Concepts in survival and demography

This is a summary relations between various quantities used in analysis of follow-up studies. They are ubiquitous in the analysis and reporting of results. Hence it is important to be familiar with all of them and the relation between them.

Probability

Survival function:

$$
\begin{aligned}
S(t) &= \mathrm{P}\{\text{survival at least till } t\} \\
&= \mathrm{P}\{T > t\} = 1 - \mathrm{P}\{T \le t\} = 1 - F(t)
\end{aligned}
$$

Conditional survival function:

$$
\begin{aligned}
S(t|t_{\text{entry}}) &= \mathrm{P}\{\text{survival at least till } t|\text{ alive at } t_{\text{entry}}\} \\
&= S(t)/S(t_{\text{entry}})
\end{aligned}
$$

Cumulative distribution function of death times (cumulative risk):

$$
\begin{aligned}
F(t) &= \mathrm{P}\{\text{death before } t\} \\
&= \mathrm{P}\{T \le t\} = 1 - S(t)
\end{aligned}
$$

Density function of death times:

$$
f(t) = \lim_{h\to 0} \mathrm{P}\{\text{death in } (t, t+h)\}/h = \lim_{h\to 0} \frac{F(t+h)-F(t)}{h} = F'(t)
$$

Intensity:

$$
\begin{aligned}
\lambda(t) &= \lim_{h\to 0} \mathrm{P}\{\text{event in } (t, t+h] \mid \text{alive at } t\}/h \\
&= \lim_{h\to 0} \frac{F(t+h)-F(t)}{S(t)h} = \frac{f(t)}{S(t)} \\
&= \lim_{h\to 0} -\frac{S(t+h)-S(t)}{S(t)h} = -\frac{\mathrm{d}\log S(t)}{\mathrm{d}t}
\end{aligned}
$$

The intensity is also known as the hazard function, hazard rate, rate, mortality/morbidity rate.

Note that f and λ are *scaled* quantities, they have dimension time^{-1}.

Relationships between terms:

$$
\begin{aligned}
-\frac{\mathrm{d}\log S(t)}{\mathrm{d}t} &= \lambda(t) \\
&\Updownarrow \\
S(t) &= \exp\left(-\int_0^t \lambda(u)\,\mathrm{d}u\right) = \exp\bigl(-\Lambda(t)\bigr)
\end{aligned}
$$

The quantity $\Lambda(t) = \int_0^t \lambda(s)\,\mathrm{d}s$ is called the *integrated intensity* or the **cumulative rate**. It is *not* an intensity, it is dimensionless.

$$\lambda(t) = -\frac{\mathrm{d}\log(S(t))}{\mathrm{d}t} = -\frac{S'(t)}{S(t)} = \frac{F'(t)}{1-F(t)} = \frac{f(t)}{S(t)}$$

The cumulative *risk* of an event (to time t) is:

$$F(t) = \mathrm{P}\left\{\text{Event before time } t\right\} = \int_0^t \lambda(u)S(u)\,\mathrm{d}u = 1 - S(t) = 1 - \mathrm{e}^{-\Lambda(t)}$$

For small $|x|$ (< 0.05), we have that $1 - \mathrm{e}^{-x} \approx x$, so for small values of the integrated intensity:

$$\text{Cumulative risk to time } t \approx \Lambda(t) = \text{Cumulative rate}$$

Statistics

Likelihood from one person:

The likelihood from a number of small pieces of follow-up from one individual is a product of conditional probabilities:

$$\begin{aligned} \mathrm{P}\left\{\text{event at } t_4|\text{entry at } t_0\right\} &= \mathrm{P}\left\{\text{event at } t_4|\text{ alive at } t_3\right\} \times \\ &\quad \mathrm{P}\left\{\text{survive } (t_2,t_3)|\text{ alive at } t_2\right\} \times \\ &\quad \mathrm{P}\left\{\text{survive } (t_1,t_2)|\text{ alive at } t_1\right\} \times \\ &\quad \mathrm{P}\left\{\text{survive } (t_0,t_1)|\text{ alive at } t_0\right\} \end{aligned}$$

Each term in this expression corresponds to one *empirical rate*[1] $(d,y) = (\#\text{deaths}, \#\text{risk time})$, i.e. the data obtained from the follow-up of one person in the interval of length y. Each person can contribute many empirical rates, most with $d = 0$; d can only be 1 for the *last* empirical rate for a person.

Log-likelihood for one empirical rate (d,y):

$$\ell(\lambda) = d\log(\lambda) - \lambda y$$

This is under the assumption that the underlying rate (λ) is constant over the interval that the empirical rate refers to.

Log-likelihood for several persons. Adding log-likelihoods from a group of persons (only contributions with identical rates) gives:

$$D\log(\lambda) - \lambda Y,$$

where Y is the total follow-up time, and D is the total number of failures.

Note: The Poisson log-likelihood for an observation D with mean λY is:

$$D\log(\lambda Y) - \lambda Y = D\log(\lambda) + D\log(Y) - \lambda Y$$

[1]This is a concept coined by BxC, and so is not necessarily generally recognized.

The term $D\log(Y)$ does not involve the parameter λ, so the likelihood for an observed rate can be maximized by pretending that the no. of cases D is Poisson with mean λY. But this does *not* imply that D follows a Poisson-distribution. It is entirely a likelihood based computational convenience. Anything that is not likelihood based is not justified.

A linear model for the log-rate, $\log(\lambda) = X\beta$ implies that

$$\lambda Y = \exp\big(\log(\lambda) + \log(Y)\big) = \exp\big(X\beta + \log(Y)\big)$$

Therefore, in order to get a linear model for λ we must require that $\log(Y)$ appear as a variable in the model for $D \sim (\lambda Y)$ with the regression coefficient fixed to 1, a so-called offset-term in the linear predictor.

Competing risks

Competing risks: If there is more than one, say 3, causes of death, occurring with (cause-specific) rates λ_1, λ_2, λ_3, that is:

$$\lambda_c(a) = \lim_{h\to 0} \mathrm{P}\left\{\text{death from cause } c \text{ in } (a, a+h] \mid \text{alive at } a\right\}/h, \quad c = 1,2,3$$

The survival function is then:

$$S(a) = \exp\left(-\int_0^a \lambda_1(u) + \lambda_2(u) + \lambda_3(u)\,\mathrm{d}u\right)$$

because you have to escape all 3 causes of death. The probability of dying from cause 1 before age a (the cause-specific cumulative risk) is:

$$\mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u \neq 1 - \exp\left(-\int_0^a \lambda_1(u)\,\mathrm{d}u\right)$$

The term $\exp(-\int_0^a \lambda_1(u)\,\mathrm{d}u)$ is sometimes referred to as the "cause-specific survival", but it does not have any probabilistic interpretation in the real world. It is the survival under the assumption that only cause 1 existed an that the mortality rate from this cause was the same as when the other causes were present too.

Together with the survival function, the cause-specific cumulative risks represent a classification of the population at any time in those alive and those dead from causes 1, 2 and 3 respectively:

$$1 = S(a) + \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_2(u)S(u)\,\mathrm{d}u + \int_0^a \lambda_3(u)S(u)\,\mathrm{d}u, \quad \forall a$$

Subdistribution hazard Fine and Gray defined models for the so-called subdistribution hazard. Recall the relationship between between the hazard (λ) and the cumulative risk (F):

$$\lambda(a) = -\frac{\mathrm{d}\log\big(S(a)\big)}{\mathrm{d}a} = -\frac{\mathrm{d}\log\big(1 - F(a)\big)}{\mathrm{d}a}$$

When more competing causes of death are present the Fine and Gray idea is to use this transformation to the cause-specific cumulative risk for cause 1, say:

$$\tilde{\lambda}_1(a) = -\frac{\mathrm{d}\log\big(1 - F_1(a)\big)}{\mathrm{d}a}$$

This is what is called the subdistribution hazard, it depends on the survival function S, which depends on *all* the cause-specific hazards:

$$F_1(a) = \mathrm{P}\left\{\text{dead from cause 1 at } a\right\} = \int_0^a \lambda_1(u)S(u)\,\mathrm{d}u$$

The subdistribution hazard is merely a transformation of the cause-specific cumulative risks. Namely the same transformation which in the single-cause case transforms the cumulative risk to the hazard.

Demography

Expected residual lifetime: The expected lifetime (at birth) is simply the variable age (a) integrated with respect to the distribution of age at death:

$$\mathrm{EL} = \int_0^\infty a f(a)\,\mathrm{d}a$$

where f is the density of the distribution of lifetimes.

The relation between the density f and the survival function S is $f(a) = -S'(a)$, and so integration by parts gives:

$$\mathrm{EL} = \int_0^\infty a\big(-S'(a)\big)\,\mathrm{d}a = -\Big[aS(a)\Big]_0^\infty + \int_0^\infty S(a)\,\mathrm{d}a$$

The first of the resulting terms is 0 because $S(a)$ is 0 at the upper limit and a by definition is 0 at the lower limit.

Hence the expected lifetime can be computed as the integral of the survival function.

The expected *residual* lifetime at age a is calculated as the integral of the *conditional* survival function for a person aged a:

$$\mathrm{EL}(a) = \int_a^\infty S(u)/S(a)\,\mathrm{d}u$$

Lifetime lost due to a disease is the difference between the expected residual lifetime for a diseased person and a non-diseased (well) person at the same age. So all that is needed is a(n estimate of the) survival function in each of the two groups.

$$\mathrm{LL}(a) = \int_a^\infty S_{\mathrm{Well}}(u)/S_{\mathrm{Well}}(a) - S_{\mathrm{Diseased}}(u)/S_{\mathrm{Diseased}}(a)\,\mathrm{d}u$$

Note that the definition of the survival function for a non-diseased person requires a decision as to whether one will consider non-diseased persons immune to the disease in question or not. That is whether we will include the possibility of a well person getting ill and subsequently die. This does not show up in the formulae, but is a decision required in order to devise an estimate of S_{Well}.

Lifetime lost by cause of death is using the fact that the difference between the survival probabilities is the same as the difference between the death probabilities. If several causes of death (3, say) are considered then:

$$\begin{aligned}S(a) = 1 &- \mathrm{P}\left\{\text{dead from cause 1 at } a\right\}\\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a\right\}\\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a\right\}\end{aligned}$$

and hence:

$$\begin{aligned}S_{\text{Well}}(a) - S_{\text{Diseased}}(a) = &\ \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Diseased}\right\}\\ &+ \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Diseased}\right\}\\ &+ \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Diseased}\right\}\\ &- \mathrm{P}\left\{\text{dead from cause 1 at } a|\text{Well}\right\}\\ &- \mathrm{P}\left\{\text{dead from cause 2 at } a|\text{Well}\right\}\\ &- \mathrm{P}\left\{\text{dead from cause 3 at } a|\text{Well}\right\}\end{aligned}$$

So we can conveniently define the lifetime lost due to cause 2, say, by:

$$\begin{aligned}\mathrm{LL}_2(a) = \int_a^\infty &\mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\right\}\\ &-\mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Well \& alive at } a\right\}\,\mathrm{d}u\end{aligned}$$

These will have the property that their sum is the years of life lost due to total mortality differences:

$$\mathrm{LL}(a) = \mathrm{LL}_1(a) + \mathrm{LL}_2(a) + \mathrm{LL}_3(a)$$

The terms in the integral are computed as (see the section on competing risks):

$$\mathrm{P}\left\{\text{dead from cause 2 at } u|\text{Diseased \& alive at } a\right\} = \int_a^u \lambda_{2,\text{Dis}}(x) S_{\text{Dis}}(x)/S_{\text{Dis}}(a)\,\mathrm{d}x$$