# Applied Survival Analysis Using R
# Chapter 2: Basic Principle of Survival Analysis

Qi Guo

Department of Mathematical Sciences
The University of Texas at Dallas

April, 9 2019

1 Two key ways of specifying a survival analysis

2 Connections

3 Mean and Median Survival Time

4 Parametric Survival Distributions

5 A Brief Introduction to Maximum Likelihood Estimation

# The Hazard and Survival Functions

## The Survival Function

Survival analysis is defines the probability of surviving up to a point $t$. Formally,

$$S(t) = 1 - P(T \leq t) = 1 - F(t),\ 0 < t < \infty$$

And it decreases over time

## The Hazard Function

It is the instantaneous failure rate. It is the probability that, given that a subject has survived up to time t, he or she fails in the next small interval of time, divided by the length of that interval. Formally,

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

# Connections

- Cumulative risk function: $F(t) = P(T \geq t),\ 0 < t < \infty$
- Probability density function: $f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}$
- So we can deduce the following two formulas:
  - $h(t) = \frac{f(t)}{S(t)}$
  - $H(t) = \int_0^t h(u)du$
  - $S(t) = exp(-\int_0^t h(u)du) = exp(-H(t))$

# Mean and Median Survival Time

- The mean survival is the expected value of the survival time
$$\mu = E(T) = \int_0^\infty tf(t)dt$$
- or
$$\mu = \int_0^\infty S(t)dt$$
- The mean survival also cannot be computed with the Kaplan-Meier survival curve when the curve does not reach zero, an issue we will discuss in the next chapter.
- The median survival time is defined as the time t such that $S(t) = 1/2$

# Exponential Distribution

- Commonly we analyze the non-parametric survival distribution due to more flexible and more applicable.
- The exponential distribution is the simplest but classical survival distribution, has a *constant* hazard, $h(t) = \lambda$
- $H(t) = \lambda t$
- $S(t) = e^{-H(t)} = e^{-\lambda t}$
- $f(t) = S(t)h(t) = \lambda e^{-\lambda t}$
- $E(T) = 1/\lambda$
- $t_{med} = log(2)/\lambda$

# Weibull Distribution

- The exponential distribution is easy to work with, but the constant hazard assumption is not often appropriate for describing the lifetimes of humans or animals.
- $h(t) = \alpha\lambda^{\alpha}t^{\alpha-1}$
- $H(t) = (\lambda t)^{\alpha}$
- $S(t) = e^{-(\lambda t)^{\alpha}}$
- $E(T) = \frac{\Gamma(1+1/\alpha)}{\lambda}$
- $t_{med} = \frac{[log(2)]^{1/\alpha}}{\lambda}$

# In R

- We may generate random variables from the exponential or Weibull distribution using the functions "**`rexp`**" and"**`rweib`**".
- The functions "**`dweibull`**" and "**`pweibull`**" compute the p.d.f. and c.d.f., respectively, of the Weibull distribution.
- These functions use the arguments "shape" and "scale" to represent the parameters $\alpha$ and $1/\lambda$.

```
> tt.weib <- rweibull(1000, shape=1.5, scale=1/0.03)
> mean(tt.weib)
[1] 31.35497
> median(tt.weib)
[1] 26.84281
```

# In R

- Plot the *Weibull survival function* with $\alpha = 1.5$ and $\lambda = 0.03$ by first defining a function "**weibSurv**" with these parameters and then using the "**curve**" function to plot the curve as follows:

```
> weibSurv <- function(t, shape, scale) pweibull(t, shape=shape,
scale=scale, lower.tail=F)
> curve(weibSurv(x, shape=1.5, scale=1/0.03), from=0, to=80,
ylim=c(0,1), ylab='Survival probability', xlab='Time')
```

- Plot the *hazard function* with p.d.f. divided by the survival function:

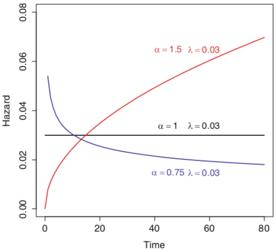
```
> weibHaz <- function(x, shape, scale) dweibull(x, shape=shape,
scale=scale)/pweibull(x, shape=shape, scale=scale,
lower.tail=F)
> curve(weibHaz(x, shape=1.5, scale=1/0.03), from=0, to=80,
ylab='Hazard', xlab='Time', col="red")
```

# In R

- The other two curves obtained using $\alpha = 1$ and $\alpha = 0.75$ when calling the "**`curve`**" function. To place the additional curves on the plot, add "**`add=T`**" as an option to the "**`curve`**" function.

```
> curve(weibHaz(x, shape=1, scale=1/0.03), from=0, to=80,
ylab='Hazard', xlab='Time', col="black", add = TRUE)
> curve(weibHaz(x, shape=0.75, scale=1/0.03), from=0, to=80,
ylab='Hazard', xlab='Time', col="blue", add = TRUE)
```

**Fig. 2.4** Weibull hazard functions



# M.L.E

- No Censoring

$$L(\lambda; t_1, t_2, ..., t_n) = f(t_1, \lambda) \cdot f(t_2, \lambda) \cdot \cdots \cdot f(t_n, \lambda) = \prod_{i=1}^{n} f(t_i, \lambda)$$

- Right-censoring

$$L(\lambda; t_1, t_2, ..., t_n) = \prod_{i=1}^{n} f(t_i, \lambda)^{\delta_i} \cdot S(t_i, \lambda)^{1-\delta_i} = \prod_{i=1}^{n} h(t_i, \lambda)^{\delta_i} \cdot S(t_i, \lambda)$$

- This expression means that when $t_i$ is an *observed death*, the censoring indicator is $\delta_i = 1$, and we enter a *p.d.f. factor*. When $t_i$ is a *censored observation*, we have $\delta_i = 0$ we enter a *survival factor*.

# Example

- Exponential M.L.E
  $L(\lambda) = \prod_{i=1}^{n} [\lambda e^{-t_i/\lambda}]^{\delta_i} [e^{-\lambda t_i}]^{1-\delta_i}$
  $d = \sum_{i=1}^{n} \delta_i$,means the total of the death
  $V = \sum_{i=1}^{n} t_i$,means the total of the time of patients on study
- Log-Likelihood
  $l(\lambda) = d\ log\lambda - \lambda V$
- First Derivative(*Score Function*)
  $\mathfrak{l}'(\lambda) = \frac{d}{\lambda} - V \Rightarrow \hat{\lambda} = \frac{d}{V}$

# Example

- Second Derivative(*Information*)
  $l''(\lambda) = -\frac{d}{\lambda^2} = -I(\lambda)$

- The inverse of the (*Information*) is approximately the variance of the m.l.e
  $var(\hat{\lambda}) \approx I^{-1}(\lambda) = \lambda^2/d$

- Substitute the $\hat{\lambda}$ for $\lambda$ to obtain the observed information $I(\hat{\lambda})$, and get an estimate of the variance
  $\hat{var}(\hat{\lambda}) \approx I^{-1}(\hat{\lambda}) = \hat{\lambda}^2/d = d/V^2$

- Use this formula to carry out hypothesis tests or find a confidence interval for $\lambda$.