Bridging the micro-macro gap in population forecasting

Contract no. SP23-CT-2005-006637

Deliverable D9

Report on Input Data Requirements of MIC

Work Package 2
Micro-simulation Model

May 2007

Authors:

Frans Willekens
Joop de Beer
Nicole van der Gaag

Netherlands Interdisciplinary Demographic Institute
P.O. Box 11650
2502 AR the Hague
The Netherlands

Contents

MicMac
Report on Input Data Requirements of MIC

1. Introduction

The aim of MicMac is to develop a new methodology that complements demographic projections with projections of the way people live their lives. MicMac offers a bridge between aggregate projections of cohorts (Mac) and projections of the life courses of individual cohort members (Mic). Both Mic and Mac are multistate models with transition rates as parameters. Mac focuses on transitions among functional states by age and sex; Mic addresses demographic events and other life transitions at the individual level. The new methodology MicMac can be used to monitor and forecast demographic change as well as lifestyles and life courses of the population.

In deliverable D2: 'MicMac, Report on Data Input Requirements of Mac' the state space of MicMac has been defined. The state space is the set of possible states a person can occupy. The states are categories of a variable, *e.g.* marital status. Also a combination of variables may be used, resulting in a composite variable. For example, if marital status consists of four states (never-married, married, divorced or widowed) and living arrangement of five (living in the parental home, living alone, living with a partner, living with other persons, living in an institution) the combination of both variables results in twenty states, although the number can be reduced as some categories may be excluded (e.g. married while living at the parental home). At any age, an individual occupies one of the multiple states. States are mutually exclusive and exhaustive. A change in value of the state variable implies a transition from one state to another. The state space does not only define the states, but also the possible transitions between states. In general, the life course is viewed as a sequence of state occupancies and events that result in state transitions.

MicMac is a single multistate model, not a combination of two separate models. The basic component that integrates the micro and the macro level is the transition rate between functional states by age, sex and other personal characteristics. The characteristics of the model are determined by the state space. In Deliverable D2 the following 12 variables have been defined for MicMac:

1. age
2. year of birth
3. sex
4. level of educational attainment
5. age at leaving school
6. marital status
7. living arrangement
8. number of children ever born
9. number of children in household
10. disability status
11. smoking
12. body mass index (BMI)

In addition to these variables, also the entries into the population due to birth and immigration, and the exits from the population due to death and emigration are included in the model.

Broadly speaking, the data requirements of Mic are not much different from the data requirements of Mac. In Mac, the fundamental parameters are the transition rates (occurrence-exposure rates) by age and one or a few personal characteristics. The transition rates are translated into transition probabilities that are applied to a base population to determine future populations. Rates are estimated from data. Two types of data are distinguished:

Aggregate data (also referred to as grouped data): counts of events and durations of exposure for groups of people, e.g. people of a given age, sex and marital status. Occurrence-exposure rates are ratios of number of events and durations of exposure. This approach is illustrated in traditional demographic methods. Sources of aggregate data are tabulations based on censuses and vital statistics.

Individual observations or micro data (e.g. survey data) on occurrence and timing of events. If observations on individuals are available, transition rates are generally predicted (estimated) by a transition rate model. The transition rate model is a regression model with personal characteristics as predictor variables. The regression model can be a simple univariate model or a complex multivariate model that includes several interaction effects, unobserved heterogeneities, variables measured at different levels of analysis, etc.

In principle, the same variables can be used in the analysis of grouped data as in the analysis of individual observations. In case of grouped data, the size of cross-classification of variables may become very large when all variables are included in a single cross-classification. As a consequence, event counts and/or exposures in an individual cell of the cross-classification may be very small, too small to reliably estimate a transition rate. For instance, if individuals with a high level of educational attainment tend to have different probabilities of entering into a relationship compared to individuals with a low level of educational attainment, then the transition rates between different states of educational attainment and between different states of household positions will be interrelated. Consequently, if we want Mac to be able to handle all possible interactions, the base population for the projections will contain an enormous number of cells. A cross-classification of all categories of the twelve variables of the state space of MicMac results in a very large number of categories and even if only a few variables will be used, stratification will imply that a large number of categories have to be distinguished.

If the number of variables and categories in the cross-classification is prohibitive large, one of two approaches may be followed. First, the number of variables and/or categories may be reduced by combining categories. The second approach is to omit higher-order interactions between variables. What it means is that instead of considering the full cross-classification including the higher-order interactions, only the marginal distribution (marginal totals) are considered. This approach is equivalent to specifying a regression model that excludes higher-order interaction effects. By implication not all interdependencies of events are considered. The first approach is the common approach in macro-level analysis of tabulated data, e.g. in demographic projections. The second

approach is common in regression analysis of individual observations. Few regression models include all possible interaction effects because higher-order interaction effects are likely not to be statistically significant, i.e. significantly different from zero. The second approach allows a considerably larger number of variables and categories to be considered at the price of limited interdependencies. That approach is common in microsimulation. In most practical applications, a main advantage of micro-simulation over macro simulation is the level of detail. In microsimulation all possible cross-classifications can be included, introducing heterogeneity (at the price indicated). Whereas macro models (based on cross-classified data) for instance project the number of people in certain living arrangements at different ages, micro projections (based on individual observations) provide information on the number of people that experience a certain succession of household positions during their life cycle and the duration in each of these states.

If the micro-model Mic includes more detailed information than the macro-model Mac the outcomes of both simulations will diverge: life trajectories or lifepaths generated by Mic will differ from the state occupancies at consecutive ages produced by Mac. In order to reach consistency between Mic and Mac some constraints have to be imposed either on Mic or on Mac, using the outcomes of the other part of the model. As Mic contains more information than Mac, the outcomes of Mic may be used to reconsider the assumptions of Mac, while at the same time the more robust results of Mac may put some constraints to the aggregated results of Mic. Most microsimulation models apply alignment techniques to bridge the gap between macro-level and micro-level analysis. For a technical description, see e.g. Bækgaard (2002).

Mic and Mac differ in the use of the transition rates. In Mac transition rates and the associated transition probabilities are applied to the base population to calculate future populations. In Mic, on the other hand, the transition rates are used to determine whether a given individual member k (at age x and time t) of a sample population will actually experience a specific transition and what the waiting time is to the transition. The transition rate is viewed as an expected value (given the predictors) and Monte Carlo methods (random number generators) are used to determine whether k experiences an event and when the event will occur. Whereas Mac uses expected values of transition rates to determine aggregate measures (e.g. state occupancies at a future date, number of transitions during a period of a given length, expected sojourn times in different states), Mic uses distributions around expected values to determine the lifepaths of all members of a (sample) population. As long as the macro and micro projections of MicMac are based on exactly the same set of transition rates or probabilities, the outcomes at the macro and micro level simulations are mutually consistent. Mic produces individual histories whereas Mac produces aggregations of individual histories. Since Mic is individual-based, it may take into account many more characteristics of the individual than Mac. For example, for a woman aged 45 the transition rate from household position ‘married with one child’ to household position ‘married without children’ Mic may account for the age of the child and the number of years the woman has lived in a household with children. If two consecutive transitions are considered, e.g. marriage and birth of the first child, Mac may give the expected interval between the two events whereas Mic gives the distribution of the intervals between the events. For instance, Mic may be used to determine the proportion of women in the sample population having the first child in the year of marriage.

An important difference between Mic and Mac is how to link lives. Persons are related to each other, e.g. partners or parents and children. In Mac, restrictions may be imposed at macro level, for instance the number of women living with a partner needs to be consistent with the number of men living with a partner (leaving same-sex couples aside). The problem is generally known as the two-sex problem. In Mic, linking lives means linking individuals: the way the rates change depends on the rates or events of the partner, the parents and/or the children. For instance, in case of partnerships, linked lives are generated using matching rules that apply or are assumed to apply in marriage markets. In principle, in MicMac linked lives will be dealt with using pointers, however, this will only be captured where we know how rates of linked persons will change and this will not be implemented in the first release of the program.

The current deliverable (D9: 'Report on input requirements of MIC') pays attention to additional data aspects compared to those described in D2 'Report on input requirements of MAC'. Chapter 2 focuses on the life course paradigm that is the rationale behind the perspective on events and experiences as being embedded in the entire life course, also referred to as the biographical method. The quantitative study of biographies leads to several data and measurement problems. These are discussed in Chapter 3. Chapter 4 deals with single transitions and single episodes and focuses on the estimation of the expected waiting time to the event from incomplete observations.

2. The life course paradigm

The life course paradigm is inextricably bound up with projections of individual cohort members. Life-course theory tries to discover the mechanisms by which events and conditions at one stage of life or during an extended period of life influence life at a later age. In recent years, the life course paradigm emerged significantly in behavioural and social sciences and epidemiology (Giele and Elder, 1998 and Ben-Shlomo and Kuh, 2002). In health sciences for instance, there is considerable interest in understanding the effects of early-life experiences and lifestyle on diseases at older ages and the factors and interventions that prevent, postpone or slow down these diseases. In sociology on the other hand, the lifetime consequences of early life transitions constitute a major subject of study. For instance, the transition from initial education to work may affect the employment career for the rest of ones life (e.g. Scherer, 2001).

2.1 The life course as a composite process

The interpretation and understanding of human life is facilitated when life is viewed as a set of interdependent evolving processes in an organic setting. In this view, processes may be seen as lines connecting events (Salmon, 1998). The origin of the process is an event, generally known as event-origin. The event-origin triggers the process. For instance, an infection is an event that triggers a process, the infectious disease. The process may have several outcomes (e.g. recovery, death, transition to a chronic status). The outcome is an event, which, in the epidemiological literature, is known as end-point. An end-point is the outcome of interest. End-points can be of multiple types. Death, the end-point of a mortality process, has various causes and each cause of death represents a type of death. Marriage dissolution may be either divorce or death of the spouse. An end-point may be viewed as an intersection that ends a road and continues in a new direction. The direction taken at the intersection is not known in advance although the possible directions are generally given. Since the direction taken cannot be predicted in advance, it is uncertain. The uncertainty is bounded, however. We know what directions are possible and that some directions are taken more often than others. The different directions of change are not equally likely. We may even know the likelihood that a given direction is taken, i.e. we know the probabilities with which directions are taken. These probabilities must add to one. Event-origin and end-point are fundamental characteristics of processes. Many processes are characterized by the event that triggers the process (e.g. infection) or the event that represents the outcome (e.g. mortality process, divorce as a process).

2.2 The life course as an outcome of interacting processes

Salmon's view of human life is particularly useful in life course research. A useful cognitive structure of the life course is that of a multitude of processes organized by domain of life and arranged in sequence and parallel. In that perspective, the life course evolves because processes trigger other processes and interact with parallel processes occurring simultaneously. The basic building block is an *elementary process*, which is a process that generates at most one event. The event is the outcome of the process and the process is the mechanism that may lead to the event.

Three types of interactions between parallel processes may be distinguished: status dependence, event dependence and resource dependence.

Process B is *status-dependent* on process A if the likelihood of an event in B is dependent on the status of process A. For instance, if married women are more likely to leave their job than non-married women, then the marital status (status in the marriage process or marital career) influences the rate of leaving a job (employment career). If smokers are at higher risk of cardiovascular disease than non-smokers, then smoking status influences the biological process causing cardiovascular disease. If women with small children are less likely to work than women without small children, then the presence of small children is the cause of the (temporarily) lower labour force participation.

Process B is *event-dependent* on process A if the occurrence of an event in process A changes the likelihood of an event in B. It may enable, facilitate, enhance, prevent or inhibit the event in B. For instance if at the time of a divorce a person is more likely to change jobs than at other times, then the job change is causally related to the divorce. The divorce has triggered the job change.

Two remarks are warranted. They relate to the conditions of temporal ordering and contiguity, two conventional criteria for causation. First, 'at the time of a divorce' introduces the condition of simultaneous events. Simultaneity is a fuzzy concept. Courgeau and Lelievre (1992, p. 96) consider two events as simultaneous when they occur in a chosen time interval. They introduce the notion of 'fuzzy time' to capture the time lag between a decision to act and the actual event. When a divorce triggers a job change, the actual time at which a job is changed is conditioned by many factors that are not related to the divorce, such as the availability of a vacancy and the duration of the search. As a consequence, event dependence does not necessarily imply that the events are simultaneous. The condition that cause and effect are contiguous may also not be satisfied because of an 'incubation period' during which the effect is latent rather than manifest. The contiguity requirement is untenable in life course analysis.

The second remark is related to the temporal ordering of the events. Conventionally, A must be prior to B in order to be a cause of B. But human beings anticipate consequences of actions and plan accordingly. For instance, one may change jobs in anticipation of a divorce, even before a separation. As a result, the job change is prior to the divorce although the divorce is the cause of the job change. The conventional condition of temporal ordering would indicate that the job change causes the divorce. Simon (1979, p. 73) argues that not future events but expectations and anticipations are the cause of actions. The causal priority is established in the mind in a way that is not reflected in the temporal sequence of behaviour (Marini and Singer, 1988, p. 377). It is the presence of a causal link in the mental schema that counts. If the causal link exists in the mind of the person, the two events are causally related. Marini and Singer argue that it is not sufficient to consider the temporal sequence of behavioural intentions because of perceived constraints and incompatibilities. Causal priority can be identified only by asking people about the causal mechanism (Marini and Singer, 1988, p. 378). That requires in-depth interviewing.

2.3 Developmental and chronological time

The interpretation of human life as a set of interdependent evolving processes is connected with a *timetable*. The timetable is a coordinating mechanism that assures harmonious development or maturation of the organism along the axis of chronological time. The timetable may require events to occur during relatively narrow periods of time, in particular events that trigger subprocesses. The understanding of the proper timing of events and the developmental impact of events that are off-time, i.e. too early or too late, is an important subject of life course research. For instance, children born prematurely have an increased risk of respiratory problems because lungs develop in the final stage of gestation. Young adults who are unable to get employment during extended time after graduation from school are likely to experience problems entering the job market. Teenage pregnancies endanger the development of mother and child because maturation is not sufficiently advanced. Short birth intervals result in depletion of the mother. The concept of developmental readiness captures the true meaning of the timetable in the context of development. Each event in life has a proper position in developmental time. Because developmental processes proceed at different pace, the location in chronological time may not coincide with that in developmental time.

Development proceeds in stages and early stages of development are often crucial stages. Many chronic diseases can be treated successfully at early stages of development and deviant behaviour can be corrected more effectively with early intervention. The timing of an intervention, therefore, is of crucial importance. In addition to the search for the timetable of development, the detection of critical periods of development is a subject of life course research. Ben-Shlomo and Kuh, working in the field of epidemiology, define a critical period as a limited time window in which an exposure to an exogenous factor can have adverse or protective effects on development and subsequent disease outcome. The exogenous factor leads to a substantially increased or reduced risk. The 'critical period model' is one of the important aspects of the life course perspective. Events or experiences during critical periods may cause effects that are irreversible. They produce an 'imprint' that may be expressed only many years later. Periods of rapid change (periods of transition, such as transition to adulthood) are critical periods too. The effect of events and experiences during these periods are more significant than if they occur during other periods of human development.

2.4 Cumulative causation

A characteristic of any developmental process is cumulative causation. At the end of the 19th century, Veblen (1898) published a paper entitled 'Why is economics not an evolutionary science?' In it, he proposed cumulative causation as the key concept in the study of unfolding processes in terms of causes and effects. In his view, causal processes engender effects that provide the starting point for subsequent causal processes, which in turn generate effects that provide the material for subsequent causal processes, and so on (Vromen, 1995, p. 1). Veblen's view on economic change is consistent with that of a development process composed of elementary processes organized in sequence and parallel. The effects of processes may depend on their duration and the number of events they generate. When an event or experience occurs repeatedly, or an exposure is prolonged, the effects accumulate. Examples include the effects of repeated spells of unemployment, breakdown of partnerships, losses of

significant others, abuse, etc., or effects during prolonged periods of malnourishment, poor diet, growth retardation, unemployment, poverty, stress, smoking, etc. The effect of number of occurrences of an event on life at higher age is known as *occurrence dependence*. Many causes do not arise suddenly but they cumulate over time and many consequences of an event or an experience unfolds over multiple periods. The effect of radiation and the effect of job experience are cases in point. The effect of duration (of exposure) is called *duration dependence*.

A distinction can be made between duration of exposure and degree of exposure. They are two aspects of exposure that influence the impact. The impact of exposure is often expressed in terms of an event. Examples include cancer following prolonged periods of radiation, lung cancer or cardiovascular disease following periods of smoking, marital disruption following prolonged periods of domestic violence or addiction to alcohol or drugs. Most people can stand short episodes of detrimental influences but not long durations.

Part of the cumulative causation may be attributed to the fact that, when developmental processes are interrupted or modified, subsequent processes are affected. This is related to the fact that one event may trigger another event or may prevent another event from occurring. As a result, the remaining life course may be affected for a long time. When the processes that are interrupted or modified are critical in the life course, the effect is severe and lasts a lifetime.

In a developmental process, not only events have a timetable. Effects have a timetable too. Many reactions extend over multiple periods. Some effects may be immediate while others are delayed. Effects may also change over time. A sequence of experiences or interventions may have a cumulative effect. For instance, major life events such as the death of a spouse or a divorce have effects that decay gradually with time. Other events generate effects that follow another pattern. For instance, the immediate effect of myocardial infarction (MI) is a substantially increased risk of death. After one day, the excess mortality risk is much lower although it is higher than without the MI. The time paths of effects differ by event or experience. The time path may also vary by effects. Consider the effect of smoking and in particular of a quitting. Smoking increases the risk of cardiovascular disease (CVD) and lung cancer. After quitting, the increased risk may remain (in case of lung cancer) or decline (in case or cardiovascular disease).

2.5 From life course theory to life course models

In order to operationalize the theory and to confront the theory with empirical observations, we need a modelling framework and a measurement scheme or observational plan to register the events, experiences and conditions during the course of life.

Mathematical representations of life paths and the processes they entail constitute powerful interpretative schemes, modes of explanation and predictive devices. A complicating factor is that cause-effect relationships are not deterministic but probabilistic. When an event triggers a process, the outcome is uncertain although the possible outcomes may be known. Models are needed to determine the likelihood of each possible outcome and the expected outcome. Life course models, therefore, are

part of probability theory. The critical link between theory and models is the rate and probability of an event, conditional on past experience and contemporary factors. Models that describe events and transitions in life are called transition rate models.

Point of departure for the life course framework is that the entire life course can be described in terms of personal attributes[1] and that behaviour can be described in terms of changes in personal attributes. Attributes may be biological, physical, psychological and social. Examples are sex, marital status, employment status, health status, income category, and region of residence. At any age or point in time, an individual can be characterized by a set of attributes. With time, attributes may change. The variables that denote attributes may be referred to as *attribute variables*, or simply attributes. An attribute variable has usually a limited number of categories. It is therefore a discrete variable. For instance, marital status has usually four categories: never married, married, divorced and widowed. Additional categories may be distinguished, e.g. cohabitation and separation. The health status may have two categories only: healthy and not healthy.

In multistate modelling, the attribute variable is referred to as *state variables*. Employment status, marital status, health status and region of residence are state variables. An individual who is married at a given age is said to occupy the state 'married'. In multistate analysis, the attribute of an individual at a given age is denoted by the state occupied. The set of possible states the person can occupy is the *state space*. The attribute variables distinguished in MicMac are listed in chapter 1; a detailed description of the state space of MicMac is presented in deliverable D2: 'MicMac, Report on Data Input Requirements of Mac'.

A change in attribute is an event. Events in the life course are also called life events to distinguish them from other events, e.g. historical events. In multistage modelling, the event is usually referred to as transition. Migration and marriage are transitions. Infection, i.e. onset of infectious disease, and recovery are transitions too. They are events that change the health status. The terms transition and event are used interchangeably.

Attributes and events are two basic concepts that are required to bridge theories and models. The life course is adequately described when for each relevant personal attribute the value is known at each age. The sequence of attributes constitutes the life history or biography. Multistate models use this status-based approach. It is often not practical, however, to report the attributes for every age. An alternative is to report the initial attributes, the life events and the age at which the events occur. This is called the event-based approach. Event history models approach the life course as a sequence of events. The life course of an individual is fully documented when at all ages the relevant attributes are known.

The overall objective of life course models is to predict attributes and transitions. In MicMac the basic component that integrates the micro and macro level is the transition rate between functional states by age, sex and other attributes. The transition rates change as a result of changes in the occurrence and timing of transitions people make during the course of life, i.e. as a result of events in individual biographies. While Mac works at the aggregate level, Mic works at the individual level. To simulate individual

---

[1] Synonyms of attributes are properties, characteristics, and traits.

biographies microsimulation models are designed (see e.g. Zaidi and Rake, 2001). Examples include SOCSIM, one of the oldest but continuously updated microsimulation packages developed at the University of California at Berkeley, and LifePaths developed by Statistics Canada to analyse the effects of government policies on health expenditures and public pension sustainability. A detailed description of the model used in MicMac is given in deliverable D8: 'Description of the Microsimulation model'.

2.6 Incomplete observations and synthetic biographies

The description of the entire biography requires that an individual is under observation from birth to death and the relevant attributes are recorded at each age or the initial attributes are recorded as well as the events that occur between birth and death. A comprehensive description of the entire life course of individuals is not practical as it requires a follow-up from birth to death. Information on the life course is almost by definition incomplete. For instance, attributes may be recorded in discrete time rather than continuous time. More often, people are observed during a restricted period of time only. As a consequence, a segment of the life course is observed rather than the entire life course. The period during which an individual is observed is referred to as the *observation window*. The observation window plays an important role in the generic model since the parameters of the generic model must be estimated from the incomplete data. Different types of incomplete data are distinguished. At the start of the observation window, most subjects are already alive and occupy particular states of existence and stages of life. Usually, the ages at which they entered the states remain unknown. The observation is therefore left censored or left truncated. At the end of the observation, many subjects did not experience the events under study and continue to be at risk. The observation is right censored or truncated. Some subjects may not experience any event during the observation window. At the end of the observation, they are in the same state as at the start of the observation. In this case the observation is left and right censored. Subjects may also enter the observation after the onset of the observation window and/or leave the observation during the observation window for reasons unrelated to the event under study. The first case is called 'late entry' and the second case is 'attrition'. In all cases, the observation window affects the calculation of the transition rates (see further chapter 3).

Incomplete data are omnipresent in life course research. One response is to restrict the analysis to the segments of life that are observed for every individual in the study. It severely restricts the understanding of life trajectories and the prediction of outcomes events later in life. An alternative is to combine information from different individuals. When the study includes individuals of different age, the observation window covers different segments of live. The partial information on several individuals may be combined to construct the life course or biography of a hypothetical individual. It is an individual who experiences events at a rate that is determined on the basis of the partial information on individuals of different age. The biography of the hypothetical individual is a *synthetic biography*. It is a biography that is based on partial information on different individuals. The synthetic biography is a powerful instrument in life course research and longitudinal data analysis since it represents the course of life that is consistent with the data, i.e. that encompasses the empirical evidence. It summarizes the evidence that is collected from different individuals in a single biography or a set of representative biographies. The instrument to construct synthetic biographies is the

multistate life table (Willekens, 1999). The life table constructs the life history from birth to death for a virtual individual or a synthetic cohort on the basis of *age-specific transition rates* that are estimated from the available data. The rates are ratios of numbers of events during a given age interval and total durations of exposure during the same interval. Because of the incomplete data, all age-specific rates are not based on the same individuals followed from birth to death but on different individuals at different stages of life. The rates are based, however, on observations covering a limited period of time. These observations are generally more recent than observations on the entire life span. In periods of important historical changes, predictions based on synthetic biographies may be more accurate than predictions based on entire life histories. The concept of synthetic biography is similar to the concept of synthetic longitudinal data (Joshi, 2001, p. 10). As Joshi argues, if one wants to predict over a whole lifetime (e.g. predict long-term consequence of some antecedent) but the longitudinal data are incomplete, it is better to generate synthetic longitudinal data than waiting until a whole cohort dies (Joshi, 2001, p. 10).

The age-specific transition rates play a pivotal role in the construction of synthetic biographies. They are obtained by a careful measurement of (1) the events that occur during a specified period of observation and (2) the duration of exposure during the same period. Left and right censoring is accounted for. The transition rates are the basis for the transition probabilities and the state probabilities that characterize the biography. A single synthetic biography that uses information on all individuals under observation (e.g. in the sample) is usually too crude a representation of reality. It is often necessary or desirable to stratify the population and to construct a separate biography for each stratum. The most important stratification variable is the birth cohort. The population may also be stratified by individual attributes such as level of education, place of residence or religion. The rationale of stratification is that individuals who live in different historical eras (cohorts) or who have different background are likely to have different life paths.

The multistate life table is a generic model for synthetic biographies. The advantage of a generic model that is sufficiently abstract is that it may be applied in different domains of life. Events may be demographic (e.g. marriage, divorce, childbirth, migration, death), economic (e.g. entry into the labour force, retirement, entry and exit from poverty spells, purchase of durable goods such as a house or a car), social (e.g. entry into or exit from significant social organizations or social status category). Events may be related to health (e.g. infection, onset of chronic disease, recovery from an illness, onset of impairment or handicap), lifestyle (e.g. starting of smoking, change in food habits, transition from risk-taking behaviour to risk-aversive behaviour), education (e.g. graduation, school drop-out) or other domains of life.

The next chapter deals with data and measurement issues related to the construction of individual biographies.

3. Data and measurement issues

To study the life course of individuals, or in other words developmental processes, requires information and repeated measurements. Repeated measurements are observations of the same characteristic that are made several times (Lindsey, 1999, p. 3). They are often referred to as longitudinal data. Longitudinal data can be collected prospectively, following subjects forward in time, or retrospectively by extracting multiple measurements on each subject from historical records and autobiographic memory (Diggle *et al*., 1995, p. 1).

One of the key characteristics of longitudinal data is chronological information. In a prospective study, the sample units (individuals in the sample) are chosen on the basis of a set of attributes and then followed up in time to see what response is obtained. In a retrospective study, a sample is chosen according to the current characteristics and the values of previous explanatory and response variables are investigated. Prospective studies include panel studies, clinical trials, and cohort studies. Retrospective studies include cross-sectional life history surveys and case-control studies.

Longitudinal data are collected to examine behavioural changes and to determine the part of the change that may be attributed to experiences and interventions in earlier stages of life. An overview of longitudinal data sets is not feasible neither desirable. Examples of longitudinal (micro-level) data sources that are of relevance for MicMac are the longitudinal Labour Force Survey (LFS), the European Community Household Panel (ECHP), and the Fertility and Family Survey (FFS).

For estimating transition rates, numbers of events and exposure times have to be measured. The different ways of collecting information on events result in different data types. Several types of data are distinguished in Section 3.1. The time dimension of the data is discussed in Section 3.2. As discussed in chapter 2, the observation of the life course is generally limited to a segment of life. The measurement issues introduced by this limitation are topic of Section 3.3. Section 3.4, finally, contains a general discussion of the age-time framework that is useful in studies of changes in the life course, in particular the inter-cohort comparison of biographies. The age-time framework is visualized by the Lexis diagram.

3.1 Data types

Data on life histories may be of different types. A first distinction is between micro data and grouped data. Micro-data provide information on individuals, households, or other unites of observation (subjects). Grouped data provide information on groups of persons. The grouping along the duration variable (time or age) is of particular significance. Time is continuous and hence it is logical to assume the duration variable to be a continuous variable. In many situations, this assumption is not realistic for two reasons (Vermunt, 1997, p. 87). Firstly, the events of interest can sometimes only occur at particular points in time. Voting in elections is an example. Since elections take place at particular points in time, changes in voting behaviour can be measured in discrete time only. Secondly, in many cases, time or age is not measured accurately enough to be treated as continuous. The dates of demographic events are often measured in month or year. The date is measured in terms of the time period or time interval in which the

event occurs. These data are known as *interval-censored data* or *grouped duration data*. Interval censoring indicates that the exact date t or age x of the event is not known, but only the time interval in which the event took place: $t_1 < t < t_2$ or $x_1 < x < x_2$. This implies the absence of exact information on the time to the event and hence the exposure time. In this case the measurement of the exact time can be replaced by an assumption about when exactly during the interval the event occurs. Assumptions on the timing of the event of interest and censoring constitute an essential aspect of life-table analysis. For a discussion, see for instance Namboordiri and Suchindran , 1987, pp. 58ff; Vermunt, 1997, p. 87; and Hougaard, 2000, p. 302.

The grouping of duration data needs particular attention when multiple time scales are used. In demography, two time scales are often combined, calendar time (historical time) and age (individual time). Dates of events may be grouped along the age scale but not along the calendar time scale. Dates may also be grouped along both the age and time scales. Different groupings result in different data types. The Lexis diagram is a convenient graphical representation of different groupings.

A second distinction is between status data and event data. Status data give information on the status of a person at one point in time or different points in time. Event data give information on the events that occur and imply a transition from one status to another (see also chapter 2). For example, the survey question on current marital status results in status data. The question on current place of residence and place of residence of the parents at time of birth results in transition data, which belong to the category of status data. The question whether in the household a child died in the past 12 months leads to data on events, although it may also be approached as status data. The number of children ever born is of the event data type. It is the number of events (childbirths) since the birth of the respondent. The following data types may be distinguished (with illustrative survey questions):

Micro-data: data on subjects
Event data
- Information on occurrence or non-occurrence of an event in the observation period
  "Did you have a child in the past 12 months?"
- Time to event (measured in particular time scale)
  "In what year and month were you born?"
  "What was your age at marriage?"
  "How old were you when your first marriage dissolved?"
- Number of events experienced by a subject during a period of observation
  "Number of changes of residence in the past 3 years?"

Status data
- Current status: status at one point in time (at survey)
  "Current place of residence?"
  "Which contraceptive method are you currently using?"
- Status at a previous point in time
  "What was your place of residence 5 years ago?"
  "What contraceptive method were you using in April 1991?"
  "For most of the time until you were 12 years old, did you live in a city or in a village?"
  "What was your previous place of residence?"

- Transition data: status at two points in time. Transition data are obtained by comparing the status at two points in time. The two points are generally the survey (current status) and a point in time prior to the survey. The prior time point may be fixed or variable.
- Origin-dependent transition data: status at several points in time
  "For ever-married women: how many children did you have in the past 2 years?"
  "For foreign-born population: what is you current place of residence and the place of residence five years prior to the survey?"

Grouped data: data on groups of subjects (count data)
Events
- Number of events during a given period (e.g. period of observation)

Transitions
- Number of subjects by current status and status prior to the survey.

In this report, we focus on *micro-data on events*.

3.2 Measurement of time

The time to event is measured in continuous time or, more often, in discrete time. In most surveys, e.g. demographic surveys and health surveys, the exact dates of events are not recorded. What is recorded is the week, month or year in which an event occurs. Consequently, we know that an event occurred but we do not know the exact date. The observation is referred to as *interval censored* observation. In interval censoring, the exact time to event is not observed, but the interval in which an event occurs is recorded. All empirical observations are interval-censored since the timing of the events is recorded to the nearest hour, day, month or year (Lindsey, 1999, p. 361). When the interval is sufficiently small, the indicators that are derived are not much affected by the length of the interval. In demographic analysis, it is common to use the year or the month as the unit of time. Sometimes the dates are given in integer-truncated years.

To enhance the measurement of intervals between events, dates are recoded. Two cases are considered. The first assumes that the month of the event is recorded. The second assumes that the day of event is recorded.

3.2.1 Dates in month and year

In case the dates of events are given in months and years, the dates are recoded in *Century Month Code* (CMC). In that coding scheme, dates of events are measured in months elapsed since the beginning of the 20th century. January 1900 is month 1, February 1900 is month 2, January 1901 is months 13, March 1946 is month 555 and April 1991 is month 1096. The CMC is determined from the year and month of occurrence of the event, using the expression:

DATECMC = (YEAR-1900)*12+MONTH

where DATECMC is the CMC of occurrence. The major reason for using the CMC is that the intervals between events can easily be determined. When the CMC is given, the month and year of the event may be determined using the following expression (see e.g. Blossfeld and Rohwer, 1995, p. 39):

YEAR = INT(DATECMC – 1)/12 + 1900
MONTH = DATECMC – [(YEAR – 1900)*12]

where INT denotes the integer value of a real number.
For instance, when the CMC is 555, YEAR is 1946 and MONTH is March.

The CMC may be negative. A negative value indicates a month before 1900.

The age of the respondent at the time of an event is derived from the CMC at event and the CMC at birth of the respondent. The age in months is simply the difference between the CMC at the event and the CMC of birth. The exact age in years is calculated as
AGE = [CMCE – CMCB] / 12

where CMCE is the CMC at the event of interest and CMCB is the CMC of birth.

The age in completed years is

IAGE = TRUNC[[CMCE – CMCB] / 12]

3.2.2 Dates in day, month and year

The time at event is recoded in days, months or years since a reference date. We consider year as the time unit. Suppose an event occurs on May 4, 1988 and the reference date is January 1, 1900.

The exact number of years since 1st January 1900 is obtained by the following transformation (Mamun, 2001, p. 98):

EY = YEAR + (MONTH-1)/12 + (DAY-1)/(30.437*12)

May 4, 1988 is 88.3415 years since the beginning of the 20th century:
88 + (5-1)/12 + (4-1)/(30.437*12) = 88.34154702

The date in exact years may be converted back in year, month and day of occurrence, using the following formula:

YEAR = TRUNC(EY)
MONTH = TRUNC[(EY – YEAR)*12] + 1
DAY = TRUNC[(EY – YEAR – (MONTH-1)/12)*30.437*12] + 1

For instance, 88.3415470 is
YEAR = TRUNC[88.3415470] = 88
MONTH = TRUNC[(88.3415470-88)*12] + 1 = 5 (MAY)
DAY = ROUND[(88.34154702-88 – (5-1)/12)*30.437*12] + 1 = 4

The conversion is not always perfect because it does not account for the different numbers of days in a month and the changing number of days in the month of February. *Table 3.1* illustrates the method.

The dates in exact years may be converted in dates in CMCs:

DATECMC = (EY – 1900) * 12

Note that numeric DATECMC is not an integer value but a real value. Consider May 4, 1988. The date in CMC is 88.34154702 * 12 = 1060.098564. The month is CMC 1060 and the day is ROUND[0.098564 * 30.437] + 1 = 4.

Sometimes, the reference date is different. For instance, in the Framingham Heart Study, which is a longitudinal study that started in 1948-50 and is widely used in epidemiology, the dates of the exams are measured in number of days since 1st January 1960.

Table 3.1. Conversion of dates in exact years and vice versa

Exact				Re-estimated			
Year	Month	Day	Years	Year	Month	Day	Difference
88	5	4	88.34155	88	5	4	0
80	1	1	80.00000	80	1	1	0
80	12	31	80.99880	80	12	31	0
16	3	2	16.16940	16	3	1	1
13	9	19	13.71595	13	9	19	0
27	12	23	27.97690	27	12	23	0
47	7	28	47.57392	47	7	27	1
13	11	8	13.85250	13	11	8	0
17	8	29	17.65999	17	8	29	0
23	3	27	23.23785	23	3	27	0
21	6	11	21.44405	21	6	11	0
26	2	9	26.10524	26	2	8	1
20	8	22	20.64083	20	8	21	1
15	7	9	15.52190	15	7	9	0
14	2	22	14.14083	14	2	21	1
14	9	16	14.70774	14	9	16	0
14	12	4	14.92488	14	12	3	1
19	5	16	19.37440	19	5	15	1
48	6	29	48.49333	48	6	28	1
11	12	27	11.98785	11	12	26	1
9	10	4	9.75821	9	10	4	0
13	10	19	13.79928	13	10	19	0
25	9	29	25.74333	25	9	29	0
26	6	5	26.42762	26	6	5	0
10	12	2	10.9194046	10	12	1	1

An event that occurs on 30th January 1960 occurs at 29 days (1+29). The date can be negative. For instance, an event that occurred on 3rd December 1959 occurred at day -29

(December 31 is day -1 and December 3 is day -29). These figures can be converted into exact number of years since the beginning of the 20th century:

EY = 1960 + DATE/365.25

where DATE is the date of the event in days since 1st January 1960 and EY is the date of the event in exact years (Mamun, 2001, p. 97). For instance, if the DATE of an event is -460, the event occurs in 1958 and more specifically at 1958.741.

3.2.3 Multiple events during time interval

When dates are measured in discrete time, the timing of events during the interval becomes an issue in the estimation of exposure time or duration at risk. If at most one event occurs during the interval, it is assumed that events occur at the beginning of the discrete time interval, at the end of the interval or in the middle of the interval. When the interval is small, e.g. a month, the assumption is not very important. But when the interval is large, i.e. a year, the assumption may have a significant effect on the exposure time. If the event occurs at the beginning of the interval, the state occupied at the beginning of the month is occupied for the entire interval.

A further complication arises when more then one event may occur during the interval. Multiple transitions are often not permitted, e.g. in the Life History Calendar (for a discussion, see Khatun and Willekens, 2001, pp. 18ff). When multiple transitions are permitted, their timing must be determined. Yamaguchi (1991, pp. 97ff) considers the case of two events during the interval and assumes that events are uniformly distributed over the interval. Consider a unit interval and suppose that the first event occurs at time x ($0 < x < 1$). The occurrence signifies the entry into the period of being exposed to the risk of the second event. For convenience, the event may be considered as the entry into a state of interest (e.g. immigration or labour force entry). Let f denote the density of the second event, the probability of an event during the unit interval. The density is constant and consequently the events are uniformly distributed during the interval. The probability of a second event between the entry into the risk period and the end of the unit interval is (1-x)f. People who experience the second event during the unit interval enter the risk period on average at

$$\bar{x} = \frac{\int_0^1 [x(1-x)f]dx}{\int_0^1 [(1-x)f]dx} = \frac{\int_0^1 xdx - \int_0^1 x^2 dx}{\int_0^1 dx - \int_0^1 xdx} = \frac{x(\frac{1}{2} - \frac{1}{3}x)}{1 - \frac{1}{2}x}\Bigg|_0^1 = \frac{1}{3}$$

or four months. The probability of not experiencing the second event before the end of the unit interval is 1 – (1-x)f for a person who enters at time x. The average proportion of the year passed before these individuals enter the risk period is

$$\bar{x} = \frac{\int_0^1 x[1-(1-x)f]dx}{\int_0^1 [1-(1-x)f]dx} = \frac{\frac{1}{2}x^2 - \frac{1}{2}x^2 f + \frac{1}{3}x^3 f}{x - xf + \frac{1}{2}x^2 f}\Bigg|_0^1 = \frac{\frac{1}{2}(1-f) + \frac{1}{3}f}{1 - f + \frac{1}{2}f}$$

$$\bar{x} = \frac{1}{2}\left[\frac{1-\frac{1}{3}f}{1-\frac{1}{2}f}\right]$$

In Yamaguchi (1991, p. 99) f = 1/a and[2]

$$\bar{x} = \frac{6a-2}{12a-6} = \frac{6a-3}{12a-6} + \frac{1}{12a-6} = \frac{1}{2} + \frac{1}{12a-6}$$

When entries are uniformly distributed, the value of $\bar{x}$ varies from 1/2 to 1/3 as a function of the density f. It becomes close to 1/2 when f becomes small.

3.3 Observation window

The complete life history is rarely observed. Only periods or segments of life are observed because the duration of observation is necessarily limited in time. As a result, some respondents do not experience the event under study during the period of observation. For instance, in a study of marriage dissolution, the dissolution may occur before the observation starts, during the period of observation, or after the observation is terminated. The period of observation is referred to as the observation window.

3.3.1 Specification of the observation window

The observation window is defined by the start of observation and the end of observation. Observation may start at a fixed point in time, at a given age, or upon the occurrence of an event. It may end at a fixed point in time, at a given age, at an event, or after the occurrence of a given number of events. *Table 3.2* shows the type of onset and end of observation. To delineate the observation window, we define two fictitious events. The first fictitious event is the onset of observation and the second is the termination. The state occupied before the onset of observation is 'Not yet in observation' and is coded as 0. The state occupied at the end of observation, i.e. after the termination of observation, is 'Censored' and is coded as S+1, where S is the number of states in the state space. The length of the observation period is CMC at the end of observation minus the CMC at onset of observation.

Table 3.2. Characterization of observation window

Onset of observation	End of observation	Type of censoring
Fixed time t	Fixed time t	Type I
Fixed age	Fixed age	
Event	Event	
	Number of events	Type II
Entry during observation period	Exit during observation period	

[2] Note the error in Yamaguchi (1991, p. 99).

The period that is observed may be one year, five years or longer. For instance, in some retrospective surveys such as the Demographic and Health Survey, all relevant events since birth are recorded, i.e. the onset of observation is the event of birth, and the observation window is the period between birth and the survey date (fixed calendar time t, variable age). The period is longer for older respondents in the sample than for younger respondents. Older subjects contribute more person-years of observation than younger subjects. One implication is that retrospective surveys of that kind provide information over a much longer period for younger age groups (e.g. adolescence) than for older age groups. Older age groups are observed not only among a relatively small subsample, since few older persons are included in the survey, but the observation is limited in time to a relatively short period before the survey (when these persons were already old). In contrast, the behaviour at younger ages is observed over a much longer period. For older respondents, the period of observation is relatively long ago. For young respondents, it is recent. That is important to remember in interpreting the age-specific occurrence-exposure rates that are calculated from survey data. These rates that are used as input to the life table, are based on a longer period of observation for younger ages than for older ages. To avoid that problem, one may define an observation window of a given length (duration) by omitting the events that occurred prior to a given date, e.g. five years prior to the survey (fixed t at start and fixed t at end). In that case, we omit the information on early stages of the life course of older respondents. An alternative is to stratify the sample by period of birth and generate life tables by birth cohort. These life tables indicate the intergenerational shifts in behaviour during adolescence and young adulthood.

Two diagrams are in use to demonstrate the observation window. The first considers one time scale (the time diagram), the other two time scales (age and calendar time, the Lexis diagram).

Time diagram

The time diagram locates events and processes on a single time scale. It is often used to denote different types of incomplete observation on events and processes (of subjects). *Figure 3.1* and *figure 3.2* are time diagrams. Figure 3.1 shows an observation window and different types of incomplete observations on processes. The types of censoring will be discussed later in this section. Figure 3.2 shows observations collected in a Life History Calendar (LHC). The LHC is a technique to collect life history data retrospectively. It situates retrospective measurements within the prospective panel design. The calendar shows the marital status (MS) and contraceptive practice status (CC). Two marital states are distinguished: not married (0) and married (X). The following contraceptive practice states are considered: 1-Pill use, 3-Injection, 0-Non use, B-Birth, P-Pregnant.

Lexis diagram

The Lexis diagram offers a useful instrument in exposure analysis. It situates events in an age-time framework. The horizontal axis represents the chronological time or calendar time. The vertical axis represents age. *Figure 3.3* shows an individual lifeline. Event A occurs at exact time t and exact age x.

Figure 3.1. Types of censoring and episodes in the calendar records

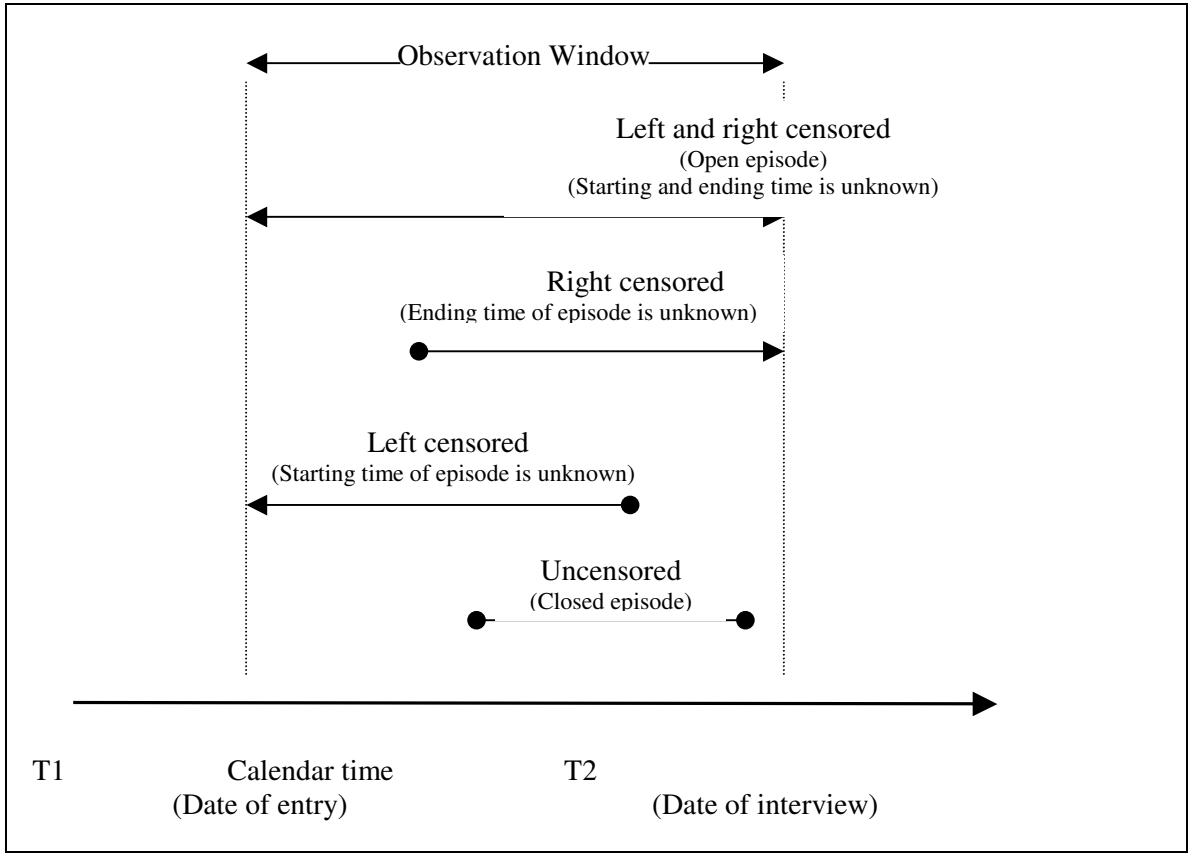


Figure 3.2. The life history calendar

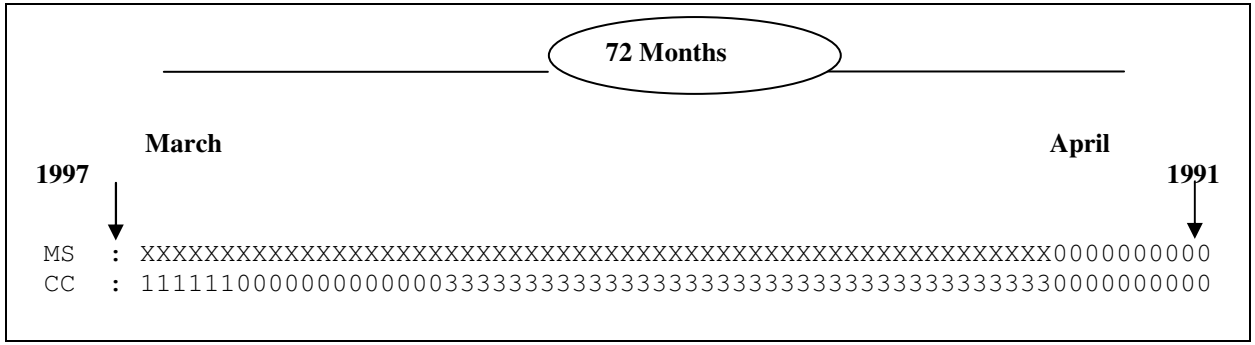


Figure 3.3. Lexis diagram

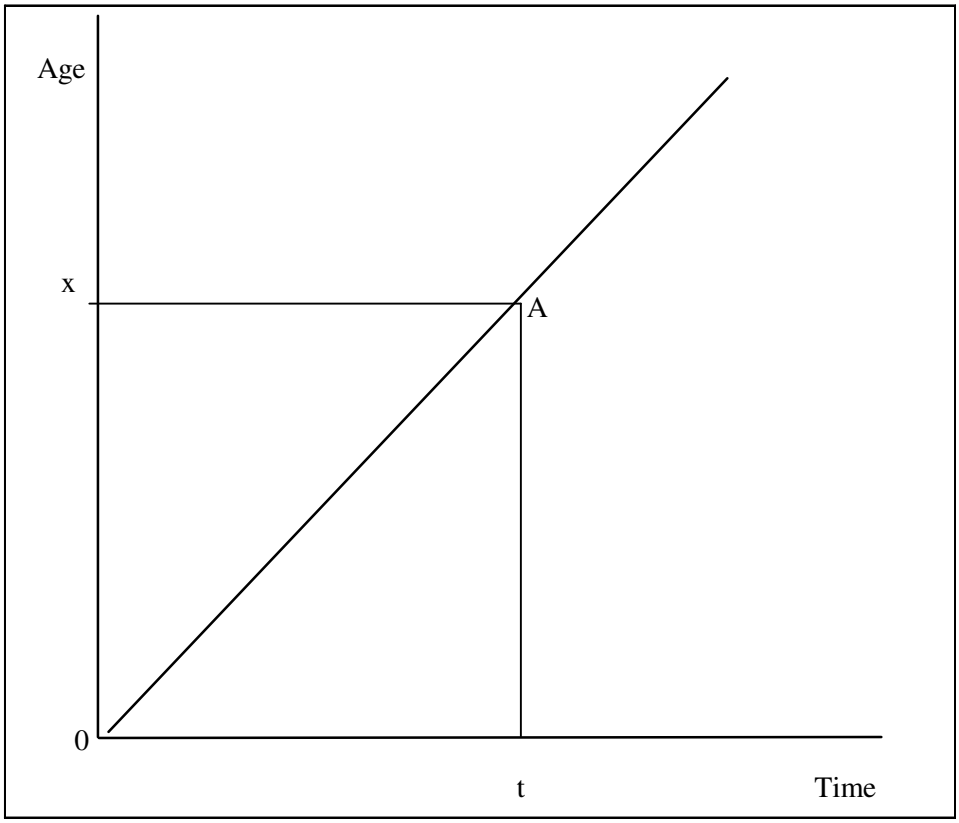
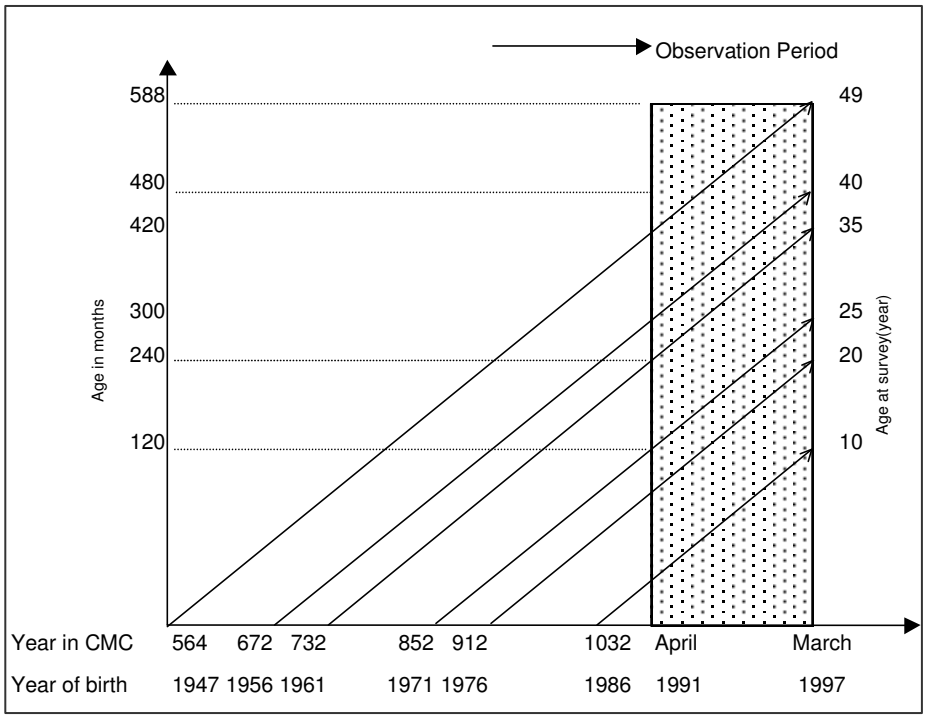


Figure 3.4 shows the Lexis diagram applied to the observation window used in the Bangladesh Demographic and Health Survey (BDHS) (from Khatun and Willekens, 2001, p. 11). The horizontal axis shows the year of birth of respondents and the CMC at birth. It also shows the year and the month of survey (March 1997). The calendar starts in April 1991. The vertical axis shows the age in years (right axis) and month (left axis). The rectangular area represents the observation window associated with the calendar. The 45-degree lines denote lifelines of respondents. They start at birth and end at survey date.

The lines are longer for older respondents indicating that retrospective surveys of life histories starting at birth record more information for older respondents than for younger respondents. The lifelines of younger respondents are truncated at an age at which several life events under study have not occurred yet.

The observation window is crucial for the recording of events and the estimation of the duration of exposure to the risk of experiencing the events. Events that occur outside of the observation window escape the observer and generally no information is available on risk periods before or after observation. Since the reconstruction of life histories is entirely based on transition rates or occurrence-exposure rates that are estimated from the number of events and the lengths of exposure recorded, the life histories that are reconstructed may be sensitive to the observation window.

Figure 3.4. Lexis diagram of exposure time by year of birth and age, in year and century month code, BDHS 1996-97



Using the Lexis diagram, the observation window may be characterized further. The onset of observation may be a point in time, an age, or an event. The Life History Calendar mentioned above start recording at a fixed point in time (April 1991).

Observation often starts at the time of a life event, such as birth, marriage, or entry into the labour market. The events occur at different points in time and, except for birth, at different ages. The event that defines the onset of observation will be referred to as the *event-origin.* The end of an observation is again a fixed point in time, a given age, or an event. Retrospective surveys record life histories from a given start until the survey date. Persons who leave the study population, because of death or emigration, are excluded from observation. In prospective studies, the sample is chosen and then followed up in time to study behaviour. The observation may be discontinued before the end of the study because of death, emigration, or withdrawal by the respondent. The event that leads to discontinuation of observation for reasons other than end of study is referred to as *attrition*.

3.3.2 Type of episodes and types of censoring

It is important to distinguish a process from the observation of the process. A process starts at an event-origin and ends at the end-point with the occurrence of the event under study. The observation window defines a segment of ongoing processes. Processes may start before the onset of observation in which case the start is not observed and the time to event cannot be determined. The end-point of processes may also be after the observation is terminated. The time to event cannot be determined either. The observation on a process may also start somewhere during the observation window or

the observation may end prematurely. The time to event can only be determined for processes that start and end during the observation window. The observation on processes leads to two types of episodes, closed and open episodes. *Closed episodes* are episodes that begin and end during the observation window (see also Figure 3.1). The observation covers the entire process. The time (age) at onset and the time at termination are known. *Open episodes* are episodes with unknown duration. They are observations on processes that either start or end outside of the observation window.

Instead of processes, we may consider persons exposed to the risk of experiencing an event. Exposure time and process time coincide. At the start of the observation, some persons may already be at risk of an event of interest and the duration at risk is not known. At the end of the observation, some persons may still be at risk, i.e. have not experienced the event of interest. The duration of exposure may be unknown for a number of reasons. First, exposure (episode) starts before the onset of observation. Second, the respondent enters observation during the observation period (and is already exposed at entry). Third, the observation is discontinued while the person is still exposed. Fourth, the respondent leaves the observation before the end of the study period due to a reason unrelated to the events being studied (drop-out). The fourth reason is particularly relevant in prospective studies, such as follow-up studies and panel studies.

Open episodes are examples of incomplete data. The exact length of an open episode remains unknown. However, the length is known to exceed a given duration; namely, the duration under observation. That information is important and should not be rejected by omitting all or some of the open episodes. In survival models and the life table, the information is considered in the estimation of the hazard rates or occurrence-exposure rates.

In the literature, incomplete observation on processes is referred to as censoring and truncation (see e.g. Klein and Meischberger, 1997, pp. 64ff and Mills, 2000, pp. 111ff). A censored observation is an incomplete observation. Censoring occurs when the process or exposure is only partially observed. An observation is truncated if the process is omitted from observation because the observer is unaware of its existence. Truncation involves a selection process. If a process starts before the period of observation or if it enters observation somewhere during the observation window (i.e. a respondent enters observation during the period), the observation is *left censored*. If a process continues when observation ends, either because the observation period ended or the respondent left, the end of the episode is not observed and the observation is *right censored*. If an observation is left censored, we do not know the start of the episode and hence the duration of exposure to the event of interest. For example, in life history calendars, observation starts at a particular point in time (April1991 in case of the Bangladesh Demographic and Health Survey 1996-97). From the calendar, we may know that a respondent had a first child at a particular date, and that she had a second child some time thereafter; however, we do not know when exactly she married. The duration of marriage at time of birth of the first child is not known from the calendar. It may be known from other data, e.g. when the year and month of marriage is recorded. In case of right- censored observation, we do not know the ending time of the episode and whether the episode resulted in the event under study. Some observations may be both left and right censored. In that case, the episode started before the onset of observation and was still intact when observation was discontinued. The length of that episode exceeds the

size of the observation window. Hence, only a segment or part of the episode is observed. That information is used in the estimation of the occurrence-exposure rates. When an episode is left and right censored, no event occurred during the entire observation period.

Interval censoring is a form of incomplete observation of survival time or time to event that can involve left and right censoring as well as left and right truncation. Interval censoring refers to a situation where a subject's survival time is known only to lie between two values. Interval censoring typically arises in panel studies where follow-up is done at fixed intervals, e.g. every six months. Events that occur between visits may be recorded but the exact time of occurrence may remain unknown.

Truncation involves a selection of processes or respondents to be included in the observation. The selection procedure applies to all subjects (Hosmer and Lemeshow, 1999, p. 256). *Right truncation* occurs when, by design of the study, there is a selection process such that data are available only on subjects who have experienced the event. An individual who has not yet experienced the event is excluded from observation (Klein and Meischberger, 1997, p. 65). For instance if a mortality study is based on data from a cancer registry, only subjects with cancer are included in the study. The time to diagnosis of cancer involves right truncation. The selection of ever-married women to be included in the Demographic and Health Survey is another example of right truncation. To be included in the survey, women must have experienced the event of marriage. Women who did not marry yet by survey date are excluded from the study. No information on these women is considered. *Left truncation* occurs when individuals with certain characteristics are selected for observation. For instance, a sample necessarily includes survivors only and generally also omits persons who migrated out of the region of observation. Left truncation also occurs when subjects of a given age are selected for observation and other subjects, who may also be at risk, are excluded (age truncation).

The types of censoring may be displayed in the Lexis diagram. Consider the observation period from April 1991 to March 1997. Events that occur during the interval are observed. Events that occur before April 1991 or after the survey date March 1997 are not observed. Events define episodes. Closed episodes begin and end during the observation window, i.e. they begin and end during the period April 1991 and March 1997. Left censored episodes start before April 1991, whereas right-censored episodes start before or in March 1997 but end after the survey month. Episodes that start before April 1991 and end after March 1997 are left *and* right censored.

A distinction is made between type of censoring. *Type I censoring* describes the situation when the observation is terminated at a fixed point in time. Subjects that did not experience the events under study are known to have 'survived' until the end of study or observation. In that case, the number of events observed is a random variable. In *Type II censoring*, the observation would be continued until a fixed proportion of subjects have experienced the event (e.g., we stop the experiment after exactly 200 marriages have failed). In this case, the number of marriage dissolutions is fixed, and the time of observation is the random variable. In *Type III censoring*, subjects enter the study at different time, but the study ends at a pre-determined time.

Age-time framework

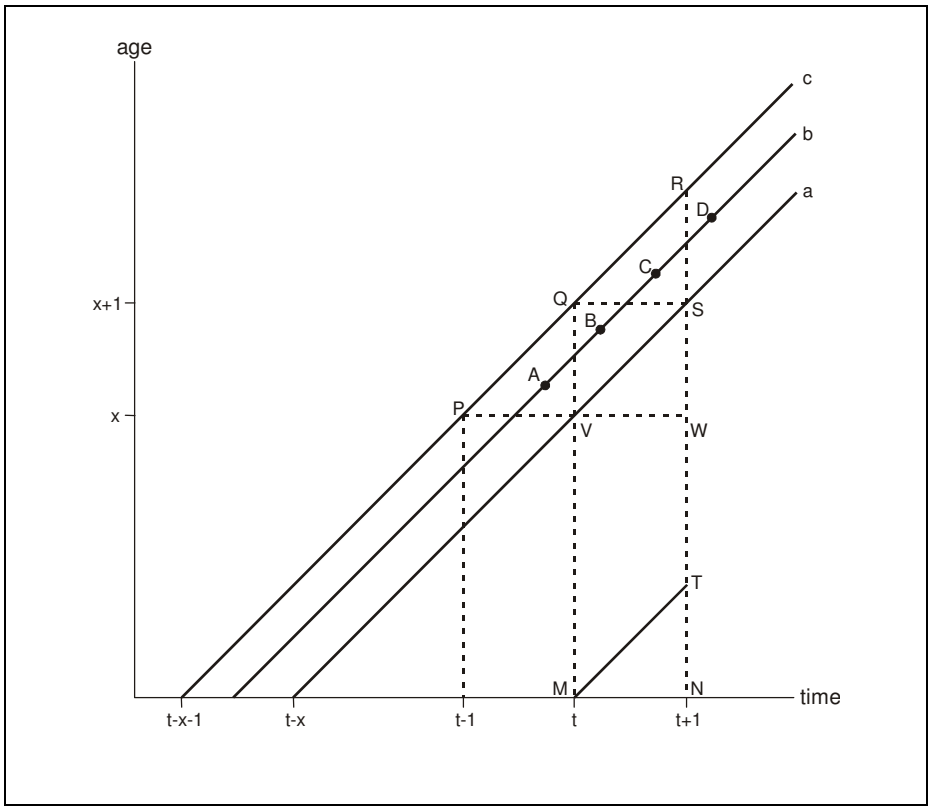
The Lexis diagram visualizes the age-time framework that is central to the biographic method or life course analysis. Individual lifelines are generally grouped into age intervals and time intervals. In other words, age and time become discrete variables. The grouping has important implications for the estimation of transition rates. Each grouping results in a different type of rate that affects further analysis, such as life-table analysis. In this section, we discuss the age-time observational plans that should be considered in life course analysis.

The grouping of individual lifelines on the basis of the time at event-origin results in a cohort. A cohort is commonly defined as a group of individuals born during the same period, usually one year or five years. The cohort concept may be broadened to a group of people that experienced any given event-origin during the same period. People who marry during the same period constitute a marriage cohort. People who enter the labour force during the same period constitute a labour force cohort. Lifelines may also be grouped on the basis of the time to event. Individuals experiencing the event during the same period are grouped. The combination of cohorts and discrete duration intervals defines characteristic grouping of the individual data on events and exposures. The grouping results in observation intervals that are delineated by age and time. This is important since each grouping results in a different type of occurrence-exposure rate. The observation interval may be viewed as a special case of an observation window, one that is defined in an age-time framework. The intervals may be visualized in a Lexis diagram (*figure 3.5*). The figure shows three lifelines (a, b and c). On lifeline b, four events are shown (A, B, C and D). The observation interval is delineated by the cohort, time (period) and seniority (age) segments considered. The figure shows an age interval (x, x+1), a time interval (t, t+1), and a cohort interval (t-x-1, t-x). If the cohort, time and age segments are fixed, the observation interval is a triangle. Data that are classified by age, period and cohort are referred to as age-period-cohort data or doubly-classified data. In general, only two of the three measures of time are available. Four types of observations exist depending on the observation intervals.

Period observation (or period-age observation). A period observational plan records the calendar year in which an event occurs as well as the age of the person at the time of the event (age at last birthday). The age is recorded in completed years. In figure 3.5, the period observation interval is represented by QSWV. This interval covers two cohorts. The rates based on the period observational plan are referred to as period rates.

Cohort observation (or cohort-age observation). A cohort observational plan records for a person experiencing an event the cohort to which the person belongs as well as the age in completed years at the time of the event. In figure 3.5, it is parallelogram QSVP. The observation period extends over two calendar years. The rates based on the cohort observation plan are cohort rates.

Figure 3.5. Lexis diagram



Period-cohort observation. A period-cohort observational plan records the calendar year in which an event occurs as well as the cohort to which the person belongs. In figure 3.5, it is parallelogram QRVS. The observation interval covers two age groups. Recording the cohort in the period-cohort observational plan is equivalent to recording the age at the beginning of the interval (i.e. at time t) or at the end of the interval (i.e. at time t+1). It is also equivalent to recording age in period difference. The latter is obtained by subtracting the year of birth from the year of occurrence of the event under study (Wunsch and Termote, 1978, p. 10). Rates based on the period-cohort observational plan are period-cohort rates.

Age-period-cohort observation. An age-period-cohort observational plan records for a person experiencing the event the calendar year in which the event occurs, the year of birth of the person, and the age of the person. In figure 3.5, the APC observational plan is represented by the triangle QVS. The observational interval covers only one cohort, one calendar year and one age. Rates based on the age-period-cohort observational plan are age-period-cohort rates.

Consider for instance the event of leaving the parental home. The cohort observational plan measures how many members of the cohort (t-x-1, t-x) leave the parental home at age x in completed years (i.e. between their birthdays x and x+1). The period-cohort observational plan measures how may members of the cohort (t-x-1, t-x) leave the parental home in calendar year (t, t+1). The period observational plan measures the number of persons leaving the parental home at age x (in completed years at time of

leaving home, i.e. aged x to x+1) in year (t, t+1). Note that in the cohort and period observational plans, age is measured at time of the event. In the period-cohort observational plan, age is measured at the beginning of the interval (time t). The latter measurement is equivalent to the measurement of the birth cohort.

For life-table analysis, the cohort observational plan is to be preferred. For projection purposes, a period-cohort observational plan is the ideal one. The period observational plan often serves as an approximation to one of the other plans. Many texts in demography implicitly assume that the data that are recorded for discrete age and time intervals are period data. See e.g. Keyfitz, 1968, p. 9; Rogers, 1975, 1995, p. 41 and Schoen, 1988, p. 11).

Transition rates are ratios of occurrences over exposure time. With each observational plan is associated a type of transition rate. We distinguish cohort rates (age-cohort rates), period rates and period-cohort rates. The estimation of transition rates require the measurement of exposure time. Consider the transition rate for persons aged x to x+1. The timing of transition is given in CMC. The age at transition is the CMC at event minus CMCB (CMC at birth) and is denoted by AGE if age is the exact age and IAGE if the age is measured in completed years (integer value). To determine the length of the episode, we consider each age interval from exact age x to exact age x+1 as an observation window. The observation starts at exact age x and is discontinued at exact age x+1. The age may be expressed in terms of CMC. The age x is reached at CMCX1 = CMCB + x*12 and the age x+1 is reached at the beginning of month CMCX1+12. Hence the end of the observation period is CMCX2 = CMCX1 + 11. The interval between ages x and x+1 is the interval between CMCX1 and CMCX2. The number of events experienced by respondents of age IAGE is easily determined. It is the number of events that occur during the interval from CMCX1 and CMCX2. The exposure time is the number of months an individual is at risk of experiencing the event of interest. Several situations may be distinguished (assuming an interval of 12 months):

a. The individual is present at x and at x+1 and does not experience the event during the observation window from x to x+1. The duration of exposure is equal to 12 months.
b. The individual is present at x and experiences the event during the interval. The duration of exposure is CMC – CMCX1.
c. The individual enters the population at risk during the interval (x, x+1), does not experience the event, and is present at the end of the interval. The duration at risk is CMCX2 – CMCI where CMCI is the CMC at entry into the population at risk.
d. The individual enters the population at risk during the interval (x, x+1) and experiences the event before reaching age x+1. The duration at risk is CMC – CMCI.
e. The individual is present at x and leaves the population at risk before reaching age x+1 without experiencing the event. The duration at risk is CMCE – CMCX1 where CMCE is the CMC at leaving the population at risk.

The next chapter focuses on the estimation of the expected waiting time to the event from incomplete observations on process time.

4. Single transitions and single episodes

Life is a sequence of states, transitions and episodes. The subject of this chapter is a single episode and the expected duration of that episode. An episode is an uninterrupted period in a given state or with a given (primary) attribute. It starts with an event (event-origin) and ends with an event (end-point). The event at end-point implies a transition to a different state. In its most simple form, life may be viewed as a single episode starting at birth and ending at death (end-point). It is an uninterrupted period in the living state. Life may also be viewed as consisting of three episodes: childhood, adulthood and old age. Childhood starts at birth and ends with the transition to adulthood, whilst adulthood ends with the transition to old age. Episodes are commonly defined for each domain of life. The employment career consists of employment episodes, unemployment episodes and episodes out of the labour force. The professional career consists of job episodes and periods out of a job (e.g. for training). The fertility career consists of episodes without children, with one child, with two children, etcetera.

A focus on episodes follows from an event-based approach to the life course and the data are episode data. The study of single episodes involves the definition of two states and the event that results in a transition from the origin state to the destination state. The length of the episode of interest is equal to the sojourn time in the origin state which is identical to the duration of a process.

The duration of an episode or process is often not known, as discussed in detail in chapter 3. The measurement of the length of an episode requires the observation of the starting time and ending time. An episode with both starting and ending times observed is known as a *closed episode*. In this case the observation is complete. The duration of closed episodes can be observed directly since the event-origin and the terminating event are in the observation window. More often, however, the observation is incomplete as either the starting or the ending time is situated outside of the observation window and remains unobserved. That situation is known as an *open episode*. The true process time of an open episode remains unknown. What is known, however, is that the process time exceeds the duration of observation on the process. The length of the episode cannot be observed but must be estimated using information on other open and closed episodes. This chapter presents techniques to determine the expected length of episodes or process time. The expected process time is also known as the (expected) waiting time to the event. If the event is death, the process time is the lifetime and the expected process time is the life expectancy or expectation of life.

Two approaches exist to determine the expected length of episodes that include open episodes: the parametric method and the non-parametric method. The *parametric* method involves the use of a model that predicts the likelihood of an event on the basis of the process time, i.e. duration of the process. The model describes the duration dependence. Parametric models of duration dependence include the exponential distribution. It assumes that the transition rate does not vary with process time. As a result, the probability that the length of an episode exceeds successive values declines exponentially. The Gompertz distribution assumes that the transition rate varies exponentially with process time. The Weibull distribution assumes that the

transition rate is a power function of process time. Some distributions describe transition rates that first increase with process time, reach a peak and then decline as the process time further increases. Examples include the log-normal distribution. Two distributions are popular in demography: the Heligman-Pollard model to describe mortality by age and the Coale-McNeil model of first marriage by age. Each distribution imposes a particular duration dependence of the process being studied.

The advantage of parametric models is parsimony. The number of parameters to be estimated is small. A disadvantage is that the model does not capture some particularities of duration dependence. For instance, human lifetimes show an increased hazard around the legal driving age (18 years in many countries). Another disadvantage is what Hougaard (2000, p. 41) calls the *conditioning aspect*. Consider a 75-year old male and suppose we are interested in the probability of surviving to age 76. In a parametric model, this conditional survival probability is based on all the information in the data set since the shape of the distribution is estimated using information on all ages, including males who died at age 20.

The *non-parametric* approach does not impose a duration dependence onto the data. Since it does not make any assumptions about the duration dependence, it is particularly useful when interest centres on conditional probabilities. Non-parametric methods are distinguished on the basis of the time scale adopted and the statistic estimated. The process time can be measured in continuous time or in discrete time (time intervals). In case of continuous time, the risk of experiencing the event or transition is measured for every moment of the episode. In case of discrete time, the episode is divided into discrete time intervals and the risk of an event is determined for each interval. In order to estimate the risk of an event, the number of persons at risk, i.e. the *risk set*, must be determined. The risk set at a given point in time is the number of subjects at risk at that point in time. The measurement of the risk set is a cumbersome process when time is continuous. Since the risk set does not change unless an event occurs, the measurement of the risk set is limited to moments just before events occur. For instance consider a group of students who start a study at the beginning of the academic year. We want to know the probability that students complete the academic year. The study is known for its drop-out. Consider ten students starting. In the 3rd week, one student quits the course. In the 5th week, two students decide to continue the study but at a different school. Since they are out of sight, their fate remains unknown. During the 6th week another student drops out. The risk set varies from ten students to nine students at the beginning of the 4th and 5th week, to seven students at the beginning of the 6th week and six students at the beginning of the 7th week. A student belongs to the risk set as long as he does not discontinue the study. In other words, students are at risk as long as they occupy the origin state. During the first six weeks, two episodes of study are terminated by a drop-out and two observations are censored. At the beginning of the seventh week, six students are left. Two dropped out and two switched to another school (and may have dropped out there). The probability of drop-out is more than 20 percent but less than 40 percent. It is estimated to be 23 percent. The censoring must be taken into account when estimating the risk of drop-out.

To estimate the risk of an event in the presence of open intervals (censoring or truncation) and in continuous time, two approaches exist. The first estimates the survival probability directly from the data. The second estimates the duration-specific

transition rates from the data and the survival function from the cumulative transition rates. The probability or rate is determined at every point in time when an event occurs. The first approach is the Kaplan-Meier method. It was developed by Kaplan and Meier (1958) to handle incomplete observations. The second approach estimates the cumulative transition rate at every point in time when an event occurs. It is the Nelson-Aalen estimator. It was developed by Nelson (1969) and reformulated by Aalen (1975). Both approaches require that all episodes are sorted according to their ending times.

Non-parametric methods that use discrete time by dividing episodes in time intervals are known as the life table method or actuarial method. The intervals should be sufficiently small to accurately describe the duration dependence and to estimate the average length of episodes. When the length of the intervals tends to 0, discrete time approaches continuous time and the life table method tends to the Kaplan-Meier method and the Nelson-Aalen method. Note that no assumption is made about the duration dependence of the process across duration intervals. However, an assumption is made about duration dependence of the process within the intervals. In the empirical estimation of the life table, two techniques are distinguished. The first estimates transition *probabilities* directly from the data by dividing the number of events during the interval by the risk set at the beginning of the interval. Censored observations are taken into account in the risk set. The second technique estimates transition *rates* by dividing the number of events during the interval by the total duration of exposure. Transition probabilities are obtained from the transition rates. The Kaplan-Meier estimator, the Nelson-Aalen estimator and the life table method are described in this chapter.

When the population is stratified into subpopulations, an intermediate class of methods exists. It does not impose a pattern of duration dependence onto the data but it imposes a relation between the duration dependence in the different strata. For instance, at any duration, the transition rate in one stratum is proportional to the transition rate in another stratum. The ratio of transition rates, i.e. the risk ratio or relative risk, does not vary with duration since the event origin. This method is the *semi-parametric method*. The well-known Cox proportional hazard model is a semi-parametric method. The baseline hazard does not follow a predefined pattern and is said to remain unspecified.

4.1 Process time is continuous time

4.1.1 The Kaplan-Meier estimator of survival function

The simplest non-parametric estimate of the survival function is the empirical distribution. The empirical distribution is the observed distribution. The distribution is easily obtained if the observations are complete, i.e. if all episodes are closed episodes. Consider the example on students in the previous section. If no student drops out or changes school, and all students are followed till they complete the education, the time to completion is observed for all students and the empirical distribution of completion time can easily be constructed. The problem arises when observations are incomplete because students drop out, change schools, or the observation ends before all students completed their education. Kaplan and Meier

(1958) extended the method to incomplete observations, when not all event times are recorded. The aim of the Kaplan-Meier procedure is to estimate a survival function non-parametrically in the presence of right censoring. The information required to derive the Kaplan-Meier estimator consists of (1) date of entry into observation, (2) date of event, and (3) the ending time of observation in case the ending time of the event is not observed. The entry into observation is assumed to coincide with the onset of the process. Left censoring is therefore not considered. We consider two examples. One uses hypothetical data and the other uses data on job episodes. The hypothetical example is presented first and is used to illustrate the method.

Example using hypothetical data
In the first example, we suppose that we have information on a sample of 22 individuals. Namboodiri and Suchindran (1987) use the data to explain how to construct a life table from observational data (survey data). The information is collected retrospectively as part of a cross-sectional survey and the date of interview is the end of the observation period. Since the data are collected retrospectively, no one drops out during observation. For each individual, three dates are given: the date at entry into observation, the date of the event under study, and the date of interview. From these dates we determine the durations of the episodes by counting the number of days between entry into observation and the event or the survey whatever comes first. The dates and the durations are shown in *table 4.1*. An indicator function is added to indicate whether the episode is a closed episode or an open episode. Closed episodes are terminated by the event. Open episodes are terminated by end of observation. Of the 22 episodes, 12 are closed and 10 are open. The duration of exposure, or the duration at risk of experiencing the event, is the time between entry into observation and event or survey whatever comes first. Note that the indicator function indicates which 'event' terminates the observation, the occurrence of the event under study or the end of observation. If t_E is the process time at the event and t_C is the process time at censoring, then the observed process time is $\min[t_E, t_C]$. Kaplan and Meier (1958, p. 458) refer to t_C as the limit of observation.

It is generally assumed that each process time or duration since the onset of the process corresponds to exactly one event, implying that no two events occur at the same time. The assumption is required for an absolutely continuous survival function. It may happen that a number of events or events and censoring occur at the same duration. For instance, process 4 ends in an event after 37 days and process 20 ends in censoring after exactly the same number of days. When there are several observations (evens and/or censoring) at the same time point, ties in the data exist. Tied process times or event times are rules rather than exception.

Table 4.1. Hypothetical survey data

ID	Date of entry in observation	Date of event	Date of interview	Duration to event Days	Duration to interview Days	Duration Exposure Days	Closed/ Open
1	Jan-02	Feb-11	May-25	40	143	40	0
2	Jan-17	May-04	May-17	107	120	107	1
3	Jan-18	-	May-10	-	112	112	0
4	Jan-22	Feb-28	May-13	37	111	37	0
5	Feb-10	May-17	May-23	96	102	96	1
6	Jan-30	Feb-12	May-15	13	105	13	1
7	Apr-04	-	May-06	-	32	32	1
8	Apr-29	-	May-27	-	28	28	0
9	May-18	-	May-29	-	11	11	0
10	May-20	-	May-31	-	11	11	1
11	May-15	-	May-18	-	3	3	1
12	Feb-05	Feb-25	May-19	20	103	20	0
13	Feb-05	Apr-18	May-10	72	94	72	1
14	Feb-06	May-18	May-28	101	111	101	1
15	Feb-26	-	May-22	-	85	85	0
16	Mar-10	-	May-25	-	76	76	1
17	Mar-11	May-08	May-12	58	62	58	0
18	Mar-28	-	May-29	-	62	62	0
19	Mar-15	Mar-23	May-10	8	56	8	1
20	Apr-13	-	May-20	-	37	37	1
21	Apr-04	May-09	May-11	35	37	35	1
22	Apr-25	May-16	May-31	21	36	21	0

- censoring.
Source: Namboodiri an Suchindran, 1987, p. 54.

In *table 4.2* the cases are arranged in ascending order of duration. The shortest episode is three days (ID 11) and the longest is 112 days (ID 3). Subject ID three entered observation on 18th January and did not experience the event yet at survey date on 10th May. The long interval is an open interval that is terminated at the end of observation at the time of interview. When a closed episode is of the same length as an open episode, the closed episode is listed first (see case ID 4 and 20). Some individuals may experience the event on the same day (tied process times). Column (3) shows the number of episodes that are under observation at the beginning of the day shown in column (2). At the beginning of the 3rd day, 22 episodes (individuals) are under observation. During that day, one episode is terminated because of the survey. In other words, one individual (ID 11) is interviewed 3 days after the start of observation. The observation start at the 15th May and ends at interview on the 18th. It is the shortest episode.

Table 4.2. Kaplan-Meier estimation of survival function

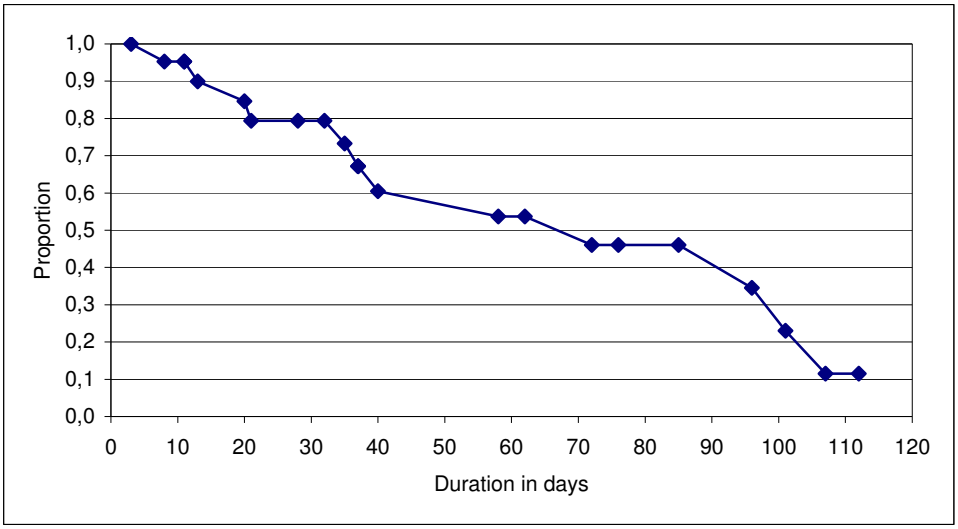
Case ID	Duration exposure Days	At risk at beginning of day	Event/ Censoring	Risk set	Prob of event	Observed survival proportions	Survival function
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
11	3	22	0				1.0000
19	8	21	1	21	0.0476	0.9524	0.9524
9	11	20	0				0.9524
10	11	19	0				0.9524
6	13	18	1	18	0.0556	0.9444	0.8995
12	20	17	1	17	0.0588	0.9412	0.8466
22	21	16	1	16	0.0625	0.9375	0.7937
8	28	15	0				0.7937
7	32	14	0				0.7937
21	35	13	1	13	0.0769	0.9231	0.7326
4	37	12	1	12	0.0833	0.9167	0.6716
20	37	11	0				0.6716
1	40	10	1	10	0.1000	0.9000	0.6044
17	58	9	1	9	0.1111	0.8889	0.5372
18	62	8	0				0.5372
13	72	7	1	7	0.1429	0.8571	0.4605
16	76	6	0				0.4605
15	85	5	0				0.4605
5	96	4	1	4	0.2500	0.7500	0.3454
14	101	3	1	3	0.3333	0.6667	0.2302
2	107	2	1	2	0.5000	0.5000	0.1151
3	112	1	0				0.1151

At the end of the 3rd day of observation, 21 episodes are left. The first event occurs after eight days of observation. It occurs to subject ID 19, who enters observation on 15th March and experiences the event on 23rd March. At that moment, 21 episodes are under observation and hence 21 individuals are at risk. The probability of the event is therefore 1/21 = 0.0476 or 4.8 percent. The observed proportion of episodes that were under observation at the beginning of the 8th day, that are still under observation at the end of the day is given in column (7). It is equal to the observed proportion of individuals that complete the day without experiencing the event among those who have not had the experience as of the beginning of the day. Column (8) shows the empirical survival function. The survival function is obtained by calculating the continued product of the entries in column (7) up to and including the one for the day shown in column (2).

The column shows, for instance, that the probability of remaining free from experiencing the event for the first 85 days is 46 percent. One episode, 11 percent of all episodes, is completed after 112 days of exposure. *Figure 4.1* displays the empirical survival function. The survival function exhibits jumps and large flat regions due to sparse data.

The Kaplan-Meier estimator is based on the calculation of the number of episodes under observation immediately before an event occurs. That number is the *risk set*. The calculation of the risk set requires that the cases are arranged in ascending order of duration. Let m denote the total number of episodes (in the illustration, m = 22) and assume that all episodes start at time zero. Let k denote an episode (k = 1, 2, …, 22).

Figure 4.1. Empirical survival function



Let $Y_k(t)$ represent a random variable that indicates the status of the i-th process at time t. It is 1 if the process or episode is still under observation at time t and it is 0 otherwise. The indicator variable is a binary variable that follows a Bernoulli distribution. Time t is the process time, i.e. the time elapsed since the onset of the process that is the date of entry into observation (in the absence of left censoring) and is expressed in days. The indicator function is 1 from onset of the process to event or censoring whatever comes first. A process that is not terminated yet at t is at risk of being terminated. The number of processes or episodes under observation is the risk set $R(t) = \sum_{k=1}^{22} Y_k(t)$. The risk set at t is equal to the total number of episodes minus the events that occurred before t and the number of censored observations before t. In other words, the risk set is the number of episodes with ending time larger than t. If an event and a censoring occur at the same time, it is assumed that the censored episode is included in the risk set. Consider it as the event occurring during the day and censoring at the end of the day. The episode 4 (ID = 4) ends with an event in day 37 and the episode 20 is censored at the same day. The censored episode contains the information that there is no event up to *and including* the observed ending time of the episode. Therefore, the censored episode is included in the risk set.

The probability of an event is determined for every time an event occurs. In the hypothetical data, no two events occur at the same time (no tied process times). The probability of an event at the time an event occurs is therefore 1 over the risk set [q(t) = 1/R(t)]. When ties in the data exist, i.e. when more than one event occurs at the same time, the equation is different. Recall that t is the process time in days. Let O(t) denote the number of events during day t since the onset of the process. If censoring occurs during the same day, it is assumed to occur at the end of the day. R(t) is the risk set at the beginning of the day. The proportion of processes at the beginning of the day that end in an event during the day is O(t)/R(t). For days without events, O(t) = 0. The proportion of initial processes that do not end in an event or censoring by day t, is the survival function. It is equal to

$$S(t) = \prod_{s=0}^{t-1}\left[1 - \frac{O(s)}{R(s)}\right] = \prod_{s=0}^{t-1}[1 - q(s)] = \prod_{s=0}^{t-1} p(s)$$

where p(s) = 1 – q(s). The survival function is a right continuous decreasing step function, with changes at event times. If the largest time value corresponds to an event, S(t) becomes 0 eventually. If the largest time value corresponds to a censoring, S(t) will have a non-zero value at that time point and be undefined afterwards. One consequence is that the waiting time to the event (or mean lifetime of the process) cannot be estimated (Hougaard, 2000, p. 71). A solution is to evaluate the median lifetime rather than the mean. That is why software packages such as SPSS show the median waiting time to the event rather than the mean waiting time.

The variance of the survival function is obtained using the Greenwood formula

$$Var[S(t)] = [S(t)]^2\left[\sum_{s=0}^{t-1}\frac{O(s)}{R(s)[R(s) - O(s)]}\right] = [S(t)]^2\left[\sum_{s=0}^{t-1}\frac{q(s)}{R(s)p(s)}\right]$$

since q(s)=O(s)/R(s) and [R(s) – O(s)] = p(s) R(s)

In the absence of censoring, the variance of the survival function is the variance of a single binomial proportion:

$$Var[S(t)] = \frac{S(t)[1 - S(t)]}{m}$$

with m the number of episodes (see also Newman, 2001, p. 174).

One consequence of that formula is that intervals without events have a transition probability of 0 and a variance of 0, suggesting that we are absolutely sure that events cannot occur during that interval. This is a problem since a larger sample size may result in events in intervals that do not have events in small samples. In other words, the 0 is not a structural 0 but a sample 0, i.e. the absence of an event at that time is related to the size of the sample.

Confidence limits are obtained assuming that the survival probability is normally distributed with mean S(t) and variance Var[S(t)]. The 95 percent confidence interval is given by

$$S(t) \pm 1.96\sqrt{Var[S(t)]}$$

One problem with the Greenwood formula is that the upper and lower bounds may be outside the 0-1 range. Kalbfleisch and Prentice (1980, p. 15; quoted by Newman, 2001, p. 174) developed an estimate of a confidence interval with upper and lower bounds always between 0 and 1. It uses the log-minus-log transformation of the survival function:

$$Var[\log(-\log S(t))] = \frac{1}{\log S(t)}\sum_{s=1}^{t-1}\frac{q(s)}{p(s)R(s)}$$

The 95 percent confidence interval for log[-logS(t)] is obtained from

$$(z_1, z_2) = \log[-\log S(t)] \mp 1.96\sqrt{Var[\log\{-\log S(t)\}]}$$

by inverting the log-minus-log transformation: exp[-exp(z_i)]. Note the $\mp$ sign rather than the usual ±.

The Kaplan-Meier estimation yields the survival function. Form it, the hazard rate or event rate can be obtained. Note that the survival function may be written as

$$S(t) = \exp[-A(t)]$$

where A(t) is the cumulative hazard function. It is the sum of all transition rates prior to duration t. In the presence of censoring, the cumulative hazard is obtained from

$$A(t) = -\ln S(t)$$

The variance of A(t) is

$$Var[A(t)] = Var[\ln S(t)] = \frac{Var[S(t)]}{[S(t)]^2} = \sum_{s=0}^{t-1} \frac{q(s)}{p(s)R(s)}$$

The cumulative hazard function for successive values of t may be used to estimate the average transition rates during unit time intervals. It is the discrete analogue of the hazard function or intensity function and is often referred to as the annual hazard rate or occurrence-exposure rate. The occurrence-exposure rate during the interval from t to t+1 is

$$m(t) = A(t+1) - A(t)$$

which implies that

$$A(t) = \sum_{s=0}^{t-1} m(s)$$

A note on the waiting time to event

The survival function is a decreasing step function. The initial value is 1 and it decreases at each event time. If the largest time value corresponds to an event, the survival function becomes 0 eventually. If the largest time value corresponds to a censoring, the function will have a non-zero value at that time point and is undefined afterwards. The censoring makes it difficult to estimate the right tail of the survival function, and this tail can have a marked influence on the mean. One consequence of this is that the expected time to event, i.e. the mean length of an episode, cannot be estimated (Hougaard, 2000, p. 71). The expected time to event, or mean length of an episode, is the area under the survival function. An alternative to the mean is to evaluate the median, which is the duration x for which S(x) = 0.5. In most packages such as SPSS, the Kaplan-Meier and life-table procedures present the median

duration. Another solution is to assume that the survival function is 0 after the largest time and to calculate the mean duration accordingly.

4.1.2 Nelson-Aalen estimator of cumulative hazard function

In the previous section, we saw that the information required to derive the Kaplan-Meier estimator consists of (1) date of entry into observation, (2) date of event, and (3) the ending time of observation in case the event did not occur.

The process time to event or censoring, i.e. the duration of the episode at the occurrence of the event or censoring, was measured from the date at entry into observation. Left censoring and delayed entry into observation was assumed to be absent. The onset of observation was assumed to coincide with the start of the process or episode under study. As a consequence, each episode was characterized by two random variables. The first indicates the length of the observation period and the other indicates whether observation ended in the event under study or end of observation (right censoring). The approach does not handle well left censoring, left truncation or delayed entry, and multiple events. To handle these situations, a different approach is required. The approach is to determine, for any duration of the process under study, whether a subject is under observation or not. Let t denote chronological time measuring the time since the onset of the process. Two random variables are introduced. One indicates the exposure status of the subject at time t and the other the number of events that occurred up to time t. Let $Y_k(t)$ represent a time-varying indicator variable that takes on the value 1 if subject k is under observation at time t and the value 0 otherwise. If a subject enters observation after the onset of the process, the observation is said to be left truncated and is also referred to as delayed entry. The total *observed* exposure time or duration at risk before time t is ${}_E Y_k(t) = \int_0^t Y_k(s)ds$ and the number of subjects at risk at t, or more formally immediately before t, is $Y_+(t) = \sum_{k=1}^{n} Y_k(t)$. The second random variable is $N_k(t)$. It denotes the number of *observed* events in the interval from 0 to t for subject k. If a subject can experience at most one event, the value is 0 or 1. The total number of events up to and including t is $N_+(t) = \sum_{k=1}^{n} N_k(t)$.

This approach handles left censoring/truncation, multiple events and multiple at –risk intervals, and is easy to generalize to more complicated situations such as multistate models of transition data (Andersen and Keiding, 1996, p. 181; Hosmer and Lemeshow, 1999, p. 254 and Therneau and Grambsch, 2000, p. 4). In the approach the emphasis is not on the estimation of the survival function but on the estimation of the hazard rate or transition rate. The estimation method was developed by Nelson (1969) and was generalized by Aalen (1975, 1978) who also provided a theoretical basis for the estimation. The theoretical basis is the theory of counting processes. $N_k(t)$ is a *counting process* The estimator became known as the *Nelson-Aalen estimator*. The theory of counting processes plays a central role in modern survival data and event history data analysis. For a brief introduction, see Hosmer and Lemeshow (1999). This section is based on Andersen *et al.*, (1993) and Therneau and Gramsbch (2000). We derive the Nelson-Aalen estimator using the hypothetical data shown in table 4.1.

Example using hypothetical data
Consider the data in table 4.1 and suppose that we start the observation on January 1st. Hence all individuals enter observation after the start of the observation (delayed entry). The first individual entering observation enters on January 2nd, the second on January 17th, etcetera. On January 31st, 3 individuals are under observation (ID 1, 2 and 3). They have a combined exposure time of (31-2) + (31-17) + (31-18) = 56 person-days. During that period, no event is recorded. The first individual (ID 1) experiences the event on February 11th, i.e. 40 days after entry into observation. By February 28, 20 individuals enter observation and 4 experience the event. The total exposure time of all individuals combined is (31-2+11) + (31-17+28) + (31-18+28) + (31-22+28) + (28-10) + (31-30+12) + (25-5) + (28-5) + (28-6) + (28-26) = 258 person-days. The rate at which the event occurs during the period January – February is 4/258 = 0.0155 per person-day, i.e. one event per person in 64 days. The total duration of exposure is 1065 person-days and the total number of events is 12. The event rate is 12/1065 = 0.01127 per day.

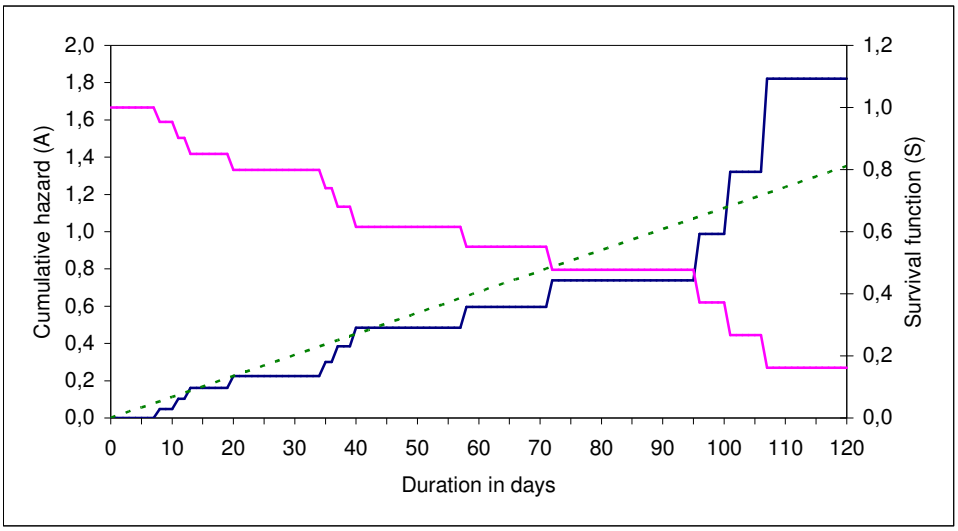
The first individual (ID 1) is under observation from day 2 to day 40. The exposure indicator variable $Y_1(t)$ is therefore 0 for the first day, 1 for day 2 to day 40, and 0 afterwards. $N_1(t)$ is 0 for the first 39 days and 1 for day 40 and later days. Individual 19, who enters observation on March 15th and experiences the event on March 23rd, experiences the event after 8 days. Individual 11 is interviewed after 3 days of observation without experiencing an event. To estimate the duration-specific hazard, the observations are arranged in ascending order of duration. The shortest episode is three days (ID 11) and the longest is 112 days (ID 3). The sorted observations are shown in *table 4.3*, columns 1 to 4. The first event is recorded after 8 days. At that moment, 21 individuals are at risk. The hazard rate is 1/21 = 0.0476. The second event is recorded after 13 days and at that time 18 individuals are at risk. The hazard rate is 1/18 = 0.0556 and the cumulative hazard is 1/21 + 1/18 = 0.1032 (Table 4.3, columns 5 and 6). This method gives the Nelson-Aalen estimator of the cumulative hazard. The cumulative hazard and the associated survival function are shown in *figure 4.2*. The figure also shows the cumulative hazard in case of a constant average hazard (and hence exponential survival function). The average hazard is obtained by dividing the total number of events (12) by the total exposure (1065 person-days). A line through 0 with slope 0.01127 is superimposed on the plot. The data are clearly not compatible with the hypothesis of a constant hazard.

As was the case with the Kaplan-Meier estimator, the Nelson-Aalen estimator is based on the calculation of the number of episodes under observation immediately before the event occurs and the process terminates. Consider n episodes (k = 1, 2, ..., n). The index k may refer to episode, process or individual. $Y_k(t)$ denotes whether the k-th process is intact at duration t or is already terminated by that duration (Andersen *et al.*, 1993, p. 124). $Y_k(t)$ is the *at risk process.* It is an *individual counting process.* The process indicates whether a process (or individual) is at risk at duration t. $N_k(t)$ denotes the number of events experienced by subject (or process) k during the interval (0,t).

Table 4.3. Nelson-Aalen estimator of cumulative hazard function

CASE ID	Duration Exposure Days	At risk beginning of day	Event/ censoring	Hazard Rate	Cumulative hazard	Survival function
(1)	(2)	(3)	(4)	(5)	(6)	(7)
11	3	22	0	0.0000	0.0000	1.0000
19	8	21	1	0.0476	0.0476	0.9535
9	11	20	0	0.0000	0.0476	0.9535
10	11	19	0	0.0000	0.0476	0.9535
6	13	18	1	0.0556	0.1032	0.9020
12	20	17	1	0.0588	0.1620	0.8504
22	21	16	1	0.0625	0.2245	0.7989
8	28	15	0	0.0000	0.2245	0.7989
7	32	14	0	0.0000	0.2245	0.7989
21	35	13	1	0.0769	0.3014	0.7398
4	37	12	1	0.0833	0.3848	0.6806
20	37	11	0	0.0000	0.3848	0.6806
1	40	10	1	0.1000	0.4848	0.6158
17	58	9	1	0.1111	0.5959	0.5511
18	62	8	0	0.0000	0.5959	0.5511
13	72	7	1	0.1429	0.7387	0.4777
16	76	6	0	0.0000	0.7387	0.4777
15	85	5	0	0.0000	0.7387	0.4777
5	96	4	1	0.2500	0.9887	0.3721
14	101	3	1	0.3333	1.3221	0.2666
2	107	2	1	0.5000	1.8221	0.1617
3	112	1	0	0.0000	1.8221	0.1617

Figure 4.2. Nelson-Aalen estimator of cumulative hazard function and survival functio. hypothetical example



In elementary processes subjects experience at most one event. $Y_k(t)$ is therefore 0 to 1. $\mathbf{N}(t) = \{N_1(t), N_2(t), \ldots, N_n(t)\}$ is a *multivariate counting process*. The number of processes that terminate during the infinitesimally small interval from t to t+dt is the increment $dN_+(t)$ which may be written as

$$dN_+(t) = \lim_{\Delta t \to 0}\left[N_+(t + \Delta t) - N_+(t)\right]$$

In case of discrete time, $\Delta N_+(t) = N_+(t+h) - N_+(t)$ where h is the length of the interval.

The expected number of processes that terminate in the interval from [t, t+dt) is $E[dN_+(t)]$ and the expected number of terminations or events per unit time interval is

$$\lambda(t) = \frac{E\left[dN_+(t)\right]}{dt}$$

Hence $E\left[dN_+(t)\right] = \lambda(t)\, dt$. In discrete time, with h the length of the interval, $E[\Delta_h N_+(t)] = \lambda(t)\, h$.

$\lambda(x)$ is the *intensity process* (Andersen *et al*., 1993, p. 51). It is a counting process involving expected values. The *cumulative intensity process* is

$$\Lambda(t) = \int_o^t \lambda(s)ds \qquad t \geq 0$$

The difference between the observed number of events during the period [0, t) and the expected number of events is the *martingale* of the counting process:

$M(t) = N(t) - \Lambda(t)$

where the subscript + is omitted. The martingale is also referred to as the compensated counting process (Andersen *et al*., 1993, p. 52).

The intensity $\lambda(t)$ depends on the hazard rate and the size of the risk set that is the number of processes at risk of terminating in the time interval [t, t+dt)

$$\lambda(t) = \sum_{i=1}^{n} \mu_i(t)\, Y_i(t)$$

where $\mu_k(t)$ is the hazard rate of process k at time t (i.e. in the interval [t, t+h)). The number of processes at risk is determined just before t. If we assume that all event times T are identical and independent, then $\mu_k(t) = \mu(t)$ for all k, and the aggregate process $N_+(t) = \sum_{k=1}^{n} N_k(t)$ is a univariate counting process with intensity process

$$\lambda(t) = \mu(t)\, Y_+(t)$$

and expected number of events during the interval [t, t+h)
$E\left[\Delta_h N_+(t)\right] = \mu(t)\, Y_+(t)\, h$

and

$$\mu(t) = \frac{E\left[\Delta_h N_+(t)\right]}{Y_+(t)\,h}$$

The hazard rate μ (t) is the probability that the process terminates during the infinitesimally small interval [t, t+dt), provided it is intact at t (did not terminate before t):

$$\mu(t) = \frac{\Pr\left\{t \le T < t + dt \,\middle|\, T \ge t\right\}}{dt}$$

An natural estimator of μ(t) is

$$\hat{\mu}(t) = \frac{dN_+(t)}{Y_+(t)} = \lim_{\Delta t \to 0} \frac{N_+(t + \Delta t) - N_+(t)}{Y_+(t)}$$

If the interval is very small, not more than one process terminates during the interval. Hence $\hat{\mu}(t) = 1 / Y_+(t)$. If more than one events occur during the interval, i.e. in the case of tied data, subintervals may be distinguished and the ties may be broken randomly and allocated to the subintervals (see e.g. Therneau and Grambsch, 2000, pp. 31-32). An alternative is to consider time as discrete and the cumulative hazard as a step function.

The cumulative hazard is $A(t) = \int_0^t \mu(s)\,ds$

A natural estimator of A(t) is

$$\hat{A}(t) = \int_0^t \frac{1}{Y_+(s)}\,ds$$

or, in case of ties in the data (Therneau and Grambsch, 2000, p. 9),

$$\hat{A}(t) = \int_0^t \frac{dN_+(s)}{Y_+(s)}\,ds$$

An alternative formulation is often used. Let ${}^{o}T_k$ denote the observed ending time of process k. The number of processes at risk of terminating at ${}^{o}T_k$ is $Y_+({}^{o}T_k)$. Hence the estimator of the cumulative hazard may be written as (Andersen *et al.*, 1993, p. 178)

$$\hat{A}(t) = \sum_{k:\,{}^{o}T_k < t} \frac{\Delta N_+(T_k)}{Y({}^{o}T_k)}$$

The estimator of the cumulative hazard is an increasing, right-continuous step-function with increments $\Delta N_+(T_k)/Y({}^oT_k)$ at the jump time ${}^oT_k$ of N_k. If the intervals are very small, at most one event occurs during the interval and $\Delta N_+(T_k) = 1$. The Nelson-Aalen estimator is essentially the sum of the scaled increments. Consider the hypothetical example in . The empirical estimate of the cumulative hazard at the 35th day (t = 35) is

$$\hat{A}(35) = \frac{1}{21} + \frac{1}{18} + \frac{1}{17} + \frac{1}{16} + \frac{1}{13} = 0.30142$$

In case one process terminates in an event and another is censored in the same day, the censoring is assumed to occur after the event. On day 37, an event and a censoring occur. The Nelson-Aalen estimator implies an assumption about the timing of the event and censoring that occur in a given interval. Consider the event occurring during the 8th day of observation. It is the event that occurs to process ID 19 the observation of which starts on March 15th. The event occurs on March 23rd. If we assume that the observation starts at the beginning of the day, the event also occurs at the beginning of the day. The same applies to censoring. An assumption that the event or censoring occurs at the end of the day implies that the observation starts at the end of the day. When a process ends in the first day of observation, the observed duration is zero. In order to avoid exposures of length zero, some authors assume that a process and an observation on a process start at the beginning of an interval and end at the end of an interval, irrespective whether the process ends in an event or the observation is censored. In that case, a one should be added to the duration in days (see below). In the remainder, I assume that events and censoring that occur on day t, occur at the end of that day.

Once the cumulative hazard is estimated, the survival function can easily be determined. The survival function is

$$\hat{S}(t) = \exp[-\hat{A}(t)]$$

The survival function *at the end* of the 35th day is exp[-0.30142] = 0.73977. The Kaplan-Meier estimator is 0.7326. The probability of surviving at a given duration is a little higher for the Nelson-Aalen estimator compared to the Kaplan-Meier estimator. A similar result is reported by Homser and Lemeshow (1999, p. 77). The relation between the two estimators is

$$S(t) = \Pr\{T > t\} = \exp\left[-\int_0^t \mu(s)ds\right] \approx \prod_0^t [1 - \mu(s)ds]$$

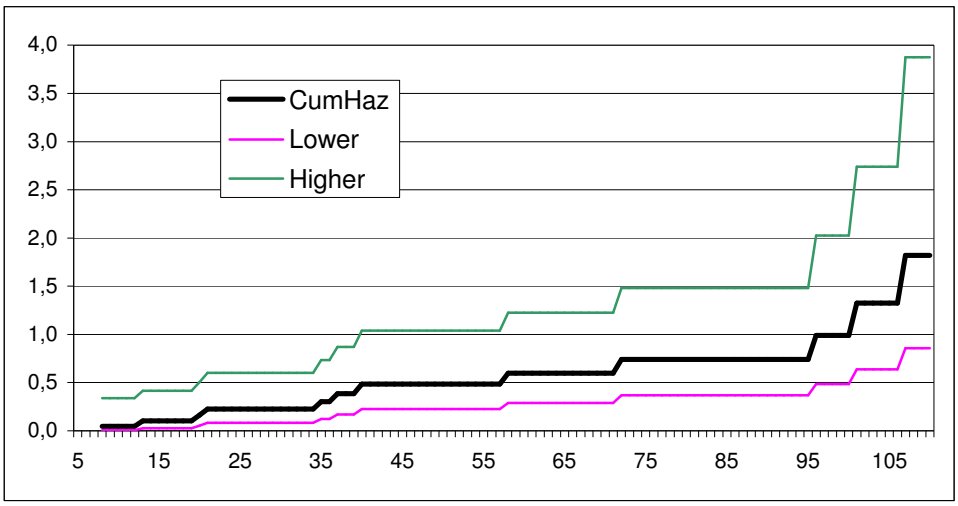
since exp(-z) ≈ 1-z for small z.

The Nelson-Aalen estimator is displayed in *figure 4.3*.

The variance of the cumulative hazard is (Andersen *et al.*, 1993, p. 180):

$$Var\left[\hat{A}(t)\right] = \int_0^t \frac{dN_+(s)}{[Y_+(s)]^2} = \sum_{k:{}^oT_k \prec t} \frac{\Delta N_+({}^oT_k)}{[Y({}^oT_k)]^2}$$

Figure 4.3. Nelson-Aalen estimator of cumulative hazard and 95 percent confidence interval, hypothetical example



It is known as the Aalen variance of the cumulative hazard. In the absence of ties, the numerator of the second term is equal to 1.

The variance estimator can be justified informally by viewing the increments as independent Poisson random variables. Virtually any counting process can be modelled as a Poisson process, at least locally over short time periods (Andersen *et al.*, 1993, p. 55). For small intervals, the increment $dN_+(t) \approx \Delta_h N_+(t) = N_+(t+h) - N_+(t)$, i.e. the number of events in the small interval (t, t+h), is approximately Poisson distributed with mean $E[\Delta_h N_+(t)] = \mu(t)\, Y_+(t)\, h$. The distribution of a Poisson random variable is fully determined by a single parameter. The mean and the variance are equal, $Var[\Delta_h N_+(t)] = E[\Delta_h N_+(t)]$. The variance of the hazard at time t is therefore

$$Var\left[\hat{\mu}(t)\right] = Var\left[\frac{\Delta_h N_+(t)}{Y_+(t)\, h}\right] = \frac{1}{h^2 \left[Y_+(t)\right]^2} \Delta_h N_+(t) = \frac{\hat{\mu}(t)}{h\, Y_+(t)}$$

Since the Poisson increments are assumed to be independent, the variance of the cumulative hazard is the sum of the variances of the hazards. Consider the hypothetical example. The variance associated with the cumulative hazard on day 35 is

$$Var\left[\hat{A}(35)\right] = \sum_{i:{}^oT_i < 35} \frac{\Delta N_+({}^oT_i)}{[Y({}^oT_i)]^2} = \frac{1}{21^2} + \frac{1}{18^2} + \frac{1}{17^2} + \frac{1}{16^2} + \frac{1}{13^2} = 0.0186$$

The standard error is

$$s.e.[\hat{A}(35)] = [0.0186]^{1/2} = 0.1365$$

The variance of the log of the cumulative hazard is (for h = 1)

$$Var\left[\log(\hat{A}(t))\right] \approx \frac{Var\left[\hat{A}(t)\right]}{[\hat{A}(t)]^2}$$

The equation is given by the delta method. It has been shown that the confidence interval for the Nelson-Aalen estimator is closer to the nominal values when the log transformation is used (see Therneau and Grambsch, 2000, p. 12). The log-based 100 * (1-α) percent confidence interval around the estimated cumulative hazard is

$$\hat{A}(t) \exp\left[\pm 1.96 \frac{\left\{Var\left[\hat{A}(t)\right]\right\}^{1/2}}{\hat{A}(t)}\right]$$

Note that this confidence interval is asymmetric, with greater spread above $\hat{A}(t)$ than below, reflecting the long right-hand tail of the distribution of $\hat{A}(t)$. Applying this expression to the cumulative hazard on day 35 is

$$0.3014 \exp\left[\pm 1.96 \frac{0.0186^{1/2}}{0.3014}\right] = 0.3014 \exp[\pm 0.8869] = (0.1241, 0.7323)$$

The confidence interval is very different from that obtained without the log transformation. The untransformed confidence interval is (0.0339, 0.5689). It is known that the untransformed interval is not satisfactory for small sample sizes, which is the case in the hypothetical example (Bie *et al*., 1987 quoted by Andersen *et al*., 1993, p. 208 and Therneau and Grambsch, 2000, p. 12).

A third method is used to estimate the variance of the cumulative hazard. If the total number of processes (e.g. sample size) is given and each process ends in an event or is censored, then the number of events during an interval follows a binomial distribution. Based on the binomial distribution, the Greenwood formula for the variance of the cumulative hazard is (Therneau and Grambsch, 2000, p. 16):

$$Var_G\ \hat{A}(t) = \int_0^t \frac{dN_+(s)}{Y_+(s)\left[Y_+(s) - dN_+(s)\right]}$$

The variance is larger than the Aalen estimate given as the first alternative. The relation between the variance of $\hat{S}(t)$ and the variance of $\hat{A}(t)$ is

$$Var\left[\hat{S}(t)\right] \approx \left[\hat{S}(t)\right]^2 Var\left[\hat{A}(t)\right]$$

In the theory of counting processes, the difference between the observed number of events during the interval [0, t), $N_+(t)$, and the expected number, $\Lambda(t)$, is important. The difference is the martingale of the counting process:

$$M(t) = N(t) - \Lambda(t)$$

Hence the observed counting process may be written as $N(t) = \Lambda(t) + M(t)$. The systematic part of a counting process is the compensator $\Lambda(t)$. The martingale is a pure noise with an expected value of zero. The number of events counted (observed)

in the sample by duration t is generally different from the expected number of events because of sample variation. The difference between the number of events counted during the [0, t)-interval and the expected number is noise due to sample variation. The counting process $N_+(t)$ is the sum of a systematic component, the cumulative intensity process, and a residual, the martingale process. A discussion of the martingale theory is beyond the scope of this report. The interested reader is referred to Andersen *et al*. (1993) for an Fleming and Harrington (1991).

Table 4.4. Variance and standard error of Nelson-Aalen estimator of cumulative hazard

Duration (days)	At risk beginning of day	Event/ censoring	Hazard Rate	Cumulative hazard	Variance hazard	Variance CumHaz	s.e. CumHaz
(1)	(2)	(4)	(3)	(4)	(5)	(6)	(7)
3	22	0	0.0000	0.0000	0.0000	0.0000	
8	21	1	0.0476	0.0476	0.0023	0.0023	0.0476
11	20	0	0.0000	0.0476	0.0000	0.0023	0.0476
11	19	0	0.0000	0.0476	0.0000	0.0023	0.0476
13	18	1	0.0556	0.1032	0.0031	0.0054	0.0732
20	17	1	0.0588	0.1620	0.0035	0.0088	0.0939
21	16	1	0.0625	0.2245	0.0039	0.0127	0.1128
28	15	0	0.0000	0.2245	0.0000	0.0127	0.1128
32	14	0	0.0000	0.2245	0.0000	0.0127	0.1128
35	13	1	0.0769	0.3014	0.0059	0.0186	0.1365
37	12	1	0.0833	0.3848	0.0069	0.0256	0.1599
37	11	0	0.0000	0.3848	0.0000	0.0256	0.1599
40	10	1	0.1000	0.4848	0.0100	0.0356	0.1886
58	9	1	0.1111	0.5959	0.0123	0.0479	0.2189
62	8	0	0.0000	0.5959	0.0000	0.0479	0.2189
72	7	1	0.1429	0.7387	0.0204	0.0683	0.2614
76	6	0	0.0000	0.7387	0.0000	0.0683	0.2614
85	5	0	0.0000	0.7387	0.0000	0.0683	0.2614
96	4	1	0.2500	0.9887	0.0625	0.1308	0.3617
101	3	1	0.3333	1.3221	0.1111	0.2419	0.4919
107	2	1	0.5000	1.8221	0.2500	0.4919	0.7014
112	1	0	0.0000	1.8221	0.0000	0.4919	0.7014

4.2 Process time is discrete time: the life table

In the previous section, we discussed two non-parametric methods for estimating duration dependence of the time to event. Both methods require individual data on elementary processes or episodes and the exact time to the event under study and censoring. The first method, the Kaplan-Meier method, estimates the survival function from data on time to event or censoring. The second, the Nelson-Aalen method, estimates the cumulative hazard from time-dependent data on exposure, where exposure starts at onset of observation or later and ends with the event occurring or the observation ending. In this section, continuous time is replaced by discrete time. Episodes are divided into discrete duration intervals and episodes that end during the same interval are grouped. The intervals do not need to be of the same size. The grouping has implications for the interpretation of the survival function and

the hazard function. When the intervals are small, the survival function tends to the empirical survival function obtained by the Kaplan-Meier method and the cumulative hazard function tends to the cumulative hazard estimated by the Nelson-Aalen method. When the intervals are large, the grouping may lead to substantially different survival functions and hazard functions. The reason is the distribution of events (and exposure time) within the intervals. The grouping of event times and censoring times into discrete time intervals does not introduce any assumption about the distribution of events across intervals. It requires however an assumption about the distribution of event times and censoring times within duration intervals. The need for that assumption is not appreciated in the statistical, demographic and epidemiological literature. It is generally assumed *implicitly* that the distribution is either uniform or exponential. The grouping of event times and censoring times into discrete time intervals results in a non-parametric method known as the life table method or actuarial method.

Two points of departure exist to start the construction of the life table. The venues differ in the definition of the population that may experience the event, i.e. the population at risk. The first approach focuses on the risk set that is the population at risk. It is the population at the beginning of the interval adjusted for the effect of censoring during the interval. Only right censoring is accounted for. This point of departure leads to the empirical probability of an event, i.e. the probability directly estimated from the data. The second approach asserts that the population at risk is a static measure that does not take into account how long people are at risk during the interval and can therefore not handle left truncation or left censoring and cannot easily be extended to more complex models such as multistate models. The second approach concentrates on exposure time or duration at risk rather than on the population at risk at one point of the interval. The two approaches are discussed. The first approach is closely related to the Kaplan-Meier estimator with its emphasis on risk set and *probabilities*. When the intervals are small, the life table survival function tends to the Kaplan-Meier estimator. For that reason, the Kaplan-Meier estimator is also called the product-limit estimator since it is obtained as a limit, when the time is split into intervals and the interval length goes to zero. The second approach is closely related to the Nelson-Aalen estimator with its emphasis on exposure and event *rates*.

Unlike the Kaplan-Meier method and the Nelson-Aalen method, the life-table method distinguishes duration intervals. Let T be a random variable denoting the process time to event. T is a discrete variable and process time is expressed in duration intervals. Suppose that the duration intervals are delineated by n time points t_0, t_1, t_2, … t_n. The first interval starts at t_0 and the last interval ends at t_n. In most cases, $t_0 = 0$ and $t_n = \infty$. The number of intervals is n: (t_0-t_1, t_1-t_2, …, t_{n-1}-t_n). Note that the j-th interval is the interval $[t_{j-1}, t_j)$, which includes t_{j-1} but not t_j. The intervals may be of equal length (equidistant) or of different length. If they are of equal length, an interval following exact duration t may be denoted by [t, t+1). Consider the hypothetical data presented in table 4.1. Following Namboodiri and Suchindran, the hypothetical data are grouped into four duration intervals, 0-29 days, 30-59 days, 60-89 days, and 90 days and over. Hence, $t_0 = 0$, $t_1 = 30$, $t_2 = 60$, $t_3 = 90$ and $t_4 = \infty$. The last interval is open-ended. The grouped data are shown in *table 4.*. The intervals are one month each. Hence, they may be expressed in months (t = 0, 1, 2, and 3+). During the first month, four events occur and four episodes are censored. During the second month three events occur and two episodes are censored. During the fourth month, one event occurs and one

episode is censored. It is the longest episode lasting 112 days. In three cases, the event occurs in the month of interview. They are case 21 (event on May 9 and interview on May 11), case 5 (event on May 17 and interview on May 23) and case 14 (event on May 18 and interview on May 28).

Table 4.5. Grouped data for life-table analysis

Duration		Number of	Events	Censoring	Events+
Days	Months	episodes at start			censoring
0-29	0	22	4	4	0
30-59	1	14	3	2	1
60-89	2	8	1	3	0
90+	3	4	1	1	2
Total			9	10	3

The number of episodes at the beginning of month t is N(t). The number of episodes that end in an event during month t is O(t) and the number that end in right censoring is C(t). It is assumed that all episodes start at the beginning of the first month. The number of episodes at the end of the month t or the beginning of month t+1 is

$$N(t+1) = N(t) - O(t) - C(t)$$

Each interval may be viewed as an observation window. Recall that an observation on an episode is left censored if the episode starts before the beginning of the interval or if it starts during the interval (delayed entry). An observation is right censoring when the observation is terminated because an event occurs that is not related to the event under study (i.e. death or emigration of respondent) or the end of the observation window is reached. In this case the end of the observation window is reached when the end of the interval is reached or when the observation ends during the interval. Although the exact date of censoring may be known, the information is not used when discrete duration intervals are considered (a consequence of grouping of duration data). The information used is whether or not censoring takes place during the interval. In this approach, the exact time of censoring is replaced by the discrete time of censoring (interval). Information that may exist on the exact timing of the event and/or censoring is replaced by an assumption about the timing of the event and/or censoring during the interval. The effect of censoring depends on the assumption about the timing of censoring. Namboodiri and Suchindran (1987, pp. 58ff) consider different ways of handling censoring times in the context of life table analysis. One assumption is that censoring takes place at the beginning of the interval. Another assumption situates censoring at the end of the interval. A third approach assumes censoring in the middle of the interval. These three cases are considered in the next section. Left truncation is not considered.

4.2.1 Empirical probabilities of events

The estimation of event probabilities and other life-table statistics in the presence of censoring is illustrated using the same hypothetical data used before.

The three types of censoring are reviewed below and are displayed in *table 4.6*.

Censoring at the beginning of the month
If censoring takes place at the beginning of the month, episodes that are censored during the month are omitted. Of the 22 episodes, four are censored during the first month of the episode. The *risk set*, i.e. the number of episodes at the beginning of the first month at risk of being terminated during that first month, consists of 18 episodes (22 – 4). The probability of an event is 0.2222, which is the number of events divided by the risk set. During the second month, three episodes are censored. In one case, an event occurs a few days before the interview (ID 21). The event is not counted because it occurs 'after' the assumed date of interview.

Censoring at the end of the month
If censoring takes place at the end of the month, episodes that are censored during the month contribute an entire month to the exposure time or duration at risk. In this case, episode ID 21 is not excluded. It contributes during the second month and it does not contribute to the censored cases in that month.

Censoring in the middle of the month
The assumption of censoring in the middle of the month is equivalent to assuming that half of the censored episodes are censored at the beginning of the month and half are censored at the end of the month. It is also consistent with the assumption of uniform distribution of censoring during the interval. If an event is recorded in the month of interview, the event is considered since the interview necessarily occurs after the event. The risk set at the beginning of the first month is equal to the number of episodes at the beginning of that month minus half the number of episodes censored during that month. It is equal to 22 – 0.5 * 4 = 20. In the second month, episode ID 21 contributes an event and does not contribute to the censored episodes. The event occurs before censoring, not because the data say so (event at day 35, censoring at day 37), but because of the assumption.

Few authors make explicit the assumption about censoring. Most assume that censoring is uniformly distributed over the interval, which is equivalent to censoring in the middle of the interval. In the calculation of survival probabilities using the Kaplan-Meier method, Selvin (1991, p. 289) considers discrete intervals and censoring at the beginning of the interval. 'Individuals withdrawn or lost from follow-up (during the interval) are included in the calculations only when they are known to be at risk for the entire interval. Otherwise, they are excluded from further calculations' (Selvin, 1991, p. 289). Episodes that terminate during an interval, as a result of censoring, are excluded from the risk set and the estimation of the mean lifetime. Therefore individuals who experience an event during an interval are assumed to have experienced the event at the end of the interval.

Table 4.6. Types of censoring determined by time of censoring

Case 1: Censoring at beginning of interval						
Duration	Risk set	Events	Censored	Prob of	Survival function	
Months				event	S	s.e.
0	18	4	4	0.2222	1.0000	0.0000
1	11	3	3	0.2727	0.7778	0.0980
2	5	1	3	0.2000	0.5657	0.1264
3	1	1	3	1.0000	0.4525	0.1431
4					0.0000	0.0000
Case 2: Censoring at end of interval						
Duration	Risk set	Events	Censored	Prob of	Survival function	
Months				event	S	s.e.
0	22	4	4	0.1818	1.0000	0.0000
1	14	4	2	0.2857	0.8182	0.0822
2	8	1	3	0.1250	0.5844	0.1149
3	4	3	1	0.7500	0.5114	0.1216
4					0.1278	0.1148
Case 3: Censoring in middle of month						
Duration	Risk set	Events	Censored	Prob of	Survival function	
Months				event	S	s.e.
0	20	4	4	0.2000	1.0000	0.0000
1	13	4	2	0.3077	0.8000	0.0894
2	6.5	1	3	0.1538	0.5538	0.1197
3	3.5	3	1	0.8571	0.4686	0.1281
4					0.0669	0.0235

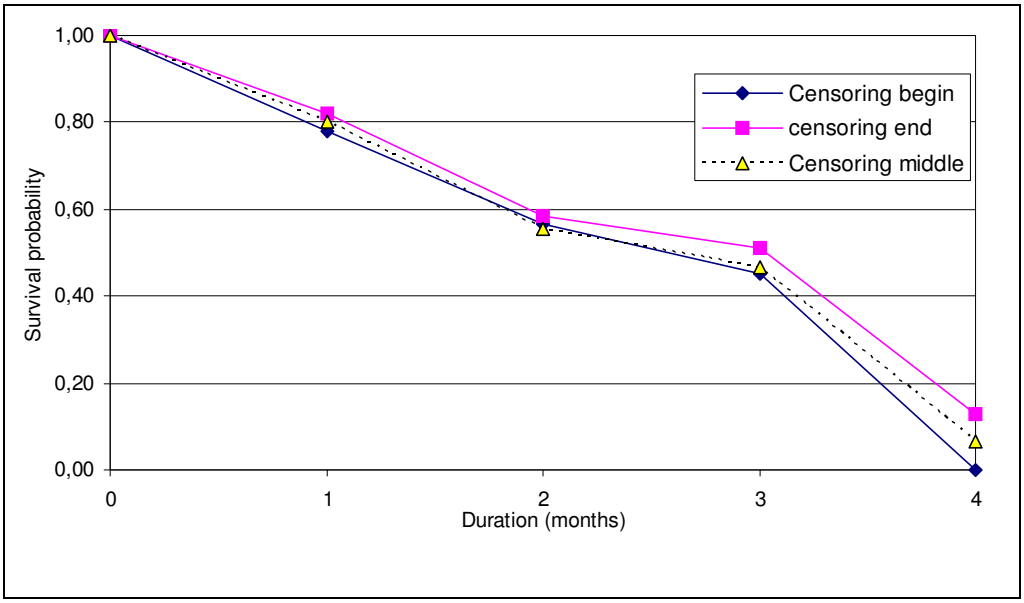
The assumption about censoring has an effect on the survival function, although a modest one. *Figure 4.4* shows the survival functions under the three cases of censoring.

For each interval an event probability is estimated. It is the probability that an episode is terminated during the interval, e.g. that an individual who did not experience the event yet at the beginning of the interval experiences the event during the interval. It is a conditional probability. Consider an interval that starts at exact time t and has a fixed length h. The interval is denoted by [t, t+h). In most applications h is equal to 1. In the hypothetical example, the interval is one month. We define two random variables: a time variable and an indicator variable. The time variable denotes the time at which the observation is terminated as a result of either the event or censoring. It is a continuous variable, denoted by X. The number of episodes that terminate during the interval [t, t+h) is

$$N(t) = \sum_{i=1}^{22} I_i(t)$$

where $I_k(t)$ is an indicator function equal to 1 if $T_k \geq t$ and 0 if $T_k < t$.

Figure 4.4. Survival function for different types of censoring



The number of episodes may also be determined from an indicator variable indicating the status of the process at exact time t, i.e. at the beginning of the interval [t, t+h). Let k denote a process, an episode or a subject, and let ${}_hY_k(t)$ be the indicator variable. The ending time of the observation on the k-th episode, T_k, is equal to a given value t if ${}_hY_k(t) = 1$ and ${}_hY_k(t+h) = 0$. The number of episodes at exact time t is

$$N(t) = \sum_{i=1}^{22} {}_hY_i(t)$$

The number of observations terminating in the interval [t, t+h) is

$$N(t) - N(t+h)$$

The indicator variable denotes the reason for termination of the observation. It is 1 if the reason for termination of the observation on the k-th episode is the event under study. If the observation is censored, it is zero. The indicator variable is denoted by δ_k.

Observations terminate because of the event under study occurs or the observation is censored. If the k-th episode ends in an event during interval [t, t+h), then

$${}_hO_i(t) = [I_i(t) - I_i(t+h)]\,\delta_i = 1$$

In case the observation is censored during the interval [t, t+h), then

$${}_hC_i(t) = [I_i(t) - I_i(t+h)][1 - \delta_i] = 1$$

The number of events during the interval [t, t+h) is

$$ {}_hO_+(t) = \sum_{i=1}^{22} {}_hO_i(t) $$

and the number of censored observations is

$$ {}_hC_+(t) = \sum_{i=1}^{22} {}_hC_i(t) $$

The number of episodes at exact time t+h is

$$ N(t+h) = N(t) - {}_hO(t) - {}_hC(t) $$

where the subscript + is omitted. In demography, the equation is known as the accounting equation or the balancing equation. It expresses the number of processes at a given point in time in terms of the number of processes at a previous point in time, the events during the interval and the observations discontinued during the interval for reasons unrelated to the event under study.

The probability of an event during the interval [t, t+h) is the ratio of the number of events during the interval and the risk set, The risk set is the population at the beginning of the interval adjusted for the effect of censoring during the interval. Only right censoring is accounted for. In addition, it is assumed that the onset of observation coincides with the start of the process being studied and hence the start of the episode. Left censoring is not considered.

The risk set R(t) is the number of processes at the beginning of the interval that are at risk of ending in an event during the interval from t to t+h. It depends on the censoring scheme and may be written as

$$ R(t) = N(t) - a\ {}_hC(t) $$

where a is a constant with a value between 0 and 1. If a is 0, all processes at the beginning of the interval contribute to the risk set for the entire interval. It is equivalent as a censoring at the end of the interval. If it is 1, only processes that are not censored during the interval contribute to the risk set. It is equivalent to censoring at the beginning of the month. The coefficient a is ½ if censoring is assumed to take place in the middle of the month.

The probability that the event occurs during the interval [t, t+h) (event probability) is estimated as the proportion of processes in the risk set that terminates in an event:

$$ {}_h\hat{q}(t) = {}_hO(t) \,/\, R(t) $$

Note that, since the first interval [0, h) start at time t = 0, ${}_h\hat{q}(0)$, is the probability that the event occurs in the first interval.

The probability that no event occurs during the month is

$$ {}_h\hat{p}(t) = 1 - {}_h\hat{q}(t) $$

The event probability is based on a sample of observations on event processes. The larger the sample, the better the estimate, i.e. the closer the sample proportion and the true proportion that is the proportion in the population. If different samples are taken from the same population, the sample proportions will vary. The variation of sample proportions between samples is expressed in terms of the variance of $_h\hat{q}(t)$. The less homogeneous the population is, the smaller the variances. If all members of a population are identical, an observation on a single member is sufficient. The variance also depends on the sample size. For a given population, larger samples lead to smaller variance. To determine the variance, we must specify the random mechanism that underlies the variation in the data. That is done below.

The occurrence of an event during the interval [t, t+h) may be viewed as the outcome of a random process. Consider a process denoted by k that is part of the risk set. During a given interval, the process has two possible outcomes: the event occurs or the event does not occur. Hence, the outcome may be represented by a binary random variable. It takes on a value 1 if the event occurs and a 0 otherwise. The variable is $_hO_k(t)$ shown above. The variable is a Bernoulli random variable, which has two possible outcomes. With each outcome may be associated a probability. The probability distribution, which consists of two probabilities only, is the Bernoulli distribution. The distribution is fully characterized by a single parameter $q_k(t)$ since $p_k(t) = 1 - q_k(t)$. The parameter is the expected value of $_hO_k(t)$.

$$_hq_i(t) = E[_hO_i(t)] = 1 * \Pr\{_hO_i(t) = 1\} + 0 * \Pr\{_hO_i(t) = 0\} = \Pr\{_hO_i(t) = 1\}$$

The variance of the Bernoulli distribution is $q_k(t)[1-q_k(t)]$. In case of n identical and independent processes, the expected outcome of a process is the mean outcome and the variance of the mean outcome is $_hq(t)[1-_hq(t)]/n$.

A Bernoulli variable can be specified for every interval [t, t+h) with t= 0, h, 2h, 3h, … A sequence of Bernoulli random variables for the successive intervals is a stochastic process, known as a Bernoulli *process* (see e.g. Çinlar, 1975).

The number of events during the interval [t, t+h) that occur among the R(t) event processes is a random variable too since its value cannot be predicted in advance. Each event process may be viewed as a trial that leads to one of two possible outcomes. Such a trial is known as a Bernoulli *trial*. The flip of a coin is a Bernoulli trial too. It has two possible outcomes (head and tail) and with each outcome may be associated a probability distribution. If in each interval, the event processes are independent, then they may be represented as independent Bernoulli trials. It is also assumed that the processes are identical. The number of events that occur among the R(t) identical and independent event processes or Bernoulli trials is $_hO(t)=_hq(t)\,R(t)$. The random variable $_hO(t)$ is known as a Binomial *random variable* and R(t) as the index. It is a discrete random variable following a Binomial distribution with parameter equal to the expected number of events during the interval [t, t+h), i.e. $E[_hO(t)]$. The sample variance is

$$Var\left[_hO(t)\right] = Var\left[_hq(t)\,R(t)\right] = [R(t)]^2\,Var\left[_hq(t)\right] = _hq(t)\left[1-_hq(t)\right]R(t)$$

Note that the sample variance of ${}_hq(t)$ is (see also Chiang, 1984, p. 79)

$$Var\left[{}_h q(t)\right] = Var\left[\frac{{}_h O(t)}{R(t)}\right] = \frac{1}{[R(t)]^2} Var\left[{}_h O(t)\right] = \frac{{}_h q(t)\,[1-{}_h q(t)]}{R(t)}$$

The probability ${}_hq(t)$ is the parameter of a binomial distribution. The parameterization corresponds to the conditioning principle discussed above. Conditional on the risk set (number of episodes at the beginning of the interval), the number of events follows a binomial distribution with parameter ${}_hq(t)$. The parameter applies to all processes at risk of termination in the event under study.

The probability that an episode is not terminated by an event at exact duration t is the survival function

$$S(t) = p(t-1)\ldots p(0) = p(t-1)\,S(t-1)$$

with p(t) = 1- S(t) is the probability that the process reaches t without experiencing the relevant event. The survival probability may also be expressed as

$$S(t) = \prod_{s=0}^{t-1}\left[1 - q(s)\right] = \prod_{s=0}^{t-1}\left[1 - \frac{O(s)}{R(s)}\right]$$

Compare this expression with the Kaplan-Meier estimator of the survival function. The life-table estimator tends to the Kaplan-Meier estimator when the intervals become smaller and smaller, i.e. when the length of the interval reduces to an infinitesimally small number denoted by dt. If the intervals are sufficiently small there will be many intervals without events and there will be no more than one event during an interval. The event that occurs during the interval [t, t+dt) is said to occur at t. This presentation of the relation between the life table and the Kaplan-Meier estimator also shows that the risk set in the Kaplan-Meier estimator includes all episodes immediately before the event occurs. Although we assume that no two observations occur at the same time point, in practice there may be several observations at the same time point. They are referred to as ties. The presence of ties does not affect the expression for the survival function.

Assuming that in the successive intervals the event processes are independent, the sample variance of S(t) is given by the Greenwood formula:

$$Var[S(t)] = \left[S(t)\right]^2 \sum_{s=0}^{t-1} \frac{q(s)}{p(s)R(s)}$$

Consider the survival function in table 4.6 case 3. The probability of surviving until the 3rd month is 46.9 percent. The variance of that survival probability is

$$Var[S(3)] = \left[0.4686\right]^2\left[\frac{0.2000}{(1-0.2000)*20} + \frac{0.3077}{(1-0.3077)*13} + \frac{0.1538}{(1-0.1538)*6.5}\right] = 0.1281$$

The *density* of events during the interval is the ratio of the number of events and the length of the interval. It is the number of events per unit interval. The density is denoted by f(t).

The event rate is the ratio of the number of transitions and the duration of exposure to the risk of the event. The occurrence-exposure rate for the interval [t, t+1) is

$$_hm(t)=_hO(t)/_hL_t(t)$$

where $_hL_0(t)$ is the exposure or duration at risk during the interval [t, t+h) by a process or episode at time t. The exposure is derived from the survival function. It is equal to the area under the survival function. The duration at risk consists of two components. The first relates to the episodes that are not terminated during the interval. Each episode has an exposure time equal to the length of the interval. The second relates to episodes that are terminated during the interval, as a result of either the event under study or censoring. If the timing of the events and censoring is known, the exposure can be calculated precisely. It is however assumed that the timing of events and censoring is not known except for the interval in which they occur. It is generally *assumed* that the events and censoring are uniformly distributed during the interval. The assumption is equivalent to occurrence in the middle of the interval. As a result, episodes that end during the interval, end in the middle of the interval. The total exposure time by episodes at the beginning of the interval is

$$_hL_t(t) = p(t) + \tfrac{h}{2}[1 - p(t)] = \tfrac{h}{2}[1 + p(t)]$$

The duration at risk in the interval [t, t+h) for a process that starts at 0 is

$$_hL_0(t) = S(t)_hL_t(t) = \tfrac{h}{2}[S(t) + S(t+h)]$$

The estimate assumes that events are uniformly distributed during the interval.

If $_hL_0(t)$ is approximately equal to R(t)-½O(t), then the rate may be expressed as

$$_hm(x) = \frac{R(t)\ _hq(t)}{R(t) - \frac{h}{2}R(t)_hq(t)} = \frac{_hq(t)}{1 - \frac{h}{2}\ _hq(t)}$$

The sample variance of the event rate is (assuming an interval of one year and uniform distribution of events) (Blossfeld and Rohwer, 2002, p. 50):

$$\mathrm{var}\left[_hm(t)\right] = \frac{[_hm(t)]^2}{_hq(t)R(t)}\left[1 - \left[\tfrac{h}{2}\ _hm(t)\right]^2\right]$$

The time spent beyond exact age t is

$$T(t) = \sum_{s=t}^{z} {}_hL_0(s)$$

where z is the highest duration interval. The expected time spent beyond t conditional on survival to x is the life expectancy

$$e(t) = T(t) / S(t)$$

The expected time is known as expected duration, expected duration of stay, expected dwelling time or sojourn time, mean lifetime, and waiting time to event.

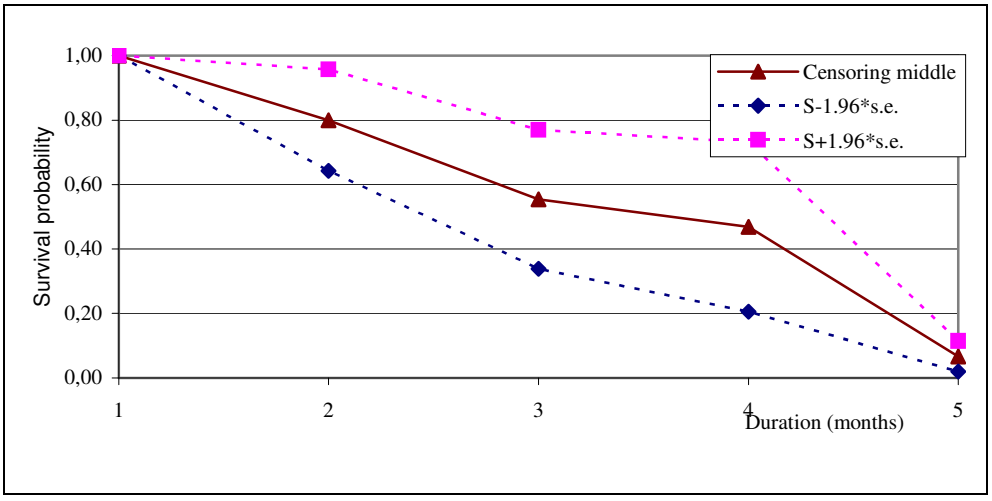
The sample variance of the life expectancy at duration t is (Chiang, 1968; Chiang, 1984, p. 163 and Biswas, 1988, p. 149), assuming uniform distribution of events during the interval, is

$$Var[e(t)] = \sum_{s=t}^{z-1}\left[\frac{S(s)}{S(t)}\right]^2 \left[\tfrac{1}{2} + e(s+1)\right]^2 Var[p(s)]$$

where $Var[p(s)] = \dfrac{p(s)[1-p(s)]}{R(s)}$

Chiang (1984, p. 163) gives the expression for the case the events are not uniformly distributed. *Figure 4.5* displays the survival function and the 95 percent confidence intervals.

Figure 4.5. Survival function and 95 percent confidence interval



In large samples it may be assumed that the value of the survival function and the event rate are approximately normally distributed. It is then possible to calculate confidence intervals.

The life table is shown in *table 4.7*. The mean duration of episode (mean lifetime) is estimated in two different ways. The first approach is based on the durations at which episodes terminate and the second approach is based on exposure times or dwelling times. The survival function shown in table 4.7 is based on the assumption that censoring and events occur in the middle of the interval. The estimation of the mean

lifetime should be based on the same assumption. Twenty percent of the episodes terminate during the first month, on average in the middle of the first month. Twenty-four percent of the episodes terminate during the second month, on average at a process time of 1.5 months. Seven percent terminates during the fourth month, on average at a process time of 4.5 months. The expected process time or duration of an episode is

$$e(0) = \sum_{s=0}^{4} \left[S(s+1) - S(y)\right](s+0.5)$$
$$= 0.200 * 0.5 + 0.246 * 1.5 + 0.085 * 2.5 + 0.402 * 3.5 + 0.067 * 4.5 = 2.39$$

Table 4.7. Life table for hypothetical data

A. Expected duration based on time to event or censoring

Duration in months	Prob of event	Survival function		
x	q(x)	S(x)	d(x)	(x+0.5)d(x)
0	0.2000	1	0.2000	0.100
1	0.3077	0.8	0.2462	0.369
2	0.1538	0.5538	0.0852	0.213
3	0.8571	0.4686	0.4017	1.406
4		0.0669	0.0669	0.301
5		0.0000		
Mean				2.389

B. Expected duration based on exposure time

Duration in months	Exposure	Event rate	Time beyond x	Expected time to event
x				
0	0.9000	0.2222	2.36	2.36
1	0.6769	0.3636	1.46	1.82
2	0.5112	0.1667	0.78	1.41
3	0.2678	1.5000	0.27	0.57
4				

The second approach is based on the duration at risk. The duration at risk is obtained assuming uniform distribution of events and that the process is truncated after four months. The expected waiting time to the event or to reaching the 5th month (month four), whatever comes first, is 2.36 months, which is approximately 70.8 days. The result is a few days larger than the 68.6 days obtained using the Kaplan-Meier method. The difference is due to the grouping of the individual data and the omission of the information on the precise timing of events and censoring during the duration intervals.

4.2.2 Empirical transition rates

The second point of departure to construct the life table is the transition rates rather than probabilities. The method that is based on probabilities has two serious shortcomings. First, the duration of exposure is approximated even when precise data are available. Second, left censoring cannot be accounted for. A method that focuses on duration of exposure rather than risk set does not have these shortcomings. Such a

method is based on empirical occurrence-exposure rates. The method moves beyond the risk set, which is an estimate of the *population at risk at the beginning* of the discrete time interval, and concentrates on *duration at risk during* the interval. An important advantage of the approach is that open episodes and closed episodes are treated similarly. Another advantage is that observation may start at any duration of the process. During the interval, subjects may move in and out of observation. There may be several periods of exposure and several events. A third advantage is that the method can easily be extended to multistate models. The key issue is to count the number of years, months, weeks or days subject are under observation and to count the number of events that occur while under observation. Processes are exposed only when they are under observation and in the origin state. The estimation of exposure time therefore reduces to the estimation of the duration a process is in the origin state and under observation. Processes may start before the onset of observation and end after observation is terminated. We assume that the dates of the events and censoring are given. The count of events and exposure time during the interval provides the information for the transition rate. The transition rate is the ratio of the number of occurrences (events) and the duration of exposure, hence the term occurrence-exposure rate. The estimation of occurrence-exposure rates may be situated within the theory of counting processes. In this section, we consider the estimation of transition rates as an extension of the Nelson-Aalen estimator to discrete duration intervals.

The method is nonparametric. The duration dependence of transition rates is not defined by a model, such as the exponential model or the Gompertz model. Transition rates are estimated for each duration interval instead. The rates are based on information on events and exposure. The required information on the events is the number of transitions that occur in the sample population during each duration interval. The required information on exposure is the total duration during which subjects are under observation and exposed to the risk of experiencing the event.

The derivation of the Nelson-Aalen estimator started by defining two random variables; one indicates whether a process is under observation at time t, and the other counts the number of events between the start of the observation and t. As before, let t denote the process time, i.e. the time since the onset of the process. Time is a continuous variable but it is divided in discrete intervals. The intervals do not have to be of equal length. For presentation purposes, an equal length of h is assumed. The discrete interval is from t to t+h and is denoted by [t, t+h). The interval is denoted by the time at the beginning of the interval. $Y_k(t)$ represents a time-varying indicator variable that takes on the value 1 if the process is under observation at time t and the value 0 otherwise. Since t is continuous, the process may enter observation at any time, may leave observation and enter again later. The second random variable is $N_k(t)$. It is a counting process; it denotes the number of observed events generated by process k during the period from 0 to t.

Suppose you want to find out how often your favourite song is played by your favourite radio music station. Continuous listening for 24 hours per day is of course not practical. The period of observation in a day may be represented by a sequence of the indicator variable $\{Y(t), t = 0, 1, 2, \dots, 23\}$, where the first hour of the day is denoted by 0 (0 hours completed since the start of the day). Suppose you listen to the radio from 6 a.m. to 10 a.m., from noon to 4 p.m. and from 7 p.m. to 11 p.m. The sequence $\{Y(t)\}$ is

000000111100111100011110

The duration of observation is 12 hours: ${}_EY(24) = \sum_{s=0}^{23} Y(s) = 12$

Suppose the song was aired three times during the hours you were listening: at 7.30 a.m., 7.15 p.m. and 10.45 p.m. Events may be represented by sequence N(t)

000000011111111111122233

The average event rate or the 'arrival rate' of your favourite song at time t is $N(t)/{}_EY(t)$. At noon, it is 1/4 =0.25 occurrences per hour (of observation) and at the end of the day it is the same, 3/12 = 0.25. The event rate is zero for every hour except for hours 7 a.m., 7 p.m. and 10 p.m. when the song is aired. At these hours the event rate is $\mu(t) = \frac{N(t+1) - N(t)}{Y(t)} = \frac{1}{1} = 1$. If the event rate would be constant throughout the day, and the producer does not have a memory, i.e. he does not remember which songs he already aired during that day, you may expect two occurrences by four p.m. instead of the single occurrence you observed. The difference between the observed occurrences (one song) and the expected occurrences (two songs) is the martingale of the counting process. If your favourite song is aired less than you expect and another song is aired more than you expect, you are implicitly applying the theory of counting processes and the concept of a martingale.

In this section, we discuss the estimation of rates that are referred to as event rates, transition rates, hazard rates and occurrence-exposure rates. We use the term transition rate to indicate the rate at which an event occurs and a transition is made from the origin state to the destination state. The estimator is derived using the hypothetical data in table 4.1 and table 4.4.

Consider *table 4.7* This table indicates the type of episode (closed or open) and the duration at risk of experiencing the event. The table gives the information for discrete duration intervals (months). The number of occurrences is the same as in case 2 of table 4.6 (Censoring at end of interval). It considers intervals of length h (from t to t+h).

Table 4.7. Occurrences, exposures, and transition rates

Duration		Count		Exposure				Rate
	At	Events	Censored	By those leaving		By survivors		(Monthly)
Months	start			Days	Months	Days	Months	
0	22	4	4	106	3.53	420	14	0.2281
1	14	4	2	239	7.97	240	8	0.2505
2	8	1	3	295	9.83	120	4	0.0723
3	4	3	1	416	13.87	0	0	0.2163
		12	10	1056	35.20	780	26	0.1961

To calculate the duration at risk during a given interval, a distinction is made between persons who leave observation during the month (because of an event or censoring) and persons who survive the interval. The duration at risk by those who experience the event or are censored is based on data on dates of events and censoring, with the date in days. No assumption is made about the duration dependence of events and censoring. The occurrence-exposure rate is the ratio of the number of events and the total exposure time. Let ${}_hO(t)$ denote the number of events during duration interval [t, t+h) where h is the length of the interval. The duration of exposure during the interval [t, t+h) is denoted by ${}_hPM(t)$ and is measured in person-days. In general, exposure is expressed in person-days, person-months or person-years. The transition rate during the [t, t+h)-interval is

$$ {}_hm(t) = \frac{{}_hO(t)}{{}_hPM(t)} $$

In our example, the transition rates are monthly rates. For the last duration category, the rate is very different from that estimated under the assumption that censoring and the events are uniformly distributed. If it is assumed that events and censoring occur in the middle of the last month, the duration at risk is 3.5 months

The occurrence of an event during the interval [t, t+h) may be viewed as the outcome of a random process and the outcome of the process may be represented by a random variable with a characteristic distribution. Since the number of processes at the beginning of the interval is not considered as processes may enter and leave observation several times during the interval, and since each process may generate more than one event during the interval, the Binomial distribution is not applicable. The random variable measuring the outcome of the process is a variable that indicates the number of events per unit period of observation (hour, day, week, month, year). The random variable is a Poisson random variable and the process generating the events is known as the Poisson process (see. e.g. Çinlar, 1975). The parameter of the Poisson process is the expected number of events during a unit duration interval, $E[{}_hO(t)]$, and the variance $Var[{}_hO(t)]$ is equal to the expected value. The expected value is

$$ {}_h\lambda(t) = E[{}_hO(t)] $$

The expected transition rate or event rate during the interval [t, t+h) is

$$ {}_hm(t) = \frac{{}_h\lambda(t)}{{}_hPM(t)} = \frac{E[{}_hO(t)]}{{}_hPM(t)} $$

The above expression assumes that the duration of exposure is not affected by a change in the number of events. That assumption is generally made and is realistic when the number of events is small relative to the total duration at risk by all subjects under observation.

Assuming that the number of events during a unit interval is a Poisson random variable, the variance of the transition rate is

$$Var\left[{}_h\hat{m}(t)\right] = Var\left[\frac{{}_hO(t)}{{}_hPM(t)}\right] = \frac{Var[{}_hO(t)]}{\left[{}_hPM(t)\right]^2} = \frac{{}_h\lambda(t)}{\left[{}_hPM(t)\right]^2} = \frac{{}_hm(t)}{{}_hPM(t)}$$

(for some discussion, see Manton and Stallard, 1988, p. 66; Courgeau and Lelievre, 1992, p. 60).

The duration-specific rates are then applied to a hypothetical process to determine how it evolves while generating events. In this illustration, the hypothetical process is a hypothetical individual experiencing an event. For each duration interval, the rates determine the number of events that occur. *Table 4.8* shows the life table.

Table 4.8. Life table based on occurrence-exposure rates

Duration		Rate (Monthly)	Probability of event in month		Survival function	Expected sojourn	Expected waiting time
Days	Months		Linear	Expon	Expon	time	to event
0-29	0	0.2281	0.20478	0.20398	1.000	0.8980	4.87
30-59	1	0.2505	0.22263	0.22161	0.796	0.7078	4.99
60-89	2	0.0723	0.06977	0.06974	0.620	0.5980	5.27
90+	3	0.2163	0.19523	0.19454	0.576	2.6643	4.62

The *probability* of an event during the interval [t, t+h), provided the process is under observation at the beginning of the interval, is

$${}_hq(t) = 1 - \exp\left[-h\ {}_hm(t)\right]$$

where exp[-h ${}_h$m(t)] is the survival function for the interval [t, t+h) and h = 1. In this hypothetical example, a process may generate at most one event. If multiple events may be generated by the process, then ${}_h$q(t) is the probability of *at least* one event during the interval.

The probability of an event during the first months is

$${}_1q(0) = 1 - \exp\left[-0.2281\right] = 0.20398$$, i.e., 20.4 percent.

The exponential expression for ${}_h$q(t) may be approximated by a linear expression:

$${}_hq(t) = \frac{h\ {}_hm(t)}{1 + \frac{h}{2}\ {}_hm(t)}$$

where ${}_h$m(t) is the transition rate during the interval [t, t+h). The linear expression is often used in demography and actuarial sciences. For example, the probability that an event occurs during the first month is

$${}_1q(0) = \frac{0.2281}{1 + \frac{1}{2}0.2281} = 0.20478$$, i.e. 20.5 percent.

The two estimates are close when the interval is short and/or the rates are small.

To determine the expected exposure time during the interval [t, t+h), note that ${}_hq(t)$ is the probability of an event provided the process is not interrupted by an event other than the event under study. The presence of censoring has been accounted for in the estimation of the transition rate, but when the rate is applied to determine the *expected* number of events and the probability of an event, the entire interval [t, t+h) is considered. The expected number of events during the interval may be estimated in two ways. The first uses the event probability and the other uses the event rate. The expected number of events is

$$ {}_h\lambda(t) = E[{}_hO(t)] = {}_hq(t)Q = {}_hm(t)\ {}_hL(t)\,Q $$

where Q is the number of processes at the beginning of the interval, i.e. at exact time t and ${}_hL(t)$ is the expected exposure time during the interval [t, t+h). The expected exposure time is also referred to as the waiting time to the event, and sojourn time or dwelling time in the origin state. The value of ${}_hL(t)$ can be estimated from ${}_hq(t)$ and ${}_hm(t)$:

$$ {}_hL(t) = \frac{{}_hq(t)}{{}_hm(t)} = \left[{}_hm(t)\right]^{-1}\left[1 - \exp[-h_hm(t)]\right] $$

During the interval [t, t+h), a process at the beginning of the interval may *expect* to be exposed to the event for ${}_hL(t)$ units of time, provided the event rate is constant during the interval and no competing events may occur. In other words, the waiting time to the event is ${}_hL(t)$. This is an unconditional measure. The waiting time to the event, provided that the event occurs, is different. The presence of competing events and censoring in the data has been accounted for in the estimation of the transition rate ${}_hm(t)$. The value of ${}_hL(t)$ may also be obtained in a different way. The waiting to an event during the interval [t, t+h) by a process at the beginning of the interval is the area under the survival function:

$$ {}_hL(t) = \frac{1}{S(t)}\int_0^h S(t+\tau)\,d\tau $$

which is equal to

$$ {}_hL(t) = \frac{1}{S(t)}\int_0^h \exp[-{}_hm(t)*(t+\tau)]\,d\tau = \int_0^h \exp[-{}_hm(t)*\tau]\,d\tau $$

$$ = \left| -\frac{1}{{}_hm(t)}\exp[-{}_hm(t)*\tau] \right|_0^h = \frac{1-\exp[-h\ {}_hm(t)]}{{}_hm(t)} $$

The waiting time to the event is the exposure irrespective whether or not the event occurs during the interval. If the event does not occur the process remains at risk for the entire interval, i.e. for h time units. The probability that a process remains at risk for the entire interval is $1 - {}_hq(t)$. If, during the interval, the event occurs, then the exposure time is shorter. Let ${}_ha(t)$ denote the *fraction of the interval*, the process is

exposed in case *the event occurs*. The conditional measure may be estimated from the unconditional waiting time to the event:

$$ {}_hL(t) = h\,[1-{}_hq(t)]+{}_ha(t)\,h\;{}_hq(t) = \frac{{}_hq(t)}{{}_hm(t)} $$

which gives

$$ {}_ha(t) = 1 - \frac{1}{{}_hq(t)} + \frac{1}{h\;{}_hm(t)} $$

Hence ${}_h$a(t) is determined by both the conditional event probability ${}_h$q(t) and the event rate ${}_h$m(t). The time to event for a process that ends in an event during the interval [t, t+h) is h ${}_h$a(t). The measure ${}_h$a(t) is generally known as Chiang's "a". In the empirical studies of mortality, Chiang (1960) found that ${}_h$a(t) is more or less invariant with respect to sex, race, geographic location, other demographic variables, and cause of death. He suggested that, if ${}_h$a(t) is determined for each age group in one population, it could be used for many populations. Chiang called ${}_h$a(t) the *fraction of the last year of life* (h = 1) (Chiang, 1968, pp 190ff; 1984, pp. 142ff).

By way of example, consider the hypothetical data and let us estimate the waiting time to event during the first month. The waiting time is

$$ {}_1L(0) = \frac{{}_1q(0)}{{}_1m(0)} = \frac{0.20398}{0.2281} = 0.8943 $$

The waiting time is composed of two parts. The first is the contribution of processes or episodes that do not end during the interval. The contribution is one month. The second is the contribution of processes that end during the interval. The expected faction of the interval exposed by processes that end during the interval is

$$ {}_1a(0) = 1 - \frac{1}{{}_1q(0)} + \frac{1}{1\;{}_1m(0)} = 1 - \frac{1}{0.20398} + \frac{1}{0.2281} = 0.4816 $$

The expected waiting time to the event, provided the event occurs during the interval [0, 1) is a little less than half a month. That could be expected given the exponentially declining survival function that is associated with a constant event rate.

${}_h$L(t) may be approximated assuming uniform distribution of events during the interval [t, t+h). If events are uniformly distributed the average waiting time to the event, provided it occurs during the interval, is half the interval. Hence ${}_h$a(t) = 0.5. The waiting time to an event for a process present at the beginning of the interval is

$$ {}_hL(t) = \tfrac{h}{2}\left[1+{}_hp(t)\right] = h\left[1-\tfrac{1}{2}{}_hq(t)\right] $$

Returning to the hypothetical data, the estimate of ${}_1$L(0) under the linear model is

$${}_1L(0) = 1 - \tfrac{1}{2}\,0.20478 = 0.8976$$

which is a little larger than under the exponential model.

In the previous expressions, the expected sojourn time during the interval [t, t+h) was expressed for processes that are present at the beginning of the interval. Processes that terminated before the beginning of the interval are excluded. The waiting time may also be expressed relative to the event origin, i.e. the onset of the process under study. That waiting time is

$${}^0_hL(t) = S(t)\ {}_hL(t)\,.$$

In the linear model, it may be written as

$${}^0_hL(t) = \tfrac{h}{2}\left[S(t) + S(t+h)\right]$$

This expression is commonly used by demographers in the construction of mortality tables (see e.g. Preston *et al.*, 2001, p. 43). Empirical research on mortality, such as that by Chiang, revealed that the assumption of uniform distribution of deaths is often not met. For instance, many infant deaths, i.e. deaths during the first year of life, occur in the neonatal period. Hence ${}_1a(0)$ is much smaller than 0.5. The value is not constant across populations but varies with the infant mortality rate. Generally, a reduction of the deaths during the first year of life first occurs in the post-neonatal period. Hence, the smaller the infant mortality rate, the smaller the value of ${}_1a(0)$. For convenience, the World Health Organization has suggested values of ${}_1a(0)$ for different levels of infant mortality. The values are given by Chiang (1984, p. 144). Other values are given by other authors (see e.g. Keyfitz and Flieger, 1990 and Preston *et al.*, 2001, p. 46 and p. 48). Preston *et al*. (2001, p. 45) assert that a traditional method of mortality table construction due to Reed and Merrell (1939) amounts to borrowing ${}_ha(t)$ values from the US mortality tables on which their statistical relation was based (24 US mortality tables covering the years 1900 to 1930; for details, see Namboodiri and Suchindran, 1987, p. 23). Several strategies exist for choosing a set of ${}_ha(t)$ values (for an overview, see Preston *et al.*, 2001, pp. 44ff); When the assumption of uniform distribution of events is not warranted, the waiting time to the event for a process at the beginning of the interval is

$${}_hL(t) = {}_ha(t)\ {}_hq(t) + [1 - {}_hq(t)] = 1 - [1 - {}_ha(t)]\ {}_hq(t)$$

The waiting time to the event for a process at the event origin is

$${}^0_hL(t) = S(t)\ {}_hL(t)$$

which may be written as

$${}^0_hL(t) = {}_ha(t)\,S(t) + [1 - {}_ha(t)]\,S(t+h)$$

The latter expression is common in demography.

The expected waiting time to the event by a process currently at process time t is the *expected lifetime* or *life expectancy* at t

$$e(t) = \sum_{s=t}^{z} {}_{h}^{t}L(s) = \frac{1}{S(t)} \sum_{s=t}^{z} {}_{h}^{0}L(s)$$

where z is the highest, open-ended duration category.

It is assumed that events may occur during the last open-ended duration category, i.e. in the 4th month and beyond. More than half of the episodes reach the 4th month (57.6 percent). The episodes that reach the open-ended interval end at a constant rate of 0.2163. The waiting time to an event, conditional on reaching the open-ended interval, is therefore 1/0.2163 = 4.623 months. The expected duration of the process beyond the 4th month, evaluated at the onset of the process, is the probability of reaching the 4th month times the waiting time to the event provided the 4th month is reached. It is 0.576 * 4.623 = 2.664 months. At the onset of the process, one may expect the process to spend 2.7 months during the 4th month and beyond.

The sample variance of the lifetime is

$$Var\left[e(t)\right] = \sum_{s=t}^{z-1} \left[\frac{S(s)}{S(t)}\right]^2 \left[h\,(1-{}_{h}a(s)) + e(s+1)\right]^2 Var\left[p(s)\right]$$

where $Var\left[{}_{h}p(s)\right] = \exp[-2h_{h}m(s)]\,h^2\,Var[{}_{h}m(s)]$

The latter expression is derived using the delta method, in which

$$Var\left[\exp(z)\right] \cong \exp(2z)\,Var\left[z\right]$$

Note that the sample variance of ${}_{h}p(s)$ differs from the sample variance derived earlier in the context of the empirical event probability:

$$Var\left[{}_{h}p(s)\right] = \frac{{}_{h}p(s)[1-{}_{h}p(s)]}{R(s)}$$

with R(s) the risk set.

References

Aalen, O.O. (1975) Statistical inference for a family of counting processes. Ph.D. dissertation, Dept. Statistics, Univ. California, Berkeley.

Aalen, O.O. (1978) Nonparametric inference for a family of counting processes. Annals of Statistics, 6:701 -726.

Andersen, P.K. and N. Keiding (1996), Survival analysis. In: P. Armitage and H.A. David (eds.). Advances in biometry. Wiley, New York, pp. 177-199.

Andersen, P.J., O. Borgan, R.D. Gill and N. Keiding (1993), Statistical models based on counting processes. Springer Verlag, New York.

Bækgaard, H. (200) Micro-macro linkage and the alignment of transition processes. Some issues, techniques and examples. Technical Paper no. 25, NATSEM, University of Canberra.

Ben-Shlomo, Y. and D.L. Kuh (2002), A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31, pp. 285-293 (ed.)

Bie, O., Ø. Borgan and K. Liestøl (1987), Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, 14, pp. 221-233.

Biswas, S. (1988), Stochastic processes in demography and applications. Wiley, New Delhi.

Blossfeld, H.P and G. Rohwer (1995), Techniques of event history modeling. New approaches to causal analysis. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Blossfeld, H.P. and G. Rohwer (2002) Techniques of event history modeling. New approaches to causal analysis. Lawrence Erlbaum, Mawah, New Jersey. Second Edition.

Brass, W. (1974) Perspectives in population prediction: Illustrated by the statistics of England and Wales. Journal of the Royal Statistical Society, 137, series A: 532-583.

Çinlar, E. (1975) Introduction to stochastic processes. Prentice-Hall, Englewood Cliffs, New Jersey

Chiang, C.L. (1964), A stochastic model of competing risks of illness and competing risks of death. In: J. Gurland (ed.). Stochastic models in medicine and biology. University of Wisconsin Press, Madison.

Chiang. C.L. (1968), Introduction to stochastic processes in biostatistics. Wiley, New York. Chapter 9 reprinted in D.J. Bogue, E.E. Arriage and E.L. Anderton (eds.) (1993) Readings in population research methodology. Vol 2, pp. 7.84-7.97.

Chiang, C.J. (1984), The life table and its applications. Robert E. Krieger Publishing Co., Malabar, Florida.

Courgeau, D. and E. Lelièvre (1992), Event history analysis in demography. Clarendon Press, Oxford.

Diggle, P.J., K.-Y. Liang and S.L. Zeger (1995), Analysis of longitudinal data. Clarendon Press, Oxford.

Giele, J.Z. and G.H. Elder Jr. (1998), Life course research: development of a field. In: J.Z. Giele and G.H. Elder Jr. (eds.) (1998), Methods of life course research.Qualitative and quantitative approaches. Sage Publications, Thousand Oaks, Ca., pp. 5-27.

Hosmer, D.W. and S. Lemeshow (1999), Applied survival analysis. Regression modelling of time to event data. Wiley, New York.

Hougaard, P. (2000), Analysis of multivariate survival data. Springer Verlag, New York.
Joshi, H. (2001), Longitudinal data as an aid to the policy maker: some British examples. Paper presented at the Annual Conference of the Economists, Perth, Australia.
Kalbfleisch, J. and R. Prentice (1980), The statistical analysis of failure time data. Wiley, New York.
Kaplan, E. L., and Paul Meier (1958) Nonparametric Estimation from IncompleteObservations. Journal of the American Statistical Association 53, 282:457-81.
Keyfitz, N. (1968) Introduction to the mathematics of population. Addison-Wesley, Reading, Massachusetts.
Keyfitz and Flieger (1990)
M. Khatun and F.J. Willekens (2001) The life history calendar. Technical aspects of data analysis using contraceptive calendar of the Bangladesh Demographic and Health Survey. Working Paper 01-2, October 2001 (80 pp.)
Klein, , J.P. and M.L. Moeschberger (1997), Survival analysis. Techniques for censored and truncated data. Springer Verlag, New York.
Lindsey, J.K. (1999), Models for repeated measurements. Second Edition. Oxford University Press, Oxford.
Mamun, A.A. (2001), Multistate models in public health. Review and application to the Framingham Heart Study. Master Thesis Series 01-3, Population Research Centre, University of Groningen.
Manton, K.G. and E. Stallard (1988), Chronic disease modeling: measurement and evaluation of the risks of chronic disease processes. Charles Griffin, London, and Oxford University Press, New York.
Marini, M.M. and B. Singer (1988), Causality in social sciences. *Sociological Methodology*, 18, pp. 347-409.
Mills, M. (2000), The transformation of partnerships. Canada, the Netherlands and the Russian Federation in the age of modernity. Thela-Thesis, Amsterdam.
Namboodiri, K. and C.M. Suchindran (1987), Life table techniques and their applications. Academic Press, Orlando.
Preston, S.H., P. Heuveline and M. Guillot (2001), Demography. Measuring and modeling population processes. Blackwell, Oxford.
Rogers, A. (1975), Introduction to multiregional mathematical demography. Wiley, New York.
Salmon, S.C. (1998), Causality and explanation. Oxford University Press, New York.
Scherer, S. (2001), Early career patterns: a comparison of Great Britain and West Germany. *European Sociological Review*, 17, pp. 119-144.
Schoen, R. (1988a), Modeling multigroup populations. Plenum Press, New York.
Schoen, R. (1988b), Practical uses of multistate population models. *Annual Review of Sociology*, 14, pp. 341-361.
Simon, H.A. (1979), The meaning of causal ordering. In: R.K. Merton, J.S. Coleman and P.H. Rossi (eds.). Qualitative and quantitative social research. Free Press, London, pp. 65-81.
Therneau, T.M. and S.M. Grambsch (2000), Modeling survival data. Extending the Cox model. Springer Verlag, New York.
Veblen, Th. (1898), Why is economics not an evolutionary science? *Quarterly Journal of Economics*, 12, pp. 373-397.
Vromen, J.J. (1995), Economic evolution. An enquiry into the foundations fo New Institutional Economics. Routledge, London.

Willekens, F.J. (1999), Reconsturction of life histories using multistate life tables. Paper presented at the international workshop "Synthetic Biographies: State of the Art and Developments", San Miniato (Pisa), June 1999.

Yamaguchi, K.(1991), Event history analysis. Sage Publications, Newbury Park, Ca.

Zaidi, A. and K. Rake (2001) Dynamic Microsimulation Models: A Review and Some Lessons, SAGE Discussion Paper no. 2. Available from the SAGE homepage http://www.lse.ac.uk/Depts/sage/conference/workshop.htm

Zeng Yi, Wang Zhenglian, Ma Zhongdong and Chen Chunjun (2000), A simple method for projecting or estimating α and β: An extension of the Brass Relational Gompertz Fertility Model. In: Population research and Policy Review 19: 525-549.