1 Basic Concepts and Measures

1.1 Meaning of "Population"

To a statistician, the term "population" refers to a collection of items, for example, balls in an urn. Demographers use the term in a similar way to denote the collection of persons alive at a specified point in time who meet certain criteria. Thus, they may refer to the "population of India on April 1, 1995," or to the "population of American black females in the Northeast on June 1, 1900." In both cases the criteria for inclusion in the population need further elaboration: do we count "legal residents" or simply those who can be found within the borders on that date? What do we mean by "black," or by "Northeast"? Do we refer to midnight or noon on the specified date? It is clear that "the population of India on April 1, 1995" is a shorthand description of what may be a rather long set of operational choices designed to minimize blurriness at the boundaries.

But demographers also use the term "population" to refer to a different kind of collectivity, one that persists through time even though its members are continuously changing through attrition and accession. Thus, "the population of India" may refer to the aggregate of persons who have ever been alive in the area we define as India and possibly even to those yet to be born there. The collectivity persists even though a virtually complete turnover of its members occurs at least once each century.

Demographic analysis focuses on this enduring collectivity. It is particularly addressed to studying changes in its size, its growth rates, and its composition. But while the emphasis is on understanding aggregate processes, demography is also attentive to the implications of those processes for individuals. Many of the indexes in common use in demography, such as life expectancy at birth and the total fertility rate, translate aggregate-level processes into

statements about the demographic circumstances faced by an average or randomly-chosen individual. In turn, a frequent concern in demography is to trace out the consequences of changes in individual-level behavior for aggregate processes. Demography is one of the social science disciplines where micro- and macro-level analyses find perhaps their most complete and satisfactory articulation.

1.2 The Balancing Equation of Population Change

No matter how a population is defined, there are only two ways of entering it: being born into it; or migrating into it. If the definition of the population includes a social element in addition to the customary geographic/temporal elements, then "migration" can include a change in the social label, a process often referred to as "social mobility." For example, the population of American high school graduates can be entered by achieving a high school diploma, a form of social migration or mobility. Note in this example that the population cannot be entered at birth since the acquisition of the label of high school graduate requires the investment of years of life. Populations defined by marital status or occupation are other examples of populations that cannot normally be entered by birth (except for the default options, unmarried and no occupation). On the other hand, populations defined by characteristics fixed at birth, such as sex, race, or nativity, cannot be entered through migration but only through birth. So there are *at most* two ways of entering a population, birth and in-migration (= immigration).

Likewise, there are at most two ways of leaving a population, death and out-migration (= emigration). All populations can be left through death, but only those defined by characteristics not fixed at birth can be exited through migration. If one is born in the United States, one cannot leave the population of persons born in the United States by migration, but one can obviously leave the population resident in the United States by migration.

Because there are only four possible ways of entering or leaving a population, we can be sure that changes in the size of the population must be attributable to the magnitude of these flows. In particular,

$$N(T) = N(0) + B[0, T] - D[0, T] + I[0, T] - O[0, T], \tag{1.1}$$

where

$N(T)$ = number of persons alive in the population at time T,
$N(0)$ = number of persons alive in the population at time 0,
$B[0, T]$ = number of births in the population between time 0 and time T,
$D[0, T]$ = number of deaths in the population between time 0 and time T,
$I[0, T]$ = number of in-migrations between time 0 and time T,
$O[0, T]$ = number of out-migrations from the population between time 0 and time T.

The unit of time in this equation, and throughout the book unless otherwise noted, is number of years. Thus, the time period in which births, deaths, and migrations are occurring is T years in length. T may be fractional and need not be an integer number.

Kenneth Boulding has called this equation the most fundamental in the social sciences. It is clearly an *identity* rather than an approximation or a hypothesized relation. However, when data are used to estimate the elements of this equation, it is no longer the case that both sides must be equal. Error in measuring any element will cause an imbalance in the equation, unless two or more errors happen to be exactly offsetting. An imbalance in the equation is sometimes

Box 1.1 *The Balancing Equation of Population Change*

$N(T) = N(0) + B[0, T] - D[0, T] + I[0, T] - O[0, T]$

Example: Sweden, 1988

Ending population Jan. 1, 1989	Starting population Jan. 1, 1988	Births between Jan. 1, 1988 and Jan. 1, 1989	Deaths between Jan. 1, 1988 and Jan. 1, 1989	In-migrations between Jan. 1, 1988 and Jan. 1, 1989	Out-migrations between Jan. 1, 1988 and Jan. 1, 1989
$N(1989.0)$	$= N(1988.0)$	$+ B[1988.0, 1989.0]$	$- D[1988.0, 1989.0]$	$+ I[1988.0, 1989.0]$	$- O[1988.0, 1989.0]$
8,461,554	= 8,416,599	+ 112,080	− 96,756	+ 51,092	− 21,461

Data source: United Nations, *Demographic Yearbook* (various years).

referred to as an "error of closure." Box 1.1 demonstrates the application of the equation to data from Sweden, which are among the world's most reliable.

1.3 The Structure of Demographic Rates

The balancing equation of population change breaks down the changes in the size of the population into four flows. Each flow is the sum of events or transitions occurring to individuals. Three of the four types of events can be related to an individual present in the population prior to the event. While death and out-migration can be related to one individual, birth can be related to two individual parents, assuming that both belong to the population of interest. Analytical insight can be gained by relating the size of these flows (number of occurrences) to the size of the population producing them. This task is normally accomplished by constructing a demographic "rate."

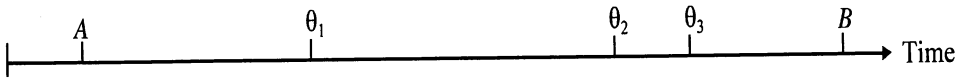
The term "rate" is used in many fields and its meaning is not consistent. An unemployment rate, for example, is simply a *ratio* of the unemployed to the total labor force at a moment in time. In demography, rates are normally (but not invariably) what are known in statistical parlance as "occurrence/exposure rates." The typical form of demographic rates reflects the fact that the frequency of occurrences can be expected to be higher in a larger population, and that the total number of occurrences can also be expected to be higher the longer the members of the population are exposed to the "risk" of the occurrence. The amount of exposure in the denominator of an occurrence/exposure rate combines these two features – the number of persons in the population and the length of the time frame in which exposure is counted. The most conventional occurrence/exposure rate in demography takes the form of:

$$Rate = \frac{Number\ of\ Occurrences}{Person\text{-}years\ of\ Exposure\ to\ the\ Risk\ of\ Occurrence}$$

Demographic rates thus contain in the numerator a count of the number of events occurring within some defined time period, and in the denominator an estimate of the number of "person-years" lived in the population during that time period. The number of person-years

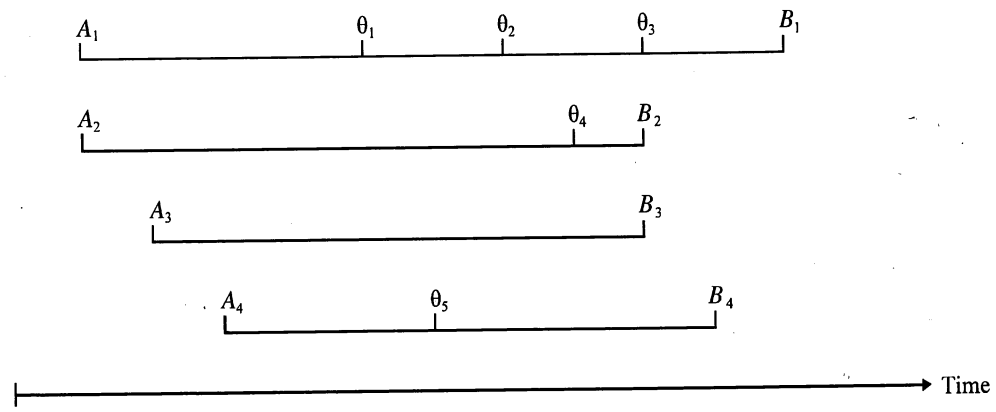
functions in part as an indicator of the population's amount of exposure to the risk of the event, hence the term *occurrence/exposure rate*. When person-years are used in the denominator, a rate is referred to as an "annualized" rate.

Unlike occurrences, the number of person-years lived is rarely directly observed or counted. Nevertheless, the concept is central in demography. To deal with the concept in a population that is continuously changing its membership, it is useful to represent individual exposures as "life-lines." A life-line extends from an individual's birth (A) to the point where he or she experiences some terminal event (B), usually death. Occurrences of interest, θ_i, can be added to the life-line, as illustrated below:



In order to better connect events and exposure to the risk of experiencing the event, a life-line is sometimes restricted: if we are interested in the risk of giving birth, for instance, we may restrict analysis of life-lines to a certain age range. In our exposition, event A and B are simply birth and death respectively, but the concept can readily be extended to other types of bounding events.

For a group of individuals, however the group might be defined, the concept of the occurrence/exposure rate can be illustrated by a set of life-lines for each member of the group G:



where θ_j are the event occurrences in group G and A_i and B_i represent the birth and death of individual i in the group. The rate for the group defined over their entire lifetimes is

$$Rate^G = \frac{\sum_{i \in G} N_i}{\sum_{i \in G} T_i}$$

where N_i is the total number of occurrences in the lifetime of individual i, T_i is the length of time between A_i and B_i, and $\sum_{i \in G}$ is an instruction to take the sum across all individuals (i) who are a member of group G.

1.4 Period Rates and Person-years

A period rate for a population is constructed by limiting the count of occurrences and exposure times to those pertaining to members of the population during a specified period of time:

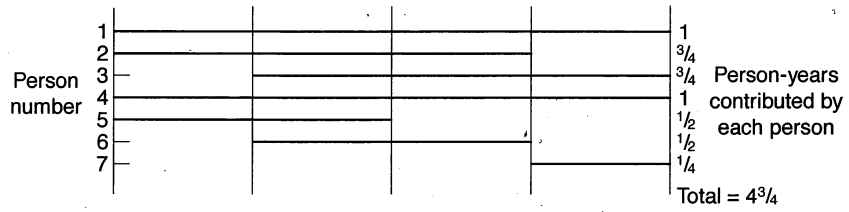
$$Rate\,[0, T] = \frac{Number\ of\ Occurrences\ between\ Time\ 0\ and\ \mathrm{T}}{Person\text{-}years\ Lived\ in\ the\ Population\ between\ Time\ 0\ and\ \mathrm{T}}$$

If a person lives one year between time 0 and time T, he or she has contributed one person-year to the denominator of the period rate. If a person lives 24 hours between 0 and T, he or she has contributed 1/365th of a person-year. The contributions from all individuals who were alive in the population at any time between 0 and T are simply added together in order to produce the denominator for our rates.

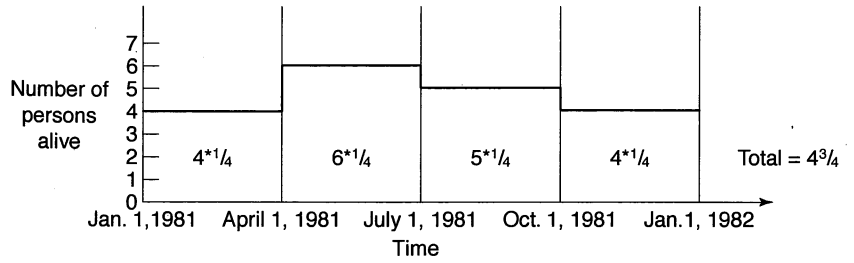
The idea is easily grasped by referring again to life-lines. If we are interested in period 0 to T, all life-lines can be truncated to the "window" 0 to T, since we will not count any occurrences outside that interval. Figure 1.1 shows the life-lines of 7 individuals in a small hypothetical population during the period from 12:00 A.M., January 1, 1981 to 12:00 A.M. on January 1, 1982.

Person 1, for example, is a member of the population for the entire year, whereas person 6 is born on April 1 and dies on October 1, thereby contributing only 6 months or one-half of a person-year to the sum of person-years. Adding exposure across individuals would be a convenient way to estimate person-years lived in country that had a population register which recorded exact dates of birth, death, and migration for each individual.

An alternative method of computing period person-years is to ignore individual histories, such as those provided by a population register, and simply record the number of persons alive



a. Life-lines for seven individuals who live in a population at any time between Jan. 1, 1981 and Dec. 31, 1981



b. Life-lines converted into numbers of persons alive at each moment

Figure 1.1 Demonstration of the equivalence of the two methods for recording person-years

at various points in time during the year. In our example, there were 4 persons alive from January 1, 1981 to April 1, 1981, so that this quarter-year contributed $4(\frac{1}{4}) = 1$ person year. The next quarter contributed $6(\frac{1}{4}) = 1.5$ person-years, and so on to a total of 4.75 person-years contributed during all of 1981. This value is of course the same number derived by following personal histories, as demonstrated in figure 1.1.

In this alternative approach, what we have done is to estimate the area under the $N(t)$ curve between January 1, 1981 and January 1, 1982. $N(t)$ is defined as the number of persons alive at time t. An area is found by taking the height of a figure times its width. In our case, $N(t)$ is the height and the proportion of the year that corresponds to our measurement of $N(t)$ is the width. Since height represents persons and width represents fractions of a year, it is natural to measure the product in units of person-years.

In our example the number of person-years was:

$$PY[1981.00, 1982.00] = 4(.25) + 6(.25) + 5(.25) + 4(.25) = 4.75$$

This sum can be written in conventional notation as:

$$PY[1981.00, 1982.00] = \sum_{i=1}^{4} N_i \cdot \Delta_i$$

where N_i is the number of persons alive in the ith quarter and Δ_i is the fraction of a year represented by that quarter. Had we measured the size of the population each *day* instead of each quarter, the sum would be represented as:

$$\begin{aligned} PY[1981.00, 1982.00] &= N(Jan.\,1, 1981) \cdot \frac{1}{365} \\ &\quad + N(Jan.\,2, 1981) \cdot \frac{1}{365} \\ &\quad + \cdots \\ &\quad + N(Dec.\,31, 1981) \cdot \frac{1}{365} \\ &= \sum_{i=1}^{365} N_i \cdot \Delta_i \end{aligned}$$

If we were able to measure the height, $N(t)$, in tiny intervals of time dt, where dt represents the width of the interval, the area under the curve could be represented more accurately as:

$$PY[1981.00, 1982.00] = \int_{1981.00}^{1982.00} N(t) \cdot dt$$

Here an integral sign has replaced the summation sign and for the fraction of a year represented by the time interval, dt has replaced Δ_i.

We have seen that areas under a curve can be represented in two ways, using either algebraic or calculus notation. In demography, algebraic notation satisfies a practical need that arises when measurement occurs in discrete intervals. But calculus is often preferred for its compact notation and for its far more extensive body of theorems having direct applicability

to population processes. We will use algebra and calculus interchangeably in this volume. One of the most frequent uses of calculus will occur in the issue we have already encountered, representing the area under a curve.

1.5 Principal Period Rates in Demography

We can now apply the concept of period rate to demographic events of interest, in particular the four components of the balancing equation of population change. When the elements of equation (1.1), the balancing equation of population growth, are each divided by the number of person-years lived between 0 and T, we define four rates:

The Crude Birth Rate between times 0 and T:

$$CBR[0, T] = \frac{\text{Number of births in the population between times 0 and } T}{\text{Number of person-years lived in the population between times 0 and } T}$$

The Crude Death Rate between times 0 and T:

$$CDR[0, T] = \frac{\text{Number of deaths in the population between times 0 and } T}{\text{Number of person-years lived in the population between times 0 and } T}$$

The Crude Rate of In-migration between times 0 and T:

$$CRIM[0, T] = \frac{\text{Number of in-migrations into the population between times 0 and } T}{\text{Number of person-years lived in the population between times 0 and } T}$$

The Crude Rate of Out-migration between times 0 and T:

$$CROM[0, T] = \frac{\text{Number of out-migrations from the population between times 0 and } T}{\text{Number of person-years lived in the population between times 0 and } T}$$

We could label the crude birth rate as we have defined it as the "true" crude birth rate, since it includes the actual births and actual person-years in the numerator and denominator, respectively. Throughout the book, the term "rates" will refer to the true or actual rates prevailing in a population. These should be distinguished from the "recorded" or "estimated" rates that are produced when data are used to estimate the value of the true rate.

A person is normally counted as having migrated during the period 0 to T if he or she has changed his or her principal place of residence during the period in a way that crosses the administrative boundaries defining "the population" of a region.

As is especially clear from our definition of the crude rate of in-migration, the connection between exposure and event is not always very precise in demography. No member of a population is literally exposed to the risk of in-migrating into that same population; those at risk are all outside of the population. Like any definitions, these contain an element of arbitrariness, and we could have chosen to put another element in the denominator. What the crude rate of in-migration expresses is the rate at which the population is growing as a result of in-migration. The other rates also indicate the rate at which the population is changing as a result of births, deaths, or out-migration. Using person-years as the denominator for all the major rates in demography provides a firm basis for developing and integrating many different functions and formulas involving population growth. This advantage should become evident in the course of this volume.

It is important to keep in mind the distinction between the reference period to which a rate pertains (i.e., the period for which the values are calculated) and the unit in which exposure time is measured. As noted, the conventional practice is to count exposure in the form of person-years lived, thus creating "annualized" rates. They express the number of events occurring *per year* of exposure. But a period rate need not refer to a single year of the population's experience. For example, we can readily define a crude death rate for 1990–1. Here the number of events in the numerator would include all deaths for calendar years 1990 and 1991, and the denominator would include all person-years lived in 1990 as well as those lived in 1991. Since both the numerator and denominator are, in size, approximately double what they would be if they referred to only a single calendar year, defining the rate over a 2-year period does not affect the scale of the rate. It is still an annualized rate, expressing the number of events per person-year. Likewise, we could define a crude death rate for May 1992, in which both numerator and denominator would be approximately one-twelfth of their value for all of 1992. The scale of the rate, and its annualized nature, is preserved.

Although a period rate in demography apparently can accommodate any length of reference period, it is important to recognize that it must have *some* reference period. The phrase, "the crude birth rate of the United States," has no meaning and there is no way to calculate its value. We must know in what period to count births for the numerator and person-years for the denominator.

1.6 Growth Rates in Demography

1.6.1 Crude growth rate

Let us rearrange the balancing equation of population change (1.1), by subtracting $N(0)$ from both sides and then dividing both sides by the total of person-years lived between 0 and T, $PY[0, T]$:

$$\begin{aligned} \frac{N(T) - N(0)}{PY[0,T]} &= \frac{B[0,T]}{PY[0,T]} - \frac{D[0,T]}{PY[0,T]} + \frac{I[0,T]}{PY[0,T]} - \frac{O[0,T]}{PY[0,T]} \\ CGR[0,T] &= CBR[0,T] - CDR[0,T] + CRIM[0,T] - CROM[0,T] \\ &= \quad CRNI[0,T] \quad + \quad CRNM[0,T] \end{aligned} \tag{1.2}$$

Here we have defined the crude growth rate between 0 and T, $CGR[0, T]$, as the change in the size of population divided by person-years lived between 0 and T. If $N(T)$ exceeds $N(0)$, then the growth rate will be positive; if $N(0)$ exceeds $N(T)$, it will be negative. Clearly, the crude growth rate as we have defined it is simply equal to the crude birth rate minus the crude death rate plus the crude rate of in-migration minus the crude rate of out-migration.

The difference between the crude birth rate and the crude death rate is usually termed the crude rate of natural increase (*CRNI*); also, the difference between the crude rate of in-migration and the crude rate of out-migration is usually termed the crude rate of net migration (*CRNM*). So the crude growth rate will equal the crude rate of natural increase plus the crude rate of net migration. Box 1.2 illustrates the calculation of crude demographic rates, again using the Swedish data in box 1.1 and estimating the person-years lived in 1988 by the population size on July 1, 1988. Table 1.1 presents the estimated value of demographic rates for major regions of the world.

Box 1.2 *Principal Period Rates in Demography*

$$\frac{N(T) - N(0)}{PY[0,T]} = \frac{B[0,T]}{PY[0,T]} - \frac{D[0,T]}{PY[0,T]} + \frac{I[0,T]}{PY[0,T]} - \frac{O[0,T]}{PY[0,T]}$$

$$CGR[0,T] = CBR[0,T] - CDR[0,T] + CRIM[0,T] - CROM[0,T]$$

$$= \quad CRNI[0,T] \quad + \quad CRNM[0,T]$$

Example: Sweden, 1988

Person-years lived in Sweden between January 1, 1988 and January 1, 1989 = 8,438,477 (mid-year population)

$$\frac{N(1989.0) - N(1988.0)}{PY[1988.0, 1989.0]} = \frac{B[1988.0, 1989.0]}{PY[1988.0, 1989.0]} - \frac{D[1988.0, 1989.0]}{PY[1988.0, 1989.0]} + \frac{I[1988.0, 1989.0]}{PY[1988.0, 1989.0]} - \frac{O[1988.0, 1989.0]}{PY[1988.0, 1989.0]}$$

$$CGR[1988.0, 1989.0] = CBR[1988.0, 1989.0] - CDR[1988.0, 1989.0] + CRIM[1988.0, 1989.0] - CROM[1988.0, 1989.0]$$

$$\frac{8{,}461{,}554 - 8{,}416{,}599}{8{,}438{,}477} = \frac{112{,}080}{8{,}438{,}477} - \frac{96{,}756}{8{,}438{,}477} + \frac{51{,}092}{8{,}438{,}477} - \frac{21{,}461}{8{,}438{,}477}$$

$$0.00533 = 0.01328 - 0.01147 + 0.00605 - 0.00254$$

$$CGR[1988.0, 1989.0] = CRNI[1988.0, 1989.0] + CRNM[1988.0, 1989.0]$$

$$0.00533 = 0.00182 + 0.00351$$

Data source: United Nations, *Demographic Yearbook* (various years).

The crude growth rate is only one of several types of growth rate encountered in demography. The term "growth rate" is used to refer to other measures as well, and it is important to distinguish the various forms.

1.6.2 Instantaneous growth rate

As any rate, the crude growth rate can be computed for any period of time. What happens when we compute the growth rate during a very short period of time, between time t and $t + \Delta t$, as Δt approaches 0? Denote the population change, $N(t + \Delta t) - N(t)$, as $\Delta N(t)$ and the growth rate as $r(t)$. Since the person-years lived over the period $[t, t + \Delta t]$ is now $N(t)\Delta t$, the crude growth rate for the period is $r(t) = \Delta N(t)/N(t)\Delta t$. But the limit of $\Delta N(t)/\Delta t$ when Δt approaches 0 is simply the derivative of the $N(t)$ function, designated $dN(t)/dt$. Therefore:

$$r(t) = \lim_{\Delta t \to 0} \frac{\Delta N(t)}{N(t)\,\Delta t} = \frac{\dfrac{dN(t)}{dt}}{N(t)} = \frac{d \ln[N(t)]}{dt} \tag{1.3}$$

where "ln" refers to the natural logarithm. The time interval is very short, dt years, so that $r(t)$ pertains to the tiny interval of time between t and $t + dt$. Because it is measured in time units of years, $r(t)$ continues to be an annualized rate. It is referred as "the growth rate at time t" or "the instantaneous growth rate at time t." It is, of course, also the crude growth rate in the tiny interval of time from t to $t + dt$.

Table 1.1: *Population size and components of change in major areas of the world, 1995–2000*

Major area	*Population size (thousands)*		*Births (thousands)*	*Deaths (thousands)*	*Net international migrants (thousands)*	*Crude growth rate (percentage)*	*Crude birth rate (per 1000)*	*Crude death rate (per 1000)*	*Crude rate of natural increase (per 1000)*	*Crude rate of net migration (per 1000)*
	1995	2000	1995–2000	1995–2000	1995–2000	1995–2000	1995–2000	1995–2000	1995–2000	1995–2000
World	5,666,360	6,055,049	649,050	260,360	0	1.33	22.1	8.9	13.2	0.0
Africa	696,963	784,445	140,575	51,655	−1,435	2.37	38.0	13.9	24.1	−0.4
Asia	3,436,281	3,682,550	389,765	137,460	−6,035	1.38	21.9	7.7	14.2	−0.3
Europe	727,912	728,887	37,465	41,240	4,750	0.03	10.3	11.3	−1.0	1.3
Latin America and the Caribbean	479,954	519,143	57,770	16,225	−2,355	1.57	23.1	6.5	16.6	−0.9
Northern America	296,762	309,631	20,860	12,640	4,650	0.85	13.8	8.3	5.5	3.1
Oceania	28,488	30,393	2,635	1,135	405	1.30	17.9	7.7	10.2	2.8

Source: United Nations, 1999.

The concept of the instantaneous growth rate enables us to develop a new expression for population change over a longer time interval. Integrating formula (1.3) between exact times 0 and T (also measured in years), gives:

$$\int_0^T r(t)\,dt = \int_0^T \frac{d \ln N(t)}{dt} dt = \ln N(t)\Big]_0^T$$

So:

$$\int_0^T r(t)\,dt = \ln\left(\frac{N(T)}{N(0)}\right) \tag{1.4}$$

Taking exponentials on both sides we have:

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

or

$$N(T) = N(0)e^{\int_0^T r(t)\,dt} \tag{1.5}$$

Formula (1.5) is extremely important in demography. It appears in many guises in many different applications. It expresses the change in population size during a particular discrete time period (in this case from 0 to T) as a simple function of the set of instantaneous growth rates that prevailed during that period. Note that the proportionate growth in population over the period, $N(T)/N(0)$, is a simple function of the sum of growth rates. The order in which those growth rates are applied is immaterial; all that matters is their sum.

Viewing $r(t)$ as a continuously varying function raises questions about the commonly encountered term, "exponential growth." Any growth that occurs, including zero growth or negative growth, must obey equation (1.5). An exponential appears in that formula because we have *defined* our measure of growth – the growth rate – in proportionate terms. In this sense the term "exponential growth" is a redundancy; all growth is exponential by our measure of growth as the proportionate rate of change in population size. When people use the term "exponential growth" they are often (but not invariably) referring to an $N(t)$ sequence produced by a *constant positive* growth rate within some time interval. Such a sequence is probably more precisely characterized by the term Malthus chose for it, "geometric growth," or by "constant growth rate." If the instantaneous growth rate is in fact constant between time 0 and time T at some value r^*, then equation (1.5) simplifies to:

$$N(T) = N(0)e^{r^* \cdot T} \tag{1.6}$$

This formula follows from the fact that:

$$\int_0^T r^*\,dt = r^*\Big]_0^T = r^* \cdot T - r^* \cdot 0 = r^* \cdot T$$

Rearranging equation (1.6) and taking natural logarithms gives:

$$r^* = \frac{\ln\left(\frac{N(T)}{N(0)}\right)}{T} \tag{1.7}$$

Equation (1.7) shows that, if the instantaneous growth rate is constant during the interval 0 to T, one can solve for its value by observing the population size at the beginning and end of the interval.

1.6.3 Mean annualized growth rate

If we divide both sides of equation (1.4) by T, the length of the time interval over which growth is occurring, we have:

$$\frac{\int_0^T r(t)\,dt}{T} = \frac{\ln\left[\frac{N(T)}{N(0)}\right]}{T}$$

The left-hand side of this equation is simply the mean value of the instantaneous growth rate over the period 0 to T, which we will designate as $\bar{r}[0, T]$. It is the area under the $r(t)$ function between 0 and T, divided by the length of the time interval. Thus:

$$\bar{r}[0, T] = \frac{\ln\left[\frac{N(T)}{N(0)}\right]}{T} \tag{1.8}$$

Note that the right-hand side of equation (1.8) is identical to that of (1.7); if the growth rate is constant between 0 and T, equation (1.8) provides a way of estimating its value. But (1.8) is clearly a more general expression since it requires no assumption of constancy. Performing the simple operation given by the right-hand side of equation (1.8) provides the "mean annualized growth rate between 0 and T."

1.6.4 Doubling time

If population size doubles between time 0 and time T, then $N(T)/N(0) = 2$ and:

$$\ln[N(T)/N(0)] = \ln[2] = .693$$

A population thus doubles in size beyond some initial date whenever the sum of its annualized growth rates beyond that date equals 0.693. If the growth rate is constant at r^*, the population will double whenever the product of r^* and T, the length of time (in years) over which it is applied, is 0.693.

So with constant growth rate r^*,

$$\textit{Doubling time} = \frac{.693}{r^*}$$

Under a constant annual growth rate of 0.03, the population will double in $.693/.03 = 23.1$ years. With a constant growth rate of 0.01, it will double in $.693/.01 = 69.3$ years. Since $e^{-.693} = 1/e^{.693} = 0.5$, a population will be reduced to half of its initial size whenever the sum of annual growth rates equals $-.693$.

1.6.5 Comparison of crude growth rate and mean annualized growth rate

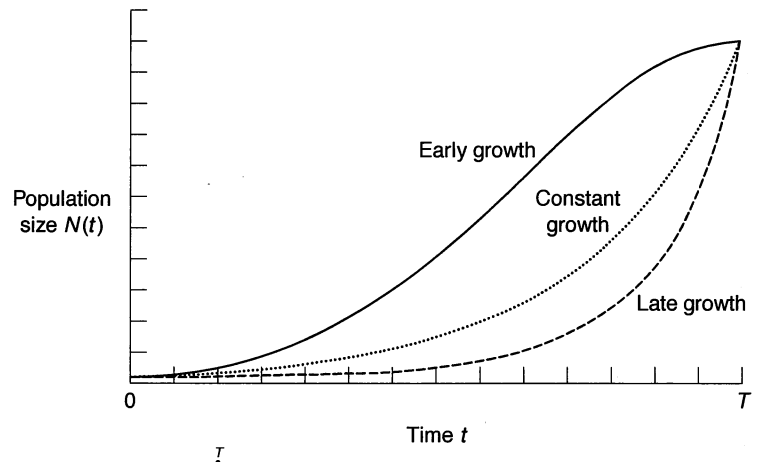
We have now developed two formulas for period growth rates over the discrete interval between 0 and T: the crude growth rate and the mean annualized growth rate. This section, which is included for completeness and can be skipped by many readers, compares the two rates. The basic lesson is that the two growth rates will be the same when the instantaneous growth rate is constant during the period 0 to T. Otherwise, the two rates will not, in general, have the same value. However, differences between them can usually be ignored for practical purposes unless the period of measurement is very long (say, longer than 10 years) *and* the growth rate function, $r(t)$, is very irregular.

From (1.2), the crude growth rate between 0 and T can be written as:

$$CGR[0, T] = \frac{B[0, T] - D[0, T] + I[0, T] - O[0, T]}{\int_0^T N(t)\,dt} = \frac{N(T) - N(0)}{\int_0^T N(t)\,dt} \tag{1.9}$$

As is clear in (1.8), $\bar{r}[0, T]$ does not depend on the order in which growth rates occur between 0 and T. The numerator of $CGR[0, T]$ in (1.9) is also independent of the order in which growth rates occur. But the denominator of $CGR[0, T]$ in (1.9), person-years lived between 0 and T, *does* depend on the order in which growth rates occur. A distribution of positive growth rates that is heavily skewed toward the beginning of the period will raise person-years lived relative to a distribution that is skewed toward the end of the period. This tendency is illustrated in figure 1.2.

So it is clear that, in general, there can be no equality between CGR and $\bar{r}$. An "early" distribution of growth rates will lower CGR relative to $\bar{r}$, and a "late" distribution will



The sum of growth rates, $\int_0^T r(t)dt$, is the same in the three cases, since $N(0)$ and $N(T)$ are the same.

Person-years lived – the area under the $N(t)$ curve – are different, however.

Figure 1.2 Population growth sequences between times 0 and T under three different assumptions about the time sequence of growth rates

raise *CGR* relative to $\bar{r}$. There is, however, one circumstance in which *CGR* will equal $\bar{r}$. This occurs when the growth rates are constant between 0 and T. Suppose that $r(t) = r^*$ for $0 \leq t \leq T$. Then:

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

$$= N(0) \cdot \frac{1}{r^*} \cdot e^{r^*t}\Big]_0^T = \frac{N(0) \cdot e^{r^*t}]_0^T}{r^*}$$

$$= \frac{N(T) - N(0)}{r^*} \qquad \textbf{(1.10)}$$

Substituting expression (1.10) for person-years lived between 0 and T into equation (1.9) gives:

$$CGR[0, T] = \frac{N(T) - N(0)}{\left[\dfrac{N(T) - N(0)}{r^*}\right]} = r^*$$

In the case of a constant growth rate, we also have:

$$\bar{r}[0, T] = \frac{1}{T}\int_0^T r^*\,dt = r^*$$

So in the case of constant growth rates – and, except for rare circumstances, only in this case – the crude growth rate will equal $\bar{r}$. Differences between the two will normally be trivial in size unless the growth rate sequence is extremely erratic and the time period (0 to T) very long, say a decade or more.

If one wants to ensure that the crude growth rate calculated by (1.9) is in fact equal to the mean of annualized growth rates, then a simple rule for computing person-years is indicated: compute person-years lived during the period as though the growth rate were constant throughout. Under this circumstance, the denominator for calculating all crude rates would be:

$$\int_0^T N(t)\,dt = \begin{cases} \dfrac{N(T) - N(0)}{\bar{r}[0, T]} = \dfrac{[N(T) - N(0)] \cdot T}{\ln\left(\dfrac{N(T)}{N(0)}\right)}, & \text{if } \bar{r} \neq 0 \\ T \cdot N(0), & \text{if } \bar{r} = 0 \end{cases}$$

Although we defined the "mean annualized growth rate" as the average of period rates, in equation (1.8) it does not have person-years in the denominator, which was said to be a typical feature of a demographic rate. In this format, it shares the characteristic of many rates in common usage, such as a mean rate of speed or mean rate of inflation. But under the simplifying assumption that the "mean annualized growth rate" is constant during the interval of measurement, its value is in fact identical to that of the crude growth rate, which does explicitly contain person-years in the denominator.

1.7 Estimating Period Person-years

The above argument suggests that, if one knew nothing about the path of $N(t)$, or $r(t)$, during a particular year, one should assume constancy of the growth rate during the period and estimate person-years lived during the year as:

$$PY[0, 1] = \frac{N(1) - N(0)}{r[0, 1]} = \frac{N(1) - N(0)}{\ln\left[\dfrac{N(1)}{N(0)}\right]}$$

More generally, when the period is not necessarily one year long,

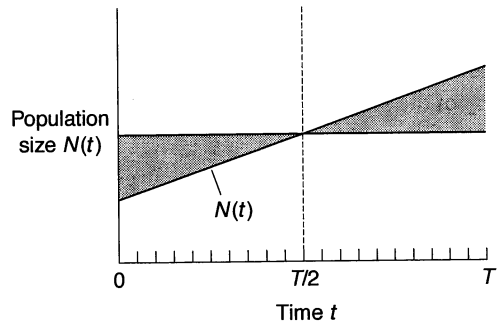
$$PY[0, T] = \frac{[N(T) - N(0)] \cdot T}{\ln\left[\dfrac{N(T)}{N(0)}\right]} \qquad \textbf{(1.11)}$$

Using equation (1.11) to estimate person-years has the advantage of forcing consistency between the crude growth rate for the period and the mean annualized growth rate for that period, and it would be exactly correct if the growth rate were constant during the period. But it does require observations on population size at the beginning and end of the period. It is often the case (e.g., in the United States) that population size estimates are only available at mid-year. It will usually be perfectly acceptable to use the mid-year population size as an estimate of person-years lived during the year. The mid-year approximation to person-years will be exactly correct if the $N(t)$ sequence is linear between the beginning and end of the year, as demonstrated in figure 1.3. Even if the $N(t)$ sequence is a product of a constant growth rate, the error in using the mid-year approximation will be very small. For example, if $r = 0.03$ (rapid by historical standards), the ratio of true person-years lived in a year to the mid-year population will be 1.00004. The mid-year population will always underestimate the true number of person-years lived if the population is changing at a constant rate, whether positive or negative.

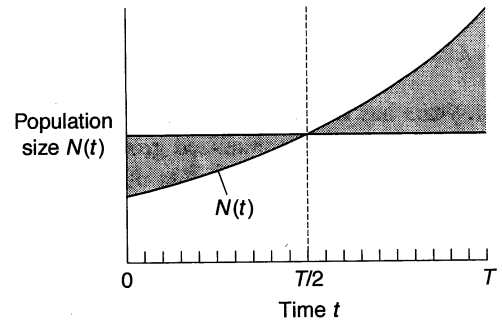
More caution is necessary in using mid-period approximations to estimate person-years when the interval of time for which an estimate is sought extends far beyond a year. For example, if we estimate the person-years lived during a 10-year period in a population growing at 3 percent a year by taking the mid-period population times 10 (i.e., mid-height times width), then the ratio of true person-years lived to our estimated person-years will be 1.0038. This error of about four-tenths of 1 percent is too large to ignore for most purposes. Note that if we had used the arithmetic mean of beginning and end-period populations (times 10) as our estimate of person-years lived in this example, we would have *overestimated* true person-years by the factor 1.0075. So this procedure provides an even poorer estimate of person-years than does the mid-period population in a population having a constant positive growth rate.

If mid-year population estimates are available for each year during a 10-year period, a sensible way to estimate person-years lived during the period would be simply to add up the 10 estimates. If observations are available at the beginning, middle, and end of the period, then it is possible to ascertain whether growth is more nearly linear or exponential and to use the corresponding approximation for each half-period.

Although it is convenient and fairly accurate to estimate person-years lived during a particular year as the population size in the middle of the year, it is important to remember that the resulting demographic rate should not be expressed as a number of occurrences divided by a



When $N(t)$ follows a linear growth pattern, the estimate of person-years lived using the mid-period population times period length will be accurate because the overestimate for the first half-period is exactly offset by the underestimate for the second half-period, i.e., the two triangles have equal areas.



When $N(t)$ follows an exponential growth pattern, the two shaded surfaces have different areas and the mid-year approximation, $N(T/2) \cdot T$, will underestimate person-years lived during the period.

Figure 1.3 Approximation of person-years lived by midperiod population times period length

number of people. The unit in which exposure-time is measured (usually, person-years) must not disappear, or confusion is inevitable. We are using the mid-year population as an *estimate* of person-years lived during the period, and not as a *substitute* for person-years. The risk of confusion is greatest when an annualized rate is being estimated for a period that is not one year in length. Box 1.3 illustrates the computation of growth rates and person-years lived during a 10-year period in a hypothetical population with a constant annualized growth rate of 0.03.

1.8 The Concept of a Cohort

Almost as important to demography as the concept of a population is the concept of a cohort. A cohort is the aggregate of all units that experience a particular demographic event during a specific time interval. As in the case of a population, a cohort always has some specific geographic referent, whether it is explicit or implicit. A cohort usually consists of people, but it may also consist of entities (e.g., marriages) formed by a demographic event. The cohort is usually identified verbally both by the event itself and by the time period in which it is experienced. Some examples of cohorts are:

"US birth cohort of 1942," which refers to all persons born as US citizens in calendar year 1942;

Box 1.3 *Illustration of Calculation of Growth Rates and Person-years*

Suppose that a population had 100,000 persons at time 0 and that it grew at a constant annualized growth rate of 0.03. Then:

$N(0) = 100,000$

$N(5) = 100,000 \cdot e^{5 \cdot .03} = 116,183$

$N(10) = 100,000 \cdot e^{10 \cdot .03} = 134,986$

1. Calculating the mean annualized growth rate between $t = 0$ and $t = 10$:

$$\bar{r}[0, 10] = \frac{\ln\left(\frac{N(10)}{N(0)}\right)}{10} = \frac{\ln\left(\frac{134,986}{100,000}\right)}{10} = 0.0300$$

2. Estimating person-years lived between $t = 0$ and $t = 10$:

   a) Assuming a constant growth rate:

$$PY[0, T] = \frac{N(T) - N(0)}{\bar{r}[0, T]} = \frac{N(10) - N(0)}{\bar{r}[0, 10]} = \frac{134,986 - 100,000}{0.03} = 1,166,200$$

   b) Assuming growth is linear and using the mid-period approximation:

$$PY[0, T] = N(T/2) \cdot T$$
$$PY[0, 10] = N(5) \cdot 10 = 116,183 \cdot 10 = 1,161,830$$

   c) Assuming growth is linear and using the mean of initial and final population sizes:

$$PY[0, T] = \left[\frac{N(0) + N(T)}{2}\right] \cdot T$$
$$PY[0, 10] = \left[\frac{N(0) + N(10)}{2}\right] \cdot 10$$
$$= \left[\frac{100,000 + 134,986}{2}\right] \cdot 10 = 1,174,930$$

3. Calculating crude growth rates based upon various estimates of person-years lived:

   a) $CGR[0, T] = \frac{34,986}{1,166,200} = 0.0300$

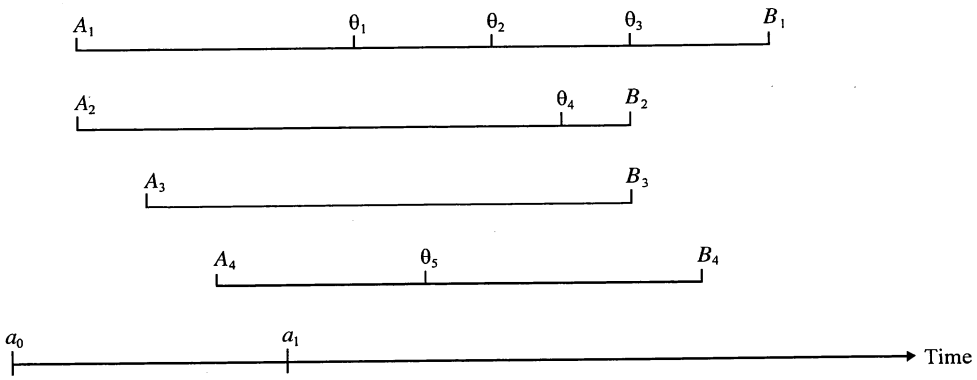
   b) $CGR[0, T] = \frac{34,986}{1,161,830} = 0.0301$

   c) $CGR[0, T] = \frac{34,986}{1,174,930} = 0.0298$

"French marriage cohort of 1990," which refers to all marriages contracted in France during the calendar year 1990;
"French female marriage cohort of 1990," which refers to all women who married in France in 1990;
"Austrian immigrant cohort of 1995," which refers to all immigrants into Austria in 1995.

The most frequently encountered type of cohort is a birth cohort. Persons who are born during the same period are destined to pass through life together, in the sense that they will reach their xth birthday during a period exactly x years beyond that which defined their cohort membership. For the US birth cohort of 1942, all would reach their 10th birthday (assuming that they survived) in 1952, their 15th birthday in 1957, and so on. The time period that circumscribes the cohort need not be one year in length; it is common to deal with such entities as the US birth cohort of 1918–22, for example.

To calculate a rate for a cohort, we simply restrict the counting of occurrences and person-years of exposure to people who were born during the period that defines membership in the cohort. The lines below show the counting schema for a birth cohort defined by birth in the period a_0 to a_1:



Although those life-lines refer to a birth cohort, the concept can clearly be extended to other types of cohorts.

1.9 Probabilities of Occurrence of Events

We can define an additional concept for a cohort that is impossible for a population: the concept of a probability. The term is used in demography in a manner similar to its usage in statistics. It refers to the *chance* that some event will occur, rather than to the rate at which it occurs. Thus, for example, we may compute the probability that a marriage would end in a divorce for a given birth cohort by counting, over all members of the cohort, the number of marriages and the number of divorces over the cohort's lifetime:

$$q^D = \frac{\textit{Number of Divorces}}{\textit{Number of Marriages}}$$

In doing so, we have used a "relative frequency" approach to estimating the probability of divorce. We have said, in effect, that our best guess about the true underlying probability

of divorce in the cohort is the observed frequency of divorce. The situation is analogous to drawing balls out of a very large urn. If we draw a sample of 10 balls and 2 of them are red, then the relative frequency of red balls in that drawing is 0.2. This relative frequency is also the maximum likelihood estimator of the true proportion of balls in the entire urn that are red, assuming that the outcome of drawings is independent. That is, a proportion in the entire urn of 0.2 is more likely than any other proportion to have given rise to the observed sample of 10 balls of which 2 are red. Many introductory statistics texts contain a clear discussion of maximum likelihood estimation.

The structure of a probability in demography is thus quite different from the structure of a rate:

$$\textit{Rate} = \frac{\textit{Number of Occurrences}}{\textit{Number of Person-years Lived}}$$

$$\textit{Probability} = \frac{\textit{Number of Occurrences}}{\textit{Number of Preceding Events or Trials}}$$

The denominator of the probability indicates that it is not possible to define a probability unless there is some *event* or *trial* (equivalent to the act of drawing balls out of an urn). Since each occurrence in the numerator (e.g., divorce) must be preceded by an event in the denominator (marriage), the number of occurrences cannot exceed the number of preceding events. Thus the probability cannot exceed one and, since we are only dealing with positive quantities, probabilities cannot be negative.

Populations do not have probabilities except insofar as they pertain to cohorts that are included in the population. Although we could count the number of marriages in a population during some calendar year and the number of divorces during that year, the two numbers combined do not give a sensible estimate of the probability of divorce because they don't apply to the same cohort. We are, in effect, counting events (or trials) in one urn and occurrences in another. If we happened to choose a year in a small population where no one married but there was a divorce, our population's probability of divorce q^D would be $1/0 = \infty$, an obviously absurd outcome. Only when we count the events pertaining to the cohort at risk of the event can we properly define a probability.

The concepts of cohorts and of probabilities that certain events will occur to cohorts can be applied to a vast number of situations extending well beyond demography's customary range. They are central to all analysis of longitudinal data in the social and health sciences. Perhaps their major utility derives from the fact that they translate aggregate-level measures into implications for individuals. They help "locate" the individual in an otherwise amorphous and undifferentiated population.

Despite its conceptual simplicity, analysis of data on actual cohorts suffers from several major practical limitations. First, computing cohort rates and probabilities requires complete information on each individual until he or she has died (or at least has ceased to be "at risk" of the event of interest). We may lose track of some individuals, for instance, when they move out of the area of the study. Out-migration is part of a more general problem called "loss to follow-up." We deal with one way of coping with this problem in chapter 4. A more serious practical problem is that, by the time the cohort's experience is completely observed, much of the experience may be ancient. In order to provide more timely information, demographers rely primarily on data for recent periods. The measures that are constructed from period rates

include life expectancy, expected years to be lived in the single state, total fertility rate, net reproduction rate and gross reproduction rate. They also include probabilities of dying, giving birth, migrating, and so on. In constructing these and other measures, demographers rely on the concept of a cohort, but adapt that concept to deal with data pertaining to a period. The principal adaptation is the introduction of "hypothetical cohorts," a concept that will be encountered frequently in the remainder of this volume.

2 Age-specific Rates and Probabilities

In nearly every population, the rate of occurrence of demographic events varies very sharply with age. In fact, the rates defined in chapter 1 are called "crude" rates precisely because they fail to account for age variation in the underlying rate schedules. In the case of mortality and fertility, this variation mainly reflects age differences in physiological capacity. Age variation in migration rates seems to reflect primarily age differences in the economic and social gains from movement.

Because of this age variation, it is common to define and study age-specific rates. These have the same structure as crude rates, with a count of events in the numerator and of person-years in the denominator. However, the age range within which the events and person-years are to be tallied is restricted.

2.1 Period Age-specific Rates

The following notation is conventional for defining a period age-specific death rate:

$$ {}_nM_x[0, T] = \frac{\text{Number of deaths in the age range } x \text{ to } x+n \text{ between time 0 and } T}{\text{Number of person-years lived in the age range } x \text{ to } x+n \text{ between time 0 to } T} $$

Note that, just like the crude death rate, a period age-specific death rate must pertain to some specified time period.

It is clear from the definition that x, the right subscript of ${}_nM_x[0, T]$, refers to the age at the beginning of the age interval and n, the left subscript, to the length of the interval. Both are measured in *exact number of years*. That is, they refer to the elapsed time since one's birth in years, including decimal or fractional years. So ${}_5M_{30}$, the death rate between ages 30 and 35, refers to events occurring to and person-years lived by persons who are aged 30.0000 to 34.9999 in exact years since birth. This concept of exact years of age differs from the one in normal use in most countries. When asked their age, most people give a number indicating

how many years of life they have *completed*. That is, they omit the decimals altogether. This latter concept of age is sometimes termed "age last birthday." If the data one uses are classified in terms of age last birthday, then the ages (at last birthday) to which ${}_5M_{30}$ pertains are 30, 31, 32, 33, and 34. Often the analyst will have to determine which age grouping is being used in published data by observing whether the age ranges are stated as 30–5, 35–40, 40–5... (i.e., in exact age) or as 30–4, 35–9, 40–4... (age at last birthday).

Table 2.1 displays the number of deaths by age for females in Sweden, 1992, as well as the estimated mid-year population by age. The format uses age last birthday. The data are converted into age-specific death rates in the fourth column (M_i^{Sw}), using the mid-year population as the estimate of person-years lived in an age interval. The table also shows the same information for Kazakhstan, 1992.

Note that the crude death rate, shown at the bottom of the table as the death rate for "all" ages, is higher in Sweden than in Kazakhstan (0.01055 vs. 0.00742). This result seems on the face of it inconsistent with the fact that the age-specific death rates in Sweden are lower than those in Kazakhstan *at every age*. To understand this apparent anomaly, let us show explicitly how the crude death rate is related to age-specific death rates. Designate ${}_nN_x$ as the number

Table 2.1: *Comparison of crude death rates and age-specific death rates in two populations*

Sweden, females, 1992					*Kazakhstan, females*, 1992				
Age group i	*Mid-year population*	*Deaths during year*	*Death rate*	*Proportion in age category*	*Age group i*	*Mid-year population*	*Deaths during year*	*Death rate*	*Proportion in age category*
	N_i^{Sw}	D_i^{Sw}	M_i^{Sw}	C_i^{Sw}		N_i^{K}	D_i^{K}	M_i^{K}	C_i^{K}
0	59,727	279	0.00467	0.0136	0	174,078	3,720	0.02137	0.0200
1–4	229,775	42	0.00018	0.0524	1–4	754,758	1,220	0.00162	0.0868
5–9	245,172	31	0.00013	0.0559	5–9	879,129	396	0.00045	0.1011
10–14	240,110	33	0.00014	0.0548	10–14	808,510	298	0.00037	0.0929
15–19	264,957	61	0.00023	0.0604	15–19	720,161	561	0.00078	0.0828
20–4	287,176	87	0.00030	0.0655	20–4	622,988	673	0.00108	0.0716
25–9	311,111	98	0.00032	0.0709	25–9	733,057	752	0.00103	0.0843
30–4	280,991	140	0.00050	0.0641	30–4	732,312	965	0.00132	0.0842
35–9	286,899	197	0.00069	0.0654	35–9	612,825	1,113	0.00182	0.0704
40–4	308,238	362	0.00117	0.0703	40–4	487,996	1,405	0.00288	0.0561
45–9	320,172	643	0.00201	0.0730	45–9	284,799	1,226	0.00430	0.0327
50–4	242,230	738	0.00305	0.0552	50–4	503,608	2,878	0.00571	0.0579
55–9	210,785	972	0.00461	0.0481	55–9	301,879	3,266	0.01082	0.0347
60–4	216,058	1,640	0.00759	0.0493	60–4	374,317	5,212	0.01392	0.0430
65–9	224,479	2,752	0.01226	0.0512	65–9	256,247	6,866	0.02679	0.0295
70–4	222,578	4,509	0.02026	0.0508	70–4	154,623	6,182	0.03998	0.0178
75–9	184,102	6,745	0.03664	0.0420	75–9	149,917	8,199	0.05469	0.0172
80–4	140,667	9,587	0.06815	0.0321	80–4	88,716	9,013	0.10159	0.0102
85+	110,242	17,340	0.15729	0.0251	85+	58,940	10,627	0.18030	0.0068
All	4,385,469	46,256	0.01055	1.0000	All	8,698,860	64,572	0.00742	1.0000
CDR		10.55 p. 1,000			*CDR*		7.42 p. 1,000		

Data source: United Nations, *Demographic Yearbook* (various years).

of persons aged x to $x+n$ at mid-year and use it as an estimate of person-years lived in the age interval x to $x+n$ during the year. N is the size of the total population and functions as an estimate of total person-years lived. D is the total number of deaths during the year. To simplify the notation, we will not use any indicator of the time period to which the rate pertains.

The crude death rate, using this simplified notation, is:

$$CDR = \frac{D}{N} = \frac{\sum_{x=0}^{\infty} {}_nD_x}{N} = \frac{\sum_{x=0}^{\infty} \frac{{}_nD_x}{{}_nN_x} {}_nN_x}{N}$$
$$= \sum_{x=0}^{\infty} \frac{{}_nD_x}{{}_nN_x} \cdot \frac{{}_nN_x}{N} = \sum_{x=0}^{\infty} {}_nM_x \cdot {}_nC_x \qquad (2.1)$$

where ${}_nC_x = {}_nN_x/N$ = the proportion of total population that belongs to the age interval x to $x+n$.

This equation says that the crude death rate is determined by two functions: the set of age-specific death rates (${}_nM_x$) and the proportionate age distribution of the population (${}_nC_x$). In particular, the crude death rate is a weighted average of age-specific death rates, where the weights are supplied by a population's proportionate age distribution (strictly speaking, the proportionate distribution of person-years lived). The sum of these weights, of course, must be unity:[1]

$$\sum_{x=0}^{\infty} {}_nC_x = \sum_{x=0}^{\infty} \frac{{}_nN_x}{N} = \frac{N}{N} = 1.000$$

Now it is easy to see how Kazakhstan could have a lower crude death rate than Sweden even though Sweden had a lower death rate at each age: Sweden's age distribution gives greater weight to the older ages, where age-specific death rates are higher, than did Kazakhstan's.

An equation equivalent to (2.1) can be written with regard to any categorization of the population into subgroups. For example, we could express the crude death rate in terms of height-specific death rates and the proportion of the population that falls into various height classes. There are four reasons for emphasizing the role of age composition:

1. Death rates show very great variation with age, as demonstrated in table 2.1;
2. Human populations differ very considerably from one another in age composition, also as illustrated in table 2.1;[2]
3. The age distribution of the population is itself a demographic variable, being uniquely determined by a population's history of birth, death, and migration rates by age;
4. Data on age-specific deaths and population size are commonly available.

Just as there is nothing unique to age in the derivation in equation (2.1), neither is there anything in it that restricts its applicability to the crude death rate. The development there shows explicitly how any rate (or proportion) in a population is determined by category-specific rates (or proportions) weighted by the proportions of the population that fall into various categories.

Another common way of writing a sum over different ages in demography is to use i to denote the ith age group. So the age group used for the youngest age group becomes $i = 1$; the next youngest becomes $i = 2$, and so on. The sum can go to the highest interval or simply

to ∞, since beyond the highest interval the values of any age series are zero. So equation (2.1) can also be written as:

$$CDR = \sum_{i=1}^{\infty} M_i \cdot C_i$$

The main advantage of using the i notation instead of notation with x and n subscripts is that the i notation can accommodate age groups of irregular size. It is common for deaths to be tabulated in age (last birthday) intervals of 0, 1–4, 5–9, 10–14... A series of death rates in such intervals cannot be represented using the summation sign with ${}_nM_x$ because n is of variable length (1, 4, and 5 years in the first three age intervals). To show explicitly how the age-specific death rates and population proportions shown for Sweden in table 2.1 combine to produce its crude death rate, we used the i notation.

2.2 Age-standardization

The example of Sweden and Kazakhstan showed that differing age structures in the two populations were having a major influence on the comparison of crude death rates. In comparing the levels of mortality in two populations, it is often desirable to eliminate or at least minimize the influence of age composition. One way of making such a comparison would be to assume that Kazakhstan, for example, had the same proportionate age composition as Sweden. The formula for the crude death rate that would result under these circumstances is straightforward:

$$CDR^* = \sum_i M_i^{\mathrm{K}} \cdot C_i^{\mathrm{Sw}}$$

CDR^* is the estimated death rate in Kazakhstan if it retained its own age-specific death rates but had the age distribution of Sweden. In making this estimate we have assumed that adopting the age distribution of Sweden would have no influence on the age-specific death rate schedule in Kazakhstan. CDR^* is a special case of what is commonly termed an age-standardized rate. An age-standardized crude death rate for population j, which we will denote as $ASCDR^j$, has the following structure:

$$ASCDR^j = \sum_{i=1}^{\infty} M_i^j \cdot C_i^s$$

where C_i^s is the proportion of the population that falls in the ith age interval in some population chosen as a "standard." Of course,

$$\sum_{i=1}^{\infty} C_i^s = 1.00$$

What we have done by choosing some population's age distribution as a standard is simply to weight the age-specific rate schedule in population j not by its own age distribution (such a weighing would just reproduce its observed crude death rate), but by that of another, "standard," population.

Standardization is normally used to control or "standardize" the effects of "extraneous" influences when comparing conditions among populations. In the case of age standardization, the extraneous influence that is "standardized" among the populations involved in the

comparisons is their age composition. The procedure is applicable to any rate or proportion. For example, the age-standardized proportion literate ($ASPL$) in population j would be:

$$ASPL^j = \sum_{i=1}^{\infty} L_i^j \cdot C_i^s$$

where L_i^j is the proportion literate in the ith age interval in population j. This index indicates what population j's proportion literate for all ages combined would be if it had the standard age distribution.

As noted above, there is nothing about standardization that restricts its applicability to age. We might, for example, want to standardize the effects of differences in birth-order distributions between two populations whose infant death rates we are comparing. Infant death rates usually vary with birth order and for some purposes it is desirable to control for differences in birth order distributions in making infant death rate comparisons. The birth-order standardized infant death rate in population j is:

$$BOSIDR^j = \sum_{i=1}^{\infty} {}_1M_{0,i}^j \cdot C_i^s$$

where ${}_1M_{0,i}^j$ = death rate between exact ages 0 and 1 in population j for births of order i,
and C_i^s = proportion of all births in a "standard" population which are of order i.

Note that the i index now refers to the birth order rather than to age. As before:

$$\sum_{i=1}^{\infty} C_i^s = 1.00$$

Most price indexes, such as the Consumer Price Index computed by the US Bureau of Labor Statistics, have the form of a standardized rate. They are weighted averages of prices of different goods, with the weights supplied by a "standard market basket of goods."

In performing a standardization, the question arises of what population structure to adopt as a standard. To illustrate that this selection can be consequential, let us examine Mexican and English crude death rates standardized using two different standards (table 2.2). One is a young population age distribution, the other old.[3]

When a young standard is used, both countries' crude death rates decline; when an old standard is used (with relatively high fraction in the older ages), both rates rise. But the curious result is that when a young standard is used, England has a lower age-standardized crude death rate than Mexico's; but when an older standard is used, Mexico has the lower rate. Obviously,

Table 2.2: *Comparison of crude death rates and age-standardized crude death rates using a "young" and an "old" age distribution as standard*

Female population	*Crude death rate (per* 1000 *persons)*	*Age-standardized crude death rate (per* 1000 *persons) by standard age distribution used*	
		"Young" distribution	*"Old" distribution*
Mexico, 1964	9.30	9.20	11.50
England and Wales, 1931	11.61	8.76	13.13

Data source: Preston, Keyfitz, and Schoen, 1972: 254 and 458.

the choice of standard here affects not only the *amount* of difference between standardized rates but even the *direction* of that difference. Such a result could occur only if England had higher death rates at older ages and Mexico had higher death rates at younger ages; the age-specific death rates functions must cross at least once on the age axis. In this case, Mexico's death rates are higher than England's at young ages and lower at old ages.

In view of the possible sensitivity of results to the choice of a standard, it is regrettable that there are no simple rules for selecting one. In fact, the selection inevitably has a large element of arbitrariness. Arbitrariness is scientifically unhealthy, since it allows the researcher to manipulate results to his or her own taste. So let us set down a pair of rules:

a) When comparing only two populations, A and B, use the average of the two population compositions as the standard:

$$C_i^s = \frac{C_i^A + C_i^B}{2}$$

Since both C_i^A and C_i^B sum to unity, so must C_i^s. This procedure for selecting a standard has some important interpretive advantages, as we will see in the next section. Box 2.1 illustrates the application of standardization for a two-population comparison by applying the procedure to the data for Sweden and Kazakhstan shown in table 2.1. Once age-standardized, the CDR for Kazakhstan becomes higher than the CDR for Sweden, reflecting the higher mortality conditions in Kazakhstan.

b) When comparing many populations, use a standard that is close to the mean or median of population structures in the populations under investigation. The only instance where this rule should be ignored is when some peculiarity in structures makes the average or median somehow unrepresentative of human experience. For example, population compositions may be quite distorted by a recent war and a more "normal" structure might be sought for a standard.

It should be clear that the technique of standardization is useful when three conditions are met:

1. One is comparing an aggregate-level variable (usually a rate or proportion) among two or more populations, or in the same population over time;
2. The variable takes on different values from subgroup to subgroup within each population (e.g., from age group to age group);
3. One wishes to minimize the effect on the comparison of differences in the composition of the population according to these subgroups.

Standardization requires data by subgroup both on the composition of population and on the number of events of interest, e.g. on both population and deaths by age group. It is clear from equation (2.1) that an operation closely related to age standardization can be performed. Instead of asking what population A's crude death rate would be if it had population B's age distribution, we could ask what it would be if it had population B's age-specific death rates.

This type of question is frequently asked if data are lacking on age-specific death rates in population A itself. The answer provides a means of indirectly comparing the (unknown) rate schedule in A to the (known) schedule in B. A ratio of the actual number of deaths in A to the expected number based on B's rate schedule is sometimes called a Comparative Mortality Ratio (CMR):

$$CMR = \frac{\sum_i N_i^A \cdot M_i^A}{\sum_i N_i^A \cdot M_i^B} = \frac{D^A}{\sum_i N_i^A \cdot M_i^B}$$

Box 2.1 *Example of Age-standardization*

Formulas:

$$ASCDR^{Sw} = \sum_{i=1}^{\infty} M_i^{Sw} \cdot C_i^s = \text{Age-standardized crude death rate for Sweden}$$

$$ASCDR^{K} = \sum_{i=1}^{\infty} M_i^{K} \cdot C_i^s = \text{Age-standardized crude death rate for Kazakhstan}$$

$$C_i^s = \left(\frac{C_i^{Sw} + C_i^{K}}{2}\right) = \text{Average age distribution}$$

Example: Sweden and Kazakhstan, females, 1992

Age group i	*Age distribution of Sweden*	*Age distribution of Kazakhstan*	*Average age distribution*	*Age-specific death rate in Sweden*		*Age-specific death rate in Kazakhstan*	
	C_i^{Sw}	C_i^{K}	$\frac{C_i^{Sw}+C_i^{K}}{2}$	M_i^{Sw}	$M_i^{Sw} \cdot \frac{C_i^{Sw}+C_i^{K}}{2}$	M_i^{K}	$M_i^{K} \cdot \frac{C_i^{Sw}+C_i^{K}}{2}$
0	0.0136	0.0200	0.0168	0.00467	0.00008	0.02137	0.00036
1–4	0.0524	0.0868	0.0696	0.00018	0.00001	0.00162	0.00011
5–9	0.0559	0.1011	0.0785	0.00013	0.00001	0.00045	0.00004
10–14	0.0548	0.0929	0.0738	0.00014	0.00001	0.00037	0.00003
15–19	0.0604	0.0828	0.0716	0.00023	0.00002	0.00078	0.00006
20–4	0.0655	0.0716	0.0686	0.00030	0.00002	0.00108	0.00007
25–9	0.0709	0.0843	0.0776	0.00032	0.00002	0.00103	0.00008
30–4	0.0641	0.0842	0.0741	0.00050	0.00004	0.00132	0.00010
35–9	0.0654	0.0704	0.0679	0.00069	0.00005	0.00182	0.00012
40–4	0.0703	0.0561	0.0632	0.00117	0.00007	0.00288	0.00018
45–9	0.0730	0.0327	0.0529	0.00201	0.00011	0.00430	0.00023
50–4	0.0552	0.0579	0.0566	0.00305	0.00017	0.00571	0.00032
55–9	0.0481	0.0347	0.0414	0.00461	0.00019	0.01082	0.00045
60–4	0.0493	0.0430	0.0461	0.00759	0.00035	0.01392	0.00064
65–9	0.0512	0.0295	0.0403	0.01226	0.00049	0.02679	0.00108
70–4	0.0508	0.0178	0.0343	0.02026	0.00069	0.03998	0.00137
75–9	0.0420	0.0172	0.0296	0.03664	0.00108	0.05469	0.00162
80–4	0.0321	0.0102	0.0211	0.06815	0.00144	0.10159	0.00215
85+	0.0251	0.0068	0.0160	0.15729	0.00251	0.18030	0.00288
Sum	1.0000	1.0000	1.0000		0.00737		0.01188

$ASCDR^{Sw} = 7.37$ p. 1,000 $ASCDR^{K} = 11.88$ p. 1,000

Data source: United Nations, *Demographic Yearbook* (various years).

where

D^A = recorded deaths at all ages combined in A,
N_i^A = number of persons in the ith age interval in A,
M_i^B = death rate in the ith age interval in B.

This index was used for many years by the Registrar-General of Great Britain to compare the death rates of different occupational groups. If the ratio is greater than one, the implication is

that the (unknown) age-specific death rates are in general higher in A than in B, though strictly speaking this need be true only in one age interval. This procedure is part of a demographic method called "indirect standardization" that is now rarely used in its complete form (see Shryock and Siegel, 1973: 421–2). The truncated portion of the procedure just described finds extensive application in historical studies of fertility (see section 5.1).

2.3 Decomposition of Differences between Rates or Proportions

A closely-related question is, "How much of the difference between death rates in A and B is attributable to differences in their age distributions?" This latter question is addressed through a technique known as decomposition (Kitagawa, 1955).

We should note at the outset that there is no unique answer to the question addressed by decomposition. There are many ways to decompose a difference and the choice among them is, to an important extent, arbitrary. However, one technique has an advantage of economy and expositional cleanness, and that is what we shall develop here. Let us suppose that we are interested in decomposing the difference between crude death rates in populations A and B. Define the original difference as Δ.

$$\Delta = CDR^B - CDR^A = \sum_i C_i^B \cdot M_i^B - \sum_i C_i^A \cdot M_i^A$$

Now we will divide each of these terms into two equal parts and add and subtract certain additional terms, thereby keeping the difference (Δ) constant:

$$\Delta = \frac{\sum_i C_i^B \cdot M_i^B}{2} + \frac{\sum_i C_i^B \cdot M_i^B}{2} - \frac{\sum_i C_i^A \cdot M_i^A}{2} - \frac{\sum_i C_i^A \cdot M_i^A}{2} + \frac{\sum_i C_i^B \cdot M_i^A}{2} - \frac{\sum_i C_i^B \cdot M_i^A}{2} + \frac{\sum_i C_i^A \cdot M_i^B}{2} - \frac{\sum_i C_i^A \cdot M_i^B}{2}$$

We now combine the eight terms in Δ into four and then into two:

$$\Delta = \sum_i C_i^B \cdot \left[\frac{M_i^B + M_i^A}{2}\right] - \sum_i C_i^A \cdot \left[\frac{M_i^B + M_i^A}{2}\right] + \sum_i M_i^B \cdot \left[\frac{C_i^A + C_i^B}{2}\right] - \sum_i M_i^A \cdot \left[\frac{C_i^A + C_i^B}{2}\right]$$

$$= \sum_i \left(C_i^B - C_i^A\right) \cdot \left[\frac{M_i^B + M_i^A}{2}\right] + \sum_i \left(M_i^B - M_i^A\right) \cdot \left[\frac{C_i^A + C_i^B}{2}\right]$$

$$= \begin{matrix}\text{difference in age} \\ \text{composition}\end{matrix} \cdot \left[\begin{matrix}\text{weighted by average} \\ \text{age-specific mortality}\end{matrix}\right] + \begin{matrix}\text{difference in rate} \\ \text{schedules}\end{matrix} \cdot \left[\begin{matrix}\text{weighted by} \\ \text{average age} \\ \text{composition}\end{matrix}\right]$$

$$= \begin{matrix}\text{contribution of age compositional} \\ \text{differences to } \Delta\end{matrix} + \begin{matrix}\text{contribution of rate schedule} \\ \text{differences to } \Delta\end{matrix}$$

We have decomposed the difference into two terms, one of which is clearly interpretable as the contribution of age distributional differences and the other as the contribution of rate schedule differences. Between them, they completely account for the original difference. Note that the "contribution of rate schedule differences" term is precisely the difference between age-standardized death rates in B and A, when the "standard" population age composition applied to both populations is the *average* age composition in A and B.

Interpreting this version of decomposition is straightforward. Any other decompositional procedure introduces a residual, or interaction, term whose meaning is not always clear-cut. For example, by including a different set of terms in the expansion of Δ and then rearranging and simplifying again, we can develop an alternative formula:

$$\Delta = \sum_i C_i^B \cdot \left(M_i^B - M_i^A\right) + \sum_i M_i^B \cdot \left(C_i^B - C_i^A\right) - \sum_i \left(M_i^B - M_i^A\right) \cdot \left(C_i^B - C_i^A\right)$$

The right-most summation term in this expression looks something like a covariance term; it is positive if M_i^A tends to be high relative to M_i^B at ages where C_i^A is high relative to C_i^B. Such a pattern would contribute *negatively* to Δ (since the sign of the last term is negative), because Δ was expressed as the crude death rate in B minus the crude death rate in A.

But it is awkward and unnecessary, in general, to deal with residual terms. The earlier procedure obviated the need for them. And it used an approach to decomposition that is perhaps least arbitrary, since it accepted the average of their rate schedules to weight their age-compositional differences and the average of age structures to weight their rate differences. For most applications, it seems preferable.

Box 2.2 demonstrates the application of the recommended procedure to the decomposition of differences between the crude death rates in France, 1991, and Japan, 1992. France's crude death rate is higher than Japan's by 0.003116. Differences in age composition account for 75 percent (.002333/.003116) of the difference between crude death rates and differences in rate schedules account for the remaining 25 percent. In this case, both factors contribute in the same direction to the difference. But in many applications, one of the factors will account for more than 100 percent of the original difference. This happens when the two factors work in opposite directions, and there is no reason to expect that they will normally work in concert.

Both standardization and decomposition procedures can be applied simultaneously to more than one variable (see Das Gupta, 1993, for a thorough development of multivariate standardization and decomposition). The same standard can also be applied to many populations to produce standardized rates. However, the decompositional procedure described above must be limited to the two populations being directly compared. Comparisons among more than two populations require more complex procedures (Das Gupta, 1993; Smith et al., 1996).

Note that, as in the case of standardization, there is nothing to require that age be one of the variables involved in decomposition. For example, one could decompose a difference between two nations' infant death rates into differences due to birth-order distributions and differences due to rate-schedule differences (that is, differences in their death rates for children of the same birth order). When age is one of the variables in a standardization or decomposition of demographic rates, it is strongly recommended that age categories be no wider than 5 years when data permit. Age variation in vital rates is sufficiently great that the age composition *within* a 10-year age interval can have a substantial effect on the value of an age-specific rate pertaining to that interval.

Box 2.2 *Decomposition of Differences between Rates*

$\Delta = CDR^{F} - CDR^{J}$ = difference between crude death rates in France and Japan

$$= \sum_i \left(C_i^F - C_i^J\right) \cdot \left[\frac{M_i^F + M_i^J}{2}\right] + \sum_i \left(M_i^F - M_i^J\right) \cdot \left[\frac{C_i^F + C_i^J}{2}\right]$$

Example: France, 1991 and Japan, 1992, females

Age group i	C_i^F	C_i^J	M_i^F	M_i^J	$(C_i^F - C_i^J) \cdot \left[\frac{M_i^F + M_i^J}{2}\right]$	$(M_i^F - M_i^J) \cdot \left[\frac{C_i^F + C_i^J}{2}\right]$
0	0.0133	0.0089	0.0061	0.0040	0.000022	0.000023
1–4	0.0467	0.0349	0.0004	0.0004	0.000005	0.000000
5–9	0.0508	0.0734	0.0002	0.0001	−0.000003	0.000006
10–14	0.0541	0.0720	0.0002	0.0001	−0.000003	0.000006
15–19	0.0746	0.0811	0.0003	0.0002	−0.000002	0.000008
20–4	0.0686	0.0674	0.0005	0.0003	0.000000	0.000014
25–9	0.0730	0.0703	0.0006	0.0003	0.000001	0.000021
30–4	0.0749	0.0618	0.0007	0.0004	0.000007	0.000021
35–9	0.0794	0.0581	0.0009	0.0007	0.000017	0.000014
40–4	0.0768	0.0789	0.0014	0.0011	−0.000003	0.000023
45–9	0.0533	0.0677	0.0022	0.0016	−0.000027	0.000036
50–4	0.0507	0.0649	0.0029	0.0024	−0.000038	0.000029
55–9	0.0551	0.0602	0.0042	0.0037	−0.000020	0.000029
60–4	0.0544	0.0554	0.0064	0.0056	−0.000006	0.000044
65–9	0.0528	0.0470	0.0096	0.0090	0.000054	0.000030
70–4	0.0317	0.0365	0.0184	0.0158	−0.000083	0.000089
75–9	0.0360	0.0286	0.0279	0.0303	0.000216	−0.000078
80–4	0.0298	0.0197	0.0589	0.0587	0.000596	0.000005
85+	0.0240	0.0132	0.1605	0.1356	0.001599	0.000462
Sum	1.0000	1.0000			0.002333	0.000783

$$CDR^{F} = \sum_i C_i^F \cdot M_i^F = 0.008996$$

$$CDR^{J} = \sum_i C_i^J \cdot M_i^J = 0.005880$$

Original difference $= CDR^{F} - CDR^{J} = 0.008996 - 0.005880 = 0.003116$
Contribution of age compositional differences $= 0.002333$
Contribution of age-specific rate differences $= 0.000783$
Total contribution $= 0.002333 + 0.000783 = 0.003116$
Proportion of difference attributable to differences in age composition $= 0.002333/0.003116 = 0.749$
Proportion of difference attributable to differences in rate schedules $= 0.000783/0.003116 = 0.251$

Data source: United Nations, *Demographic Yearbook* (various years).

2.4 The Lexis Diagram

We can define cohort age-specific rates by restricting occurrences and exposures to the relevant ages, exactly as we did for period age-specific rates. Thus, the age-specific death rate between ages 25 and 30 for a cohort born in 1940 (denoted with a 1940 superscript) is:

$$ {}_5M_{25}^{1940\,c} = \frac{\text{Number of deaths to the 1940 cohort between ages 25 and 30}}{\text{Number of person-years lived by the 1940 cohort between ages 25 and 30}}$$

Note that counting deaths and person-years for this cohort requires including experience that stretches from calendar year 1965 (when they all reach age 25) through calendar year 1970 (when they all reach age 30), or over a span of 6 calendar years (1965, 1966, 1967, 1968, 1969, and 1970).

As noted already, cohort rates and period rates have the same structure but take into account different exposure segments. The Lexis diagram (Lexis, 1875) is a useful device to clarify relations between exposure segments for cohorts and exposure segments for periods. It is simply a two-dimensional figure in which age (in this case) is one dimension and calendar time the other. Units of age and time are normally the same (e.g., years), and these units are displayed in equal increments along both axes. What goes onto the diagram varies from one application to the next. Sometimes it is counts of events; sometimes it is symbols that represent counts; and sometimes it is life-lines.

On a Lexis diagram with the same time unit on both the time and age axis, a cohort advances through life along a 45° line. The exposure of interest in a cohort rate is thus delineated by two 45° lines that demarcate the time interval that defines membership in the cohort. Figure 2.1 delineates the age-time exposure region pertaining to the cohort born between 1.000 and 3.999.

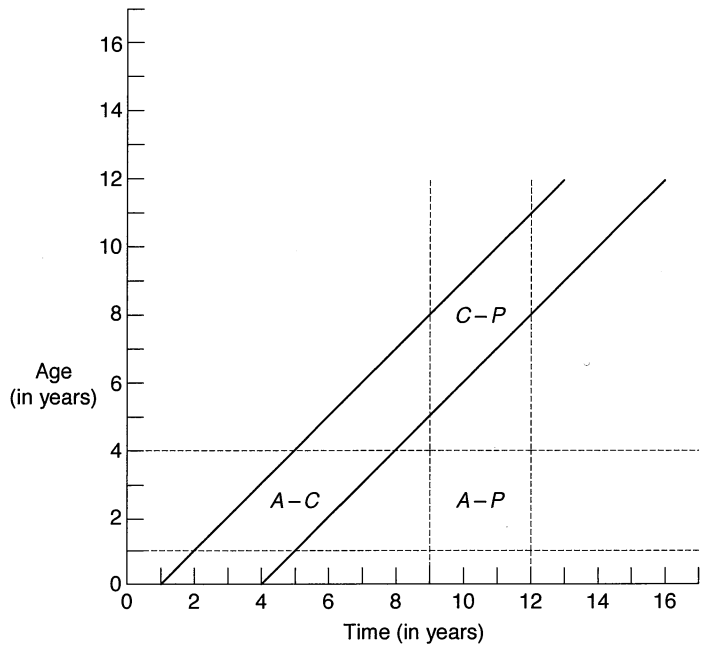


Figure 2.1 Lexis diagram representations of exposure for age (A), period (P), and cohort (C)

Period rates would be constructed from regions of exposure delineated by two vertical lines, shown on the figure for the period between 9.000 and 11.999.

An age-specific cohort rate thus restricts measuring exposure and counting occurrences to a parallelogram formed by two 45° lines defining the cohort and the two horizontal lines defining the age range (A–C on figure 2.1). An age-specific period rate restricts the measurement of exposure and occurrences to a rectangle formed by the two vertical lines defining the period and the two horizontal lines defining the age-range (A–P on figure 2.1). One can also define a cohort-specific period rate, which restricts exposure and occurrences of interest to a parallelogram delineated by two vertical lines defining the period and two 45° lines defining the cohort (C–P in figure 2.1). This latter construction is rarely encountered.

2.5 Age-specific Probabilities

Just as in the case of rates, the computation of probabilities can also be restricted to a certain age range. The conventional notation for a probability of dying between age x and $x+n$ (with both x and n measured in exact age) is ${}_nq_x$. The probability that a birth in the 1940 cohort would die before reaching age one is thus:

$$ {}_1q_0^{1940\,c} = \frac{\text{Number of deaths to 1940 birth cohort between ages 0 and 1}}{\text{Number of births in the 1940 birth cohort}} $$

In the above example, the events (or "trials") that were counted in the denominator were the number of births in the 1940 cohort. If we had been dealing instead with the probability that a member of the 1940 birth cohort who reached age 25 died before he or she reached age 30, we would have:

$$ {}_5q_{25}^{1940\,c} = \frac{\text{Number of deaths to 1940 birth cohort between ages 25 and 30}}{\text{Number of persons in the 1940 birth cohort who reached their 25th birthday}} $$

Recall that the calculation of a probability requires having a number of events in the denominator. The number of events in the denominator of ${}_5q_{25}^{1940\,c}$ is the number of 25th birthdays achieved by the 1940 cohort.

The infant deaths occurring to the birth cohort of 1940 will stretch over two calendar years, 1940 and 1941 (since the cohort will reach its first birthday, on average, about halfway through 1941). Likewise, the infant deaths occurring in 1941 will pertain to two annual birth cohorts, those born in 1940 and those born in 1941. The counting rules for calculating probabilities are also usefully displayed on a Lexis diagram.

Figure 2.2.a is a Lexis diagram containing life-lines of 6 persons during a two-year segment of age and time that begins with birth in calendar year 1995. The 1995 cohort's life-lines clearly must all fall within a parallelogram formed by two 45° lines that originate on January 1, 1995 and on January 1, 1996. A line ends when a person dies. Two persons out of the original cohort of 6 persons die before reaching age 1, so the probability of infant death for the cohort born in 1995 is

$$ {}_1q_0^{1995\,c} = \frac{2}{6} = .3333 $$

We cannot calculate, on the basis of the information presented, the probability that a person who reaches age 1 in 1996 dies before reaching age 2 because some of the members of the

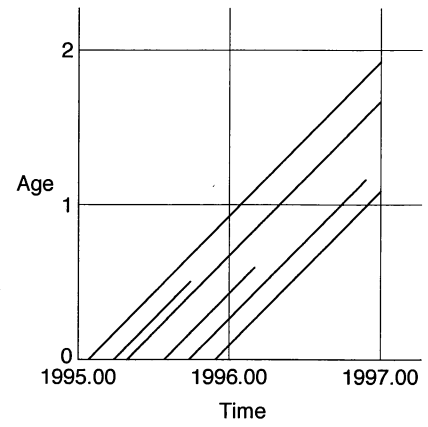


Figure 2.2a Lexis diagram containing life-lines for a birth cohort of 1995
Note: Time = 1995.00 refers to January 1, 1995.

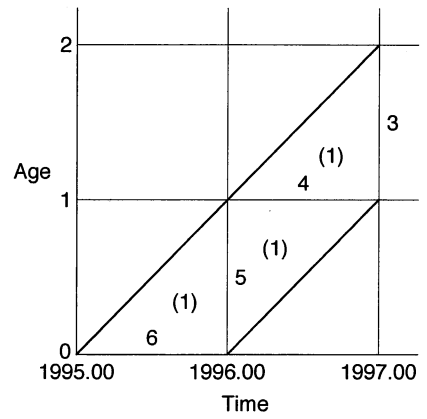
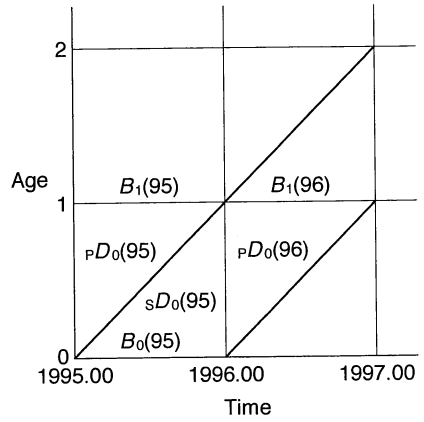


Figure 2.2b Lexis diagram containing counts of events pertaining to a birth cohort of 1995
Notes: From a birth cohort of 6 births in 1995: (1) death in 1995 and 5 survivors to the beginning of calendar year 1996; (1) death at age 0 in 1996 and 4 survivors at age 1 (4 first birthdays in the cohort, all occurring in 1996); (1) death to the cohort at age 1 during 1996 and 3 survivors to the beginning of calendar year 1997.

cohort may have died, before reaching age 2, in calendar year 1997. That year is not shown on the Lexis diagram.

The mortality experience represented by those 6 life-lines is summarized in a series of counts presented on figure 2.2.b, where the interpretation of the various numbers is also presented. The counts are placed within the same parallelogram that contains the cohort's experience. If a census were taken on January 1, 1996, it should have counted 5 persons aged 0 (last birthday); the only persons aged 0 at that date would have to have been born during calendar year 1995. A census taken at any time other than the beginning of a year would mix persons from two different birth cohorts at age 0. Counts adjacent to horizontal lines show the numbers arriving at a particular age in the cohort (6 births and 4 first birthdays).



$B_0(95)$ = Number of births in 1995.
$B_1(95)$ = Number of first birthdays in 1995.
$B_1(96)$ = Number of first birthdays in 1996.
${}_SD_0(95)$ = Number of deaths at age 0 in 1995 to people who reached age 0 in the same year.
${}_PD_0(95)$ = Number of deaths at age 0 in 1995 to people who reached age 0 in the previous year.
${}_PD_0(96)$ = Number of deaths at age 0 in 1996 to people who reached age 0 in the previous year.

Figure 2.2c Lexis diagram containing main symbols used to represent counts of events

Part c of figure 2.2 presents symbols that can be used to represent the counts presented in part b. For example,

$${}_SD_0(95) = \text{number of deaths at age 0 in 1995 to persons who reached age 0 (i.e., who were born) in the } \textit{same} \text{ year that they died}$$
$$= 1$$

So $D_x(Y)$ is the total number of deaths at age x (last birthday) in year Y. The S and P subscripts on the left divide those deaths between those occurring to the birth cohorts who reached age x during the *same* year in which they died (S) and those occurring to the cohort that reached age x during the *previous* year (P). The "separation factor" at age x, year Y, separates the deaths at age x last birthday into two birth cohorts to which they occur. The separation factor at age x, year Y, is the proportion of deaths at age x (last birthday) in year Y occurring to persons who reached age x during year Y:

$$SF_x(Y) = \frac{{}_SD_x(Y)}{D_x(Y)} = \frac{{}_SD_x(Y)}{{}_SD_x(Y) + {}_PD_x(Y)}$$

In terms of these symbols, the probability of death before reaching age one for the birth cohort of 1995 is:

$${}_1q_0^{1995\,c} = \frac{{}_SD_0(95) + {}_PD_0(96)}{B_0(95)}$$

Box 2.3 *Calculating Rates and Probabilities*

$${}_nM_x^{\text{cohort c}} = \frac{\text{Number of deaths to the cohort c between ages } x \text{ and } x+n}{\text{Number of person-years lived by the cohort c between ages } x \text{ and } x+n}$$

$${}_nM_x^{\text{year t}} = \frac{\text{Number of deaths in the age range } x \text{ to } x+n \text{ during year t}}{\text{Number of person-years lived in the age range } x \text{ to } x+n \text{ during year t}}$$

$${}_nq_x^{\text{cohort c}} = \frac{\text{Number of deaths to the cohort c between ages } x \text{ and } x+n}{\text{Number of persons in the cohort c who reached their } x^{th} \text{ birthday}}$$

The probability of dying before reaching age 2 for the birth cohort of 1995 is:

$${}_2q_0^{1995\,c} = \frac{{}_SD_0(95) + {}_PD_0(96) + {}_SD_1(96) + {}_PD_1(97)}{B_0(95)}$$

The probability of dying before reaching age 2 for a member of the 1994 birth cohort who survived to age 1 is:

$${}_1q_1^{1994\,c} = \frac{{}_SD_1(95) + {}_PD_1(96)}{B_0(94) - {}_SD_0(94) - {}_PD_0(95)}$$

This type of measure cannot be computed without data that separate deaths occurring during a certain calendar year at a particular age into the two birth cohorts that contribute those deaths.

Box 2.3 summarizes the main cohort and period mortality indexes developed in this and previous sections.

2.6 Probabilities of Death Based on Mortality Experience of a Single Calendar Year

For many purposes it is desirable to have measures of mortality that pertain to a particular time period rather than to a particular cohort. But we have seen that two annual birth cohorts contribute to the deaths recorded during any year at any particular age. How are these two cohorts' experiences to be synthesized in producing an estimate of age-specific mortality for that calendar year? Such a synthesis for infants is facilitated by writing the probability of death before age 1 as:

$${}_1q_0 = \begin{array}{l}\text{probability that a}\\ \text{child dies in his}\\ \text{calendar year of birth}\end{array} + \begin{array}{l}\text{probability that a}\\ \text{child survives his}\\ \text{calendar year of birth}\end{array} \cdot \begin{array}{l}\text{probability that if a child}\\ \text{survives his year of birth, he}\\ \text{dies in the next calendar year}\\ \text{before reaching age 1}\end{array}$$

Let us insert the appropriate elements in this formula and show that it produces a correct formula for a cohort born in year Y:

$${}_1q_0^{Yc} = \frac{{}_SD_0(Y)}{B_0(Y)} + \frac{B(Y) - {}_SD_0(Y)}{B_0(Y)} \cdot \frac{{}_PD_0(Y+1)}{B_0(Y) - {}_SD_0(Y)} = \frac{{}_SD_0(Y) + {}_PD_0(Y+1)}{B_0(Y)}$$

The idea underlying the synthesis is to take all the death terms in the numerator from the same calendar year, rather than from two different years (as required for calculating a cohort's probability). Thus we can write the probability of dying between ages 0 and 1 for calendar year Y as:

$$ {}_1q_0(Y) = \frac{{}_SD_0(Y)}{B_0(Y)} + \frac{B(Y) - {}_SD_0(Y)}{B_0(Y)} \cdot \frac{{}_PD_0(Y)}{B_0(Y-1) - {}_SD_0(Y-1)} \qquad \textbf{(2.2)} $$

A more general formula appropriate for any age interval x to $x+1$ would replace age 0 with age x in equation (2.2) and would use $B_x(Y)$ as the number of xth birthdays achieved in year Y.

A closely related concept is the "infant mortality rate," one of the most common indexes used in demography. The conventionally defined infant mortality rate for year Y is defined as infant deaths in year Y divided by births in year Y:

$$ IMR(Y) = \frac{D_0(Y)}{B_0(Y)} = \frac{{}_SD_0(Y) + {}_PD_0(Y)}{B_0(Y)} \qquad \textbf{(2.3)} $$

Unfortunately, this infant mortality "rate" is structured as a probability rather than as a conventional demographic rate, since it has a count of events in the denominator. But in fact it not only fails as a rate but as a probability: it is counting trials in one urn (births in year Y) but events from parts of two (deaths to births that occurred in years Y and $Y-1$). It will equal the probability of dying before age one for the *cohort* born in year Y only if ${}_PD_0(Y) = {}_PD_0(Y+1)$. Such equivalence would occur if births were constant from year to year and if age-specific mortality conditions were also constant. Under these restricted circumstances, the infant mortality rate will also equal the period probability of dying before age 1 given by (2.2).

But the infant mortality rate is simple to define and materials for its calculation do not require the division of infant deaths by calendar year of birth. Its value should not be seriously misleading as an estimate of the probability of dying before age 1 (assuming that data are accurate) unless the number of births varies greatly from year to year. As a simple expedient, it probably deserves tolerance more than condemnation. Table 2.3 presents estimates of infant mortality rates in major regions of the world in recent years, and box 2.4 defines other conventional measures of fetal and early-life mortality.

An alternative procedure for converting data on mortality in a particular period into estimated probabilities of dying is used more frequently than the method described in this section. It is developed in the next chapter.

Table 2.3: *Infant mortality rates in major areas, 1995–2000 (deaths per 1,000 live births)*

Major area	*IMR* 1995–2000
Africa	87
Asia	57
Europe	12
Latin America and the Caribbean	36
Northern America	7
Oceania	24

Source: United Nations, 1999.

Box 2.4 *Conventional Measures of Fetal and Early-infancy Mortality*

Fetal mortality rate:

$$ \frac{\text{Fetal Deaths during year } t}{\text{Fetal Deaths} + \text{Births during year } t} $$

Perinatal mortality rate:

$$ \frac{(\text{Fetal Deaths} \geq 28 \text{ weeks of pregnancy}) + (\text{Deaths} < 1 \text{week of age}) \text{ during year } t}{\text{Births} + (\text{Fetal Deaths} \geq 28 \text{ weeks}) \text{ during year } t} $$

Neonatal mortality rate:

$$ \frac{\text{Deaths} < 1 \text{ month of age during year } t}{\text{Births during year } t} $$

Post-neonatal mortality rate:

$$ \frac{\text{Deaths 1–11 months of age during year } t}{\text{Births during year } t} $$

Infant mortality rate:

$$ \frac{\text{Deaths} < 1 \text{ year of age during year } t}{\text{Births during year } t} $$

NOTES

1. In equation (2.1), the *CDR* can be seen as the sum of a mortality level indicator and of a covariance between two distributions: the population by age and the age-specific death rates. Indeed, (2.1) can be rewritten as:

$$ CDR = \sum_{x=0}^{\omega} [({}_nM_x - \overline{M}) \cdot ({}_nC_x - \overline{C})] + \overline{M} $$

where $\overline{M}$ is the (unweighted) mean of age-specific death rates and $\overline{C}$ is the (unweighted) mean of proportions of the population in an age interval. For a given set of age-specific death rates, the *CDR* is thus higher the higher is the covariance of the population by age with these age-specific rates.

2. Age composition only matters as long as the variable of interest varies with age. If age-specific mortality rates were constant by age, the *CDR* would not depend on the age structure of the population. If ${}_nM_x = M$ at all ages, then equation (2.1) becomes: $CDR = \sum M \cdot {}_nC_x = M \cdot \sum {}_nC_x = M$ (since $\sum {}_nC_x = 1$). The assumption of constant age-specific rates is an unrealistic assumption for mortality.

3. In particular, the "young" structure is that of a model "West" female stable population with $r = .02$ and life expectancy at birth of 45 years; the old structure has an $r = .01$ and life expectancy at birth of 65 years (Coale and Demeny, 1983: 46 and 64). The concept of a stable population is developed in chapter 7 below and model life tables are described in chapter 9.