1 Definitions

- State[1] vs. Event[2] [the *transition* (rapid) from one state to another] [3]
- Population An aggregate of people, defined by a membership-defining...
  - event → "cohort" (closed population i.e., closed for exit)
    or
  - state – one is a member just for duration of state → Open population (open for exit) / dynamic / turnover

[1]Google: The way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state". State of matter: (chemistry) the three traditional states of matter are solids (...) liquids (...) and gases (...).

[2]Most of the definitions below are adapted from the glossary in the textbook *Theoretical Epidemiology: Principles of Occurrence Research in Medicine* by O.S. Miettinen (Wiley 1985).
Google: something that happens at a given place and time | a phenomenon located at a single point in space-time; the fundamental observational entity in relativity theory | In the Unified Modeling Language, an event is a notable occurrence at a particular point in time. Events can, but do not necessarily, cause *state transitions* from one state to another ... | An event in computer software is an action which can be initiated either by the user, a device such as a timer or Keyboard (computing), or even by the operating system. | ~~In probability theory, an event is a set of outcomes and a subset of the sample space where a probability is assigned. Typically, when the sample space is finite, any subset of the sample space is an event (i.e. all elements of the power set of the sample space are defined as events)~~. | An occurrence. | A runtime condition or change of state within a system. | A thing which happens, like a button is pressed. Events can by low-level (such as button or keyboard events), or they can be high level (such as when a new dataset is available for processing). | A means by which the server notifies clients of *changes of state*. An event may be a side effect of a client request, or it may have a completely asynchronous cause, such as the user's pressing a key or moving the pointer. In addition, a client may send an event, via the server, to another client.

[3]In epidemiology, some authors reserve the word "*occur*" for an event (Google: happen; take place; come to pass; "Nothing occurred that seemed important") But, both in epidemiology and in lay use, it is and can also be used for a *state* (to be found to exist; "sexism occurs in many workplaces"; "precious stones occur in a large area in Brazil"). Miettinen [European J of Epi. (2005) 20: 11-15] makes this point in his reply to one of the several authors who commented on his article *Epidemiology: Quo vadis?* ibid, 2004; 19: 713718.

> Walker's commentary was devoted to teaching me that the concept of occurrence has to do with outcome events only; that it thus does not encompass outcome *states*; and that etiologic occurrence research therefore does not encompass the important study of causal *prevalence* functions. As I now consult The New Oxford Dictionary of English (1998 edn), I find as the meanings of occurrence (as a mass noun) these: 'the fact or frequency of something happening' and 'the fact of something *existing* or being found . . .,' as in 'the occurrence of natural gas fields.' And in my Perspective article I find 'state' or 'prevalence' occurring as many as eight times, 'event' or 'incidence' no more than nine times. The verb 'occur,' I might need to add, means 'happen; take place; *exist or to be found to be present* . . . ,' as in 'radon *occurs* naturally in rocks . . . ' [italics added by JH]

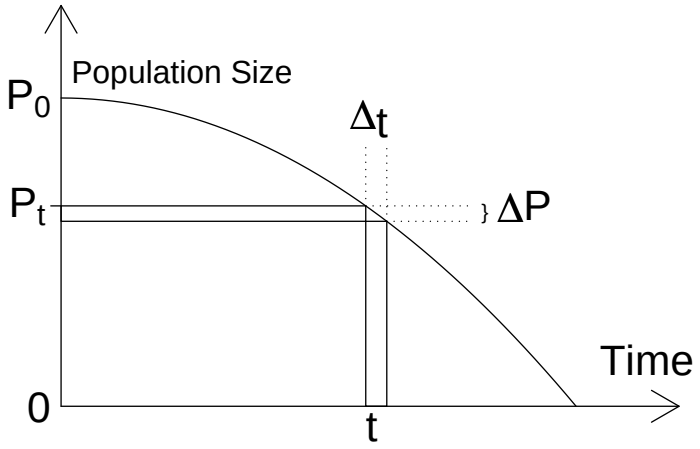
- Prevalence (of a state) : The existence (as opposed to the inception or termination) of a particular state among the members of the population.
- Prevalence Rate: the proportion of a population that is in a particular state.
- Population-time: The amount of population experience in terms of the integral of population size over the period of observation.
- Incidence: The appearance of events of a particular kind in a population (of candidates over time)
  - *Incidence density (ID)*: The ratio of the number of events to the corresponding population time (candidate time). If we subdivide time into very short spans, *ID* becomes a function of time, $ID(t)$; otherwise *ID* refers to the average over the entire span of time.
  - *Hazard*: limiting case of *ID* as we narrow the span of time. More commonly used w.r.t. *closed* population, with a natural "t_0."
  - *Force of morbidity/mortality (Demography).*
- Case: Medicine – episode of illness, ("a case of gonorrhea"). Epidemiology – a person representing a case (in medical sense) of some state or event.[4]
- Incident cases: Cases that appear (as against those that exist or prevail).
- Cumulative Incidence (CI): The *proportion* of a cohort (of candidates) experiencing the event at issue over a particular risk period if time-specific incidence density is considered to operate over that period.
  - The relation between ID and CI can be expressed mathematically as
$$CI_T = CI_{0\to T} = 1 - \exp\left\{ -\int_0^T ID(t)dt \right\}.$$
  - As a function of t, the *complement*, $1 - CI_{0\to t}$ is called the "*Survival*" function, $S(t)$, since it is the proportion of the cohort that, at time t, remains (continues, "survives") in the initial state.
- Risk: The *probability* that an event (untoward) will occur.
- Case Fatality Rate: (Rothman 1986, p31) The cumulative incidence of death among those who develop an [acute] illness [e.g., SARS, influenza]. The time period for measuring the case fatality rate is often unstated.

[4] *Google*: an occurrence or instance of something; "a case of bad judgment"'; "another *instance* occurred yesterday"; *Merriam-Webster*: noun, Middle English cas, from Anglo-French, from Latin *casus* fall, chance, from *cadere* to fall. 1 a: a set of circumstances or conditions b (1): a situation requiring investigation or action 6 a: an *instance* of disease or injury <a *case* of pneumonia> .

2 Link between ID and CI

2.1 Version adapted from Rothman 1986, pp 29-31.

Despite the interpretation that can be given to incidence rate, it is occasionally more convenient to use a more readily interpretable measure of disease occurrence. Such a measure is the *cumulative incidence*, which may be defined as the proportion of a fixed population that becomes diseased in a stated period of time. If risk is defined as the probability of an individual developing disease in a specified time interval, then cumulative incidence is a measure of average risk. Like any proportion, the value of cumulative incidence ranges from zero to 1 and is dimensionless. It is uninterpretable, however, without specification of the time period to which it applies. A cumulative incidence of death of 3 percent may be low if it refers to a 40-year period, whereas it-would be high if it applies to a 40-day period. It is possible to derive estimates of cumulative incidence from incidence rate. Consider a fixed population. The figure shows the size of this fixed population, by time, indicating a small decrement at time t



At time t, $CI_t = (P_0 - P_t)/P_0$; in words, the cumulative incidence at time t equals the number of people who have exited the fixed population by time t because of disease $(P_0 - P_t)$, divided by the initial number of people in the population. The incidence rate at time t is the ratio of new cases to the person-time observation experience; thus

$$I_t = \frac{\Delta P}{P_t \Delta t}, \quad \text{i.e., in calculus notation,} \quad I_t = \frac{-dP}{P_t dt} \;\; ; \;\; -I_t dt = \frac{dP}{P_t}.$$

(The minus sign is used because the change in P is negative in relation to t; without the minus sign, the incidence measure would be negative.) Integrating both sides, $-\int_0^T I_t dt = \ln(P_T) - \ln(P_0)$. Taking antilogs, $\exp\{-\int_0^T I_t dt = P_T/P_0$. And, since $CI_T = (P_0 - P_T)/P_0$, we have

$$CI_{0\to T} = 1 - \exp\left\{ -\int_0^T I_t dt \right\}.$$

This is estimated as

$$CI_{0\to T} = 1 - \exp\{-\sum I_i \Delta t_i\},$$

where the summation of the index, i, is over categories of time covering the interval $[0, T.]$

For a constant incidence rate,

$$CI_{0\to T} = 1 - \exp\{-I \times T\}.$$

Because $\exp\{x\} \approx 1 + x$ for $|x| <$ about 0.1, a good approximation for a small cumulative incidence (less than 0.1) is

$$CI_{0\to T} \approx \sum I_i \Delta t_i.$$

or

$$CI_{0\to T} \approx I \times T.$$

if the rate is constant with time. Thus, to estimate small risks, one can simply multiply the incidence rate by the time period. The above approximation offers another interpretation for the incidence rate; it can be viewed as the ratio of a short-term risk to the time period for the risk as the duration of the time period approaches zero.

The cumulative incidence measure is premised on the assumption that there are no competing risks of death. Thus, if an individual at age 40 faces a cumulative incidence, or risk, of 35 percent in 30 years for cardiovascular death, this is interpreted as the probability of dying from cardiovascular disease given that the individual is free from other risks of death. Because no one is actually free from competing risks, the cumulative incidence measure for any outcome other than death from all causes is a hypothetical measure. In principle, cumulative incidence for lengthy periods is unobservable and must be inferred because of the influence of competing risks.

2.2 Link between ID and CI – other derivations

Many epidemiologic textbooks give the mathematical expression that links the cumulative incidence (CI) or "risk" function, or its complement the "survival" function, with the integral of the incidence density (ID) function. Of the 15 modern texts JH has examined, only Rothman 1986 derives the relationship. Unfortunately, the formal geometric and calculus-based derivation used does not provide any insight into 'why' or 'how' the exp function comes into it, so epidemiologists are forced to accept it as a mere mathematical 'fact.

In his simpler 2002 book, in pp 33-38, he uses heuristic arguments, but does not show the formula itself. Below I derive the formula heuristically. By working through a simple example, I try to make clear the difference between rate and risk, and the units involved, and when one is numerically close to the other.

a. Simplest case

I begin with an exercise which, unless explicitly given in the context of this formula, tends to perplex many first year epidemiology trainees. I base it on data from Ayas et al (2006). In a large study, the observed rate of reported percutaneous injuries (PIs) among residents/interns in obstetrics/gynecology (ob/gyn) programs was 94 injuries in 964 intern-months, or (to the first 2 significant digits) 0.10 injuries per intern-month. I ask students to assume uniform 250-work-hours each month, with injury rates of 0.1 per intern-month that are constant, both within and across the hours and months in question. I then ask them to "calculate the probability that an average-risk ob/gyn resident would suffer at least one PI by the end of 1, 6 and 12 months of experience." I do not explicitly describe each of the probabilities as a 'cumulative incidence' or 'risk, but I do tell them that if they prefer, they may calculate the (complementary) probability of '*surviving*' these lengths of work-time *without* a PI.

Many students readily volunteer answers of $0.1 \times 1 = 0.1 = 10\%$ and $0.1 \times 6 = 0.6 = 60\%$ for the 1- and 6-month risks, before realizing when they try to calculate the 12-month risk that it cannot be $0.1 \times 12 = 1.2 = 120\%$. And while they are unable to now give an exact 12-month risk, many are confident that the 1-month risk is indeed 0.1 or 10%.

They have all been taught very early on how 'person-time' rates are calculated, and that a rate (or ID), which has dimension events/person-time, is entirely different conceptually from a risk, which is a (dimensionless) proportion. It is interesting to try to understand why there is such difficulty going back and forth between the two, in appreciating whether the one-month risk is less than or more than 10%, and in estimating how much less than 120% the 12-month risk is!

b. More generally

[Note: This section was written a few years ago, before JH came across Edmond's definition of force of mortality – in the Appendix to the article on the Bridge of Life, Turner & Hanley re-use this analogy of computers acting as servers.]

One heuristic way to begin might be to imagine a physical or human system consisting of say 100 workstations, each one in *continuous operation*. The Figure overleaf shows the 12-month log for a system in which the physical devices (humans) failed (were injured), independently of each other and of the duration they had been operating, and where, if such events occurred, they were immediately replaced. The expected failure rate (incidence or incidence density) is the expected number of events (120) per 1200 device-months or person-months, 0.1 per device-month or person-month, or 1.2 per device-year or person-year of operation.

As we will show below, one would expect approximately 70 of the 100 *initial* devices or operators to fail before the end of the year, so that the one-year risk is in fact considerably less than 100%. The 120 failures or injuries in that first year of the system occur in an average of 70 of the 100 *first generation* members, and in 34 of their 70 *replacements*, and in 12 of *their* 34 *replacements*, and so on. In all, it takes an average of 220 different (100 *initial*, plus 120 *replacement*) devices or humans to keep the 100 workstations in continuous operation for 1 year.

Some of the reasons for the disconnect is our propensity to think in terms of *individual devices* rather than the continuous device-time or person-time needed to maintain the service. In effect, the device-moments or person-moments are entirely *interchangeable*. We tend to draw person time as separete parallel lines, as if a station belonged to a device or person, but the 'up-time' can be generated by having some replacement devices or persons use the same stations as others.

1-month Risk (CI)

If one understands the Poisson distribution, and how exactly it is derived, it is easy to move from a failure rate (ID) to a 1-month or x-month *risk*: the number of device failures in a period of 1 device-month of operation (up-time) is a Poisson random variable, with possible values 0, 1, 2, .. , and the expected (mean) number of failures is $\mu = 0.1$. Thus the probability of no (zero) PI injuries or failures is $P[0] = \exp(-0.1) = 0.90484$, so the 1-month risk or cumulative incidence is $1 - 0.90484 = 0.09516$ or 9.516%; the x-month risk is obtained similarly, using $\mu = 0.1 \times 6 = 0.6$, to arrive at a risk of $1 - P[0] = \exp(-0.6) = 1 - 0.54881 = 0.45118 = 45.115\%$.

However, just as with the relationship between incidence density (failure rates) and risk, the Poisson distribution is seldom well explained in introductory or epidemiology biostatistics texts, and so many would not be further enlightened by this 'explanation.'

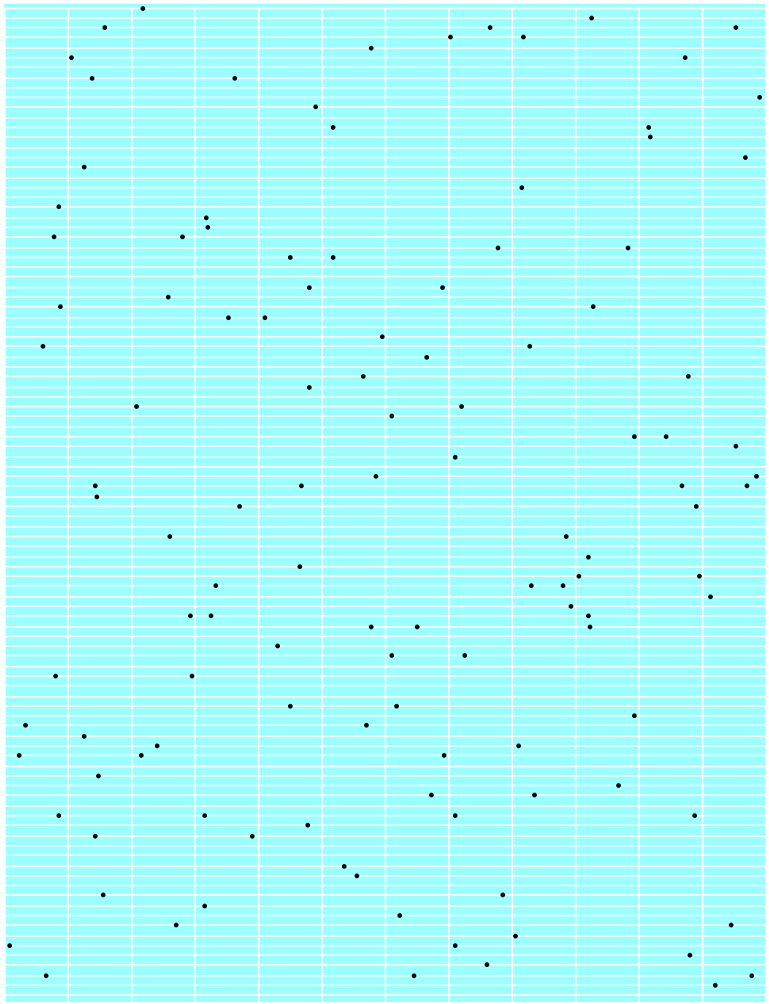


Figure Legend: 12-month log for a computer system (workplace) consisting of 100 workstations, represented by 100 horizontal white lines. The dots – if applicable – for a station represent the times at which the devices at that station failed (workers at that station were injured). Failed devices (injured workers) were immediately replaced, so that each station remained in continuous operation. Devices failed (workers were injured) independently of each other and of the duration they had been operating. On average, some 120 failures (injuries) occurred in 1200 operator-months of operation. Thus, the failure(injury) rate was 0.1/operator-month, or 1.2/operator-year.

The key to understanding how the exp function is involved in the transition from PI *rate* to PI *risk* is to express the injury rate not as 0.1 per intern-month, but as 0.0004 injuries/intern-hour, or an average of 1 injury per 2500 intern-hours. (we could equally use the rate of failures of the physical devices). The number of events in such a small time unit is again a random variable with possible values $0, 1, 2, \ldots$ but because one intern-hour is so small, the chance of 1 event in that amount of experience is already very small, and the chance of 2 or more is less than 1 in 10 million. Thus, one can very accurately regard the 1-hour risk as $0.1 \times 0.004 = 0.1/250 = 0.0004$, and its complement, the 1-hour 'survival' probability, as $1 - 0.1/250 = 0.9996$. Thus, the 1-month survival probability can be approximated by $(1 - 0.1/250)^{250} = 0.90482 = 90.482\%$ and its complement, the risk or cumulative incidence, by 9.518%.

An *even more accurate approximation* to the survival probability can be obtained by further dividing the 250 hours into 15000 minutes, so that the injury rate is 0.1 per 15000 intern-minutes, and calculating $(1 - (0.1/15000)^{15000} = 9.484\%$ so that the 1-month risk is 9.516%. Subdividing the time units further does not change these decimal places; the function $(1 - 0.1/LargeNumber)^{LargeNumber}$ converges to a constant which is solely a function of the 0.1. The function is the exp function. Indeed, one formal definition of exp x is that it is the limit,

$$\lim_{N \to \infty} (1 + x/N)^N.$$

In our example, $x = -0.1$, and the exact survival probability, to 6 decimal places, is $\exp(-0.1) = 0.904837$.

6-month and 12-month Risk (CI)

As in standard survival calculations, the 6-month survival probability $S_{0 \to 6}$ is the product of 6 conditional probabilities:

$$S_{0 \to 6} = S_{0 \to 1} \times S_{1 \to 2} \times S_{2 \to 3} \times S_{3 \to 4} \times S_{4 \to 5} S_5 \to 6.$$

In our example the constant PI rate implies that each $S_{t \to (t+1)}$ equals $\exp[-0.1 \times 1]$ and so the 6-month survival probability is

$$S_{0 \to 6} = \exp[-0.1 \times 1] \times \cdots \times \exp[-0.1 \times 1] = \exp[-0.6] = 0.548 = 54.8\%,$$

so that the 6-month risk is $100 - 54.8 = 45.2\%$.

Had the PI rate *varied over the period at risk*, say as an (equal-) step-function, starting at 0.05 PI/intern-month in month 1 and rising to 0.10 PI/intern-month in month 6, then the 6-month survival probability is again obtained by summing the area under the ID curve to obtain $\int_{t=0}^{6} ID[t]\,dt = 0.45$, and by then calculating

$$S_{0\to 6} = \exp\left[-\int_{t=0}^{6} ID[t]\,dt\right] = \exp[-0.45] = 0.64$$

If, as appears to be the case, the injury rate is closer to 0.16 PI/intern-month when working an extended shift, and 0.08 PI/intern-month when working regular shifts, then the risk for a resident over 3 months of extended shift is $1-\exp[-(0.16\times 3)] = 38\%$. The corresponding PI risk for the 9 months on regular shifts is $1-\exp[-(0.08\times 9)] = 51\%$. The chance of escaping injury-free for the entire 12 months is $\exp[-\{(0.16\times 3)+(0.08\times 9)] = \exp[-1.2]$.

The above calculations further illustrate the '*interchangeability*' of the contributions to the integral involved in the cumulative incidence (CI), and the fact that the CI only depends on the integral itself: the overall 6- or 12-month risk is the same whether the higher- and lower-risk blocks of time are interspersed or contiguous: the overall risk is determined by the integral, also called the cumulative hazard $H[T] = \int_{t=0}^{t=T} h[t]\,dt$.

> You can think of $H[T]$, the integral of $ID(t)$ over the time span in question, as the expected number of events, μ, if there was **always 1** (*not necessarily the same*) **individual at risk for the full time span involved in the integral**. It is easier to think of the continuous time at risk in the context of a work station where whenever the machine/individual fails, it is immediately replaced by another one. Then, the CI is 1 minus the Poisson probability of 0 events in $[0,T]$ when $\mu = H[T]$, i.e.
>
> $$CI = 1-\exp\{-\mu\} = 1-\exp\{-H[T]\} = 1-\exp\left\{ -\int_{t=0}^{t=T} h[t]\right\}.$$

This exponential formula for $S[.]$ is the same as the one for the depreciation/appreciation of a financial fund, where $A_{t=0}$ is the amount at $t=0$, and $\delta(t)$ / $\alpha(t)$ is the rate of depreciation/appreciation, expressed as a smooth function.

Depreciation: $A_{t=T} = A_0 \times exp[-\int_{t=0}^{t=T} \delta[t]\,dt]$.

Appreciation: $A_{t=T} = A_0 \times exp[\int_{t=0}^{t=T} \alpha[t]\,dt]$.

2.3 Approximation to CI

The fact that the risk function is a 1:1 function of the integral of the incidence-density function has implications for when one can obtain acceptably accurate approximations to the risk. The $1-\exp[-H]$ function can be closely approximated by H over the range $H=0$ to $H=0.1$, but this approximation becomes less accurate thereafter. As is shown by the following table

H:	0.05	0.10	0.20	0.30	0.50	1.00	1.50
$(1-\exp[-H])$:	0.049	0.095	0.181	0.259	0.393	0.632	0.777
% over	3	5	10	16	27	58	93

The percentage over-estimation by using $\text{CI}_{approx} = H$, rather than $\text{CI}_{exact} = 1-\exp[-H]$, is close to $50\times H$.

Large values of H can arise from a low rate operating over a longer time-interval, or higher ones over a shorter one.

2.4 T.R. Edmonds' definition of the force of mortality using the concept of a '*person-year*'. (taken from Appendix of Turner-Hanley article on Bridge of Life

We use a modern example involving the 'blue screen of death' for our own depiction of a person-year, and couple this with the Poisson distribution to provide an alternative derivation of the formula linking the hazard (force of mortality) function and the survival function:

Whatever the magnitude or pattern of $h(u)$ over that range, the complement of $S(t)$ is a proportion – epidemiologists call it a 'risk' or 'cumulative incidence.'

In his 1832 book, Edmonds begins his theoretical treatment of the link with the words (emphasis ours)

> The force of mortality at any age is measured by the number of deaths in a given time, *out of a given number constantly living.* The given time has been here assumed to be one year, and the given number living to be one person; consequently, the algebraic sign for the force of mortality represents—the quantity of death in one year for a unit of life at the assumed age; or rather (since the force is changing continually) represents—the quantity of death on a unit of life which would occur by the action of this force continued uniform for the space of one year.

In his example, the 'given number constantly living' is a number of *person-years*

(1 p-y in his calculations). Heuristically, imagine a setup in which one of a large number of personal computers, all of the same model and age, acted as a server for the others. In order to maintain virtually continuous server operation, a server that fails is immediately replaced by another available computer. Let $h(u)$ be the hazard function, which can be influenced only by the age (u) of the computer. If over the first year $h(u)$ is – in the simplest case – assumed to be independent of age, and constant at say $h = 0.1$ failures per computer-month of operation, then of 100 such servers, one would expect approximately 70% of the 100 *initial* machines, 34 of their *replacements*, 12 of *their* replacements, and so on, to fail before the end of the year. In all, it would take an average of 2.2 different computers to keep a server in virtually continuous operation for 1 year, and the expected number of failures is 0.1 events/server-month $\times$ 12 months = 1.20 events in a server-year of operation. The chance that the year of operation is accomplished with a *single* computer is the Poisson probability of 0 failures when $\mu = \int_0^t h(u)du = 1.2$ are expected, namely $\exp[-1.2] = 0.30$. The complement of this is the one year 'risk' or 'cumulative incidence' of 0.70.

Since the computers were assumed to be exchangeable, this example emphasizes that the 1 year of virtually continuous 'server-time' ('up-time') – and observed number of failures – could equally well have been accomplished by having a different computer be the server for different days, or months, or randomly selected periods. As Edmonds would put it, "the failure rate is measured by the number of failures in a given time (here 1 year), out of a given number of servers (here, 1) constantly serving for that time." **Thus for human populations, we can think of the integral of $h(u)$ over the time span in question as the *expected* number of events if there was always 1 (not necessarily the *same*) individual at risk for the full time span involved in the integral.** And even if $h(u)$ *varies* over the time span – as it realistically would – we can still use the integral in the risk / cumulative incidence formula as the expected value of a Poisson random variable. We can do so because of the seldom-emphasized 'closure under addition' property of the Poisson family: the sum of independently distributed Poisson random variables with different expected values is again a Poisson random variable with expected value equal to the sum of these expected values.

3 Relationship between Prevalence, Incidence, and Duration in a State

In a steady state situation,

$$Prevalence = Incidence \times Average\ Duration.$$

Two of the clearest examples of this are admissions to and stays in hospital, and in graduate programs, assuming no change over time in the admission rates, or in the durations in the state they are admitted to.

If the average number of hospital admissions is 55 per day, and the average stay is 10 days, then the average number of beds occupied is 550 beds. Note the units

$$Average\ Prevalence = 550\ beds\ occupied = \frac{55\ beds}{1\ day} \times 10\ days.$$

If the average number of admissions to a program is 10 students per year, and the average stay is 3 years, then

$$Average\ Prevalence = 30\ students = \frac{10\ students}{1\ year} \times 3\ years.$$

Example:

> Although no free-living population is likely to meet the steady state criteria, the qualitative relation embodied in the preceding equation applies widely. A study of HLA types (a class of genetic markers) among children with acute lymphocytic leukemia (ALL) who attended an oncology clinic found that the prevalence of type A2 was higher than that in the general population.[9] The observation raised considerable interest, implying as it did that susceptibility to acute leukemia might be mediated by genetic factors. A follow-up study of a series of newly diagnosed leukemics found identical prevalences of the "high risk" type A2 in patients and in the general population.[10] The discordance between the two findings was due to an effect of HLA type on the mean duration of ALL. Far from being at high risk of ALL, children with HLA type A2 were at no increased risk, responded better to chemotherapy, had longer survivals, and were therefore overrepresented in the (prevalent) clinic population. The lesson is that if you want to study the determinants of incidence rate, you need incident rather than prevalent cases of disease.

9. Rogentine GN et al. HLA antigens and disease: acute lymphocytic leukemia: J Clin Invest. 1972;61:2420-8. 10. Rogentine GN et al. HLA antigens and acute lymphocytic leukemia: the nature of the association. Tissue Antigens 1978;3:470-6.

Alexander Walker, *Observation and Inference*, pp 11-12.

4 Length-biased sampling

See the "Length Bias" entry, written by Mei-Cheng Wang, in the 2005 Encyclopedia of Biostatistics. The Encyclopedia is available online as an eBook through the McGill Libraries.

4.1 Definition

Consider a non-negative[5] random variable Y that – for illustration, but without loss of generality – takes on the *integer* values $y = 0,\ 1, \ldots,$ has a probability distribution with a probability mass function $p_0,\ p_1, \ldots,$ and has expected value $\mu = \sum y \times p_y$ and variance $\sigma^2 = \sum (y-\mu)^2 \times p_y$. Denote the coefficient of variation, σ/μ, by "CV."

We can estimate μ by taking a simple random sample $\{y_1, \ldots, y_n\}$ and calculating $\hat{\mu} = \bar{y} = (1/n)\sum y_i$. Since $Prob[y_i = y] = p_y$, it is easy to show that $E[\bar{y}] = \mu$.

Length-Bias Sampling refers to a form of sampling where the probability of selecting a unit with $Y = y$ is not p_y, but rather $p_y \times y$. Thus, larger Y values will be over-represented in the sample. One of the early instances of this sampling was in the sampling of strands of wool in the textile industry, where *longer* strands has a higher chance of being selected – thus the name *length* bias.

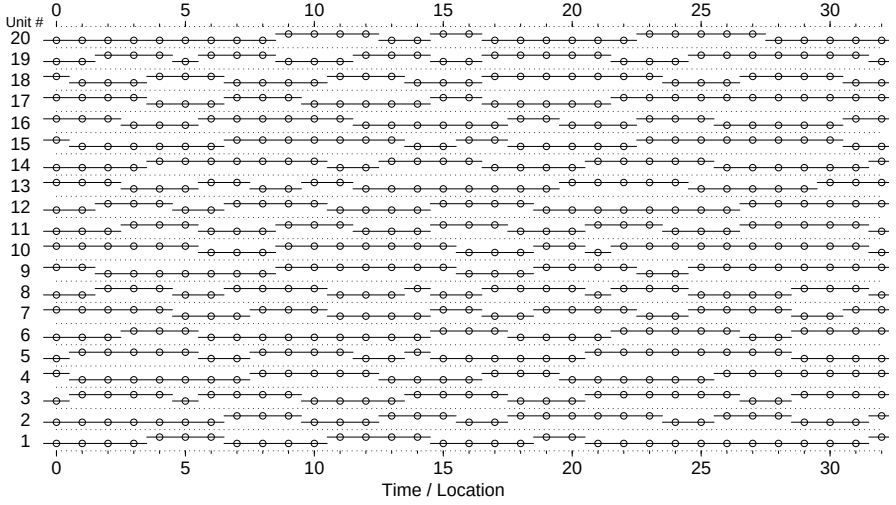
4.2 Examples

Some examples are:

Units	Y	Method of Selecting Units
Words in a text	# letters	Random locations on pages
Homes in a town	# children in elem. sch.	S.R.S. of children in the school
Hospital Admissions	Length of Stay	Pts. in hospital on random day(s)
PhD Students	# Years to Obtain PhD	Cross-sectional/Prevalence survey
Pts. $\bar{c}$ Alzheimer's	Disease Duration	Cross-sectional/Prevalence survey
Cancers	Dur'n of Pre-clinical Phase	Screening with Imaging Test
Outpatients	# visits to facility	S.R.S. of all visits
BRCA1 Carriers	Lifetime risk of Breast Ca.	Case proband studies
Inter-event Intervals	Length	Randomly selected time $\in$ Interval

4.3 Length-bias, illustrated

Imagine the horizontal lines below are the lengths of stay of patients in 20 hospital rooms over a 30 day period, or the length and locations of words in 20 lines of text, or the duration of dementia, from onset until death, for patients with dementia. Imagine selecting a sample of units (admissions, words, dementia patients) by drawing a vertical (cross-sectional) line and selecting those admissions or words or patients that this vertical line intersects. Clearly, the line has a higher chance of selecting longer stays or words or durations. The data were generated from the distribution of word-lengths.[6]



[5] Most texts restrict attention to *positive* random variables. JH included non-negative r.v.'s to highlight examples, such as those in rows 2 and 7 of the table, where even though there is a non-zero probability mass at $Y = 0$, there would no instances of $Y = 0$ in the sample used to estimate μ.

[6] In the Book of Genesis, readily available online.

4.4 Magnitude of the Bias

Let $Y_{l.b.}$ be the random variable representing the value of Y in a unit selected by Length-Biased Sampling. One can easily show that

$$E[Y_{l.b.}] = \frac{\mu^2 + \sigma^2}{\mu} = \mu \times (1 + CV^2).$$

Thus, the greater the relative variation, the greater the over-estimation of μ.

So, if the ***average word length*** is $\mu = 4.5$ letters, and the SD is $\sigma = 2.25$, or 50% of the mean, the expected number of letters in words selected by sticking pins at random in the text is $4.5 \times (1 + 0.5^2)$, i.e., 25% higher.

In the ***number of children per household*** example, say that 50% of households have $Y = 0$ school-age children, that 30% have $Y = 1$, and 20% have $Y = 2$. Thus $\mu = 0 \times 0.5 + 1 \times 0.3 + 2 \times 0.2 = 0.7$ children per household, with $\sigma^2 = (0 - 0.7)^2 \times 0.5 + (1 - 0.7)^2 \times 0.3 + (2 - 0.7)^2 \times 0.2 = 0.61$. In a sample of 70 schoolchildren, one would expect 30 to be singletons and to answer that there is 1 school-age child in their house, and 40 to answer that there are 2 school-age children in their house; thus the mean in this length-biased sample is $(1 \times 30 + 2 \times 40)/70 = 11/7$. This agrees with the $0.7 \times (1 + 0.61/0.49)$ given by the formula above.

Suppose that the ***inter-arrival intervals of buses*** on a certain route during a particular portion of the day are highly variable from day to day, say $\mu = 15$ min and $\sigma = 15$ min. Under these conditions, the average *wait for the next bus* by a person who arrives at the bus-stop each day at a randomly selected time during that portion of the day will be $(1/2) \times 15 \times (1 + 1^2) = 15$ mins. One distribution that has a mean of 15 and a SD of 15 is one where 1/2 the intervals are 30 minutes, and 1/2 are 0, i.e., where two buses show up together every 30 minutes! Another distribution with this same is if one shows up at a random time and waits for radio-active disintegrations whose inter-event intervals have an exponential distribution with mean μ and $\sigma = \mu$. Here again, the average wait until the next event is again $(1/2) \times \mu \times (1 + 1) = \mu$. On the other hand, *if there is no variation*, as in the Metro during certain hours, when *each* inter-arrival interval is say μ minutes, with $\sigma^2 = 0$, one can obtain an unbiased estimate of μ by entering the station at random times, averaging the waiting times until the next Metro train, and doubling the mean of these "forward recurrence times."

In the (N Engl J Med 2001;344:1111-1116.) article "A reevaluation of the ***duration of survival after the onset of dementia***," Christina Wolfson et al. used follow-up data from the Canadian [cross-sectional i.e., Prevalence] Study of Health and Aging to estimate survival from the onset of symptoms of dementia...

> In the 821 subjects, the *unadjusted* median survival was *6.6 years* (95 percent confidence interval, 6.2 to 7.1). *After adjustment for length bias*, the estimated median survival was *3.3 years* (95 percent confidence interval, 2.7 to 4.0).
>
> *Median survival after the onset of dementia is much shorter than has previously been estimated.*

4.5 Correcting for the Length-Bias

One option is to, in the *design* of the study, only select 'new' admissions or onsets, i.e., select an '*inception*' sample or cohort.

But what if the data have already been collected by length-biased sampling? Some options at the *data-analysis stage* include:

- Throw away the data. This was the action taken by a research group at the community health department associated with the MGH in the 1980's, after they realized that their survey of use of outpatient facilities by psychiatric patients over-represented those with greater use. The sampling was from a file-drawer where each patient-visit had generated one card. The investigators simply took a systematic sample of the cards in the drawer. JH only found out about this solution in the 1990's, long after the answers would have been of any value.

- Re-weight the data. This was the action suggested by JH to a research group in the 1990s. They had used a cross-sectional survey to estimate how many days (Y) patients on i.v. medications had to spend in hospital just because there was no outpatient facility that could have managed them – i.e., the patients were well enough to be discharged by day X but had to wait in hospital until day $(X + Y)$ when their i.v. line could be removed. The sample of patients was assembled by having a nurse visit the hospital on randomly selected days, and select those for whom the physician considered that they had already been there for $> X$ days.

  The procedure re-weights the observed frequency, $f_{l.b.}[y]$, of patients with $Y = y$ days in the length-biased sample to the frequency of $Y = y$ one would expect in an (unbiased) inception sample, i.e. $f_{unbiased}[y] = f_{l.b.}[y] \div y$.[7]

[7] This will not work if one of the possible values of Y is 0, as in the number of school-age children per household example. The problem becomes even more acute at the smaller Y values when the data values are recorded on a 'continuous,' rather than discrete scale.

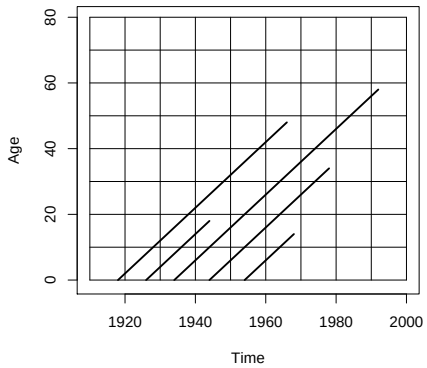
- Use only the data from the inception sample. Unfortunately, this wastes a lot of the data.
- Use parametric models for the distribution of Y.
- (For censored survival-type data, as in the dementia study which got them interested) consult with statisticians in McGill's Department of Mathematics and Statistics.
- ...

5 Lexis Diagram - 2 time axes

The following is excerpted/adapted from the entry "Lexis Diagram" by N. Keiding in the Encyclopedia of Biostatistics.

A **Lexis**[8] **diagram is a (time, age) coordinate system, representing individual lives by line segments of unit slope, joining (time, age) of birth and death** (see Table and Figure).

A Lexis diagram representing the five lives in the Table.



Born	Died	Age at death
1918	1966	48
1926	1944	18
1934	1992	58
1944	1978	34
1954	1968	14

The **Lexis diagram** is an important descriptive tool in epidemiology and demography. However, it also has several applications in survival analysis and analytical epidemiology as a tool for several classes of statistical models, as surveyed by Keiding [ref.]. These uses of the Lexis diagram are less common and it is the aim of this article to indicate some recent developments. [...] Despite its long history, the Lexis diagram is still being rediscovered among statisticians, cf. Goldman, A.I. (1992, Eventcharts: Visualizing survival and other timed-events data, *American Statistician* 46, 13-18) for the standard Lexis diagram.

Applications of the Lexis Diagram in Survival Analysis and Analytical Epidemiology

Clinical Trials with Staggered Entry / Disease Incidence Studies / Prevalent Cohort Studies

Statistical Inference in the Lexis Diagram

Piecewise Constant Intensity Models / Point Processes, Continuous Time

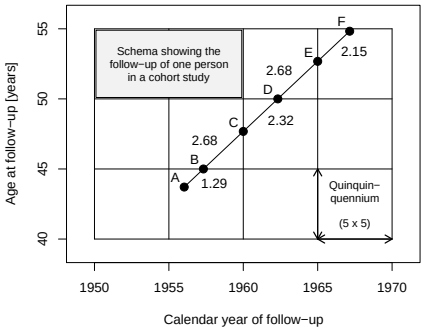
[8] Lexis, W. (1875). *Einleitung in die Theorie der Bevölkerungsstatistik.* Trübner, Strassburg.

The following is excerpted/adapted from Breslow & Day Volume I, section 2.2.

The basic feature of cohort studies that distinguishes them from cross-sectional, case-control or other types of investigation is that, at least in principle, each subject is kept under continuous surveillance for a defined interval of time. If the study endpoint is death, we assume that each subject is 'at risk' of death during the entire interval from his entry into the study until his exit. This means that the study period should contain no interval during which the subject is known to be alive as a condition of cohort membership. If the cohort is defined to consist of all workers with at least five years of employment in a certain factory, therefore, the first five years of their employment history would be excluded from the observation period. A second critical assumption is that any death that actually occurs during the study period will be recorded. For cohorts defined on the basis of past records, this implies that adequate mechanisms exist for tracing individuals from their date of entry into the study until death or until the study's closing date. If no record exists of someone's whereabouts after a certain point in time, he should generally be considered as having left the study at that point. Obvious problems of selection bias exist if such losses are at all frequent, since the causes of and ages at death for 'lost-to-follow-up' subjects may well differ from those for persons who are successfully traced.

The basic method used to estimate age-time-specific mortality rates is to determine each individual the amount of observation time contributed to a given age × calendar period category and to sum up those contributions for all cohort members so as to obtain the total number of person-years of observation in that category. These person-years form the denominators of rates the numerators of which are simply the numbers of deaths due to a given disease, likewise classified by age and calendar year of death.

In some applications, particularly when the observation period is relatively short, the calendar-year axis is ignored and the rates are determined by age interval alone. Computer programs for performing the calculations have been developed by Hill (1972), Monson (1974), Waxweiler et al. (1983), Gilbert and Buchanan (1984) and Coleman et al, (1986), among others, Sometimes the exact dates of birth and of entry and exit from study, which are needed to draw the Figure, will not be available. Then, approximate numbers of standardized person-years may be calculated as shown in the right-hand column of the Table, using the three integer variables, age at entry, year of entry and year of exit. The approximation is based on the notion that a person aged 43 in 1956 will be 44 in 1957, 45 in 1958 and 54 in 1967. He contributes 0.5 years of observation time to the calendar year of entry (1956), 0.5 years to the year of exit (1967), and a full 1.0 year to each intervening year. There would be a single 0.25-year contribution for someone who enters and leaves the study in the same calendar year. The discrepancies between the exact and approximate figures tend to be averaged out when cumulated over individuals, so that the approximate method is sufficiently accurate for most practical purposes.



The process is illustrated in the Figure [**Lexis Diagram**], which shows schematically the course of one worker who was entered on study (point A) at age 43.71 in year 1956.03 and left 11.12 years later (F). He contributed observation time to five separate cells, boundary crossings being made at points B through E. The duration of time spent in each cell is easily determined, as shown in the Table below.

Table. Calculation of exact and approximate age- and year-specific person-years at risk

Point (See Fig.)	Coordinates (Cal. year, age)	***Quinquinquennium*** Year	Age	***Person-years*** Exact	Approx.
A	(1956.03, 43.71)				
		1955-1959	40-44	1.29	1.50
B	(1957.32, 45.00)				
		1955-1959	45-49	2.68	2.00
C	(1960.00, 47.68)				
		1960-1964	45-49	2.32	3.00
D	(1962.32, 50.00)				
		1960-1964	50-54	2.68	2.00
E	(1965.00, 52.68)				
		1966-1969	50-54	2.15	2.50
F	(1967.15, 54.83)				
Total				11.12	11.00

Cause-specific national death rates are typically published by five-year intervals of age and calendar year. Such '***quinquinquennia***' are widely used in cancer epidemiology, and our example of the calculation of age- and calendar period-specific rates illustrates this standard breakdown. Analogous methods may be used if the age/time intervals are longer or shorter than five years.

6 Appendix 1. Rate Measures of Occurrence

Miettinen, O.S. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine.*

A.1.1. PREVALENCE RATE

Prevalence is a phenomenon in populations only, not in individuals. It has to do with occurrence in the sense of the existence of individuals with some particular state (condition or trait). It is quantified in terms of a *prevalence rate*, which is the proportion of individuals (in the population at issue) who are in that state. Examples of prevalence rate thus include the proportions of people who have bloodtype AB, "hypertension," or a congenital malformation, respectively, and the proportion of patients with "hypertension" who have manifested complications of it. As a proportion, prevalence rate is a dimensionless quality, a "pure number."

In the context of a prevalent belief to the contrary, it may be helpful to note that a prevalence rate is not inherently momentary, any more than velocity is in physics. Of course, the prevalences of many conditions or traits are functions of time (age or other), as are incidence and velocity. When prevalence is a function of time, one may still address the average prevalence (expressly) or even ignore this dependence (along with many others, known and unknown).

[...] More to the point, one need, and should, not insist on a latter day mathematician's definition of "rate." The original Latin word takes direct manifestation in today's "unemployment rate," "tax rate," and so on.

A.1.2. INCIDENCE RATES

Incidence is not as singularly a population phenomenon as prevalence, because it refers to changes *within* individuals (prevalence has to do with differences *among* them). Thus, an incidence rate could, in principle at least, characterize the intraindividual frequency of some recurrent event (seizure, arrhythmia, infection, intoxication, etc.). It is, however, characteristic of the epidemiologic outlook to make a sharp distinction between individual and population characterizations (cf. Examples 1. 1 1. 4), and in this spirit the recurrence pattern within individuals is viewed as an aspect of the condition itself, a basis for quantification of severity perhaps, and the incidence concept is confined to events that occur among individuals.

Because the concept of incidence is divorced from that of recurrence, it generally refers to *first* events only, but there is subtlety to this. The first event may be first only in the sense of *first new* event or *first recurrence* (of heart attack, for example). For individuals with one previous event, the first new event (first recurrence) is obviously the second event overall, and the occurrence of a possible third event would be ignored in the incidence rate. Similarly, for those who have already had two events, one considers the incidence of the third event, ignoring any potential subsequent recurrences. The point is that *no more than one event* (new) is properly tallied for any given individual toward an incidence measure for a population.

Although individuals who have had an event are candidates for a subsequent event, those who have had a first (kth) event are no longer candidates for a first (kth) event (ever). To say in the latter case that the incidence is zero is not to express knowledge about reality; it is, instead, a statement that involves an absurd mental construct. To avoid such problems, incidence is *defined* only for populations of *candidates* for the event individuals who could experience the event in principle (as a matter of logic). Thus

1. Incidence of cervical cancer for those who have had this disease but do not have it any more is defined.
2. Incidence of cervical cancer for men and for women without a cervix is defined. (Indeed, even its magnitude is known a priori, not by mere logic but through substantive insight into the prerequisites for its occurrence.)
3. Incidence of cervical cancer for those who have this disease is not defined (as only noncases are logically candidates for it).
4. Incidence of cervical cancer is not defined for a mixture of cases and noncases (because it is undefined for the cases).
5. The incidence of death among the dead is not defined. (Only the living are logically candidates for death, by the very nature of the concept of death, that is, death is construed to occur among the living only.)

It should be noted that for incidence to be defined, that is, for a population to be one of candidates, the population need *not* be "*at risk*" in the sense of having a nonzero incidence (just as velocity need not be nonzero for it to be defined). Thus it is consonant with proper conceptualization of incidence to say that the incidence of cervical cancer in men is zero. For a candidate population to manifest incidence (events) it must move over time. In this regard, there is a need to distinguish between two basic types of population experience over time:

1. *Cohort* experience (Section 3.2. 1. 1), in which an enumerable set of individuals, all candidates initially (at T = to), moves over the *risk period* (Section 3.3) at issue. *Dynamic population* experience (Section 3.2.1.2), in

which a population of a given size but with turnover of membership moves over *calendar time*, with all members being candidates throughout (so that the event at issue is among the mechanisms of removal of individuals from the candidate population).

These two types of population experience may be viewed in terms of different types of incidence rate.

The availability of a cohort experience may suggest its direct characterization in terms of *cohort*, or *cumulative*, incidence rate. This type of incidence rate is a proportion, the proportion of the population of candidates, defined as of some zero time ($T = t_0$), who experience the event during the risk period at issue. It is an attractive direct measure of observed occurrence predominantly, though not exclusively, in situations in which each member of the cohort is followed up to the event at issue or to the end of the risk period, without attrition of the cohort due to loss to follow up or extraneous mortality. This type of incidence rate is often of theoretical and practical interest as well.

The definition of the risk period to which a cohort (cumulative) incidence rate (an incidence proportion) refers is either substantive (and variable) or arbitrary (and fixed). Examples of the former type of incidence rate in epidemiology include the following:

1. "Hospital mortality" in myocardial infarction (proportion of patients entering a hospital who die before discharge).
2. "Fetal death rate" (proportion of youngest fetuses dying any time in the prenatal period).
3. "Life time incidence" of breast cancer (proportion of young people who will ever develop breast cancer).
4. "Incidence" of postpartum depression (proportion of deliveries followed by maternal depression attributable to the delivery).
5. "Incidence" of venereal disease in subsequent contacts of active cases, or "secondary attack rate" (proportion of contacts who develop the disease from the contact).

The cohort type incidence rate with an arbitrary, fixed risk period is exemplified by the following:

1. "Neonatal death rate" (proportion of live born babies dying within 28 days).
2. "Five year mortality" among survivors of first myocardial infarction (proportion of cases of first, nonfatal infarction dying within 5 years).

Cohort incidence rates of this latter type translate immediately to fractiles of the *precurrence* period (waiting time) to the event. For example, if 35- and 75-year incidence rates of death for a birth cohort are 5 and 50%, respectively, these ages represent the 5th and 50th centiles of the waiting time (at birth) to death, that is, of the duration of death's precurrence period (life).

Whereas the amount of experience in an empirical cohort incidence rate (for a given span of time) is characterized in terms of the size (S) of the cohort, dynamic population experience is measured in terms of time (T) or more specifically *candidate time*, which is the integral of the size of the (dynamic) candidate population over the observation period.

For an experience of this latter type, with a certain number c of events occurring in it (cases "emitted" from it), the incidence rate is, naturally, that number divided by the candidate time, c/T. This rate is not a proportion, as the numerator is not a subset of the denominator. Rather, it is of the form of denisty measures in general. The dimensionality of this particular measure – *incidence density* – is inverse time. For example, if 15 cases arose from an experience of 5000 candidate years, the incidence density was $3/(10^3 yr)$.

There is a direct relation between incidence density (ID) and cohort (cumulative) incidence (CI) of the second type. Specifically, incidence density determines for a cohort (defined at $T = t_0$) the proportion which *in the absence of attrition* experiences the event before some common, quantitatively (nonsubstantively) defined subsequent point in the time ($T = t_1$). With ID_t the ID at $T = t$, the CI for the interval t_0 to t_1 is (Chiang, 1968), Miettinen 1976a)

$$CI_{t_0,t_1} = 1 - \exp\left[- \int_{t_0}^{t_1} (ID_t) dt \right].$$

If the incidence density is known for categories of time (e.g., age categories) in the interval at issue, the cohort (cumulative) incidence may be derived as follows:

$$CI_{t_0,t_1} = 1 - \exp\left[- \sum_i (ID_i) d_i \right],$$

where the summation is over the categories in the interval (t_0 to t_1), and d_i is the duration of the ith interval.

It should be noted that this relation does not obtain between ID and CI referring to a substantively defined period of time. The latter depends not only on ID_t but also on the distribution of the risk period among individuals,

that is, on the density of risk-time terminations for reasons unrelated to the event at issue.

A.1.3. RISK

The concept of *risk* is related to that of incidence proportion. It is the probability of a particular event, especially an untoward one, such as the inception of a particular illness. Thus, the risk of an (adverse) event is akin to its incidence in the sense that it has to do with its inception. As a probability, however, risk is inherently a *theoretical*, nonempirical entity, whereas incidence can be either theoretical or empirical. Moreover, it refers to *individuals* (of a given kind), whereas incidence characterizes populations. Thus, in a given kind of surgical situation, an empirical incidence of operative death among a series of patients serves as an estimate of the theoretical incidence among patients of that kind, operated on in that manner. By the same token, it serves as an estimate of the risk for a patient of that kind, not otherwise specified, to die after being operated on in that manner. The individual risk of operative death, a theoretical value, estimated from all the relevant experience, is not revealed by the actual, empirical outcome of the operation. Thus, if the risk was 5% before the operation, it remains 5% even in light of and regardless of the outcome.

Analogously with incidence, risk is not a singular parameter of nature. Its value depends on the specifications of the situation, on *determinants* of risk. In a further analogy, such a conditional risk remains quantitatively nebulous, because it depends on yet other, unspecified determinants.