AMS 588 – Biostatistics

Instructor: Prof. Song Wu

Spring, 2013

Part I:
Analysis of Survival Data

Some Basics

- Binomial distribution

The Binomial distribution is based on the idea of a Bernoulli trial, which is an experiment with two, and only two, possible outcomes.

Y_1,…Y_n iid Bernoulli(π), $p(Y_i = 1) = \pi = 1 - \mathrm{p}(Y_i = 0)$

$Y_i = 1$ is often termed as a "success" and π is referred to as the success probability.

$$E(Y_i) = \pi; \quad Var(Y_i) = \pi(1 - \pi)$$

The total number of "success" in n trails follows a *Binomial* distribution:

$Y = \sum_{i=1}^{n} Y_i \sim Bin(n, \pi)$

pmf: $p(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$ for y=0,1,…,n

$E(Y) = \mu = n\pi; \quad Var(Y) = \sigma^2 = n\pi(1 - \pi)$

Binomial Theorem: For any real numbers$n$and y and integer $n \geq 0$,

$$(x + y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

Some Basics

- Hypergeometric distribution

Suppose we have a large urn filled with N balls that are identical in every way except that M are read and $\mathrm{N}-\mathrm{M}$ are green. We reach in, blindfolded, and select K balls at random. The number of red balls X in a sample of size K follows a *Hypergeometric* distribution:

$$P(X=x|N,M,K)=\frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}},\ x=0,1,\dots,K$$

$$\mathrm{E(X)}=\mu=\frac{KM}{N};$$

$$Var(X)=\sigma^2=\frac{KM}{N}\left(\frac{(N-M)(N-K)}{N(N-1)}\right)$$

Example: Let X_1 and X_2 be independent observations with $X_1 \sim binomial(n_1,p_1)$ and $X_2 \sim binomial(n_2,p_2)$. Consider testing $H_0: p_1=p_2$ vs $H_1: p_1>p_2$. It is easy to show that under $H_0: p_1=p_2=p$, $X=X_1+X_2$ is a sufficient statistics for p, and

$$X_1|X=x \sim hypergeometric(n_1+n_2,n_1,x)$$ -- **Fisher's exact test**

The data can also be formulated into a 2x2 table

$\boldsymbol{x_1}$	n_1-x_1	n_1
x_2	n_2-x_2	n_2
x	$n-x$	n

Under $H_0: p_1=p_2$

$$f(x_1|n_1,n_2,x)=\frac{\binom{n_1}{x_1}\binom{n_2}{x-x_1}}{\binom{n}{x}}$$ **(hypergeometric)**

Some Basics

- Iterative expectation & variance

If X and Y are any two random variables, then

$$EX = E(E(X|Y))$$

$$Var(X) = E\big(Var(X|Y)\big) + Var\big(E(X|Y)\big)$$

Provided that the expectations exist

Example: $X|P \sim \text{binomial}(n, P)$ and $P \sim \text{beta}(\alpha, \beta)$

$$EX = E(E(X|P)) = E(nP) = n\frac{\alpha}{\alpha+\beta}$$

$$Var(X) = E\big(Var(X|P)\big) + Var\big(E(X|P)\big)$$

$$= E(P(1-P)) + Var(nP)$$

$$= n\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + n^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

$$= n\frac{\alpha\beta(\alpha+\beta+n)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Some Basics

Central Limit Theorem:

Let $X_1, X_2, \cdots$ *be iid random variables* $\ni$. $M_{X_i}(t) < \infty$ *for* $|t| < h$, *for some* $h > 0$. *Let* $EX_i = \mu$ *and* $Var(X_i) = \sigma^2 > 0$. *Define* $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. *Then*

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

Delta method:

Let $\{Y_n\}$ *be a sequence of random variables and* $\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$.

For a given function g, *suppose* $g'(\theta)$ *exists and is not 0. Then*

$$\sqrt{n}[g(Y_n) - g(\theta)] \xrightarrow{d} N\left(0, [g'(\theta)]^2\sigma^2\right).$$

Proof. *Taylor expansion:* $g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R_2(\theta)$

$R_2(\theta) \longrightarrow 0$ *as* $Y_n \longrightarrow \theta$.

Since $Y_n \xrightarrow{p} \theta$, $R_2(\theta) \xrightarrow{p} 0$.

By Slutsky's theorem, $\sqrt{n}[g(Y_n) - g(\theta)] \approx g'(\theta)\sqrt{n}(Y_n - \theta)$.

Survival Analysis

In many biomedical applications the primary endpoint of interest is time to a certain event, such as time to death; time it takes for a patient to respond to a therapy; time from response until disease relapse (i.e., disease returns); time to eradicate an infection after treatment with antibiotics; etc.

We may be interested in

- characterizing the distribution of "time to event" for a given population ;
- comparing this "time to event" among different groups (*e.g.*, treatment vs. control in a clinical trial or an observational study);
- modeling the *relationship* of "time to event" to other covariates (sometimes called prognostic factors or predictors).

Difficulties:

- Censoring: Since the data are collected over a finite period of time, the "time to event" may not be observed for all the individuals in our study population (sample). This results in what is called censored data.
- Differential follow-up: It is also common that the amount of follow-up for the individuals in a sample vary from subject to subject.

The standard statistical methods cannot properly handle these difficulties in the analysis of the "time to event" data. A new research area in statistics has emerged which is called Survival Analysis or Censored Survival Analysis.

"Time To Event"

Let the random variable *T* denote time to the event of our interest. Of course, *T* is a positive **random variable** which has to be unambiguously defined; that is, we must be very specific about the **start** and **end** with the length of the time period in-between corresponding to *T*.

EX: Survival time of a treatment for a population with certain disease: measured from the time of **treatment initiation** until **death**.

cdf of *T*: $F(t) = P[T \leq t], \quad t \geq 0$ Right continuous: $\lim_{u \to t^+} F(u) = F(t)$

When T is a survival time, $F(t)$ is the probability that a randomly selected subject from the population will die **before** time t.

pdf of *T*: $f(t) = \frac{dF(t)}{dt}, F(t) = \int_0^t f(u)du$

In biomedical applications, it is often common to use the **survival function**

$$S(t) = P[T \geq t] = 1 - F(t^-) \qquad where, F(t^-) = \lim_{u \to t^-} F(u)$$

When *T* is a survival time, *S*(*t*) is the probability that a randomly selected individual will **survive** to time *t* or beyond.

Note: Sometimes, a survival function may be defined as $S(t) = P[T > t] = 1 - F(t)$. This definition will be identical to the above one if *T* is a continuous random variable, which is the case we will focus on in this course.

$S(t)$

- Non-increasing;
- $S(0) = 1$ and $S(\infty) = 0$ for a proper random variable, which means that everyone will eventually experience the event, e.g. death.
- However, we may also allow the possibility that $S(\infty) > 0$. This corresponds to a situation where there is a positive probability of not "dying" or not experiencing the event. For example, if the event of interest is the time from response until disease relapse and the disease has a cure for some proportion of individuals in the population, then we have $S(\infty) > 0$, where $S(\infty)$corresponds to the proportion of cured individuals.
- If T is continuous r.v.,

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

That is, there is a one-to-one correspondence between $f(t)$ and $S(t)$.

Definitions

Mean Survival Time: $\mu = E(T)$. Due to censoring, sample mean of observed survival times is no longer an unbiased estimate of $\mu = E(T)$. If we can estimate $S(t)$ well, then we can estimate $\mu = E(T)$ using the following fact:

$$E(T) = \int_0^\infty t dF(t) = -\int_0^\infty t dS(t) = \int_0^\infty S(t) dt$$

Median Survival Time: Median survival time m is defined as the quantity m satisfying $S(m) = 0.5$. Sometimes denoted by $t_{0.5}$. If $S(t)$ is not strictly decreasing, m is the smallest one such that $S(m) \leq 0.5$.

***p*th quantile of Survival Time** (100*p*th percentile): t_p such that $S(t_p) = 1 - p$ $(0 < p < 1)$. If $S(t)$ is not strictly decreasing, t_p is the smallest one such that $S(t_p) \leq 1 - p$

Mean Residual Life Time(mrl):

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

i.e., $mrl(t_0) =$ average remaining survival time **given** the population has survived beyond t_0. It can be shown that

$$mrl(t_0) = \frac{\int_{t_0}^\infty S(t) dt}{S(t_0)}$$

EX:

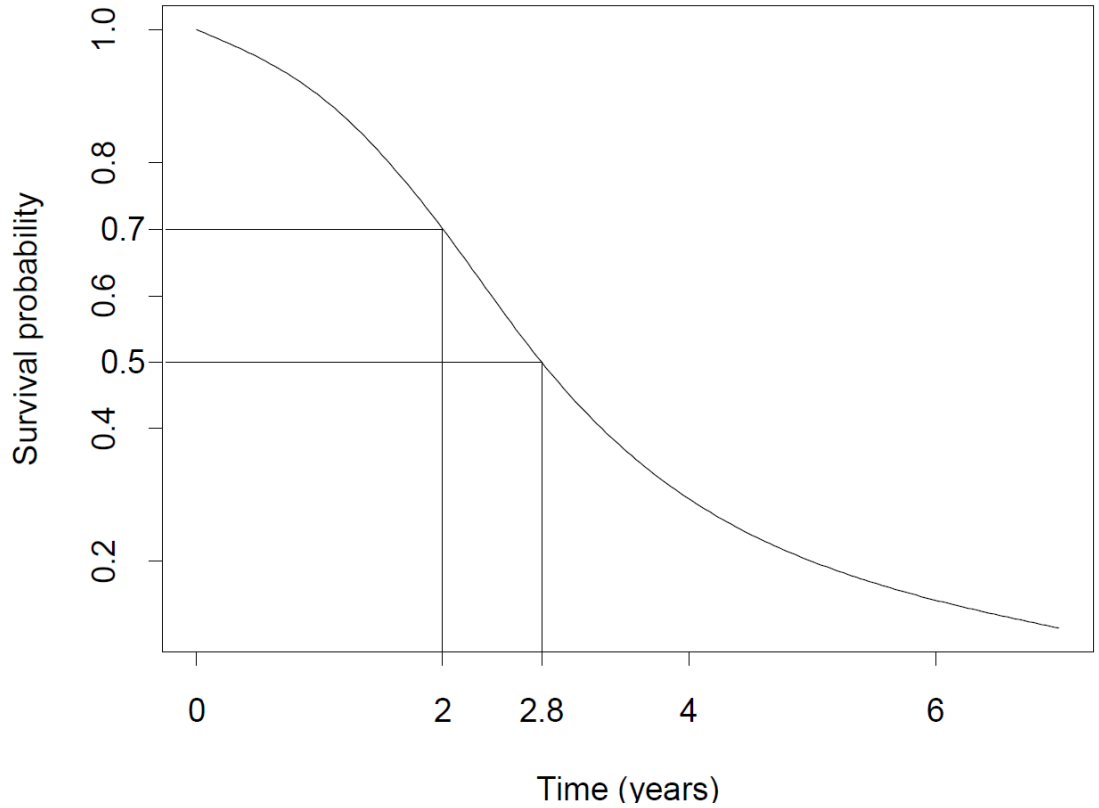


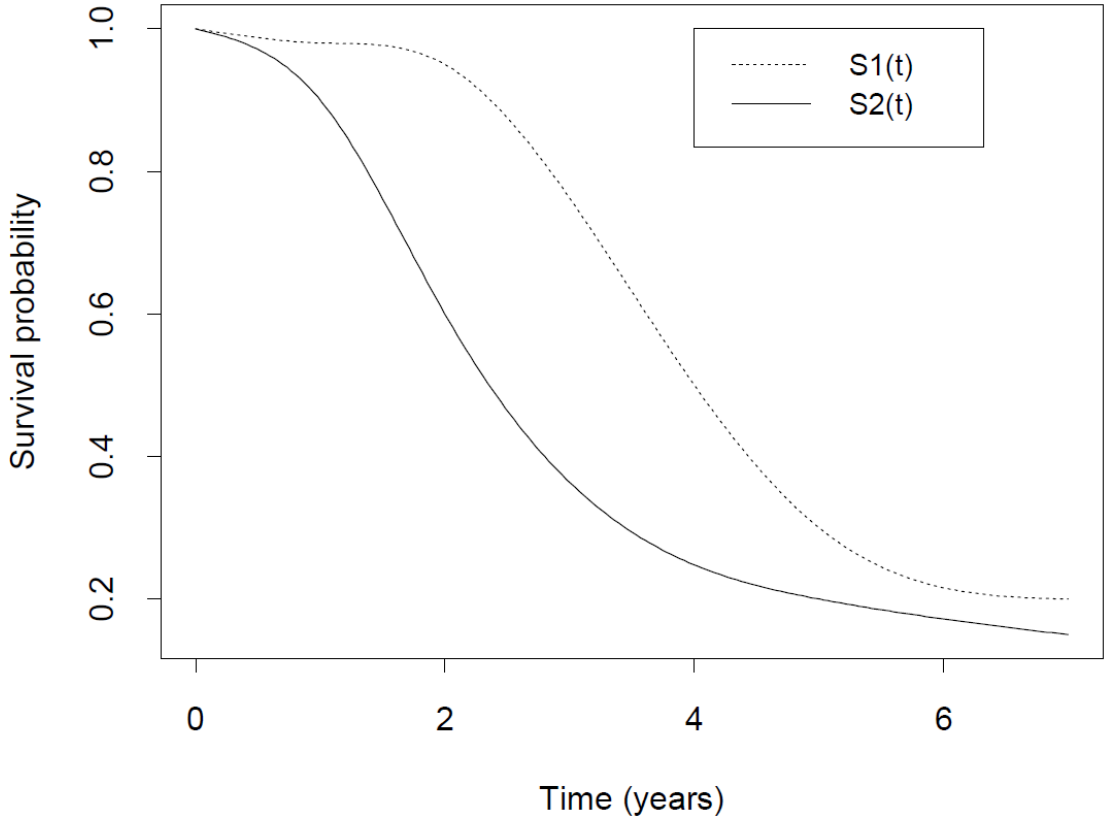
Figure 1.1: *The survival function for a hypothetical population*

In this population:

- 70% of the individuals will survive at least 2 years (i.e., $t_{0.3}$ = 2) and
- the median survival time is 2.8 years (*i.e.*, 50% of the population will survive at least 2.8 years).

T_1 is stochastically larger than T_2

Figure 1.2:



We say that the survival distribution for group 1 is stochastically larger than the survival distribution for group 2 if $S_1(t) \geq S_2(t)$, for all $t \geq 0$, where $S_i(t)$is the survival function for group i. If T_i is the corresponding survival time for groups i, we also say that T_1 is stochastically (not deterministically) larger than T_2. Note that T_1 being stochastically larger than T_1 does NOT necessarily imply that $T_1 \geq T_2$.

Mortality Rate

The **mortality rate** at time t, where t is generally taken to be an integer in terms of some unit of time (*e.g.*, years, months, days, etc), is the proportion of the population who fail (die) between times t and $t+1$**among** individuals **alive** at time t, , *i.e.*,

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

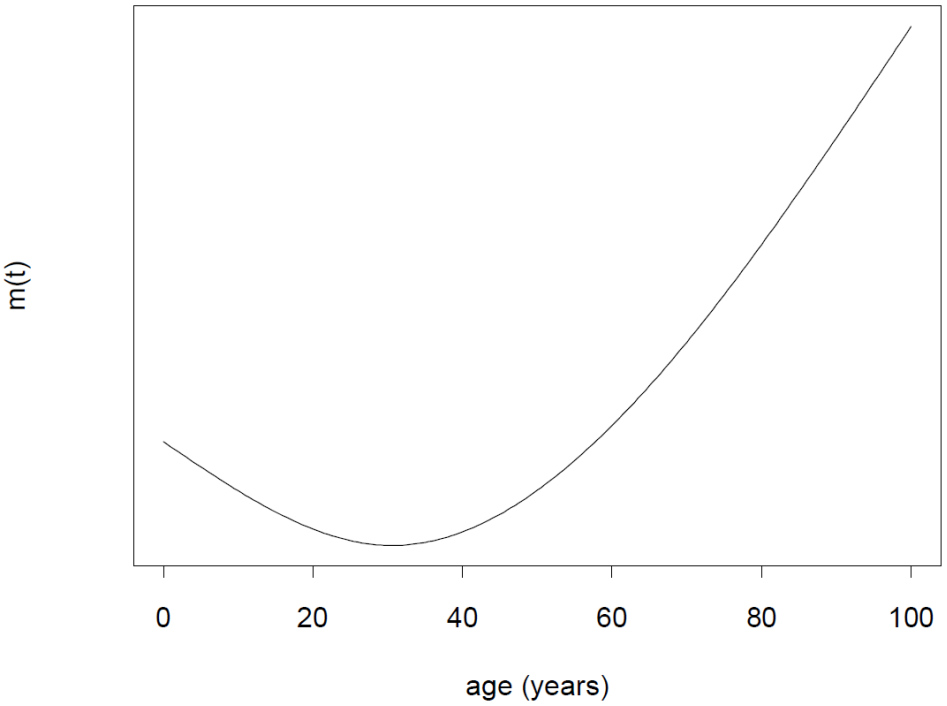


Figure 1.3: *A typical mortality pattern for human*

Hazard Rate

The **hazard rate** is the limit of a mortality rate if the interval of time is taken to be small (rather than one unit). The hazard rate is the instantaneous rate of failure (experiencing the event) at time t given that an individual is alive at time t.

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

$$= \frac{\lim_{h \to 0} \frac{P[t \le T < t + h]}{h}}{P[T \ge t]}$$

$$= \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{dlog\{S(t)\}}{dt}$$

- A useful way of describing the distribution of "time to event" because it has a natural interpretation that relates to the aging of a population;
- Very popular in biomedical community.
- If h is very small, we have $P[t \le T < t + h | T \ge t] \approx \lambda(t)h$
- $\Lambda(t) = \int_0^t \lambda(u)\, du = -log\{S(t)\}$, where $\Lambda(t)$ is referred to as the ***cumulative hazard*** function. Here we used the fact that *S*(0) = 1. Hence,

$$S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u) du}$$

Note:

1. There is a one-to-one relationship between hazard rate $\lambda(t), t \geq 0$, and survival function $S(t)$, namely,

$$S(t) = e^{-\int_0^t \lambda(u)du}, \qquad \lambda(t) = -\frac{dlog\{S(t)\}}{dt}$$

2. The hazard rate is NOT a probability, but a probability rate. Therefore it is possible that a hazard rate can exceed one in the same fashion as a density function $f(t)$ may exceed one.

Common Parametric Models for survival data:

Distribution	$\lambda(t)$	$S(t)$	density $f(t)$	$\mathrm{E}(T)$
Exponential	$\lambda(>0)$	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	$\frac{1}{\lambda}$
Weibull	$\alpha\lambda t^{\alpha-1}(\alpha,\lambda>0)$	$e^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma	$\frac{f(t)}{S(t)}$	$1-I(\lambda t,\beta)$	$\frac{\lambda^{\beta}t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$

$I(t,\beta) = \int_0^t \frac{u^{\beta-1}e^{-u}}{\Gamma(\beta)}du.$

(See page 38 of Klein and Moeschberger (textbook) for more distributions.)

Exponential Distribution

$$\lambda(t) = \lambda,\ S(t) = e^{-\lambda t} \text{ and } f(t) = \lambda e^{-\lambda t}$$

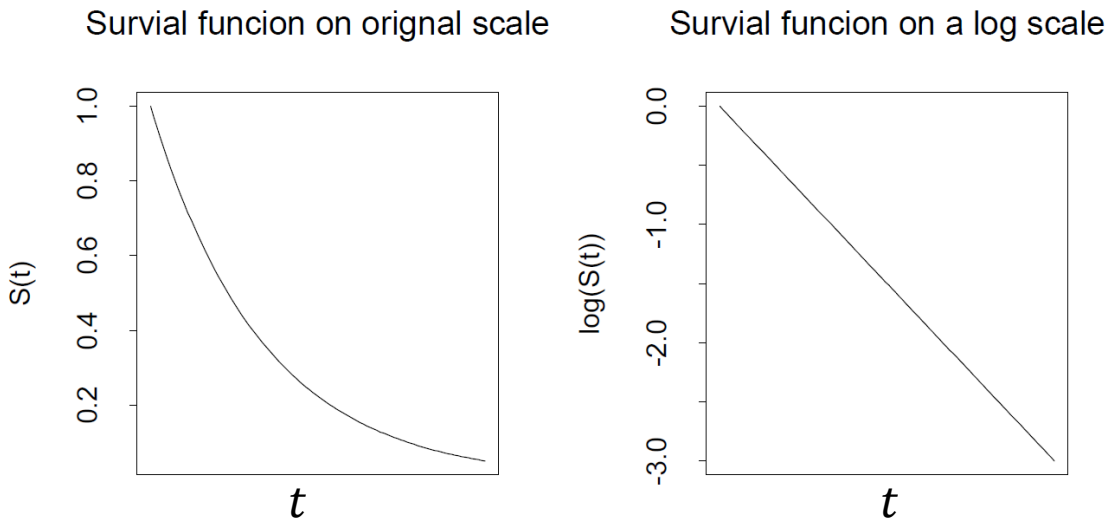
The **Mean Survival Time**: $\mu = E(T) = \int_0^\infty t f(t)dt = \int_0^\infty S(t)dt = \int_0^\infty e^{-\lambda t}dt = \frac{1}{\lambda}$.

The **Median Survival Time**: $t_{0.5} = \frac{log2}{\lambda}$, since $S(t_{0.5}) = e^{-\lambda t_{0.5}} = 0.5$.

The **Mean Residual Life Time** after t_0

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

Notice that $\log[S(t)] = -\lambda t$. Therefore, sometimes it is useful to plot the survival distribution on a log scale, which can be used to check if the underlying true distribution of the survival time is exponential or not given a data set.



Check Exponential Distribution

1. Suppose we can have an estimate $\hat{S}(t)$of $S(t)$without assuming any distribution of the survival time (the Kaplan-Meier estimate to be discussed later is such an estimate). Then we can plot $\log\left[\hat{S}(t)\right]\ vs\ t$ to see if it is approximately a straight line. A (approximate) straight line indicates that the exponential distribution may be a reasonable choice for the data.

2. Alternatively, we can assume the exponential distribution for the data and get the estimate of $S(t) = \ e^{-\lambda t}$(we only need to estimate λ; this kind of estimation will be discussed in Chapter 3). Denote this estimate by $\hat{S}_1(t)$ and Kaplan-Meier estimate by $\hat{S}_{KM}(t)$. If the exponential distribution assumption is correct, both estimates will be good estimates of the same survival function $S(t) = \ e^{-\lambda t}$. Therefore, $\hat{S}_1(t)$ and $\hat{S}_{KM}(t)$ should be close to each other and hence the plot $\hat{S}_1(t)\ \ vs\ \hat{S}_{KM}(t)$ should be approximately a straight line. A non-straight line indicates that the exponential distributional assumption is not appropriate.

Weibull Distribution

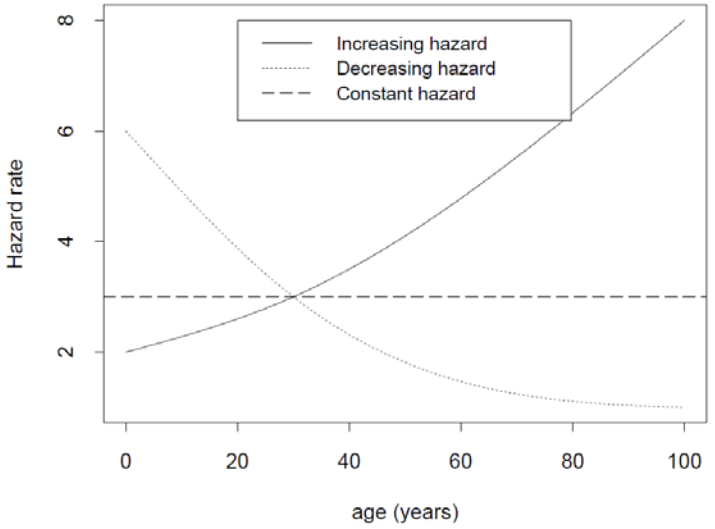
$$\lambda(t) = \alpha\lambda t^{\alpha-1},\ S(t) = e^{-\lambda t^{\alpha}}$$

The **Mean Survival Time**: $\mu = E(T) = \int_0^{\infty} S(t)dt = \int_0^{\infty} e^{-\lambda t^{\alpha}} dt = \frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$.

The **Median Survival Time**: $t_{0.5} = \left[\frac{log2}{\lambda}\right]^{1/\alpha}$, since $S(t_{0.5}) = e^{-\lambda t_{0.5}^{\alpha}} = 0.5$.

The Weibull model allows :

- Constant hazard: $\alpha = 1$;
- increasing hazard: $\alpha > 1$;
- decreasing hazard: $\alpha < 1$.



Notice that:

$$\log\{\log[S(t)]\} = \log(\lambda) + \alpha\log(t).$$

Therefore a straight line in the plot of $\log\{\log[S(t)]\}\ vs\ \log(t)$ indicates a Weibull model. We can use the above equation to check if the Weibull model is a reasonable choice for the survival time given a data set.

Alternatively, we can assume a Weibull model for the survival time and use the data to estimate $S(t)$and plot this estimate against the Kaplan-Meier estimate as we proposed for the exponential distribution. A (approximate) straight line indicates the Weibull model is a reasonable choice for the data.

Other Parametric Models for Survival Data

- Log-normal; Exponential Power; Gompertz; Inverse Gaussian; Pareto; Gamma; Generalized Gamma…
- See section 2.5 of Klein and Moeschberger (textbook) for more details.