CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

1

Rates and their properties

Rates are ratios constructed to compare the change in one quantity to the change in another. For example, postal rates are the price per unit weight for mailing a letter (price per ounce); miles divided by time produces a rate of speed (miles per hour). However, to understand and clearly interpret a rate applied to human survival data, a more detailed description is necessary. This description begins with Isaac Newton, who in the 17th century mathematically defined a rate and derived many of its properties.

The key to describing human survival, measured by rates of death or disease, is a specific function, traditionally denoted by $S(t)$, called the *survival function*. A survival function produces the probability of surviving beyond a specific point in time (denoted t). In symbols, a formal definition is

$$\begin{aligned} S(t) &= P(\text{surviving from time} = 0 \text{ to time} = t) \\ &= P(\text{surviving during interval} = [0, t]) \end{aligned}$$

or, equivalently,

$$S(t) = P(\text{surviving beyond time } t) = P(T \geq t).$$

Because $S(t)$ is a probability, it is always between zero and one for all values of t $(0 \leq S(t) \leq 1)$.

A simple survival function, $S(t) = e^{-0.04t}$, illustrates this concept (Figure 1.1). Perhaps such a function describes the pattern of 18th-century mortality for any age t (probability of living beyond age t). The probability of surviving beyond $t = 20$ years is, for example, $S(20) = P(T \geq 20) = e^{-0.04(20)} = 0.449$ (Figure 1.1). Similarly, this survival function dictates that half the population lives beyond 17.327 years. Thus,

$$P(\text{surviving beyond 17.327 years}) = S(T \geq 17.327) = e^{-0.04(17.327)} = 0.50.$$

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

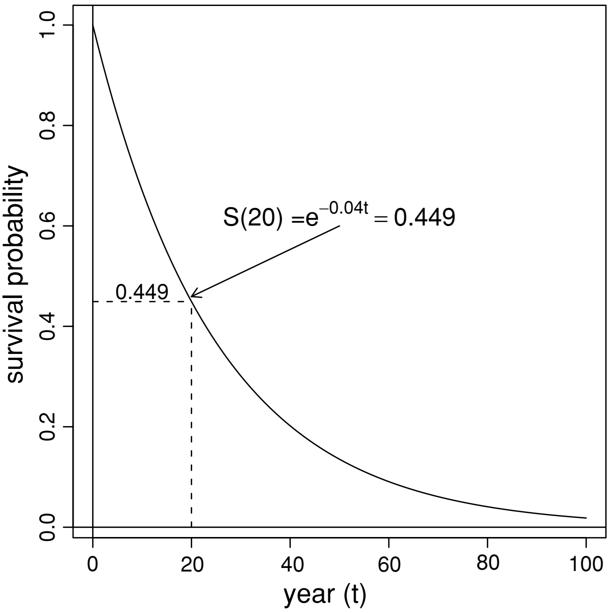

Figure 1.1. A simple survival function—$S(t) = P(T \leq t) = e^{-0.04t}$.

To create a rate that does not depend on the length of the time interval, Newton defined an instantaneous rate as the change in $S(t)$ as the length of the time interval (denoted δ) becomes infinitesimally small. This version of a rate is *the derivative of the survival function* $S(t)$ *with respect to t* or, in symbols,

$$\text{the derivative of } S(t) = \frac{d}{dt}S(t).$$

The derivative of a function is a rich concept and a complex mathematical tool completely developed in a first-year calculus course. From a practical point of view, the derivative is closely related to the slope of a line between two points (an appendix at the end of the chapter contains a few details). That is, for two points in time (t and $t + \delta$), the derivative is approximately

$$\frac{d}{dt}S(t) \approx \frac{S(t+\delta) - S(t)}{\delta}$$
$$= \text{slope of a straight line between } S(t) \text{ and } S(t+\delta).$$

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

When the change in the survival function $S(t)[S(t)$ to $S(t+\delta)]$ is divided by the corresponding change in time t (t to $t+\delta$), one version of a rate becomes

$$\text{rate} = \frac{\text{change in } S(t)}{\text{change in time}} = \frac{S(t+\delta) - S(t)}{(t+\delta) - t} = \frac{S(t+\delta) - S(t)}{\delta}.$$

The proposed rate, constructed from two specific values of the survival function $S(t)$ and the length of the time interval δ, consists of the change (decrease) in the survival function $S(t)$ relative to the change (increase) in time (δ). For small values of δ, this rate (the slope of a line) hardly differs from an instantaneous rate. In the following, the slope of a line (one kind of rate) is frequently used to approximate the derivative of the survival function at a specific point in time, an instantaneous rate.

Newton's instantaneous rate is rarely used to describe mortality or disease data, because it does not reflect risk. A homicide rate, for example, of 10 deaths per month is easily interpreted in terms of risk only when it refers to a specific population size. A rate of 10 deaths per month in a community of 1,000 individuals indicates an entirely different risk than the same rate in a community of 100,000.

When the instantaneous rate $(d/dt)S(t)$ is divided by the survival function $S(t)$, it reflects risk. To measure risk, a relative rate is created, where

$$\text{instantaneous relative rate} = h(t) = -\frac{\frac{d}{dt}S(t)}{S(t)}.$$

Multiplying by -1 makes this relative rate a positive quantity, because $S(t)$ is a decreasing function (negative slope). An instantaneous relative rate $h(t)$ is usually called a *hazard rate* in human populations and a *failure rate* in other contexts. The same rate is sometimes called *the force of mortality* or *an instantaneous rate of death* or, from physics, *relative velocity.*

Two properties of a hazard rate complicate its application to collected data. The exact form of the survival function $S(t)$ must be known for all values of time t and the hazard rate is instantaneous. Knowledge is rarely available to unequivocally define $S(t)$ completely, instantaneous quantities are not intuitive, and interpretation frequently requires special mathematical/statistical tools.

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

Instead of an instantaneous rate, an average rate is typically used to measure risk, particularly from epidemiologic and medical survival data. Formally, a rate averaged over a time interval from t to $t + \delta$ is

$$\text{average rate} = \frac{S(t) - S(t+\delta)}{\int_t^{t+\delta} S(u)du}.$$

In more natural terms, an average rate over a specified time period is simply the proportion of individuals who died ("mean number of deaths") divided by the mean survival time for all individuals at risk during that period. Equally, an average rate is the total number of individuals who died divided by the total accumulated time at risk. Geometrically, the value in the numerator of an average rate is the decrease in the survival probability between the two points t and $t + \delta$. The value of the integral in the denominator is the area under the survival curve $S(t)$ between the same two points and equals the mean survival time of individuals who lived the entire interval or died during the interval.

For the survival function $S(t) = e^{-0.04t}$ and the time interval $t = 20$ to $t = 25$ years ($\delta = 5$ years), the proportion of individuals who died (mean number of deaths) is $S(20) - S(25) = e^{-0.80} - e^{-1.00} = 0.449 - 0.368 = 0.081$ (Figure 1.1). The mean survival time for all individuals at risk (area) during the interval 20 to 25 years ($\delta = 5$) is

$$\begin{aligned}\text{area} &= \int_t^{t+\delta} S(u)du = \int_{20}^{25} e^{-0.04u}du\\ &= \frac{e^{-0.04(20)} - e^{-0.04(25)}}{0.04} = \frac{0.449 - 0.368}{0.04} = \frac{0.081}{0.04}\\ &= 2.036 \text{ person-years.}\end{aligned}$$

Thus, the mean survival time lived by individuals who survived the entire five-year interval and those who died during the interval (20–25 years) is 2.036 years. A mean time at risk of 2.036 years makes the average mortality rate

$$\begin{aligned}\text{average rate} &= \frac{\text{mean number of death}}{\text{mean survival time}} = \frac{e^{-0.80} - e^{-1.00}}{2.036} = \frac{0.081}{2.036}\\ &= 0.040 \text{ deaths per person-year}\\ &= 40 \text{ deaths per 1,000 person-years.}\end{aligned}$$

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

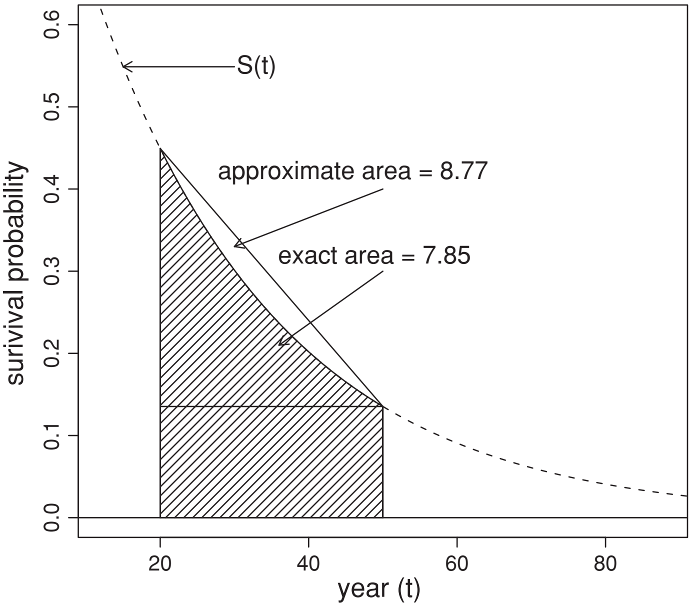

Figure 1.2. The geometry of an approximate average rate for the interval $t = 20$ to $t + \delta = 50$ (approximate rate = 0.036 and exact = rate = 0.040).

In many situations, particularly in human populations, the area under the survival curve is directly and accurately approximated without defining the survival function $S(t)$, except at two points. When the survival function between the two points t and $t + \delta$ is a straight line, the area under the curve has a simple geometric form. It is a rectangle plus a triangle (Figure 1.2). Furthermore,

$$\text{area of the rectangle} = \text{width} \times \text{height} = ([t + \delta] - t) \times S(t + \delta)$$
$$= \delta S(t + \delta)$$

and

$$\text{area of the triangle} = \tfrac{1}{2}\,\text{base} \times \text{altitude}$$
$$= \tfrac{1}{2}([t + \delta] - t) \times [S(t) - S(t + \delta)]$$
$$= \tfrac{1}{2}\delta[S(t) - S(t + \delta)],$$

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

Table 1.1. Approximate and exact areas for the time interval $t = 20$ and $t + \delta = 20 + \delta$ for the survival function $S(t) = e^{-0.04t}$ (exact rate = 0.04).

δ	t to $t + \delta$	$S(t)$	$S(t + \delta)$	$d(t)$	area*	area**	*rate***
30	20 to 50.0	0.449	0.135	0.314	7.850	8.770	0.036
20	20 to 40.0	0.449	0.202	0.247	6.186	6.512	0.038
10	20 to 30.0	0.449	0.301	0.148	3.703	3.753	0.039
5	20 to 25.0	0.449	0.368	0.081	2.036	2.043	0.039
1	20 to 21.0	0.449	0.432	0.018	0.440	0.441	0.040
0.1	20 to 20.1	0.449	0.448	0.002	0.045	0.045	0.040

* = exact $S(t)$

** = approximate (straight line).

making the total area

$$\begin{aligned}\text{area} &= \text{rectangle} + \text{triangle}\\ &= \delta S(t+\delta) + \tfrac{1}{2}\delta[S(t) - S(t+\delta)] = \tfrac{1}{2}\delta[S(t) + S(t+\delta)].\end{aligned}$$

Figure 1.1 displays the geometry for the survival function $S(t) = e^{-0.04t}$. For the interval $t = 20$ to $t + \delta = 25 (\delta = 5)$, the area of the rectangle is $\delta S(25) = 5(0.368) = 1.839$ and the area of the triangle is $\frac{1}{2}\delta[S(20) - S(25)] = \frac{1}{2}(5)[0.449 - 0.368] = 0.204$, making the total area $1.839 + 0.204 = 2.043$ (mean time-at-risk during the interval). Again, the mean number of deaths is 0.0814. A measure of risk becomes the approximate average rate = $0.0814/2.043 = 0.039$ (exact = 0.04) or 39 deaths per 1,000 person-years.

The approximate area is usually an accurate estimate of the exact area because the human survival curve in most situations is approximately a straight line over a specific and moderately small time interval. More simply, when a straight line and part of a survival function $S(t)$ are not very different, using an approximation based on a straight line works well [straight line $\approx S(t)$]. Table 1.1 and Figure 1.2 illustrate this similarly for $t = 20$ years, where the exact average rate is 0.04 for all time intervals.

Because $S(t)$ represents the probability of surviving beyond time t, the difference $S(t) - S(t + \delta) = d(t)$ represents the probability of dying in the interval from t to $t + \delta$. In addition, the approximate area under the survival curve $S(t)$ has three equivalent forms, $\delta[S(t) - \frac{1}{2}d(t)]$ or $\delta[S(t + \delta) + \frac{1}{2}d(t)]$ or $\frac{1}{2}\delta[S(t) + S(t + \delta)]$, for the time interval t to $t + \delta$. All three

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

expressions are the sum of the mean time lived by those who survived the entire interval [rectangle $= \delta S(t+\delta)$] plus the mean survival time lived by those who died [triangle $= \frac{1}{2}\delta d(t)$]. Therefore, to calculate the mean number of deaths and to approximate the mean time at risk, all that is needed is the values of $S(t)$ at the two points in time, namely t and $t+\delta$. The ratio of these two mean values is the average approximate mortality rate.

Example

Suppose that out of 200 individuals at risk, 100 individuals were alive January 1, 2004, and by January 1, 2006, suppose 80 of these individuals remained alive. In symbols, $t = 2004$, $t + \delta = 2006$ ($\delta = 2$ years), $S(2004) = 100/200 = 0.50$, and $S(2006) = 80/200 = 0.40$, making the proportion of the original 200 at-risk individuals who died during these two years $d(2004) = S(2004) - S(2006) = 0.50 - 0.40 = 0.10$ or $20/200 = 0.10$. The approximate area enclosed by the survival curve for this $\delta = 2$-year period is $\frac{1}{2} \cdot 2(0.50 + 0.40) = 0.90$ person-years (area). The average approximate rate becomes $R = (0.50 - 0.40)/0.90 = 0.10/0.90 = 0.111$ or, multiplying by 1,000, the rate is 111 deaths per 1,000 person-years. Rates are frequently multiplied by a large constant value to produce values greater than one (primarily to avoid small fractions). The mortality rate R reflects the approximate average risk of death over the period of time from 2004 to 2006 experienced by the originally observed 200 individuals. In addition, the total accumulated person-years lived by these 200 individuals during the two-year period is $200(0.90) = 180$ person-years because the mean years lived by these 200 individuals during the interval is 0.90 years. Therefore, the number who died $(100 - 80 = 200(0.10) = 20)$ divided by the total person-years (180) is the same approximate average rate,

$$\text{average rate} = R = \frac{\text{total deaths}}{\text{total person-years}} = \frac{20}{180} = \frac{0.10}{0.90} = 0.111.$$

The example illustrates the calculation of an approximate average rate free from the previous two constraints. It is not necessary to define the survival function $S(t)$ in detail and the rate is not instantaneous. The only requirements are that the two values $S(t)$ and $S(t+\delta)$ be known or accurately estimated and the survival curve be at least close to a straight line over the time period considered. Both conditions are frequently fulfilled by routinely collected human data providing a huge variety of mortality and disease rates

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

(see the National Center for Health Statistics or the National Cancer Institute Web site—http://www.cdc.gov/nchs/ or http://www.nci.nih.gov).

It is important to note (or review) the equivalence of two ways to calculate a rate. An approximate average rate is calculated by dividing the mean number of deaths (the proportion of deaths) that occur during an interval by the mean survival time for that interval. That is, the ratio of means is

$$\begin{aligned}\text{approximate average rate} &= \frac{\text{mean number of deaths}}{\text{mean survival time}}\\ &= \frac{d(t)}{\frac{1}{2}\delta[S(t)+S(t+\delta)]}.\end{aligned}$$

Or, more usually but less intuitively, the same rate calculated from a specific number of individuals (denoted l) in terms of deaths and total person-years is

$$\begin{aligned}\text{approximate average rate} &= \frac{\text{total number of deaths}}{\text{total person-years at-risk}}\\ &= \frac{l d(t)}{l\left\{\frac{1}{2}\delta[S(t)+S(t+\delta)]\right\}}.\end{aligned}$$

These two rates are identical.

An approximate average rate is sometimes calculated by dividing the observed number of deaths by the number of individuals alive at the midpoint of the interval considered. For example, for the year 2000 in Marin County, California, there were 247,653 women alive halfway through the year and 494 deaths from cancer for the entire year. The annual average cancer mortality rate becomes 494 deaths divided by the midinterval count of 247,653 persons, and the approximate average rate $= (494/247{,}653) \times 100{,}000 = 199.5$ deaths per 100,000 person-years. This “short cut” is no more than an application of the fact that the midinterval population for l individuals is approximately the total accumulated person-years at risk or, in symbols, the midinterval population $l \times \delta S(t+\frac{1}{2}\delta)$ is approximately $l \times \frac{1}{2}\delta[S(t)+S(t+\delta)]$ and is exact when $S(t)$ is a straight line.

A number of ways exist to calculate an approximate average rate from mortality data based on the assumption that a straight line closely approximates the survival function. The following example illustrates three methods using

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

Table 1.2. U.S. mortality rates (all causes of death) age 65–74 for the years 1999, 2000, and 2001.

i	Year	Deaths (d_i)	Person-years (pyrs$_i$)	Rate/100,000
1	1999	387,437	16,167,771	2396.4
2	2000	376,986	16,100,428	2341.5
3	2001	367,128	15,969,452	2298.9
	Total	1,131,551	48,237,651	2345.8

U.S. mortality data for individuals aged 65 to 74 during the years 1999–2001 (Table 1.2).

Method 1:

$$\text{rate} = \frac{\sum d_i}{\sum \text{pyrs}_i} = \frac{1,131,551}{48,237,651} = 2{,}345.8 \text{ deaths per 1000,000 person-years}$$

Method 2:

$$\text{rate} = \frac{d_2}{\text{pyrs}_2} = \frac{376,986}{16,100,428} = 2{,}341.5 \text{ deaths per 100,000 person-years}$$

and

Method 3:

$$\text{rate} = \frac{\sum d_i}{3 \times \text{pyrs}_2} = \frac{1,131,551}{3 \times 16,1000,428}$$
$$= 2,342.7 \text{ deaths per 100,000 person-years.}$$

The three methods produce essentially the same average mortality rate because the change in human mortality over short periods of time is usually close to linear.

Another frequent measure of risk is a probability. A probability, defined in its simplest terms, is the number of equally likely selected events (a subset) that might occur divided by the total number of all equally likely relevant events that could possibly occur (the entire set). In symbols, if $n[A]$ represents the number of selected events among a total of n equally likely events, then

$$\text{probability of event } A = P(A) = \frac{n[A]}{n}.$$

© Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-71937-7 - Survival Analysis for Epidemiologic and Medical Research: A Practical Guide
Steve Selvin
Excerpt
More information

For example, the probability of death (denoted q) is $q = d/n$, where $n[A] = d$ represents the number of deaths among n individuals who could possibly have died. The complementary probability of surviving is $1 - q = p = (n - d)/n$. Notice the explicit requirement that all n individuals be members of a population with a proportion of q deaths and p survivors (next topic). Other, more rigorous definitions of probability exist, but this basic definition is sufficient for the following applications to survival analysis.

A probability is always zero (impossible event) or one (sure event) or between zero and one. In addition, a probability is unitless and does not depend directly on time. On the other hand, a rate can be any positive value, is not unitless (per person-time), and depends directly on time. Nevertheless, these two quantities are closely related. For an average approximate rate R and a probability q,

$$R = \frac{S(t) - S(t+\delta)}{\delta[S(t) - \frac{1}{2}d(t)]} = \frac{S(t)/S(t) - S(t+\delta)/S(t)}{\delta[S(t)/S(t) - \frac{1}{2}d(t)/S(t)]} = \frac{q}{\delta(1 - \frac{1}{2}q)}$$

and thus

$$q = \frac{\delta R}{1 + \frac{1}{2}\delta R},$$

where probability of death q is $d(t)/S(t)$ for the interval $(t, t + \delta)$. The probability of survival becomes $1 - q = p = S(t+\delta)/S(t)$. Note that q, and necessarily p, are conditional probabilities, conditional on being alive at time t. More specifically,

$$\begin{aligned}\text{probability of death} = q &= P(\text{death between } t \text{ and } t+\delta \mid \text{alive at time } t)\\ &= \frac{P(\text{death between } t \text{ and } t+\delta)}{P(\text{alive at time } t)} = \frac{d(t)}{S(t)}.\end{aligned}$$

The probability of death or disease in human populations is almost always small ($p \approx 1$ or $q \approx 0$), making the relationship between a rate and a probability primarily a function of the length of the time interval δ. In symbols, the rate $= R \approx q/\delta$ when $\frac{1}{2}\delta q \approx 0$. When the period of time considered is one year, an average annual mortality rate and a probability of death typically produce almost identical values ($R \approx q$). These two quantities are more or less interchangeable and, particularly in the study of human mortality and disease, it often makes little practical difference which measure of risk is used.

© Cambridge University Press www.cambridge.org