SURVIVAL AND EVENT HISTORY ANALYSIS I

Period IV (11.3-29.4.2019)
Sangita Kulathinal, HY

Course contents

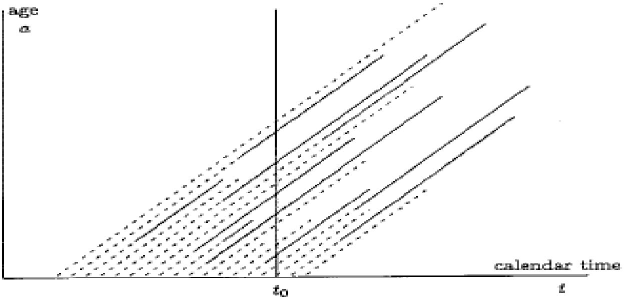
- Part I Characteristics and description of survival data
  - Basic concepts
  - Kaplan-Meier estimator of the survival function
  - Log-rank test for comparing survival functions
  - Nelson- Aalen estimate of the cumulative hazard
- Part II: Parametric survival models
  - Exponential model, Weibull model, more complex models
  - Survival likelihood
  - Incomplete data (right censoring and left truncation)
- Part III: Survival regression models
  - Proportional hazards and accelerated failure time models
  - Parametric: Weibull and exponential regression models
  - Semiparametric: Cox proportional hazards model
  - Partial likelihood
- Part IV: More advanced topics
  - Lexis diagram, more than one time variable
  - Piecewise-constant hazards and Poisson regression
  - Competing risks and event history models
  - Counting process formulation

Part I

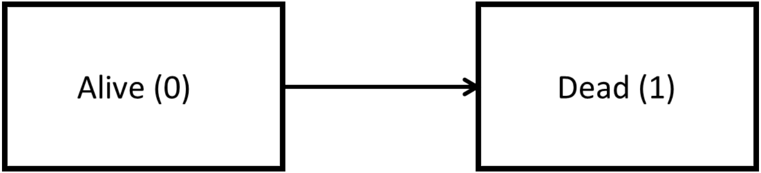
- Characteristics of survival analysis and survival data
- Basic concepts in modelling survival (i.e. time-to-event) data
  - Survival function, hazard function, density function
- Non-parametric estimation of survival functions
  - Kaplan-Meier estimate of the survival function
  - Comparing survival between two groups
  - Nelson-Aalen estimate of the cumulative hazard

Lexis diagram

Lexis diagram of individuals born healthy ($\cdots$), possibly becoming ill ($-$) and dying. A cross-section is taken at time t_0.



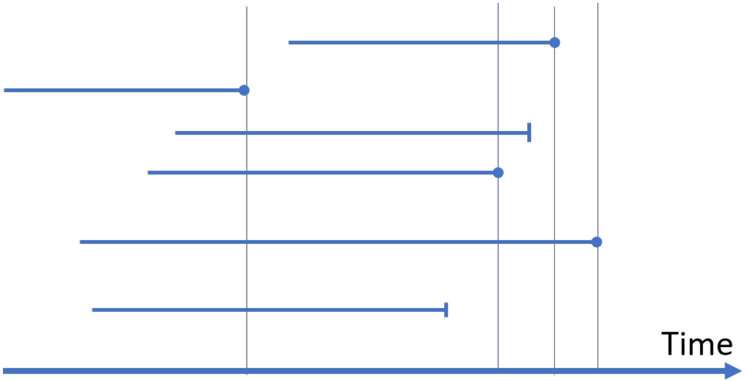
Example 1: Survival model



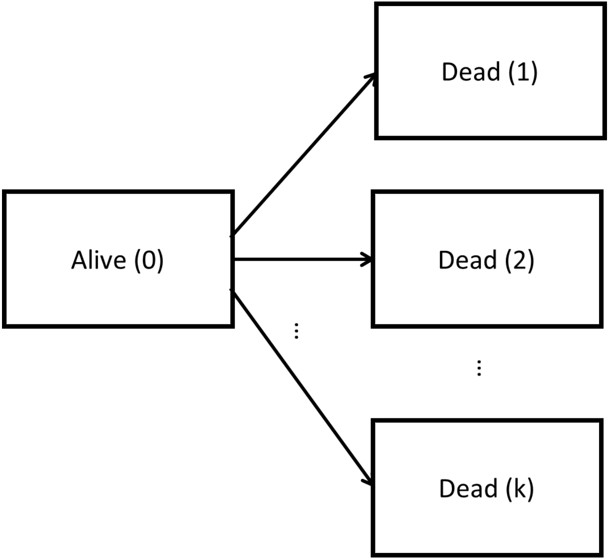
Alive (0)

Dead (1)

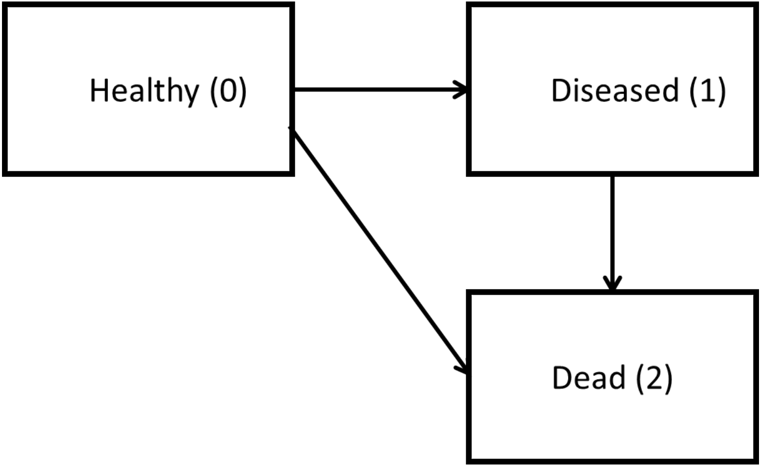
Analysis 1: Risk set



Example 2: Competing risks model



Example 3: Event history model



Survival analysis and related models

- In *survival analysis*, the outcome is the time to event
  - Study of a cohort of individuals, each moving from one state (0) to another state (1)
  - 0 = alive/unemployed/functioning/infected/...
    1 = dead/employed/out of order/uninfected/...
- By contrast, in *time series analysis* or *longitudinal data analysis*, time is used to index measurements (i.e. time itself is *not* the outcome)
- *Event history analysis* generalises survival analysis to situations in which there are more than two states (e.g. healthy/ill/dead) and therefore more than one possible transition between the model states
  - Analysis of life histories
- In survival and event history analysis, exact transition (i.e. event) times between the model states provide the basis for modelling the phenomena
- Applications in medicine, epidemiology, sociology, engineering, economics , ...

Characteristics of survival analysis

- Survival data contain measurements of duration times until the occurrence of a specific event of interest
  - The event is often generically called *failure*, even if no actual 'failure' would be attached to the event
  - Times to the event are also called lifetimes, survival times or durations
  - Depending on the application, we still assume that the time origin and the time scale (calendar time, age, duration since diagnosis,...) is understood from the context, at least for the time being; see the next page
- Technically, survival analysis means statistical modelling of positive random variables (=times), i.e., observations that can assume only positive values
- However, much of the terminology and ways to build and intepret survival models exploit the special nature of time itself and the special patters in which data on durations arise in practice

Setting the clock

- When the outcome is ‘time to event’, one needs to decide when to start the clock
- Different questions and time variables require different choices

Starting point	Time variable
Birth	Age
Any fixed date	Calendar time
First exposure	Time exposed
Entry into the study	Time in study
Disease onset	Time since onset
Start of treatment	Time on treatment

- See e.g. Clayton and Hills, or Bull and Spiegelhalter

Characteristics of survival data (2)

- Very often we do not observe the event of interest in every subject of the study cohort
  - If the subject's actual event time is only known to be larger than a specific time point, the event time is said to be right-censored or, simply, censored
- Apart from right censoring, there are other patterns of incomplete survival data
  - In some studies, subjects are not followed from time 0 (in the study time scale), but only from a later entry time. This leads to delayed entry or left-truncation.
  - There are also situations with left censoring, interval censoring, or even right truncation
  - In particular, right censoring and left truncation will be discussed later

Characteristics of survival data and analysis

In describing the distribution of failure times or durations, special attention is needed because

- failure times are often not normally distributed
- failure times may be censored
  - the individual leaves the study cohort (lives until the end of follow-up, migrates, quits,...)
  - the individual may leave the study cohort for another event than what is being studied (competing risks)
  - it is crucial that censoring does not selectively remove subjects at particular high or low risk of experiencing an event. This is the independent censoring assumption.
- comparison of survival across different study cohorts needs to adjust for possibly different amounts of total follow-up time
- model specification and interpretation are often more convenient in terms of hazard rates, i.e. *conditional* failure rates

Introduction to modelling survival: concepts (1)

- We denote the time to failure by X
  - This is taken to be a non-negative (continuous) random variable; for a discrete-time model, see page 19
- *Survival function* $S(t) = \mathrm{P}(X > t)$, $t \geq 0$, is a non-increasing right-continuous function of time t with $S(0) = 1$ and $\lim_{t \to \infty} S(t) = 0$
  - Cumulative incidence function $F(t) = 1 - S(t)$
- *Density function* $f(t) = \lim_{h \to 0} \frac{1}{h} \mathrm{P}(t \leq X < t + h) = -\dfrac{dS(t)}{dt}$
  - Knowing the density function, the survival function is naturally found by integration:

$$S(t) = \int_t^\infty f(u) du$$

Concepts (2)

- *Hazard function*, (hazard, hazard rate) is the conditional failure rate

$$\begin{aligned}\lambda(t) &= \lim_{h\to 0}\frac{1}{h}\mathrm{P}(t \leq X < t+h \mid X \geq t) \\ &= \lim_{h\to 0}\frac{f(t)h}{hS(t)} = \frac{f(t)}{S(t)} = -\frac{d\log S(t)}{dt} \qquad (1)\end{aligned}$$

- It follows from the definition on the first line that the probability for an individual that has survived up to time t to encounter the event in the next time interval of length h is $\lambda(t)h$
- Cumulative hazard over an interval $[0, t)$ is defined as the hazard integrated over that interval:

$$\Lambda(t) = \int_0^t \lambda(u)du, \;\; t \geq 0$$

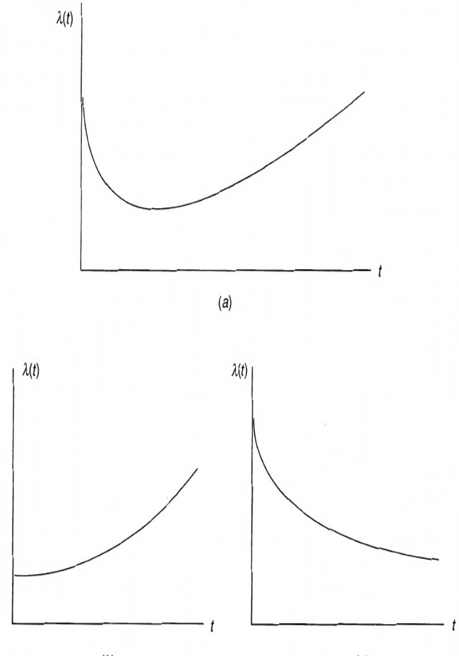
- It follows from eq. (1) that $S(t) = \exp(-\Lambda(t))$ and so $f(t) = -S'(t) = \lambda(t)S(t) = \lambda(t)\exp(-\Lambda(t))$

Concepts (3)

- Remark 1: Any nonnegative function $\lambda(t)$ that satisfies $\int_0^t \lambda(u)du < \infty$ for some $t > 0$ and $\int_0^\infty \lambda(u)du = \infty$ can be used as a hazard function of a continuous random variable

- Remark 2: On the two previous pages, we have shown that given any one of the three alternative defining characteristics of a survival distribution (survival function $S(t)$, density function $f(t)$, or hazard function $\lambda(t)$), the two other functions become fully determined (i.e. can be calculated)

Shapes of hazard function



(a)

Concepts (4)

- Yet another representation of a failure time distribution is the expected residual life at time t

$$r(t) = \mathrm{E}(X - t \mid X \geq t] = \frac{\int_t^\infty (u-t)f(u)du}{S(t)}$$

- This uniquely determines a continuous survival function with finite mean

$$S(t) = \frac{r(0)}{r(t)} \exp\{-\int_0^t \frac{du}{r(u)}\}$$

- See Kalbfleisch and Prentice, p. 7-8

Concepts (5)

- Assume T is a discrete failure time, taking values $a_1 < a_2 \ldots$ with probabilities $f(a_1), f(a_2), \ldots$
- The survival function is $S(t) = 1 - \sum_{j; a_j \leq t} f(a_j),\ t \geq 0$
- The corresponding discrete hazard function λ_i, $i = 1, 2, \ldots$, is defined as follows:

$$\lambda_i = \mathrm{P}(T = a_i \mid T \geq a_i) = \frac{f(a_i)}{1 - \sum_{j=1}^{i-1} f(a_j)}$$

- It can be easily shown that

$$\begin{aligned} S(t) &= \prod_{j; a_j \leq t} (1 - \lambda_j),\ t \geq 0, \quad \text{and} \\ f(a_i) &= \lambda_i \prod_{j=1}^{i-1} (1 - \lambda_j),\ i = 1, 2, \ldots. \end{aligned}$$

Concepts (6)

- Assuming that T has both a continuous component $\lambda_c(t)$, and a discrete component (λ_i), $i = 1, 2, \ldots$ for discrete times $a_1 < a_2 < \ldots$, the three functions are specfied as follows:

$$
\begin{aligned}
S(t) &= \exp\{-\int_0^t \lambda_c(u)du\} \prod_{j;a_j \leq t} (1 - \lambda_j) \\
\Lambda(t) &= \int_0^t \lambda_c(u)du + \sum_{j;a_j \leq t} \lambda_j \\
d\Lambda(t) &= \lambda_i, \quad t = a_i, \\
&= \lambda_c(t)dt, \quad \text{otherwise}
\end{aligned}
$$

Concepts (7)

- In the most general formulation, the survival function can be defined as

  $$S(t) = \mathcal{P}_0^t[1 - d\Lambda(u)]$$

  where the product integral $\mathcal{P}$ is defined by

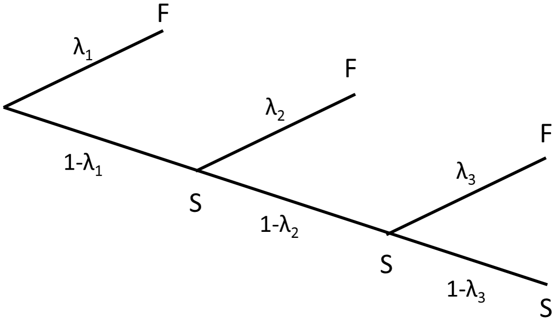
  $$\mathcal{P}_0^t[1 - d\Lambda(u)] = \lim_{r \to \infty} \prod_{k=1}^{r} \{1 - [\Lambda(u_k) - \Lambda(u_{k-1}]\},$$

  where $0 = u_0 < u_1 \ldots < u_r = t$ and $\max(u_i - u_{i-1}) \to 0$ as $r \to \infty$

The product representation can be thought of as describing a coin-tossing experiment in which the probability of heads varies over time. The coin is tossed repeatedly and failure corresponds to the first occurrence of a tail.

Introductory example

- Break the total follow-up period into shorter time intervals (*bands)*
- For example, from a follow-up of one individual over three consecutive bands (next slide), there are four possible observations:
  - failure (F) during the 1st band
  - failure during the 2nd band
  - failure during the 3rd band
  - survival (S) until the end of follow-up


F
λ_1
F
λ_2
F
$1-\lambda_1$
λ_3
S
$1-\lambda_2$
S
$1-\lambda_3$
S

Hazards (conditional failure probabilities)

- The three consecutive Bernoulli trials (coin tossing trials) are described in terms of three hazards, i.e., conditional probabilities of failure:
  - Probability λ_1 of failure during the 1st band
  - Probability λ_2 of failure during the 2nd band, given survival until the end of the 1st band
  - Probability λ_3 of failure during the 3rd band, given survival until the the end of the 2nd band
- N.B. These are conditional probabilities because each is conditioned opon the individual not having failed before the respective band

Probabilities of failure

- The marginal probabilities for failure occurring during each of the three intervals can be expressed in terms of the conditional failure probabilities:

$$\begin{array}{l} \lambda_1 \\ (1\text{-}\lambda_1)\lambda_2 \\ (1\text{-}\lambda_1)(1-\lambda_2)\lambda_3 \end{array}$$

- In addition, the probability to survive the entire follow-up is $(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)$
- N.B. The four marginal probabilities sum up to one as they should: one of the four outcomes must happen in the above model

Survival probabilities

- The probabilities to survive, i.e., to escape failure up to the end of each time band:

$$\begin{array}{l}(1-\lambda_1)\\(1-\lambda_1)(1-\lambda_2)\\(1-\lambda_1)(1-\lambda_2)(1-\lambda_3)\end{array}$$

- Survival probabilities, conditional probabilities, and marginal probabilities of failure are three equivalent ways to describe survival distributions
  - These characterisations of a discrete survival distribution were introduced on page 19

Inference

- Statistical inference questions:
  - How to estimate the survival function and the hazard rate?
  - Are there parametric distributions describing the data?
  - How to use the data collected during at the baseline to estimate their effects on survival? Survival regression models
  - How can survival data be explored and presented?
  - What non-parametric or semi-parametric approaches are there?

Estimating survival based on life tables

- When only grouped failure times are available, censorings can be taken to occur sometime during the band
- Assume that for band i (time interval $[a_{i-1}, a_i[$) the observations are (d_i, l_i, y_i), where

$$\begin{aligned} d_i &= \text{number of failures during time band } i \\ l_i &= \text{number of censorings during time band } i \\ y_i &= \text{the size of the risk set at the beginning} \\ &\qquad \text{of time band } i \text{ (how many subjects still being followed)} \end{aligned}$$

- The effctive size of the risk set at band i is taken to be

$$\begin{aligned} r_i &= y_i - 0.5 l_i \\ &= n - \textstyle\sum_{j=0}^{i-1} d_j - \sum_{j=0}^{i-1} l_j - 0.5 l_i \end{aligned}$$

- We thus assume that a half of censorings took place at the beginning and another half at the end of the interval
- The table presents life times since diagnosis in two cancer treatment groups (next slide)

Example data: two stages of cancer

Year i	Stage I			Stage II		
	y_i	d_i	l_i	y_i	d_i	l_i
1	110	5	5	234	24	3
2	100	7	7	207	27	11
3	86	7	7	169	31	9
4	72	3	8	129	17	7
5	61	0	7	105	6	13
6	54	2	10	85	6	6
7	42	3	6	73	5	6
8	33	0	5	62	3	10
9	28	0	4	49	2	13
10	24	1	8	34	4	6

Life tables: conditional and cumulative survival

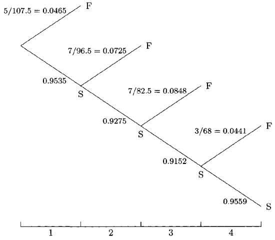
- For each time band i (i.e., year since diagnosis), estimate the conditional probability to survive the band, given survival over the previous band:

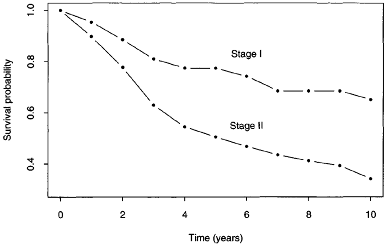
$$\mathrm{P}(T > i | T > i-1) = 1 - d_i / r_i$$

- The cumulative survival probability to survive beoynd band j is estimated as a product of the conditional probabilities of surviving band j as well as each of the preceding bands:

$$\hat{S}_j \equiv \mathrm{P}(T > j) = \prod_{i=1}^{j} (1 - d_i / r_i)$$

Estimated conditional probability for Stage 1





Kaplan-Meier estimate

- The Kaplan-Meier estimate is a non-parametric estimate of the survival function $S(t) = \mathrm{P}(T > t)$
    - Non-parametric means that we do not have to make any assumptions on how the hazard (instantaneous rate) of failure varies over time
- The Kaplan-Meier estimate generalises the life-table method to the case where the event times $t_1 < t_2 < t_3 < \ldots$, including those of censorings, can be assumed to be exact (i.e. not grouped)
    - In particular, we divide the time into a large numer of tiny bands, so called *clicks*, so that each contains at most one (or in practice, only few) events

The risk set

- The risk set at time t is defined as all those study subjects who have not encountered failure or censoring by time t
    - Initially, the size of the risk set if n (number of individuals in the cohort at the study onset)
    - The risk set remains unchanged over all clicks at which there is no failures
    - The risk set is updated at the event times t_i, the censorings are not a problem any more because of the very short duration of the click*:

$$y_i = n - \sum_{j=0}^{i-1} d_j - \sum_{j=0}^{i-1} l_j$$

where

$$d_j = \text{number of failures at time } t_j$$
$$l_j = \text{number of censorings at time } t_j$$

* In practice, there can be several failures and/or censorings at the same time. Censorings are assumed to take place after failures.

Kaplan-Meier estimate cont.

- The conditional survival probabilities are 1 for any click at which there are no failures (i.e. events of interest)
- The conditional survival probabilities for the event times:

$$\mathrm{P}(T > t_i | T > t_{i-1}) = 1 - d_i / y_i$$

- The (cumulative) survival probablity to survive beoynd time t is estimated by the probability of surviving over all previous (observed) event times $t_i < t$:

$$\hat{S}(t) = \mathrm{P}(T > t) = \prod_{i; t_i \leq t} (1 - d_i / y_i), \ t \geq 0$$

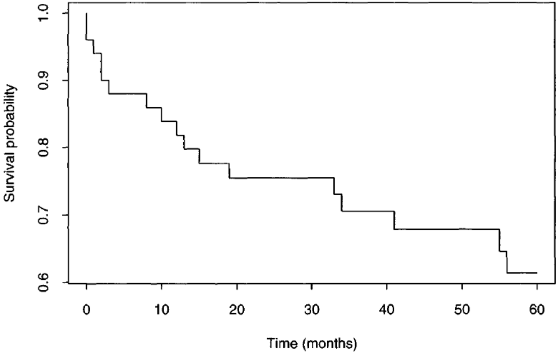
Example

- Data: time until melanoma death since diagnosis in 50 patients (time in months)
- The table shows the first six rows in the data matrix as well as examples of calculating the cumulative survival probabilities for the Kaplan-Meier curve

t_i	y_i	d_i	l_i	d_i/y_i	$1 - d_i/y_i$	$P(T > t_i)$
0	50	2	0	0.0400	0.9600	0.9600
1	48	1	0	0.0208	0.9792	0.9400
2	47	2	0	0.0426	0.9574	0.9000
3	45	1	1	0.0222	0.9778	0.8800
8	43	1	0	0.0233	0.9767	0.8595
10	42	1	0	0.0238	0.9762	0.8391
...	...	...	...	...	...	...

Example cont.

- A Kaplan-Meier estimate of the survival function



Properties of the Kaplan-Meier estimate

- non-parametric maximum likelihood estimator (see e.g., Kalbfleisch and Prentice, p. 14-16)
- piecewise constant: jumps only at the observed times of the events of interest
- non-parametric: does not assume any particular form of the survival function
- in the absence if any censoring, the size of the jump is d_j/n and the curve is $(1 -$ the cumulative distribution function$)$
- the precision of the estimate becomes poorer towards the end of the follow-up
- confidence limits can be derived (next slide)

Confidence limits for Kaplan-Meier estimates

- Taking the logarithm of the K-M estimate

$$\log \hat{S}(t) = \sum_{j;t_j \leq t} \log(1 - d_j/y_j)$$

- Considering observations at each failure as binomial experiments (”drawing failures from the risk set”):

$$\mathrm{Var}(\log \hat{S}(t)) = \sum_{j;t_j \leq t} \frac{d_j}{y_j(y_j - d_j)}$$

- The asymptotic variance of $\hat{S}(t)$ thus is

$$\mathrm{Var}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{j;t_j \leq t} \frac{d_j}{y_j - d_j}$$

- This is the so called Greenwood formula (see e.g., Kalbfeisch and Prentice, p. 17-18)
- The derivation above is only heuristic as it did not take into account that also the risk sets are random

Comparison of two survival functions

- Next we consider the log rank test, a non-parametric test to compare two (or more) survival curves
- The test is based on comparing the expected and observed numbers of failures at the observed failure times the two groups
- In case of two groups, the test is based on a sequence of data summaries, one one for each (separate) failure time t_i:

	Group 0	Group 1	Total
Failures	d_{0i}	d_{1i}	d_i
At risk	y_{0i}	y_{1i}	y_i

Comparison ... cont.

- Given that there were d_i failures at time t_i and assuming (according to the null hypothesis!) that survival is equal in the groups, the *conditional* distribution of D_{1i} is hypergeometric, with the probability function given by

$$\mathrm{P}(D_{1i} = d_{1i} | D_i = d_i) = \frac{\binom{y_{0i}}{d_{0i}} \binom{y_{1i}}{d_{1i}}}{\binom{y_i}{d_i}},$$

- The expected value of D_{1i} is $E_{1i} = d_i(y_{1i}/y_i)$ and the variance

$$v_i = d_i \frac{y_{1i}}{y_i} \frac{y_{0i}}{y_i} \frac{y_i - d_i}{y_i - 1}$$

- The random variable $(D_{1i} - E_{1i})^2 / v_i$ therefore has approximately a χ^2 distribution

Log rank test

- It follows from above that the different of the total numbers of observed and expected failures in group 1, $D - E := \sum_i (D_{1i} - E_{1i})$ is a random variable with mean 0
- The terms in the above sum appear to be dependent. However, it can be shown that the variance of $D - E$ is in fact $\sum_i v_i$
- It thus holds asumptotically that $(D - E)^2 / \sum_i v_i \sim \chi^2$
- If there are p groups to be compared, the multivariate generalisation of the above argument holds (see e.g. Kalbfleisch and Prentice, p. 21-22)

Alternative formulation

- The log rank test is often presented in the following form, comparing the observed and expected numbers under a Poisson assumption

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

- This is based on the idea that O_i is a Poisson random variable with mean and variance E_i (note, however, that the two terms are not independent as $O_1 + O_2 = n$)
- This means that the actual test is based on the hypergeometric distribution as explained above

Nelson–Aalen estimate

- Recall that the cumulative hazard is defined as $\Lambda(t) = \int_0^t \lambda(u) du$
- An estimate can be calculated as follows
  - $\Lambda(t)$ jumps upwards at failure times t_j
  - the size of the jump can be estimated as

$$\hat{\lambda}^{(j)} h = \left(\frac{d_j}{y_j h} \right) h = d_j / y_j$$

where y_j is the size of the risk set and d_j the number of failures at t_j. We obtain

$$\hat{\Lambda}(t) = \sum_{j; t_j \leq t} \frac{d_j}{y_j}$$

Nelson-Aalen estimate (2)

There is a close relation between the Kaplan-Meier and Nelson–Aalen estimates: when $d_j/y_j \simeq 0$,

$$\exp(-\overbrace{\sum_{j;t_j\le t} d_j/y_j}^{Nelson-Aalen})$$

$$= \prod_{j;t_j\le t} \exp(-d_j/y_j) \simeq \overbrace{\prod_{j;t_j\le t}(1-d_j/y_j)}^{Kaplan-Meier}$$

- Note the close connection between the above relation between the (non-parametric) estimates of survival function and cumulative hazard and the general definition of survival function on page 18

References

- Andersen, Abildstrom, Rosthøj: Competing risks as a multi-state model. Statistical Methods in MedicalResearch 2002; 11: 203-215.
- Andersen, Borgan, Gill, Keiding Statistical Models Based on Counting Processes. Springer-Verlag.
- Andersen and Keiding: Multi-state models for event history analysis, Stat Methods in Med Research (2002); 11:91-115
- Bull and Spiegelhalter: Tutorial in biostatistics - survival analysis in observational studies, Statistics in Medicine (1997); 16: 1041-1074
- Clayton and Hills: Statistical Models in Epidemiology, Oxford University Press
- Collett: Modelling Survival Data in Medical Research, Chapman and Hall
- Kalbfleisch and Prentice: The Statistical Analysis of Failure Time Data, Wiley
- Keiding: Event history analysis. Annual Review of Statistics and its Applications 2014; 1: 333-360.