

# LECTURE 1

## Rates and their properties

Rates come in a variety of forms. There are insurance rates, postal rates, tax rates as well as rates of speed, rates of births and rates of weight gain. In all situations, a rate measures a change in one quantity relative to a change in another. For example, postal rates are the price per unit weight for mailing a letter (price per ounce) and miles divided by time produce rates of speed (miles per hour). However, to understand and clearly interpret a rate calculated from mortality and disease data, a more extensive approach is necessary. The approach begins with Issac Newton who in the 17<sup>th</sup> century mathematically defined a rate and derive many of its properties.

A definition of a rate starts with mathematically describing a changing pattern over time, represented by the symbol  $S(t)$ . One version of a rate is created by dividing the change in the function  $S(t)$  [ $S(t)$  to  $S(t + \delta)$ ] by the corresponding change in time  $t$  ( $t$  to  $t + \delta$ ) producing the

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

Rates with respect to time apply to a variety of situations but a specific function,

traditionally denoted by  $S(t)$ , is central to the analysis of survival data. It is called the *survival function* and is defined as the probability of surviving beyond a specific point in time (denoted  $t$ ). In symbols,

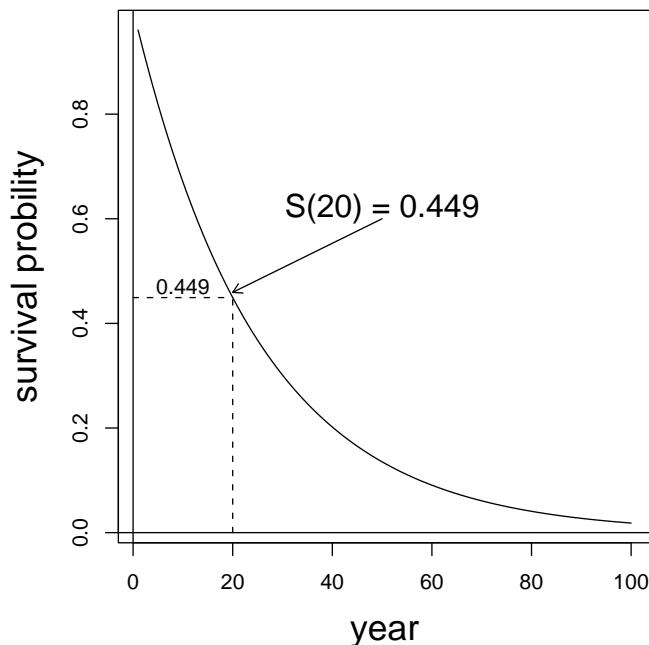
$$S(t) = P(\text{surviving from time } = 0 \text{ to time } = t) = P(\text{surviving during interval } = [0, t])$$

or equivalently

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

To illustrate, consider a simple survival function  $S(t) = e^{-0.04t}$ . Perhaps,  $S(t)$  describes a decrease in population size over time due to mortality. The probability of surviving beyond  $t = 20$  years is then  $S(20) = P(T \geq 20) = e^{-0.04(20)} = 0.449$  (Figure 1.0).

**Figure 1.0 Simple survival curve --  $S(t) = e^{-0.04t}$**



The proposed version of a "rate" depends on both the specific survival function  $S(t)$  and the length of the time interval  $\delta$ . To create a rate that does not depend on the length of the interval, Newton defined a "pure rate" as the change in  $S(t)$  as the length of the interval  $\delta$  became infinitesimally small and called this quantity the derivative of  $S(t)$  with respect to  $t$  or, in symbols,

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

The derivative of  $S(t)$  is an instantaneous rate.

A derivative is a rich concept and a complex mathematical tool described 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. That is, approximately the derivative is

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

for small values of  $\delta$  and becomes exactly the derivative when  $\delta$  approaches zero. In the following, the slope of a line (one kind of "rate") is frequently used to describe approximately the derivative of the survival function (an instantaneous rate) at a specific point in time ("pure rate").

Newton's instantaneous rate is rarely used to summarize mortality or disease data because it does not directly reflect risk. A homicide rate, for example, of 10 deaths per month is easily interpreted in terms of risk when it refers to a specific population size. A rate of 10 deaths per month in a community of 1000 individuals indicates an entirely

different risk than the same rate in a community of 100,000. When the instantaneous rate

$\frac{d}{dt} S(t)$  is divided by  $S(t)$ , it reflects risk. To measure risk, a relative rate is defined as

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

Multiplying by -1 makes this relative rate a positive quantity when  $S(t)$  is a decreasing function (negative slope). When  $S(t)$  refers to mortality or disease or other kinds of failures, the 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 *an instantaneous rate* or *the force of mortality*.

Two elements of Newton's pure rate and a hazard rate that complicate the application to collected data are: the exact form of the function  $S(t)$  must be known and both kinds of rates are instantaneous. Knowledge is rarely available to unequivocally define  $S(t)$  and instantaneous quantities are conceptually difficult and interpretation requires special mathematical/statistical tools.

Instead of an instantaneous rate, an average rate is a frequently used measure of risk. A rate averaged over a time interval  $t$  to  $t + \delta$  is

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

In more intuitive terms, an average rate over a defined period is the proportion of

individuals who died divided by their mean time at risk or, equally, the total number of individuals who died divided by their total-time-at-risk. Geometrically, the value of the integral in the denominator is the area under the curve  $S(t)$  between the two points  $t$  and  $t + \delta$ . For the survival function  $S(t) = e^{-0.04t}$  and the interval  $t = 20$  to  $t = 25$  ( $\delta = 5$ ), the proportion who died is  $S(20) - S(25) = e^{-0.80} - e^{-1.00} = 0.0814$ . The average time at risk (area) is

$$\int_t^{t+\delta} S(u)du = \int_{20}^{25} e^{-0.04u} du = \frac{e^{-0.04(20)} - e^{-0.04(25)}}{0.04} = \frac{0.449 - 0.368}{0.04} = 2.036 \text{ years}$$

making the average mortality rate

$$\text{average rate} = \frac{e^{-0.80} - e^{-1.00}}{2.036} = \frac{0.0814}{2.036} = 0.040 \text{ deaths per person - years.}$$

In most cases, particularly in human populations, the area under the survival curve  $S(t)$  is accurately approximated without defining the function  $S(t)$  in detail. When the survival curve 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 2.0). Furthermore,

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

and

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

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

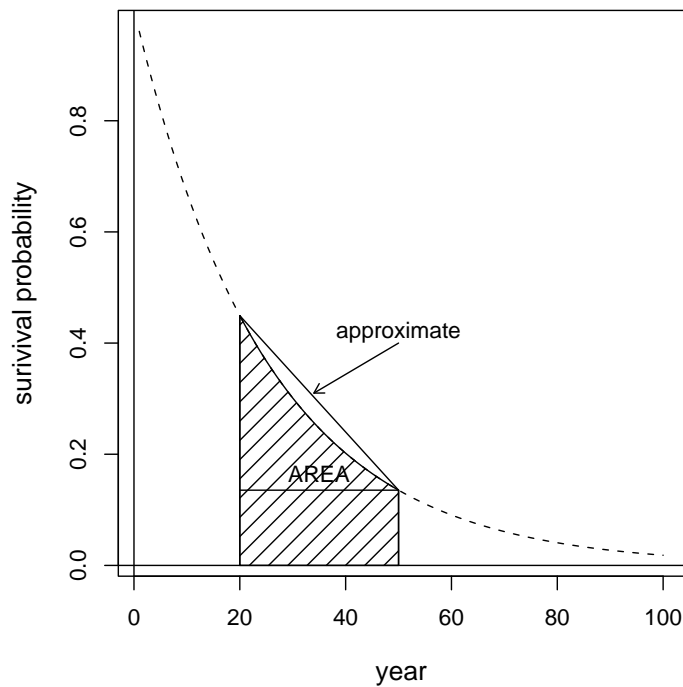
makes the total area

$$\text{area} = \text{rectangle} + \text{triangle}$$

$$= \delta S(t + \delta) + \frac{1}{2} \delta[S(t) - S(t + \delta)] = \frac{1}{2} \delta[S(t) + S(t + \delta)].$$

Figure 2.0 displays the geometry for the survival function  $S(t) = e^{-0.04t}$ .

**Figure 2.0 The geometry of an approximate average rate**



For  $t = 20$  to  $t + \delta = 25$ , 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 = 2.043. The approximate rate becomes *average rate* =  $0.0814/2.043 = 0.039$ .

When  $\delta$  is moderately small, the approximate area is usually an accurate estimate of the

exact area. That is, a straight line and the survival function  $S(t)$  are not very different over a short interval (Figure 2.0). Table 1.0 illustrates for  $t = 20$  years.

**Table 1.0 Approximate and exact areas between  $S(t)$  and  $S(t + \delta)$  for  $S(t) = e^{-0.04t}$**

$\delta$	$t$ 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	0.040
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  $S(t)$  and \*\* = approximate (straight line)

Returning to the mortality example, where  $S(t)$  represents the probability of surviving beyond time  $t$ , then  $S(t) - S(t + \delta) = d(t)$  represents the proportion who died in the interval  $t$  to  $t + \delta$ . The area under the survival curve  $S(t)$  is approximately  $\delta[S(t) - \frac{1}{2} d(t)]$  or  $\delta[S(t + \delta) + \frac{1}{2} d(t)]$  or  $\frac{1}{2} \delta[S(t) + S(t + \delta)]$ . To approximate this area, all that is required is the value  $S(t)$  at the two points in time,  $t$  and  $t + \delta$ . In addition, the area of the rectangle is approximately the mean years lived by those who survived the entire interval and the area of the triangle is approximately the mean years lived by those who died within the interval. Therefore, the total area under the survival curve is approximately the mean years lived by all individuals during the interval  $t$  to  $t + \delta$  who were alive at the beginning the interval (at risk at time  $t$ ). When the survival probabilities are described by a straight line these approximate quantities are exact.

Suppose that out of 200 individuals at-risk, 100 individuals were alive January 1, 2002 and by January 1, 2004 suppose 80 of these individuals remained alive. In symbols,

$t = 2002$ ,  $t + \delta = 2004$  ( $\delta = 2$  years),  $S(2002) = 100/200 = 0.50$ ,  $S(2004) = 80/200 = 0.40$  and the proportion who died is  $d(2002) = 0.50 - 0.40 = 0.10$  or  $20/200 = 0.10$ . The approximate area enclosed by the survival curve for the  $\delta = 2$  year period is  $\frac{1}{2}2(0.50 + 0.40) = 0.90$  years. The average approximate rate becomes  $R = (0.50 - 0.40)/0.90 = 0.10/0.90 = 0.111$  or, multiplying by 1000, the rate becomes  $R$  is 111 deaths per 1000 person-years. Rates are frequently multiplied by a large constant value to produce values greater than one (primarily to avoid small fractions with many zeros). The rate  $R$  reflects the approximate average risk over the period of time 2002 to 2004 for the observed 200 individuals. In addition, the total person years lived by the 100 individuals between the years 2002 and 2004 is  $100(0.90) = 90$  person-years because the mean years lived is 0.90 years. Therefore, the number who died ( $200(0.10) = 20$ ) divided by the total person-years (90) is also the approximate average rate

$$\text{average rate} = R = \frac{\text{total deaths}}{\text{total - person - years}} = \frac{20}{90} = 0.111.$$

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 interval considered. Both the conditions are frequently fulfilled by routinely collected human data providing a huge variety of accurate mortality and disease rates (see the National Center for Health Statistics or the National Cancer

Institute websites -- <http://www.cdc.gov/nchs/> and <http://www.nci.nih.gov>).

Another measure of risk is a probability. A probability is simply defined as the number of equally likely specific events (a subset) that could occur divided by the total number of all equally likely relevant events (the entire set) that could possibly occur. In symbols, if  $n[A]$  is the number of specific events among a total of  $n$  equally likely events, then

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

For example, the probability of death (denoted  $q$ ) is  $q = d/n$  where  $n[A] = d$  represents the number of deaths among  $n$  at-risk individuals. The complementary probability of surviving is  $1 - q = p = (n - d)/n$ . Notice the explicit requirement that all  $n$  individuals are equally likely to die (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 a unitless quantity that 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 on time. Nevertheless, these two quantities are related. For an average rate  $R$  and a probability  $q$ ,

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

or

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

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

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

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 a function primarily of the time interval length  $\delta$ . In symbols,  $\text{rate} = R \approx q/\delta$ . When the period of time considered is one year, an average annual rate and a probability typically produce almost identical values. These two quantities are more or less interchangeable and, particularly in the study of human mortality and disease, it usually makes little practical difference which measure of risk is used. For example, a ratio of rates and a ratio of probabilities hardly differ when applied to the same time interval. In symbols,

$$\text{rate ratio} = \frac{R_1}{R_0} \approx \frac{q_1/\delta}{q_0/\delta} = \frac{q_1}{q_0}.$$

Under rather extreme conditions a rate and a probability can differ considerably. For example, when among 100 individuals, 80 die in the first month during a disease epidemic and the remaining 20 survive the rest of the year ( $\delta = 1$ ), the probability of death is

$q$  is  $80/100 = 0.8$  and the approximate average mortality rate  $R$  is  $80/[20 + 0.5(80)] = 1.33$  deaths per person-years. However, for the year considered, the probability of death is not small and the survival curve is not close to a straight line.

## Statistical properties of the probability of death

When a rate is estimated from survival data, a fundamental assumption frequently made about the underlying sampled population is that the probability of death (represented again by  $q$ ) is at least approximately constant. Constant in this context means that the probability  $q$  refers to a population made up of two outcomes (for example, died/survived or diseased/disease-free) with a proportion of individuals  $q$  of one kind and a proportion of individuals  $1 - q$  of another kind. Under this condition, the properties of a sample of  $n$  individuals are described by a binomial probability distribution. Therefore, the probability that the sample of  $n$  independent individuals contains exactly  $d$  individuals who died and  $n - d$  who survived is

$$P(D = d) = \binom{n}{d} q^d (1 - q)^{n-d} \quad d = 0, 1, 2, \dots, n$$

when the probability of death  $q$  is constant.

These  $n + 1$  probabilities, determined completely by the two parameters  $n$  and  $q$ , generate all the properties of a binomially distributed variable represented by  $D$ . For example, the mean of the distribution of  $D$  is  $nq$  and its variance is  $nq(1 - q)$ . The estimate of the binomial parameter  $q$  is the number of sampled individuals who died divided by the total number sampled,  $\hat{q} = d/n$ . The properties of this estimate also follow directly from the binomial probability distribution. For example, the variance of the distribution of the estimate  $\hat{q}$  is  $q(1 - q)/n$  and is naturally estimated by  $\text{variance}(\hat{q}) = \hat{q}(1 - \hat{q})/n$ .

Note: the " $\hat{\quad}$ " placed over a symbol means the value is calculated from data (subject to the

influences of sampling variation).

The variability associated with the distribution of the estimate  $\hat{q}$ , estimated by the expression  $\hat{q}(1 - \hat{q})/n$ , is due to sampling variation that accompanies all estimates. That is, another sample likely produces another value of  $\hat{q}$  because another sample will be made up of different individuals. It is this sample to sample variation that is measured by  $q(1 - q)/n$ . It is this variation that is described by a binomial distribution. Occasionally the variation associated with the estimate  $\hat{q}$  is erroneously attributed to the fact that individuals within the population vary with respect to the probability of death. Variation of the probability  $q$  among the population members (heterogeneity) is an issue (to be discussed) but it is not the variation associated with a binomial distribution, which requires that the quantity  $q$  to be constant.

Two notable issues arise in applying a binomial distribution to summarize a sample of survival data: the use of the normal distribution as an approximation and the consequences of assuming that the probability  $q$  is the same for all individuals within the sampled population (constant) when it is not.

### Normal approximation

Statistical tests and confidence intervals derived from a normal distribution are basic statistical tools used to assess the influence of sampling variation on an estimated value. In many situations they apply to the assessment of the estimated binomial probability  $\hat{q}$ . For example, an approximate 95% confidence interval is  $\hat{q} \pm 1.960\sqrt{\text{variance}(\hat{q})}$  but requires the distribution of the estimate  $\hat{q}$  to be at least approximately normal. The

accuracy of this approximate approach is best when  $q$  is in the neighborhood of 0.5 and the sample size exceeds 30 or so ( $n > 30$ ). For survival data, particularly human survival data, the probability  $q$  typically refers to mortality or disease risk and is almost always small and in many cases extremely small. A consequence of a small probability is that the binomial distribution is no longer symmetric and has a limited and positive range in the neighborhood of zero. Because the normal distribution is symmetric and likely produces negative values near zero, it is no longer a useful approximation for a binomial distribution. Alternative approaches to evaluate an estimate  $\hat{q}$  when  $q$  is small employ exact methods or transformations.

Exact methods are conceptually complicated and numerically difficult but are available as part of several statistical computer packages. Transformations, however, require only a bit of calculation and, unlike exact methods, are conceptually simple. Transformations are created to make asymmetric distributions (such as the binomial distribution with small  $q$ ) more or less symmetric. Using the transformed variable, the normal distribution once again becomes an accurate approximation and normal-based tests and confidence intervals once again apply.

Such a transformation of a small value of  $\hat{q}$  is the *logistic transformation*. A logistic transformation of an estimated probability  $\hat{q}$  (denoted  $\hat{l}$ ) is

$$\hat{l} = \log \left[ \frac{\hat{q}}{1 - \hat{q}} \right] = \log \left[ \frac{d}{n - d} \right].$$

The estimate  $\hat{l}$  has an approximate normal distribution. The value  $\hat{l}$  is the logarithm of

the odds, sometimes called the *log-odds* or *logit*. Note: all logarithms used in the following are the natural logarithms (base  $e = 2.718282$ ). The odds are defined as the probability an event occurs divided by the probability that the event does not occur (the complementary event). The odds are a popular measure of risk used most often in gambling and epidemiology. To improve the accuracy (reduced bias) of the logistic transformation a value of one-half is added to the numerator and denominator making the log-odds

$$\hat{l} = \log \left[ \frac{d + \frac{1}{2}}{n - d + \frac{1}{2}} \right].$$

The estimated variance of the normal-like distribution of the estimate  $\hat{l}$  is given by the expression

$$\text{variance}(\hat{l}) = \frac{(n+1)(n+2)}{n(d+1)(n-d+1)}.$$

The variance of  $\hat{l}$  is approximately  $\text{variance}(\hat{l}) \approx 1/(d+1)$  when  $n$  is much larger than  $d$ , which is frequently the case for mortality and disease data ( $q$  is small).

The estimated probability of death from cancer among females in Marin county, CA is  $\hat{q} = 494/247,900 = 0.001993$  or 199.3 per 100,000 women ( $d = 494$  deaths among  $n = 247,900$  women who were residents of Marin county during the year 2000). Construction of a confidence interval from this estimate provides an example of applying a logistic transformation to mortality data (small  $q$ ). The log-odds value is  $\hat{l} = \log(494.5/24706.5) = -6.215$  with estimated variance of  $\hat{l}$  given by  $\text{variance}(\hat{l}) =$

0.00202. The bounds of an approximate 95% confidence interval based on the estimated log-odds  $\hat{l} = -6.215$  and the normal distribution, as usual, are

$$A = \text{lower bound} = \hat{l} - 1.960\sqrt{\text{variance}(\hat{l})} = -6.215 - 1.960\sqrt{0.00202} = -6.303$$

and

$$B = \text{upper bound} = \hat{l} + 1.960\sqrt{\text{variance}(\hat{l})} = -6.215 + 1.960\sqrt{0.00202} = -6.127.$$

A little algebra shows that  $1/(1 + e^{-\hat{l}}) = \hat{q}$ . Therefore, the bounds for the log-odds confidence interval constructed from the normal distribution are the identically transformed into the bounds associated with the estimated probability  $\hat{q}$ . The 95% bounds become  $\text{lower bound} = 1/(1 + e^{-A}) = 1/(1 + e^{6.303}) = 0.00183$  and  $\text{upper bound} = 1/(1 + e^{-B}) = 1/(1 + e^{6.127}) = 0.00217$  or (182.7, 217.8) per 100,000 at-risk women. As required, the probability  $\hat{q} = 1/(1 + e^{6.215}) = 0.001993$  or 199.3 deaths per 100,000 women. (Details of the construction of confidence intervals based on transformed estimates are reviewed at the end of the lecture).

The same logistic transformation can be applied to compare estimated probabilities from different populations (sometimes called the *two-sample problem*). For example, the probability of a cancer death in Marin county compared to the same probability for the entire state of California indicates the magnitude of the excess risk experienced in this specific county. The Marin county probability is again 199.3 cancer deaths per 100,000 women and the probability for the entire state is 147.6 deaths per 100,000 women ( $[51,186/34,689,000] \times 100,000$ ). The corresponding logistic transformed estimates are

$\hat{l}_{marin} = -6.215$  and  $\hat{l}_{state} = -6.517$ . Again the normal distribution provides an approximate but accurate assessment of the influence of sampling variation on the observed difference in log-odds transformed values. Specifically, the comparison takes the form

$$z = \frac{\hat{l}_{marin} - \hat{l}_{state}}{\sqrt{\text{variance}(\hat{l}_{marin} - \hat{l}_{state})}} = \frac{-6.215 - (-6.517)}{\sqrt{0.00202 + 0.00002}} = \frac{0.302}{0.045} = 6.680$$

where the estimated  $\text{variance}(\hat{l}_{marin} - \hat{l}_{state}) = \text{variance}(\hat{l}_{marin}) + \text{variance}(\hat{l}_{state})$ . For the comparison of Marin county to the state as a whole, the estimated variance is  $\text{variance}(\hat{l}_{marin} - \hat{l}_{state}) = 0.00204$ . The test statistic  $z$  has an approximate standard normal distribution when the underlying Marin county and the state cancer mortality risks are the same and the estimated log-odds values then differ by chance alone. A significance probability (usually called a  $p$ -value) of  $P(|Z| \geq 6.680 | \text{no difference}) < 0.001$  leaves little doubt that random variation is not an adequate explanation of the observed difference. This statistical test is consistent with the confidence interval for Marin county. The 95% confidence interval (182.5, 217.6) defines a range of likely values of the underlying probability  $q$  and does not include the estimated probability of death for the entire state (147.6). Thus, the California value (probability or log-odds) is not a plausible value for Marin county from either perspective in light of the observed Marin county mortality rate. The two approaches rarely give very different answers, particularly when the number of sampled observations in each group is large.

#### Homogeneity of the binomial probability $q$

Human populations are never perfectly homogeneous with respect to the probability of death or disease. Age-, race-, location-, sex-specific samples of data are collected but the underlying probability of death typically remains heterogeneous to some extent even in these more homogeneous subpopulations. The consequence of ignoring this residual heterogeneity is seen by an example.

Suppose a population of 160 ( $n = 160$ ) individuals is made up of four groups heterogeneous for the probability  $q$  (defined in Table 2.0).

**Table 2.0 Four hypothetical groups ( $n = 160$ ) heterogeneous for the probability  $q$**

group	$n_i$	$d_i$	$q_i$	$v_i^*$
group 1	60	2	0.033	1.933
group 2	50	4	0.080	3.680
group 3	30	6	0.200	4.800
group 4	20	8	0.400	4.800
combined	160	20	0.125	--

\* variance =  $v_i = n_i q_i (1 - q_i)$

A natural estimate of  $q$  is  $\hat{q} = d/n = 20/160 = 0.125$  ( $d = \sum d_i$  and  $n = \sum n_i$ ), combining the four groups. The estimated variance of this estimate is  $\hat{q}(1 - \hat{q})/n = (0.125)(0.875)/160 = 0.0007$ . Both estimates ignore the heterogeneity of  $q$ . An estimate accounting for the heterogeneity is the weighted average  $\hat{q} = \sum n_i \hat{q}_i / \sum n_i = \sum d_i / n$  and is also 0.125. The estimated variance of  $\hat{q}$  accounting for the heterogeneity among the four groups is, however, reduced. The estimated variance is  $\sum v_i / n^2 = 0.0006$ . In symbols, the perhaps not very intuitive result emerges that

$\text{variance}(\hat{q} | \text{accounting for heterogeneity}) \leq \text{variance}(\hat{q} | \text{not accounting for heterogeneity})$ .

Not accounting for the heterogeneity of the probability  $q$  produces a conservative estimate of the variability. Conservative in the sense that the variance is likely too large, producing a statistical test with a larger  $p$ -value or a confidence interval with wider bounds than would be produced if the heterogeneity was known and taken into account. This difference in variability is entirely due to the heterogeneity among the  $\hat{q}_i$ -values. Specifically, the difference between the two estimated variances  $\hat{q}(1 - \hat{q})/n$  and  $\sum v_i/n^2$  is strictly a function of the differences in the  $\hat{q}_i$ -values among the subgroups. Or, in symbols, the difference is  $\sum n_i(\hat{q}_i - \hat{q})^2/n^2$ . For the hypothetical example (Table 2.0), the heterogeneity of  $q$  among the four groups measured by  $\sum n_i(\hat{q}_i - \hat{q})^2/n^2$  is 0.0001. That is,  $\hat{q}(1 - \hat{q})/n - \sum v_i/n^2 = 0.0007 - 0.0006 = 0.0001$ . Only when  $\hat{q}_i$  equals  $\hat{q}$  in all groups sampled is the variance estimated by  $\hat{q}(1 - \hat{q})/n$  strictly correct; otherwise it is biased upward. That is, the variance of  $\hat{q}$  ignoring heterogeneity is always larger because it is the sum of the variances of  $\hat{q}_i$  within each group plus the variance in the values of  $\hat{q}_i$  among the groups.

This simple example is realistic in the sense that the bias arising from ignoring heterogeneity is not only conservative but it is typically small. Therefore, not completely accounting for heterogeneity in a sampled population, a necessity in most applied situations, leads to statistical tests and estimated confidence intervals that are understated (lower power) but not likely misleading. The Marin county cancer data is an example of an estimate that does not account for the heterogeneity of the probability of death  $q$

producing a confidence interval and a statistical test that are slightly conservative.

## Simplest case: average rates, survival probabilities and hazard rates

Suppose a population of 90 year-olds is identified and their survival experience during the next ten years is the focus of interest. In this simplest case, all individuals are envisioned as dying at random over the next ten years. Thus, the average number of years lived is five because these individuals are equally likely to die at any time during the ten year period. Postulating such a mortality pattern is not entirely unrealistic and has an important application to life table calculations (a future topic). In other situations, suggesting a completely random risk of death produces an approximate description of the survival pattern. Furthermore, even this simple illustration characterizes the fundamental relationships among an average mortality rate, a survival probability and a hazard rate. More realistic situations differ in technical details but differ little in principle.

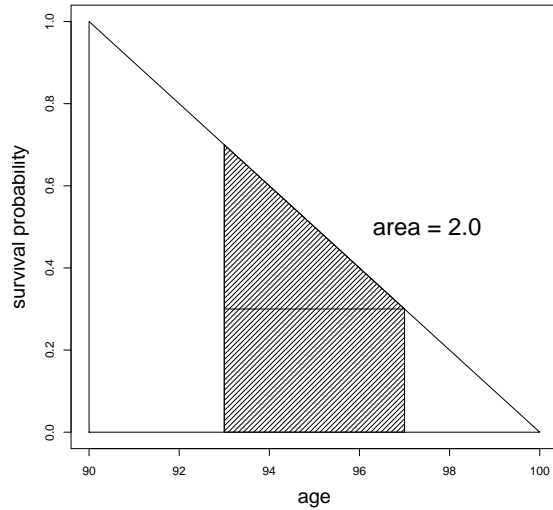
If all individuals are equally likely to die at any time during the age interval 90 to 100 years, then half the original individuals will live beyond age 95 and half die before age 95. In fact, at any time (denoted  $t$ ) during the 10 year interval,  $[100 \times (100 - t)/10]$ -percent will be alive and  $[100 \times (t - 90)/10]$ -percent will have died. For example, at age 97, 30% remain alive (70% have died). In symbols, the formal survival probability function  $S(t)$  is

$$S(t) = P(T \geq t) = \frac{b - t}{b - a} \quad \text{where} \quad a \leq t \leq b.$$

In the present case,  $a = 90$  and  $b = 100$  years. Geometrically, the survival function (a

continuous series of survival probabilities) is a straight line running from 1.0 at age 90 to 0.0 at age 100 with slope of  $-1/(b - a) = -1/10 = -0.1$ . As with all survival functions, the maximum  $S(a)$  is 1.0 at age =  $a$  90 and the minimum  $S(b)$  is 0.0 at age =  $b = 100$ . Figure 3.0 displays this survival function.

**Figure 3.0 Survival function  $S(t)$ : individuals dying uniformly during the age interval 90 to 100 years**



The average mortality rate of these individuals follows directly from the survival probabilities. Suppose  $l$  represents the total number of 90-year olds at risk, then the number of deaths between times  $t_1$  and  $t_2$  is

$$\text{number of deaths} = l \times [S(t_1) - S(t_2)] = \frac{l}{b - a} (t_2 - t_1)$$

and the total person-years these individuals lived is

$$\text{person - years - at - risk} = l \times \text{area} = l \times \{ \text{rectangle} + \text{triangle} \}$$

$$= l \times \left\{ (t_2 - t_1) S(t_2) + \frac{1}{2} (t_2 - t_1) [S(t_1) - S(t_2)] \right\}$$

$$= l \times \left\{ \frac{(t_2 - t_1) [b - \frac{1}{2} (t_1 + t_2)]}{b - a} \right\}.$$

Note that the total person-years of life between two points in time is the number of individuals at risk multiplied by the area under the survival curve (Figure 3.0). For example, for  $l = 1000$  individuals between ages 93 and 97,

$$area = \frac{(97 - 93)[100 - \frac{1}{2}(93 + 97)]}{10} = \frac{4(100 - 95)}{10} = 2.0$$

and the total person-years-at-risk becomes  $l \times area = 1000(2.0) = 2,000$  person-years-at-risk. In other words, the mean years of life lived between the age 93 and 97 is 2.0 years.

The average mortality rate becomes

$$R = \text{average mortality rate} = \frac{\text{number of deaths}}{\text{person - years - at - risk}} = \frac{1}{b - \frac{1}{2}(t_1 + t_2)}.$$

For example, the average mortality rate for individuals between the ages of 93 and 97 is

$$R = \frac{1}{100 - \frac{1}{2}(93 + 97)} = 0.2 \text{ or } 200 \text{ deaths per } 1000 \text{ person-years.}$$

An average mortality rate measures risk over an interval. As the interval becomes smaller, the average mortality rate more accurately reflects the instantaneous hazard rate.

When the length of the interval  $t_2$  to  $t_1$  ultimately becomes 0 or  $t_1 = t_2 = t$  ( $t_1 - t_2 = 0$ ), the two rates become identical. For the uniform mortality case, the average rate (now the hazard rate) is then

$$\text{hazard rate} = h(t) = \frac{1}{b - \frac{1}{2}(t + t)} = \frac{1}{b - t}.$$

This expression is the same when derived by directly applying the definition of a hazard rate where

$$h(t) = -\frac{\frac{d}{dt} S(t)}{S(t)} = \frac{\frac{1}{b-a}}{\frac{b-t}{b-a}} = \frac{1}{b-t}.$$

The geometry of a hazard rate for this random mortality illustration is displayed in Figure 4.0.

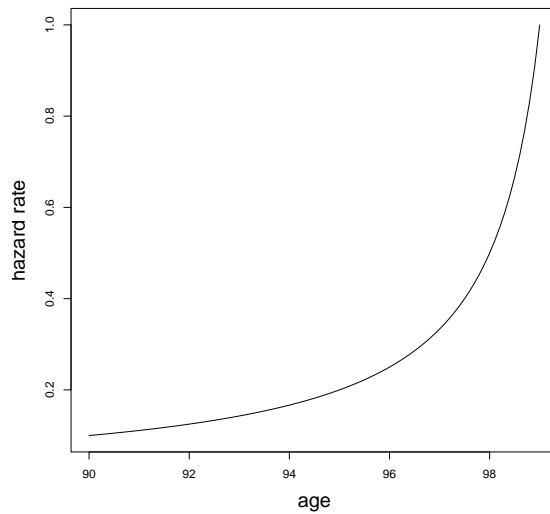
The mean survival time for these 90 year-olds (entire population) is

$\text{mean years lived} = \frac{1}{2}(b - a) = 0.5(100 - 90) = 5$  years. It is the total area under the survival "curve" from ages 90 to 100 (area of the triangle -- Figure 3.0). The entire area under the survival curve is also related to the total person-years of survival (denoted  $L$ ).

In symbols,  $L = l \times \text{mean} = l \times \frac{1}{2}(b - a)$ . In general, the total area under the survival curve is equal to the mean survival time or

$$\text{mean survival time} = \text{total area} = \frac{\text{total - person - years}}{\text{the number of persons - at - risk}} = \frac{L}{l}.$$

**Figure 4.0 Hazard function  $h(t)$ : individuals dying uniformly during the age interval 90 to 100 years**



The mean survival time is calculated in essentially the same way as any mean value. It is the total time lived ( $L$ ) divided by the total number of individuals ( $l$ ) who lived it.

The crude mortality rate (mortality rate for all individuals for the entire time interval -- age 90 to 100 years) is

$$\begin{aligned} \text{crude rate} &= \frac{\text{total number of deaths}}{\text{total - person - years}} = \frac{d}{L} = \frac{l}{L} \\ &= \frac{1}{\text{mean survival time}}. \end{aligned}$$

Higher risk causes less survival time and visa versa. Specifically for the 90-year olds, the mean survival time is five years making the crude mortality rate  $1/5 = 0.2$  deaths per person-year. Notice that the total number of deaths ( $d$ ) equals the total number of individual at risk ( $l$ ) when the entire 10 year interval is considered ( $l = d$  -- everyone dies). A rate is

called crude not because it is lacking or rudimentary. The term "crude rate" applies to a rate that is unadjusted for the influences of other factors.

The median survival time is that age where half the  $l$  individuals have survived and half have died (denoted  $t_m$ ). In symbols, when the survival probability  $S(t_m) = 0.5$ , the median is  $t_m$ . For the case of uniform mortality,

$$S(t_m) = \frac{1}{2} = \frac{b - t_m}{b - a} \quad \text{and} \quad \text{median value} = t_m = b - \frac{1}{2}(b - a) = \frac{1}{2}(b + a).$$

For the age interval 90 to 100 years of age, the median age is  $t_m = \frac{1}{2}(100 + 90) = 95$  years, making the median survival time equal to five years. The mean and median are expected to be equal for this symmetric pattern of survival times.

The probability of death during a specific age interval (again denoted  $q$ ) is the number of individuals who died in the interval divided by the number of individuals who could have died (at-risk individuals). For the random uniform mortality case, this probability is

$$\text{probability of death} = q = \frac{l \times [S(t_1) - S(t_2)]}{l \times S(t_1)} = \frac{\frac{l}{b - a}(t_2 - t_1)}{\frac{l}{b - a}(b - t_1)} = \frac{t_2 - t_1}{b - t_1}.$$

For example, the probability of death during the age interval 93 to 97 is  $q = 4/7 = 0.571$ .

Again, the symbol  $q$  represents the conditional probability of death (conditional on being alive at  $t = 93$ ). Of course, if  $t_2 = b$ , then the probability of death is 1.0.

As with an average rate and a survival probability, the probability of death (not surprisingly) is related to the hazard rate. Specifically, the hazard rate is

$$h(t_1) = \frac{1}{b - t_1} = \frac{q}{t_2 - t_1}.$$

The relationship between a conditional probability of death and a hazard rate, as well as several other relationships illustrated by this simplest case, are useful in understanding more complex survival and hazard functions. Three recurring relationships are:

1. An average rate is approximately equal to a hazard rate over a short interval or

$$R = \text{average rate} \approx \text{hazard rate} = h(t).$$

2. The hazard rate and the conditional probability of death are related or

$$h(t_1) = \frac{1}{b - t_1} = \frac{q}{t_2 - t_1}.$$

for this example and will have approximately the same relationship in many other situa-

3. The mean survival time is geometrically the area under the survival curve or

$$\text{mean survival time} = \frac{L}{l} = \frac{\text{total - person - years}}{\text{total persons - at - risk}}$$

*= area under the survival curve.*

---

## Statistical Tools: properties of confidence intervals

Consider as estimate of a generic parameter  $g$ , denoted  $\hat{g}$ . A normal distribution based 95% confidence interval is

$$P[\hat{g} - 1.960S_{\hat{g}} \leq g \leq \hat{g} + 1.960S_{\hat{g}}] = P[a \leq g \leq b] = 0.95$$

where  $S_{\hat{g}}$  represents the estimated standard error of the (at least approximate) normal distribution of the estimate  $\hat{g}$ .

It is then true that a function of the lower ( $a$ ) and upper ( $b$ ) bounds is also a 95% confidence interval for the same function of the parameter  $g$ , or

$$P[f(a) < f(g) \leq f(b)] = 0.95.$$

For example,

$P[\log(a) \leq \log(g) \leq \log(b)]$  is a 95% confidence interval for the logarithm of  $g$ ,

$P[a^2 \leq g^2 \leq b^2]$  is a 95% confidence interval for the squared value of  $g$  and

$P[\sqrt{a} \leq \sqrt{g} \leq \sqrt{b}]$  is a 95% confidence interval for the square root of  $g$ .

The reverse is also true. A normal based confidence interval constructed for a function of the parameter can be transformed to be a confidence interval for the parameter itself. Specifically, if

$$P[f(\hat{g}) - 1.960S_{f(\hat{g})} \leq f(g) \leq f(\hat{g}) + 1.960S_{f(\hat{g})}] = P[A \leq f(g) \leq B] = 0.95,$$

then an algebraic manipulation of the function  $f$  yields a 95% confidence interval for the parameter  $g$ . For example,

$$P[A \leq \log(g) \leq B] = P[e^A \leq e^{\log(g)} \leq e^B] = P[e^A \leq g \leq e^B] = 0.95,$$

$$P[A \leq g^2 \leq B] = P[\sqrt{A} \leq g \leq \sqrt{B}] = 0.95 \text{ and}$$

$$P[A \leq \sqrt{\log(g)} \leq B] = P[A^2 \leq \log(g) \leq B^2] = P[e^{A^2} \leq g \leq e^{B^2}] = 0.95.$$

In short, a 95% confidence interval for  $g$  can be transformed to be a 95% confidence interval of  $f(g)$  and a 95% confidence interval for  $f(g)$  can be transformed to be a 95% confidence interval for  $g$ .

---