Chapter 1

Fundamental concepts

Introduction

It has long been acknowledged that descriptive epidemiology is primarily characterized by its exploratory goals. It is seen as a first approach aimed at defining the scope of a research problem, at best generating hypotheses without aspiring to verify them. When descriptive epidemiology is seen in this light, the fact that no important developments in methodology had taken place until recently is less surprising. Its basic techniques were borrowed from demography: mortality and morbidity rates were seen as the key descriptive tools, with their comparison and standardization being the only methodological sophistication required. Statistical variability was rarely taken into account, sometimes producing serious errors in interpretation.

Several factors seem to have inspired the development of the techniques which make up modern descriptive epidemiology. The first is probably the proliferation and improvement of epidemiological data. In the area of cancer research these developments have undoubtedly been greater for incidence data than for mortality data. Cancer registries have multiplied and worked to standardize their definitions and registration procedures. The collection of demographic data, which provides the denominators of rates, has also seen a marked improvement, notably in the frequency of their publication.

The accumulation of incidence and mortality data over time has led to a focus on the analysis of time series. New techniques, mainly based on mathematical modelling, have been developed to distinguish between the different factors that underlie changes in rates. These methods have had both explanatory as well as predictive goals.

Descriptive epidemiology have also benefited from a more rigorous definition of its concepts, and from a more satisfactory incorporation in its methodology of the basic ideas developed in the context of stochastic process analysis. Appropriate mathematical and statistical methods have been developed, largely due to the contribution of epidemiologists. These advances follow a similar development of statistical methods in other areas of medicine. It is significant that published reports of epidemiological investigations now have a readership which includes specialists from other areas of research. The new approaches have led to better solutions to the

problems posed, in particular through more appropriate definition of hypotheses and the construction of suitable models for their evaluation.

The integration into descriptive epidemiology of spatial analysis and a more critical consideration of ecological studies are two examples of the increasing interaction between the improvement in data collection and the need for more sophisticated methods. Thus, the collection of increasingly detailed morbidity and mortality data, and the creation of data systems which allow cases and deaths to be located in time and space, have provided a basis for evaluating real or supposed environmental hazards, requiring in turn the development of appropriate statistical methods.

In the same way, when suspected exposures are easier to define at a group level rather than at an individual level, it is the role of descriptive epidemiology to assess the relationship between these exposures and the risk of cancer. Techniques to better control for potential confounding factors have thus been added to the classical methods of geographical correlation.

Traditionally, epidemiology is defined as the study of the distribution of diseases over time and place and according to individual characteristics. For the purpose of this book, descriptive epidemiology can be defined by replacing this last term with 'group characteristics'. This definition encompasses the intended contribution of descriptive epidemiology to etiologic research, as well as emphasising that data known only at a group level are the basis of the discipline. Inference is made from the group to the individual, in contrast to analytical epidemiology, in which risk is studied in groups formed *a posteriori* from data collected at an individual level. Throughout this text, it will be seen that the formation of groups on which the analysis is ultimately based is one of the crucial problems confronting descriptive epidemiology.

Apart from the methods of data collection, both for defining populations at risk and identifying risk factors, descriptive epidemiology utilizes exactly the same methodology as that of cohort studies in analytical epidemiology. Moreover, it will be seen that the concepts used are exactly the same. This resemblance is especially obvious when descriptive epidemiology has the task of describing the survival of cancer patients according to group characteristics. In this situation, data are available for individuals and the distinction between analytical and descriptive epidemiology becomes somewhat artificial. Survival studies have progressively found their place as an activity appropriate to cancer registries, and their goals are mainly descriptive in this context. Presentation of the methods of incidence analysis and then of survival analysis in the same text is in any case justified both mathematically and statistically. These two forms of analysis both concern the occurrence of an event (diagnosis or death respectively) in the presence of competing risks which lead to incomplete observation (also known as censoring). The estimation and modelling of the probability of occurrence of such an event leads to analytical methods requiring mathematical concepts rarely taught in medical schools.

In this first chapter, our goal will be primarily to convince the reader of the need for such ideas, then to present them as simply as possible through examples, while also providing the appropriate theoretical background. The subsequent chap-

ters offer a more user-oriented description of the methods, so that the reader can carry out the calculations and tests presented. It should be emphasized that the reader who does not wish to become involved in the theoretical developments of the first chapter can by-pass them, without compromising an understanding of the rest of the book.

Basic concepts of descriptive epidemiology

Time and the concept of incidence

While no-one would dispute that cancer incidence varies with time, there is less agreement over the causes of its evolution. Public opinion readily seizes upon the idea that the disease is a modern-day plague. Some people maintain that the increase in the incidence of cancer is simply due to the ageing of Western populations and to the fact that the other diseases from which people used to die are being controlled. On the other hand, there are others who will state that it is a curse, linked to atmospheric pollution, nuclear energy or the use of new chemicals. Epidemiology allows us to establish that, in any given age group, the frequency of cancer (apart from those associated with tobacco) is remaining almost constant or, in some countries, is even decreasing (see Chapter 3, page 174).

These contradictory statements may seem to be an illustration of the saying that statistics are a sophisticated form of lying. In fact, they result from the difficulty of differentiating between the effects of many variables which are acting simultaneously on the phenomenon being studied: at the end of the twentieth century, the 'educated layman' does not necessarily have available the tools needed to make an objective analysis of the effects of these variables. The first step towards an understanding of the problem is an accurate definition of the concept of incidence.

In epidemiology as in demography, time can be located by two indices: date and age. Cancer incidence can only be described properly by taking into account both of the indices which play parallel roles and are in fact measures of time with respect to two different origins.

Figure 1.1 (the Lexis diagram) illustrates this duality: a segment of oblique line in this graph represents the observable fraction of an individual's life, that is, the interval of time and age during which an event of interest (e.g., incidence of cancer or complications of diabetes or AIDS in a seropositive patient) can occur. The left extremity of the segment is the start of observation: it is for example the date of birth in the descriptive study of cancer incidence, or the date of first employment in an industrial cohort aimed at measuring the risk of a suspected exposure. It could also be the date of the start of treatment in a study designed to measure the risk of relapse after illness or the chance of survival after the occurrence of a serious disease. The other extremity is the end of observation, characterized by the date

and age at which either the event under study took place or the individual stopped being observed. This second possibility can be due to death (when we are interested in the incidence of some other event), loss to follow-up of the subject or the end of the study. In these three situations, it is said that the observation is *censored* because the event had not yet taken place at the end of the observation period. It is only known that the time necessary for the event to happen to the individual is greater than the duration of the observation period.

In some studies it is the death from a given disease which is the event of interest, either because incidence data are not available or because the probability of surviving from this diseases is the subject of the analysis. The censored observations comes in this context from subjects who died from other causes, who were lost to follow-up or for whom the diagnostic of the disease was too recent.

Depending on the point of view adopted, we can look at different segments of an individual's trajectory on the Lexis diagram. In a study of survival, the origin of the time scale is most often the date of diagnosis or of first treatment. The duration of *time at risk of death* is therefore measured as the time elapsed from this date, age being considered as an additional prognostic variable. Conversely, in an industrial cohort study, the basic measure of time is usually age, the time since the first entry being taken as an explanatory covariate. But, in both situations, the time

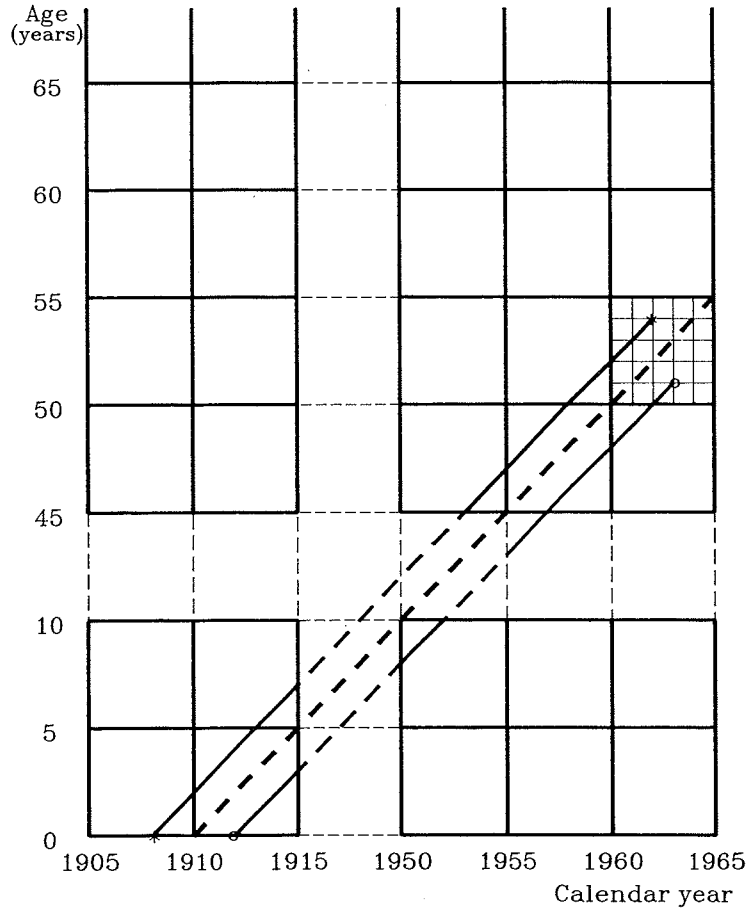

Figure 1.1 The Lexis diagram

origin is specific to each individual in the study; for a correct statistical analysis, we must 'synchronize the clocks' governing each individual's life events.

The aim of the epidemiologist is to draw conclusions about the different levels of risk to which groups of individuals are subjected. This requires well defined measures of risk in order to make objective comparisons between one situation and the next, or between one country and another. These measures might take the form of the probability of developing cancer or of dying from it, or they might be the survival rate or the probability of relapse. In all cases, the measures are based on a ratio between the number of observed events (the numerator) and the number of individuals at risk within a given period of time (the denominator). Alternative choices of the latter can lead to widely divergent results.

A few simple analogies will show how important the problem is. For example, if we want to compare the safety of different types of transport, should we measure the number of passenger deaths per kilometre travelled, per passenger x kilometre, or per passenger x time travelled? It is obvious that the definition of risk depends on the method of calculation. To compare the incidence of cancer in two cohorts, should we base our results on the observed proportion of cancer in each group, or should we take into account the number of years for which each individual was actually observed and at risk of developing cancer? If the two cohorts have the same average age and have been observed for the same time period and if the only reason for stopping observation was the onset of cancer (or, more generally, the event under study), the proportion is a good index of comparison. If, as more often happens, other events prematurely bring some individual observations to an end and if, in addition, these events do not occur in the same way in the two cohorts, it is likely that more cancers will be seen in the group which has, on average, been observed for longer. Conversely, if we take the duration of employment as an approximate measure of exposure in a study of lung cancer mortality in an asbestos mine, we should be aware that remaining employed for a given duration means having survived this number of years. Thus, if we want to assess the risk of people employed for more than twenty years, only the period beginning after twenty years of employment and the corresponding cases of cancer would be taken into account for the evaluation of this risk.

These examples lead to the following principles:

- the calculation of the denominator should take into account the number of years of observation relevant to the proposed study; it should take into account the continuous modification of the population actually 'at risk' throughout the duration of the study. By definition, a subject is no longer at risk after the occurrence of the event or after the censoring time.[1]
- incidence rate should be defined as the number of events per *person-year*, that is, per person and per year of observation relevant to the risk being analysed.

---

[1] Note however that cancer registries record second primary cancers. Strictly speaking, the period at risk starts in this situation immediately after the first tumour as if a new subject was added to the population at risk at this point.

• A given period of observation of a subject will contribute to the person years in the denominator only if this subject would have been counted in the numerator had he experienced the event being studied over that period of time.

Here it is appropriate to introduce *the instantaneous rate*, a concept which is crucial to epidemiology. Intuitively, this parameter measures the probability that an individual in a defined population becomes a victim of the event at a specified time point, given that the individual is still living and under observation at that time. In the same way that the speed at a given moment can be approximated by an average speed, so an instantaneous rate can be approximated by an average rate. In Figure 1.1, the squared cell shows individuals who were 50 to 55 years old between the years 1960 and 1965, that is, individuals who were born between 1905 and 1915. If the asterisk represents the end of observation due to the occurrence of cancer and the point represents the termination of observation for all other reasons, the risk of developing cancer between 50 and 55 years of age for individuals born around 1910 is then measured by the number of asterisks observed in the square divided by the number of years accumulated in the same space by the individuals born between 1905 and 1915. Only individuals born in 1910 will be able to accumulate five years of observation; the further the birth date is from this date, in either direction, the smaller the individual's contribution to the calculation of the denominator in this square. The resulting ratio, generally called the *average annual rate* of cancer between 50 and 55 years for the generation born around 1910, or else the *specific rate* for the age group 50-55 years, is an approximation to the instantaneous rate.

Figure 1.2 shows the evolution with age of lung cancer mortality in France; it can be seen that, for successive generations, those born more recently have suffered the highest lung cancer mortality. In such a situation, the cross-sectional curve obtained by plotting age-specific rates at a given time point (for example, the curve obtained by joining the points corresponding to the period 1950-1954) would be an incorrect description of the phenomenon if it was interpreted as a representation of the effect of age. Actually, the observed decrease in risk for higher ages corresponds to a *generation effect*: it has been shown that the lung cancer risk in the older French population is lower only because the corresponding generation has had less exposure to tobacco. The phenomenon is clearly seen in Figure 1.3, where the evolution of mortality from cancers of the lung, the oesophagus and the larynx in France is shown for successive generations. The lung cancer risk increases regularly with date of birth, whereas the risk of cancers of the oesophagus and larynx, which are much more dependent on alcohol consumption, have both been smaller for those generations subjected to rationing related to the second world war.

Group characteristics and place

By revealing the large variability in cancer incidence throughout the world, descriptive epidemiology has shown that the prevention of cancer is, at least partially, possible; differences observed, particularly within the same ethnic groups, have

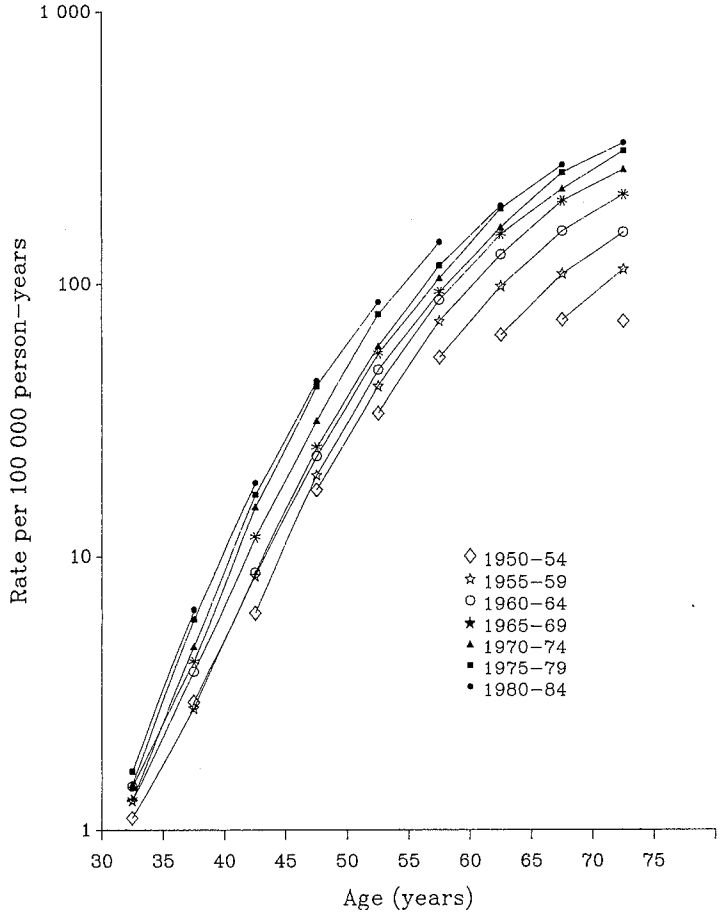

Figure 1.2 Evolution of lung cancer mortality in France, men; age-effect in successive birth cohorts drawn from cross-sectional mortality in successive time periods

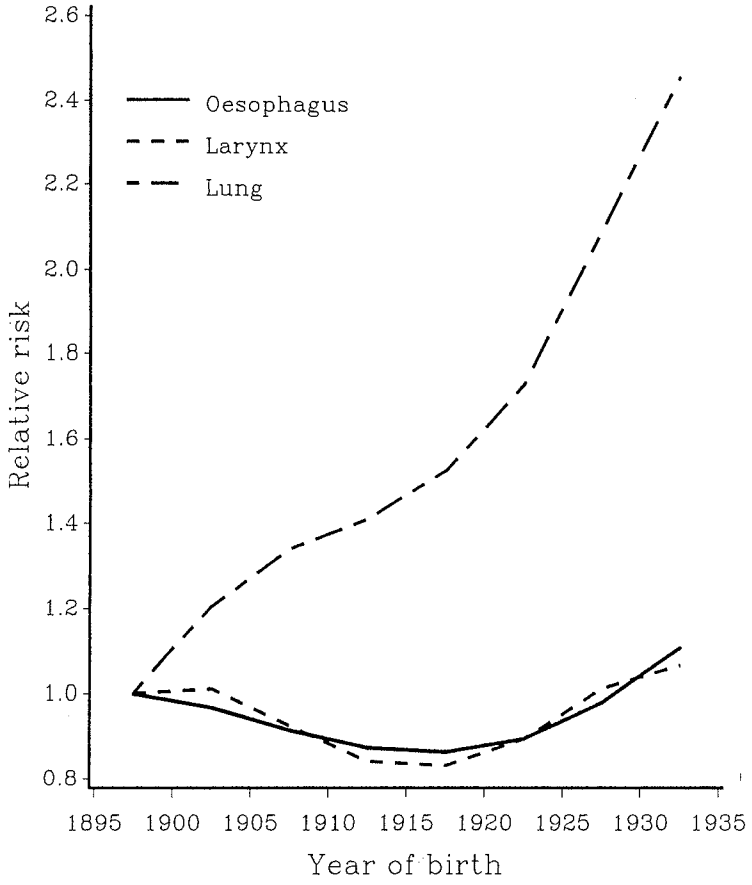

Figure 1.3 Relative risk of death from lung, larynx and oesophageal cancers for successive male birth cohorts in France compared to the cohort born around 1897

unambiguously established that environmental factors play a determining role in the development of cancer. The most striking example is probably that of oesophageal cancer for which the risk is 300 times more elevated in the north-east of Iran than it is in Nigeria. The reasons for this difference have only been partially identified, and it is quite likely that multiple factors are responsible [1]. In Europe, the risks in the regions most affected by oesophageal cancer are about a factor of 30 greater than in the regions least affected; the incidence of this cancer is highest, where the highest average alcohol consumption is reported, notably in the west of France.

Given this large variability, it is extremely tempting to try to establish causal relationships by analysing the correlation between the variation of incidence and environmental factors in different populations, and in fact many such analyses have been attempted. On the whole, however, these attempts have been rather unsuccessful: on an international scale, no substantial correlation has been demonstrated between oesophageal cancer and alcohol consumption. Undoubtedly, one of the reasons for this failure is that cancer is a multifactorial disease and that the determining factors need not be the same in two regions with very different cultural traditions. Another reason is that the degree of exposure to the factor can be distributed unequally among the individuals in the regions being compared, even if the average rate of exposure in the regions is similar. For example, it is conceivable that a country with a minority of heavy drinkers and a majority of teetotallers would report more cancer than another area with more widespread consumption at a lower level.

An absence of correlation can also be observed for less obvious methodological reasons. For example, studies on individuals show that tobacco is responsible for 85% of the lung cancer observed in populations where smoking is widespread [2]. However, if we restrict ourselves to Europe, a group of seventeen countries that is reasonably homogeneous for other factors, the correlation for the period 1970-74 between lung cancer risk (cumulative up to 80 years) and the consumption of cigarettes for the same period is only 0.56. A correlation of this size means that the variation in the consumption of tobacco explains barely a third of the variation in mortality, which is hardly compatible with the above number of 85%.

In fact, the correlation between tobacco consumption and lung cancer is slightly more impressive if we look at it correctly [3]. The first mistake in the preceding discussion is to have considered the cumulative risk cross-sectionally, thereby adding together risks over generations that had radically different tobacco exposure. The second mistake is to have considered the consumption of tobacco contemporaneously with the mortality when the latent period between exposure and the occurrence of cancer should have been taken into account. Comparing the cumulative risk for lung cancer for the seventeen countries between the ages of 35 and 50 years for people born around 1925, and cigarette consumption between 1955 and 1964 (Table 1.1 and Figure 3.9), we obtain a correlation between the two variables equal to 0.75, a much more reasonable value for data limited by substantial imprecision.

Conversely, these remarks hold true when a factor is correlated positively with a disease on a geographical level; the correlation is not always found in studies

Table 1.1 National cigarette sales and mortality from lung cancer in selected European countries

	Cumulative risk from 35 to 50 years Generation born in 1925 [a]	Number [b] of cigarettes (Rank) – 1955-64
Portugal	0.69	905 (2)
Sweden	0.75	1 147 (4)
Norway	0.86	533 (1)
Spain	0.98	1 112 (3)
Germany [c]	1.30	1 571 (9)
France	1.32	1 318 (6)
Iceland [d]	1.35	1 736 (13)
Greece	1.38	1 714 (11)
Austria	1.44	1 647 (10)
Switzerland	1.54	2 281 (15)
Denmark	1.57	1 375 (7)
Netherlands	1.73	1 716 (12)
Finland	1.77	1 975 (14)
Ireland	1.88	2 600 (16)
Italy	1.95	1 265 (5)
Belgium	2.09	1 539 (8)
UK	2.83	2 672 (17)

([a]) Average risk (per thousand) for both sexes combined; source: WHO (WHO Mortality Data Bank).
([b]) Average annual cigarette sales per adult above 15 years (1955-1964) [8].
([c]) Former Federal Republic of Germany.
([d]) Risk estimated from 14 cases, and therefore of limited reliability.

involving individuals. This situation can be illustrated by the correlation found between beer consumption and mortality from cancer of the rectum [4,5] and also the correlation between consumption of fat and breast cancer mortality [6,7].

A technical presentation of this approach and other examples will be given in Chapter 3 (see page 141), where the usefulness of this methodology will be discussed. The above examples were presented to show that the interpretation of descriptive data requires the same attention as data coming from an analytical study. Only a combined analysis of results obtained at a group and an individual level will provide the correct scientific interpretation.

Epidemiology is a science of observation, which means that it is limited to making use of natural events which simulate an experimental design. Seen from this point of view, studies of migrants and religious groups have been extremely successful. Table 1.2 provides a particularly attractive example based on the incidence of certain cancers observed in various Israeli communities and in selected western populations. The figures show that, for the given cancer sites, the incidences observed in Israel are consistently lower than those observed in the western countries used as reference, but their basic interest lies in the differences that they reveal between the Israeli communities. In fact, Jewish people not born in Israel have a risk half way between the risk of their country of origin and that of their adopted country. This tends to confirm that the observed change in risk was linked to a change in environment.

Table 1.2 Cancer and migration ([a]): Incidence rates ([b]) for selected cancer sites in Israel (1972-76), in Geneva (1973-77) and in Connecticut, USA (1973-77)

Population	Males			Females		
	All cancers except skin	Respiratory ICD8: 160-162	Digestive ICD8: 150-157	All cancers except skin	Respiratory ICD8: 160-162	Digestive ICD8: 150-157
Non-Jews born in Israel	117.3	35.7	22.9	62.8	12.3	9.0
Jews born in Africa or in Asia	167.1	32.0	42.2	137.3	30.4	18.5
Jews born in Israel	183.7	22.9	51.3	187.1	35.5	30.1
Jews born in Europe or in America	211.4	34.9	66.7	226.6	55.2	32.3
Connecticut	303.0	69.7	80.4	257.3	54.3	43.6
Geneva	328.6	81.2	88.8	225.2	46.3	42.5

([a]) Source: Cancer Incidence in Five Continents [9].
([b]) Rates standardized on world population.

Table 1.3 Standardized ([a]) incidence rates (T) and standardized incidence ratio (SIR) ([b]) for selected cancer sites in Utah, USA (1967-1975) ([c])

		Mormons		Non-Mormons	
		Urban	Rural	Urban	Rural
Males					
Tobacco-related sites ([d])	T	52.8	50.4	141.0	74.4*
	SIR	44^-	43^-	106	59^-
Lung (ICD8: 162)	T	27.1	27.3	75.7	40.3*
	SIR	37^-	39^-	96	54^-
Females					
Breast	T	63.5	55.5	90.9	80.8
	SIR	84^-	74^-	121^+	97
Uterus					
Cervix, invasive	T	8.1	9.4	17.7	18.4
	SIR	54^-	60	120^+	111
Cervix, in situ	T	15.9	12.9	45.4	34.7
	SIR	–	–	–	–
Corpus	T	21.9	19.1	27.4	24.8
	SIR	104	91	130^+	107

([a]) Standardized on 1970 US population, * significantly different from the urban rate.
([b]) TNCS Standard (Third National Cancer Survey), $^-$ significantly lower than the national rate, $^+$ significantly higher than the national rate, — data not available.
([c]) Source: Utah Cancer Registry (1967-1975) [10].
([d]) ICD8: 140 (lip), 143-150 (buccal cavity, pharynx and esophagus), 161 (larynx), 162 (lung), 188 (bladder).

Table 1.3 reproduces some results from the Utah cancer registry. A significant proportion of the population comprises Mormons, who do not consume alcohol or tobacco and who individually have less sexual partners and on average more children than the rest of the population. Taken together, this behaviour has noticeable consequences on the incidence of cancer at several sites, as shown in the Table. As these figures were not derived from a controlled experiment, it is likely that the Mormon population group differs from the non-Mormon group for other characteristics which can be associated with cancer development. Nevertheless, it is noteworthy that the classic excess of incidence in urban populations, seen here in the non-Mormon group, disappears in the Mormon community. The urban-rural difference is thus very likely to be due to differences in individual behaviour between urban and rural inhabitants, rather than being explained by one of the urban risk factors (such as pollution) usually invoked as explanatory.

In practice, the possibility of establishing relationships such as those which we have just described largely depends on the use of the appropriate statistical methodology. In particular, the methodology should provide the means of evaluating the variability attributable only to chance, so that it can be taken into account in the interpretation of observed differences. The remainder of this chapter will be devoted to a discussion of mathematical concepts which are the basis of the analytical methods. A discussion of practical applications will be kept for the subsequent chapters.

Statistical concepts for the analysis of incidence data

Formal definition of the incidence rate

We have seen above that the identification of factors favouring or causing the occurrence of a disease or a death requires the measurement of the risk of developing the event. In other words, we need an unbiased estimate of the probability that an individual, in a given environment, might develop the event under study. Besides the factors under study, this probability depends on temporal variables such as age, in incidence and mortality studies, and duration of observation in survival studies. The mathematical concept which is fundamental to risk and survival assessment is the distribution of the time separating the beginning of observation from the occurrence of the event. From a knowledge of this distribution we can measure, for example, the risk of cancer before age t, or the risk of death t years after diagnosis. The date of the development of the event under study is often unknown because observation is interrupted before the event occurs; in this case it is necessary to use specific techniques to estimate the distribution from incomplete observations.

As we noted previously, the period for which an individual is followed is the result of two competing mechanisms, which results in two different types of obser-

vation; one produces the event under study and the other includes all the other causes which might be responsible for terminating observation. Our aim is now to show how, by taking into account these two types of observation, we can reconstruct the distribution that would have been seen if all observations had been completed.

In this way people dying, for example from a cardiac disease, before age t will contribute to the calculation of the probability of having cancer before this age; similarly the follow-up of patients who were only diagnosed 1, 2, 3 or 4 years ago will contribute to the calculation of the survival probability at 5 years.

The mathematical concept used for this reconstruction is the instantaneous rate, which was defined intuitively above. We now adopt a more formal approach, which will allow further mathematical developments.

Let T denote the time period between the start and the end of observation for an individual, whether terminated by the end-point under study (for example, the occurrence of cancer) or by any other circumstance which might interrupt the follow-up. Furthermore, let δ be the indicator function of the end-point: $\delta = 1$ when the event has taken place and $\delta = 0$ when the observation is censored.

The following definitions characterize the random distribution of the couple of variables (T,δ). Let

- $R(t) = \text{Prob}\ (T < t)$ be the probability distribution of T
- $S(t) = 1 - R(t)$ denote the probability that the subject is still under observation (surviving) at the time-point t without the event having taken place,
- $p_1 = \text{Prob}\ (\delta = 1)$ be the probability that the event take place and
- $R_1(t) = \text{Prob}\ (T < t \mid \delta = 1)$ be the conditional distribution of the event, that is the probability that the event takes place before the time t, given that it has taken place.

Thus, the probability that the event occurs before the time t may be written

$$\pi(t) = \text{Prob}\ (T < t,\ \delta = 1) = p_1 R_1(t).$$

The probability that the event occurs on a given date, while the subject is still being followed-up, defines the force of incidence (or mortality) at this point in time. The following expression, which is directly derived from this probability, will be referred to as the *instantaneous rate*,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \text{Prob}\,(t < T < t + \Delta t\,,\ \delta = 1 \mid T > t) \tag{1.1}$$

It should be noted that $\lambda(t)$ is not, strictly speaking, a probability, but a probability per unit of time, also known as a probability rate. Application of the rules of probability immediately gives

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{p_1 R_1(t + \Delta t) - p_1 R_1(t)}{1 - R(t)} \tag{1.2}$$

The numerator is the probability that the event occurs at time t of the subject's follow-up; the denominator indicates that only the subjects who have been followed-up at least until t are taken into account. Furthermore,

$$\lambda(t)\, S(t) = p_1 \lim_{\Delta t \to 0} \frac{R_1(t+\Delta t) - R_1(t)}{\Delta t} = p_1 R'_1(t) \tag{1.3}$$

where $R'_1(t)$ is defined as the conditional probability density of T, derivative of $R_1(t)$.

In this way we obtain the relationship between $\lambda(t)$, $S(t)$ and the distribution function of T when the event occurs. The probability that the event occurs before time t can be written

$$\pi(t) = \int_0^t p_1\, R'_1(u)\, du$$

that is,

$$\pi(t) = \int_0^t \lambda(u)\, S(u)\, du \tag{1.4}$$

In the situation where there are no censored observations, $p_1 = 1$ and $R_1(t) = R(t)$; this would be the case for example in a study of mortality from all causes, if every individual in the cohort was under observation until death. In this situation, formula (1.3) leads to

$$\lambda(t) = \frac{R'(t)}{1 - R(t)} \tag{1.5}$$

which is a differential equation with solution

$$\Lambda(t) = -\,\mathrm{Log}\,[S(t)] \quad \text{and} \quad R(t) = 1 - e^{-\Lambda(t)} \tag{1.6}$$

where

$$\Lambda(t) = \int_0^t \lambda(u)\, du \tag{1.7}$$

When there are censored observations, the distribution function defined by formula (1.6) is in fact that which would govern the observations if they were all complete, that is, if none was censored. The probabilities generated by this distribution are called *net probabilities* as opposed to *crude probabilities* defined by $\pi(t)$ in formula (1.4).

If T is age and if the end-point is the occurrence of cancer, the following terminology is used:

- $\lambda(t)$ is the *force of incidence*,
- $\pi(t)$ is the *crude probability* of developing cancer by age t,
- $\Lambda(t)$ is the *cumulative incidence rate* at age t.

The net probability of developing cancer by age t, $R(t) = 1 - e^{-\Lambda(t)}$, is a measure of cancer risk when there are no censored observations, that is, in the absence of mortality. Therefore, the net probability is not affected by the structure

of the mortality pattern in the population under study and it can be used to compare several populations. This measure of risk is known as *cumulative risk*; its properties, and methods for its calculation are presented in Chapter 2 (page 66).

If T is the interval between diagnosis and the end of follow-up, the censored observations are essentially those for which the diagnosis is too recent; in this case, $\pi(t)$ is of little interest. The net probability of survival given in formula (1.6) is usually the parameter of interest in survival analysis.

Estimation of the instantaneous incidence rate

Having established a framework in which incidence can be defined, we must now consider methods for its calculation, or rather, for its estimation. The age-specific rates are usually calculated from the number of cases observed in the different age groups and from demographic statistics which enable the person-years of observation in each age group to be evaluated. Only the justification of the method will be given at this stage; the practical details will be left until Chapter 2.

In the case of a cohort of limited size in which each individual history is known and stretches over a long time period, the estimation of the age-specific rate requires an exact calculation of the person-years of observation. The estimation would be straightforward if the rate were independent of time and if each individual observation were complete; this situation is described on page 15. When instead some observations are censored, this fact has to be taken into account in the calculation (see page 18). The discussion will lead us to explain why the random fluctuations in the number of observed cases can be described by the Poisson distribution.

An approximation useful in descriptive epidemiology

We saw in formula (1.4) that the crude probability π of developing cancer between age t_0 and age t_1 depends on the age-specific rate $\lambda(u)$ and the probability of surviving without cancer $S(u)$, that is

$$\pi = \int_{t_0}^{t_1} \lambda(u)S(u)\,du$$

In principle, this probability can be easily estimated from data on a population with a given date of birth (a *birth cohort*). In this situation, the birth date is the natural time origin for all individuals in the cohort and the variable t is simply their age. Therefore, to estimate π we simply divide the number of cases occurring between age t_0 and t_1 by the initial size of the cohort. However, the survival to age t_0 will influence the result more than the value of the age-specific rate between t_0 and t_1. Thus, this probability is of no use in estimating $\lambda(u)$; in contrast, the conditional probability π_c of having the disease between age t_0 and t_1, given that the subject was still at risk at age t_0, is obviously not influenced by survival up until

that age and is very little influenced by survival between t_0 and t_1 if this interval is short. We can write

$$\pi_c = \int_{t_0}^{t_1} \lambda(u) \frac{S(u)}{S(t_0)} du \tag{1.8}$$

If $t_1 - t_0$ is sufficiently small so that $\lambda(u)$ can be considered constant in $[t_0,t_1]$ and $S(u)$ roughly equal to $S(t_0)$ in the interval, then

$$\pi_c \approx \lambda(t_0)\,(t_1 - t_0)$$

If on the other hand, n_{t_0} denotes the number of subjects at risk at t_0, and k is the number of cases observed between t_0 and t_1, then the estimate of π_c is

$$\hat{\pi}_c = \frac{k}{n_{t_0}}$$

and, therefore, the estimate of λ is

$$\hat{\lambda}(t_0) \approx \frac{k}{n_{t_0}(t_1 - t_0)} \tag{1.9}$$

In other words, the instantaneous rate estimated at t_0 is obtained by dividing the number of cases observed by the number m of person-years of observation for the cohort between t_0 and t_1, where $m = n_{t_0}(t_1 - t_0)$,that is

$$\hat{\lambda}(t_0) \approx \frac{k}{m}$$

Formula (1.8), which is simply the application of the definition of λ in Formula (1.1) above, shows that the approximation will not be good if $\lambda(u)$ varies sharply within the interval $[t_0, t_1]$ or when a large number of subjects die from other causes or are lost to follow-up between t_0 and t_1; in this situation, the ratio $S(u)/S(t_0)$ would become too far from unity for the approximation being valid. If there is a substantial proportion of censored observations, survival time must be explicitly taken into account for each of the n_{t_0} individuals in the interval $[t_0, t_1]$. In other words, the number of person-years of observation appearing in the denominator of formula (1.9) must be calculated exactly, by taking into account the date of the end of follow-up for each individual.

In order to understand the procedure to be used when individual observations are available, we shall first study the situation where $\lambda(u)$ remains constant and all observations are complete. Although this is rarely the case in practice, it will help us to understand the more complicated situation where observations may be censored. This simple example will also allow the principle of the maximum likelihood estimation to be introduced.

When individual observations are available and complete

Let t_1, t_2... t_n be the time elapsed between the start of the observation and the occurrence of the event under study for a random sample of n individuals subject

to the same constant hazard rate $\lambda = \lambda_0$. The cumulative hazard rate is then $\Lambda(t) = \lambda_0 t$ and the probability distribution of the t_i is defined by the function (see formula (1.6)):

$$P(T < t) = R(t) = 1 - e^{-\lambda_0 t} \tag{1.10}$$

This distribution known as the exponential distribution has the density

$$r(t) = R'(t) = \lambda_0 e^{-\lambda_0 t}$$

According to the principle of maximum likelihood, the estimate of the unknown value of λ is the value $\hat{\lambda}$ which maximizes the probability density of n observations written as a function of λ

$$V(\lambda) = \prod_{i=1}^{n} r(t_i) = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i} \tag{1.11}$$

It is intuitively clear that the higher the probability of the observed data for a given λ, the more likely it is that λ will be close to the unknown value λ_0.

For technical reasons, the function to be maximized is not $V(\lambda)$ but Log $[V(\lambda)]$, which in this example may be written:

$$L(\lambda) = \text{Log}\,[V(\lambda)] = n\,\text{Log}\,(\lambda) - \lambda \sum_{i=1}^{n} t_i \tag{1.12}$$

The value $\hat{\lambda}$ is then obtained by equating the derivative of L to zero:

$$\frac{d\,L(\lambda)}{d\lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} t_i = 0 \tag{1.13}$$

that is,

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\overline{T}} \tag{1.14}$$

In other words, in this situation, the instantaneous rate is estimated by the number of observed events divided by the time taken to produce all of them. This estimate can also be described as the inverse of the mean duration of the n observation times. The random variable $\hat{\lambda}$ which is a function of the t_i is known as the *maximum likelihood estimator.*

It has been shown that this approach generally produces a good estimator in that $\hat{\lambda}$ becomes numerically closer to λ_0 as n increases (*consistency*).

Study of the probability distribution of the difference between the log-likelihood of $\hat{\lambda}$ and that of λ_0 has shown that:

$$2\,[L(\hat{\lambda}) - L(\lambda_0)] \rightsquigarrow \chi_1^2 \tag{1.15}$$

It is therefore possible to state that this difference will rarely (less than 1 in 20 times) exceed the critical value 3.84 of a χ^2 with one degree of freedom. More generally using the result (1.15) allows the construction of a $(1 - \alpha)\%$ confidence interval for λ_0

$$[\lambda_I, \lambda_S] = \{\lambda \mid 2[L(\hat{\lambda}) - L(\lambda)] < Z^2_{\alpha/2}\} \tag{1.16}$$

In order to illustrate this method, 20 observations t_i of an exponential distribution with mean $\lambda_0 = 1$ were simulated. The sum of the observations was $\sum_{i=1}^{20} t_i = 19.36$. Thus $\bar{T} = 0.9680$ and $\hat{\lambda} = 1.033$. Figure 1.4 shows the function $2L(\lambda)$ in the neighbourhood of $\hat{\lambda}$ and the 95% confidence interval obtained from the above method. The quadratic approximation of $2L(\lambda)$ is also shown on the same graph as a dotted line. Since $\frac{dL(\hat{\lambda})}{d\lambda} = 0$, this approximation may be written according to Taylor's formula:

$$2L(\lambda) \approx 2L(\hat{\lambda}) + \frac{d^2 L(\hat{\lambda})}{d\lambda^2} (\lambda - \hat{\lambda})^2 \tag{1.17}$$

From this expression, it can be seen that the horizontal line $Z^2_{\alpha/2}$ units (3.84 units if $\alpha = 0.05$) below the maximum of the curve will intersect the dotted line at two points defined on the x-axis by:

$$[\lambda_I^* ; \lambda_S^*] = \hat{\lambda} \pm Z_{\alpha/2} \left\{ \sqrt{-\left[\frac{d^2 L(\hat{\lambda})}{d\lambda^2}\right]^{-1}} \right\} \tag{1.18}$$

This interval provides an approximate $(1 - \alpha)\%$ confidence interval for λ_0.

It may in fact be shown that, when n is large, the probability distribution of $\hat{\lambda}$ is normal with mean λ_0 and variance equal to the quantity under the square root sign in formula (1.18). This result, which can be generalized to more complex situations, will be used later in this book. In the simpler context of the exponential distribution presented here, the derivation of (1.13) gives:

$$\frac{d^2 L(\hat{\lambda})}{d\lambda^2} = -\frac{n}{\hat{\lambda}^2} \tag{1.19}$$

which is equal to −18.74 in the present numerical example. Then:

$$[\lambda_I^* ; \lambda_S^*] = \hat{\lambda} \pm 1.96 \frac{\hat{\lambda}}{\sqrt{n}} \tag{1.20}$$

which is equal to [0.58 ; 1.49] as shown in Figure 1.4.

The above method provides a simple means of constructing a confidence interval for an exponential distribution using a sample of independent observations. The negative of the second derivative in (1.19) may be considered as a measure

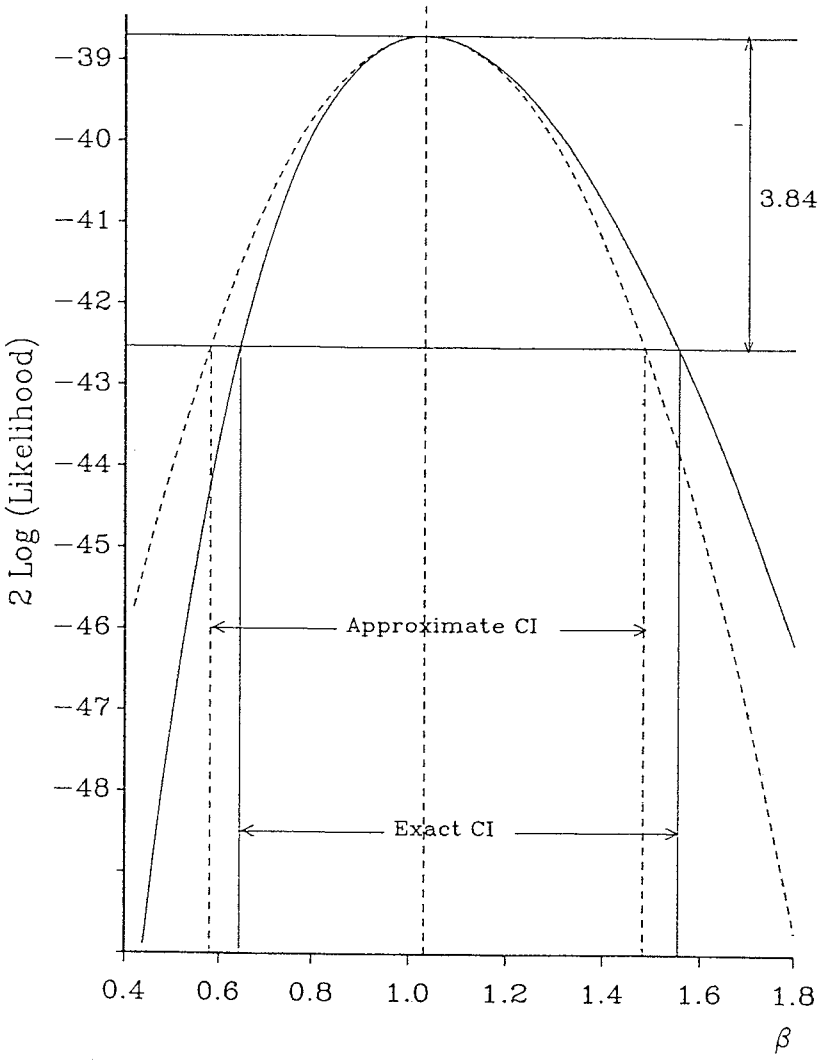

Figure 1.4 Maximum likelihood estimate of a parameter (β) and its confidence intervals (CI)

of the *information* provided by the sample with respect to the parameter λ: the larger the information, the more precise the estimate will be.

When individual observations are available but possibly censored

Let us now consider a cohort in which n individuals undergo the same force of incidence λ(u) and the same survival S(u) in the interval of observation 0, t (where the origin, 0, represents the beginning of observation, which may be, for example, the start of a five-year age interval for subjects born around the same time, see Lexis diagram Figure 1.1). Each individual observation is characterized by the value of two variables t_i and δ_i, where $\delta_i = 1$ if the event has taken place at the time t_i for individual i, and $\delta_i = 0$ if the event has not taken place at the time-point t_i when individual i ceases to be under observation, either because he has not survived or because $t_i = t$, i.e., the subject is alive at the end of observation.

If the function $\lambda(t)$ is defined by a finite number of parameters, these can be estimated from the sample of observations by choosing as previously values for the parameters which maximize their likelihood. Although the principle is the same, the situation becomes more complicated because of the presence of censored observations. The random variable (T,δ) does not have a probability density, therefore the density of complete observations ($\delta_i = 1$) and that of censored observations ($\delta = 0$) should be written separately.

- When $\delta_i = 1$, the contribution of the individual to the likelihood is given by formula (1.4),

$$P(t_i < T < t_i + dt, \delta = 1) = \lambda(t_i)\, S(t_i)\, dt$$

- When $\delta_i = 0$, the contribution is:

$$P(t_i < T < t_i + dt, \delta = 0) = c(t_i)\, S(t_i)\, dt,$$

where $c(u)$ is the analogue of $\lambda(u)$ for censored observations.

Thus, the likelihood may be written

$$V = \prod_{i=1}^{n} S(t_i)\, \lambda(t_i)^{\delta_i}\, c(t_i)^{1-\delta_i}$$

where $S(t)$ is the probability of still being followed up at time t without the event having occurred. Writing this probability as a function of incidence and censoring rates, we have

$$V = \prod_{i=1}^{n} e^{-\int_0^{t_i} [\lambda(u) + c(u)]\, du}\, \lambda(t_i)^{\delta_i}\, c(t_i)^{1-\delta_i}$$

$$V = \prod_{i=1}^{n} e^{-\Lambda(t_i)}\, \lambda(t_i)^{\delta_i}\, e^{-C(t_i)}\, c(t_i)^{1-\delta_i} \qquad (1.21)$$

where $C(t) = \int_0^t c(u)\, du$

If the mechanism which leads to censored observations is independent of incidence, $c(t)$ does not depend on the parameters that determine $\lambda(t)$. To maximize V with respect to these parameters, we can therefore ignore the last two factors. In fact the contribution of a censored observation to the likelihood becomes the probability that T is greater than t_i in the absence of risks other than the one under consideration. Therefore, the logarithm of the function to be maximized is

$$L(\lambda) = \mathrm{Log}\left[\prod_{i=1}^{n} e^{-\Lambda(t_i)}\, \lambda(t_i)^{\delta_i}\right]$$

$$L(\lambda) = -\sum_{i=1}^{n} \Lambda(t_i) + \sum_{i=1}^{n} \delta_i\, \mathrm{Log}\,[\lambda(t_i)] \qquad (1.22)$$

As an example, if $\lambda(t)$ is constant, this function becomes

$$L(\lambda) = -\sum_{i=1}^{n} \lambda\, t_i + \sum_{i=1}^{n} \delta_i \,\mathrm{Log}\,(\lambda)$$

$$L(\lambda) = -\lambda \sum_{i=1}^{n} t_i + k\,\mathrm{Log}(\lambda) = -\lambda m + k\,\mathrm{Log}\,(\lambda) \qquad (1.23)$$

where $k = \sum_{i=1}^{n} \delta_i$ is the number of events observed in the interval [0,t], and $m = \sum_{i}^{n} t_i$ is now the exact number of person-years of observation of the cohort within the interval [0,t]. The quantity m may also be written $n\bar{T}$ where $\bar{T}$ is the mean duration of observation.

The function reaches a maximum for

$$\hat{\lambda} = \frac{k}{\sum_{i=1}^{n} t_i} = \frac{k}{n\,\bar{T}} = \frac{k}{m} \qquad (1.24)$$

The comparison of formulae (1.9), (1.14) and (1.24) shows that the principle governing the estimation of λ is unique. The only variation is in the way in which the mean observation time is calculated. Furthermore, as above, the precision of $\hat{\lambda}$ obtained from the second derivative of the likelihood (1.23) is:

$$\mathrm{Var}\,(\hat{\lambda}) = \frac{\hat{\lambda}^2}{k} = \frac{\hat{\lambda}}{m}$$

At this point, it should be noted that the function to be maximized in formula (1.23) is, to within a constant, the logarithm of the likelihood of a single observation k having a Poisson distribution with parameter λm:

$$\mathrm{Log}\left[e^{-\lambda m}\, \frac{(\lambda m)^k}{k!} \right] = L(\lambda) + M(k\,,\, m) \qquad (1.25)$$

where $M(k,m) = -\mathrm{Log}(k!) + k\,\mathrm{Log}(m)$ refers to all the constant terms independent of λ.

Consequently, when estimating an instantaneous rate, although the numerator and denominator are both random variables, we are led to the same estimation procedure as if the numerator alone were random and followed a Poisson distribution. Therefore, the precision of the estimate of the incidence (or mortality) rate is judged exclusively from the variability of the numerator described by a Poisson distribution. We will use this equivalence throughout Chapter 2. The distribution of k is actually more complicated; however, there are no disadvantages and many benefits in making this approximation as long as the analytical methods are based on the likelihood. It is often stated that the true distribution of k is binomial; this would

only be the case if each of n individuals exposed to a given constant risk where observed for the same duration t defined *a priori* (see formula (1.9)). In this situation, the probability that the event (disease or death) occurs in a given individual would be $R(t) = 1 - e^{-\lambda t}$ (see formula (1.10)) and the number of observed events would follow the binomial law with parameters n and R(t). Actually, when R(t) is small, this distribution is close to the Poisson distribution of parameter $n(1 - e^{-\lambda t}) \approx n\lambda t = \lambda m$. However, the argument for the binomial law has little weight in practice since the contribution of the individuals to person-years is random and varies widely from subject to subject. We shall therefore consider that the Poisson distribution is the best compromise to describe the random fluctuation of the number of cases and that it remains adequate as long as the number of events (k) is small compared to the number of individuals at risk (n).

In practice, formula (1.24) is mainly used in cohort studies [11], since its use requires knowing the time t_i for each individual in the population under study. This information is available in a survival study and the terminology traditionally used in this context will be presented in the following sections.

Conversely, in a descriptive study, individual dates are never available and, as we have previously stated, the denominators of the age-specific rates must be estimated from demographic data. The most simple method of calculation is to multiply, for each age group x, the number of individuals recorded at the mid-point of the case-registration interval, by the number of years in the interval. If the local statistical office provides annual population data, the calculation of the denominators can be made in a way which is more precise. If we know the number of cases k_x which arise in the age group x during the year t and the total number of individuals $n_x(t)$ in the age group on the 1st January of the years t and t + 1, then we can estimate λ_x by using the average of these two totals for the number of years lived during the year t by the individuals in the groups:

$$\hat{\lambda}_x = \frac{k_x}{[n_x(t) + n_x(t+1)]/2} \qquad (1.26)$$

When the cases have been recorded between 1 January of year t and 31 December of year t + h, the sum of the annual average totals (the denominator of (1.26)) can be used to estimate the years lived for the period (see also page 27).

Statistical concepts in survival analysis

Follow-up studies

In the preceding section, basic principles for the analysis of event occurrence in the presence of censoring were discussed. These principles were illustrated by the examples of incidence or mortality where time is explicitly accounted for only in

the form of age. In this context, the observation of a large number of individuals over a short period is the basis for the analysis, but it could also involve a cohort in which the individual follow-up extends over several decades; therefore the ageing of the individuals is the principal factor which modifies the instantaneous rate of occurrence of the event under study. In survival studies, on the other hand, the rate is suddenly modified by the occurrence of the disease and tends to return to normal as the time since diagnosis increases; age becomes simply a covariable which can if necessary be taken into account in comparisons of the survival of several groups. Despite the similarities of the underlying principles, each of these situations has generated its own terminology and sometimes requires specific approaches; those used in the framework of survival studies will be reviewed below.

There are three fundamental notions on which the calculation of survival depends. The first is the *group* (or *cohort*), defined by a common event whose date marks the beginning of the observation period. In the context of cancer epidemiology, this date is usually the time when the risk of death is considered to be increased by the existence of the tumour, that is, the date of diagnosis. In clinical trials, as a general rule the point chosen is the date of randomization when the force of mortality should start to decrease as a result of treatment.

The second notion is the *follow-up* of each of the individuals in the cohort, from the date of the common event which defines the cohort; this procedure enables the status (living or deceased) of cohort members to be ascertained. It ensures in particular that those for whom death has not been notified are still living and under observation.

Finally, we require the *follow-up time* of each subject, defined as the time between the date of the common event characterizing the cohort and the date at which observation ends (the variable T of page 12). There are three ways in which observation of a subject ceases: by *death*; by the subject's being *lost to follow-up*, in which case the end of observation is considered to be the date of the last information on vital status; and by *withdrawal* from the follow-up of patients who have been diagnosed recently and therefore have a duration of observation shorter than the maximum time for which survival probability will be calculated.

Any observation that terminates by death is a *complete observation*. All others are *censored observations*. Two further terms will be defined. A *closed group* consists of a group of individuals in which there are only complete observations. An *open group* is a group where observations may be incomplete. In practice, it is rare to find a closed group except in the artificial situation of the construction of a life table. In most real situations, the group is open because there are subjects either lost to follow-up or withdrawn from follow-up.

When only one cause of mortality is taken into consideration, the group should also be treated as an open group. Observations which are interrupted by death from other causes can in fact be considered, under certain conditions, in the same way as other censored observations.

A further possibility which would imply an open group is the entry into the study of subjects subsequent to the occurrence of the disease which characterizes

the cohort. Such patients are by definition those who have survived at least up to their date of entry; their inclusion in the cohort would clearly lead to an overestimation of survival if this possible bias is not appropriately taken into consideration. In fact, the situation of a study cohort that accepts such subjects after the original group has been defined is rather uncommon and will not be considered further here.

In Chapter 4, we will discuss in detail different methods of follow-up which are being used at present in cancer registries.

Survival probability

If the group is closed, survival at time-point t can be calculated directly by the ratio between the total number of living subjects at time-point t and the original number of subjects, that is, n_t/n_0. In this context, the probability of survival has been termed *direct survival probability*. In this situation, survival can be estimated by the above ratio, and the statistical precision of this estimate can be assessed by noting that the numerator n_t obeys a binomial probability distribution law with index n_0, size of the cohort, and parameter S(t), survival probability at time t.

In practice, as previously explained, it is rare to find a closed group for several reasons. Diagnoses occur gradually over time and information brought to the study by cases which recently join the study is useful. Alternatively, there may be a number of subjects lost to the study whose observations could contribute to the final analysis. In these circumstances, survival probability can only be properly estimated by utilizing the idea of instantaneous rate. An alternative approach, especially appropriate in dealing with discrete data, is based on the concept of conditional probabilities of death.

If s(t) is the conditional probability that the subject is living at date $t + \Delta t$, given that he or she was living at t, then the probability that this subject is living at date $t + \Delta t$ is

$$S(t + \Delta t) = S(t)\ s(t) \tag{1.27}$$

Therefore, the calculation of survival depends on dividing the observation time into successive intervals $(0, t_1, t_2 \ldots t_k)$, and on making a separate calculation of the conditional probabilities $s(t_i)$ for each one of them.

If we know for each interval $[t_j, t_{j+1}]$ the number of subjects n_{t_j} who are at risk at the beginning of the interval t_j, as well as the number of deaths d_{t_j} occurring in the interval, we can estimate the values of $s(t_j)$ by $1 - d_{t_j}/n_{t_j}$, and, from them, we can deduce $S(t_{j+1})$ for successive intervals. Thus the probability of surviving until the end of the ith interval is

$$S(t_i) = \prod_{j=0}^{i-1} s(t_j) \tag{1.28}$$

with $t_0 = 0$.

The actuarial method and the Kaplan-Meier method described in detail in Chapter 4 are both based on this principle. These two methods actually differ only in the definition of the intervals used for calculating $s(t_j)$. The choice of intervals is linked to the assumption that we make about the instantaneous death rate. For the actuarial method, we assume that the instantaneous rate is constant in the intervals which are defined *a priori*; in the second situation, no assumption is made about the instantaneous rate, which leads us to assume that it is zero in the interval between two deaths; the dates of death are then the end-points of the intervals (Kaplan-Meier method).

Suppose that λ is constant in the interval $(t, t + \Delta t)$ and that d_t deaths have been observed among n_t subjects under observation at time t. Then the estimate of the instantaneous rate is $\hat{\lambda} = \dfrac{d_t}{m_t}$ where m_t is the number of person-years of survival of the n_t subjects in the interval (see (1.24)). If we assume that d_t deceased individuals and r_t subjects with censored observations had been living on average for half of the interval, the estimate of λ is

$$\hat{\lambda} = \frac{d_t}{\Delta t\left(n_t - \dfrac{r_t}{2} - \dfrac{d_t}{2}\right)} = \frac{d_t}{\Delta t\left(N_t - \dfrac{d_t}{2}\right)} \tag{1.29}$$

where

$$N_t = n_t - \frac{r_t}{2} \tag{1.30}$$

Therefore, we make the calculation as if N_t subjects were at risk at the beginning of the interval and that d_t deaths were observed among them. The probability of death is then

$$\hat{q}_t = \frac{d_t}{N_t} = \frac{2\hat{\lambda}\Delta t}{2 + \hat{\lambda}\Delta t} \tag{1.31}$$

The above formula (1.31) which links rate and probability has been used in the context of the construction of the life table (see page 26). The assumption that $\lambda(t) = \lambda$ remains constant in the interval should in fact imply

$$\hat{q}_t = 1 - e^{-\hat{\lambda}\Delta t} \tag{1.32}$$

The expressions (1.31) and (1.32) differ only by a term of the order of $(\Delta t)^2$, which is usually negligible.

In the actuarial method, it is the number N_t (*effective number at risk*), which is used as the denominator to calculate the probability of death. Therefore, the survival probability is calculated at the end of each interval by the formula

$$S(t + \Delta t) = S(t)\left(1 - \frac{d_t}{n_t - \dfrac{r_t}{2}}\right) = S(t)\,\frac{N_t - d_t}{N_t} \tag{1.33}$$

and, furthermore, by using the approximation

$$S(u) = S(t)\left(1 - \hat{\lambda}\,\frac{u-t}{\Delta t}\right)$$

in each interval, the function is linear between t and t + Δt.

The Kaplan-Meier method is much simpler as no assumption is made about λ; the dates of death are now the only information available to estimate the survival probability and it cannot be excluded that λ is zero in the interval between two deaths. Accordingly, survival probability is estimated as constant between two deaths. In other words, if all the dates of death are distinct and if t_i and t_{i+1} are the dates of two successive deaths, survival probability just after t_{i+1} is

$$S(t_{i+1}) = S(t_i)\left(1 - \frac{1}{n_{i+1}}\right) \tag{1.34}$$

where $n_{i+1} = n_i - 1 - r_i$ is the number of subjects remaining under observation just before t_{i+1} if r_i observations are censored between t_i and t_{i+1} (inclusive); function S is now constant between t_i and t_{i+1} and changes its value at the time of each death. Furthermore, it can be shown that S is the maximum likelihood estimate of theoretical survival. In practice, if several deaths occur on the same date, we use the formula

$$S(t_{i+1}) = S(t_i)\left(1 - \frac{d_{i+1}}{n_{i+1}}\right) \tag{1.35}$$

where d_{i+1} is the number of deaths observed on date t_{i+1}, and $n_{i+1} = n_i - d_i - r_i$. When censoring and death occur at the same time, it is considered that death occurs first; in other words, the censored observations at time t_{i+1} are counted in the denominator n_{i+1}.

Note that, in the actuarial method, the exact dates of death or loss to follow-up are not necessarily needed in the calculation; in fact, it is sufficient to know the subjects' status at the limits of the intervals. In the Kaplan-Meier method, the date of each death needs to be known but not the dates when subjects are censored, as only the number of censored observations between two deaths plays a role in the calculation.

Theoretically, the actuarial method could be improved if the exact dates of death and censoring were known; this information would enable the exact computation of the person-years of observation m_j in each interval $[t_j, t_{j+1}]$ to be carried out. If the death rate is constant in each interval and if Δt_j is the length of the interval $[t_j, t_{j+1}]$, survival would be estimated by the function

$$S(t) = e^{-\left[\sum_{j<i} \frac{d_j}{m_j}\Delta t_j + \frac{d_i}{m_i}(t - t_i)\right]} \qquad t_i < t < t_{i+1} \tag{1.36}$$

The argument of the exponential is the estimate of the cumulative rate, which we defined above (see page 13); each d_j/m_j is the estimate of the instantaneous rate in the interval $[t_j, t_{j+1}]$.

If d_j is the number of deaths from a given cause, all the methods estimate the net probability of survival from this cause, to the extent that risks which are related to other causes are independent. In practice, the possibility of estimating this probability presents several problems which will be discussed in Chapter 4; in particular, in the situation which arises when we are interested in deaths due to the disease under study among individuals diagnosed with the disease. The related concept of competing risk will be discussed on page 34 after we have introduced the necessary tools to construct a life table for a given population and discussed a few classical models for survival distributions (see page 29).

The life table

The *life table* is an example of calculating survival in a closed group. It describes, for each sex, the survival of a fictitious cohort of new-borns from one birthday to the next up until the complete extinction of the group, under the hypothesis that it is subject to the force of mortality of the population for which the table is constructed. As only one risk is operating, the group is closed, that is, subjects cannot leave the group for other reasons (such as departures or loss to follow-up); likewise, the group is closed to new entries (new arrivals) and the total number of subjects at each birthday is consequently the same as the number of surviving subjects at the preceding birthday minus the deaths which have occurred between the two birthdays.

The construction of the table is based on mortality rates by age; the rates are calculated from counts of deaths and census results, which explains why most of the tables refer to a period around the census date. The annual mortality rate is in fact often calculated over several calendar years in order to avoid large random fluctuations.

The table is built from a fictitious cohort whose initial total membership is arbitrarily fixed at 100 000 or 10 000 individuals (the radix of the table); it gives the number of surviving individuals at each birthday until a terminal age w at which, by convention, all members of the cohort have died (i.e., the number of cohort survivors at age w + 1 is zero).

The following terms, referred to as biometric functions, describe the principal information which is tabulated on a life table (see Appendix 1)

- x (column 1) indicates the beginning of the age interval, that is, birth and then successive birthdays. For most tables, x is used for males and y for females.
- $\hat{q}_x$ (column 2) is the proportion of individuals who die during the interval out of those who were living at the beginning of this interval. This proportion is the estimate of the probability of dying in the interval; it is obtained from vital statistics as described in (1.42). In the Swiss table given in Appendix 1, $\hat{q}_{25} = 0.001532$ is the proportion of those who died between their 25th and 26th birthdays.

• The quantity $\hat{p}_x = 1 - \hat{q}_x$ is the estimate of the conditional probability of survival between x and x + 1 given that the subject was alive at age x (column 3).

• $\hat{\lambda}_x$ (column 4) is the estimate of the mortality rate (see page 12).

• ℓ_x (column 5) is the number of survivors at the xth birthday, when the mortality at each age is defined by the series q_x. The series ℓ_x is called the *survivor function*. For example, the cohort of 100 000 births still includes ℓ_{25} = 97 155 survivors at the 25th birthday; the probability of survival which corresponds to this age is equal to 0.97155. As the group is closed, the probability of survival between x and x + h is given by the ratio of the number of survivors at these two birthdays:

$$\hat{p}_{x,h} = \frac{\ell_{x+h}}{\ell_x} \tag{1.37}$$

• L_x (which is not shown in the Table in Appendix 1) denotes the total number of years lived by the members of the cohort between x and x + 1 (the person-years), taking account of the fraction of years lived by those who died between the two birthdays. If the ages at death are spread uniformly over the interval, it may be written

$$L_x = \ell_x - \frac{1}{2} d_x = \frac{(\ell_x + \ell_{x+1})}{2} \tag{1.38}$$

showing that L_x is equal to the average number of individuals of age x.

• d_x (column 6) is the number of deaths which occurred in the cohort between age x and age x + 1.

• $\overset{\circ}{e}_x$ (column 7) is the life expectancy (or average number of remaining years of life) at the beginning of each age interval, that is, at each birthday x. (The $^\circ$ symbol above e indicates that deaths occurring at age x did not take place on the day of the xth birthday but, on average, between birthdays). Life expectancy is calculated by adding the remaining years of life of the ℓ_x survivors up to the terminal age of the table (age w) and by dividing this total by ℓ_x:

$$\overset{\circ}{e}_x = \frac{L_x + L_{x+1} + L_{x+2} + \ldots + L_w}{\ell_x} \tag{1.39}$$

From formula (1.38), we obtain

$$\overset{\circ}{e}_x = \frac{1}{\ell_x} \sum_{t=x}^{w} \frac{\ell_t + \ell_{t+1}}{2} = \frac{1}{\ell_x} \sum_{t=x}^{w} \ell_t - \frac{1}{2} = e_x - \frac{1}{2} \tag{1.40}$$

where e_x is obtained directly from the survivor function ℓ_x.

From the table in Appendix 1, the life expectancy on the day of the 25th birthday is

$$\overset{\circ}{e}_{25} = 49.28 \text{ years}$$

From the preceding definitions, the estimate of the mortality rate at age x is

$$\hat{\lambda}_x = \frac{d_x}{L_x} = \frac{2d_x}{\ell_x + \ell_{x+1}} = \frac{2\hat{q}_x}{2 - \hat{q}_x} \tag{1.41}$$

The equation is classically used to pass from the annual observed mortality rate to the annual probability of death on which the table is based. However, some authors prefer to calculate the annual probability of death directly without first estimating λ_x. This latter parameter is in fact obtained from $\hat{q}_x$ using (1.41).[2] In this approach, the estimate of $\hat{q}_x$ is obtained by dividing the number of deaths at age x observed in a given birth cohort by the number of persons at risk at the beginning of year t.

$$\hat{q}_x = \frac{d'_t + d''_{t+1}}{n_x(t+1) + d'_t} \tag{1.42}$$

Figure 1.5 presents the various elements required to calculate the annual probability of death on a Lexis diagram where d'_t is the number of deaths which have taken place in year t, and d''_{t+1} is the number of deaths which have taken place in year t + 1 in a cohort whose members have their xth birthday in year t; $n_x(t + 1)$ is the number of persons of age x in the population alive on 1 January of year t + 1.

In published tables, probabilities of death are usually smoothed, by using various analytical and graphical procedures, in order to attenuate the effect of random fluctuation [12].

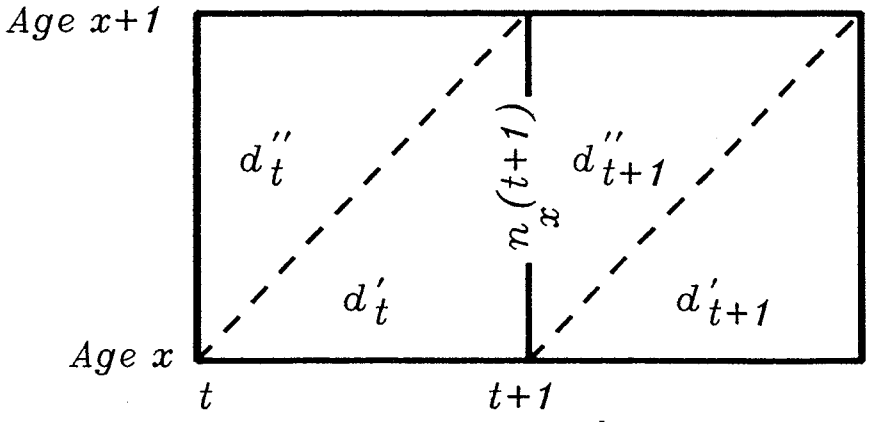

Figure 1.5 Representation of the data needed for the calculation of the annual probability of death on a Lexis diagram

[2] In the life table provided in Appendix 1, this formula gives a result which is correct only up to the first two decimal places as the published results have been smoothed.

National or regional tables made by statistical bureaux consider mortality over a short period of time (current life table), that is, as it is observed at a given time (or over one short period) across a range of ages. Mortality at various ages is estimated from different birth cohorts and the table which is constructed in this way thus refers to a fictitious force of mortality made up of the mortality experience of several successive birth cohorts. Cohort life tables can sometimes be constructed retrospectively; they describe the actual mortality experience for successive ages of a given birth cohort by combining the mortality information from several censuses. In the calculation of expected survival of a cohort which is followed for a relatively long time, the change of mortality of the general population must be taken into consideration. It is then advisable to apply the proper mortality rate to the different cohorts instead of using the cross-sectional force of mortality.

Classical models for survival distribution

It was seen on page that a survival distribution may be completely specified by the instantaneous mortality rate. There are several families of distributions which have played an important role in medical applications and whose definition depends on a parametric expression of $\lambda(t)$. Two of these families lead to a simple expression for the survival distribution $S(t)$:

- The Weibull distribution, for which

$$\lambda(t\,;\,\alpha,\,\theta) = \theta\alpha\,(\theta t)^{\alpha-1}$$

$$S\,(t\,;\,\alpha,\,\theta) = e^{-(\theta t)^{\alpha}} \qquad (1.43)$$

- The log-logistic distribution for which

$$\lambda(t\,;\,\alpha,\,\theta) = \frac{\theta\alpha(\theta t)^{\alpha-1}}{1+(\theta t)^{\alpha}}$$

$$S(t\,;\,\alpha,\,\theta) = \frac{1}{1+(\theta t)^{\alpha}} \qquad (1.44)$$

It is simple to estimate the parameters θ and α that define respectively the scale and the shape of the survival distribution by using the maximum likelihood method. The log-likelihood may be written

$$L(\alpha,\,\theta) = \sum_{i=1}^{n} \text{Log}\,[S(t_i\,;\,\alpha,\,\theta)] + \sum_{i=1}^{n} \delta_i\,\text{Log}\,[\,\lambda(t_i\,;\,\alpha,\,\theta)] \qquad (1.45)$$

Note that the exponential distribution discussed on page 16 is a particular case of the Weibull distribution with $\alpha = 1$. In fact, $\lambda(t;1,\theta) = \theta$ and $S(t) = e^{-\theta t}$. The Weibull

hazard rate may also be used to describe cancer incidence rate and could in particular be used in the framework of the multistage theory of carcinogenesis. In this context, α would be the number of stages needed for a cell to become malignant. The Weibull distribution is in fact the paradigmatic survival distribution and the starting point for the definition of more complex models which include prognostic factors $\mathbf{z} = (z_1, \ldots z_p)$.

First, by writing $\mu = -\text{Log}(\theta)$ and $\sigma = 1/\alpha$, it can be shown that the logarithm of survival duration $Y = \text{Log}(T)$ is $\mu + \sigma W$, where W has the same distribution as the minimum of a sample of continuous variables (extreme value distribution [13]). An analogous property holds for the log-logistic distribution, with W having in this case a distribution defined by the logistic probability density $e^W/(1 + e^W)^2$. A natural extension is to model the expectation μ of Log(T) with a linear function of the prognostic factors $\mathbf{z}$ ($\mu = \boldsymbol{\beta}\mathbf{z}$). This model supposes that the factors $\mathbf{z}$ act on survival by multiplying (or dividing) the mean duration of survival by a constant ($e^{\beta z}$).

A second approach more commonly used in medical applications starts from the observation that the hazard rates defined by the Weibull family are proportional. Writing $\mu = \theta^{\alpha}$, the hazard rate of the Weibull distribution becomes

$$\lambda(t) = \mu\alpha t^{\alpha-1}.$$

Considering that each prognostic factor acts on the instantaneous rate by multiplying (or dividing) it by a constant ($\mu = e^{\beta z}$), we obtain an example of a proportional rates model $\lambda(t) = \alpha t^{\alpha-1} e^{\beta z}$. The most general model of this class is the Cox model [14] defined by the relation

$$\lambda(t\ ,\ \mathbf{z}) = \lambda_0(t)\ e^{\beta z} \tag{1.46}$$

where $\lambda_0(t)$ is left unspecified.

Estimation of the parameter vector $\boldsymbol{\beta}$ in the model (1.46) is made difficult by the presence in its equation of the arbitrary function $\lambda_0(t)$. The likelihood of the observations given by formula (1.22) depends explicitly on λ_0 and is impossible to maximize without parameterizing $\lambda_0(t)$. However, as one of the goals of the Cox method is to specifically avoid such a parametric distribution, this approach would not be satisfactory. Full mathematical development of the likelihood function under the Cox model is beyond the scope of this text. It is however useful to understand the principles underlying its development in simple situations. In the framework of this model, only the ranks of the observed survival times are informative for the estimation of $\boldsymbol{\beta}$: as the rate $\lambda_0(t)$ is *a priori* an arbitrary function, it could be zero between two deaths. Another set of values of survival times with the same rank order should provide the same estimate of $\boldsymbol{\beta}$. More precisely, it is simple to check that a change in time scale defined by a monotonic function $\tau = u^{-1}(t)$ would give survival time τ_i with a distribution specified by the same model. The background hazard rate would simply be replaced by $\lambda_0[u(\tau)u'(\tau)]$. As a result, the estimate of $\boldsymbol{\beta}$ will be the vector of numerical values which maximize the probability that the ranks of the survival time are as observed.

Consider first two complete observations t_1 and t_2 for which **z** equals **u** and **v** respectively. The probability that the death of the subject having covariate value **u** comes first is

$$\Pr(t_1 < t_2) = \frac{e^{\beta u}}{e^{\beta u} + e^{\beta v}} \tag{1.47}$$

This intuitive result can be checked from the joint probability distribution of t_1, t_2. This principle may be generalized easily to m complete observations. If $\mathbf{u}_i \ldots \mathbf{u}_m$ are the values of **z** for the $m - i + 1$ subjects still alive just before the ith death, the probability that the death of the subject with covariate value $\mathbf{u}_i$ comes first is given by

$$\frac{e^{\beta u_i}}{e^{\beta u_i} + \ldots + e^{\beta u_m}}$$

The extension of this approach to n observations among which $n - m$ are censored leads to the likelihood

$V(\beta) = \Pr[(t_1 < \ldots < t_m)$ and ($t_i <$ censored observations in t_i, t_{i+1} ; $1 \leq i \leq m$)]

that is

$$V(\beta) = \prod_{i=1}^{m} \frac{e^{\beta u_i}}{\sum_{j \in R_i} e^{\beta u_j}} \tag{1.48}$$

where

- i indexes the m dates of death t_i ranked in increasing order;
- $\mathbf{u}_i$ is the covariate value of the subject who died at time t_i; and
- R_i is the set of subjects still at risk at time t_i of the ith death.

The log-likelihood is

$$L(\beta) = \text{Log}[V(\beta)] = \sum_{i=1}^{m} \left\{ \beta u_i - \text{Log}\left[\sum_{j \in R_i} e^{\beta u_j} \right] \right\} \tag{1.49}$$

The estimate of β is the value $\hat{\beta}$ which maximizes $L(\beta)$, obtained by equating to zero its derivatives with respect to the coordinates β_k of β:

$$C_k(\beta) = \frac{\partial L(\beta)}{\partial \beta_k} = \sum_{i=1}^{m} [u_{ki} - \bar{u}_{ki}(\beta)] \tag{1.50}$$

where

$$\bar{u}_{ki}(\beta) = \frac{\sum_{j \in R_i} u_{kj} e^{\beta u_j}}{\sum_{j \in R_i} e^{\beta u_j}} \tag{1.51}$$

is the average of the covariate values of the subjects still at risk just before time t_i weighted by their respective relative rates. C_k is known as the *score function* and is used to construct the *score test* described below.

The observed information matrix having as elements the negative of the second derivatives of the log-likelihood $L(\boldsymbol{\beta})$ (see page 17)

$$I_{k\ell}(\boldsymbol{\beta}) = -\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_k\,\partial\beta_\ell} = \sum_{i=1}^{m} \iota_{k\ell i}(\boldsymbol{\beta}) \tag{1.52}$$

where

$$\iota_{k\ell i}(\boldsymbol{\beta}) = \frac{\sum_{j\in R_i} u_{ki}\,u_{\ell i}\,e^{\boldsymbol{\beta}u_j}}{\sum_{j\in R_i} e^{\boldsymbol{\beta}u_j}} - \bar{u}_{ki}(\boldsymbol{\beta})\,\bar{u}_{\ell i}(\boldsymbol{\beta}) \tag{1.53}$$

will be used to carry out the maximization using the Newton-Raphson method. This algorithm constructs a sequence of $\boldsymbol{\beta}$ values which lead by successive iterations to the value $\hat{\boldsymbol{\beta}}$ where $\mathbf{C}(\boldsymbol{\beta}) = \mathbf{0}$:

$$\boldsymbol{\beta}_0 = \mathbf{0}\ ; \quad \boldsymbol{\beta}_n = \boldsymbol{\beta}_{n-1} + \mathbf{I}^{-1}(\boldsymbol{\beta}_{n-1})\ \mathbf{C}(\boldsymbol{\beta}_{n-1})$$

This method is used by most computer programs, which estimate the Cox model. The inverse of the observed information matrix $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$ provides an estimate of the covariance matrix of the maximum likelihood estimates of the parameter vector $\boldsymbol{\beta}$.

The application of these principles leads to unmanageable formulae when the number of deaths d_i occurring at time t_i exceeds more than a few. The likelihood may then be approximated [15] by

$$V(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{e^{\boldsymbol{\beta}s_i}}{\left(\sum_{j\in R_i} e^{\boldsymbol{\beta}u_j}\right)^{d_i}} \tag{1.54}$$

where :

$$\mathbf{s}_i = \sum_{j=1}^{d_i} \mathbf{u}_j \tag{1.55}$$

is the sum of the covariate values of the d_i subjects dying at date t_i. Expressions (1.49) (1.50) and (1.52) then become

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left[\boldsymbol{\beta}\mathbf{s}_i - d_i \operatorname{Log}\left(\sum_{j\in R_i} e^{\boldsymbol{\beta}u_j}\right)\right] \tag{1.56}$$

$$C_k(\boldsymbol{\beta}) = \sum_{i=1}^{m} [s_{ki} - d_i \bar{u}_{ki}(\boldsymbol{\beta})] \tag{1.57}$$

$$I_{k\ell}(\boldsymbol{\beta}) = \sum_{i=1}^{m} d_i \upsilon_{k\ell i}(\boldsymbol{\beta}) \tag{1.58}$$

It is worth noting that the above approximation may be obtained directly from the assumption that the hazard function is constant within intervals, as is the case with the actuarial method. The estimation of the nuisance parameters λ_i, $1 \leq i \leq m$ and their substitution in the likelihood lead to (1.54) [16].

Several tests for comparison of survival distributions may then be obtained from the likelihood or from the score function [17]. The practical aspects of these methods are described in detail in Chapter 4. Here, we simply note that the test of the hypothesis $\boldsymbol{\beta} = 0$ by the *likelihood ratio test* is based on the statistic

$$T_1 = 2[L(\hat{\boldsymbol{\beta}}) - L(0)] \tag{1.59}$$

which has a χ^2 distribution with r degrees of freedom, dimension of $\boldsymbol{\beta}$, under the null hypothesis $\boldsymbol{\beta} = 0$. The *score test* is based on the evaluation of the score function at $\boldsymbol{\beta} = 0$ which should be close to zero under the null hypothesis since, at the true value of $\boldsymbol{\beta}$, the derivative of L should be close to zero, its value at the maximum likelihood estimate. After standardization by its variance, the score statistic is written

$$T_2 = \mathbf{C}(0)' \, \boldsymbol{\Gamma}^{-1} \, \mathbf{C}(0) \tag{1.60}$$

and also has a χ^2 distribution with r degrees of freedom under the null hypothesis. The *Wald test* is based on the evaluation of $\hat{\boldsymbol{\beta}}$ itself which should be close to zero under the null hypothesis. After standardization by its variance, we obtain the statistic

$$T_3 = \hat{\boldsymbol{\beta}} \, \mathbf{I} \, \hat{\boldsymbol{\beta}}' \tag{1.61}$$

which also has the χ^2 distribution with r degrees of freedom under the null hypothesis.

Similar tests exist when the null hypothesis does not completely specify the value of $\boldsymbol{\beta}$. In this context the null hypothesis is usually defined by one or several constraints on the coordinates of $\boldsymbol{\beta}$ (e.g., $\beta_i = 0$). T_1 and T_2 are then calculated by replacing zero in (1.59) and (1.60) by the maximum likelihood estimate of $\boldsymbol{\beta}$ under the null hypothesis. When the null hypothesis specifies that some coordinates of $\boldsymbol{\beta}$ are zero, this approach is equivalent to setting the other coordinates to their maximum likelihood estimates under the null hypothesis. In this case, the test T_3 is restricted to the coordinates being tested. The number of degrees of freedom of these three tests is equal to the number of coordinates of $\boldsymbol{\beta}$ which specify the null hypothesis. Applications of this methodology are presented in Chapter 4, page 268.

In the preceding discussion, we have concentrated on the properties of the proportional hazards model which enable group comparisons to be carried out. In other words, the problem of estimating β has been seen as more important by considering λ_0 as a nuisance function. In practice, it is often necessary to provide an estimate of the survival distribution for some given value of the covariate $\mathbf{z}$. The same principle as used previously for the Kaplan-Meier procedure (1.34) may be used here, taking into account the fact that the subjects are not at the same risk of death at the time when one of them dies. For a subject with covariate $\mathbf{z}_j$, this risk is characterized by the relative rate of mortality $\hat{\theta}_j = e^{\hat{\beta}\,\mathbf{z}_j}$ where $\hat{\beta}$ is the maximum likelihood estimate of β. Thus, in estimating $\lambda_0(t_i)$, each subject at risk at that time will account for θ_j units instead of one. Therefore the cumulative rate and survival distribution will be given by:

$$\hat{\Lambda}_0(t) = \sum_{t_i \le t} \frac{d_i}{\sum_{j \in R_i} e^{\hat{\beta}\mathbf{z}_j}} \tag{1.62}$$

$$\hat{S}_0(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{\sum_{j \in R_i} e^{\hat{\beta}\mathbf{z}_j}} \right) \tag{1.63}$$

The estimate of the survival distribution for a given value of $\mathbf{z}$ is then obtained from the fact that the hazard rates are proportional. Therefore

$$\hat{S}(t\,,\,\mathbf{z}) = \hat{S}_0(t)^{e^{\hat{\beta}\mathbf{z}}} \tag{1.64}$$

Interactive risks

Competing risks

We can see from the preceding sections that it is relatively simple to estimate the distribution of survival times while taking account of information provided by incomplete or censored observations. The method which has been discussed depends on the assumption of independence between risk of death and the mechanism which leads to censored observations. In Chapter 4, we will discuss situations in which this assumption can be questioned, most notably when not all the members of the cohort are followed up in the same way. However, the assumption is usually quite reasonable. In fact, the survival time which corresponds to a withdrawal is clearly defined. It could be observed by prolonging the study; it would then be possible to check statistically that censored observations are not associated with

either longer or shorter survival times, or equivalently, that survival does not change with time of diagnosis.

The problem presents itself in different terms when our aim is no longer to measure overall mortality but to establish the impact of a specific cause of death, usually corresponding to the diagnosis by which the cohort is defined. Therefore, we should consider that the individuals under observation are subject to other risks of death besides the one which forms the basis of the study. Since the realization of one of the risks excludes the possibility of the realization of the others, the risks are then said to be *competing*. It is tempting to consider deaths due to other causes as censored observations; survival related to the cause under study could then be estimated by simply using the method described earlier. Discussion of the practical problems raised by this approach will be left until Chapter 4; in this section we will treat succinctly the theoretical aspects of competing risks and problems raised by the definition and especially by the evaluation of the independence of risks.

The crude probability of death from a given cause is the probability of death from this cause in the presence of other risks.

The net probability is the probability of death from the given cause when all other risks of death have been eliminated.

The partial crude probability is defined as the probability of death from a given cause when the potential effects of another cause (or group of causes) are eliminated.

The third concept is obviously crucial to competing risk theory. Its recognition probably goes back to the controversy over the efficiency of the smallpox vaccination; in 1760, Bernouilli [18], d'Alembert [19] and other authors were each attempting to evaluate the consequences of eliminating the risk of death from smallpox on the composition and life expectancy of the population. Today, it is relatively straightforward to construct life tables based on probabilities of death after a cause has been eliminated, in order to estimate the cause's impact on life expectancy [20]. For example, it has been calculated that, if mortality from cancer (all sites combined) was totally eliminated, the consequent lengthening of the expectation of life would be about two years. However, these statistics only tell part of the truth: the improvement in survival for patients who suffer from the disease is much more significant both qualitatively and quantitatively. Our goal is to define the survival probability as a measure of the consequence of a specific disease. This concept corresponds better to the net probability, that is, the survival probability from cancer in the absence of mortality from other causes.

The data which are generally available for the study of mortality by cause can be summarized by the three variables T, Δ and $\mathbf{z}$, where T is survival time, Δ the indicator of the cause and $\mathbf{z}$ the vector of covariables which influence the risk of death. Δ varies between 1 and m + 1 when m causes of death are studied and the number m + 1 indicates withdrawals other than those due to death. If the withdrawals are independent of death, the same argument that was used on page 19 shows that the contribution of observation t_i, $\mathbf{z}_i$ of subject i to the likelihood may be written

- $\lambda_j(t_i, \mathbf{z}_i)\ S(t_i, \mathbf{z}_i)$ if death resulted from the jth cause,
- $S(t_i, \mathbf{z}_i)$ if the observation is censored at t_i,

where

$$S(t_i, \mathbf{z}_i) = e^{-\sum_{i=1}^{m} \Lambda_j (t_i, \mathbf{z}_i)}$$

formulae in which λ_j and Λ_j are the instantaneous and cumulative rates of death for the jth cause. The likelihood is thus a product of m terms of the form

$$\prod_{j=1}^{m} \left[\prod_{i=1}^{n} \lambda_j (t_i, \mathbf{z}_i)^{\delta_j} e^{-\Lambda_j(t_i, \mathbf{z}_i)} \right]$$

where $\delta_j = 1$ if death results from the jth cause ($\Delta = j$) and $\delta_j = 0$ otherwise ($\Delta \neq j$); each of these terms represents the likelihood which would be obtained in the study of the jth cause of death if all deaths from other causes could be considered to be independent censored observations.

From this discussion it is clear that the methods previously described for estimating λ and assessing the effect of covariates $\mathbf{z}$ on mortality are appropriate in the presence of competing risks. It is also clear that they are describing a particular risk of mortality within a complex of risk interactions, rather than the risk that would prevail if one or several causes of mortality were eliminated. It was previously stated that it is generally valid to assume that the instantaneous rate of death observed in the presence of censored observations is that which would prevail if the censored observations were eliminated (or completed). However, this assumption may well be questionable when a specific cause of death is being studied in the presence of other risks of death. Indeed, it is very likely that the removal of one cause of death would have noticeable consequences for the risk of death from one or several other causes. We should remember that some individuals can be subject to increased risk of death from several diseases, either because the diseases have similar etiology or because they are linked to the same innate susceptibility. When one of these diseases tends to occur earlier in life or to be associated with a shorter survival probability, it will more often be the cause of death. Any action taken to eliminate one disease or to reduce its associated mortality will tend, therefore, to modify the instantaneous mortality rate of associated competing diseases. For example, it has been suggested that coal miners who survived pneumoconiosis were subject to a reduced risk of lung cancer, as a result of the selection of the most resistant. If this assumption is true, an improvement in the treatment for pneumoconiosis resulting in better survival could lead to an increase in the lung cancer mortality rate. On the other hand, a measure aimed at reducing exposure to coal dusts might result in a decrease in the two risks under consideration.

This example shows that the probabilities of death that are calculated in a given context of risk interaction need not correspond to the instantaneous rates which would prevail if other causes of death were eliminated; it also shows that the

direction of the interaction between risks can be modified by the intervention used on one of the risks.

In practice, the existence of a statistical link between competing risks is difficult to identify, and the strength of a link is hard to measure. Independence of risks cannot be verified simply from survival data, since, by definition, the occurrence of death from cause j excludes the possibility of death from all other causes. In fact, data of the type (T, Δ, **z**) do not contain the information necessary to assess risk interaction. Moreover, it has been shown that, for a given set of such observations, a compatible model of independent risks can always be constructed [21-23]. Some empirical models have been proposed to assess the interaction between risks of death using additional information such as concomitant causes of death [24]. It is however possible that the concomitant causes of death have a direct link with the disease primarily responsible for death or are a consequence of its diagnosis or treatment. In such a situation, information on concomitant causes is of little use in assessing risk dependence and may even lead to a biased evaluation. So far, these models have not proved to be usable. By definition, there is absolutely no information which could allow the correct estimation of the joint distribution of potential survival times for multiple causes. Consequently, the formal specification of this joint distribution cannot be verified and is therefore of little practical value.

For lack of a better alternative, we therefore restrict our discussion to the net probabilities of mortality (and consequently the net survival) with respect to a given environment of risks, while remaining aware of the limitations in their interpretation (see Chapter 4). These difficulties probably explain why life tables routinely published by official statistical services only rarely present net probabilities, and generally restrict themselves to crude probabilities by cause or group of causes.

Relationship between incidence, mortality, survival and prevalence

The most widely available information describing the risk of cancer as a function of space and time are age- and sex-specific mortality statistics. In many countries, these data have been recorded systematically over long time periods for most cancer sites. In some countries, they may even be available for small geographical areas such as census or administrative districts. However, mortality data are frequently of uneven quality and inadequate for the descriptive study of site-specific cancer occurrence.

Information on cancer incidence is provided by the number of new cases of cancer occurring each year, and is generally available from cancer registries. This information is much more reliable than mortality statistics but, except for Nordic countries, it is limited in space and time. Cancer registries may also have information on survival of cancer patients when they have established routine procedures of follow-up (see Chapter 4). Thus, in a region where cancer incidence is recorded, it is possible to estimate the empirical relation which links incidence, mortality and survival and then use this observed relation to estimate cancer incidence in regions

where cancer registries do not exist [25-28]. The goal of the present section is to give some insight into the theoretical relationships which link incidence, mortality and survival and to assess the feasibility of estimating one of them from the other two.

This discussion will also introduce the concept of *prevalence*, the number (or proportion) of subjects with a specific condition in a population at a given time. This measure of disease frequency depends on incidence and duration of disease, that is, on survival probability, in the case of a 'non reversible' disease such as cancer, for which an incident case is considered prevalent up to death, even if treatment is effective. In contrast to incidence, which is a concept with a natural link to age and therefore logically described in the context of birth cohorts, prevalence is related to the time period of observation. Incidence is better assessed in a *longitudinal* study, whereas prevalence is measured on a cross-sectional basis. For this reason, the relations between incidence, survival and prevalence are simple only in *stationary* populations in which longitudinal and cross-sectional measures are identical. In this section, the meaning of the term 'stationary' will be explained and the usage of the relationship 'prevalence is the product of incidence and the duration of the disease' will be discussed.

Although it is rarely estimated in cancer registries, prevalence is important to public health planning. When incidence data are not systematically recorded (as for HIV and diabetes), it is often from prevalence surveys that incidence will be estimated.

In order to understand the relationships between these concepts, a fictitious cohort of size ℓ_0 born in year $t = u_0$ and subject to cancer incidence rate λ_y is described in Table 1.4. We assume that the number of years L_y lived without cancer by each individual of the cohort is known for each age y.

In the absence of migration, the number of cancer deaths at age x occurring in year $t = u_0 + x$ among incident cases in the cohort is given by the formula

$$d_x(t) = \sum_{y=0}^{x} L_y \lambda_y \left[S_y\left(x - \frac{1}{2}\right) - S_y\left(x + \frac{1}{2}\right) \right] \qquad (1.65)$$

where $S_y(x)$ is the probability that a subject diagnosed at age y survives to age x; $S_y\left(x - \frac{1}{2}\right) - S_y\left(x + \frac{1}{2}\right)$ is then the probability that death occurs at age x. Similarly the cases of age x prevalent in the population during year $t = u_0 + x$ come from the cohort born in year u_0. Their number is given by the formula

$$n_x(t) = \sum_{y=0}^{x} L_y \lambda_y S_y(x) \qquad (1.66)$$

which shows that they are calculated from the cases in the cohort diagnosed before age x and still surviving at age x.

Table 1.4 Incident cases, deaths and prevalent cases

Age	Time period					
	$t = u_0$	$t = u_0 + 1$	.....	$t = u_0 + y$	.....	$t = u_0 + x$
0	$L_0\ \lambda_0$					
1	.....	$L_1\ \lambda_1$				
.						
.						
.						
y	.....	.....	.....	$L_y\ \lambda_y$	.....	$k_y(t),\ d_y(t),\ n_y(t)$
.						
.						
.						
.						
x	.....	.....	.....	.....	.....	$k_x(t)=L_x\lambda_x,\ d_x(t),\ n_x(t)$

If ℓ_0 is equal to 100 000, the figures are obtained per 100 000 births in year u_0. The numbers of deaths and cases (incident or prevalent) actually observed are obtained by multiplying the figures in Table 1.4 by the actual number of births $B(u_0)$ in year u_0.

The figures in column $u_0 + x$ of Table 1.4 are generated by successive birth cohorts and depend on factors which change with time, either period– or cohort-wise. Most often, survival probability changes with period, whereas age-specific incidence depends substantially on birth cohort. Each line in column $u_0 + x$ must therefore be calculated from the parameters L_y, λ_y and $S_y(x)$ by taking their evolution over time into account.

The prevalence at age x or *age-specific prevalence* is the proportion $p_x(t) = \dfrac{n_x(t)}{\ell_x(t)}$ where $\ell_x(t)$ is the number of survivors at time t among the individuals of the cohort born in u_0. This figure depends only on the risk environment experienced by this cohort up to age x while the overall prevalence depends on the experience of several successive cohorts:

$$p(t) = \frac{\sum_{x=0}^{g} n_x(t)}{\sum_{x=0}^{g} \ell_x(t)} \qquad (1.67)$$

where each term in the above sums is generated by different birth cohorts for each of the g age groups.

A deeper understanding of the relationship linking the survival, incidence and prevalence requires more detailed modelling which involves explicit mathematical definitions. Let

- $\lambda(t,x)$ be the incidence rate at age x and time t for the disease under study,
- $\mu(t,x)$ be the mortality rate at age x from causes other than the disease, for individuals without the disease,
- $\nu_y(t,x)$ be the mortality rate at age x of patients diagnosed with the disease at age y and time t,
- $\beta(u)$ be the average annual number of births, considered to be a Poisson process and depending on year u.

It is then possible to write formulae similar to (1.65) and (1.66) as well as formulae for the number of individuals with and without a particular health condition living in the population at time t.

The probability of being alive and free of cancer at age x and time t for an individual born in year $u = t - x$ can be written

$$H(t,x) = e^{\left\{ -\int_0^x [\mu(u+y,y)+\lambda(u+y,y)]\, dy \right\}} \tag{1.68}$$

This expression shows that to be alive and without cancer, an individual must escape both the force of cancer incidence and the force of mortality in the interval between birth and age x. Therefore, the number of individuals of age x without cancer at time t is on average

$$h(t,x) = \beta(t-x)\, H(t,x) \tag{1.69}$$

In the same way, the probability that an individual with cancer is alive at age x and time t may be written

$$\pi(t,x) = \int_0^x H(u+y,y)\, \lambda(u+y,y)\, S_y(u+y,x)\, dy \tag{1.70}$$

where :

$$S_y(u+y,x) = e^{-\int_y^x \nu_y(u+y,z)\, dz} \tag{1.71}$$

is the probability of surviving up to age x when diagnosed at time u+y and age y. The number of individuals of age x who have been diagnosed with the disease in the population is therefore at time t

$$n(t,x) = \beta(t-x)\, \pi(t,x) \tag{1.72}$$

The prevalence p(t), the proportion of individuals with the disease living in the population at time t, is then obtained in a simple way from the ratio of the number

of individuals diagnosed with the disease to the number without the disease (prevalence odds). This ratio can be written from (1.69) and (1.72):

$$\frac{p(t)}{1 - p(t)} = \frac{\int_0^\infty \beta(t - x)\, \pi\,(t, x)\, dx}{\int_0^\infty \beta(t - x)\, H(t, x)\, dx} \tag{1.73}$$

In a stationary population where λ, μ and β are all independent of t, the above formulae lead to simple relationships. It is however important to realize how restrictive the stationary hypothesis is; it implies that the birth rate, the cancer incidence rate, the cancer survival probability and mortality rate from other causes all remain constant with time. We will nevertheless give the main results which are obtained under the stationary hypothesis, since most epidemiological textbooks define prevalence in this situation.

In a stationary population, the various rates do not depend on time so that formulae (1.68) and (1.70) simplify to

$$H(x) = e^{\left\{-\int_0^x [\,\mu(y) + \lambda(y)\,]\, dy\right\}} \tag{1.74}$$

$$\pi(x) = \int_0^x H(y)\, \lambda(y)\, S_y(x)\, dy \tag{1.75}$$

The integrals of these functions which no longer depend on t are respectively

$$\int_0^\infty H(x)\, dx = E_0(X) \tag{1.76}$$

which is the mean duration of life for individuals who remain without the given disease over their lifetime, since H(x) is their survival distribution. By exchanging the order of integration

$$\int_0^\infty \pi\,(x)\, dx = \int_0^\infty [H(y)\, \lambda(y) \int_y^\infty S_y(x)\, dx]dy \tag{1.77}$$

which is, except for division by $R = \int_0^\infty H(y)\, \lambda(y)\, dy$, the mean duration of disease for those who have contracted it. In other words

$$\int_0^\infty \pi\,(x)\, dx = R\ E_1(X - Y) = R\ E_1(V) \tag{1.78}$$

is the product of the crude risk of disease [see (1.41)] and the mean survival of the patients which, in the case of a 'non-reversible' disease (see above), is the duration of the disease.

Since β is constant, formula (1.73) simplifies to become the ratio of (1.78) and (1.76), that is

$$\frac{p}{1 - p} = \frac{R\, E_1(V)}{E_0(X)} \tag{1.79}$$

Furthermore, the ratio of $R/E_0(X)$ may be written as a function of H and λ, using their respective definitions:

$$\frac{R}{E_0(X)} = \frac{\int_0^\infty H(x)\,\lambda(x)\,dx}{\int_0^\infty H(x)\,dx} = \bar{\lambda} \tag{1.80}$$

This last result leads to the classical statement that 'prevalence is the product of incidence and the duration of the disease' since we can write from (1.79) and (1.80)

$$\frac{p}{1-p} = \bar{\lambda}\,E_1(V) \tag{1.81}$$

The above approach to the concept of prevalence is taken from a paper by Keiding [38]. We will also describe the method of Verdecchia and Capocaccia [39], who showed that under certain conditions the information needed to carry out the various calculations is contained in the net probability distributions of age at the occurrence of cancer and at death from cancer.

Let X, Y and V be respectively the age at death from the disease of interest, the age at diagnosis and the survival time up to death from cancer. We may then write

$$X = Y + V$$

Consequently, the age at death from cancer has the probability density

$$d(x) = \int_0^x i(y)\,s_y(x-y)\,dy \tag{1.82}$$

where i and s_y are the probability densities of Y and V. As explained previously, this function is known only from the corresponding incidence rates because of censoring. (1.82) must therefore be written

$$d(x) = \int_0^x \lambda(y)\,e^{\left\{-\int_0^y [\lambda(u)+\mu(u)]\,du\right\}}\,v_y(x-y)\,e^{\left\{-\int_y^x [v_y(u-y)+\mu_y^*(u)]\,du\right\}}\,dy \tag{1.83}$$

where

- $v_y(v)$ is the mortality rate from the disease v years after diagnosis for a patient diagnosed at age y; and
- $\mu_y^*(u)$ is the mortality rate from causes other than the disease at age u for a patient diagnosed at age y.

Denoting by $\mu_y^e(x)$ the difference $\mu_y^*(x) - \mu(x)$ which represents the excess death from other causes for a patient diagnosed at age y, we can write

$$d(x) = \int_0^x \lambda(y)\,e^{-\Lambda(y)}\,v_y(x-y)\,e^{\left\{-\int_y^x [v_y(u-y)+\mu_y^e(u)]\,du\right\}}\,dy\,e^{-\int_0^x [\mu(y)\,dy]} \tag{1.84}$$

The rate $\nu^*_y(u-y) = \nu_y(u-y) + \mu^e_y(u)$ is the mortality rate experienced by patients after diagnosis taking into account the excess (or the reduction) of the hazard of death from other causes. This rate is connected to the relative survival rate (see Chapter 4, page 231), whereas ν_y corresponds to net survival. If most deaths of cancer patients are in fact certified as due to cancer, we can replace ν_y by ν^*_y in (1.84) and obtain

$$\delta(x)\, e^{-D(x)} = \int_0^x \lambda(y)\, e^{-\Lambda(y)}\, \nu^*_y(x-y)\, e^{[-N^*_y(x-y)]}\, dy \qquad (1.85)$$

where δ is the mortality rate for the given cancer and D, Λ and N^* denote the cumulative rates associated respectively with δ, λ and ν^*. Formula (1.85) results from the fact that $d(x) = \delta(x)\, e^{-M(x)}$ since $M(x) = D(x) + \int_0^x \mu(y)\, dy$ is the cumulative mortality rate from all causes.

The relationship initially given for the crude probability density in (1.82) therefore remains true for net density in (1.85) if net and relative survivals are identical. However, this relationship is only simple when survival probability does not depend on age at diagnosis or depends on it according to a simple model. In this situation, the relationship between mortality, incidence and survival distribution can be written as a convolution and corresponding mathematical tools are available to carry out its analysis.

The probability $\pi(x)$ of having cancer and still being alive at age x may be calculated in the same way. Thus

$$\pi(x) = \int_0^x \lambda(y)\, e^{-\Lambda(y)}\, e^{[-N^*_y(x-y)]}\, dy\; e^{-\int_0^x \mu(y)\, dy} \qquad (1.86)$$

The number of cancer cases of age x in the corresponding birth cohort is $n(x) = B\,\pi(x)$ where B is the number of births in this cohort. Furthermore, the number of survivors without cancer of age x is $h(x) = B\,H(x)$, where H(x) is obtained from (1.74). Therefore the age-specific prevalence is given by

$$\frac{p(x)}{1-p(x)} = \frac{n(x)}{h(x)} = \frac{\pi_c(x)}{1-R_i(x)} \qquad (1.87)$$

where $\pi_c(x)$ is the first integral of the right-hand side of formula (1.86) and $R_i(x)$ is the net risk of disease before age x.

Denoting the net probability densities of age at death, age at diagnosis and survival time by d, i and s_y, the following two equations can be written:

$$d(x) = \int_0^x i(y)\, s_y(x-y)\, dy \qquad (1.88)$$

$$\pi_c(x) = \int_0^x i(y)\, S_y(x-y)\, dy \qquad (1.89)$$

The derivative with respect to x of $\pi_c(x)$ is by definition

$$\pi'_c(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} [\pi_c (x + \Delta x) - \pi_c (x)]$$

$$= \lim_{\Delta x \to 0} \int_0^x i(y) \frac{S_y(x + \Delta x - y) - S_y(x - y)}{\Delta x} dy + \frac{1}{\Delta x} \int_x^{x+\Delta x} i(y) S_y(x + \Delta x - y) dy$$

which may be written, using the rules of calculus and the fact that the derivative of S_y is $-s_y$:

$$\pi'_c(x) = -\int_0^x i(y) s_y(x - y) dy + i(x)$$

$$= -d(x) + i(x)$$

and therefore

$$\pi_c(x) = \int_0^x i(y) dy - \int_0^x d(y) dy \tag{1.90}$$

which expresses the fact that the numerator of the prevalence odds in (1.87) is the difference between the net risk of having cancer before age x and the net risk of dying from this cancer before age x. Thus the age-specific prevalence of cancer may be obtained from:

$$\frac{p(x)}{1 - p(x)} = \frac{R_i(x) - R_d(x)}{1 - R_i(x)} \tag{1.91}$$

where $R_d(x)$ denotes the net risk of dying from the given cancer.

This result is obtained under some fairly general assumptions about the interactions between the risk of dying from the given cancer and the risk of dying from other causes. When the cancer risk is not stationary, the formula 1.91 must be used in conjunction with the modelling of the time trend in incidence and mortality by birth cohort (see page 189)

Preston has provided a useful and intuitive approach to calculate prevalence when the population is not stationary [40].

Bibliographical notes

Mathematical arguments used in basic epidemiological texts, and in particular those which form the theoretical basis of descriptive epidemiology, are often approximate, and for a good reason: a satisfactory mathematical approach, based on the statistical analysis of stochastic processes, quickly leads to advanced mathematics [29] in even the simplest situations. Moreover, this level of sophistication is

rarely required to meet the real problems of descriptive epidemiology, which are more often of a different kind. The approach we have taken in this first chapter is similar to that used in demography [30], where data of this type were first analysed rigorously. The modern trend in mathematical statistics is to treat the analysis of censored data using the concepts of stochastic processes, which provide very general results on the convergence and the speed of convergence of the estimates. It is not surprising that, in medical research, most effort in this direction has been in the context of clinical trials, because these often have relatively few subjects and the validity of the statistical conclusions is a paramount requirement. Readers interested in this approach can find the necessary concepts in Hill et al. [31], particularly the appendix. Anderson [32] has published a fairly complete and non-technical introduction to this method. To the extent that the fundamental principles of descriptive epidemiology do not differ from those of cohort studies, several sections of chapters 2, 3 and 4 in the book by Breslow and Day [11] make profitable reading, and give a more complete bibliography of the various formalizations.

Chapter 9 in Pressat [30] provides a complete presentation of the concepts involved in the life table. The estimation of the life table is discussed in depth by Chiang [20] in chapter 9. Classical survival models are described in detail in Kalbfleisch and Prentice [33] [see pages 21-30] and in Cox and Oakes [34] [see pages 13-28]. Since Cox [14] was first published, the proportional hazards model has had so many applications, that even an abridged list would be difficult to provide. References [35,36 and 37] provide a clear discussion of its application in epidemiology.

The theory of competing risks is discussed in Chiang [20], chapter 2; a monograph has also been written on this subject [41]. Makeham [42] is generally recognized as having originated the concept of multiple decremental forces, from which the essentially similar idea of latent survival time was largely derived. In this approach, the observed survival of a subject is the smallest of the [unobserved] latent survival times, with each of these times corresponding to the causes of death under study. This approach is described in the monograph by David and Moeschberger [41], and discussed in reference [23]. The problem of estimating mortality when the competing risks cannot be assumed to be independent is reviewed in an article by Duchene [43].

The concept of prevalence and its calculation has been discussed by many authors. The texts by MacMahon and Pugh [44], and Kleinbaum and co-workers [45] can be consulted, and an article by Freeman and Hutchison [46] gives a detailed overview. Reference [38] also provides a full bibliography on the subject.

REFERENCES

[1] Day NE, Muñoz N. Oesophagus. *In* D Schottenfeld, JF Fraumeni (eds) : *Cancer Epidemiology and Prevention,* (2nd Edition). Philadelphia, WB Saunders, in press

[2] *IARC Monographs on the Evaluation of the Carcinogenic Risk of Chemicals to Humans,* Vol. 38, *Tobacco Smoking.* Lyon, IARC, 1986

[3] DOLL R, PETO R. The causes of cancer. *J Epidemiol Commun Health* 1981, **66**

[4] BRESLOW NE, ENSTROM JE. Geographic correlations between cancer mortality rates and alcohol – Tobacco consumption in the United States. *J Nat Cancer Inst* 1974, **53** : 631-39

[5] JENSEN OM. *Cancer morbidity and causes of death among Danish brewery workers.* 1980, Lyon, International Agency for Research on Cancer

[6] ARMSTRONG BK, DOLL R. Environmental factors and cancer incidence and mortality in different countries with special reference to dietary practices. *Int J Cancer* 1975, **15** : 617-31

[7] WILLETT WC, STAMPFER MJ, COLDITZ GA et al. Dietary fat and risk of breast cancer. *N Engl J Med* 1987, **316** : 22-8

[8] BEESE DH, Ed. *Tobacco consumption in various countries.* London, Tobacco Research Council, 1972, research paper 6, 3rd edition

[9] WATERHOUSE J, MUIR C, SHANMUGARATNAM K, POWELL J. *Cancer incidence in five continents*, Vol. IV (IARC Scientific Publications No 42), Lyon, IARC, 1982

[10] LYON JL, GARDNER JW, WEST DW. Cancer in Utah : Risk by religion and place of residence. *J Nat Cancer Inst* 1980. **65** : 1063-71

[11] BRESLOW NE, DAY NE. *Statistical methods in cancer research. Vol II : The Design and analysis of cohort studies.* (IARC Scientific Publications No. 82), Lyon, IARC, 1987

[12] DUCHÊNE J. *Un essai de modélisation de la répartition des décès selon l'âge et la cause dans les pays industrialisés.* Louvain-la-Neuve, Cabay, 1980

[13] JOHNSON NL, KOTZ S. *Distribution in statistics. Continuous univariate distribution.* New York, Wiley, 1970, chapter 21

[14] COX DR. Regression models and life tables. *J Roy Stat Soc B* 1972, **34** : 187-220

[15] PETO R. Contribution to the discussion of paper by DR Cox. *J Roy Stat Soc B* 1972, **34** : 205-7

[16] BRESLOW NE. Covariance analysis of censored survival data. *Biometrics* 1974, **30** : 89-99

[17] RAO CR. *Linear statistical inference and its applications.* New York, Wiley, 1973, pp. 417-8

[18] BERNOULLI D. Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir. *Mémoire de l'Académie Royale des Sciences,* 1760, pp. 1-45

[19] D'ALEMBERT. Sur l'application du calcul des probabilités à l'inoculation de la petite vérole. Opuscules II, 1761, pp. 26-95

[20] CHIANG CL. *Introduction to stochastic process in biostatistics.* New York, Wiley, 1968

[21] COX DR. The analysis of exponentially distributed life times with two types of failures. *J Roy Stat Soc B* 1959, **21** : 411-21

[22] TSIATIS A. A non-identifiability aspect of the problem of competing risks. *Proc Nat Acad Sci USA* 1975, **72** : 20-2

[23] PRENTICE RL, KALBFLEISCH JD, PETERSON AV, FLOURNOY N, FAREWELL TT, BRESLOW NE. The analysis of failure times in the presence of competing risks. *Biometrics* 1978, **34** : 541-54

[24] WONG O. A competing risk model based on the life table. Procedures in epidemiological studies. *Int J Epidemiol* 1977, **6** : 153-60

[25] DOLL R. The geographical distribution of cancer. *Br J Cancer* 1969, **23** : 1-8

[26] BENHAMOU E, LAPLANCHE A, WARTELLE M, FAIVRE J, GIGNOUX M, MÉNÉGOZ F, ROBILLARD J, SCHAFFER P, SCHRAUB S, FLAMANT R. *Incidence des cancers en France, 1978-1982.* Paris, Les Éditions INSERM, 1990

[27] JENSEN OM, ESTÈVE J, MØLLER H, RENARD H. Cancer in the European Community and its member states. *Eur J Cancer* 1990, **26** : 1167-1256

[28] VERDECCHIA A, CAPOCACCIA R, EGIDI V, GOLINI A. A method for estimation of chronic disease, morbidity and trends from mortality data. *Stat Med* 1989, **8** : 201-16

[29] BRILLINGER DR. The natural variability of vital rates and associated statistics. *Biometrics* 1986, **42** : 693-734

[30] PRESSAT R. *L'analyse démographique.* Paris, Presses Universitaires de France, 1973

[31] HILL C, COM-NOUGUÉ C, KRAMAR A et al. *Analyse statistique des données de survie.* INSERM/Médecine-Sciences Flammarion, 1990, Paris

[32] ANDERSEN PK. Counting process for life history data : A review. *Scand J Stat* 1985, **12** : 97-158

[33] KALBFLEISCH JD, PRENTICE RL. *The statistical analysis of failure time data.* New York, Wiley, 1980

[34] COX DR, OAKES D. *Analysis of survival data.* London, Chapman and Hall, 1984

[35] BRESLOW NE. The proportional hazards model : Applications in epidemiology. *Commun Stat,* (Ser A) 1978, **7** : 315-32

[36] BERRY G. The analysis of mortality by the subject years method. *Biometrics* 1983, **39** : 173-84

[37] BRESLOW NE, LUBIN JH, MAREK P, LANGHOLZ B. Multiplicative models and cohort analysis. *J Am Stat Assoc* 1983, **78** : 1-12

[38] KEIDING N. Age-specific incidence and prevalence : a statistical perspective. *J R Statist Soc* A 1991, **154** : 371-412

[39] VERDECCHIA A, CAPOCACCIA R. Discussion of the paper by Keiding. *J R Statist Soc A* 1991, **154** : 405-6

[40] PRESTON SH,. Relations among standard epidemiologic measures in a population. *Am J Epidemiol* 1987, **126** : 336-45

[41] DAVID HA, Moeschberger MC. *The theory of competing risks.* London, Charles Griffin, 1978

[42] MAKEHAM WM. On an application of the theory of the composition of decremental forces. *J Hist Actuaries* 1874, **18** : 317-22

[43] DUCHÊNE J. Dépendances entre processus morbides et mesures de la mortalité par cause de décès. *In* Chaire Quételet : 82 : *Morbidité et mortalité aux âges adultes dans les pays développés.* Louvain-la-Neuve, Cabay-Jezierski, 1983

[44] MACMAHON B, PUGH TF. *Epidemiology : principles and methods.* Boston, Little-Brown, 1970

[45] KLEINBAUM DG, KUPPER LL, MORGENSTERN H. *Epidemiologic research : Principles and quantitative methods.* Belmont, Life-time Learning, 1982

[46] FREEMAN J, HUTCHISON GB. Prevalence, incidence and duration. *Am J Epidemiol* 1980, 112 : 707-23