Announcement

- TA office hours
  - Tuesdays 1-2pm **(657 classroom)**
  - Wednesdays 3-4pm **(657 conference room)**
  - Thursdays 10:30-11:30am **(627)**

- 17 groups are formed

- 5% attendance in evaluation is removed

- Lecture recording will be offered

Syllabus

1. Introduction
   - Survival data
   - Censoring mechanism
   - Application in medical field
2. **Concepts and definitions**
   - **Survival function**
   - **Hazard function**
3. Non-parametric approach
   - Life table
   - Kaplan-Meier survival estimate
   - Hazard function
   - Median and percentile survival time
4. Hypothesis testing
   - Overview – hypothesis, test statistics, p-values
   - Log-rank
   - Wilcoxon
   - Gehan test
5. Study design and sample size estimation
   - Overview
   - Survival sample size estimation
   - Accrual time and Study duration
6. Semiparametric model – proportional hazard model
   - Partial likelihood
   - Inference
   - Time varying covariates
   - Stratification
7. Model checking in the PH model
   - Model checking
   - Residuals
8. Parametric model
   - Parametric proportional hazard model
   - Accelerate failure model
9. Other topics
   - Competing risk
   - Recurrent events
   - Non-proportional hazard ratio
   - Interval censoring

Recap

- Previous lecture
  - Survival data
  - Censoring
  - Impact of survival data in medical research

- Variables in survival data sets
  - Subject ID
  - Treatment
  - Time (start, end)
  - Censoring status
  - Covariates

Time of Remission (Weeks) of Leukaemia Patients

Subject ID	Time	Censor	Treatment
1	6	1	New
2	6	0	New
3	6	0	New
4	6	0	New
5	7	0	New
6	9	1	New
7	10	1	New
8	10	0	New
9	11	1	New
10	13	0	New
11	16	0	New
12	17	1	New
13	19	1	New
14	20	1	New
15	22	0	New
16	23	0	New
17	25	1	New
18	32	1	New
19	32	1	New
20	34	1	New
21	35	1	New

22	1	0	Control
23	1	0	Control
24	2	0	Control
25	2	0	Control
26	3	0	Control
27	4	0	Control
28	4	0	Control
29	5	0	Control
30	5	0	Control
31	8	0	Control
32	8	0	Control
33	8	0	Control
34	8	0	Control
35	11	0	Control
36	11	0	Control
37	12	0	Control
38	12	0	Control
39	15	0	Control
40	17	0	Control
41	22	0	Control
42	23	0	Control

Censor=0: events
Censor=1: censor

Topics

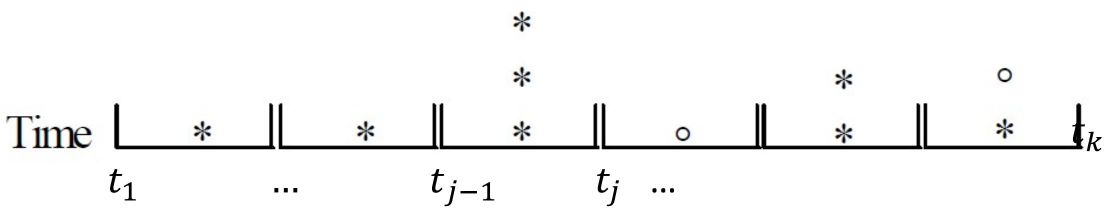
2. Concepts and definitions relate to survival function

    - Density function $f(t)$
    - Survival function $S(t)$
    - Hazard function $h(t)$
    - Cumulative hazard function $H(t)$

- Will see that
    - Hazard function determine the distribution of survival data

Discrete Survival Time

- T right censored survival time and takes discrete values $t_1 < t_2 < \cdots < t_k$
- Assuming noninformative censoring



- The probability distribution function at t_j

$$f\left(t_j\right) = P\left(T = t_j\right) \quad j = 1, 2, \ldots, \mathrm{k}$$

$$P(T \geq t_1) = 1$$

$$P\left(T \geq t_j\right) = P\left(T > t_{j-1}\right)$$

Discrete Survival Time

The Survival function

$$S\left(t_j\right) = P(T > t_j) = \sum_{i>j} f(t_i)$$

Hazard Function

- Conditional probability of failure at t_i given that the individual has survived to t_i

$$h(t_i) = P(T = t_i | T \geq t_i)$$

- Represent the risk at interval t_i, a ratio of
  - Number of events observed at t_i
  - Number of subject at risk right before t_i, t_i^-
  - The length of the time interval
- An important quantity in survival analysis, also known as
  - Hazard rate, risk rate, conditional failure rate, intensity function

Hazard Rate

- The unit of hazard is
  - Number of events per person-time
    - Patient-year
    - Patient-month
- Examples : 0.2 events per 100 patient-month
    - 2.4 events per 100 patient-year
    - 0.05 events per 100 patient-week

Hazard Function and the Survival Distribution

For discrete survival time

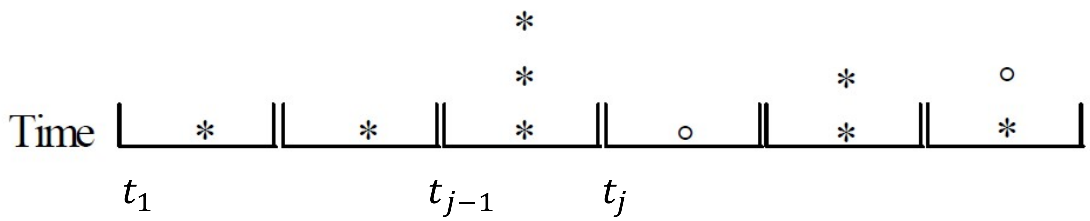
$$h(t_j) = P\left(T = t_j \middle| T \geq t_j\right)$$

$$= \frac{P(T = t_j \cap T \geq t_j)}{P(T \geq t_j)}$$

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

$$= \frac{f(t_j)}{S(t_{j-1})}$$

Discrete Survival Function (2)



$P(T \geq t_1) = 1$ $\qquad$ $P(T \geq t_j) = P(T > t_{j-1})$

$$\begin{aligned} S(t_j) &= P(T > t_j) \\ &= P(T > t_j | T \geq t_j) P(T \geq t_j) \\ &= P(T > t_j | T \geq t_j) P(T > t_{j-1}) \end{aligned}$$

$$S(t_j) = P(T > t_j | T \geq t_j) \times P(T > t_{j-1} | T \geq t_{j-1}) \times \cdots \times P(T > t_1 | T \geq t_1)$$

Discrete Survival Time

$$S(t_j) = P\left(T > \ t_j \middle| T \geq \ t_j\right) \times P\left(T > \ t_{j-1} \middle| T \geq \ t_{j-1}\right) \times \cdots \times P(T > \ t_1 | T \geq \ t_1)$$

- Since

$$h_i = P(T = t_i | T \geq t_i) = P(T \geq t_i | T \geq t_i) - P(T > t_i | T \geq t_i)$$

$$P(T > t_i | T \geq t_i) = 1 - h_i$$

$$S(t_j) = \prod_{i: t_i \leq t_j} (1 - h_i)$$

Continuous Survival Time

- Again, consider right censored
- Censoring is non-informative
- The density function

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

- The survival function

$$\begin{aligned} S(t) &= P(T > t) \\ &= 1 - P(T \leq t) \\ &= \int_t^{\infty} f(x)dx \end{aligned}$$

- Therefore

$$f(x) = -\frac{dS(t)}{dt}$$

Continuous Survival Time

- Hazard definition for continuous survival time

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

- $\lim\limits_{\Delta t \to 0}$
  - indicates instantaneous risk after t
- $P(t \le T < t + \Delta t | T \ge t)$
  - survival probability in the interval of $(t, t + \Delta t)$ given that the person has survived up to t
- $h(t)$ >0

Hazard Function and Survival Function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, T \geq t)}{\Delta t P(T \geq t)}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, T \geq t)}{\Delta t} \cdot \frac{1}{P(T \geq t)}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \cdot \frac{1}{P(T \geq t)}$$

$$= \frac{f(t)}{S(t^-)}$$ where $S(t^-) = \lim_{t \to t^-} S(t^-)$

Hazard Function and Survival Function

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t, T \geq t)}{\Delta t P(T \geq t)}$$

$$= \frac{f(t)}{S(t)}$$

$$= -\frac{dS(t)}{dt} \frac{1}{S(t)}$$

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

- Remember

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

$$\frac{d}{dx} \log x = \frac{1}{x}$$

Cumulative Hazard Function

$$H(t) = \int_0^t h(x)dx$$

$$= \int_0^t -\frac{dlogS(x)}{dx}dx$$

$$= -logS(t)$$

$$S(t) = e^{-H(t)}$$

Hazard Function with Dependent Censoring

- Observed survival data, $(T_i\,,\Delta_i)$, $i = 1,2,\ldots,n$
  - $T_i = \min(X_i\,, C_i)$
  - Δ_i - I($X_i \;\le C_i$) – event indicator
  - X_i - event time

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T_i < t + \Delta t | T_i \ge t, C_i \ge t)}{\Delta t}$$

$$= \lim_{\Delta t \to 0} \frac{P(t \le X_i < t + \Delta t, X_i \ge t, C_i \ge t)}{\Delta t P(X_i \ge t, C_i \ge t)}$$

Mean Survival

- The expected survival (mean survival)

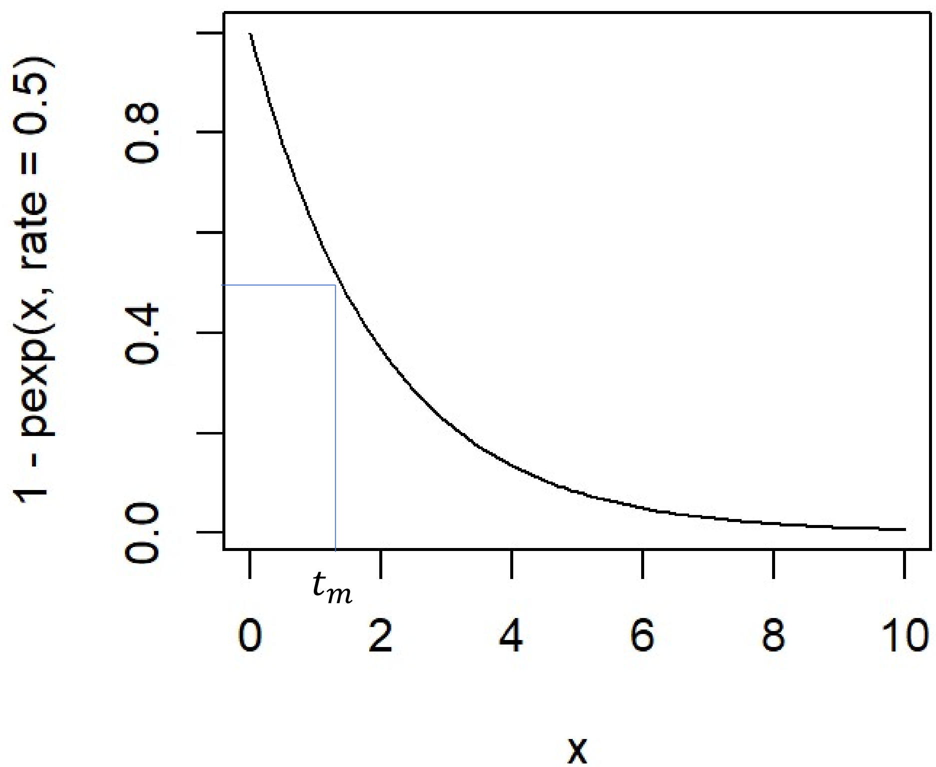
$$E(T) = \int_0^{\infty} u f(u) du$$

$$= \int_0^{\infty} S(u) du$$

Survival Quantiles

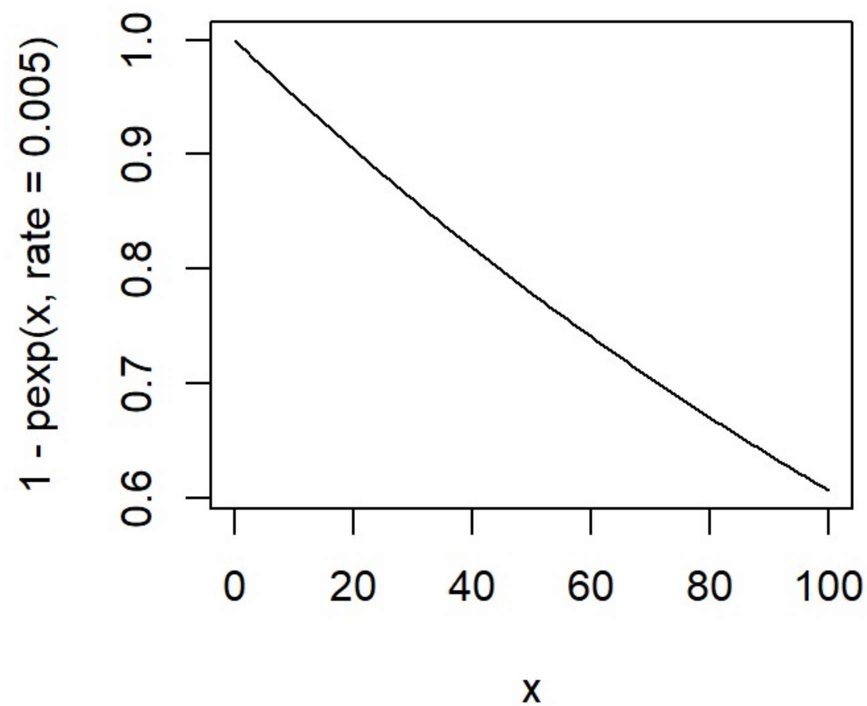
- Median Survival time t_m

- $S(t_m) = P(T > t_m)$=0.5
- $t_m = \inf\{\mathrm{t}: \mathrm{S(t)} \leq 0.5\}$



Survival Quantiles

- Sometimes, the survival curve does not reach to median
- The p^{th} quantile
- $S(t_p) = P(T > t_p) = p$
- $t_p = \inf\{\mathrm{t}: \mathrm{S(t)} \leq p\}$
- Example
  - The survival probability at 102 days is 60%.



R code

```
curve(1-pexp(x,rate=0.005),from=0, to=100)
```

Distributions of Survival Data

- Exponential
- Piecewise exponential
- Weibull
- Gamma
- Log-logistic
- Lognormal

Exponential Distribution

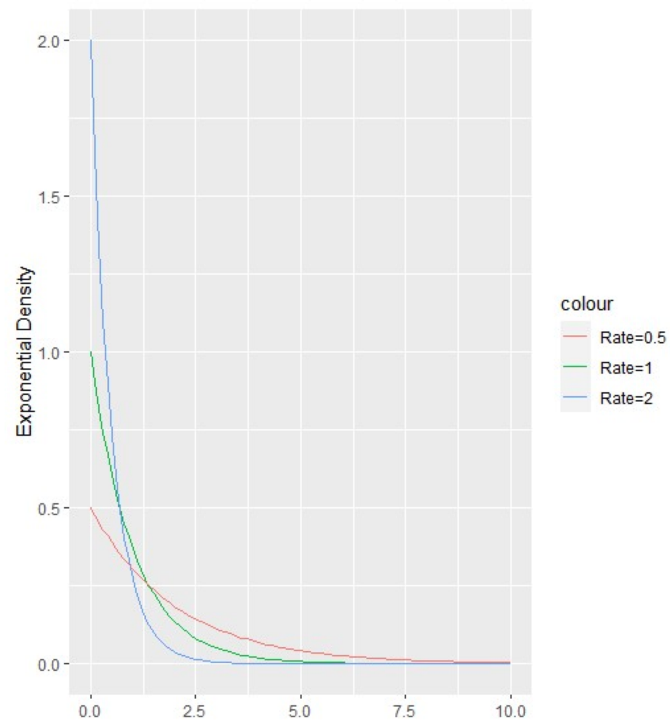
- The most used
  - Owning to its simplicity - constant hazard rate
  - In real world,
    - The hazard rate may change
      - Overtime
      - Depending composition of enrolled patient population
    - The constant hazard rate can represent a weighted average hazard
  - A necessary assumption in study design
- Easy to interpret
  - Help communicate with clinicians
- Foundation of advanced analysis methods

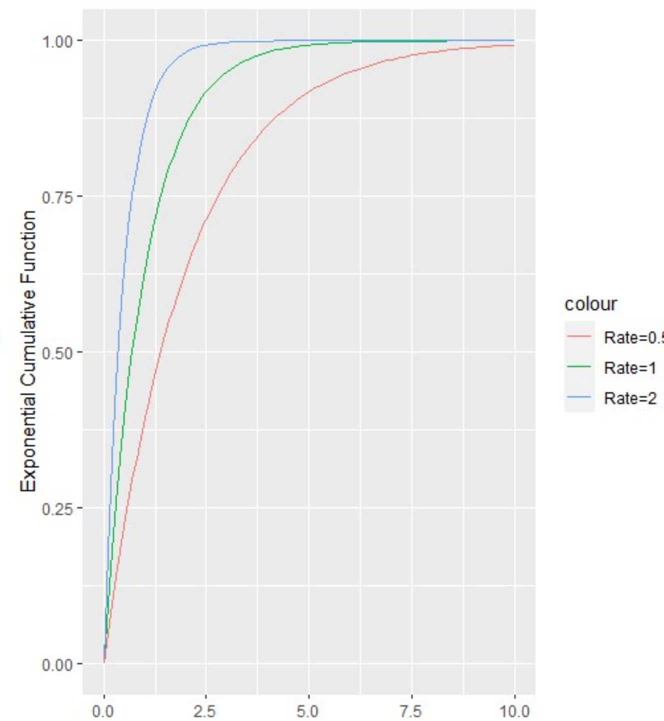
Exponential

- $T \sim Exp(\lambda)$

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

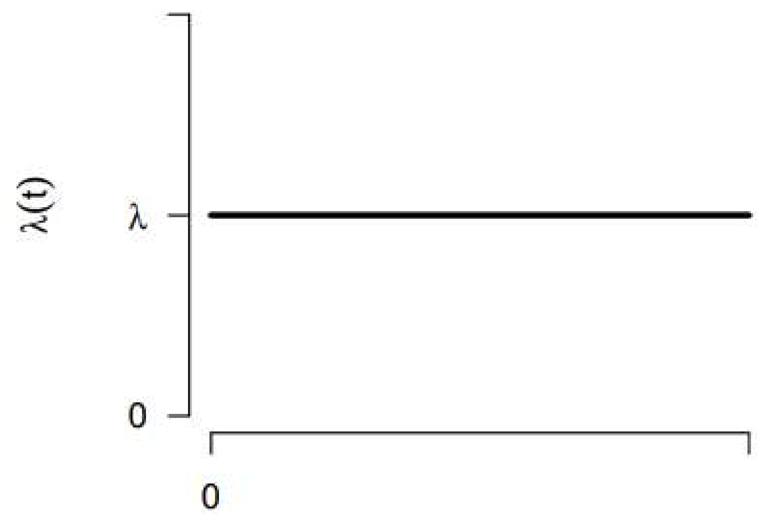
- $S(t) = P(T > t)$
  $= \int_0^t \lambda e^{-\lambda x} dx$
  $= e^{-\lambda t}$
- $E(T) = \frac{1}{\lambda}$
- $Var(T) = 1/\lambda^2$





Constant Hazard Rate

- Hazard function
$$\lambda(t) = \lambda$$
- Cumulative hazard
$$\Lambda(t) = \lambda \mathrm{t}$$
- Memoryless property
$$P(T > t) = P(T > t + s | T > s)$$



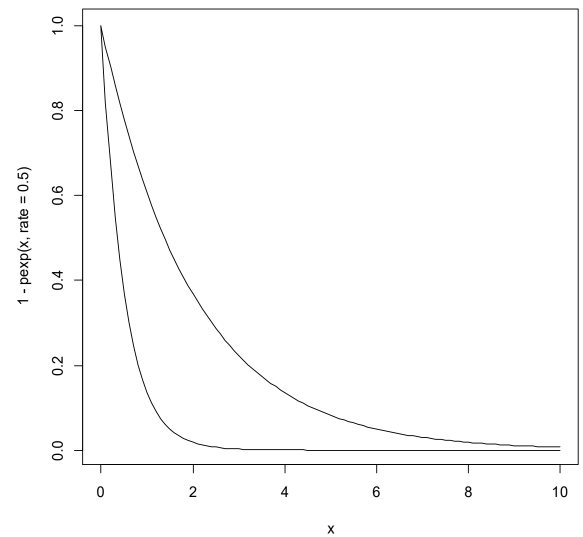
Exponential Distribution

- λ is the rate of event
  - $\lambda_1 = 0.5$
    - 50 deaths per 100 patient-year
    - 4.17 deaths per 100 patient-month
  - $\lambda_2 = 2$

- Hazard ratio

$$\frac{\lambda_1}{\lambda_2} = 0.25$$

A 75% risk reduction



R code

```
curve(1-pexp(x,rate=0.5),from=0, to=10)
curve(1-pexp(x,rate=2),from=0, to=10,
add=TRUE)
```

Exponential Distributions - Survival Quantiles

Examples

- Median survival time λ=0.5 per person-day

$$S(t_m) = 0.5 = e^{-\lambda t_m}$$

$$t_m = \ln 2/\ \lambda = \ln 2/0.5$$

The median survival time is 1.38 days

- 60% survival time λ=0.005 per person-day

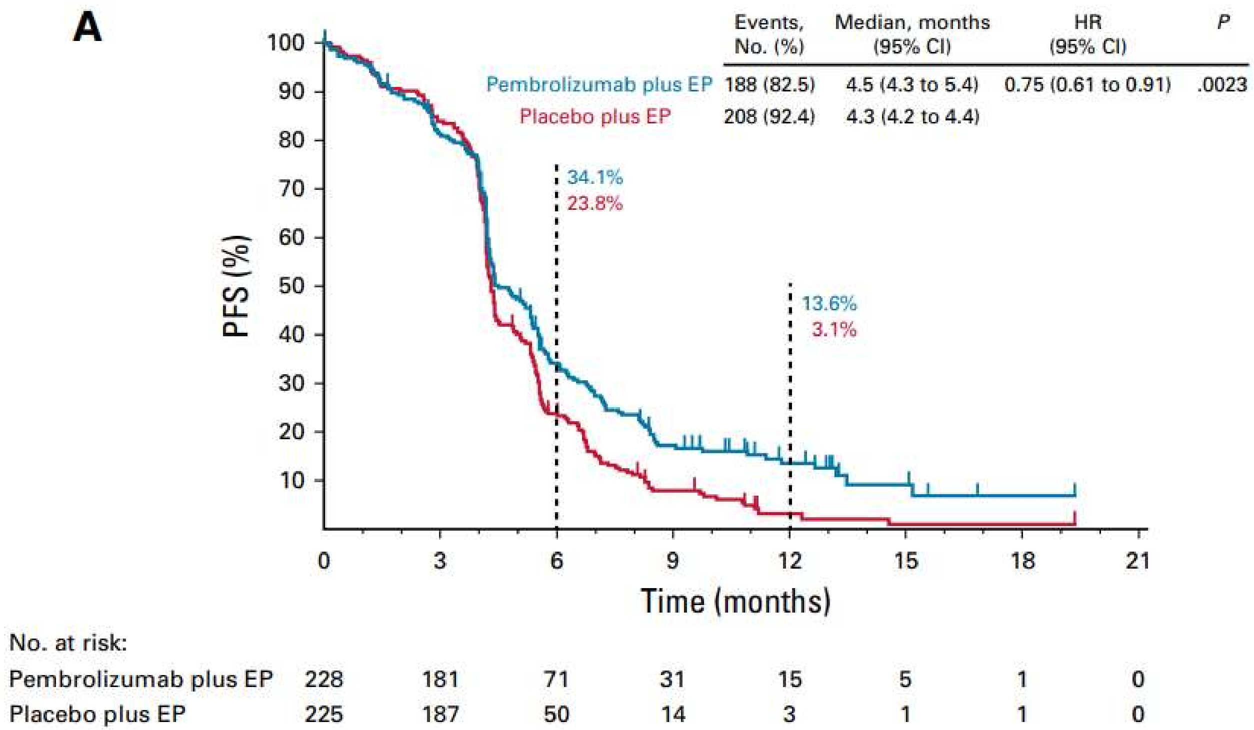
$$S(t_{0.6}) = 0.6 = e^{-\lambda t_m}$$

$$t_{0.6} = -\ln 0.6/\ 0.005 = 102 \text{ days}$$

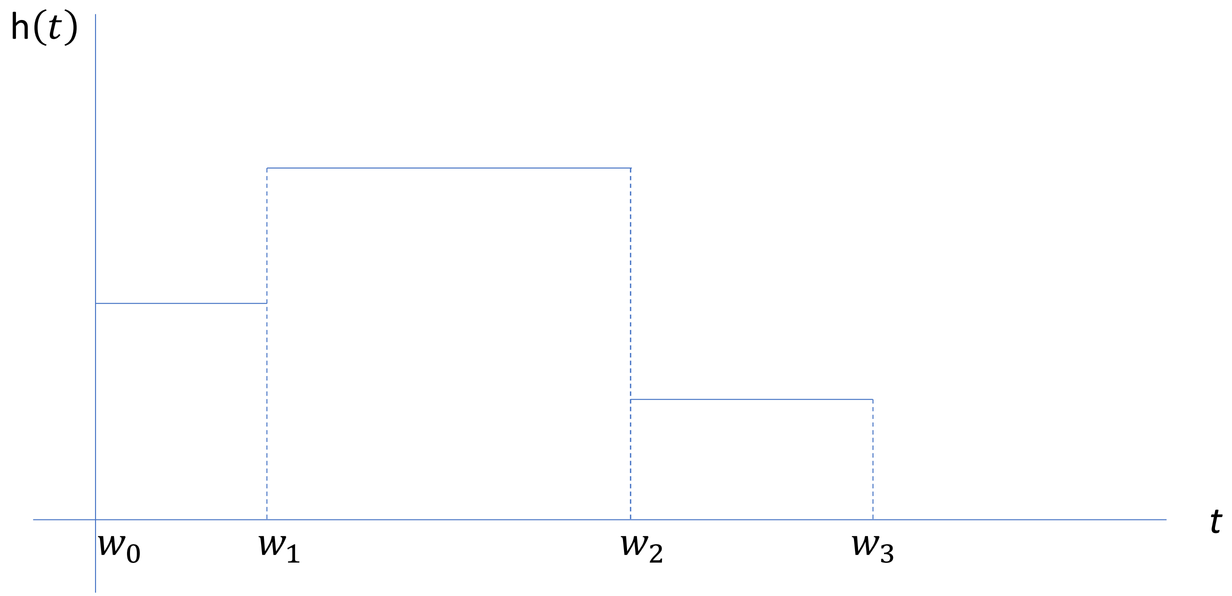
Piecewise Exponential

- Instead of a constant event rate λ in an exponential distribution
- The event rate can change overtime
  - Subject’s risk may increase or decrease
    - Change at certain age threshold
  - Treatment may change the risk, however not uniform overtime
    - PD1 inhibitors delayed treatment effect – no benefit at early treatment stage
  - Sick patients may develop events early
    - The risk set contains relatively healthy subjects
- Approximated by pieces of several exponential distributions

Delayed Treatment Effec



Piecewise Hazard Function



Piecewise Exponential –The Cumulative Hazard Function

- The hazard function

$$\mathrm{h}(t) = h_1 I(t \le w_1) + h_2 I(w_1 < t \le w_2) + \cdots$$

where $w_1, w_2, \dots,$ are fixed time intervals, $w_0 = 0$

- At time $t \in (w_{j-1}, w_j)$, the cumulative hazard function can be written as

$$H(t) = \sum_{i<j} h_i (w_i - w_{i-1}) + h_j \left(t - w_{j-1}\right) \mathrm{I}(t \in (w_{j-1}, w_j))$$

Piecewise Exponential – The Survival Function

- Recall $S(t) = e^{-H(t)}$
- Therefore, for $t \in (w_{j-1}, w_j)$

$$S(t) = e^{-\{\sum_{i<j} h_i(w_i - w_{i-1}) + h_j(t - w_{j-1})\}}$$

$$= \prod_{i<j} e^{-h_i(w_i - w_{i-1})} \, e^{-h_j(t - w_{j-1})}$$

Piecewise Exponential - PDF

- Recall $f(t) = -\frac{dS(t)}{dt}$
- Therefore, for $t \in (w_{j-1}, w_j)$

$$f(t) = -\frac{d}{dt} e^{-\{\sum_{i<j} h_i(w_i - w_{i-1}) + h_j(t - w_{j-1})\}}$$

$$= h_j \prod_{i<j} e^{-h_i(w_i - w_{i-1})} \, e^{-h_j(t - w_{j-1})}$$

Weibull Distribution

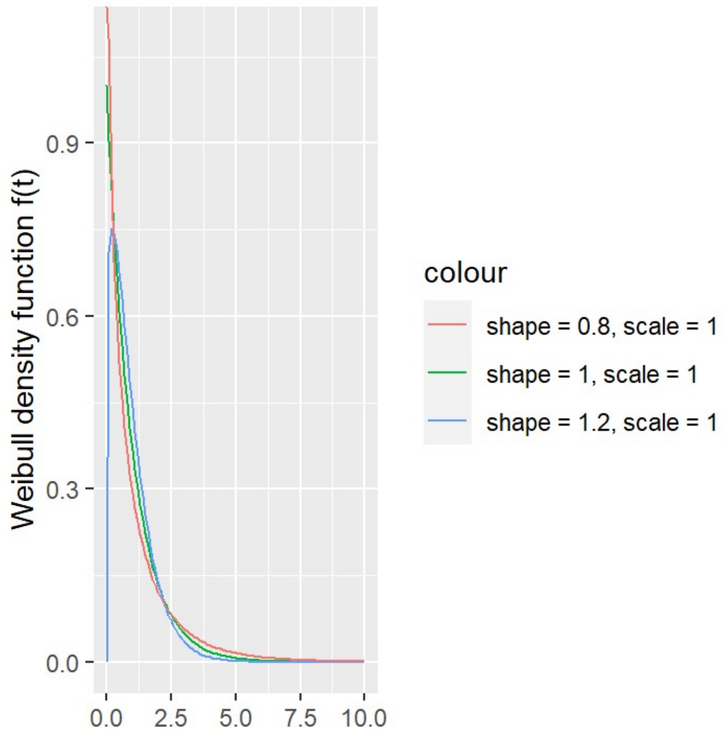
- Generalized exponential distribution
- Accelerated hazard functions
- Convenient to model
  - Non-constant hazard rate for baseline
  - Constant hazard ratio

Weibull

- $T \sim weibull(\alpha, \lambda)$
  - $\alpha > 0$ is the shape parameter
  - $\lambda > 0$ is the scale parameter
  - $\alpha = 1 \implies T \sim Exp(\lambda)$

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

- $E(T) = \lambda^{-\frac{1}{\alpha}}\Gamma(1 + \frac{1}{\alpha})$
- $Var(T) = \lambda^{-\frac{2}{\alpha}}[\Gamma\left(1 + \frac{2}{\alpha}\right) - \{\Gamma\left(1 + \frac{1}{\alpha}\right)\}^2$



Weibull Hazard Function

- The hazard function

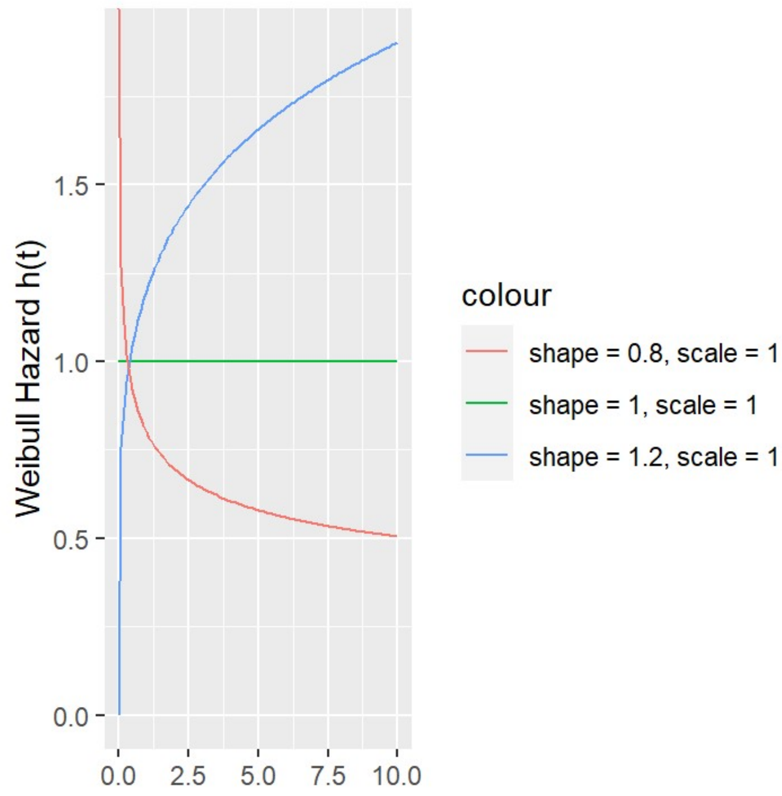
$$h(t) = \frac{f(t)}{S(t)} = \lambda\alpha t^{\alpha-1}$$

- $\alpha > 1$ – Accelerate $\alpha < 1$ – De-accelerate

- Constant hazard ratio

$$h_1(t) = \lambda_1\alpha t^{\alpha-1} \text{ and } h_0(t) = \lambda_0\alpha t^{\alpha-1}$$

$$\frac{h_1(t)}{h_0(t)} = \frac{\lambda_1}{\lambda_0}$$



R-code Plotting Weibull Hazard functions

```
library(epa)

base <- ggplot() + xlim(0, 10) + ylab("Weibull Hazard h(t)")
base +
   geom_function(aes(colour = "shape = 1, scale = 1"),
  fun = hweibull, args = list(shape = 1, scale = 1)) +
   geom_function(aes(colour = "shape = 1.2, scale = 1"),
  fun = hweibull, args = list(shape = 1.2, scale = 1)) +
   geom_function(aes(colour = "shape = 0.8, scale = 1"),
  fun = hweibull, args = list(shape = 0.8, scale = 1))
```

Log-normal

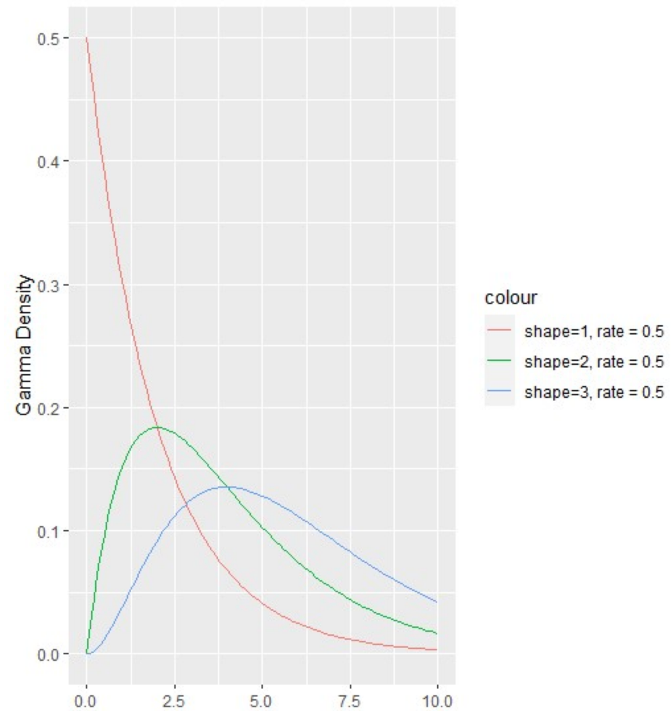
- Another commonly used parametric distribution for survival time
  - $T \sim LN(\mu, \sigma^2) = \exp(N(\mu, \sigma^2))$
  - $S(t) = 1 - \Phi\left(\frac{\ln t - \mu}{\sigma}\right)$
  - $f(t) = \frac{\exp(-\left(\frac{\ln t - \mu}{\sqrt{2}\sigma}\right)^2)}{\sqrt{2\pi} t \sigma}$
  - $E(T) = \exp(\mu + \sigma^2/2)$
  - $Var(T) = \exp(2\mu + \sigma^2/)(\exp(\sigma^2) - 1)$

Gamma

- $T \sim gamma(\alpha, \beta)$
  - $f(t) = \frac{\beta^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\beta t}$
  - $S(t) = P(T > t)$

    no closed form

  - $\alpha > 0$ is the shape parameter
  - $\beta > 0$ is the scale parameter



Log-logistic

- $T \sim Log - logistic(\alpha, \beta)$
  - $f(t) = \frac{\beta}{\alpha^{\beta}(1+\left(\frac{t}{\alpha}\right)^{\beta})^{2}} t^{\beta-1}$
  - $S(t) = P(T > t) = \frac{t^{\beta}}{\alpha^{\beta}+t^{\beta}}$

- $\alpha > 0$ is the shape parameter
- $\beta > 0$ is the scale parameter

Homework 2

1. For discrete survival time, show $f(t_j) = h_j \prod_{i \leq j-1}(1 - h_i)$

2. Show that if $S_1(t) = \{S_0(t)\}^\lambda$, then $h_1(t) = \lambda h_0(t)$

3. Show that the survival function $S(t) = \exp(-H(t))$, where $H(t)$ is a cumulative hazard function

4. In a two-arm randomized and controlled clinical trial, the median survival time in the control and new treatment arms are 9 months and 14 months, respectively. Assuming the survival time follows exponential distribution, what is the hazard rate for the control and new treatment arms? What is the risk reduction in the new treatment in comparison to the control arm?

5. Plot $h(t)$ for log-normal, Gamma distributions with your choice of parameters.