(a)$ is the average effectiveness of methods used at age a, and $r^*(a)$ is a fecundity adjustment (1.18 in the original model). In practice the use and effectiveness corrected for overlap are combined in a single proportion, as we'll see in the application.

The fecundity adjustment is based on a regression equation which reflects the higher fecundity of users, and also picks up the fact that younger users have lower effectiveness. The required values are given in Table 2 below. When calculating the index any values of $c_c^*(a)$ lower than 0.1 are set to 0.1, as a simple way to correct some anomalies noted in countries with high levels of sterilization.

The aggregate index C_c is a weighted average of the age-specific indices, with weights proportional to age-specific fecundity rates $f_f^*(a)$. These are also unknown, but are estimated from the same regression equation used above, and are given in Table 2 below.

Abortion

The estimate of births averted by abortion was based on a model with "limited analytic foundation", and has been replaced by the ratio of average reproductive time associated with abortions and live births:

$$b(a)^* = 14/(18.5 + i(a))$$

where $i(a)$ is the length of post-partum insusceptibility. In practice $i(a)$ varies little by age and one uses an average over all ages. The age-specific index is then

$$c_a^*(a) = \frac{f(a)}{f(a) + b^* \, \mathrm{ab}(a)}$$

where $f(a)$ is the ASFR at age a. The aggregate index is calculated as $C_a^* \approx \mathrm{TFR}/(\mathrm{TFR} + b^* \, \mathrm{TAR})$ where TAR is the total abortion rate, often estimated as 30 times the general abortion rate at ages 15-44.

Infecundity

There are no changes in the age-specific index of post-partum infecundity:

$$c_i^*(a) = \frac{20}{18.5 + i(a)}$$

where $i(a)$ is the average duration of post-partum infecundity at age a. As noted earlier $i(a)$ varies little by age, so the aggregate index is computed using the average duration of the non-susceptible period.

Illustrative Calculation: Colombia 2010

We will illustrate the calculation of the revised indices using DHS data for Colombia. I am very grateful to John Bongaarts for providing a spreadsheet with the data used in his paper. The age-specific inputs needed appear in Table 2. For measures that require interpolation we provide the data for 2005 and 2010.

TABLE 2. Inputs for Proximate Determinants Model in Colombia

Age group	ASFR	Input Data from DHS 2005 and 2010				Regression-based fecundity adjustment	
		Married-exposed		Use-effectiveness			
	2010	2005	2010	2005	2010	r	$\bar{f}_f(a)$
15-19	84	.3235	.3816	.4611	.5244	.6167	679
20-24	122	.7274	.7777	.5765	.6027	.8099	631
25-29	100	.8252	.8729	.6421	.6634	.9902	588
30-34	70	.8661	.8906	.6959	.7207	1.0767	514
35-39	38	.8519	.8868	.7462	.7613	1.1359	380
40-44	12	.8248	.8695	.7640	.7938	1.2550	192
45-49	2	.8001	.8168	.7213	.7446	1.6187	60

The married-exposed columns are calculated from the DHS surveys using six standard variables: V501 for current marital status, V536==1 for sexually active in the last 4 weeks, V405==1 for women in post-partum amenorrhea, v406==1 for those in post-partum abstinence, V203==1 for currently pregnant and v312 >=1 for women currently using a method, all of whom are considered exposed.

The use-effectiveness columns are computed using standard variable V312, the method currently used. The calculation first sets this variable to zero when V405 or V406 is one, which avoids any overlap between use and post-partum insusceptibility period, and then computes a proportion using contraception weighted by effectiveness, based only on exposed women. For lack of better data the

calculation assigns effectiveness 1 to male and female sterilization, 0.95 to IUD and Norplant, 0.90 to the pill and injection, and 0.70 to every other method.

We also know that the mean duration of post-partum infecundity was 8.9 months in 2005 and 9.8 months in 2010. Finally the general abortion rate is estimated as 32 per 1000 using data from Sedge and Singh published in *The Lancet* in 2012. These are all the survey-specific inputs needed, plus of course the regression weights shown in the last two columns of Table 2.

For convenience I divide the ASFR by 1000 to obtain

$$f(a) = (0.084\ 0.122\ 0.100\ 0.070\ 0.038\ 0.012\ 0.002).$$

The TFR is the 5 times the sum, 2.14.

To estimate the proximate determinants 2.5 years before the 2010 survey we use linear interpolation, with a weight $w = 2.25/5 = 0.45$ for 2005 and $1 - w = 0.55$ for 2010. The proportions exposed in 2007.75 are then estimated as

$$c_m^*(a) = (0.3555, 0.7551, 0.8514, 0.8796, 0.8711, 0.8494, 0.8093)$$

A weighted average of these using exposed fertility $f(a)/c_m^*(a)$ as weight gives the exposure index

$$C_m^* = 0.653$$

The effectiveness-weighted average proportions using contraception by age in 2007.75 are estimated as

$$ue(a) = (0.4959, 0.5909, 0.6538, 0.7095, 0.7545, 0.7804, 0.7341)$$

The age-specific contraception index is obtained as one minus the product of these proportions times the fecundity corrections r given in Table 2, or 0.1 (whichever is greater), to obtain

$$c_c^*(a) = (0.6942, 0.5214, 0.3526, 0.2360, 0.1430, 0.1000, 0.1000)$$

The weighted average of these indices using the fecundity rates in Table 1 as weights yields

$$C_c^* = 0.397$$

The interpolated average length of the non-susceptible period is 9.395 months, leading to the index

$$C_i^* = 0.717$$

Finally the total abortion rate is estimated as 30 times 0.032 or 0.96 abortions. When computing the number of births averted by one abortion, Bongaarts uses the most recent estimate of post-partum infecundity, not the interpolated value used above. The number of births averted by one abortion is then 14/(18.5 + 9.8) = 0.495, leading to

$$C_a^* = 0.818$$

The product of the four indices times the average total fecundity rate of 15.36 in the 36 countries in Bongaarts's analysis yields a predicted TFR of 2.34, a bit higher than the observed rate. It is clear that in Colombia contraceptive use is the proximate determinant with the largest fertility-reducing effect by far.

Bongaarts's Results for 36 Countries

As noted earlier, Bongaarts applied the revised framework to 36 countries with two recent DHS surveys. He also calculated his original index as well as Stover's revision. The average bias of the latest model is only 0.04 births, compared with for 0.18 Stover and 1.19 for the original model. A better measure is the standard deviation of error, which is 0.61 compared to 0.76 for Stover and 1.47 for the original.

The figure below plots the observed and predicted TFR's for all 36 countries using a log scale. I also include a line reflecting the expected linear relationship with slope one.

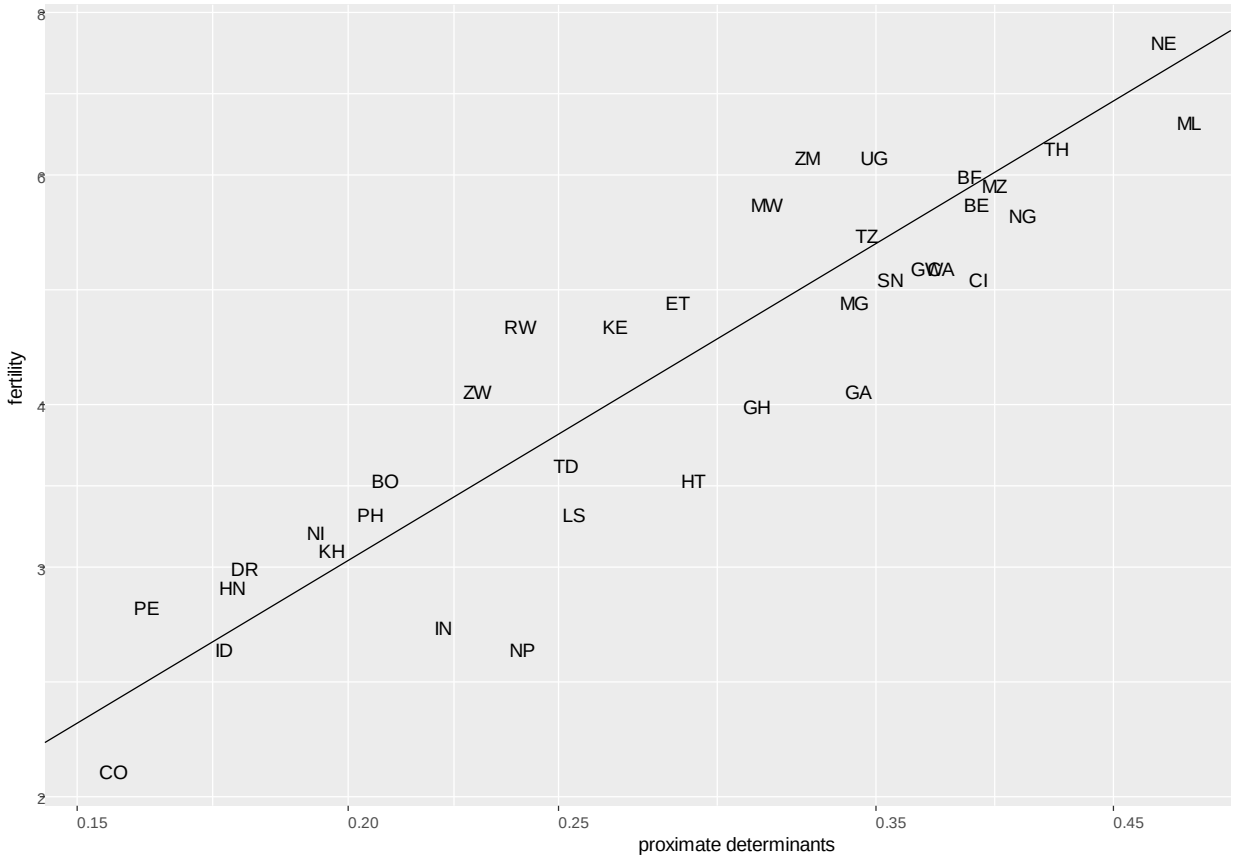


FIGURE 3. Proximate Determinants Fit to 36 DHS Countries

Overall the revised proximate determinants model explains 81.5% of the variation in observed log TFRs, and almost the same in the original scale, in a model where only the constant is estimated.

Our last figure shows the same type of decomposition we used for Korea and the U.S. for the 36 countries in Bongaarts's analysis, ordered from highest to lowest observed fertility and showing the reductions from maximum natural fertility that can be attributed to post-partum infecundity, abortion, contraception, and exposure.

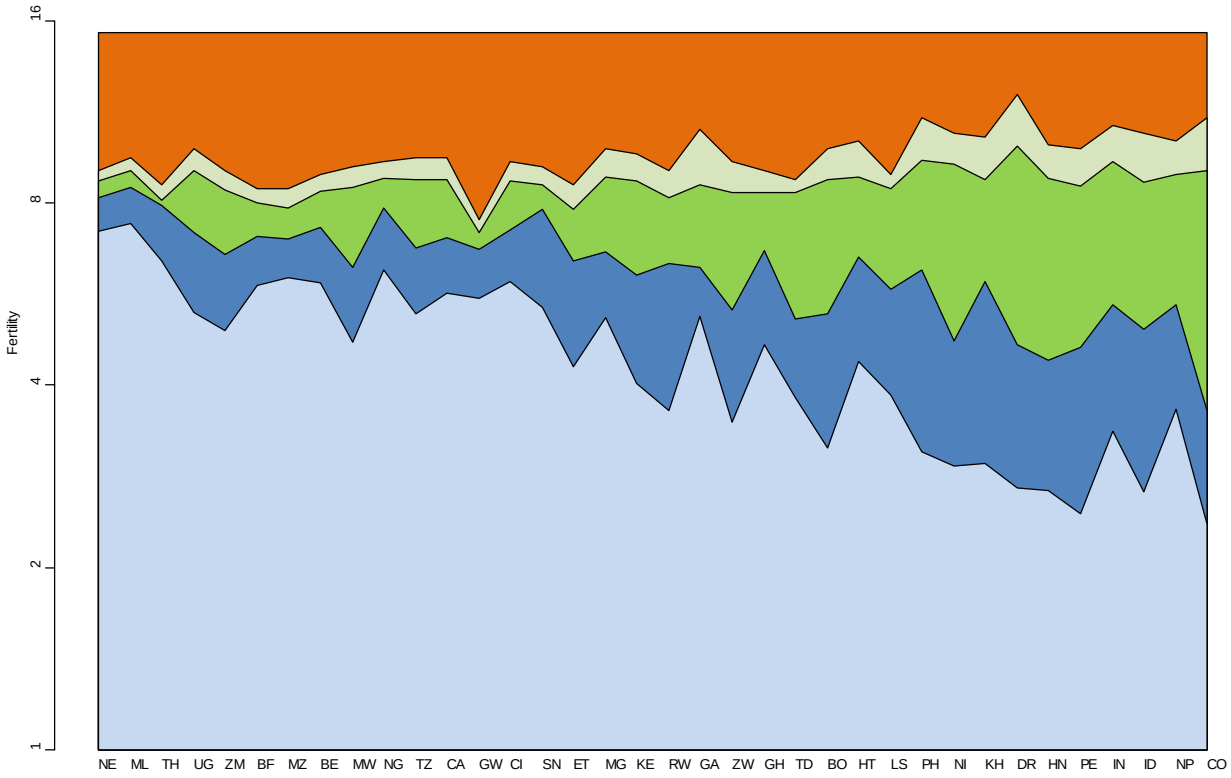


Figure 2. Proximate Determinants in 36 DHS Countries

We see how lactational infecundity plays a larger role in high-fertility countries, while contraceptive use and reductions in sexual exposure are much more determinant in low fertility countries, with abortion generally playing a more limited role.

Tempo Effects

POP 502 / ECO 572/ SOC 532 • SPRING 2017

A hot topic in Demography concerns the need to adjust period measures of fertility, nuptiality and mortality for so-called "tempo distortions". We will review quickly some of the main ideas, including Ryder's demographic translation formula and the Bongaarts-Feeney tempo-adjusted measures.

Fertility

Imagine a surface $f(a,t)$ of fertility rates by age and period. A period summary is obtained by summing over ages for a fixed time. In particular, the period TFR for year t is

$$\text{TFR}(t) = \int f(a,t)da$$

It is also useful to define the mean age $\mu_p(t)$ of the fertility schedule as

$$\mu_p(t) = \int af(a,t)da/\text{TFR}_p(t)$$

Cohort summaries are obtained by summing across a diagonal, where age and time vary together. In particular, the cohort TFR for the cohort born in year t is

$$\text{TFR}_c(t) = \int f(a,t+a)da$$

and the cohort mean age of childbearing is

$$\mu_c(t) = \int af(a,t+a)da/\text{TFR}_c(t)$$

Ryder

Ryder's chief concern was that period summaries provide a distorted view of the behavior of cohorts when fertility is changing. In particular, if women delay childbearing the period TFR will drop even if the cohorts have the same number of children as before, so the cohort TFR stays constant. Thus, a cohort change in tempo would look from the period perspective as a change in the quantum of fertility!

The following artificial example may help fix ideas. Consider cohorts having children in three age groups as follows:

	3.0	3.0	3.0	3.0	3.0	3.0	←Cohort TFR
35-44	0.6	0.6	0.6	0.7	0.8	0.8	
25-34	1.6	1.6	1.7	1.8	1.8	1.8	
15-24	0.8	0.6	0.4	0.4	0.4	0.4	
Period TFR→	3.0	2.8	2.7	2.9	3.0	3.0	

Initially cohorts have an average of 0.8, 1.6 and 0.6 births in each age group, for a total of three children. But then a cohort delays childbearing and has 0.6, 1.7 and 0.7 birth per age group, for a total of three.

Subsequent cohorts further delay childbearing and have 0.4, 1.8 and 0.8 in each age group, for a total of three. The cohort TFR is always 3.0. But as shown in the table the period TFR drops from 3.0 to 2.8 and 2.7 before it recovers to 2.9 and bounces back to 3.0. A more accurate cohort-to-period conversion would use Lexis triangles, but the simple calculation shown here suffices to show that a cohort change in tempo appears as a period change in quantum!

Ryder used a first order Taylor series expansion to relate period and cohort TFRs. He showed that for the cohort that reaches its mean childbearing age μ at time t (the cohort born at $t-\mu$)

$$\mathrm{TFR}_c(t-\mu) \approx \frac{\mathrm{TFR}_p(t)}{1-r_c(t-\mu)}$$

where $r_c(t-\mu)$ is rate of change or time derivative of cohort mean age of childbearing for the cohort reaching it mean childbearing age at time t. This remarkable formula shows that, to a first order of approximation, if cohorts postpone childbearing the period TFR will fall *below* the cohort TFR by an amount that depends on how fast the mean age of childbearing was increasing. This actually happened during the baby boom, as you can see in the computing logs.

Bongaarts-Feeney

In 1998 Bongaarts and Feeney proposed a tempo-adjusted total fertility rate, usually denoted TFR^*, based on an expression that looks remarkably like Ryder's translation formula

$$\mathrm{TFR}^*(t) = \frac{\mathrm{TFR}(t)}{1-r_p(t)}$$

There are, however, a few important differences. First, $r_p(t)$ is the rate of change or time derivative of the *period* mean age of childbearing at time t. This is much easier to calculate from available data. It is usually estimated by averaging the "in and out" changes between $t-1$ and t and between t and $t+1$. This turns out to be exactly the same as half the change beween $t-1$ and $t+1$.

Second, TFR^* is not a cohort rate, but rather a pure-period measure representing tempo-corrected fertility. This raises issues of interpretation that we discuss below.

A third difference is that B-F recommend applying the procedure separately by birth order, using rates that divide births of a given order by all women. The reasoning behind this approach is that as women have fewer high-order births the overall mean age of childbearing will decline without any changes in the timing of earlier births, so order-specific means provide a better measure of tempo changes. In my own opinion, order-specific fertility is best analyzed using true hazard rates, a point made by van Imhoff and Keilman in comments to the original B-F paper.

There has been a lot of discussion of TFR^* and some confusion about its meaning. B-F argue that they are not trying to estimate the TFR for any particular cohort and that TFR^* is just a "period measure purged of tempo distortions". The best way to think about this is as a counterfactual estimate of what the period TFR would be if women were not delaying childbearing. Zeng and Lang show that it can also

be interpreted at the TFR that would be observed if women followed a period schedule that is shifting constantly to older ages. A simple derivation of these results may be found in my tempo paper, which also provides a result for the mean age of childbearing under the implied period-shift model, which is shown as equivalent to an accelerated failure time model of cohort behavior.

The online computing logs show an application of these ideas to U.S. fertility using the Heuser cohort fertility tables. For a much more detailed analysis see the paper by Schoen in *Demography* in 2004.

Nuptiality

The same phenomenon we have noted with fertility can happened with nuptiality. If women postpone first marriage, then period estimates of the proportion who eventually marries will decline even if the same fraction of each cohort ends up marrying. Thus, a cohort change in tempo can masquerade as a change in period quantum.

Bongaarts and Feeney apply their procedure working with period marriage frequencies, obtained by dividing first marriages by the total number of women in an age group (not just those single). They accumulate these frequencies to obtain a Total First Marriage rate (TFMR), and also use them to calculate a period Mean Age at Marriage. They then define a tempo-adjusted TFMR as

$$\text{TFMR}^* = \frac{\text{TFMR}(t)}{1 - r_p(t)}$$

where $r_p(t)$ is the rate of change or time derivative of period mean age at first marriage. The procedure is formally identical to the adjustment used for fertility.

Note that this approach relies on frequencies rather than true event-exposure rates, and this causes some technical difficulties. If 60% of a cohort marries before age A, and 60% of the next cohort marries after age A, and we combine these frequencies, we would get a synthetic cohort where 120% marry! This couldn't happen with hazard rates, but the model is predicated on a shift of the period schedule of frequencies (or equivalently, cumulative proportions married).

The approach also assumes that women postpone first marriage by the same amount of time at all ages. If we observe fewer women marrying in a given year it could be because some will marry later and/or because some will forego marriage, and it is hard to determine the relative weight of these two explanations. Bongaarts and Feeney can separate the two effects by assuming a uniform delay at all ages. The quality of the adjustment depends on the validity of this assumption.

Mortality

In a more recent series of papers Bongaarts and Feeney extended their proposed tempo adjustment to mortality. They claim that conventional period life expectancy is a biased measure of longevity when mortality is declining, with a bias of up to 2 years in developed countries. Needless to say, they created quite a stir in the demographic community. With fertility (and nuptiality) we could all understand the risk of confusing changes in quantum and tempo, but with mortality the quantum is fixed, only tempo

can change, and no one would mistake one for the other. In other words if mortality rates decline, we know it is because people are delaying death.

B-F claim, however, that period measures of mortality suffer from the same "tempo distortions" as period measures of fertility, and propose an adjustment based on an estimate of the rate at which mortality is declining. To motivate the need for adjustment they use an example along the following lines. Suppose in a given year we all took a pill that made us immune to death (and aging) for three months. Clearly such a pill would add exactly three months to our life. Yet the death rates for that year would decline 25%, and in a country such as the U.S. in 2002 period life expectancy would rise by about 3.6 years, overestimating the gain in longevity.

We should remember, however, that conventional life expectancy is a counterfactual estimate of how long we would live if the rates observed in a given year remained in effect through our lives. Effectively that assumes that we would get a magic pill every year, in which case we would indeed live quite a bit longer. However, the example serves to illustrate a key feature of the B-F approach, the assumption that adult mortality declines because we all receive "increments to life", not because "rates decline". The underlying model is formally identical to the model for fertility and nuptiality, assuming a uniform shift in the survival curve to older ages. This is not realistic for the youngest ages, but Bongaarts and Feeney restrict their discussion of tempo effects in mortality to adult ages, say above age 25 or 30.

The gist of the method relies on death frequencies, computed by dividing deaths in an age group by the original size of the cohort (not just those alive). Accumulating these leads to the Total Mortality Rate (TMR). The rate of change or time derivative of the TMR is used to compute an adjustment factor, that is then used to inflate the age-specific mortality rates before calculating life expectancy. The result is the B-F tempo-adjusted life expectancy. Other interesting measures that come up are the cohort average length of life (CAL), and the standardized mean age at death; in addition, of course, to conventional life expectancy.

We will not discuss these further, as the dust has not settled. The book *How long do we live?* edited by Barbi, Bongaarts and Vaupel, has a collection of papers giving different views on this issue, including my own. It turns out that when adult mortality follows a Gompertz model it is impossible to distinguish a period shift to older ages from a proportionate decline in rates at every age. However, the two models—"reduction in rates" as opposed to "increments to life"—have different implications for the future, with the latter implying that if gains in longevity were to stop age-specific mortality rates would rise.

Conclusion

Everyone agrees that tempo effects exist. Mortality is a pure tempo phenomenon, as death is bound to occur and the only question is when. Things are different with fertility and nuptiality because the event in question may or may not occur. Under these circumstances a period change in quantum may reflect a change in cohort quantum, a change in cohort tempo, or an unknown mixture of the two. Distinguishing the two while the cohorts are still "making up their minds" is a tall order. The B-F adjustment removes the tempo component of the change under a model that assumes a uniform delay at all ages. Whether this adjustment is meaningful in the case of mortality remains a hotly debated issue.

Population Projections

POP 502 / ECO 572/ SOC 532 • SPRING 2017

We now review in some detail the cohort component method of population projection. We assume that we have population counts by age at time t and we want to project the population to time $t + n$ given a life table and a set of fertility rates. For simplicity we take the length of the projection to be the same as the width of the age groups. (We work with five-year intervals, but we could work with single years just as well.) We assume initially that the population is closed to migration.

Application to Sweden in 1993

The basic inputs are the initial population, the life table person-years lived, and the maternity rates. Box 6.1 in the textbook has the basic inputs for Sweden in 1993, and these are reproduced below and in the companion computing log. The data are given in terms of fertility rates, but I converted to maternity rates dividing by 2.05.

Age	1993 ${}_5N_x^F$	${}_5L_x^F$	${}_5F_x^F$	Survival Ratio[1]	Maternity Rate[2]	1998 ${}_5N_x^F$
0	293,395	497,487	0	0.9950[1]	0	293,574
5	248,369	497,138	0	0.9993	0	293,189
10	240,012	496,901	0	0.9995	0.0029	248,251
15	261,346	496,531	0.0059	0.9993	0.0250	239,833
20	285,209	495,902	0.0443	0.9987	0.0587	261,015
25	314,388	495,168	0.0731	0.9985	0.0639	284,787
30	281,290	494,213	0.0549	0.9981	0.0382	313,782
35	286,923	492,760	0.0215	0.9971	0.0126	280,463
40	304,108	490,447	0.0036	0.9953	0.0019	285,576
45	324,946	486,613	0.0001	0.9922	0.0001	301,731
50	247,613	480,665	0	0.9878	0	320,974
55	211,351	471,786	0	0.9815	0	243,039
60	215,140	457,852	0	0.9705	0	205,109
65	221,764	436,153	0	0.9526	0	204,944
70	223,506	402,775	0	0.9235	0	204,793
75	183,654	350,358	0	0.8699	0	194,419
80	141,990	271,512	0	0.7750	0	142,324
85+	112,424	291,707	0	*0.5179*[1]	0	131,768

[1]Entries are ${}_5L^F{}_x / {}_5L^F{}_{x-5}$ except for the first and last rows, which are ${}_5L^F{}_0/5l_0$ and $T^F{}_{85}/T^F{}_{80}$.

[2]Entries are average of ${}_5F^F{}_x$ and ${}_5F^F{}_{x+5}$ ${}_5L^F{}_{x+5} / {}_5L^F{}_x$. See the text for details

Projecting the Existing Population

Projecting the population above age 5 at time $t+5$ is relatively easy because in a closed population they are just the survivors of the population at time t. All we need is the probability of surviving from age $(x, x+5)$ to age $(x+5, x+10)$, which is ${}_5L_{x+5}/\,{}_5L_x$ from the life table, so

$$ {}_5P_{x+5}^{t+5} = {}_5P_x^t \frac{{}_5L_{x+5}}{{}_5L_x} $$

All rates and survival probabilities are for the female population. I omit the superscript F to avoid clutter.

The only slight complication concerns the open-ended age group, say 85+. This group consists of two subgroups, those 80-84 at baseline who would then be 85-89, and those 85+ at baseline who would be 90+ five years later. The survival probabilities are ${}_5L_{85}/\,{}_5L_{80}$ for the first group and T_{90}/T_{85} for the second. If the last age group is 85+, however, we do not have T_{90}. It is then customary to combine the last two age groups at baseline and project them together (so 80-84 and 85+ are treated as 80+, who then become 85+) using the survival ratio T_{85}/T_{80}, where $T_{80} = {}_5L_{80} + T_{85}$, so

$$ {}_\infty P_{85}^{t+5} = \left({}_5P_{80}^t + {}_\infty P_{85}^t \right) \frac{T_{85}}{{}_5L_{80} + T_{85}} $$

(The textbook describes on page 121 the procedure to be followed when T_{90} is available, but uses the combined projection in Box 6.1, so we'll stick to that.)

The resulting survivals ratios are shown in the first column of the second panel in the table. (Ignore the first entry for now.) This is all you need to project the population above age 5 in 1998.

Projecting Births

The population under age 5 at $t+5$ consists of births during the period from t to $t+5$ who survive to the end of the projection period. Female births, in turn, are obtained by applying the maternity rates to the women exposed to the risk of having a child between t and $t+5$. Thus, the number of girls under 5 at the end of the projection period depends on the initial population of women in the reproductive ages, their survival probabilities, the maternity rates, and the survival probabilities for female births.

There are essentially two ways to proceed, and both yield exactly the same answer. Here I will focus on women and average the rates that apply to them over the projection interval. (The textbook describes an alternative approach that focuses on the rates and averages the women exposed to them.)

Consider, then, women age 15-19 at the start of the projection, who will become 20-24 if they survive. This cohort is exposed to the rates at 15-19 and to the rates at 20-24, with the later discounted by the probability of surviving to be 20-24. The relevant maternity rates are then ${}_5F_{15}$ and ${}_5F_{20}\;\; {}_5L_{20}/\,{}_5L_{15}$, which we average and multiply by 5, the width of the period. The resulting births have to survive from birth to age 0-4, which occurs with probability ${}_5L_0/5l_0$. (This ratio is shown in the first row of the table.) Putting all of this together, the contribution from women age $(x, x+5)$ to girls under 5 at $t+5$ is

$$ {}_5P_x \frac{5}{2}\left({}_5F_x + {}_5F_{x+5}\frac{{}_5L_{x+5}}{{}_5L_x}\right)\frac{{}_5L_0}{5l_0} $$

Note that the 5's cancel out; but I left them in for clarity. To obtain the total number of girls under 5 we sum these contributions over all reproductive ages. The results are shown in the last two columns of the table. You now have a complete projection for 1998.

The Leslie Matrix

The calculations required can be laid out conveniently as the result of multiplying a projection matrix by a population vector:

$$ \boldsymbol{p}_{t+5} = \boldsymbol{L}\,\boldsymbol{p}_t $$

Here $\boldsymbol{p}_t$ is a column vector with the population at time t in each of the k age groups, and $\boldsymbol{L}$ is a k by k projection matrix known as the *Leslie* matrix, with entries given by the coefficients we have just derived. I will not try to write the entire matrix in symbols (the textbook does so in equation 6.10, although it treats the open-ended group differently by using one more value of T_x than we have here). Instead, you can see it in full glory in the companion computing logs.

The Leslie matrix can be used to project the population repeatedly. If you do that you will find that the population will continue to grow (or decline until it becomes extinct), but the age distribution eventually will stop changing. What's more interesting, the final age distribution depends on the Leslie matrix (and hence on the fertility and mortality rates) but not on the initial age distribution. This result is the jewel in the crown of mathematical demography.

Stable Populations

Stability means that the population at time $t+5$ is just proportional to the population at time t. In symbols, for sufficiently large t

$$ \boldsymbol{p}_{t+5} = \boldsymbol{L}\,\boldsymbol{p}_t = \lambda\boldsymbol{p}_t $$

If we write $\boldsymbol{p}$ for the population vector in a stable population, the equation becomes simply

$$ (\boldsymbol{L} - \lambda\boldsymbol{I})\boldsymbol{p} = \boldsymbol{0} $$

where $\boldsymbol{I}$ is the identity matrix. This is a well-known equation in matrix algebra. It has a solution only if the determinant of $\boldsymbol{L} - \lambda\boldsymbol{I}$ is zero, that is, if

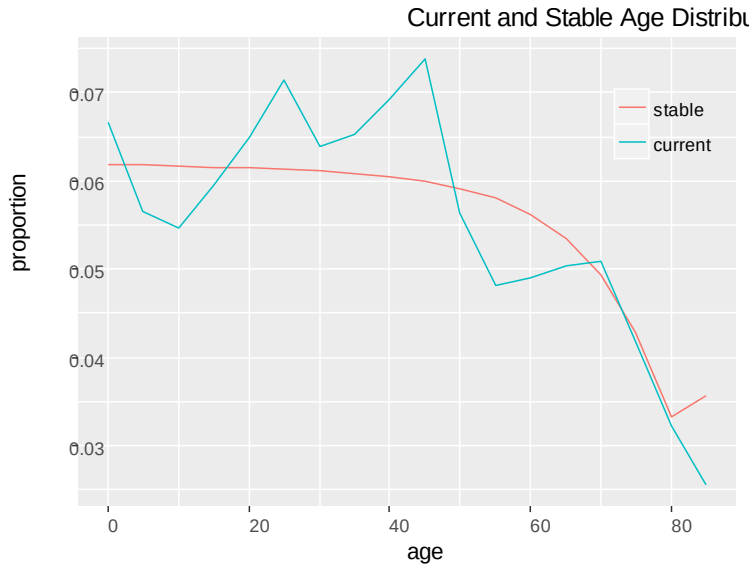
$$ |\boldsymbol{L} - \lambda\boldsymbol{I}| = 0 $$

This equation is called the *characteristic equation* of the matrix $\boldsymbol{L}$. For a real matrix $\boldsymbol{L}$ the equation has n roots λ_i called the *eigenvalues* of $\boldsymbol{L}$. The corresponding vectors $\boldsymbol{p}_i$ are called the *eigenvectors* of $\boldsymbol{L}$.

In demography we are mostly interested in the first eigenvalue and eigenvector. For most (reasonable) Leslie matrices the first eigenvalue and the corresponding eigenvector are real. The eigenvalue is one if

the population is stationary; values above and below one indicate growth and decline respectively. The implied instantaneous growth rate is called the *intrinsic growth rate* or *Lotka's* r. The corresponding eigenvector is proportional to the stable age distribution.

Using a matrix algebra package one can compute eigenvalues and eigenvectors directly. For the Leslie matrix in our example the first eigenvalue is λ =1.00111253, which over 5 years is equivalent to a growth rate of $r = \ln(\lambda)/5$ =0.000222383; so at 1993 rates Sweden would end up growing by 0.02% per year. The first eigenvector, scaled so the entries add to one, is shown on the sidebar and the graph below. If you project the Swedish population for about 100 years the age distribution will look very similar to this vector. (And that's about as close as we'll get to proving any of these results.)



	Current	Stable
0	0.0667	0.0619
5	0.0565	0.0618
10	0.0546	0.0617
15	0.0594	0.0616
20	0.0649	0.0614
25	0.0715	0.0613
30	0.0640	0.0611
35	0.0652	0.0608
40	0.0692	0.0605
45	0.0739	0.0600
50	0.0563	0.0592
55	0.0481	0.0580
60	0.0489	0.0562
65	0.0504	0.0535
70	0.0508	0.0494
75	0.0418	0.0429
80	0.0323	0.0332
85+	0.0256	0.0356

When we return to stable population theory we will learn simple yet accurate methods for estimating Lotka's r and the stable age distribution from standard demographic data.

Open Populations

Migration complicates the picture. Out-migration can be handled by introducing death and emigration as competing risks. In-migration is harder to model (particularly if you assume that immigrants may be subject to different fertility and mortality schedules, at least initially) and is usually handled by making assumptions about absolute numbers rather than rates. See Section 6.3.3 in the textbook for a description of how the closed-population procedures can be adapted to handle immigration.

Stable Populations

POP 502 / ECO 572/ SOC 532 • SPRING 2017

A closed population subject to constant age-specific birth and death rates eventually becomes *stable,* with a constant growth rate and a constant age-distribution. We have already encountered the idea of a *stationary* population, where death rates are constant and there is a steady stream of births. In a stable population the birth stream growths exponentially over time. We review briefly some of the main ideas, focusing on the standard female dominant model. The textbook has an excellent discussion in Chapter 7.

The Renewal Equation

Let $B(t)$ denote the number of female births at time t, and let $N(a,t)$ be the number of women age a at time t. (These are both densities, so strictly speaking the number of births in a short interval of time is the product of $B(t)$ times the width of the interval.) Suppose that starting at time zero the age-specific birth and death rates become constant.

Let $m(a)$ denote the maternity function at any time after zero. If the reproductive span runs from ages α to β the number of births at time t is

$$B(t) = \int_{\alpha}^{\beta} N(a,t)m(a)da\,, \qquad t > 0$$

Between time zero and α the only women having children were already born at time zero, so they are products of whatever fertility and mortality regimes existed before. Between time α and β we have a mix of old timers and women who have been born in the new regime. But when we reach time β and beyond, all women in the reproductive ages have been born in the new regime; they are the product of the fertility and mortality schedules in the model.

Consider then $N(a,t)$, the number of women aged a at time t for $t > \beta$. These are the survivors of the cohort born at time $t-a$, which had initial size $B(t-a)$. Let $p(a)$ the probability of surviving to age a for someone born after time zero. We can then write

$$N(a,t) = B(t-a)p(a), \qquad t > \beta$$

and the stream of births becomes

$$B(t) = \int_{\alpha}^{\beta} B(t-a)p(a)m(a)da\,, \qquad t > \beta$$

This is an integral equation (an equation involving a function and its integral). We would like to solve it, by which we mean finding a function $B(t)$ that satisfies it. Lotka tried an exponential form, where

$$B(t) = Be^{rt}, \qquad t > 0$$

For $t > \beta$ we can also write $B(t-a) = Be^{r(t-a)} = Be^{rt}e^{-ra}$. If we substitute these results into the integral equation and cancel Be^{rt}, which appears on the left and right hand sides, we get after time β

$$1 = \int_{\alpha}^{\beta} e^{-ra}p(a)m(a)da$$

This is *Lotka's equation*. Note that instead of integrating from α to β we can integrate from 0 to t as the textbook does in equation 7.5. The integral from 0 to α is zero, and as long as $t > \beta$ the integral from β to t is also zero because $m(a)$ is zero outside the reproductive ages, so we only need to integrate over the reproductive ages. The textbook uses this fact in equation 7.10.

The next step is to see if this equation has a solution. Write ρ for r in the right-hand side and view that as a function of ρ. We'll assume that the survival and maternity schedules are well-behaved, so the integral is a continuous differentiable function of ρ. The function is always positive and declines monotonically as ρ increases, going from ∞ down to zero as ρ goes from $-\infty$ to $+\infty$. This means that there will be a value of ρ for which the function is one. This is Lotka's r, the *intrinsic growth rate.*

Estimating Lotka's r

Let us write the right hand side of Lotka's equation as a function of ρ

$$f(\rho) = \int_{\alpha}^{\beta} e^{-\rho a}p(a)m(a)da$$

We want a value of ρ such that $f(\rho) = 1$. The first derivative of this function w.r.t. ρ is

$$f'(\rho) = -\int_{\alpha}^{\beta} a\, e^{-\rho a}p(a)m(a)da$$

This derivative looks like a weighted mean age with weights $e^{-\rho a}p(a)m(a)$, except that we haven't divided by the sum of the weights, which is of course $f(\rho)$. Let us define the *mean age of childbearing*

$$A(\rho) = \frac{\int_{\alpha}^{\beta} a\, e^{-\rho a}p(a)m(a)da}{\int_{\alpha}^{\beta} e^{-\rho a}p(a)m(a)da}$$

We can then write the derivative as

$$f'(\rho) = -f(\rho)A(\rho)$$

This, by the way, shows that the function declines monotonically, as the derivative is always negative.

Coale proposed an iterative procedure for solving this equation. The method can be justified starting from a Taylor series expansion of $f(\rho)$ around the *solution*, where

$$f(\rho) \approx f(r) + (\rho - r)f'(r)$$

Solving for Lotka's *r* we obtain

$$r \approx \rho - \frac{f(\rho) - f(r)}{f'(r)}$$

Using the fact that $f'(r) = -A(r)f(r)$ and $f(r) = 1$ this equation simplifies to

$$r \approx \rho + \frac{f(\rho) - 1}{A}$$

where A is an approximation to the mean age of childbearing in the stable population, for example 27. This is the equation in Box 7.1 in the textbook.

An alternative direct application of Newton's method is to expand $f(r)$ around a trial value, so that

$$f(r) \approx f(\rho) + (r - \rho)f'(\rho)$$

Solving for r this equation and recalling that $f(r) = 1$ leads to

$$r \approx \rho + \frac{1 - f(\rho)}{f'(\rho)}$$

The two expansions are in fact equivalent but Newton uses the exact derivative at the trial value while Coale approximates it using an estimate of the derivative at the solution. Using the actual derivative often speeds convergence but, more importantly, yields the mean age of childbearing as a by-product.

Lotka himself used a quadratic approximation that requires no iteration but is less accurate than the iterative procedures.

In practice we need to use discrete data. With age groups of width n we will usually approximate the integral using midpoints and the usual survival and maternity functions, so

$$f(\rho) \approx \sum_{\alpha}^{\beta-n} e^{-\rho\left(x+\frac{n}{2}\right)} \frac{{}_nL_x}{l_0} \, {}_nF_x^F$$

and $f'(\rho) = -f(\rho)A(\rho)$, where the mean age of childbearing is estimated as

$$A(\rho) \approx \sum_{\alpha}^{\beta-n} \left(x + \frac{n}{2}\right) e^{-\rho\left(x+\frac{n}{2}\right)} \frac{{}_nL_x}{l_0} \, {}_nF_x^F / f(\rho)$$

(In both cases we are approximating the integral inside an age group $(x, x+n)$ by evaluating the integrand at the mid point $(x + n/2)$ and multiplying by the width of the interval n. The survival ratios are estimated as $p\left(x + \frac{n}{2}\right) = {}_nL_x/nl_0$. But the two n's cancel out, so I didn't show them.)

Box 7.1 in the textbook and the computing logs obtain an intrinsic r of 0.01424 for Egypt in 1977 after three iterations of Coale's method. The alternative procedure described here gives the same result and as a bonus gives the mean age of childbearing in the stable population, which is 29.47

The Stable Equivalent Age Distribution

Once we have an estimate of Lotka's r we can compute the stable age distribution. The population age a at time t for sufficiently large t (so that everyone has been born in the new regime) is

$$N(a,t) = B(t-a)p(a) = Be^{rt}e^{-ra}p(a)$$

The total population at time t can be obtained by integrating over all ages

$$N(t) = \int_0^\infty N(a,t)da = Be^{rt}\int e^{-ra}p(a)da$$

The proportion of the population age a at time t is then

$$c(a,t) = \frac{N(a,t)}{N(t)} = \frac{e^{-ra}p(a)}{\int e^{-rx}\,p(x)dx}$$

and *doesn't depend on* t, so we'll simply write $c(a)$.

We can simplify this a bit further if we think in terms of the instantaneous birth rate at time t, which is births divided by population:

$$b(t) = \frac{B(t)}{N(t)} = \frac{1}{\int e^{-ra}p(a)da}$$

where I have cancelled Be^{rt} in the numerator and denominator. The birth rate *doesn't depend on* t, so I will now write simply b. Moreover, the denominator of b is the same as the denominator of $c(a)$, so we can write

$$c(a) = b\,e^{-ra}p(a)$$

When working with discrete data we employ the usual mid-point approximations, so having obtained r we compute

$${}_nc_x = b\;\; e^{-r\left(x+\frac{n}{2}\right)}\frac{{}_nL_x}{l_0}$$

For the open-ended group one uses $x + e_x$ as the 'midpoint' and T_x instead of ${}_nL_x$ to approximate the integral. The birth rate is obtained as a normalizing constant such that the relative age distribution adds to one.

Box 7.2 in the textbook and the online computing logs calculate Lotka's r and the stable age distribution for U.S. females in 1991 using these procedures. Very similar results can be obtained from the first eigenvalue and eigenvector of the Leslie matrix. The online supplements also include a graph comparing the current and stable equivalent age distributions. The intrinsic growth rate is (slightly) negative, but the age structure is relatively young. As a result, we find that at 1991 rates the U.S. female population would have continued to growth for about 45 years before heading into extinction.

Why a Population Converges to Stability

My favorite proof of the basic theorem of stable population theory is due to Brian Arthur. His article is very clear and has the great merit of revealing the mechanism involved. While his proof is in discrete time, the gist of the argument can be conveyed equally well in continuous time.

Recall from page 1 that after time β the age distribution is given by

$$c(a,t) = \frac{B(t-a)p(a)}{\int_x B(t-x)p(x)dx}$$

Arthur notes that it is sufficient to show that the birth sequence eventually becomes exponential, say $B(t) \to Be^{rt}$, because if that is the case the age distribution becomes

$$c(a,t) = \frac{Be^{r(t-a)}p(a)}{\int_x Be^{r(t-x)}\,p(x)dx} = \frac{e^{-ra}p(a)}{\int_x e^{-rx}\,p(x)dx}$$

as Be^{rt} cancels out and we obtain an expression that doesn't depend on t!

The next insight comes from the observation that the birth sequence will become exponential if the ratio of $B(t)$ to Be^{rt} becomes a constant. Dividing the left and right-hand sides of the renewal equation by Be^{rt} we obtain

$$\frac{B(t)}{Be^{rt}} = \int_\alpha^\beta \frac{B(t-a)}{Be^{rt}} p(a)m(a)da$$

We can view the left-hand side as a growth-corrected birth sequence $\hat{B}(t) = B(t)/e^{rt}$. To obtain a similar expression on the right-hand side we multiply and divide by e^{-ra}, which leads us to

$$\hat{B}(t) = \int_\alpha^\beta \hat{B}(t-a)\,e^{-ra}p(a)m(a)da$$

So far the algebra holds for any value of r, but now we pick Lotka's r, which has the nice property that $\int e^{-ra}p(a)m(a)da = 1$. The reason why this is important is that we can now write

$$\hat{B}(t) = \int_\alpha^\beta \hat{B}(t-a)\,w(a)da$$

where $w(a)$ represents weights that integrate to one. In other words, the growth-corrected births at time t are a weighted average of the growth-corrected births in the past, where averaging is over a sliding window determined by the reproductive ages.

We can view this successive averaging as a form of smoothing because $\hat{B}(t)$, being a mean, is always *inside* the range of values between α and β years ago (unless they are all equal and the sequence has already converged). As time goes by and the window shifts we discard old values in favor of averages, until the range inevitable collapses and the sequence converges to a constant.

Population Momentum

POP 502 / ECO 572/ SOC 532 • SPRING 2017

This unit focuses on population momentum, the notion that most of the world population would continue to grow even if fertility dropped suddenly to replacement level.

The Preston-Guillot Method

The textbook illustrates a method due to Preston and Guillot. The calculations are reproduced in the computing logs. We start from a maternity function ${}_nm_a$ and divide it by the NRR of 1.7028, assuming a proportionate decline to replacement level. We then compute the mean age of the net maternity schedule (new or old) which turns out to be 26.6. These quantities are computed as

$$NRR = {}_nL_x^F\ {}_nF_x^F \quad \text{and} \quad \bar{a} = \sum_x (x + n/2)\ {}_nL_x^F\ {}_nF_x^F\ /NRR$$

The next step is to compute the weight function, representing the ratio of the expected number of births that would occur above (the mid-point of) each age to the mean age,

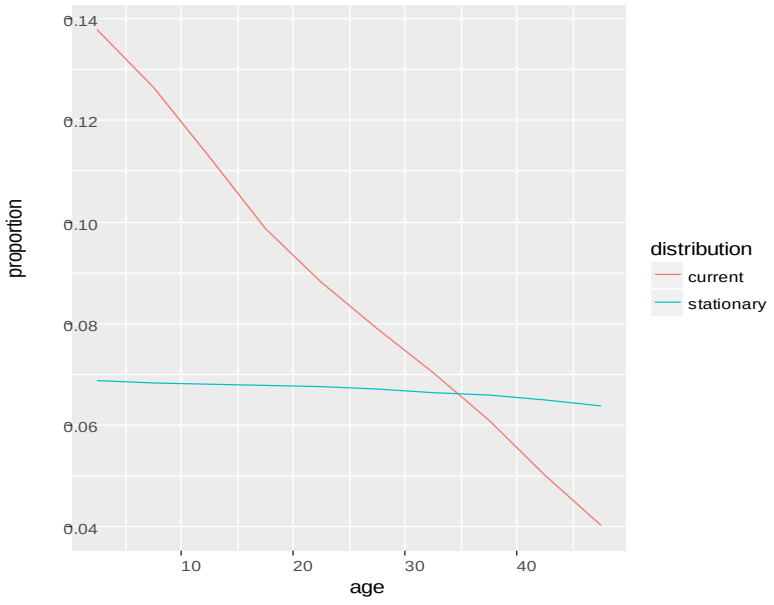
$${}_nw_x = (0.5\ {}_nL_x^F\ {}_nF_x^F\ + \sum_{x+5}\ {}_nL_a^F\ {}_nm_{ax})/\bar{a},$$

Finally we multiply by the ratio of the current and stationary equivalent age distributions and sum, so

$$M = \sum_a\ {}_nw_a\ {}_nc_a/\ {}_ns_a$$

where ${}_nc_a = {}_nN_a^F/N$ is the current age distribution and ${}_ns_a = {}_nL_a^F/e_0^F$ is the stationary equivalent.

We find that even if fertility dropped immediately to replacement level the female population would still growth 61%. The reason is apparent if we plot the current and stationary age distributions. The large proportions at young ages in the current age distribution compared to the stationary equivalent ${}_nL_a/e_0$ are the engine behind the momentum, particularly when weighted by the fact that most of their fertility is ahead of them.



Keyfitz's Approximation

The idea of population momentum originated with Keyfitz, and proved quite influential in policy circles. His formulation assumed that the population was stable at the outset and the reduction in rates was proportionally the same at all ages. The previous development doesn't require either assumption, and is therefore more general. Because Keyfitz's method is still popular, however, we apply it to Western Asia.

In the original 1971 paper momentum is given by

$$M = b\frac{e_0}{r\,m}\left(\frac{NRR-1}{NRR}\right)$$

where b is the birth rate, r is the rate of natural increase, m is the mean age of childbearing in the stationary population, and NRR is the net reproduction ratio, all before the fall to replacement level.

James Frauenthal pointed out that the formula was very nearly

$$M = b\frac{e_0}{\sqrt{NRR}}$$

and it is this simpler formula that is often used. (See Keyfitz and Caswell, pp. 197-198 for details.)

For our example the CBR is 0.029 and the formula gives $0.029 \times 79.3/\sqrt{1.7028} = 1.57$, which is pretty close to the more exact value of 1.61 computed above. Not bad for a simple calculation.

Reproductive Value

In class we provided some additional background on these results, starting from Fisher's notion of reproductive value. We discount future births at an annual rate r, so that the present value of a newborn 's future childbearing over the reproductive ages is

$$\int_\alpha^\beta e^{-ra}\,p(a)m(a)da$$

Setting the "interest rate" to Lotka's r gives a present value of 1. You may think of this as a newborn repaying her own life. The amount "owed" by a woman age x is known as her *reproductive* value and is

$$v(x) = \int_x^\beta e^{-r(a-x)}\,\frac{p(a)}{p(x)}m(a)da$$

Moving the terms on x out of the integral we can also write this as

$$v(a) = \frac{1}{e^{-rx}p(x)}\int_x^\beta e^{-ra}p(a)m(a)da$$

The Stable Equivalent Population

Each population has a *stable equivalent*, the population that would emerge if fertility and mortality stayed constant for a long time. That population has a growth rate r given by Lotka's r, intrinsic birth and death rates b and d, and a constant age structure $c(a)$, see equations 7.8, 7.9 and 7.10 in the textbook. What we don't know yet is its size. We will define the size as

$$Q = \lim_{t\to\infty} \frac{P_t}{e^{rt}}$$

One way to think about this construction is to project the population until it becomes stable and then reverse-project it at a constant rate r. The resulting population is stable and will eventually become indistinguishable from the target population, with the same size, age structure and vital rates.

Keyfitz shows that we can write the size of the stable equivalent population as

$$Q = \frac{\int_0^\beta n(x)v(x)dx}{b_r\, A_r}$$

where $n(x)$ is the female density at age x, $v(x)$ is Fisher's reproductive value, b_r is the birth rate and A_r is the mean age of childbearing in the stable population. This is an important formula because we can obtain the population momentum results from it, so let us write it in full glory as

$$Q = \int_0^\beta \frac{n(x)}{e^{-rx}p(x)} \int_x^\beta e^{-ra}p(a)m(a)da\, dx/(b_r A_r)$$

All we have done here is plug in Fisher's reproductive value.

Population Momentum

Suppose the maternity rates $m(a)$ were to change to replacement-level rates $m_0(a)$; for example we could set $m_0(a) = m(a)/\mathrm{NRR}$ changing all rates by the same proportion, but this restriction is not necessary. Suppose further that fertility and mortality then stay constant. Eventually the population will become stationary, and we can obtain its size from the general formula with $r = 0$

$$S = \int_0^\beta \frac{n(x)}{p(x)} \int_x^\beta p(a)m_0(a)da\, dx/(b_0 A_0)$$

where b_0 is the birth rate and A_0 the mean age of childbearing in the stationary population. Write $n(x) = Pc(x)$ where P and $c(x)$ are the current population size and age distribution, and note that $p(x) = c_0(x)/b_0$ where $c_0(x)$ is the stationary age distribution. Dividing both sides by P we then obtain momentum or S/P as

$$M = \int_0^\beta \frac{c(x)}{c_0(x)} \int_x^\beta p(a)m_0(a)da\, dx/A_0$$

This is the Preston-Guillot formula 7.21 in a slightly different guise, as taking A_0 inside the integral makes it their weight function $w(x)$. Note that they use a subscript s and a superscript * for the stationary population where I use a subscript 0. The main point here is that the mysterious weight function comes from Fisher's reproductive value.

Keyfitz Momentum

As noted earlier, Keyfitz's formula assumes that the population is already stable and fertility is reduced by the same proportion at all ages, so that $m_0(a) = m(a)/\text{NRR}$. We can obtain his result starting from the general formula. Write $n(x) = Pc_s(x)$ where $c_s(x)$ is the current stable age distribution and $p(x) = c_0(x)/b_0$ as we did before to obtain

$$M_k = \int_0^\beta \frac{c_s(x)}{c_0(x)} \int_x^\beta p(a)m(a)da\,dx/(A_0\text{NRR})$$

Recall that the stable age distribution is $c_s(x) = b_r e^{-rx}p(x)$ and the stationary is $c_0(x) = b_0 p(x)$ so the survival probabilities cancel. We can also write the stationary birth rate as $b_0 = 1/e_0$, so

$$M_k = \frac{b\ e_0}{A_0\text{NRR}} \int_0^\beta e^{-rx} \int_x^\beta p(a)m(a)da\,dx$$

Changing the order of integration and noting that $\int_0^a e^{-rx}dx = \frac{1}{r}(1 - e^{-ra})$ we obtain

$$M_k = \frac{b\ e_0}{rA_0\text{NRR}} \int_0^\beta (1 - e^{-ra})p(a)m(a)da$$

We then recognize the two terms in the integral as the NRR and Lotka's equation, so the result becomes

$$M_k = \frac{b\ e_0}{r\ A_0} \frac{\text{NRR} - 1}{\text{NRR}}$$

which is exactly Keyfitz's formula, writing A_0 instead of m for the stable mean age of childbearing.

Stable and Non-stable Momentum

Tom Espenshade and collaborators have a nice paper on population momentum in *Demography* in 2011 which I found very helpful. They call Keyfitz's formula the stable momentum, driven by differences between the stable and stationary age distributions, and Preston-Guillot's formula the total momentum, driven by differences between the current and stationary age distributions. They derive a formula for non-stable momentum, driven by differences between the current and stable age densities, and show that to a very close approximation total momentum is the product of stable and non-stable momentum.

In our example the Keyfitz formula worked well because most of the momentum in western Asia was stable. For the world as a whole, momentum around 2005 was 1.398, with 1.173 stable and 1.193 non-stable (for a product of 1.399), so the approximation would not be adequate. For the least developed countries, however, most of the momentum is stable (1.468 of 1.513).

Indirect Estimation

POP 502 / ECO 572/ SOC 532 • SPRING 2017

Our last topic is indirect estimation, a subject covered in Chapter 11 of the textbook and in much greater detail in the United Nation's Manual X, recently updated in *Tools for Demographic Estimation*. We focus on some of the seminal contributions of Bill Brass.

Fertility and P/F

Let $f(a)$ denote fertility at age a and $F(a) = \int_{15}^{a} f(x)dx$ cumulative fertility up to age a. If fertility has been relatively constant in the recent past one could estimate $f(a)$ from age-specific fertility rates for the last year and $F(a)$ from questions on children ever born by age of mother, and the two sets of estimates should be consistent.

Brass postulated, however, that these two sources of data are subject to different sources of error. Specifically, children ever born is subject to recall errors that increase with age, so $F(a)$ may be considered reliable only for younger ages. On the other hand, reports of births in the past year are subject to time scale errors, referring to periods longer or shorter than a year; if these errors are independent of age then $f(a)$ will have the wrong level but the right shape.

The basic idea of the procedure is to get the level from parity and the shape from fertility, hence the name P/F, usually read as "P over F". Specifically, the method starts with mean children ever born at a young age, typically 20-24, as the estimate of P. It then accumulates age-specific fertility rates up to the same age, for example 5 times the 15-19 rate plus 2.5 times the 20-24 rate, as the estimate of F. The third step is to calculate the P/F ratio and use this to inflate the age-specific fertility rates, effectively correcting the level while preserving the shape.

In countries with good data, such as England and Wales in 1951, Brass finds P/F ratios very close to one. In Africa, however, he finds more dispersion, for example 0.8 in Guinea and 1.13 in Uganda, suggesting reference period errors.

The table on the right shows an illustrative calculation using Brass's interpolation factors to accumulate ASFRs (for example for the age group 20-24 we use 2.695 instead of 2.5). The P/F ratio at 20-24 suggests that fertility is about 33% *higher* than reported, and leads to a revised TFR of 5.1 instead of 3.9.

Age	f	k	F	P	P/F	f*
15-19	0.021	1.345	0.028	0.038	1.345	0.028
20-24	0.170	2.695	0.563	0.747	1.326	0.225
25-29	0.195	2.865	1.514	1.892	1.250	0.259
30-34	0.172	3.085	2.461	2.884	1.172	0.228
35-39	0.124	3.200	3.187	3.560	1.117	0.164
40-44	0.067	3.405	3.638	3.868	1.063	0.089
45-49	0.022	4.020	3.833	3.868	1.009	0.029
TFR	3.855		3.855			5.114

There are better interpolation factors to accumulate fertility up to the middle of an interval, depending on the ratio of the rates at 20-24 and 15-19. A variant uses P/F ratios for first births to derive a correction factor. An adjustment is also needed when fertility rates are based on births last year by current age of woman, rather than true event-exposure rates. Models may play a role here. Manual X uses the Coale-Trussell model, and *Tools for Demographic Estimation* emphasizes the use of relational Gompertz models, and has a worked example with average parity and period fertility rates from the Malawi 2008 Census.

Schmertmann and collaborators have a nice 2013 paper in *Population Studies* on "Bayes plus Brass" to estimate total fertility for many small areas using sparse census data, with an application to 2000 Brazilian Census data for over five thousand municipalities. Their algorithm first uses Bayesian techniques to smooth local age-specific rates, and then applies a variant of Brass's P/F method that is robust under conditions of rapid fertility decline.

Child Mortality from Reports of Children Surviving

Brass proposed a method for estimating child mortality from mother's reports of children ever born and children surviving, that quickly became the main source of child mortality estimates in the developing world. The basic idea is to ask a mother how many children she has given birth to, and how many are still alive. These questions have been added in many censuses, and are usually known as "the Brass questions".

Let $f(a)$ denote fertility at age x and $p(a)$ denote the probability of surviving from birth to age a. If a mother is now age a, she is expected to have had $F(a) = \int_{15}^{a} f(x)dx$ children. A child born when the mother was age x was born $a - x$ years ago and has a probability $p(a - x)$ of being alive today. The expected number of children surviving is then

$$S(a) = \int_{15}^{a} f(x)p(a - x)dx.$$

By the mean value theorem, we should be able to approximate the last integral by

$$S(a) = p(a - x^*) \int_{15}^{a} f(x)dx$$

where x^* is some value between 0 and a. Under these assumptions, the ratio of children surviving to children ever born is

$$\frac{S(a)}{F(a)} = p(a - x^*)$$

and is a direct estimate of the life table probability of surviving to some age $a - x^*$, that depends on the age of the mother and the average time since her children were born. (Compare this with equation 11.3 in the text, noting that I used a for the age of the mother in what I think is a simpler explanation.)

Obviously young women must have had their children relatively recently. Brass noted that women aged 15-19 had their children on average a year ago, so the ratio estimates $p(1)$, the probability of surviving to age one. For women 20-24 it estimates $p(2)$, and for women 30-34 it estimates $p(5)$, see the table on page 228 of the textbook for more details. These values, however, need to be adjusted depending on the age pattern of fertility.

Brass developed a set of correction factors using simulation. These were later revised by Sullivan and then by Trussell. The factors are in the form of regression coefficients that take as inputs the ratios of mean parities 15-19/20-24 and 20-24/25-29. The result is a correction factor that is multiplied by the proportion dead to yield an estimate of the appropriate ${}_nq_0$ for each age of mother. (See Table 11.1 in the textbook. Note that b_i should be $-.5381$ at age 20-24.).

A second problem is that the ratio of children surviving to children ever born depends on mortality conditions in the past. If mortality has not been constant, then estimates for older ages refer to periods further in the past.

Coale and Trussell developed formulas for estimating the period to which a set of estimates refer, based on an assumption of linearly declining mortality. These are also in the form of regression coefficients that take as input the ratio of mean CEB at ages 15-19/20-24 and 20-24/25-29. The result is an estimate of the period to which the estimates apply. (See Table 11.2 in the textbook.)

The table below shows calculations for the data from Zimbabwe in 1994 found in Box 11.1 in the textbook. The basic inputs are the mean parities and the proportions of children dead by age of mother.

Age	Parity	Prop dead	Child age	Mort adj	q(x)	Ref period
15-19	0.170	0.0560	1	1.080	0.0605	1.0
20-24	1.100	0.0817	2	1.050	0.0858	2.3
25-29	2.360	0.0760	3	1.001	0.0760	4.2
30-34	3.890	0.0847	5	1.009	0.0855	6.5
35-39	5.130	0.0935	10	1.027	0.0960	9.0

The relevant ratios of mean parities are 0.17/1.1=0.155 and 1.1/2.365=0.465. Using these as inputs in the regression equation for adjusting the data for women aged 15-19 we get a correction factor of 1.08. Multiplying this by the proportion dead for women 15-19 we get an estimate of ${}_1q_0 = 0.0605$. Plugging the same two ratios in the regression equation for the reference period we get 1.0, so we estimate that the probability of infant death was about 60 per thousand, approximately one year before the survey. Calculations for the other age groups proceed along the same lines. The age group 30-34 leads to an estimate of ${}_5q_0 = 0.0855$, so under five mortality was about 86 per thousand, and we time this estimate around 6.5 years before the survey.

Tools for Demographic Estimation uses data on children ever born and children surviving from the 2008 Census of Malawi to illustrate the method, and relies on relational logits to convert estimates to ${}_5q_0$ at various times in the past.

Adult Mortality from Data on Orphanhood

Essentially the same logic used to estimate child mortality can be used to estimate adult mortality from reports of orphanhood. Let $B(t)$ denote the birth density at time t and $p(a)$ the probability of surviving to age a. The number of people age a at time t is

$$N(a,t) = B(t-a)p(a),$$

the number of births $t-a$ years ago times the probability of surviving a years.

The probability that a person age a at time t will not be a maternal orphan is $p_M(a)$, the probability that a woman would survive a years after giving birth. The density of people age a at time t whose mother is alive is then

$$NO(a,t) = B(t-a)p(a)p_M(a)$$

and the ratio of non-orphans to the total population age a at time t is

$$\frac{NO(a,t)}{N(a,t)} = p_M(a)$$

a direct estimate of the probability that a mother would survive a years after giving birth.

The next question is how to relate this ratio to life table survival probabilities. The answer depends on the average age of mothers given the age of their offspring. To a first approximation $p_M(a)$ is the probability of surviving a years from the mean age of childbearing M^* to age M^*+a , or l_{M^*+a}/l_{M^*} . If mean age of childbearing was 27.5, then the proportion non-orphan among respondents 15-19 would estimate the probability of surviving from age 27.5 to 45 (or 27.5+17.5).

Just as was done for the children surviving method, Hill and Trussell developed a set of regression equations for converting the proportions non-orphaned by age into survival probabilities using simulation, based on model schedules of fertility and mortality. The equations take as input the mean age of mothers at childbirth, and the proportion of people in the age group whose mothers are alive. The survival ratios estimated range from l_{45}/l_{25} for the age group 15-19 to l_{60}/l_{25} for 30-34.

Unfortunately, the method depends on the assumption of constant mortality and there is no reliable procedure to date the estimates if mortality has been declining. A potential source of bias is selectivity, induced by the fact that only surviving children can report their orphanhood status. For younger respondents there may also be an "adoption effect", where children report the survivorship of their adopted rather than their biological mother. Alternative methods rely on the survival of siblings or spouses, but they tend to be less accurate for adult mortality.

Tools for Demographic Estimation has an application of the orphanhood method to data from Iraq. They also have an extensive discussion of the impact of the HIV/AIDS on mortality estimation, with an illustration from Kenya.

The Sisterhood Estimate of Maternal Mortality

The last indirect method we will mention estimates maternal mortality from reports of sisters. The following brief description borrows heavily from a note I wrote with Trussell. The textbook describes the procedure in more detail in section 11.3 and has an example using data from Gambia in 1987.

The basic idea of the method is to take a sample of women and ask how many sisters have ever married and, of these, how many (if any) have died during pregnancy, childbirth or puerperium. In populations where sexual relations outside marriage are common, or where marriage itself is not well defined, inquiries can be made about sisters past menarche or past age 15; the basic idea remains the same.

If the sample consists of women aged 60 or over, the simple fraction of sisters who died of maternal causes turns out to be an estimator of the lifetime risk of maternal mortality in the presence of other causes of death. If the sample includes women under 60, however, some of their sisters are still at risk of maternal mortality, so an inflation factor must be used to convert the fraction dead to a lifetime risk. Appropriate adjustment factors have been computed using standard fertility and mortality schedules. Estimates typically refer to mortality conditions about 12 years before the survey.

A key assumption of the method is independence between the number of siblings and their survival probabilities, as well as independence of the mortality experiences of adult sisters. An interesting feature of the method is the fact that the sampling frame appears to count the experience of some women multiple times. In fact, an early DHS survey restricted the sampling frame so only one sister per household was allowed to answer the maternal mortality note. It turns out, however, that restricting the sampling frame introduces biases; multiple reporting is not only simpler, by not requiring linking sisters who may live in different households, but essential for the success of the technique.

Tools for Demographic Estimation has a worked example using data from the Malawi 2004 DHS to estimate pregnancy-related mortality. They note that sampling uncertainty is very large compared to estimates of under-5 mortality, so while estimates of levels may be useful, interpretation of differentials is hazardous, and any conclusions about trends should be based on estimates from two or more surveys.

Demography General

Useful Formulae

Spring 2017

Population growth:

$$N(T) = N(0)e^{\int_0^T r(t)dt},$$

$$N(T) = N(0)e^{rT}, \quad r = \log\left(\frac{N(T)}{N(0)}\right)/T$$

Hazard and survival:

$$\mu(x) = \frac{d(x)}{l(x)}, \qquad l(x) = l(0)e^{-\int_0^x \mu(a)da},$$

$$e(x) = \int_x^\infty l(a)da/l(x)$$

Rates to probabilities:

$${}_nq_x = \frac{n\,{}_nm_x}{1 + (n - {}_n a_x)\,{}_nm_x}, \qquad {}_nq_x = 1 - e^{-n\,{}_nm_x}$$

Time lived:

$${}_nL_x = n\,l_{x+n} + {}_na_x\,{}_nd_x, \qquad {}_nL_x = \frac{l_x - l_{x+n}}{{}_nm_x}, \qquad {}_\infty L_x = \frac{l_x}{{}_\infty m_x}$$

Gompertz:

$$\log\mu(x) = \alpha + \beta x$$

Brass:

$$Y_x = \alpha + \beta Y_x^s \quad \text{where} \quad Y_x = 0.5\log\frac{l_0 - l_x}{l_x} \quad \text{so} \quad l_x = l_0\frac{1}{1 + e^{2Y_x}}$$

Unobserved heterogeneity:

$$\mu(x|\theta) = \mu_0(x)\theta, \qquad \mu(x) = \mu_0(x)E(\theta|X > x)$$

Lee-Carter:

$$\log\,{}_nM_{x,t} = a_x + b_x\,k_t$$

Coale-McNeil:

$$G(a) = cGs(\frac{a - a_0}{k}) = cG_0(\frac{a - \mu}{\sigma})$$

Hernes:

$$g(a) = Ae^{-ra}G(a)[1 - G(a)]$$

Coale and Coale-Trussell:

$$r(a) = Mn(a)e^{-mv(a)}, v(a) \geq 0 \quad \text{and} \quad f(a) = G(a)r(a)$$

Page:

$$r(a, d) = n(a)e^{\alpha+\beta d}$$

Bongaarts:

$$\mathrm{TFR} = \mathrm{TN}\, C_m\, C_c\, C_i\, C_a$$

Leslie matrix:

$$\frac{{}_nL_0}{l_0}({}_nF_x + \frac{{}_nL_{x+n}}{{}_nL_x}\,{}_nF_{x+n})/2$$

$$\frac{{}_nL_{x+n}}{{}_nL_x}, \qquad \frac{T_{x+n}}{T_x}$$

Net reproduction rate:

$$\mathrm{NRR} = \int_\alpha^\beta p(a)m(a)da \approx \mathrm{GRR}\, p(A_M)$$

Stationary population:

$$be_0 = 1, \qquad c(a) = p(a)/e_0$$

Renewal equation:

$$B(t) = \int_\alpha^\beta B(t-a)p(a)m(a)da, t > \beta$$

Lotka's equation:

$$1 = \int_\alpha^\beta e^{-ra}p(a)m(a)da$$

Stable age distribution:

$$c(a) = be^{-ra}p(a)$$

"Sheer Poetry":

$$\frac{d\log c(a)}{dr} = A_P - a, \quad \text{where} \quad A_p = \int ac(a)da / \int c(a)da$$

Stable population birth rate:

$$b = 1/\int_0^\infty e^{-ra}p(a)da$$

Mean length of a generation (T):

$$r = \frac{\log(\mathrm{NRR})}{T}, \qquad T \approx \frac{A_B + \mu}{2}$$

Population momentum

$$M = \int_0^\beta \frac{c(a)}{c_s(a)}w(a)da \quad \text{(P-G)} \qquad M = \frac{be_0}{\sqrt{\mathrm{NRR}}} \quad \text{(K-F)}$$

Tempo-adjustment:

$$\mathrm{TFR}^* = \frac{\mathrm{TFR}}{1-r}$$