Lecture 1 Basic Concepts

Study "time to event" T, where event

- age @ death
- age @ recurrence
- cure (time to cure after treatment)
- time to pregnancy after fertility treatment
- time to equipment failure

⋮

Characteristics of 'survival data'

- $T \geq 0$
- often 'incomplete' data due to censoring

Note – not the same as 'missing data' (bad measurement,

– here 'missingness' provides partial information about the outcome

Typical data, 2 treatment groups
- control
- treatment (receive screening for breast cancer)

Follow for 18 years

During this time, some women die from breast cancer
Some are still alive after 18 years

So... for women who died from BC we know the exact time to event. For women still alive we only know $T > 18$ years.

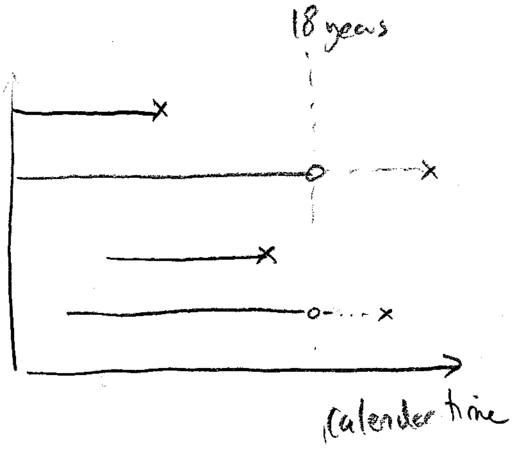
Possible complications

- Censoring during study due to other (random) causes
  - loss to follow-up (perhaps moved)

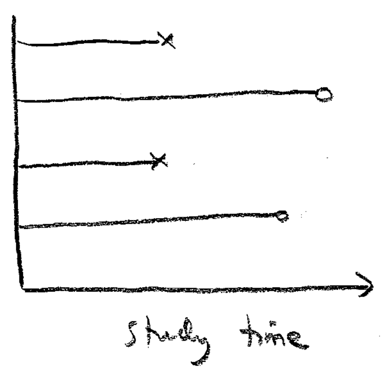
- Censoring due to death due to other causes
  - car accident – unrelated
  - heart disease – perhaps related ⇒ competing risks

- different length of study for different women
  → perhaps recruitment to study took ~ 5 years.
  Then some women have been followed for only 13 years, some for more.

Important to define start and end of study properly.



⇒



(more later)

Goals

1. Estimate & interpret survival characteristics
2. Compare survival between groups
3. Asses relationship between survival and various predictors.

Assumptions – Independent observations
– Independent censoring

Why a separate course?

Typical summary statistics, like mean, are biased when data are censored.

$\Rightarrow$ General approach : estimate distribution of T from data and base estimates of the mean etc on this

Describing T

- $S(t) = P(T > t)$ – survival function
  $= 1 - F(t) = 1 - P(T \leq t)$
  – $S(0) = 1$, $S(\infty) = 0$ (can adjust this to $S(\infty) > 0$ if there is a cure prob.)

discuss

For continuous T $\Rightarrow$ Definition $F(t) = \int_0^t f(u)\,du$

$$\Rightarrow S(t) = \int_t^\infty f(u)\,du \Rightarrow f(t) = -\frac{dS(t)}{dt}$$

– slope of survival fcn.

Sample data $\Rightarrow$ $F_n(t) = \sum_{i=1}^{n} \frac{1}{n} 1\{t_i \le t\}$

$$S_n(t) = \sum_{i=1}^{n} \frac{1}{n} 1\{t_i > t\}$$

= % (prop) of sample survival times that exceeds t.

$\longrightarrow S(t)$ as $n \to \infty$

Other quantities of interest

- mean survival time $E(T) = \int_0^\infty t f(t)\,dt$

$$= -\int_0^\infty t\,dS(t)\,dt$$

$$\overset{\text{intgr}}{\underset{\text{by parts}}{=}} \int_0^\infty S(t)\,dt$$

- median survival time $t_M : S(t_M) = \frac{1}{2}$

- mean residual life time (mrl)

$$mrl(t_0) = E[T - t_0 \mid T \geq t_0]$$

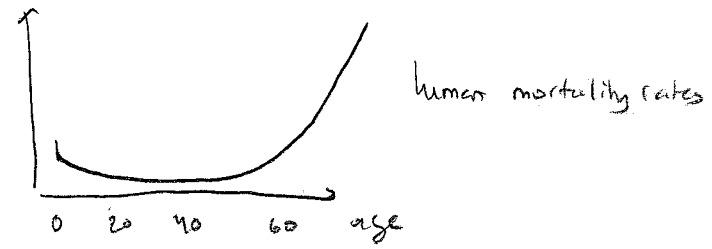
$$= \frac{\int_{t_0}^{\infty} (u-t_0) f(u)\, du}{S(t_0)} = - \frac{\int_{t_0}^{\infty} (u-t_0)\, dS(u)}{S(t_0)}$$

$$= \frac{\int_{t_0}^{\infty} S(u)\, du}{S(t_0)}$$

Hazard rate

discrete version — mortality rate

$$m(t) = P[t \leq T < t+1 \mid T \geq t]$$



hazard rate $h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t+\Delta t \mid T \geq t]}{\Delta t}$

so for small Δt $P[t \leq T < t+\Delta t \mid T \geq t] \approx h(t) \cdot \Delta t$

Note $h(t) = \frac{\lim_{\Delta t \to 0} \frac{P[t \leq T < t+\Delta t]}{\Delta t}}{P(T \geq t)} = \frac{f(t)}{S(t)} = - \frac{S'(t)}{S(t)}$

$$= - \frac{d \log S(t)}{dt}$$

Cumulative hazard, $H(t)$

hazard rate $h(t) = -\dfrac{d \log S(t)}{dt}$

$$\Rightarrow \quad H(t) = \int_0^t h(u)\,du = -\log\{S(t)\}$$

$$\Rightarrow \quad S(t) = e^{-H(t)} = e^{-\int_0^t h(u)\,du}$$

[Discrete case:

Definition of hazard rate

$$h(t) = \frac{p(t)}{S(t^-)} = P(T = t \mid T \geq t)$$

Note
- We can write

$$\frac{S(t_j)}{S(t_{j-1})} = \frac{P(T > t_j)}{P(T > t_{j-1})} = \frac{P(T > t_j, T > t_{j-1})}{P(T > t_{j-1})}$$

$$= P(T > t_j \mid T > t_{j-1})$$

- Follows that

$$S(t) = \prod_{t_j \leq t} \frac{S(t_j)}{S(t_{j-1})} = \prod_{t_j \leq t} S(t_j \mid T > t_{j-1})$$

$$= \prod_{t_j \leq t} \{1 - h(t_j)\}$$

$$\left(\text{cmp } S(t) = e^{-\int_0^t h(u)\,du}\right)$$

Cumulative hazard $H(t) = \sum_{t_j \leq t} h(t_j)$ (KM)

$\approx$ for small h

or alt

$\sum_{t_j \leq t} \ln\{1 - h(t_j)\}$ (CO)

$\Rightarrow S(t) = e^{-H(t)}$ then here as well

Remarks: • $h(t) \overset{1-1}{\longleftrightarrow} S(t)$

• hazard rate $\neq$ a probability – more like a density

Some examples

• Constant hazard $\quad h(t) = \lambda$

$$\Rightarrow S(t) = e^{-\int_0^t \lambda \, du} = e^{-\lambda t}$$

$$f(t) = \frac{-dS(t)}{dt} = \lambda e^{-\lambda t}$$

The exponential distribution!

For Exp. dist $\Rightarrow E(T) = \frac{1}{\lambda}$

$$mrl(t_0) = \frac{\int_{t_0}^{\infty} e^{-\lambda t} \, dt}{e^{-\lambda t_0}} = \frac{1}{\lambda} = E(T)$$

$$\Rightarrow P(T > t+z \mid T \geq t) = \frac{P(T > t+z)}{P(T \geq t)} = \frac{e^{-\lambda(t+z)}}{e^{-\lambda t}} = e^{-\lambda z}$$

$$= P(T > z)$$

Memoryless

In practise, this distribution is too restricted most of the time.

Reality check; plot t vs $\log \hat{S}(t)$ $\rightarrow$ should be a straight line

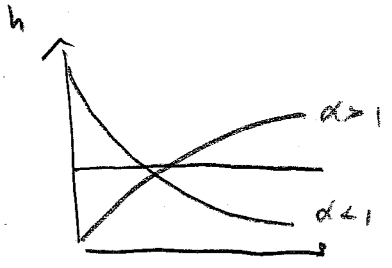
Weibull

$h(t) = \alpha \lambda t^{\alpha - 1}$, $S(t) = e^{-\lambda t^{\alpha}}$

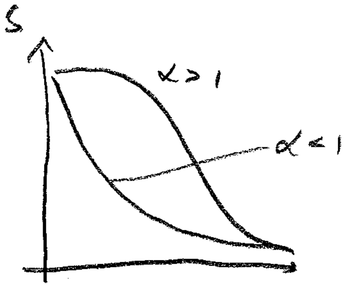
$\alpha = 1 \Rightarrow$ reverts back to exponential

$\alpha > 1 \Rightarrow h(t)$ increasing $\Rightarrow$ 'ageing'

$\alpha < 1 \Rightarrow h(t)$ decreasing (less common)



$\Rightarrow$



Note, i) Weibull true $\rightarrow \log\{-\log S(t)\} = \log \lambda + \alpha \log t$

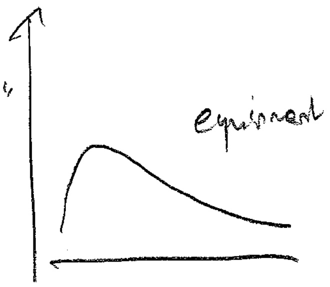
Linearity check ; plot $\ln t$ vs $\ln(-\ln \hat{S}(t))$

Other parametric forms

log-normal $S(t) = 1 - \Phi\left[\frac{\ln t - \mu}{\sigma}\right]$

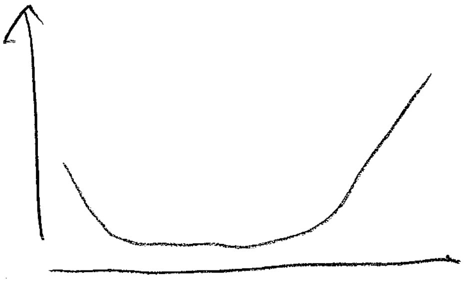
log-logistic $S(t) = \frac{e^{(y - \mu/\sigma)}}{\sigma(1 + e^{(y-\mu/\sigma)})}$, $y = \ln t$

"hump"



Other commonly observed hazards

$\Rightarrow$



bathtub hazard ~ mortality of population

Looking ahead a little — regression models

(1) $T \geq 0$ so standard regression is inappropriate

$\Rightarrow$ log transformation (accelerated failure time model)

$(*)$ $Y = \ln T = \underbrace{\mu + X\beta}_{\substack{\text{model form} \\ \text{(parametric)}}} + \varepsilon$

if specified $\rightarrow$ parametric

if not $\rightarrow$ semi-parametric

Note; $P(T > t \mid X) = ?$ quantity of interest

- $P(T > t \mid X = 0) = S_0(t) = P(e^{(\mu+\varepsilon)} > t)$

- $P(T > t \mid X) = P(Y > \ln t \mid X) = P(Y - X\beta > \ln t - X\beta \mid X)$

$= P(e^{Y - X\beta} > t e^{-X\beta} \mid X) = P(e^{(\mu+\varepsilon)} > t e^{-X\beta} \mid X)$

$= S_0\{t e^{-X\beta}\}$

Same survival function, but time-scale has been compressed / elongated (dep on sign of β)

If $\beta > 0$, $e^{-X\beta} < 1$, $t e^{-X\beta} < t$ life has been 'lengthened'

If $\beta < 0$, time has been 'shortened' by covariate

Estimation of $(*)$ models can be a bit tricky, so the more popular regression model for survival is $\Rightarrow$

(2) Proportional hazard

$h(t|X) = h_0(t)\, c(X\beta)$ where $c(X\beta)$ s.t. h positive

Usually $c(X\beta) = e^{X\beta}$

Note $\dfrac{h(t|X_1)}{h(t|X_2)} = \exp\left((X_1 - X_2)\beta\right)$ <u>not</u> a function of t

Example $X = \begin{cases} 0 & \text{control} \\ 1 & \text{treatment} \end{cases} \Rightarrow$ $\boxed{\text{relative risk RR}}$

$= \dfrac{h_0(t)\, e^{\beta}}{h_0(t)} = e^{\beta}$

<u>constant</u> across time

If $\beta > 0 \Rightarrow$ RR of group $X=1$ > 1

If $\beta < 0 \Rightarrow$ RR of group $X=1$ < 1

(treatment effective)

- Estimation (as we shall see) — can estimate β's w/o knowing $h_0(t)$ $\Rightarrow$ then use $\hat{\beta}$ and $\hat{S}(t)$ to estimate $h_0(t)$.

- $S(t|X) = \exp\left(-\int_0^t h_0(u)\, e^{X\beta}\, du\right) = \exp\left(-e^{X\beta} \int_0^t h_0(u)\, du\right) = \exp\left(-e^{X\beta} H_0(t)\right)$

$= \exp\left(-e^{X\beta} \ln S_0(t)\right) = \exp\left(-\ln S_0(t)^{e^{X\beta}}\right) = S_0(t)^{e^{X\beta}}$

- Proportional hazard = assumption $\Rightarrow$ Need to check!
(more in diagnostics later)