Chapter 2

Techniques for the analysis of cancer risk

Measurement of the risk of cancer

Age- and sex-specific rates

The annual incidence rate for a specific tumour, for a group and for a given time period is equal to the ratio between the number of new cases of the tumour observed in the group over the given time and the number of *person-years* accumulated by the members of the group in the same time interval.

The calculation of an incidence rate is more meaningful when the group is homogeneous and when there is a constant risk during the time period. Moreover, it is only under these conditions that the observed incidence rate can be considered as an estimate of the underlying instantaneous rate which plays a key role in the definition of the risk of cancer (see Chapter 1, page 11). The homogeneity condition justifies the calculation of rates separately by age and sex, known as *specific incidence rates* because they refer to subgroups of the population and not to the population as a whole.

In the following, we first describe methods for calculating specific incidence rates, and then examine techniques of estimating their precision since, like all indexes calculated from observed data, the incidence rate is subject to random variation. Finally, we describe some typical incidence curves.

The calculation of a specific rate

The only problems involved in the determination of the numerator are the completeness of registration and respect for whatever guidelines have been adopted to define new cases. We will return to this point later in detail with the study of time trends, which are particularly vulnerable to changes in the definition adopted (see Chapter 3, page 176).

The determination of the denominator depends on available demographic statistics. In theory, the calculation of the exact number of person-years of observation requires individual data, but statistical offices provide at best reports including cross-sectional characteristics of the population at periodic intervals, obtained from cen-

suses or other population estimates. Thus, the denominator can be estimated only by making assumptions about the evolution of the population between two of these points, that is, about the way in which individuals traverse the age × time rectangle of the Lexis diagram (see Figure 1.1). Let us suppose, for example, that we wish to estimate the annual incidence of breast cancer for women aged 45 to 49 years in Zaragoza (Spain) between the beginning of 1973 and the end of 1977. Theoretically, we should add up the number of years lived in this age group by each woman of the population of Zaragoza during the period 1973-1977: thus, a woman who turned 45 years of age on 1 January 1977 will contribute one year to the person-years, in the same way as a woman who turned 49 on 1 January 1973 will contribute one year. In reality, it is known only that 27 699 women were between 45 and 49 years of age in 1975, the year of the census. It is supposed that there are as many women each year joining the age group as there are leaving it and that the number counted at the mid-point is consequently an estimate of the average number throughout the interval. Therefore, the estimate of the number of person-years accumulated between 1973 and 1977 is obtained by multiplying the number at the mid-point by five (27 699 × 5). Then, as the Cancer Registry recorded 109 cases of breast cancer for women between 45 and 49 years of age in the interval under consideration, the specific rate of breast cancer in this age group is

$$109/(27\ 699 \times 5) = 78.7 \text{ cases per } 100\,000 \text{ women per year.}$$

In most situations, this method for approximating the denominator is acceptable. However, the example below shows that the method can sometimes lead to aberrant results.

In Calvados, France, the resident population in the age group 60 to 64 years at the first of January evolved as follows from 1977 to 1982:

	Number in age group 60 to 64 years at 1 January
1977	20 790
1978	18 592
1979	16 886
1980	15 643
1981	18 757
1982	22 106

To calculate the incidence rate in the interval between 1 January 1977, and 31 December 1981, using the previously described method, we would take as the denominator five times the average population for the year 1979, that is

$$5 \times \frac{(16\,886 + 15\,643)}{2} = 81\,323 \text{ person-years}$$

However, a careful examination of the annual figures reveals fluctuations due to the effects of the decline in the birth rate during the first world war. Therefore, the calculation of incidence rate should take the figures for each year of the interval

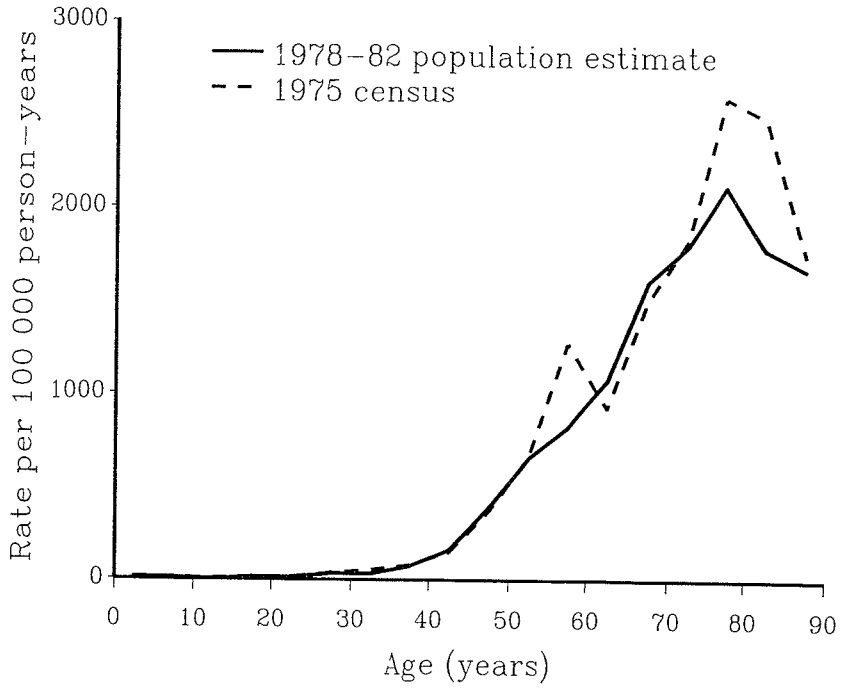


Figure 2.1 Influence of the denominator estimates on the age-specific incidence curve. All cancer sites, Calvados (France), males, incident cases 1978-1982
Source : Robillard [1]

into consideration; supposing that, on average, the number of individuals at risk in the group under study can be estimated each year by the arithmetic average of the number of individuals in the age group at the beginning and the end of the year, then we have

$$20\,790/2 + 18\,592 + 16\,886 + 15\,643 + 18\,757 + 22\,106/2,$$

that is, a total of 91 326 person-years [1].

In this example, the previous approximation under-estimated the calculation of person-years accumulated in the interval by 11%. The solution which takes as denominator a demographic estimate that does not correspond to the mid-point of the interval being considered can lead to even more serious inaccuracies. Figure 2.1 shows, again in Calvados, biases in the age-specific incidence curve when the number of cases observed for the interval 1978-82 (males) is related to data from the 1975 census. Even if variations from one year to the next are rarely as marked as those in our example, successive annual estimates should be used in the calculations when they are available.

The accuracy of the estimate of a rate

Regardless of the bias that a wrong evaluation of the denominator causes, we should question the accuracy of the estimate of the rate being calculated.

For reasons that were discussed earlier (see Chapter 1, page 20), the denominator can be considered as a non-random quantity; thus, the accuracy of a rate

only depends on the variability of the number of cases observed (K). We can therefore suppose that K is a variable that follows a Poisson distribution whose expectation and variance are equal to the theoretical rate (λ) that we are attempting to estimate, multiplied by the number of person-years (m) accumulated within the period of the study:

$$K \rightsquigarrow P(\lambda m)$$

$$E(K) = \lambda m$$

$$Var(K) = \lambda m$$

Therefore, the variance of the rate estimator (K/m) is

$$Var\left(\frac{K}{m}\right) = \frac{Var(K)}{m^2} = \frac{\lambda}{m}$$

Its estimate is obtained by replacing λ by k/m in the above formula, k being the observed value of K; it is given by

$$\hat{Var}\left(\frac{K}{m}\right) = \frac{k}{m^2} = \frac{\hat{\lambda}^2}{k}$$

an expression which has already been obtained in Chapter 1, page 20. It is then possible to construct a confidence interval of level $1 - \alpha$ for λ. When k is large, we can consider that the distribution of K/m is normal with mean λ and standard deviation $\frac{\sqrt{k}}{m}$, therefore

$$Prob\left[\left|\frac{\frac{K}{m} - \lambda}{\frac{\sqrt{k}}{m}}\right| < Z_{\alpha/2}\right] = \alpha$$

hence the confidence interval:

$$\left[\frac{k}{m} - Z_{\alpha/2}\frac{\sqrt{k}}{m} ; \frac{k}{m} + Z_{\alpha/2}\frac{\sqrt{k}}{m}\right]$$

The usual value of α is 0.05 and $Z_{\alpha/2} = 1.96$. As an example, if nine cases have been observed in a population of 10 000 persons followed up during three years, the incidence rate is 30 per 100 000; its variance is $9/(30\,000)^2$, and its standard error is 10/100 000. Therefore, the confidence interval may be written:

$$30/100\,000 \pm (1.96 \times 10/100\,000) = [10.40/100\,000 ; 49.6/100\,000]$$

It is also possible to use directly a confidence interval for the expectation of K as calculated from the Poisson distribution (see page 64). Table 2.3 below gives the values [4.12 ; 17.08], which leads to a confidence interval for the rate equal to [13.70/100 000 ; 56.93/100 000]. This exact interval is fairly different from the above

conventional interval. It is therefore recommended to use the Poisson distribution when the number of cases observed is less than 50.

In practice, it is usual to assess the accuracy of the rate on a relative scale. The relative error in the estimation of a rate is given by the coefficient of variation of the estimated rate, which is defined as the ratio between its standard error and its mean:

$$CV = \frac{\sqrt{\lambda/m}}{\lambda} = \frac{1}{\sqrt{\lambda m}}$$

The expected value of K being λm, $1/\sqrt{k}$ provides a simple estimator of the accuracy of the rate measured on a relative scale. In the previous example, the relative error in the calculated rate is $1/\sqrt{9} = 33\%$. If we had observed four cases, the relative error would have been $1/\sqrt{4} = 50\%$. These examples reveal the substantial inaccuracies which can affect measures of rare cancers.

The coefficient of variation that we defined above has a natural interpretation when it is appropriate to consider the rates after logarithmic transformation (see next page). In fact, in this case, variability is measured by the standard error of the logarithm of the specific rate which can be calculated in the usual way:

$$\mathrm{Var}\left[\mathrm{Log}\left(\frac{K}{m}\right)\right] = \mathrm{Var}\left[\mathrm{Log}(K) - \mathrm{Log}(m)\right] = \mathrm{Var}\left[\mathrm{Log}(K)\right]$$

$$\approx \left(\frac{d\,\mathrm{Log}\left[E(K)\right]}{dK}\right)^2 \times \mathrm{Var}(K) = \left(\frac{1}{\lambda m}\right)^2 \times \mathrm{Var}(K)$$

$$\mathrm{Var}\left[\mathrm{Log}\left(\frac{K}{m}\right)\right] \approx \frac{1}{\lambda m} \qquad (2.1)$$

Thus, not surprisingly, the standard error of the logarithm of the rate is equal to the coefficient of variation. Using the same principle as before and the data from the previous example, the confidence interval of the logarithm of the rate is

$$\mathrm{Log}\ (30/100\,000) \pm (1.96 \times 0.33)$$

which leads, by taking the exponential of the interval end-points to a new confidence interval for the rate itself

$$CI_{95\%} = [15.7 / 100\,000\ ;\ 57.3 / 100\,000]$$

It is worth noting that, by improving the required normality, the logarithmic transformation has led to a result which is closer to the exact interval than the conventional interval based on the rate itself.

As the accuracy of the estimate depends only on the number of observed cases, it can theoretically be increased by lengthening the observation time. However, if incidence is not constant over time, the accumulation of cases over several years can only lead to a less meaningful result. In practice, the choice of interval

is a compromise between these two requirements. The situation is similar when we consider that a region covered by a registry is too heterogeneous to give only one estimate of the rate. If we decide to divide the area into subgroups which are more homogeneous, the accuracy of the rate estimates in each subregion is lower. Therefore, a compromise between interpretability and accuracy has to be found (see Chapter 3).

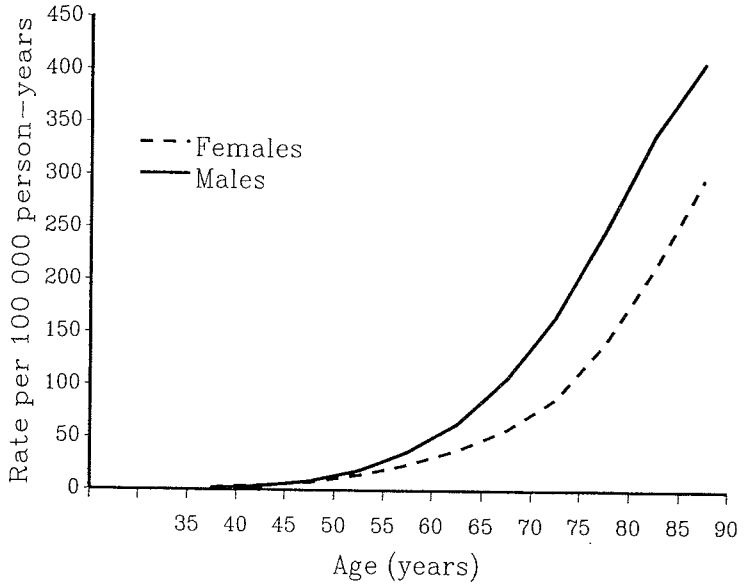
The incidence curve

Age-specific rates are usually calculated for seventeen five-year age groups between the age of 0 and 85 years, with an eighteenth group for 85 years and over. As a rule, the rates should be represented on a graph by a step-function with five-yearly increments. However, it is customary to join the points that mark the mid-point of each age group; the line obtained by doing so is called *the incidence curve*. In a population where the age-specific incidence might remain constant over a period of time, such as would occur in the absence of a cohort effect, the curve could be seen as an estimate of the function $\lambda(t)$ which we defined in Chapter 1. However, as incidence does tend to change with time, the shape of the curve is a result of the combined effect of age and observation time: incidence rates for older age-groups describe a relationship between risk and age that does not necessarily correspond to that described by incidence for the youngest individuals living at the same time. In other words, when older people today were young, they did not have the same risk as the young people of today.

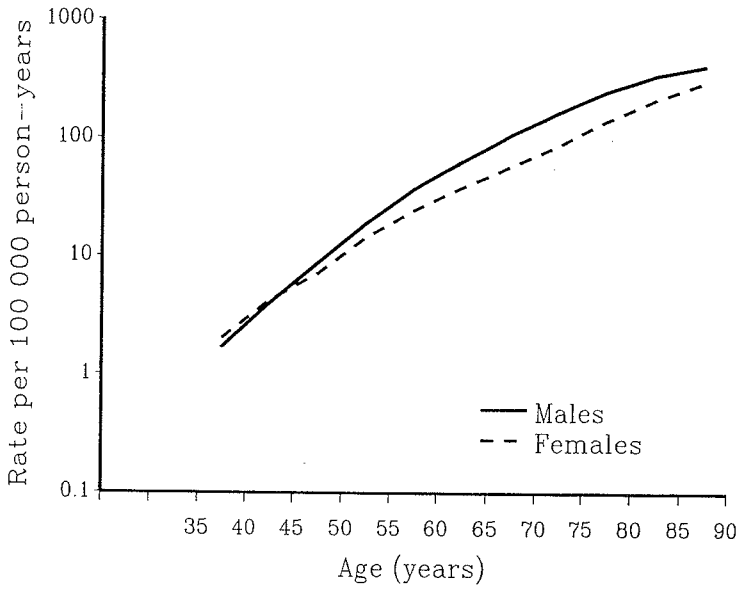
As we stated previously, incidence according to age is sometimes shown after logarithmic transformation of age-specific rates. This sort of representation is used firstly for practical reasons. Rates of very different orders of magnitude can be represented on the same graph, allowing a clear visualization of incidence levels for ages where rates are low. It is also worth noting that a constant ratio of age-specific incidence rate between two populations will produce, on a logarithmic scale, two parallel incidence curves.

A logarithmic scale may also be used on the age axis. Thus, a log-log graph is designed to place the observed data in the context of the multi-stage model of carcinogenesis [2,3]. According to this model, incidence is a power function of age and should therefore be represented by a straight line on a log-log scale. However, such a model can only be identified by this procedure in the absence of a cohort effect [4].

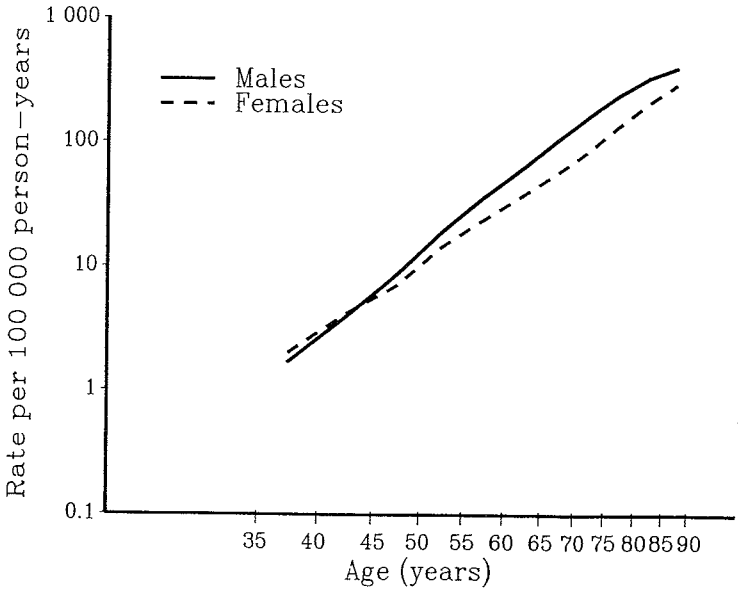
The mortality from colorectal cancer in France for the period 1978-1982 is represented in Figure 2.2 by using various scales. In this case it is clear that Figure 2.2(c) provides a remarkably concise description of the increase in risk with age. However, other more complex incidence curves are often seen (Figures 1.2 and 2.5). In particular, the incidence curve for breast cancer shows a characteristic drop in the rate of increase around 50 years of age; Clemmesen has demonstrated the universality of this phenomenon. [5]



a : arithmetic scales



b : semi-logarithmic scales



c : Log-Log scales

Figure 2.2 Influence of the choice of scale on the shape of age-specific incidence curves. Mortality from colorectal cancer in France, 1978-1982

Standardized rates

One of the principal aims of collecting incidence data is the investigation of etiological factors for the disease being considered. In order to compare observed incidence for different regions or groups or years, we should be in a position to take account of the factors which are already recognized as possible explanations of observed differences in rates. Among these factors, age is the first candidate. The effects of age are large and, in general, the various populations being compared differ in their age structures. The control of the confounding effect of a factor, by methods to be discussed below, implies that we know its distribution in the populations that we wish to compare. This is the reason why the following methods cannot be applied to biasing factors such as the quality of registration or the accuracy of diagnosis. On the other hand, when denominators are not available, the method described on page 95 could be used.

Direct standardization

The principle of this method is to determine the annual rate that would be observed in a *standard*, or theoretical, population of a given age structure, were it subjected to the force of incidence of the population under study. The procedure is based on the calculation of the expected number of cases in each age-group of this *standard population* by applying to the corresponding person-years the estimated rate of the population under study. The total number of expected cases is then divided by the total number of person-years in the theoretical population.

Let:

- g be the number of age groups under consideration, which is usually 18 but can change if we are calculating a truncated rate for a subset of adjacent age-groups, for example, 35-64 years;
- L be the size of a standard population,
- L_x be the number of individuals in the xth age-group of this standard population,
- k_x be the number of cases observed in the xth age-group of the population under study
- m_x be the number of person-years accumulated in the xth age-group of the population under study
- $t_x = k_x/n_x$ be the specific rate of the xth age-group of the population under study.

$L_x t_x$ is thus the number of expected cases that might be observed in one year in the xth age group of the standard population if it were exposed to a level of risk defined by the rate t_x. The *standardized rate* is then:

$$\bar{t} = \frac{1}{L} \sum_{x=1}^{g} L_x t_x \qquad (2.2)$$

It may also be written

$$\bar{t} = \sum_{x=1}^{g} w_x t_x \qquad (2.3)$$

where $w_x = L_x/L$ is the proportion of individuals in the xth age group in the standard population with

$$\sum_{x=1}^{g} w_x = 1$$

This expression shows that the rate $\bar{t}$ is a weighted average of age-specific rates, with the weights being the proportion of individuals in the various age groups of the standard population.

We should note that the calculation presumes that the number of person-years of observation and the number of observed cases in each age-group of the population under study (or at least the age-specific rates) are known. Furthermore, the calculation requires the choice of a standard population. In practice, this choice depends on our objective and it influences the numerical result that we obtain. The principal standard populations that have been suggested are presented in Table 2.1. For routine comparisons, it is preferable to use the world population as a standard. The European population figures are suitable when we are comparing observed incidences in countries where the age structure is similar to that usually observed in developed countries. In the same way, the African population can be used as a standard for developing countries. A *truncated population* is used to restrict the comparison to the adult age groups where the most interesting differences appear. It also has the advantage of eliminating from the standardized rate the contribution of the oldest age groups that are particularly subject to the risk of being under-registered. When we are not dealing with routine comparisons, other standards are sometimes adopted; for example, if we wish to describe the risk in several subsets of a region or a country, it is reasonable to take the total population of the region or the country as the standard population. In the particular case where we are interested in two regions or countries, the sum of their populations is sometimes taken as the standard.

Table 2.2 presents the calculation of the standardized rate of stomach cancer for males in the French region of the Côte-d'Or from 1976 to 1980, using the European population as a standard.

The calculation of a directly standardized rate uses age-specific rates that have been estimated from observations which are subject to a certain amount of random variability. This variability affects the estimate of the standardized rate and can lead to spurious conclusions if the observed difference between standardized rates is in fact mainly due to random variation. In order to evaluate the importance of this kind of variation, the standardized rate (t) should be presented with its standard error or its confidence interval.

Table 2.1 Age structure of commonly used standard populations [6] (valid for either sex)

Age group	World	African	European	World truncated
0-4	12	10	8	–
5-9	10	10	7	–
10-14	9	10	7	–
15-19	9	10	7	–
20-24	8	10	7	–
25-29	8	10	7	–
30-34	6	10	7	–
35-39	6	10	7	6
40-44	6	5	7	6
45-49	6	5	7	6
50-54	5	3	7	5
55-59	4	2	6	4
60-64	4	2	5	4
65-69	3	1	4	–
70-74	2	1	3	–
75-79	1	0.5	2	–
80-84	0.5	0.3	1	–
85 +	0.5	0.2	1	–
Total	100	100	100	31

As we saw previously when discussing the estimation of λ_x, from K_x observations resulting from m_x person-years in age group x,

$$K_x \rightsquigarrow P(\lambda_x m_x)$$

$$E(K_x) = \text{Var}(K_x) = \lambda_x m_x$$

The variance of the specific rate $t_x = K_x/m_x$ is then obtained using the classical method

$$\text{Var}(t_x) = \frac{\text{Var}(K_x)}{m_x^2} = \frac{\lambda_x}{m_x}$$

Therefore, the variance of the standardized rate is, from formula (2.3)

$$\text{Var}(\bar{t}) = \sum_{x=1}^{g} w_x^2 \, \text{Var}(t_x)$$

$$\text{Var}(\bar{t}) = \sum_{x=1}^{g} w_x^2 \left(\frac{\lambda_x}{m_x}\right) \qquad (2.4)$$

λ_x being unknown, $\mathrm{Var}(\bar{t})$ must be estimated by replacing λ_x by its estimate k_x/m_x in the above expression. Then

$$\hat{\mathrm{Var}}(\bar{t}) = \sum_{x=1}^{g} \left(\frac{w_x^2}{m_x^2} \right) k_x$$

If the theoretical standardized rate is denoted by $\mu = \sum_x w_x \lambda_x$ and if s is the estimate of its standard error, then we can consider that $(\bar{t} - \mu)/s$ is approximately a standard normal variable; the confidence interval at level $1 - \alpha$ for μ is then obtained as explained previously:

$$[\bar{t} - Z_{\alpha/2} \sqrt{\hat{\mathrm{Var}}(\bar{t})} \; ; \; \bar{t} + Z_{\alpha/2} \sqrt{\hat{\mathrm{Var}}(\bar{t})}\,]$$

In practice, rates are given per 100 000 person-years (10^5 t_x); the variance that is calculated is therefore in the form 10^{10} $\mathrm{Var}(\bar{t})$.

Table 2.2 also gives the data required to calculate the variance of the standardized rate, from which we obtain a standard error of 1.55 and its 95% confidence interval [20.49 ; 26.58].

We should note again that the procedure which enables the confidence interval to be constructed from the standard error of the estimator implies that the distribution of this estimator is reasonably close to normal. This is in fact only true in the present situation if the total number of cases is sufficiently large. It is however difficult to tell what 'sufficiently' means in the present context because the numerator of a standardized rate is no longer a Poisson variate. Its variance depends not only on the total number of observed cases but also on the weighting scheme w and the accuracy of the age-specific rates. This may be seen by writing the formula (2.4) in the following way:

$$\mathrm{Var}(\bar{t}) = \frac{1}{m} \sum_{i=1}^{g} w_x \frac{\kappa_x}{m_x} = \frac{\tau}{m}$$

where $\kappa_x = L_x \lambda_x$ and L_x, the numerator of w_x, is chosen in such a way that:

$$\sum_{i=1}^{g} L_x = \sum_{i=1}^{g} m_x = m$$

This expression shows that the variance may be badly assessed from the total number of expected cases especially if the majority of them (κ_x) originated from an age group where m_x is low (see page 100).

The quotient of two standardized rates calculated from the same standard population is known as the *comparative incidence figure* (CIF). It is a measure of the relative risk of a population compared with another population and is generally expressed as a percentage. The standardized rate in a subgroup of a population that is itself used as the standard, divided by the crude rate in the whole population

Table 2.2 Calculation of a directly standardized rate (stomach cancer in Côte-d'Or, France, males, 1976-1980, European standard)

x	Age	k_x	m_x	$10^5\ t_x$	w_x	$10^5 \bar{w}_x t_x$	$10^{10}\left(\frac{w_x^2 k_x}{m_x^2}\right)$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	0-4	0	91 228	0.00	0.08	0.000	0.0000
2	5-9	0	95 230	0.00	0.07	0.000	0.0000
3	10-14	0	95 869	0.00	0.07	0.000	0.0000
4	15-19	0	98 744	0.00	0.07	0.000	0.0000
5	20-24	0	101 131	0.00	0.07	0.000	0.0000
6	25-29	0	101 103	0.00	0.07	0.000	0.0000
7	30-34	0	83 544	0.00	0.07	0.000	0.0000
8	35-39	1	67 580	1.48	0.07	0.104	0.0107
9	40-44	3	68 577	4.37	0.07	0.306	0.0313
10	45-49	6	68 126	8.81	0.07	0.617	0.0633
11	50-54	10	63 708	15.70	0.07	1.099	0.1207
12	55-59	17	51 007	33.33	0.06	2.000	0.2352
13	60-64	27	37 695	71.63	0.05	3.582	0.4751
14	65-69	34	44 374	76.62	0.04	3.065	0.2763
15	70-74	51	36 768	138.71	0.03	4.161	0.3395
16	75-79	46	24 196	190.11	0.02	3.802	0.3143
17	80 +	42	17 491	240.12	0.02	4.802	0.5491
Total		237	1 146 371		1.00	23.537	2.4155

Columns 1 to 4 and 6 are given and columns 5, 7 and 8 are calculated.
k_x : observed number of cases of stomach cancer in Côte-d'Or from 1976 to 1980 for the xth age group.
m_x : estimate of the number of person-years for males in each age group x, obtained by summing the numbers of the Côte-d'Or population from 1976 to 1980 (INSEE, PRUDENT).
t_x : age-specific rate per 100 000 persons per year.
w_x : structure of the standard population by age.

(which in this case is equal to the standardized rate with respect to itself) is also a CIF.

The value of a CIF is independent of the standard population used only if the ratio of the age-specific incidence rates is constant, in other words, only when the two incidence curves that are being compared are parallel when the log scale is used on the rate axis. This property often holds for incidence curves (see Figure 2.3) and can be checked with a statistical test which evaluates the assumption of the homogeneity of age-specific relative rates (see page 80).

Cumulative rates

The overall incidence observed in a population can also be described by the *cumulative rate* [7] which provides, as we shall see below, an approximation of the risk of developing a disease before age b (or between two ages a and b) in the absence of mortality (see the concept of net risk in Chapter 1, page 34). The cu-

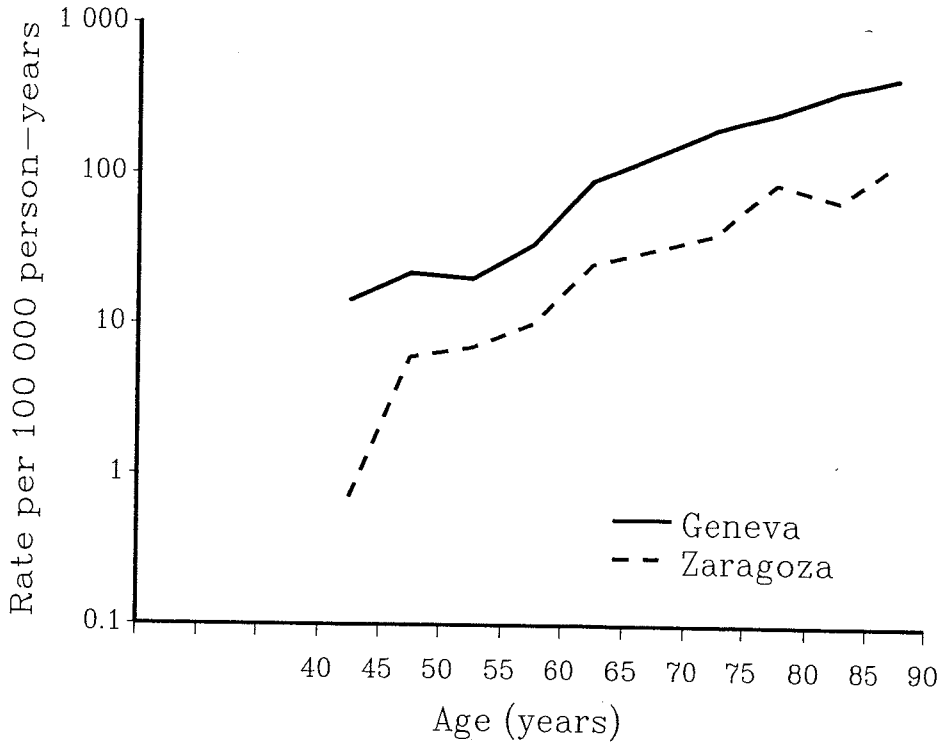


Figure 2.3 Age-specific incidence of colon cancer in Zaragoza (Spain) and Geneva (Switzerland) males, 1973-1977

mulative rate over a whole lifetime is an integral of the function represented by the incidence curve. This rate can be estimated by adding up the age-specific incidence over each year of age. Assuming that the incidence is constant within an age group (x) of five years, we will write

$$t_{0,b} = \sum_{x=1}^{j} 5 \times t_x = 5 \sum_{x=1}^{j} t_x \qquad (2.5)$$

to estimate the cumulative rate from zero to the upper limit b of age group j, and

$$t_{a,b} = 5 \sum_{x=i}^{j} t_x \qquad (2.6)$$

to estimate the cumulative rate from the lower limit a of age group i to the upper limit b of the age group j.

For example, the cumulative rate of stomach cancer between 35 and 65 years of age can be calculated from the data in Table 2.2 by adding up the numbers in column 5 from line 8 to line 13 and multiplying the result by 5/100 000, i.e.

$$t_{35,65} = 0.0068 = 0.68\%$$

Estimation of the cumulative rate over a whole lifetime presents a problem because the last age group is open and, unlike the other age-groups, does not contain five years. If the last age-group is 80 years and over, we can suppose that the estimate of the rate in this age-group is almost identical to the rate in the age group 80-85 years, and consider that the value we obtain is the cumulative rate up to 85 years. With this convention, the cumulative rate over life of stomach cancer can be estimated by

$$t_{0,85} = 0.0390 = 3.90\%.$$

In practice, it is preferable not to calculate the cumulative rates beyond the upper limit of the last closed age group. In fact, cumulative rates are rarely published above 75 years, the age at which competing causes of death begin to play a major role (see Chapter 1, page 34).

Note that the cumulative rate is proportional to the arithmetic average of the age-specific rates, that is, to a rate that would be standardized to a population in which every age-group contained the same proportion of individuals ('rectangular' population). Note also that the probabilistic interpretation mentioned above assumes that the cross-sectional incidence curve, constructed for a given time period from different cohorts, correctly represents the force of incidence applicable to an individual for whom we wish to evaluate risk; in fact, the risk obtained in this way is that of a 'fictitious' individual who synthesizes the experience of several cohorts.

The standard error and confidence interval of a cumulative rate are obtained in the same way as those for a direct standardized rate; the application of formula (2.4) with $w_x = 5$ gives

$$\text{Var}\,(t_{a,b}) = 25 \sum_{x=i}^{j} \frac{k_x}{m_x^2} \qquad (2.7)$$

for example, the standard error of the cumulative rate of stomach cancer between age 35 and 65 years is

$$\sqrt{\text{Var}\,(t_{35,65})} = 0{,}0009$$

from which we derive a confidence interval of [0.51% ; 0.85%].

Indirect standardization

While direct standardization could be called the method of the standard population, the procedure described in this paragraph could be called the method of standard incidence. The principle is based on the comparison between the total number of cases observed in the population under study and the number that could be expected if the population was subject to a given force of incidence (λ_x), the standard incidence.

The number of expected cases in the population under study is

$$E = \sum_{x=1}^{g} m_x \lambda_x \qquad (2.8)$$

where λ_x is the incidence rate of group x in the standard population, and m_x is the number of person-years accumulated by group x in the population under study.

The ratio between the total number of cases observed in the population under study (O) and the expected number (E) is called the *standardized incidence ratio* (SIR). Like the CIF, it is a measure of relative risk of the population under study compared with the standard population. It is usually expressed as a percentage,

$$SIR = \frac{O}{E} \times 100 \qquad (2.9)$$

Therefore, a value of 150 for this index means that 50% more cases were observed in the population under study than if the incidence was that of the standard population.

For reasons already discussed, the variability of the SIR depends only on the numerator, whose distribution can be considered to be Poisson. The estimate of the SIR variability can be obtained accurately from Table 2.3 which gives the 95% confidence interval of the expectation μ of a Poisson variable given an observed number of cases O.

The results in Table 2.3 are obtained by defining the lower and upper limits of the confidence interval μ_0 and μ_1 according to the formulae:

$$P[X \geq O | \mu_0] = \alpha/2 \; ; \; P[X \leq O | \mu_1] = \alpha/2$$

such an interval will contain the true value μ with probability $1 - \alpha$. On the other hand, the Poisson distribution is related to the χ^2 distribution by the relation:

$$Pr[X \geq k | \mu] = Pr[\chi^2_{2k} < 2\mu]$$

in other words, if F_{2k} is the distribution function of χ^2 with 2k degrees of freedom, we can write:

$$\sum_{x=k}^{\infty} e^{-\mu} \frac{\mu^x}{x!} = F_{2k}(2\mu)$$

$$F_{2O}(2\mu_0) = \alpha/2$$

$$F_{2(O+1)}(2\mu_1) = 1 - \alpha/2$$

therefore, if F^{-1} denotes the reciprocal function of F:

$$\mu_0 = \frac{1}{2} F^{-1}_{2O}(\alpha/2)$$

$$\mu_1 = \frac{1}{2} F^{-1}_{2(O+1)}(1 - \alpha/2)$$

Table 2.3 Exact 95% confidence interval for the expectation (μ) of a Poisson distribution according to the number of observed cases (O)

Observed cases (O)	95% Confidence interval		Observed cases (O)	95% confidence interval	
	μ_0	μ_1		μ_0	μ_1
0	0.00	3.00	31	21.06	44.00
1	0.03	5.57	32	21.89	45.17
2	0.24	7.22	33	22.72	46.34
3	0.62	8.77	34	23.55	47.51
4	1.09	10.24	35	24.38	48.68
5	1.62	11.67	36	25.21	49.84
6	2.20	13.06	37	26.05	51.00
7	2.81	14.42	38	26.89	52.16
8	3.45	15.76	39	27.73	53.31
9	4.12	17.08	40	28.58	54.47
10	4.80	18.39	41	29.42	55.62
11	5.49	19.68	42	30.27	56.77
12	6.20	20.96	43	31.12	57.92
13	6.92	22.23	44	31.97	59.07
14	7.65	23.49	45	32.82	60.21
15	8.40	24.74	46	33.68	61.36
16	9.15	25.98	47	34.53	62.50
17	9.90	27.22	48	35.39	63.64
18	10.67	28.45	49	36.25	64.78
19	11.44	29.67	50	37.11	65.92
20	12.22	30.89	51	37.97	67.06
21	13.00	32.10	52	38.84	68.19
22	13.79	33.31	53	39.70	69.33
23	14.58	34.51	54	40.57	70.46
24	15.38	35.71	55	41.43	71.59
25	16.18	36.90	56	42.30	72.72
26	16.98	38.10	57	43.17	73.85
27	17.79	39.28	58	44.04	74.98
28	18.61	40.47	59	44.91	76.11
29	19.42	41.65	60	45.78	77.23
30	20.24	42.83	61	46.66	78.36

When the number of observed cases is zero, X is greater than the observed number with probability 1 whatever μ_0 may be. To keep the correct level of confidence $1 - \alpha$, we construct the interval $[0 ; \mu_1]$ such that $P[X = 0|\mu_1] = e^{-\mu 1} = \alpha$. This interval covers the theoretical value μ with probability $1 - \alpha$. For example, when $\alpha = 5\%$, $\mu_1 = -\text{Log}\,(0.05) = 3.00$.

When O is greater than 50, we can assume that Log (O) follows a normal distribution with expectation Log(μ) and variance $1/\mu$. Thus, to obtain a 95% confidence interval we make use of the inequality

$$\frac{|\,\text{Log}\,(O) - \text{Log}\,(\mu)\,|}{1/\sqrt{\mu}} < 1.96$$

which gives after replacing μ with its estimate O

$$O\,e^{\left(-\frac{1.96}{\sqrt{O}}\right)} < \mu < O\,e^{\left(\frac{1.96}{\sqrt{O}}\right)}$$

for example, when O = 50 and E = 45.6, the 95% confidence interval of μ is

$$[\mu_0 ; \mu_1] = [37.9 ; 66.0]$$

therefore, the interval of the corresponding SIR is [83.1 ; 144.7]. If instead we use Table 2.3, the confidence intervals are respectively

[37.1 ; 65.9] and [81.4 ; 144.5].

Another more reliable approach is based on the approximation of the distribution of $\sqrt{X}$ by a normal distribution with mean $\sqrt{\mu}$ and variance 1/4 [8]; the confidence interval is then

$$[\mu_0 ; \mu_1] = \left[\left(\frac{Z_{\alpha/2}}{2} - \sqrt{O} \right)^2 ; \left(\frac{Z_{\alpha/2}}{2} + \sqrt{O+1} \right)^2 \right]$$

for O = 50, this method gives [37.1 ; 66.0].

The calculation of the SIR requires only the number of person-years accumulated in each of the different groups x in the population under study and not the number of cases occurring in these groups. It requires the choice of a standard distribution which, in practice, is dictated by the use that we intend to make of the SIR, as will be shown subsequently.

As the SIR is an estimate of relative risk with respect to a reference force of incidence, the product of the SIR and the crude rate in the standard population which provides the standard incidence rates is in fact a form of standardized rate known as the *indirectly standardized rate.*

Table 2.4 provides the data required to calculate the standardized incidence ratio of colon cancer for males in the French city of Dijon between 1976 and 1980, using rates observed in the whole region of the Côte-d'Or as a standard. We obtain

$$SIR = 100 \times \frac{123}{98.7} = 124.6$$

and the 95% confidence interval of the SIR is [104.5 ; 148.9] obtained using the first normal approximation above. We can calculate the indirect standardized rate from the crude rate of 24.3 (see Table 2.4) and we can obtain the indirectly standardized rate

$$\tilde{t} = 1.246 \times 24.3 = 30.3$$

Table 2.4 Calculation of a standardized incidence ratio (SIR) for colon cancer in the town of Dijon for the period 1976-1980 with the overall incidence in the French département of Côte-d'Or as a standard

x	k_x	m_x	$10^5 \lambda_x$	e_x
0-4	0	45 626	0.0	0.00
5-9	0	41 145	0.0	0.00
10-14	0	39 284	0.0	0.00
15-19	0	43 469	0.0	0.00
20-24	0	52 794	1.0	0.53
25-29	0	54 321	1.0	0.54
30-34	0	40 848	0.0	0.00
35-39	2	31 559	5.9	1.86
40-44	2	30 703	4.4	1.35
45-49	3	29 875	14.7	4.39
50-54	10	27 228	22.0	5.99
55-59	17	21 808	47.1	10.27
60-64	7	15 002	45.2	6.78
65-69	17	14 556	81.2	11.82
70-74	33	11 841	206.7	24.47
75-79	20	7 762	214.9	16.68
80 +	12	6 112	228.8	13.98
Total	O = 123	513 933	24.3	E = 98.70

k_x : observed number of cases in age group x in Dijon.
m_x : person-years of observation in age group x in Dijon.
λ_x : observed colon cancer rate in age group x in Côte-d'Or.
e_x : expected number of cases in age group x in Dijon if the incidence rates were λ_x (i.e., that of Côte-d'Or).

Probability of developing a specific form of cancer

The cumulative rate discussed previously is an approximation to the *net cancer risk*, that is, of the probability of developing cancer in the absence of mortality. In fact, we may also be concerned with the *crude probability* of developing a particular form of cancer; in other words, the risk actually incurred by an individual subjected not only to the risk of cancer but also to the risk of death. For a given level of incidence, this probability will be higher when the general mortality is low and vice versa.

The method of calculation of this probability is derived directly from formula 1.4 of Chapter 1. It was shown there that the raw probability of developing cancer is the sum for all ages of the product of the age-specific rate and the probability of survival without cancer up to this age. In practice, we shall estimate the probability of cancer from the life table neglecting the probability of not having cancer at age x which is close to 1 for most cancer sites.

Let:

- t_x be the incidence rate in the age group x;

- L_x be the number of years lived by the survivors of age x during the age interval starting at x if they are subject only to the force of mortality of the general population.
- ℓ_0 be the size of this population at the beginning of the first age interval under consideration (ℓ_0 and the L_X are provided by the life table, see Chapter 1, page 26).

Then the probability of developing a given cancer is

$$p = \frac{1}{\ell_0} \sum_{x=1}^{g} L_x t_x = \frac{K}{\ell_0} \qquad (2.11)$$

In fact, the summation in formula (2.11) gives the number K of expected cancers between the beginning of the first age interval and the end of the last if L_x is an acceptable approximation of person-years lived in each age group by cancer-free survivors.

When the probability of cancer (all sites) is being calculated, it might be better to construct a life table giving at each age the number of cancer-free survivors. The improvement obtained in this way is, however, somewhat illusory, as we shall see below.

When the current life table (see Chapter 1) is used in this calculation, the predictive value of this parameter should be viewed with caution. The actual mortality that will be experienced by cohorts for which the prediction is carried out may differ substantially from the reference mortality which has been used in the standard life table. This is why it is important to clarify the concept and to refer to it as being the *current probability* of developing cancer.

If we wish to compare probabilities in several regions or from several time periods, we can use the same life table; in this way, we obtain adjusted probabilities that play the same role as standardized rates. Note, however, that the standardization refers to mortality and not age, for which control is implicitly assured by the very definition of the parameter. For comparisons of this kind, it is much more simple to use the cumulative rates defined previously which provides the same type of information. When they are low, they actually provide a good approximation to the net probability R_b of developing a disease before a given age b, also known as the *cumulative risk.*

We shall give below a simple proof of this result that has previously been discussed in Chapter 1. First of all, consider an age group [x, x + Δx] in which the incidence rate is constant, and subdivide this interval into n equal parts; the probability of not developing the cancer under consideration at age x + Δx is the product of the probabilities of remaining healthy throughout each of the successive intervals thus defined. This probability is approximately

$$s_n = \left(1 - \lambda \frac{\Delta x}{n}\right)^n$$

the smaller the interval Δx/n, the more accurate the approximation will be. Now, it is known that the limit of s_n when n tends to infinity is $e^{-\lambda \Delta x}$. In other words, the probability of developing cancer between x and x + Δx is equal to $(1 - e^{-\lambda \Delta x})$.

Secondly, suppose that the age interval [0 – b] can be subdivided into j age groups of length Δx_i in which the rate λ_i is considered to be constant; the probability of not developing cancer before age b is obtained using the same principle as before:

$$1 - R_b = \prod_{i=1}^{j} e^{-\lambda_i \Delta x_i}$$

$$1 - R_b = e^{-\sum_{i=1}^{j} \lambda_i \Delta x_i}$$

If the Δ_x correspond to five-year age groups, the argument of the exponential is, except for the minus sign, the cumulative rate.

In practice, we calculate the estimate $t_{0,b}$ of the cumulative rate as was shown on page 61 and the estimate of the cumulative risk R_b according to the formula:

$$\hat{R}_b = 1 - e^{-t_{0,b}} \qquad (2.12)$$

Up to a cumulative rate of 10%, the two numbers $t_{0,b}$ and R_b are very close: the approximation of the cumulative risk R_b by the cumulative rate $t_{0,b}$ is therefore good for most cancer sites. As an example, the cumulative risk of stomach cancer between 35 and 65 years for the Côte-d'Or is 0.68%, while the life-time cumulative risk for the same region is 3.83% (the corresponding cumulative rates are respectively 0.68% and 3.90%; see Table 2.2).

Table 2.5 presents the three indexes that have been discussed, to evaluate the overall life-time cancer risk from data from New York State between 1969 and 1971 [9,10]. Note that the values of the two indexes defined by probabilities (cumulative risk and current probability) are relatively close to each other before 65

Table 2.5 Cumulative rate, cumulative risk, and current probability of cancer in New York State, USA (1969-1971) [6]

	Males		Females	
	All sites	Lung	All sites	Breast
Cumulative rate (%)				
0-65 years	12.3	3.0	12.8	4.0
0-75 years	28.8	7.0	22.6	6.2
Cumulative risk				
0-65 years	11.6	3.0	12.0	3.9
0-75 years	25.0	6.8	20.2	6.0
0-85 years	42.0	10.6	30.7	8.7
Current probability				
0-65 years	10.0	2.4	11.6	3.6
0 + ([a])	27.1	5.8	27.8	7.2

([a]) In this instance, the probability is calculated up to the terminal age of the table (see page 27).

years, particularly for females. Beyond this age, mortality has played a greater role effectively preventing incidence to manifest itself. In addition, we can see that the approximation of the cumulative risk by the cumulative rate is not-very satisfactory when incidence is high, such as occurs when all cancer sites are combined.

As an index of comparison between populations, the cumulative risk has two main advantages over the standardized rate; it avoids the arbitrary choice of the weighting and it expresses the risk on a probability scale which is interpretable immediately.

The number of years of life lost

Descriptive epidemiology is fundamental to etiologic research. In this capacity, it attempts to link characteristics of time and place to cancer development. It is therefore natural that the measurement of incidence or, failing that, the measurement of mortality will be the key instrument of the epidemiologist. But descriptive epidemiology should also provide information that could be useful in the establishment of public health priorities and policies, by addressing the consequences of cancer, the main one from a public health perspective being the amount of human life lost from the disease. This objective is already partially achieved by the determination of survival rates, but they do not provide an overall picture of the impact of cancer on the general population. In order to obtain this picture, we must measure the impact of cancer on the potential duration of life that individuals of the given population should have, on average, in the absence of the disease. The concept of *potential years of life lost* (PYLL) has exactly this objective, since it measures the average reduction of duration of life due to premature death caused by the given disease.

In order to assess the reduction in duration of life, two conceptual approaches have been proposed. The first suggests that the years lost from death due to the cause under study should only be taken into consideration up to an age limit that is arbitrarily fixed to mark the normal end of life; only deaths occurring at ages lower than this limit are then taken into account in the estimation of the reduction of duration of life. The second approach assumes that the reduction in potential life is equal to the number of years which the individual would otherwise have expected to live at the age of death. Thus this approach takes into account the force of general mortality exerted on the population under consideration. The two concepts differ in the same way as do the net and crude probabilities of dying from a certain cause of death,since the parameter is calculated respectively without and with taking other causes into account (see Chapter 1, page 34).

Several upper limits have been proposed in the context of the fixed age limit method. It has also been suggested to adopt a lower limit in order to exclude infant mortality from the definition of premature death. The approach based on life expectancy also has several variants. We will, however, only discuss the most common ones here.

Years of life lost with respect to a fixed age limit

If h is the fixed age limit, then the number of years potentially lost for an individual in age group x dying from a certain cause can be denoted by

$$h - a_x$$

where a_x is the average age of death in age group x, which is, in practice, taken as the centre of the age interval. If d_x denotes the number of deaths in age group x, then the total number of years of life lost in the population may be written

$$\text{PYLL} = \sum_{a_x < h} d_x (h - a_x) \tag{2.13}$$

and, consequently, the number of years of life lost per death on average is

$$\frac{\text{PYLL}}{\sum_x d_x}$$

which is simply $h - \bar{a}$, where $\bar{a}$ is the average age of death from the cause under consideration.

Rather than calculating this number of years per number of deaths, some authors prefer to compute years lost per number of person-years M which has produced these deaths. The number $(10^5 \times \text{PYLL})/M$ then measures the number of years of life lost in a year per 100 000 people who have the same age structure and mortality as the population under consideration. This ratio is described as the *rate of years of life lost*. Note that the index $\frac{\text{PYLL}}{M} = \frac{\sum_x d_x}{M}(h - \bar{a})$ is in fact the product of the crude mortality rate and the average number of years lost by the individuals who have died from the given cause.

The rate of life years lost can be standardized for the purpose of comparison between groups. Let m_x be the number of person-years of age x in the given group, and L_x be the number of person-years of age x in the standard population and $L = \sum_x L_x$; the standardized rate of years of life lost may then be written

$$\frac{1}{L} \sum_{a_x < h} L_x \frac{d_x}{m_x} (h - a_x) = \frac{1}{L} \left(\sum_{a_x < h} d'_x \right) (h - \bar{a}') \tag{2.14}$$

where d'_x and $\bar{a}'$ are, respectively, the age-specific number of deaths and the average age of death which would be observed in a population with the age structure of the standard population, and the mortality rate of the given group.

$$d'_x = L_x \frac{d_x}{m_x}$$

Table 2.6 Calculation of number and standardized ([a]) rate of potential years of life lost with a fixed limit at age 70 (male, lung cancer, canton of Neuchâtel, Switzerland, 1974-1976)

x	a_x	$70 - a_x$	d_x	m_x	$PYLL_x$	w_x	$10^5\ PYLL_x \frac{w_x}{m_x}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
40-44	42.5	27.5	5	15 270	137.5	0.214	192.7
45-49	47.5	22.5	4	15 102	90.0	0.214	127.5
50-54	52.5	17.5	12	14 946	210.0	0.179	251.3
55-59	57.5	12.5	18	13 044	225.0	0.143	246.7
60-64	62.5	7.5	31	10 830	232.5	0.143	307.0
65-69	67.5	2.5	38	9 843	95.0	0.107	103.3
Total			108		990.0	1.000	1 228.5

Columns 1, 4, 5, 7 are given and columns 2, 3, 6, 8 are calculated.
([a]) World population 40-69 years.

$$\bar{a}' = \frac{\sum_x d'_x a_x}{\sum_x d'_x}$$

Formula (2.14) is therefore the product of the standardized mortality rate and the average number of years of life lost in a population that would have the standard age structure and experience the mortality of the given group.

When they are calculated in this way, the rates from different causes have the advantage of being additive. In other words, the sum of the rates corresponding to several given causes is equal to the rate which is calculated from the sum of deaths due to these combined causes.

As an example, Table 2.6 presents the calculation of the years of life lost from lung cancer for Neuchâtel, Switzerland; only deaths occurring after 40 years are taken into consideration and the age limit is 70 years. Years of life lost are also expressed as rates, standardized to the European population. This example shows the weight that is given to deaths, however few in number, occurring long before the age limit.

Years of life lost with respect to life expectancy

In this situation, potential life is the number of years which would theoretically be left to live at the time of death, according to the life table.

If we let $\overset{o}{e}_x$ (see Chapter 1, page 27) be the life expectancy at the mid-point a_x of age group x, then, as previously explained, the years of life lost from a given

cause are the sum of the potential duration of life of all those who have died from this cause

$$PYLL' = \sum_{x=1}^{g} d_x \overset{\circ}{e}_x \qquad (2.15)$$

For comparisons between populations, rates and standardized rates can of course be calculated, although the justification for doing so is not obvious when the life expectancy differs among the populations being compared. Table 2.7 shows the calculation of the rates and standardized rates from data for lung cancer in Neuchâtel, taking values of life expectancy from the life table for the whole of Switzerland (see Appendix 1).

The rate obtained (2395 years per 100 000) is twice that given by the fixed limit method. The difference arises partly from the fact that deaths are taken into account at whatever age they occur, including those well after the fixed age limit. However, it also results from the fact that, for all ages less than 70 years, the life expectancy is greater than that which would be obtained with a life potential limited to 70 years. A higher fixed limit could possibly have led to the opposite conclusion.

We have stated that life expectancy implicitly took into account competing risks due to other causes that could manifest their effects at any age, including the years before the arbitrarily fixed age-limit. From this perspective, it would be more appropriate to recalculate the life expectancy at each age from a life table that excludes the deaths for which the years of life lost are calculated. This approach has some connection with the concept of additional years of life due to elimination of a cause

Table 2.7 Calculation of number and standardized rate of potential years of life lost compared to life expectancy ([a]) at age of death (male, lung cancer, canton of Neuchâtel, Switzerland – 1974-1976)

Age (x)	$\overset{\circ}{e}_x$	d_x	m_x	$PYLL'_x$	w_x	$10^5\ PYLL_x \frac{w_x}{m_x}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
40-44	32.77	5	15 270	163.9	0.188	201.8
45-49	28.24	4	15 102	113.0	0.188	140.7
50-54	23.89	12	14 946	286.7	0.156	299.2
55-59	19.85	18	13 044	357.3	0.125	342.4
60-64	16.12	31	10 830	499.7	0.125	576.8
65-69	12.78	38	9 843	485.6	0.094	463.7
70-74	9.84	29	7 656	285.4	0.063	234.9
75-79	7.35	22	5 787	161.7	0.031	86.6
80-84	5.38	13	2 721	69.9	0.015	38.5
85 +	4.59	2	1 323	9.2	0.015	10.4
Total				2 432.4	1.000	2 395.0

([a]) Swiss life table, 1978-1983. Office fédéral de la Statistique, Berne, 1985; see Annex 1.

of death (see Chapter 1, page 35). In practice, this subtlety is only necessary for the causes of death that themselves play an appreciable role in the life table, and besides, it has the disadvantage of making the procedure lose its additive property: the estimate of the years of life lost from a combination of causes could then be less than the sum of the individual estimates [11].

Methods for comparison

When we wish to compare incidence in several populations, the first step is to examine standardized rates. However, as explained in the previous section, these rates are affected by random variability. It is therefore important to know if an observed difference between two incidence curves described in this way is real or only due to chance. Knowing the confidence intervals of the rates being compared is not always sufficient to make a judgement about the difference: there exist situations in which incidence curves are significantly different even when the confidence intervals of the rates overlap.

The statistical significance of an observed difference between two rates can be roughly estimated by a method that requires only the total number of cases in both populations in addition to the two rates under study. Because it is not precise, this method, described in the next paragraph, should be reserved for use in situations in which age-specific data are unavailable. We discuss therefore in a following section the methods that are appropriate when age-specific data are available (see page 77).

Finding a statistically significant difference generally leads us to attempt to define the nature of the difference. Although age-specific rates are obtained from cross-sectional data, it is not unusual for them to differ in a constant ratio between the two populations (the proportionality assumption). When such a model (known as the *multiplicative model*) is acceptable, it is reasonable to estimate the constant factor, that is simply the *relative rate* of one population compared with another (see page 79) When it is not acceptable, the incidence ratio varies with age; this situation is known as *interaction* between group and age. On page 81, we present a general test to decide wether the assumption of proportionality is acceptable and in a following paragraph a test against the more specific assumption of increasing or decreasing trend of the incidence ratio with age; the test against the existence of a linear trend, which is the model most frequently considered, is discussed. Lastly, we give an example on page 83 to show the practical use in a complex situation of the tests that have been discussed.

In the second part of this section we deal with the problems that arise from the comparison of incidence in several populations or in different subgroups of the same population. A series of pairwise comparison of rates can actually produce contradictory results, as well as being inappropriate: by multiplying the number of comparisons that have been made, we increase the risk of concluding wrongly that

a difference is significant. We first consider an approximate method which involves the comparison of the incidence of all the subgroups of a population with a standard incidence, which is usually that of the whole population. Then the correct test for deciding wether several forces of incidence can be considered identical is introduced on page 87. In a final paragraph of this section we briefly introduce the analysis of incidence using the log-linear model which allow this type of problem to be approached in a more systematic fashion (see page 90).

Comparison of incidence of a disease in two groups

The approximate method

We can obtain a rough idea of the significance of the difference between two standardized rates when we only have these rates and the total number of individuals in the populations in which the incidence was measured.

If we were comparing crude rates, it would be sufficient to know their variances (page 51). Let t_1 and t_2 be the rates to be compared and m_1 and m_2 the person-years of observation. Since the variance of a difference of independent variables is equal to the sum of their variances, we may write

$$\text{Var}(t_1 - t_2) = \left(\frac{1}{m_1} + \frac{1}{m_2}\right)\lambda = \frac{2}{h}\lambda \qquad (2.16)$$

where λ is the theoretical common rate in the two populations and h the harmonic mean of m_1 and m_2. Then if we replace λ by its estimate under the null hypothesis

$$\hat{\lambda} = (m_1t_1 + m_2t_2) \,/\, (m_1 + m_2)$$

we can write

$$\text{Var}(t_1 - t_2) = \frac{m_1t_1 + m_2t_2}{m_1m_2} \qquad (2.17)$$

Thus, the variable

$$Z = \frac{t_1 - t_2}{\sqrt{\dfrac{m_1t_1 + m_2t_2}{m_1m_2}}}$$

has a standard normal distribution and we shall reject the hypothesis of equality of the rate in the two populations at the α = 5% significance level when |Z| is greater than 1.96.

When the rates to be compared t_1 and t_2 are standardized, the variance of the denominator calculated in this way is only an approximation to the variance of the

difference of the two rates. Writing t_1 and t_2 as an explicit function of the age-specific rates t_{1x} and t_{2x}, the expression (2.17) becomes

$$V_a(t_1 - t_2) = \sum_x w_x \frac{m_1 t_{1x} + m_2 t_{2x}}{m_1 m_2}$$

$$= \frac{2}{h} \sum_x w_x \bar{t}_x \qquad (2.18)$$

where $\bar{t}_x$ is the mean of t_{1x} and t_{2x} weighted by m_1 and m_2, the size of the groups to be compared.

The average of the $\bar{t}_x$ in (2.18) gives only a partial description of the variability of $t_1 - t_2$. Its exact variance is slightly different and is obtained from the variance of the differences of the specific rates; using formula (2.16) in each age group and replacing λ by its estimate, we get

$$V_e(t_1 - t_2) = \sum_x w_x^2 \left(\frac{2}{h_x} \hat{\lambda}_x \right) = \sum_x w_x^2 \frac{k_{1x} + k_{2x}}{m_{1x} m_{2x}} \qquad (2.19)$$

where $\hat{\lambda}_x$ is the estimate of the common rate λ_x and h_x the harmonic mean of m_{1x} and m_{2x}. Writing $w_x = L_x / h$, we get

$$V_e(t_1 - t_2) = \frac{2}{h} \sum_x w_x \frac{L_x \hat{\lambda}_x}{h_x} = \frac{2}{h} \sum_x w_x t_x^*$$

a formula which suggests that the values V_a and V_e may be close together if the structure of the standard population is not too different from that corresponding to the harmonic mean of the populations being compared.

As an example, consider the rates of stomach cancer for males in Zaragoza and Geneva, standardized to the world population restricted to the age range 35 to 74 years (see Table 2.8 and Figure 2.4). We obtain respectively $t_1 = 56.82/100\,000$ and $t_2 = 43.52/100\,000$. The approximate variance of the difference between the rates is thus (see (2.17))

$$V_a(t_1 - t_2) = \frac{(5 \times 167\,022 \times 56.82) + (5 \times 71\,298 \times 43.52)}{5 \times 71\,298 \times 5 \times 167\,022} \times 10^{-5} = 2.12 \times 10^{-9}$$

$$Z_a = \frac{|56.82 - 43.52| \times 10^{-5}}{\sqrt{2.12 \times 10^{-9}}} = 2.89$$

whereas the exact variance calculated using formula (2.18) above is

$$V_e(t_1 - t_2) = 2.11 \times 10^{-9}$$

$$Z_e = \frac{|56.82 - 43.52| \times 10^{-5}}{\sqrt{2.11 \times 10^{-9}}} = 2.89$$

Table 2.8 Cases of stomach cancer in males and population size by age group in Zaragoza, Spain, and Geneva, Switzerland. Incident cases 1973-77 [6]

x	Age	Incident cases		Population size 1975	
		Zaragoza k_{1x}	Geneva k_{2x}	Zaragoza $m_{1x}/5$	Geneva $m_{2x}/5$
1	35-39	8	10	22 801	13 506
2	40-44	8	6	27 291	12 480
3	45-49	36	7	26 762	11 012
4	50-54	54	18	25 899	9 887
5	55-59	53	17	19 853	7 010
6	60-64	96	25	17 431	6 845
7	65-69	115	35	15 024	6 066
8	70-74	145	37	11 961	4 492
Total		515	155	167 022	71 298

k_{1x} : Observed cases in age group x in Zaragoza between 1973 and 1977.
m_{1x} : number of person-years of observation in age group x in Zaragoza between 1973 and 1977.
k_{2x} and m_{2x} : similar definition for Geneva.

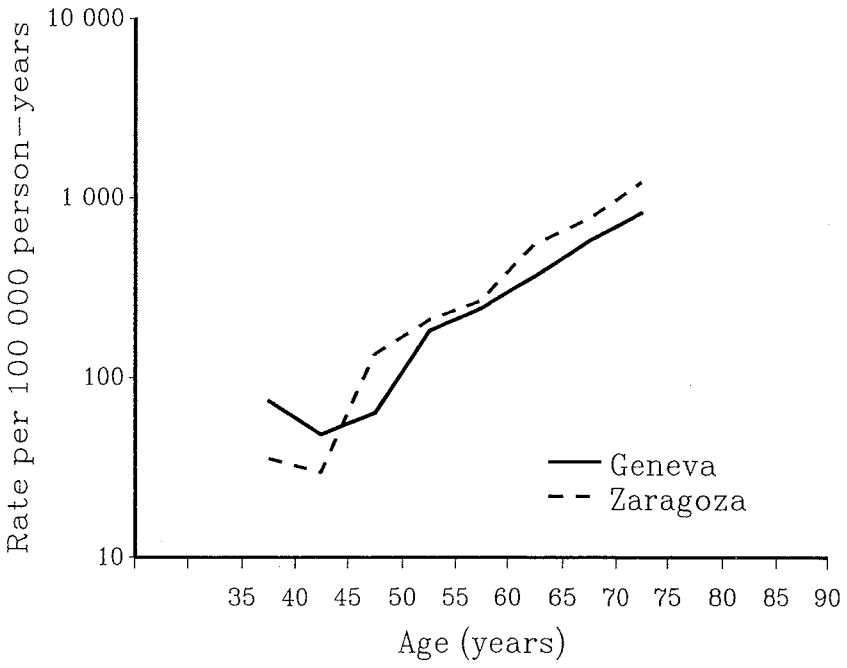


Figure 2.4 Age-specific incidence of stomach cancer in Zaragoza (Spain) and Geneva (Switzerland) males, 1973-1977

In this case, we see that the two values calculated from the variance are almost identical. The comparison of the two standardized rates by this method leads us to conclude that the incidence rate of stomach cancer observed in Zaragoza (56.82/100 000) is significantly greater than that observed in Geneva (43.52/100 000). It could, however, happen that $\bar{t}_x$ and t_x^* have different mean values.

Therefore the approximate method is not recommended when the data permit the correct calculation to be carried out.

Mantel-Haenszel test

Standardized rates have a descriptive function and the method of comparison previously proposed above is essentially aimed at avoiding gross errors in interpretation. When a comparative study of incidence is envisaged, the comparison problem should be approached in another way that requires knowing the age-specific rates and the number of person-years from which they were calculated.

Cochran [12] has shown how the performance of the χ^2 test could be improved by explicitly taking into account alternatives to the null hypothesis that we are trying to test. He proposed a method for the combination of 2×2 tables that was adapted by Mantel and Haenszel [13] in the context of case-control studies. It can also be applied with little change for the comparisons of incidence. The numerous applications of the Mantel-Haenszel method justify the amount of attention that we will give to its presentation.

Often incidence curves are approximately parallel when they are represented on a logarithmic scale. This overall shift in the curve corresponds to the fact that the ratio of the age-specific rates in the two populations being compared is more or less constant. The Mantel-Haenszel test basically involves testing the alternative assumption of proportionality of age-specific rates against the null hypothesis of equal rates.

The method involves summing the observed differences in each age group; if the differences tend to be of the same sign, as is supposed under the alternative hypothesis, their cumulative value will not be compatible with the null hypothesis of equality of age-specific rates. Small differences can thus be identified more easily whereas, if they were considered individually or incorporated into a sum of squared differences, no conclusions could be drawn.

Suppose that the hypothesis of equal rates is true. Then, apart from random variation, the total number of observed cases in each age-group is divided between the two populations in proportion to the number of person-years accumulated in each one. Summing these expected numbers over all age groups will provide the overall expected difference between the two populations which must be compared to the overall observed difference. Since the total number of expected cases is made equal to the total number of observed cases, it is sufficient, in practice, to calculate the difference between the total number of cases observed and the total number of cases expected under the hypothesis of equal rates in just one of the populations. We illustrate this method using data presented in Tables 2.8 and 2.9.

If we use data for the second population, that is, in Geneva (Table 2.8), the number of cases expected in age group x is

$$e_{2x} = K_x \frac{m_{2x}}{M_x} \qquad (2.20)$$

where $M_x = m_{1x} + m_{2x}$ and $K_x = k_{1x} + k_{2x}$. The test is then based on the overall difference between observed and expected cases in the second population, that is, if g age groups are used:

$$T = \sum_{x=1}^{g} (k_{2x} - e_{2x}) = O_2 - E_2$$

It is then evident from this latter formula that the statistic T is designed to detect systematic differences of the same sign between the observed and expected numbers in the different age-groups. In order to find out if the value of the statistic is significantly different from zero, we need to know its variability under the null hypothesis. Under this hypothesis, the total information available on the common rate λ_x in age group x is contained in the variable K_x. Therefore, K_x being fixed at its observed value, the statistical distribution of the number of cases in the age group x of the second population is independent of λ_x; it may be described as the result of K_x independent choices between the two populations with probability m_{2x}/M_x that the second population is chosen. In other words, k_{2x} has a binomial distribution with mean e_{2x} and variance

$$\text{Var}(k_{2x}) = K_x \frac{m_{2x}}{M_x} \frac{m_{1x}}{M_x} \qquad (2.21)$$

and consequently, the variance of the statistic T is

$$\text{Var}(T) = \sum_{x=1}^{g} \text{Var}(k_{2x} - e_{2x}) = \sum_{x=1}^{g} K_x \frac{m_{2x} \times m_{1x}}{M_x^2}$$

$Z = T/\sqrt{\text{Var}(T)}$ approximately follows a standard normal distribution; thus, if we observe an absolute value of Z greater than 1.96, we can reject the null hypothesis of equality of rates at the 0.05 level (two-sided test).

This statistic has low power if the alternative hypothesis is not the one specified above; for example, an incidence that is clearly higher at young ages and clearly lower in older age groups might give a result which is not statistically significant, even though the null hypothesis is not true. The test is actually much less effective the further one moves away from the assumption of proportionality of rates. We examine its use in particular situations, notably when curves cross over, on page 83.

Table 2.9 gives the various steps of the calculation of the Mantel-Haenszel test, using the data presented in Table 2.8.

The value of the statistic Z is therefore:

$$Z = \frac{155 - 188.7}{\sqrt{135.3}} = -2.90$$

The differences observed cannot therefore be attributed to random variation and we can conclude that the incidence of stomach cancer is higher in Zaragoza

Table 2.9 Comparison of incidence rates in two populations; Mantel-Haenszel test. Data from Table 2.8

x	$M_x/5$ (1)	K_x (2)	k_{2x} (3)	e_{2x} (4)	$Var(k_{2x})$ (5)
1	36 307	18	10	6.7	4.2
2	39 771	14	6	4.4	3.0
3	37 774	43	7	12.5	8.9
4	35 786	72	18	19.9	14.4
5	26 863	70	17	18.3	13.5
6	24 276	121	25	34.1	24.5
7	21 090	150	35	43.1	30.7
8	16 453	182	37	49.7	36.1
Total	238 320	670	$O_2 = 155$	$E_2 = 188.7$	135.3

than in Geneva. Note that in this case the value of |Z| only differs slightly from that obtained by the approximate method (see page 75).

Overall measure of incidence ratio

When the multiplicative model is acceptable, the rate ratio of the two populations is independent of age:

$$\frac{\lambda_{2x}}{\lambda_{1x}} = \rho$$

It is therefore natural to try to estimate ρ. Mantel and Haenszel have proposed a weighted average of the ratio of the age-specific rates which proved to be very efficient:

$$\hat{\rho} = \frac{\sum_{x=1}^{g} \frac{k_{2x}\, m_{1x}}{M_x}}{\sum_{x=1}^{g} \frac{k_{1x}\, m_{2x}}{M_x}} \qquad (2.22)$$

From data in Table 2.8 and from intermediate calculations presented in the first two columns of Table 2.10, we obtain

$$\hat{\rho} = \frac{110.73}{144.46} = 0.766$$

which means that the risk of stomach cancer is 1.3 times ($1/\hat{\rho}$) greater in Zaragoza than in Geneva. We can easily calculate a confidence level for ρ, although it would mainly be of theoretical interest in the context of most descriptive analysis.

Table 2.10 Calculation of the relative risk of stomach cancer in Geneva, Switzerland, with Zaragoza, Spain, as baseline. Data from Tables 2.8 and 2.9

x	$\frac{k_{2x}\, m_{1x}}{M_x}$ (1)	$\frac{k_{1x}\, m_{2x}}{M_x}$ (2)	$\frac{m_{1x} + \hat{\rho}\, m_{2x}}{5}$ (3)	$\frac{K_x\, m_{1x}\, m_{2x}}{M_x\,(m_{1x} + \hat{\rho}\, m_{2x})}$ (4)
1	6.28	2.98	33 146.60	4.61
2	4.12	2.51	36 850.68	3.25
3	4.96	10.49	35 197.19	9.53
4	13.03	14.92	33 472.44	15.39
5	12.56	13.83	25 222.66	14.38
6	17.95	27.07	22 674.27	26.23
7	24.93	33.08	19 670.56	32.95
8	26.90	39.59	15 401.87	38.59
Total	110.73	144.47	221 636.27	144.93

In fact, the variance of $\text{Log}(\hat{\rho})$ is approximately [14-16]:

$$V = \frac{\sum_{x=1}^{g} \text{Var}\,(k_{2x})}{\hat{\rho}\left(\sum_{x=1}^{g} \frac{K_x m_{1x} m_{2x}}{M_x(m_{1x} + m_{2x}\hat{\rho})}\right)^2} \qquad (2.23)$$

which, using the data in Table 2.9 (column 5) and 2.10, gives

$$V = \frac{135.3}{0.766\,(144.93)^2} = \frac{135.3}{16\,094.95} = 0.0084$$

from which we obtain the standard error $\sqrt{V} = 0.0917$.

Considering that $\text{Log}\,(\hat{\rho})$ has a normal distribution with mean $\text{Log}(\rho)$ and variance V, a confidence interval $[\rho_1 ; \rho_2]$ at the $(1 - \alpha)$ level can then be derived as

$$\left[\hat{\rho}\, e^{-Z_{\alpha/2}\sqrt{V}} ;\ \hat{\rho}\, e^{+Z_{\alpha/2}\sqrt{V}}\right]$$

which gives, for $\alpha = 0.05$, the lower and upper confidence bounds, respectively:

$$\left[\hat{\rho}\, e^{-1.96\sqrt{V}} ;\ \hat{\rho}\, e^{1.96\sqrt{V}}\right]$$

in the above example

$$\rho_1 = 0.766 \times 0.835 = 0.64$$

and

$$\rho_2 = 0.766 \times 1.1969 = 0.92$$

Test of a multiplicative model

The assumption of proportionality also can be tested using the same principle as before. Under the hypothesis of a constant relative risk regardless of the age group, the means of the Poisson distributions in the two populations for age group x are respectively $\lambda_x m_{1x}$ and $\rho\lambda_x m_{2x}$, where λ_x is the age-specific rate in the first population and ρ is the rate ratio. The K_x cases observed will tend to be distributed among the two populations in proportion to these values, so that,using the same principle as in the previous paragraph,

$$k_{2x} \rightsquigarrow \text{Binom}\ (K_x, p_x)$$

where

$$p_x = \frac{\rho\lambda_x m_{2x}}{\lambda_x m_{1x} + \rho\lambda_x m_{2x}} = \frac{\rho m_{2x}}{m_{1x} + \rho m_{2x}} \qquad (2.24)$$

Therefore, under the assumption of proportionality, the expectation and variance of the number of cases in age group x of population 2 are now dependent on ρ and are respectively:

$$e_{2x}(\rho) = K_x p_x$$

$$\text{Var}\ (k_{2x}\ ;\ \rho) = K_x p_x (1 - p_x)$$

and will be estimated by replacing ρ in (2.24) by $\hat{\rho}$ given by (2.22).

If the hypothesis of a constant risk ratio is not true, we will observe substantial differences between the observed and the expected numbers of cases in some age groups; overall, these differences will be detected by the sum of standardized squared differences d_x^2 in each age group,

$$X^2 = \sum_x d_x^2 = \sum_x \left[\frac{[k_{2x} - e_{2x}(\hat{\rho})]^2}{\text{Var}\,(k_{2x}\ ;\ \hat{\rho})} \right] \qquad (2.25)$$

Table 2.11 Calculation for interaction tests. Data from Tables 2.8, 2.9 and 2.10

x (1)	k_{2x} (2)	$e_{2x}(\hat{\rho})$ (3)	$\text{Var}\,(k_{2x}\ ;\ \hat{\rho})$ (4)	d_x^2 (5)	(1)× (2–3) (6)	(1)× (4) (7)	(1)× (7) (8)
1	10	5.62	3.87	4.96	4.38	3.87	3.87
2	6	3.63	2.69	2.09	4.74	5.38	10.76
3	7	10.31	7.84	1.40	− 9.93	23.52	70.56
4	18	16.29	12.60	0.23	6.84	50.40	201.60
5	17	14.90	11.73	0.38	10.50	58.65	293.25
6	25	27.98	21.51	0.41	−17.88	129.06	774.36
7	35	35.43	27.06	0.01	− 3.01	189.42	1 325.94
8	37	40.66	31.58	0.42	−29.28	252.64	2 021.12
Total	155	154.82	118.88	9.90	−33.64	712.94	4 701.46

which is approximately distributed as a χ^2 with g-1 degrees of freedom. This test is also known as the *homogeneity test.*

In the above example, its value can be calculated from Table 2.11 (column 5):

$$\sum_x d_x^2 = 9.90$$

as this value is lower than the critical value 14.07 at the significance level $\alpha = 0.05$ for a χ^2 with seven degrees of freedom, we cannot reject the null hypothesis of proportionality.

Trend test

The test with (g – 1) degrees of freedom described above is not very sensitive to small departures from proportionality; nevertheless, even small differences can be interpretable if they increase or decrease systematically with age. If such a situation is expected, it is preferable to use a trend test (with one degree of freedom) which is aimed more specifically at this alternative hypothesis. The relevant statistic is given by the weighted sum of the differences between observed and expected numbers

$$T = \sum_{x=1}^{g} u_x \left[k_{2x} - e_{2x}(\hat{\rho})\right]$$

where u_x varies with age according to a specified structure; for example, it could be assigned the age group's number if one was allowing for a linear divergence of the two curves with age.

We can show that

$$\text{Var}(T) = \sum_{x=1}^{g} u_x^2 \,\text{Var}(k_{2x}; \hat{\rho}) - \frac{\left[\sum_{x=1}^{g} u_x \,\text{Var}(k_{2x}; \hat{\rho})\right]^2}{\sum_{x=1}^{g} \text{Var}(k_{2x}; \hat{\rho})} \qquad (2.26)$$

$Z = T/\sqrt{\text{Var}(T)}$ is a standard normal variable that we will use to test for the alternative hypothesis specified by the series of coefficients u_x; this test is also known as the Armitage test [17]. Details of the calculations are presented in Table 2.11 (columns 6 to 8); from these data we obtain

$$Z = \frac{-33.64}{\sqrt{4\,701.5 - \frac{(712.9)^2}{118.9}}} = -1.63$$

The hypothesis of proportionality can therefore not be rejected even when the alternative hypothesis is more narrowly specified. However, the value of Z is rela-

tively high; this can be understood well enough by examining Figure 2.4 where we can see that, because incidence is initially higher in Geneva, there is a slight departure from the null hypothesis of proportionality.

Example : Hodgkin's lymphoma

The methods that we introduced above might seem unnecessarily sophisticated for estimating differences as obvious as those which appear between Zaragoza and Geneva with regard to stomach cancer. Their usefulness does not appear in routine contexts, but is apparent in borderline or complex situations. For example, a more precise method is needed to interpret population differences when incidence in different periods of life is described by different models. The above approach may then be extremely useful. To illustrate this idea, consider the comparison of incidence of Hodgkin's disease for males in Connecticut and the province of Zaragoza for the time period 1973 to 1977 [7] (see also Figure 2.5).

If we use the method described on page 76, we obtain a value Z = 0.56 for the Mantel-Haenszel test, which tempts us to conclude that there is no difference in incidence between the two populations. Note also that the standardized rates (respectively 3.8 and 4.0 per 100 000 in Zaragoza and Connecticut) yield the same interpretation. On the other hand, one should be warned by the high value (53.65 with seventeen degrees of freedom) obtained with the homogeneity test, suggesting that the incidence curves very likely cross; this phenomenon, which can be clearly

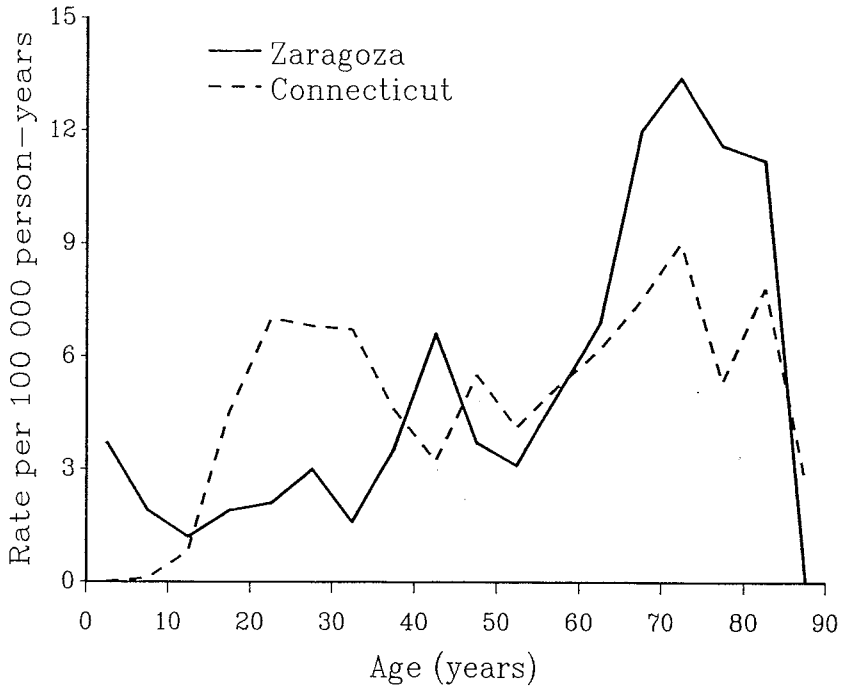


Figure 2.5 Age-specific incidence of Hodgkin's disease in Zaragoza (Spain) and Connecticut (USA) males, 1973-1977

seen on the graph in Figure 2.5 is not *a priori* surprising since we know that Hodgkin's disease has at least two forms with different etiologies. It is then natural that we should look separately at differences in younger age groups and in older age groups.

If we use the Mantel-Haenszel test on the age groups from one to 24 years, we obtain 85 observed cases as opposed to 83.22 expected in Connecticut and a value of $\hat{\rho} = 1.13$ with $Z = 0.45$. Although the test is not significant, given the appearance of the two incidence curves, it is still advisable to continue the analysis using the methods presented page and ; the homogeneity test gives a value of 38.85 with four degrees of freedom ($p < 0.001$), and the linear trend test gives 5.65 for one degree of freedom. This last value means that the difference between observed and expected numbers increased significantly with age. The observed and expected values under the hypothesis of parallel curves with $\hat{\rho} = 1.13$ are shown in Table 2.12 and, from close examination, it can be clearly seen why the hypothesis of proportionality is not justifiable. Actually, the disease is significantly more frequent in children in Zaragoza ($Z = -5.40$ with the Mantel-Haenszel test performed on the first two age groups); a reversal of risk takes place at adolescence. In Connecticut, the risk is significantly higher for young adults: if we restrict our analysis to age groups 20-34 years, the disease is three times more frequent in Connecticut ($(\hat{\rho} = 3.06$, $Z = 3.41$). On the other hand, differences between the two countries are no longer observed after 35 years: the homogeneity test gives values of $Z = -1.05$ and $\chi^2_{10} = 6.59$. These diverse results force us to suspect that Hodgkin's disease might involve a group of three pathological entities with different etiologies and not two as was previously assumed [18]. The observed difference could also originate in different definitions of the disease in the two countries.

The example demonstrates that the procedures introduced in this section can be valuable tools to help avoid erroneous interpretations when random variation are substantial and when the pattern of incidence deviates markedly from the simple shapes observed for epithelial tumours. They must nevertheless be applied with caution and their use be motivated by biological hypotheses defined *a priori*.

Table 2.12 Hodgkin's disease in Connecticut (USA) and Zaragoza (Spain). Male, 1973-1977 [7]

Age	Connecticut			Zaragoza
	Observed cases	Expected cases for $\hat{\rho} = 1.13$	Rate	Rate
1-4	0	4.74	0.00	3.67
5-9	1	3.38	0.13	1.89
10-14	6	6.65	0.85	1.23
15-19	32	29.20	4.51	1.88
20-24	46	41.07	7.02	2.10

Comparison of incidence among several populations

Often in descriptive epidemiology we have to interpret differences in incidence among a series of populations, or subgroups of the same population. This is a standard procedure when routinely published data are studied. Therefore, the analysis has no longer the goal of studying specific differences between a few given groups. Its objective is instead to find all differences which may exist. We present below the standard methods that can evaluate whether each incidence rate in a series of groups or populations is significantly different from an overall expected value. The problem with these methods, like all those which involve multiple comparisons, is that they are bound to identify some differences produced by random fluctuations as being significant. It is therefore preferable to use a test that provides an overall assessment of the homogeneity of incidence. This is introduced on page 87. We shall also discuss in Chapter 3 (see page 134) other methods which are appropriate in this context.

Comparison with an overall expected value

If the total number of cases is available in a subpopulation whose age structure is known, then it is possible to check if this observation is compatible with a given incidence rate, such as the incidence rate of the whole population. It is straightforward to use this incidence rate to calculate the number of expected cases in each age group, their total E, the SIR and its confidence interval in the subpopulation. We will take it that the SIR is different from 100 when its confidence interval does not include 100 (see page 64). When the total number O of observed cases is sufficiently large, the normal approximation to the Poisson distribution can be used. In other words, we consider that O is a normal variable with expectation E and variance E; accordingly, we can calculate the quantity: $X^2 = \frac{(O-E)^2}{E}$ which follows a χ^2 distribution with one degree of freedom.

Because of its simplicity, this method is often used systematically to find out if the incidence rate in selected subpopulations deviates significantly from the total population incidence rate, as though this incidence were known *a priori* and not calculated from the observations themselves.

To illustrate the method, let us consider the regional subdivisions of the French département of Côte-d'Or that is covered by the Burgundy Registry of digestive tract tumours. The number of cases of colon cancer observed in each five-year age group, from 1976 to 1980, as well as the number of person-years accumulated in each age-group for the same period are summarized in Table 2.13. The total number of observed cases in each region, and the calculations of expected value under the hypothesis that the rates in the whole département of Côte-d'Or apply to each region of the département, are given in Table 2.14.

Table 2.13 Colon cancer cases and person-years ([a]) in Côte-d'Or, France, Male, incident cases 1976-1980

	Dijon		Côte Viticole		Châtillonnais		Plaine de la Saône		Auxois		Morvan		Côte-d'Or (Total)	
x	k_x	m_x	k_x	m_x	k_x	m_x	k_x	m_x	k_x	m_x	k_x	m_x	k_x	m_x
0-4	0	45 626	0	9 273	0	7 309	0	19 970	0	8 094	0	905	0	91 177
5-9	0	41 145	0	9 653	0	9 683	0	24 135	0	9 500	0	1 171	0	95 287
10-14	0	39 284	0	10 255	0	10 758	0	23 297	0	10 483	0	1 819	0	95 896
15-19	0	43 469	0	11 054	0	10 360	0	21 195	0	10 870	0	1 770	0	98 718
20-24	0	52 794	0	10 073	0	8 402	0	19 034	1	9 539	0	1 186	1	101 028
25-29	0	54 321	0	10 499	0	7 984	1	19 009	0	7 936	0	1 345	1	101 094
30-34	0	40 848	0	8 588	0	7 482	0	19 040	0	6 633	0	979	0	83 570
35-39	2	31 559	1	6 459	0	6 707	0	16 017	0	6 012	1	779	4	67 533
40-44	2	30 703	0	6 860	0	7 273	0	15 762	1	6 775	0	1 120	3	68 493
45-49	3	29 875	1	7 181	0	6 869	3	15 242	3	7 377	0	1 481	10	68 025
50-54	10	27 228	2	6 955	0	7 313	1	13 265	1	7 489	0	1 422	14	63 672
55-59	17	21 808	0	5 636	1	5 211	5	10 951	1	6 066	0	1 275	24	50 947
60-64	7	15 002	2	4 041	0	4 638	5	8 249	3	4 729	0	975	17	37 634
65-69	17	14 556	4	5 203	3	6 322	7	10 287	4	6 350	1	1 595	36	44 313
70-74	33	11 841	4	4 310	7	5 129	18	8 397	13	5 764	1	1 336	76	36 777
75-79	20	7 762	3	2 425	10	3 375	11	5 625	6	4 187	2	825	52	24 199
80 +	12	6 112	10	1 894	4	2 167	11	3 913	3	2 795	0	604	40	17 485
Total	123	513 933	27	120 359	25	116 982	62	253 388	36	120 599	5	20 587	278	2 245 848

([a]) Person-years of observation were calculated by summing the mid-year populations from 1976 to 1980.

Table 2.14 Calculation of the SIRs in the different regions of Côte-d'Or (France) with the overall incidence in the département as standard, males, colon cancer, 1976-1980

Region	Observed number	Expected number	SIR	95% confidence interval ([a])
Dijon	123	98.7	124.6	[103.6 ; 148.7]
Côte viticole	27	30.6	88.2	[58.1 ; 128.4]
Châtillonnais	25	36.0	69.4	[44.9 ; 102.5]
Plaine de Saône	62	62.8	98.7	[75.7 ; 126.6]
Auxois	36	41.0	87.8	[61.5 ; 121.6]
Morvan	5	8.9	56.2	[18.2 ; 131.1]

([a]) Exact method (Poisson distribution).

As the confidence interval of the SIR for Dijon excludes 100, we conclude that the incidence of colon cancer is higher here than in the whole département. We could also have tested the observed difference by calculating χ^2 with one degree of freedom; its observed value $(123 - 98.7)^2/98.7 = 5.98$ leads to the same conclusion. However, observations in the other cantons of Côte-d'Or are compatible with the overall incidence in this département.

Note that an analysis of the SIR without an indication of its precision would not be sufficient to provide the correct conclusion about the variation of incidence in the region. For example, the value of 56.2, that appears to indicate that Morvan is a low-risk area, is actually only due to the low value of the expected number which, in turn, implies large random variation in the observed number. In this case, the probability of obtaining five or fewer cases simply by chance, when the expected number is E = 8.9, is actually 13%, therefore too high to reject the null hypothesis of equality of the incidence rate in Morvan and in the whole département.

Although interpretation of the values obtained for the different SIRs is much more convincing when their confidence intervals are taken into account, the method is still approximate. In fact, the incidence for the whole of the Côte-d'Or that is used as a standard is calculated from observations made in the different subgroups; the SIR obtained for each of the subgroups is by definition systematically closer to unity than it would be if the standard incidence had been defined *a priori*. To avoid this problem, which is more significant when the subpopulation consists of a larger proportion of the total, some authors have proposed taking as a standard the incidence in the population complementary to the subpopulation for which the SIR is calculated. In other words, to use the incidence in all of the other populations as the standard incidence. As the variability of the rates in the complementary population is not taken into account, this approach is unfortunately not much more satisfying. The first approach is conservative, as it too often tends to favour the null hypothesis, while the second method is too liberal as it often wrongly rejects the null hypothesis.

Homogeneity test for incidence

The appropriate method is actually quite similar in conception to that previously described for the situation of two populations (see page 77). Its principle has been mainly applied to survival analyses (log rank test, see Chapter 4, page 247) and case-control studies, but its application to descriptive incidence or mortality data is also straightforward.

If the theoretical incidence is the same in all groups, the total number of observed cases K_x in each age group x would be divided among the different groups in proportion to the person-years accumulated in each of them. It can then be shown that the distribution of observed cases follows a multinomial distribution. To be defined completely, the distribution should be specified by the expected number in each group and by the variance-covariance matrix which quantifies not only the variability but also the correlation of the observed numbers in these groups.

Letting

- I be the number of subgroups to be compared ($1 \leq i \leq I$),
- k_{ix} be the number of observed cases in the xth age interval of the ith subgroup,

- m_{ix} be the number of person-years accumulated in the xth age interval of the ith subgroup,
- $K_x = \sum_{i=1}^{I} k_{ix}$ and $M_x = \sum_{i=1}^{I} m_{ix}$ the total number of cases and person-years in age-group x,

the mean and the variance of the observed number of cases in each age interval of each subgroup may be written ($1 \leq i \leq I$ and $1 \leq x \leq g$):

$$e_{ix} = \frac{K_x m_{ix}}{M_x} \tag{2.27}$$

and

$$\text{Var}(k_{ix}) = \frac{K_x m_{ix} (M_x - m_{ix})}{M_x^2} \tag{2.28}$$

Furthermore, the covariance between observations in two subgroups is

$$\text{Cov}(k_{ix}, k_{jx}) = \frac{-K_x m_{ix} m_{jx}}{M_x^2} \tag{2.29}$$

As was done in the situation of two populations, we sum the quantities e_{ix} over all age groups to obtain the expected numbers E_i in subpopulation i. The variance and covariance of the observed numbers calculated under the assumption of equality of incidence are also summed over the age groups in order to obtain the variance-covariance matrix of the total number of cases in the subpopulations. The expected numbers are obviously the same as those given in Table 2.14, which were also defined by the overall incidence rate in the département of the Côte-d'Or:

$$E_i = \sum_x \frac{K_x m_{ix}}{M_x}$$

Table 2.15 gives the variance-covariance matrix **V** of the observed numbers O_i; it shows on the one hand that the variances are lower than the expected numbers. In other words, they are lower than the variance under the Poisson distribution; on the other hand, the table shows that all the covariances are negative, a predictable result since the total observed number in age group x is fixed at its observed value K_x (see (2.29)). If the observed numbers had themselves been allocated in the various populations according to a multinomial distribution, we would have the classic χ^2 test obtained from the normal approximation to the multinomial. Thus, we would calculate the test statistic

$$T_1 = \sum_{i=1}^{I} \frac{(O_i - E_i)^2}{E_i} \tag{2.30}$$

Table 2.15 Variance-covariance of the observed numbers of colon cancer cases in Côte-d'Or, France, under the hypothesis of risk homogeneity ([a]) within the département. Data from Table 2.13

Dijon	63.03					
Côte viticole	− 10.85	27.24				
Châtillonnais	− 12.59	− 3.97	31.27			
Plaine de la Saône	− 22.22	− 6.92	− 8.15	48.6		
Auxois	− 14.30	− 4.52	− 5.38	− 9.29	34.83	
Morvan	− 3.09	− 0.99	− 1.17	− 2.02	−1.34	8.61

([a]) The variance of observed numbers is on the diagonal. The covariance of one region with the regions preceding it in the first column is under the diagonal. For example, in Auxois, the variance is 34.83; the covariance of observed numbers in Auxois and Morvan is −1.34.

which, in the present example, is 12.10, a value that is greater than 11.07, the 5% critical value of χ^2 with five degrees of freedom. This leads us to reject the hypothesis of homogeneity of the incidence rates in the six cantons of Côte-d'Or.

However, as the total number of cases K_x is fixed, the O_i are distributed as the sum of multinomial variables and T_1 is on average smaller than χ^2 with I−1 degrees of freedom. The appropriate calculation is based on another quadratic function T_2 of the $(O_i - E_i)$ where these differences are weighted inversely to their variances. Calculation of this statistic therefore requires the inverse of the variance-covariance matrix of the differences $O_i - E_i$; the elements w_{ij} of this inverted matrix provide the necessary weights. The statistic can thus be written:

$$T_2 = \sum_{i=1}^{I-1} w_{ii} (O_i - E_i)^2 + 2 \sum_{i<j} w_{ij} (O_i - E_i) (O_j - E_j) \qquad (2.31)$$

Note that the restriction of the sum to the first I−1 populations is related to the same principle involved in the Mantel-Haenszel test where only one group is used for calculating the test statistic. Because the sum of O_i is fixed, the last region does not contribute any further information to the test. The matrix inversion can be computed with readily available software. In the present example, the weights are provided by the inverse of the matrix in Table 2.15 and the statistic T_2 has a value of 12.25 which follows a χ^2 distribution with five degrees of freedom and, like T_1, leads us to reject the homogeneity hypothesis. In this situation, the calculation of T_1 would have been sufficient.

In practice, we often need to find the basis for this demonstrated heterogeneity, particularly to determine whether one or a few regions are responsible for the statistical significance of the test. The appropriate tool to answer the question is similar to a trend test with one degree of freedom; $\sum_i u_i O_i$ is compared with its expectation $\sum_i u_i E_i$ where the coefficients u_i which equal + 1, −1 or zero are chosen such that

the statistic will enhance the contrast between the regions which are suspected to be different for *a priori* reasons. We thus calculate the statistic

$$T_3 = \frac{\left[\sum_i u_i (O_i - E_i)\right]^2}{\sum_i u_i^2 E_i - \sum_x \frac{\left(\sum_i u_i e_{ix}\right)^2}{K_x}} \qquad (2.32)$$

where the denominator, which is **U'VU** in matrix notation, is the variance of $\sum_i u_i O_i$.

For example, for comparing Morvan (i = 6) with the rest of Côte-d'Or, we set $u_6 = 1$ and $u_i = -1$ if i is different from 6. We obtain $T_3 = 1.77$, a value which is not significant. The use of the same principle to compare the city of Dijon with the rest of the département gives $T_3 = 9.4$, a highly significant value for χ^2 with one degree of freedom ($p = 0.002$). For Châtillonnais, we obtain a borderline value, that is, $\chi_1^2 = 3.86$. Although formally significant, a value of this kind should be treated with caution because the multiplicity of the tests carried out increases the chance of wrongly rejecting the hypothesis of equality. Strictly speaking, the test has one degree of freedom only if the comparisons result from hypotheses defined *a priori*. For example, if the subgroups could be characterized according to a sociodemographic variable, such as the average income, a test with a single degree of freedom could be carried out by choosing for the u_i the rank of the regions after ordering them according to the value of this variable. In the same manner, if we wanted to compare northern and southern areas of a region, we could perform the test choosing $u_i = 1$ for the north and $u_i = -1$ for the south.

A further hypothesis which could be considered in the context of this example is whether the rural regions (all except Dijon) are homogeneous with respect to the incidence of colon cancer. The above approach would lead to a χ^2 with four degrees of freedom with the value 3.26 for the test of homogeneity of incidence in rural areas. The conclusion of the analysis is therefore that the incidence is different in the rural and urban regions of the département (see below).

Use of the log-linear model

The analysis of descriptive incidence data can also be conducted with modelling techniques that allow for greater flexibility in interpretation. As a rule, the idea is to look for a model which provides the estimate of the parameters of interest in particular the relative rate and to select the simplest among those that are statistically compatible with the observations. This approach is particularly easy with access to modern computer software.

The linear regression, a widely used statistical tool, consists of modelling the expectation of a normal variable, using a linear function of the covariates that influence its value (see Chapter 3, page 158). it has been proposed to generalize this technique to other probability distributions, including the binomial distribution and the Poisson distribution. It can be shown that, in order to obtain the optimal statistical properties, it is more effective to model a function of expectation rather than expectation itself; thus, for a binomial distribution, the logit of the probability is modelled, and for the Poisson distribution it is the logarithm of the mean which is modelled as a linear function of the relevant covariables.

The observations in the context of this manual are most often Poisson variables, whose expectation depends on the unknown incidence rate and person-years of observation according to the formula

$$E(K_{ix}) = m_{ix}\lambda_{ix}$$

that is

$$\text{Log}\,[E(K_{ix})] = \text{Log}\,(m_{ix}) + \text{Log}\,(\lambda_{ix})$$

The aim of this section is to show how $\text{Log}(\lambda_{ix})$ can be modelled linearly to provide most of the results which have been previously presented. The hypothesis of proportional incidence rates that has been introduced on several occasions may be written

$$\lambda_{2x} = \rho\lambda_{1x}$$

thus

$$\text{Log}\,(\lambda_{2x}) = \text{Log}\,(\lambda_{1x}) + \text{Log}\,(\rho) \qquad (2.33)$$

Formula (2.33) is therefore a particular *log-linear model* which describes the incidence rate in group 1 (λ_{1x}) and the relative rate ρ of group 2 with respect to group 1. It can easily be generalized to more than two groups in the following form:

$$\text{Log}\,(\lambda_{ix}) = \text{Log}\,\lambda(_{1x}) + \text{Log}\,(\rho_i) \quad 2 \leq i \leq I \qquad (2.34)$$

where ρ_i is the relative rate of group i with respect to group 1. In practice, $\beta_x = \text{Log}\,(\lambda_{1x})$ and $\theta_i = \text{Log}(\rho_i)$ are estimated by the maximum likelihood method, then $\hat{\lambda}_{1x}$ and $\hat{\rho}_i$ are derived by exponentiation. In the present situation involving two factors, age and subgroups $\hat{\lambda}_{1x}$ and $\hat{\rho}_i$ are in fact given by close formulae

$$\hat{\rho}_i = \frac{O_i}{\sum_{x=1}^{g} m_{ix}\hat{\lambda}_{1x}} = \frac{O_i}{E_i} \qquad (2.35)$$

$$\hat{\lambda}_{1x} = \frac{\sum_i k_{ix}}{m_{1x} + \sum_{i>1} \hat{\rho}_i\, m_{ix}} \qquad (2.36)$$

where O_i is the total number of observed cases in group i and E_i is the expected number, taking $\hat{\lambda}_{1x}$ as a standard. It can be seen that this method provides a statistic related to the SIR; it serves the same purpose of locating the subpopulation on the risk scale. It is known as the internal method of standardization [19,20]. The special role given to the first subgroup is obviously the result of an arbitrary choice. An appropriate computer programme is required to estimate the parameters by the maximum likelihood method; the calculations reported below have been carried out using the program GLIM [21] and are described in detail in Appendix 2.

When the rates of stomach cancer in Geneva and Zaragoza are compared, the value of the parameter $\hat{\rho}$ is found to be 0.77 which means that there is about 30% more stomach cancer in Zaragoza. This value can be compared with results obtained from other methods previously presented in this chapter:

- SIR, using the marginal incidence rate as standard is

$$\frac{155}{188.73} \div \frac{515}{481.26} = 0.77$$

- $\hat{\rho}$ according to Mantel-Haenszel formula: 0.77
- Ratio of cumulative rates: 2.38/3.20 = 0.74
- Ratio of rates standardized to world population:

$$\text{CIF} = 43.52/56.82 = 0.77$$

When the two incidence curves are parallel, as in this example (see Figure 2.4), these various estimates are close together. It is however recommended to use the internal standardization, i.e., the log-linear model, which has optimal statistical properties in this context or to use the Mantel-Haenszel estimate which has been shown to be particularly robust.

The validity of the model (2.34) may be judged by comparing observed values k_{ix} and values $\hat{k}_{ix}$ calculated from the model itself. The ordinary goodness of fit statistic $T = \sum_{i,x} \frac{(k_{ix} - \hat{k}_{ix})^2}{\hat{k}_{ix}}$ may be used for this purpose. The measure of goodness of fit may also be based on the ratio between the likelihood of the accepted model and the likelihood of a model that would describe the observations exactly; this latter is known as a saturated model. This statistic

$$D = -2 \text{ Log}[V(\text{model}) / V(\text{saturated model})]$$

is referred to as the deviance. In the context of the classical linear model with normal error, it coincides with the above χ^2 for goodness of fit T. In the present situation, both T and the deviance D have a chi-squared distribution whose number of degrees of freedom is the number of observations h less the number of estimated parameters ν

$$T, D \rightsquigarrow \chi^2_{h-\nu}$$

When testing the goodness of fit of the proportional hazards model to the data from Geneva and Zaragoza, we obtain D = 9.392. The corresponding number of degrees of freedom is 7: 16 observations minus 9 fitted parameters (eight age groups + the relative risk). This value suggests an acceptable fit (p = 0.23): the difference between the values estimated by the model and the observed values is of an order of magnitude compatible with the random fluctuations allowed for by the Poisson distribution.

An hypothesis about the value of a parameter, for example, $\rho = 1$, can be tested by evaluating the significance of the increase in deviance which results from giving the tested value to the parameter of interest. When the increase is too large the proposed value is rejected. Thus, the comparison of deviance between the two models: (1): $\lambda_{2x} = \rho\lambda_{1x}$ and (2): $\lambda_{2x} = \lambda_{1x}$ is equivalent to the test of the hypothesis $\rho = 1$. In practice, the more general model is fitted (model 1) and the increase in deviance evaluated by fitting the restricted model (model 2). The calculations for the above examples are listed in Appendix 2.

When fitting model 2 to the present data, the deviance changes from 9.392 to 18.14. The difference of 8.75, value of a χ^2 variable with one degree of freedom, is highly significant and leads to reject the hypothesis of equality of the incidence rates ($\rho = 1$).

The variance and covariance of the parameter estimates are also derived from the likelihood (see Chapter 1, page 17). The variable

$$Z = \frac{\text{Log}(\hat{\rho}) - \text{Log}(\rho)}{\sqrt{\text{Var}(\text{Log}(\hat{\rho}))}}$$

is approximately a standard normal variate. We can then construct a $100(1-\alpha)\%$ confidence interval:

$$\text{Log}(\hat{\rho}) \pm Z_{\alpha/2} \sqrt{\text{Var}(\text{Log}(\hat{\rho}))}$$

The value of $\text{Log}(\hat{\rho})$ and its standard error are provided by the computer program GLIM (see Appendix 2) and are respectively for the current example $\text{Log}(\hat{\rho}) = -0.2651$ and $\text{Var}(\text{Log}(\hat{\rho})) = 0.00841$. Therefore, if the theoretical value of ρ were equal to one,

$$Z = \frac{-0.2651}{0.09168} = -2.89$$

a value which is too large for a standard normal deviate. We therefore conclude that ρ is significantly lower than 1 and its value is estimated at 0.77. This second way of testing the hypothesis $\rho = 1$ is known as the Wald test which here is the same as checking whether this confidence interval includes one.

The confidence interval of $\text{Log}(\rho)$ calculated as shown above is [–0.448 ; –0.0854] from which we can derive the confidence interval of ρ by exponentiation [0.64 ; 0.92]

which is identical in the present case to that obtained earlier from the Mantel-Haenszel estimate (see page 80).

As a second example, we return to previous data on colon cancer incidence in Côte-d'Or (Table 2.13). We shall describe the incidence data observed among men older than 20 years by a proportional hazards model:

$$\lambda_{ix} = \rho_i \lambda_{1x} \qquad 2 \leq i \leq 6$$

that is, since $E(K_{ix}) = \lambda_{ix} m_{ix}$:

$$\text{Log}\,[E(K_{ix})] = \text{Log}\,(m_{ix}) + \text{Log}\,(\lambda_{1x}) + \text{Log}\,(\rho_i)$$

This is an 18 parameter model (13 parameters for age and 5 for the relative rates); we have 78 observations available to carry out their estimation.

The fit of the model (see Appendix 2) leads to a deviance of 68.20 for 60 degrees of freedom; the goodness of fit is satisfactory ($p = 0.219$) showing that the proportional hazards model is acceptable. The relative rate of the 5 cantons with respect to Dijon (taken as a reference) are respectively 0.70 (côte Viticole), 0.55 (Châtillonnais), 0.79 (Plaine de la Saône), 0.70 (Auxois) and 0.45 (Morvan). However, only the risk for Châtillonnais is significantly less than 1.

The confidence intervals of these parameters, which are obtained as explained above in the context of the comparison of two populations, confirm our previous conclusion. Only the relative rate for Chatillonnais is significantly less than one (see Appendix 2). This result implies logically that the rates of colon cancer are not homogeneous; it is however preferred to test formally this hypothesis by fitting the previous model under the constraint:

$$\rho_i = 1 \qquad 2 \leq i \leq 6$$

We find a deviance of 80.78 for this new model; the increase $80.78 - 68.20 = 12.58$ is significant when compared to the critical value of χ^2 with $65 - 60 = 5$ degrees of freedom ($p = 0.03$). This confirms the heterogeneity of the rates.

The modelling approach is particularly well suited for carrying out the test of homogeneity of the rural regions made previously (see page 90). The hypothesis is then written:

$$\rho_i = \rho_j = \rho^* \qquad \rho^* \neq 1 \qquad 2 \leq i, j \leq 6$$

The fit of this model increases the deviance of 3.46 which is just below its expectation (the χ^2 in this example has $64 - 60 = 4$ degrees of freedom). The estimate of ρ^*, relative rate of rural cantons is obtained from the fit and it is equal to 0.69 (95% CI = [0.54 ; 0.88]).

We therefore conclude that Dijon has the greater risk of colon cancer and that there is no evidence of rate heterogeneity in the rural regions of Côte-d'Or.

The modelling done for the factor region may also have been done for the factor age; it is clear that 13 parameters are not needed for describing the age effect which could be smoothed by a polynomial function (the age effect estimates for younger age groups have in fact a very low precision). The resulting model would

be more parsimonous and would have the same ability for doing the above geographical comparison (see Appendix 2).

The mathematical complexity of this approach is largely compensated for by its interpretative power. The clear terms of the hypotheses, the statistical evaluation of the results, the flexibility of use and the cohesiveness of the approach are qualities that make its systematic introduction into descriptive epidemiology worthy of serious consideration.

Extension and limitations of the present methodology

Risk analyses in the absence of denominators

As we have seen in previous sections, the descriptive analysis of cancer risk requires the estimation of person-years of observation. For descriptive studies involving large areas, national bureaux of statistics are usually able to provide the necessary information. In most countries, however, the data are generally not broken down by variables of epidemiological interest, such as occupation and country of birth. In contrast, these variables are usually available for incident cases or deaths. This section will show how it is possible to take advantage of this information to carry out the analysis of risk despite the lack of corresponding denominators.

The methods which have been proposed are based on an analysis either of the distribution of cases by site (e.g., correspondence analysis) or, where the interest is mainly in cancer of a particular site, of the proportion of this cancer occurring among all other sites. These are known as relative frequency or proportional incidence (or mortality) methods. The discussion will be restricted to the situation where interest is centred on a specific cancer site.

The relative frequency of a specific cancer in a population is defined as the ratio between the number of cases of the cancer and the total number of cancer cases in the population during the same period. The comparison of relative frequencies of a given cancer between two populations is at best an indirect measure of the absolute risk difference. This comparison will be more reliable when the cancer site of interest accounts for a small proportion of all cancer cases. For example, buccal cavity and pharyngeal cancers represent only 2.1% of all cancers in men in the United Kingdom, whereas in France they represent 8.6%. The corresponding crude rates in the two countries are respectively 9.2 and 42.4 per 100 000 person-years. In this situation, the information provided by the absolute and relative indices is identical: this cancer is four times more frequent in France than in the United Kingdom.

As a rule, however, risk estimates obtained from studies of relative frequency are less precise. The methods proposed below provide only a partial remedy for

their intrinsic weakness. We will discuss briefly methods of standardization of relative frequencies and the modelling of proportional incidence in the following sections.

Standardized indices of relative frequency

The relationship between cancer incidence or mortality and age is generally site-specific. Consequently, it will generally not be the same for the site of interest and for all cancers. For example, the proportion of buccal cavity and pharyngeal cancers in France is 13.9% between 45 and 64 years and only 5.5% after 65 years [23]. The ratio of the age-specific incidence rate of the cancer under consideration and all cancers combined (λ_x / μ_x) will therefore depend on age; standardization is necessary to account for confounding by age when comparisons are carried out.

Two standardized indices have been proposed: ASCAR [24], which was initially developed for studies in developing countries, and the proportional incidence ratio (PIR). These indices are the equivalents, for relative frequencies, of the direct and indirect methods of standardization discussed previously.

ASCAR is the average of the age-specific relative frequencies, weighted by a standard distribution of age at which cancer occurs. If k_x is the number of cases of age x for the cancer of interest, K_x the total number of cancer cases and w_x the proportion of cancer of age x in the standard population ($\sum_x w_x = 1$), then

$$\text{ASCAR} = \sum_x w_x \frac{k_x}{K_x}$$

The PIR is the ratio between the total observed number of cancer cases at a given site and the number expected if the cases occurred according to a standard relative frequency p_x which was a function of age:

$$\text{PIR} = \frac{\sum_x k_x}{\sum_x K_x p_x}$$

The total number of cancer cases K_x in age group x being fixed at its observed value, the number k_x of cancer cases at a given site is distributed as a binomial variable. It is possible to make statistical inferences based on ASCAR and PIR using this distribution. This approach is however of limited interest since neither ASCAR nor PIR estimates population parameters which are interpretable in terms of risk or relative risk. The following approach overcomes this difficulty to some extent.

Modelling incidence data in the absence of the denominator

Suppose that we are studying the risk of a specific cancer C in two populations P_0 and P_1 in which cancer incidence rates are respectively λ_0 and λ_1 for cancer C

and μ_0 and μ_1 for all cancers (Table 2.16). Let $\nu_1 = \mu_1 - \lambda_1$ and $\nu_0 = \mu_0 - \lambda_0$ be the incidence rate for all cancers other than C (denoted A), and ρ and θ be respectively the relative rates of cancers C and A, that is, $\lambda_1 = \rho\lambda_0$ and $\nu_1 = \theta\nu_0$. If a cancer occurs in population P_1, the probability that it is the specific cancer C is:

$$p_1 = \frac{\lambda_1}{\mu_1} = \frac{\rho\lambda_0}{\rho\,\lambda_0 + \theta\nu_0} \tag{2.37}$$

and therefore

$$\frac{p_1}{1-p_1} = \frac{\rho}{\theta} \times \frac{\lambda_0}{\nu_0} = \frac{\rho}{\theta}\,\frac{p_0}{1-p_0} \tag{2.38}$$

The odds of cancer C occurring in population P_1 are ρ/θ times the odds of its occurring in population P_0. This odds ratio is equal to the relative risk only if $\theta = 1$, that is, if the incidence rate of other cancers A is the same in the two populations. The observed odds ratio $k_1\ell_0 / k_0\ell_1$, which is an estimate of ρ/θ, is therefore somewhat difficult to interpret. When cancer C is rare and other cancers have approximately the same incidence in the populations being compared, the method is perfectly adequate.

When a confounding variable is considered, tables similar to Table 2.16 are constructed for each category of this variable and the Mantel-Haenszel method is used to provide an estimate of ρ/θ [25], for example if the number of cases are distributed by age group (x):

$$\widehat{\left(\frac{\rho}{\theta}\right)} = \sum_x \frac{k_{1x}\ell_{0x}}{K_x} \div \sum_x \frac{k_{0x}\ell_{1x}}{K_x} \tag{2.39}$$

In practice, the logistic model is preferable, since formula (2.38) is equivalent to

$$\text{Logit}(p_1) = \text{Logit}(p_0) + \text{Log}\left(\frac{\rho}{\theta}\right)$$

More generally, if we adapt the model for confounding variables and study the risk in more than two groups, the probability of cancer C occurring in group j at age x is:

$$p_{jx} = \frac{\rho_j\,\lambda_{0x}}{\rho_j\,\lambda_{0x} + \theta_j\,\nu_{0x}}$$

Table 2.16 Distribution of cancer cases in age group x

	Number of cases		
	Population P_1	Population P_0	Total
Cancer under study (C)	k_{1x}	k_{0x}	$k_{.x}$
Other cancers (A)	ℓ_{1x}	ℓ_{0x}	$\ell_{.x}$
Total	K_{1x}	K_{0x}	$K_{.x}$

which leads to the logistic model

$$\text{Logit}(p_{jx}) = \alpha_x + \beta_j$$

where

$$\alpha_x = \text{Logit}(p_{0x}) \qquad \beta_j = \text{Log}\left(\frac{\rho_j}{\theta_j}\right)$$

The parameters of the logistic model may be estimated from data k_{jx} for cancer C and ℓ_{jx} for other cancers A over exposure categories j.

This methodology is exactly that of a case-control study in which cases are patients with cancer C and controls are all other cancer patients. Given the similarity, this approach will not be developed further. A detailed discussion can be found in Chapter 6 of Breslow and Day [25]. An example of the use of this method is found in Chapter 3, page 168 where it is applied to a study of migrants. The proportional mortality method has also been extensively used in the estimation of occupational risk [26].

Choosing between various risk measures

Describing a complex situation by a single value is inevitably a difficult exercise and the interpretation of such a numerical summary should be made with great care. Standardization is a step towards a better understanding of the phenomena under study, but it is certainly not the universal method used to solve problems of comparison of incidence. Epidemiologists should be aware of the limitations of this method and should not ignore the fact that, in extreme situations, these statistics can behave pathologically.

We have introduced three principal index classes in this Chapter: i) indices of risk that are based on probability, such as cumulative risk; ii) average rates based on standard populations that give more or less importance to different subgroups of the population under study, such as direct standardized rates; and iii) relative measures of incidence, such as the standardized incidence ratio (SIR), whose objective is to measure the risk of disease relative to a standard incidence that can be interpreted in other respects. In this section, we examine the respective advantages and disadvantages of these indices, and, in particular, the interpretability, the absence of bias and the precision of the indices, three essential requirements of statistics intended to summarize disease incidence in a population.

Cumulative risk places the population under consideration on an immediately interpretable scale of risk. Moreover, it has the advantage of being consistent, since truncated risk is less than total risk. However, a truncated standardized rate obviously does not have this property; its value is inevitably arbitrary since it provides only a rough estimate of the annual number of cases that might be observed in a fictitious population. So, in Côte-d'Or, an individual has 38 chances out of 1000 of developing stomach cancer before 85 years of age, if he does not die before this

age and he has 6.8 chances out of 1000 of developing it between the ages of 35 and 65 years. Among 100 000 persons in the same population and given the present level of risk, there would be 14.0 stomach cancers per year if the age structure was that of the world population, and 18.9 stomach cancers if the population comprised only individuals aged from 35 to 65 with the same age structure as the world population. Cumulative risk can be interpreted in a practical way by anyone who has an understanding of the concept of risk. Conversely, standardized rates appear as more abstract indices whose interpretation demands some epidemiological training and a familiarity with their orders of magnitude.

Furthermore, the situation is considerably complicated by the existence of a multitude of standards. For example, using the European standard, the same comparative rates discussed in the previous paragraph become 23.5 and 19.8, illustrating how important the choice of a standard population is in the interpretation of the number of cases observed. We should remember that a standardized rate is an average of values that varies with age in a ratio of 1:1000 for most cancers under study and it is not surprising that the weights used play a large role in the determination of the rate. In the situation where the differences of specific rates being compared do not all have the same sign, it can be shown that any desired result can be obtained by manipulating the standard population. Remember too that all the indices are summaries of the incidence curve at a given point in time and synthesize estimates of rates from various cohorts, which might have been exposed to different risk factors or to different levels of the same risk factor. One should be extremely cautious when using the indices to analyse temporal trends in cancer risk, or to examine the covariation with the level of a factor (see Chapter 1, page 8, and Chapter 3).

All these direct measures of incidence are also sensitive to random variation, and the combination of a substantial weight w_x and a very imprecise specific rate can cause surprising results (see Table 2.16 below). This is a problem to which routinely produced indices are particularly sensitive because they are not necessarily subjected to close examination before publication.

Relative measures of incidence are generally used when we want to compare subgroups of a population with its overall incidence that is considered to be free of random fluctuations. The standardized incidence ratio (SIR) is by its construction such a measure, and the comparative incidence figure (CIF) can also be used for this purpose. If the ratio of incidence rates does not depend on age, these relative measures are estimates of this ratio, and the SIR is constructed for this particular situation. Conversely, when this hypothesis does not hold, the SIR can behave pathologically.

If t_x denotes the incidence rate observed in the age group x and λ_x denotes the standard incidence, the SIR may be written

$$SIR = \sum_{x=1}^{g} u_x \frac{t_x}{\lambda_x}$$

where u_x is a weighting factor proportional to $m_x\lambda_x$, the inverse of the variance of t_x / λ_x. It is therefore a minimum variance estimator of the relative rate. Note here that this estimate can provide an absolute measure of risk if it is multiplied by the crude rate in the standard population.

With the same notation, let h_x and L_x denote the observed number of cases and the number of person-years in the standard population $\left(\lambda_x = \frac{h_x}{L_x}\right)$, and $H = \sum_x h_x$. The CIF may then be written

$$CIF = \frac{\sum_{x=1}^{g} w_x t_x}{\sum_{x=1}^{g} w_x \lambda_x} = \frac{\sum_{x=1}^{g} L_x t_x}{H} = \frac{1}{H} \sum_{x=1}^{g} h_x \frac{t_x}{\lambda_x}$$

If t_x / λ_x was *strictly* constant, the CIF would be equal to it; however, as t_x is subject to random variation, the CIF is a relative rate estimate which can be quite inaccurate, since, when it is expressed as a weighted average of the relative rates t_x / λ_x,

$$CIF = \sum_{x=1}^{g} u_x \frac{t_x}{\lambda_x}$$

the weight u_x are proportional to h_x the number of expected cases in the standard population. Once again we have the problem that has already been mentioned of heavily weighting very imprecise estimates. These difficulties are illustrated in the following example.

Suppose we study a young, healthy population such as that described in Table 2.17:

Table 2.17 Example of data distribution leading to a directly standardized rate of low precision

Age	Study population			Standard population	
	k_x	m_x	$10^3\ t_x$	w_x	$10^3\lambda_x$
15-24	196	98 000	2.00	0.24	3
25-34	2	1 000	2.00	0.20	3
35-44	2	600	3.30	0.19	7
45-54	3	300	10.00	0.19	22
55-64	2	100	20.00	0.18	62
Total	205	100 000	–	1.00	–
Crude rate	–	–	2.05	–	18

the direct standardized rate is then

$\bar{t} = (0.24 \times 2) + (0.20 \times 2) + (0.19 \times 3.3) + (0.19 \times 10) + (0.18 \times 20) = 7.01$ per 1000

consequently,

$$CIF = 100 \times \frac{7.01}{18} = 38.9\%$$

Furthermore, the expected number of cases if the population is subject to the incidence rate λ_x is:

$$E = \sum_{x=1}^{g} m_x \lambda_x = (98 \times 3) + (1 \times 3) + (0.6 \times 7) + (0.3 \times 22) + (0.1 \times 62) = 314$$

therefore, the SIR can be calculated as

$$SIR = 100 \times \frac{205}{314} = 65\%$$

We can see that the last age group (in which the incidence estimate is very imprecise) contributes 3.6 cases to the direct standardized rate, that is, more than all other age groups combined. If no cases were observed in this age-group, the CIF would be 19%; if, on the other hand, four cases were observed, the CIF would be 59%. In fact, both these possibilities are equally and reasonably likely. In contrast, under such hypotheses, the SIR would only vary from 65% to 66%.

However, it would be a mistake to believe that the SIR has only good qualities and the direct rate only faults. In reality, as we have said on a number of occasions, the strengths of the SIR depend on the hypothesis of proportionality of rates. As an illustration, consider the example in Table 2.18, where two populations with grossly different age distributions are compared.

The age-specific incidence is the same in both populations (5 and 20 per 1000) and the direct rates will therefore be the same for both populations, regardless of the standard population used. The standard rates calculated by the indirect method will also be the same if the marginal incidence rate is used as the standard incidence. However, because of the inversion of the distribution of person-years, they

Table 2.18 Example of data distribution leading to meaningless standardized incidence ratios

Age	Population 1		Population 2		Total	
	k_{1x}	m_{1x}	k_{2x}	m_{2x}	$k_{.x}$	$m_{.x}$
1	5	1 000	25	5 000	30	6 000
2	100	5 000	20	1 000	120	6 000
Total	105	6 000	45	6 000	150	12 000

can be very different for standard rates that are not proportional to the common observed rates; for example, when $\lambda_1 = 10$ and $\lambda_2 = 15$,

$$SIR(1) = 100 \times \frac{105}{10 + 75} = 124$$

and

$$SIR(2) = 100 \times \frac{45}{50 + 15} = 69$$

The difference in person-years distribution has led to an excess of expected cases in the first population and a deficit in the second. The direction of the difference will in fact depend on how the chosen standard differs from the common incidence rate. In other words, two standardized incidence ratios cannot be compared if the populations under study do not have incidence rates proportional to those of the standard population. If, however, the hypothesis of proportionality is valid as is often the case in cancer epidemiology, it is perfectly legitimate to compare two SIRs, and an appropriate test even exists for assessing their equality.

To test whether the same exposure leads to the same effect in two populations with different background incidence λ_{1x}, λ_{2x}, It is justifiable to test whether the relative rates of exposed subgroups (the SIRs) are the same in the two populations.

Let K_1 and K_2 be the observed numbers of cases in the exposed subgroups of the two populations; then K_1 follows a Poisson distribution of parameter $\rho_1 E_1$ where $E_1 = \sum_x m_{1x}\lambda_{1x}$ and, similarly, K_2 follows a Poisson distribution of parameter $\rho_2 E_2$ where $E_2 = \sum_x m_{2x}\lambda_{2x}$. Consequently, the test of equality of the SIRs ρ_1 and ρ_2 is standard and is based on similar arguments to those developed on page 81 of this chapter: the total number of observed cases $K_1 + K_2$ being fixed, K_1 has a binomial distribution with parameter $K_1 + K_2$ and $\frac{E_1}{E_1 + \theta E_2}$ where $\theta = \rho_2/\rho_1$. The hypothesis of equality of the SIRs can then be tested as the hypothesis $\theta = 1$ which is itself equivalent to a test of the parameter of the binomial distribution.

Extreme examples should not make us doubt the efficiency of standardization methods. In fact, in 80% of situations that we encounter, the SIR and the CIF are very close [22]. Nevertheless, we should remember that these indices are only summaries of a more complex situation and that they have their limitations. Sometimes it is advisable to analyse incidence data by age and if necessary by cohort in order to obtain appropriate results, and in this situation the more specific procedures introduced on page 82 and in Chapter 3 should be used.

A thorough understanding of the concepts that we have discussed should help to avoid the main pitfalls encountered in the statistical analysis of descriptive epidemiological data. It is essential that methods are kept in their proper perspective when they are used: no statistical recipe book can ever replace a good intuitive understanding obtained from practical experience.

Bibliographical notes

As we have already noted, epidemiology, and specifically, descriptive epidemiology, has borrowed a great deal from demography. Direct and indirect standardized rates, the key tools of the epidemiologist, were devised by demographers. Readers interested in referring to the source of these techniques can consult two classical works on demography which remain current in their field: those of Pressat, in French, and Benjamin, in English [28].

Breslow and Day's monograph (Volume 1) on the analysis of case-control studies provides a fundamental description, at both a theoretical and practical level, of the calculation of risk and its interpretation [25]. Volume 2 by the same authors deals with cohort studies which, as we have noted in Chapter 1, show the basic concepts and techniques of descriptive epidemiology [29].

Two articles by these authors usefully complete this bibliographical summary. The first [16] is a discussion of the statistical tests presented in this chapter, particularly, the Mantel-Haenszel and related tests. The second [30] discusses the properties of the standardized incidence ratio and its advantages and disadvantages compared to the CIF, the principles of the heterogeneity test for comparing incidence in several populations, and the use of log-linear models for this type of analysis. Once again, although the methods are presented in the context of cohort studies, they are directly applicable to descriptive studies.

In his book on rates and proportions, Fleiss [31] devotes about twenty pages to standardization, with a special focus on the case where there are several variables for which adjustment is required. In fact, most epidemiological texts consider the calculation of direct and indirect standardized rates [32]. Some discuss the problem of variability of standardized rates, but few clearly explain the conditions necessary for the application of these methods. The recent publication from the International Agency for Research on Cancer on the techniques of cancer registration devotes a chapter to basic statistical methods in this area, and discusses routine techniques for comparison when denominators are unavailable (ASCAR and PIR) [33]. An older WHO manual on mortality analysis is out-dated with respect to comparative methods, but provides a useful description of the calculation of demographic indices and an empirical approach to the analysis of all-cause mortality, when such data are available [34].

McCullagh and Nelder's monograph provides a deeper analysis of the theory of log-linear models [35] while Aitkin and coworkers' introductory work is more oriented towards practical application [36]. Finally, Healy provides an introduction to the software GLIM [37], in more detail than the brief description in Appendix 2 of this book.

REFERENCES

[1] Robillard JM. Estimation post-censitaire d'une population par projection : application dans le cas d'un registre des cancers du Calvados, 1978. *Rev Epidémiol Santé Publ* 1983, **31** : 337-340

[2] Armitage P, Doll R. Stochastic models for carcinogenesis. *In* : Proceedings of the 4th Berkeley Symposium on mathematical statistics and probability : biology and problems of health. Berkeley, University of California Press, 1961, pp. 19-38

[3] Peto R. Epidemiology, multistage models and short-term mutagenicity tests. In : HH Hiatt, JD Watson, JA Winsten (eds): *Origins of human cancer.* Cold Spring Harbor, NY, Cold Spring Harbor Laboratory, 1977, pp. 1403-1428

[4] Cook PJ, Doll R, Fellingham SA. A mathematical model for the age distribution of cancer in man. *Int J Cancer* 1969, **4** : 93-112

[5] Clemmesen J. *Statistical studies in the aetiology of malignant neoplasms : I. Review and results.* Kobenhavn, Munksgaard, 1965, pp. 249-340

[6] Waterhouse JAH, Muir C, Correa P, Powell J. (eds). *Cancer incidence in five continents,* Vol. III (IARC Scientific Publications, No. 15), Lyon, IARC, 1976

[7] Waterhouse J, Muir C, Shanmugaratnam K, Powell J (eds). *Cancer incidence in five continents,* Vol. IV. (IARC Scientific Publications, No. 42), Lyon, IARC, 1982

[8] Rao CR. *Linear statistical inference and its applications.* New York, John Wiley, 1965, pp. 426-427

[9] Zdeb MS. The probability of developing cancer. *Am J Epidemiol* 1977, **106** : 6-16

[10] Muir C, Waterhouse J, Mack T, Powell J, Whelan S (eds). *Cancer incidence in five continents,* Vol. V. (IARC Scientific Publications, No. 88), Lyon, IARC, 1987

[11] Romeder M, McWhinnie JR. Potential years of life lost between ages 1 and 70 : an indicator of premature mortality for health planning. *Int J Epidemiol* 1977, **6** : 143-151

[12] Cochran WG. Some methods for strengthening the common χ^2 tests. *Biometrics* 1954, **10** : 417-449

[13] Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst* 1959, **22** : 719-748

[14] Hauck WW. The large sample variance of the Mantel-Haenszel estimator of a common odds ratio. *Biometrics* 1979, **35** : 817-819

[15] Breslow NE, Liang KY. The variance of the Mantel-Haenszel estimator. *Biometrics* 1982, **38** : 943-952

[16] Breslow NE. Elementary methods of cohort analysis. *Int J Epidemiol* 1984, **13** : 112-115

[17] Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955, **11** : 375-386

[18] Alexander FE, McKinney PA, Williams J, Ricketts TJ, Cartwright RA. Epidemiological evidence for the « Two-disease hypothesis » in Hodgkin's disease. *Int J Epidemiol* 1991, **20** : 354-361

[19] Mantel N, Stark CR. Computation of indirect adjusted rates in the presence of confounding. *Biometrics* 1968, **24** :997-1005

[20] Breslow NE, Day NE. Indirect standardization and multiplicative models for rates, with reference to the age adjustement of cancer incidence and relative frequency data. *J Chron Dis* 1975, **28** :289-303

[21] Baker RJ, Nelder JA. *The GLIM system : release 3.77.* Oxford, Numerical Algorithms Group, 1985

[22] Occupational mortality 1970-72 : England and Wales. Series DS n° 1. London, Her Majesty's Stationery Office, 1978, 244 p.

[23] JENSEN OM, ESTÈVE J, MOLLER H, RENARD H. Cancer in the European Community and its Member States. *Eur J Cancer* 1990, **26** : 1167-1256

[24] TUYNS AJ. Studies on cancer relative frequencies (ratio studies) : a method for computing an age-standardized cancer ratio. *Int J Cancer* 1968, **3** : 397-403

[25] BRESLOW NE, DAY NE. *Statistical methods in cancer research : (Vol. 1) : The analysis of case-control studies.* (IARC Scientific Publications, No. 32), Lyon, IARC, 1980

[26] SHANMUGARATNAM K, LEE HP, DAY NE. *Cancer incidence in Singapore, 1968-1977.* (IARC Scientific Publications, No. 47), Lyon, IARC, 1983

[27] PRESSAT R. *L'analyse démographique, concepts, méthodes, résultats (2nd edition).* Paris, Presses Universitaires de France, 1969

[28] BENJAMIN B. *Demographic analysis.* London, George Allen and Unwin, 1968

[29] BRESLOW NE, DAY NE. *Statistical methods in cancer research. Vol. 2. The design and analysis of cohort studies* (IARC Scientific Publications No. 82), Lyon, IARC, 1987

[30] BRESLOW NE, DAY NE. The standardized mortality ratio. *In* : PK Sen (ed) : *Biostatistics in biomedical, public health and environmental sciences.* Amsterdam, Elsevier North Holland, 1985

[31] FLEISS JL. *Statistical methods for rates and proportions.* New York, Wiley, 1981 (2nd ed)

[32] RUMEAU-ROUQUETTE C, BRÉART G, PADIEU R. *Méthodes en épidémiologie.* Paris, Flammarion, 1985

[33] JENSEN OM, PARKIN DM, MACLENNAN R, MUIR CS, SKEET RG. *Cancer registration : principles and methods.* (IARC Scientific Publications No. 95), Lyon, IARC, 1991

[34] World Health Organisation. *Manual of mortality analysis.* Geneva, WHO Div. of Health Statistics, Dissemination of statistical information, 1977

[35] MCCULLAGH P, NELDER JA. *Generalized linear models.* Chapman and Hall, London 1983

[36] AITKIN M, ANDERSON D, FRANCIS B, HINDE J. *Statistical modelling in GLIM.* Oxford Science Publications (OUP), Oxford, 1989

[37] HEALY MJR. *GLIM : An introduction.* Oxford Science Publications (OUP), Oxford, 1989