BI476: Biostatistics - Case Studies

Lec07: Survival Data Analysis

Maoying,Wu
ricket.woo@gmail.com

Dept. of Bioinformatics & Biostatistics
Shanghai Jiao Tong University

Spring, 2018

Outline

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Phase II Trial of Stage-2 Breast Cancer

As we know, the objective of Phase II trial was to assess the relative efficacy of themotherapy (CT) and immunotherapy (IT) alone and in combination, as adjuvant to surgery in the treatment of patients with stage-2 breast carcinoma.

The design of the trial was parallel with randomized assignment to the three treatment groups:

- `S+CT`: Surgery plus one year of chemotherapy
- `S+IT`: Surgery plus one year of immunotherapy
- `S+CT+IT`: Surgery plus one year of combination treatment of chemotherapy and immunotherapy.

The outcome is "time-to-death" with two columns:

- `Time`: Time-to-death in weeks;
- `Status`: Censoring status with 0-died; 1-censored (i.e., alive at the last-follow-up)

The primary objective is to assess whether immunotherapy will improve survival.

Importing the dataset

```
library(psych)
carcinoma <- read.table("data/carcinoma-ct-it.txt", header=T)
headTail(carcinoma)
```

```
##          TRT Time Status Age
## 1       S+CT   48      1  26
## 2       S+CT   55      0  65
## 3       S+CT   58      1  48
## 4       S+CT   63      0  53
## ...     <NA>  ...    ... ...
## 28   S+CT+IT  239      1  55
## 29   S+CT+IT  239      1  56
## 30   S+CT+IT  240      1  61
## 31   S+CT+IT  242      1  35
```

```
table(carcinoma$Status)
```

```
##
##  0  1
## 14 17
```

```
table(carcinoma$TRT)
```

Next Section ...

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Survival: Primary functions and definitions

Let T be random survival/failure time

- **Density function (死亡密度函数)** $f(t)$
- **Survival function (生存函数)** $S(t)$
- **Hazard (rate) function (风险函数)** $h(t)$
- **Cumulative hazard function (累积风险函数)** $H(t)$
- **Cumulative distribution function (累积密度函数)** $F(t)$

The density function

The **death density function** is usually employed for modeling survival data using a likelihood approach. The mathematical defintion of the **death density function** $f(t)$ is

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

$f(t)$ may be thought of as

- unconditional probability of dying at time t;
- the instantaneous unconditional risk of death at time t

The Survival Function

The mathematical definition of the **survival function** $S(t)$ is

$$S(t) = Pr(T > t)$$

where T denotes the random variable survival time.

The survial function $S(t)$ can be interpreted as

- the **probability** of surviving to at least time t
- the cumulative proportion surviving to at least t,
- the proportion of a population surviving to at least t

The hazard function

The mathematical definition of the **hazard function** $h(t)$ is

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

where T denotes the random variable - survival time.

- When Δt is small, $P(t \leq T \leq t + \Delta t | T \geq t) \approx h(t)\Delta t$
- $h(t)$ is the instantaneous risk of failure at time t given that death did not occur prior to t.

According to the hazard function definition,

$$\begin{aligned} h(t) &= \frac{\lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t]}{\Delta t}}{P[T \geq t]} \\ &= \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} \\ &= -\frac{d \log S(t)}{dt} \end{aligned}$$

The parametric hazard functions

A hazard function may

- remain constant w.r.t time (**exponential density**)
- increase as a function to time according to some power function (**Weibull density**)
- increase linearly with time (**Rayleigh density**)
- increase exponentially with time (**Gompertz density**)
- etc.

Cumulative hazard function

累积风险函数

$$H(t) = \int_0^t h(u)du$$

a.k.a integrated hazard function, is widely used in survival modeling.

Cumulative distribution function

Cumulative death distribution function $F(t)$ is the complement of the survival function

$$F(t) = 1 - S(t)$$

Survival Data: Summary

- The death density is the negative differential of the survival function:

$$f(t) = -\frac{d[S(t)]}{dt}$$

- There is one-to-one relationship between the hazard rate function $h(t), t \geq 0$ and survival function $S(t)$, namely

$$S(t) = \exp\{-\int_0^t h(u)du\}, h(t) = -\frac{d \log S(t)}{dt}$$

- The hazard rate is NOT a probability, but a **probability rate**. Therefore it is possible that a hazard rate function can exceed 1.
- $S(0) = 1$ and $S(\infty) = 0$.
- **Median survival time**: $m = (t|S(t) = 0.5)$; If $S(t)$ is not strictly decreasing, $m = \min\{t : S(t) \leq 0,5\}$.
- We will focus on right-censored data.

Statistical Methods for Right-Censored Data

- Parametric models
- Nonparametric models: Kaplan-Meier Estimator
  - no distributional or specific model assumptions about the observed survival times.
  - distribution-free methods
- Semi-parametric models: Cox Proportional Hazards Regression
  - accounting for covariate information that might be correlated with survival

Next Section ...

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Parametric models for survival

Survival follows some parametric distribution or model (assumption).

- Exponential distribution
- Weibull distribution
- Rayleigh distribution
- Gompertz distribution
- Lognormal distribution

The Exponential Model

The exponential model is used to model a constant hazard function:

$$h(t) = \lambda > 0$$

The exponental death density function is:

$$f(t) = \lambda \exp(-\lambda_0 t)$$

The exponential survival function is:

$$S(t) = \exp(-\lambda t)$$

The cumulative death distribution function is

$$F(t) = 1 - \exp(-\lambda t)$$

The Weibull Model

The Weibull model derives from a power law hazard function:

$$h(t) = \lambda_0 \lambda_1 t^{\lambda_1 - 1}$$

where $\lambda_0 > 0$ and $\lambda_1 > 0$. If $\lambda_1 > 1$, $h(t)$ is guaranteed to be monotonely increasing.

The Weibull death density function is

$$f(t) = \lambda_0 \lambda_1 t^{\lambda_1 - 1} \exp(-\lambda_0 t^{\lambda_1})$$

The Weibull survival function is

$$S(t) = \exp(-\lambda_0 t^{\lambda_1})$$

The Weibull cumulative death density function is

$$F(t) = 1 - \exp(-\lambda_0 t^{\lambda_1})$$

The Rayleigh Model

The Rayleigh model is used to model a linear hazard function:

$$h(t) = \lambda_0 + 2\lambda_1 t$$

where $\lambda_0 > 0$ and $\lambda_1 \geq 0$. It may be noted that if $\lambda_1 > 0$ then $h(t)$ is monotone increasing.

The Rayleigh death density function is

$$f(t) = (\lambda_0 + 2\lambda_1 t) \exp[-(\lambda_0 t + \lambda_1 t^2)]$$

The Rayleigh survival function is

$$S(t) = \exp[-(\lambda_0 t + \lambda_1 t^2)]$$

The Rayleigh cumulative death distribution function is

$$F(t) = 1 - \exp[-(\lambda_0 t + \lambda_1 t^2)]$$

The Gompertz Model

The Gompertz model derives from an exponential hazard function:

$$h(t) = \exp(\lambda_0 + \lambda_1 t)$$

Note that if $\lambda_1 > 0$, then $h(t)$ is monotone increasing.

Gompertz death density function

$$f(t) = \exp(\lambda_0 + \lambda_1 t) \exp\left\{ \frac{1}{\lambda_1} [\exp(\lambda_0) - \exp(\lambda_0 + \lambda_1 t)] \right\}$$

The Gompertz survival function is

$$S(t) = \exp\left\{ \frac{1}{\lambda_1} [\exp(\lambda_0) - \exp(\lambda_0 + \lambda_1 t)] \right\}$$

The cumulative death distribution function is

$$F(t) = 1 - \exp\left\{ \frac{1}{\lambda_1} [\exp(\lambda_0) - \exp(\lambda_0 + \lambda_1 t)] \right\}$$

The Lognormal Model

The lognormal death density function is given by

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left[-\frac{(\log(t) - \mu)^2}{2\sigma^2}\right]$$

The lognormal survival function is

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

where $\Phi(\cdot)$ is the standard cumulative normal distribution.

The hazard function is

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

The cumulative death distribution function is

$$F(t) = \Phi\left[\frac{\log(t) - \mu}{\sigma}\right]$$

Next Section ...

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Survival Data: Likelihood Principle

- In a survival data, some failure times are observed, while others are only partially observed.
- They can be right-censored (右删失), left-censored (左删失), or interval-censored (区间删失)
- For censored data, estimation of moments (矩估计) is not possible.
- Likelihood provides a highly versatile fashion for quantifying the consistency of parameters with the observed data, which makes it particularly well-suited to survival analysis.

Likelihood: Definition

- Let X denote observable data, and suppose that we have a probability model that relates potential values of X to an unknown parameter θ
- Given observed data $X = x$, the **likelihood function** for θ is defined as

$$L(\theta) = \mathcal{P}(x|\theta)$$

- Note that this is a function of θ, not x; now that we have observed the data, x is fixed.
- Also, note that a likelihood function is not a probability distribution - for example, it does not have to integrate to 1.

The Likelihood for a Set of Observations

- The likelihood for a set of observations ($x = \{x_1, \ldots, x_n\}$) is

$$L(\lambda) = \prod_{i=1}^{n} L_i$$

- Likelihoods provide only a relative measure of preference for one parameter value vs. another.
- $L(\lambda)$ is not meaningful, but the relative quantity $L(\lambda_1)/L(\lambda_2)$ is meaningful

Likelihood for Fully Observed Data

- Suppose that the survival time follow a **exponential** distribution:

$$T_i \sim \exp(\lambda)$$

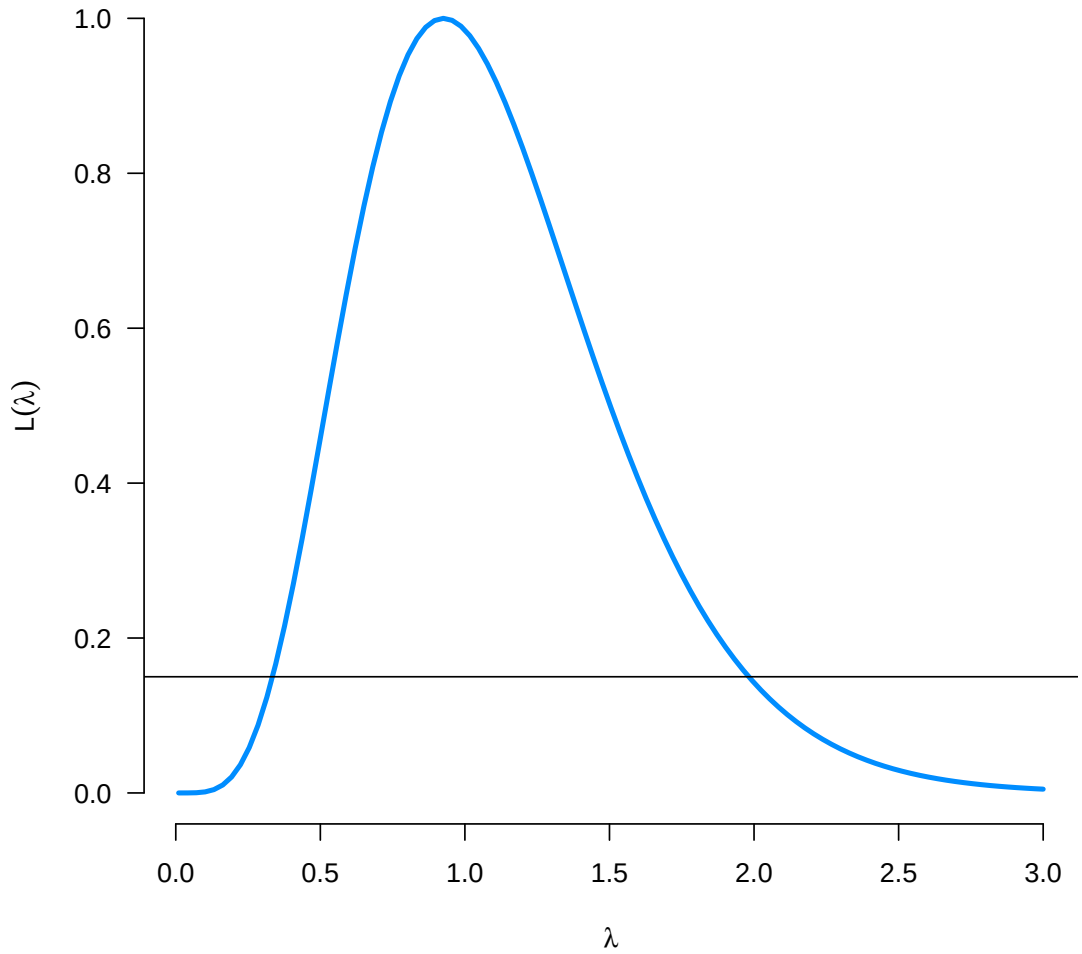
- Suppose we observe the following data:

$$t = \{0.1, 0.5, 0.5, 1.6, 2.7\}$$

- The likelihood is therefore

$$L(\lambda) = \prod_i f(t_i|\lambda) = \prod_i \lambda \exp(-\lambda t_i)$$

Likelihood function: Fully observed data



Likelihood Function for Right-Censored Data

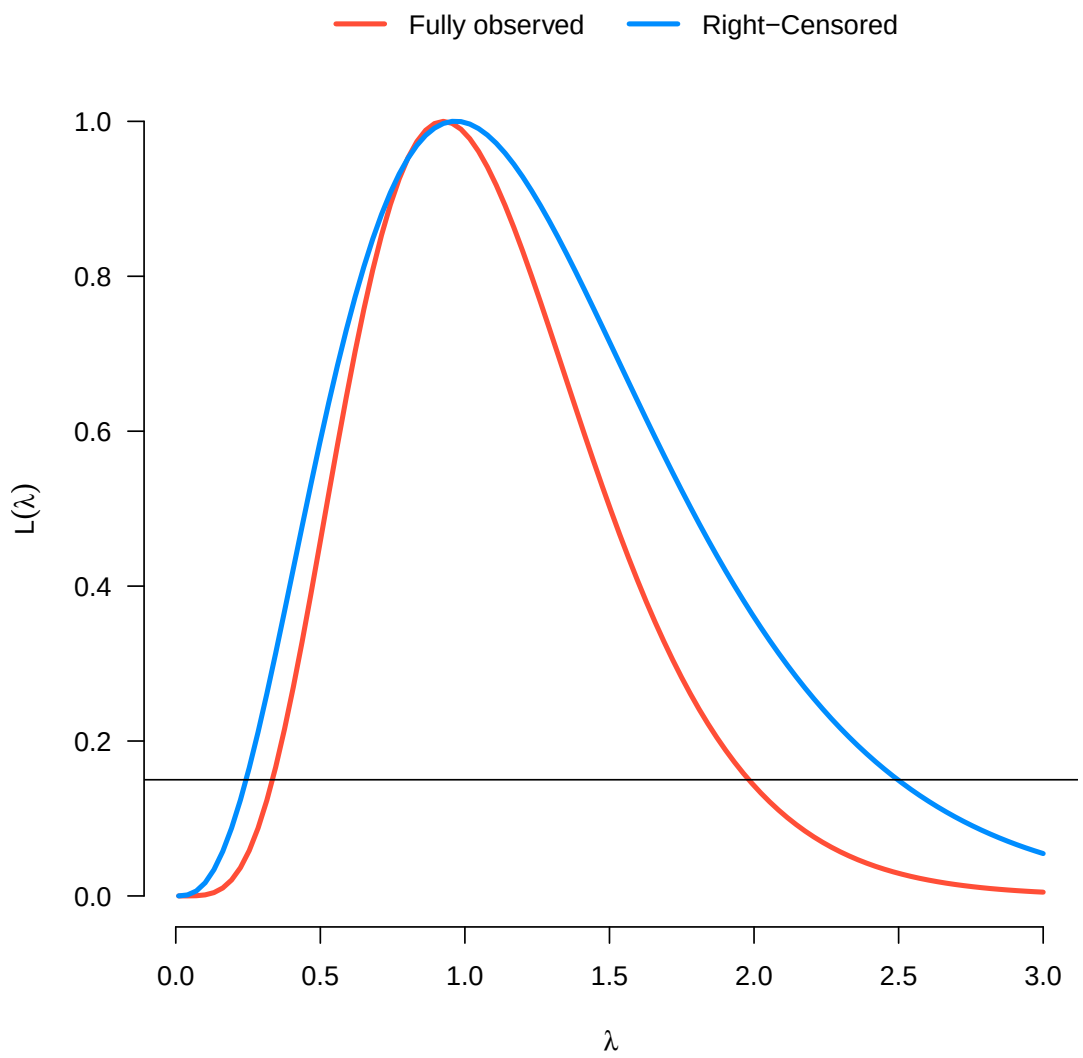
- Suppose that the study was stopped at time $t = 1$
- Therefore for $\{t_1, t_2, t_3\} = \{0.1, 0.5, 0.5\}$, the likelihood remains the same
- While for t_4 and t_5, the likelihood is now:

$$\mathcal{P}(T > 1|\lambda) = S(1|\lambda) = \exp(-\lambda)$$

- The likelihood becomes:

$$L(\lambda) = \prod_{i=1}^{3} f(t_i|\lambda) \prod_{i=4}^{5} S(1|\lambda)$$

Likelihood with right-censored data



Likelihood for Left-censored Data

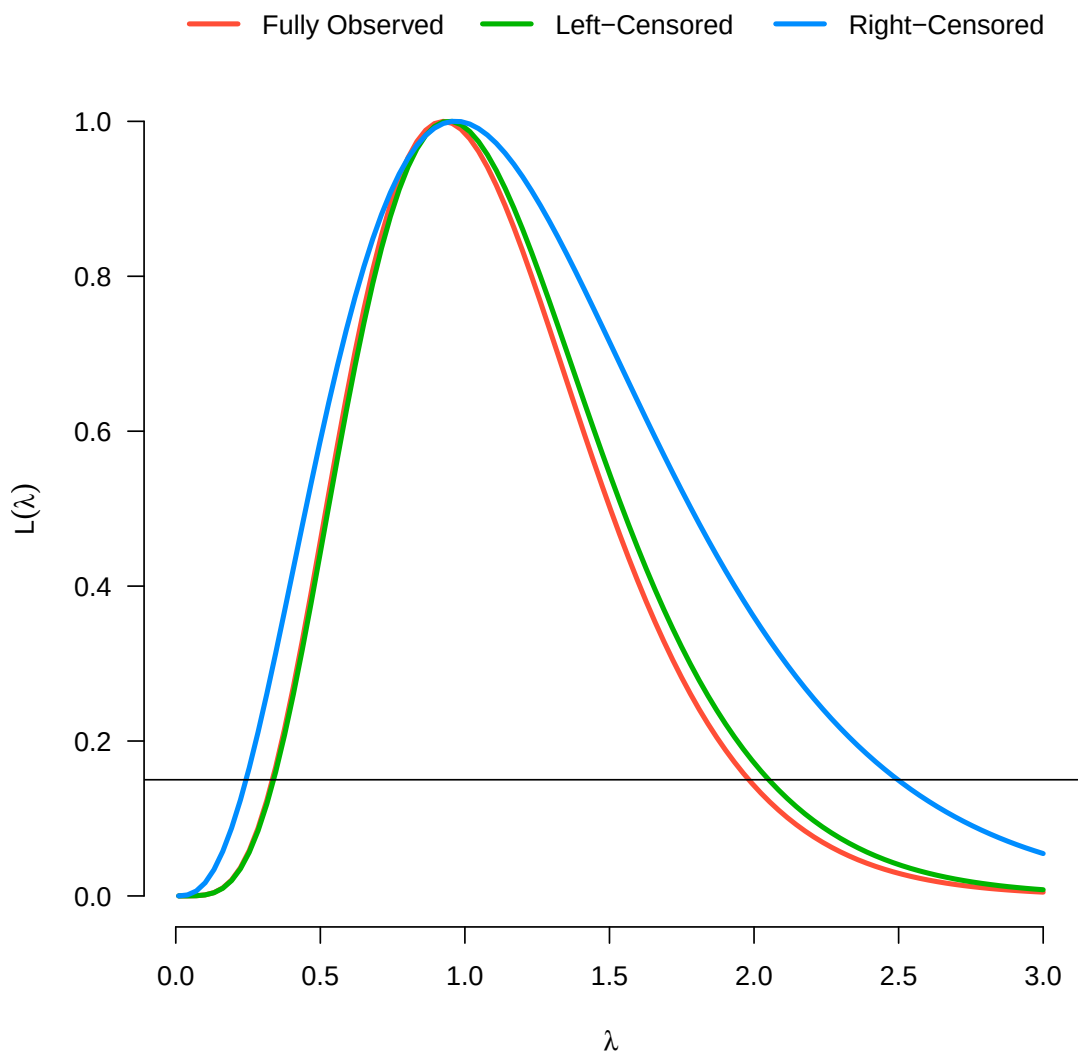
- Suppose that we start observation from time $t = 0.75$
- That is, left-censored
- In this case, the contribution to the likelihood from an left-censored observation at time t would be

$$L_i(\lambda) = F(t|\lambda) = 1 - \exp(-\lambda t)$$

- Therefore the likelihood becomes:

$$L(\lambda) = \prod_{i=1}^{3}(1 - \exp(-0.75 \times \lambda)) \prod_{i=4}^{5} \lambda \exp(-\lambda t_i)$$

Likelihood function for left-censored data



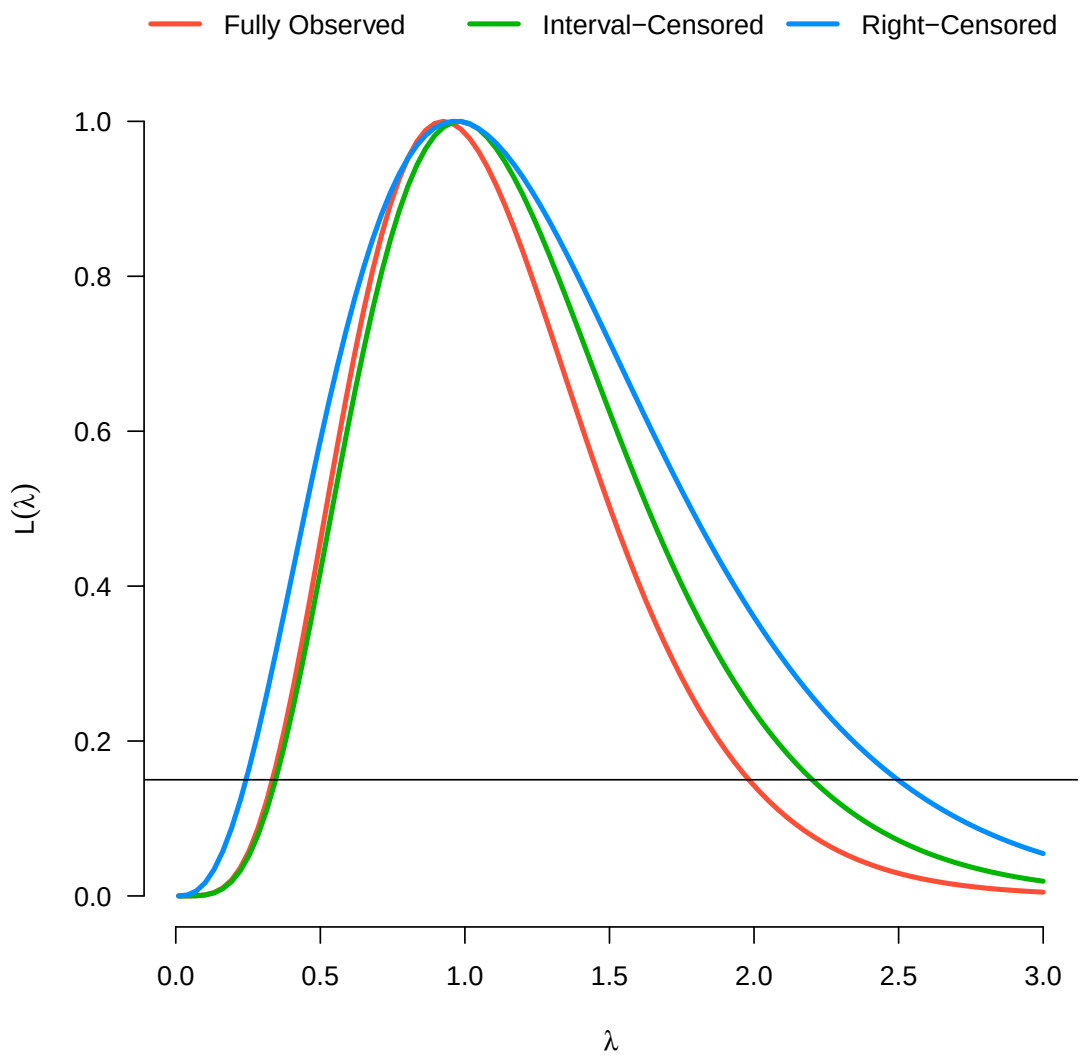
Interval Censored Data

- Interval censored means that for each time T, we only know an interval $[L, U]$ such that $L < T < U$
- In this case the contribution of the likelihood becomes

$$L_i(\lambda) = F(U|\lambda) - F(L|\lambda) = \exp(-\lambda L) - \exp(-\lambda U)$$

- In the previous example, suppose we only observe the times within intervals $[0, 1], [0, 1], [0, 1], [1, 2], [2, 3]$

Likelihood with interval censored data



Likelihood: Summary

Table : Likelihood function for non-censored and censored data

Type	T	L_i
Full observation	$T = t_i$	$f(t_i)$
Right-censored	$T > t_i$	$S(t_i)$
Left-censored	$T < t_i$	$F(t_i)$
Interval-censored	$l_i < T < u_i$	$F(u_i) - F(l_i)$

Next Section ...

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Fit Weibull Parametric Model

```
fit.weibull <- survreg(Surv(Time, Status==0) ~ TRT, data=carcinoma,
                       dist="weibull")
summary(fit.weibull)


##
## Call:
## survreg(formula = Surv(Time, Status == 0) ~ TRT, data = carcinoma,
##     dist = "weibull")
##               Value Std. Error     z      p
## (Intercept)   5.564      0.168 33.17 <2e-16
## TRT1         -0.301      0.200 -1.51 0.1312
## TRT2          0.310      0.237  1.31 0.1913
## Log(scale)   -0.627      0.234 -2.68 0.0074
##
## Scale= 0.534
##
## Weibull distribution
## Loglik(model)= -92.2   Loglik(intercept only)= -93.6
##  Chisq= 2.77 on 2 degrees of freedom, p= 0.25
## Number of Newton-Raphson Iterations: 5
## n= 31
```

Fit Exponential Model with Covariates

```
# fit exponential model +Age
fit.exp.age <- survreg(Surv(Time, Status==0) ~ TRT + Age,
                       data=carcinoma, dist="exponential")
summary(fit.exp.age)


##
## Call:
## survreg(formula = Surv(Time, Status == 0) ~ TRT + Age, data = carcinoma
##     dist = "exponential")
##               Value Std. Error     z       p
## (Intercept) 11.3781     2.2136  5.14 2.7e-07
## TRT1        -0.3221     0.3652 -0.88  0.3779
## TRT2         0.4113     0.4352  0.95  0.3446
## Age         -0.0966     0.0366 -2.64  0.0084
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -91   Loglik(intercept only)= -96
##  Chisq= 9.91 on 3 degrees of freedom, p= 0.019
## Number of Newton-Raphson Iterations: 5
## n= 31
```

Fit Weibull Model with Covariates

```
# fit Weibull model+Age
fit.weibull.age <- survreg(Surv(Time, Status==0) ~ TRT + Age,
                           data=carcinoma, dist="weibull")
summary(fit.weibull.age)


##
## Call:
## survreg(formula = Surv(Time, Status == 0) ~ TRT + Age, data = carcinoma
##     dist = "weibull")
##                Value Std. Error     z       p
## (Intercept)  8.7531     1.3216  6.62 3.5e-11
## TRT1        -0.1646     0.1801 -0.91  0.3609
## TRT2         0.2253     0.2144  1.05  0.2933
## Age         -0.0569     0.0217 -2.62  0.0088
## Log(scale)  -0.7294     0.2291 -3.18  0.0015
##
## Scale= 0.482
##
## Weibull distribution
## Loglik(model)= -87.2   Loglik(intercept only)= -93.6
##  Chisq= 12.8 on 3 degrees of freedom, p= 0.0051
## Number of Newton-Raphson Iterations: 7
## n= 31
```

Next Section ...

1. Survival concepts
2. Parametric models
3. Likelihood Functions
4. Parametric Regression Method
5. Nonparametric Method
   - Life-table method
   - Kaplan-Meier method
   - Nonparametric test

Life table method

- There are no changes in survivorship over calendar time
- The experience of individuals who are lost to follow-up is the same as the experience of those who are followed.
- Withdrawal occurs uniformly within the interval.
- Event occurs uniformly within the interval.

Ordered failure times ($t_{(i)}$)	failures (d_i)	censored (q_i)	Risk set $R(t_{(i)})$
$t_{(0)} = 0$	m_0	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(n)}$	m_n	q_n	$R(t_{(n)})$

Noncensored Survival Data Estimator

We can estimate $S(t)$ from a sample of n observations as

$$\hat{S}(t) = \frac{\text{number of patients with observed times} \geq t}{n}$$

when there are no censored survival time in the sample.

The variance estimate can be obtained by

$$Var\left[\hat{S}(t)\right] = \frac{\hat{S}(t)[1 - \hat{S}(t)]}{n}$$

The estimation of survival function

Let $t_1 < t_2 < \cdots < t_D$ represent the distinct event times, and let n_j be the size of the risk set prior to t_j, and d_j be the number of failures at t_j.

- **Breslow estimate** of the survival funciton:

$$\hat{S}(t_j) = \exp\left(-\sum_{i=1}^{j} \frac{d_i}{n_i}\right)$$

  where $\sum_{i=1}^{j} d_i/n_i$ is the Nelson-Aalen estimate of the cumulative hazard function $\hat{H}(t_j)$.
- **Fleming-Harrington** estimate of the survival function:

$$\hat{S}(t_j) = \exp\left(-\sum_{k=1}^{j}\sum_{i=0}^{d_k-1} \frac{1}{n_k - i}\right)$$

  only for integer.
- **Kaplan-Meier product limit** estimate of the survival function:

$$\hat{S}(t_j) = \prod_{i=1}^{j}\left(1 - \frac{d_i}{n_i}\right)$$

- **Greenwood estimate** of the standard error:

Kaplan-Meier Estimator for Censored Survival Data

When there are censored observations, the survival function is estimated by the Kaplan-Meier estiator:

(1) Sort the observed survival times from smallest to the largest as $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ where $t_{(j)}$ is the j^{th} ordered survial time.

(2) Compute the survival function by

$$\hat{S}(t) = \prod_{R(j)} \left(1 - \frac{e_j}{r_j}\right)$$

where

- r_j is the dimension of the risk set of $R(j) = \{j : t_{(j)} \leq t\}$
- e_j is the number of patients with events at time t_j.

(3) Estimate the variance of $\hat{S}(t)$

$$Var\left[\hat{S}(t)\right] = \left[\hat{S}(t)\right]^2 \prod_{R(j)} \frac{e_j}{r_j(r_j - e_j)}$$

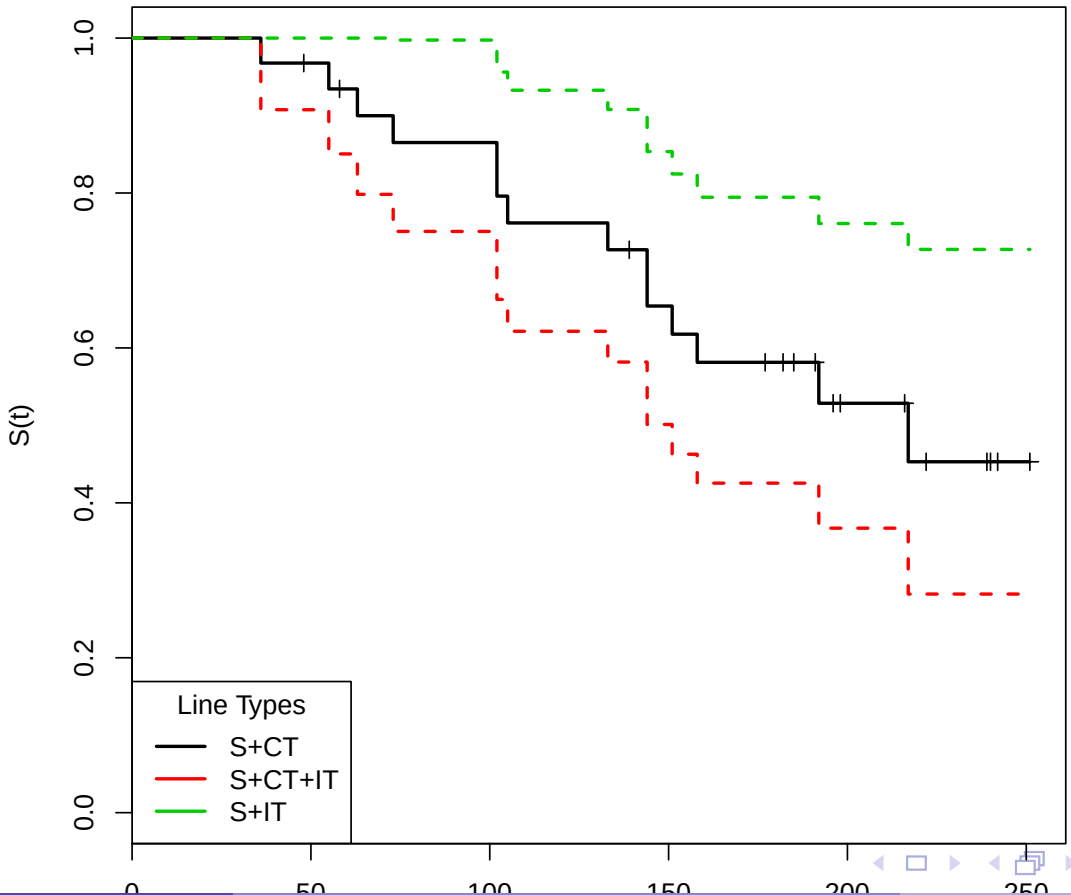
Fit Kaplan-Meier

```
library(survival)
# fit Kaplan-Meier
fit.km <- survfit(Surv(Time, Status==0) ~ 1,
                  type="kaplan-meier", data=carcinoma)
# print the model fitting
fit.km


## Call: survfit(formula = Surv(Time, Status == 0) ~ 1, data = carcinoma,
##     type = "kaplan-meier")
##
##       n  events  median 0.95LCL 0.95UCL
##      31      14     217     151      NA
```

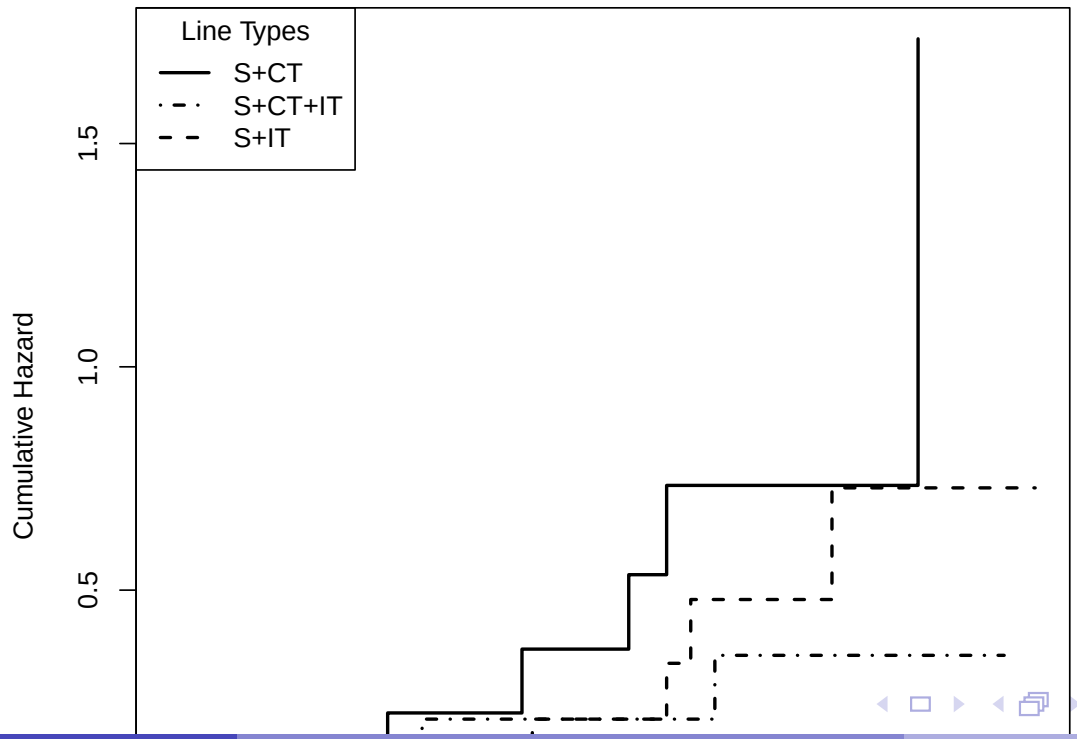
Kaplan-Meier Survival Curve

```
plot(fit.km, col=1:3, lwd=2, mark.time=TRUE,
        xlab="Time in Weeks", ylab="S(t)")
legend("bottomleft", title="Line Types",
        c("S+CT", "S+CT+IT", "S+IT"), col=1:3, lwd=2)
```



Estimating the cumulative hazard function

```
fit.fleming <- survfit(Surv(Time, Status==0) ~ TRT,
                       data=carcinoma, type="fleming")
# Plot the estimated cumulative hazard function
plot(fit.fleming, lty=c(1,4,8), lwd=2, fun="cumhaz",
        xlab="Time in Weeks", ylab="Cumulative Hazard")
legend("topleft", title="Line Types",
        c("S+CT", "S+CT+IT", "S+IT"), lty=c(1,4,8), lwd=2)
```



Statistical testing of the survival times

- Log-rank test
- Gehan-Breslow Wilcoxon's test
- Peto-Prentice Wilcoxon's test

Log-rank test (1)

Comparing two groups

Suppose that at time t_j

	Group-1	Group-2	Total
Failures	d_{1j}	d_{2j}	d_j
Survivors	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
Risk set	n_{1j}	n_{2j}	n_j

Log-rank test (2)

To test

$$H_0 : S_1 = S_2$$

Under H_0, the random variable D_{1j} follows a hypergeometric distribution with

$$\text{mean } e_{1j} = n_{1j}\frac{d_j}{n_j}, \text{variance } v_{1j} = n_{1j}\left(\frac{d_j}{n_j}\right)\left(\frac{n_j - d_j}{n_j}\right)\left(\frac{n_{2j}}{n_j - 1}\right)$$

with this, we can get an approximate normal distribution under H_0:

$$w_j = d_{1j} - e_{1j} \overset{d}{\sim} N(0, v_j = v_{1j})$$

Therefore,

$$W \sim N(0, V)$$

where $W = \sum_j w_j$ and $V = \sum_j v_j$, provided that failures are conditionally independent.

Or equivalently,

$$\frac{W^2}{V} = \frac{(\sum_j w_j)^2}{\sum_j v_j} \sim \chi_1^2$$

General log-rank test

Comparing multiple groups

Assume that we have $(K+1)$ groups, and use the $(K+1)$-th group as the reference,

$$\begin{aligned} w_j &= (d_{1j} - e_{1j}, \ldots, d_{Kj} - e_{Kj}) \\ (V_j)_{ik} &= -\frac{n_{ij} n_{kj} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \\ W &= \sum_j w_j \\ V &= \sum_j V_j \\ W V^{-1} W^T &\sim \chi^2_K \end{aligned}$$

Note that the degrees of freedom K is the number of groups minus 1.

Weighted log-rank test

We can extend the log-rank test to weighted log-rank test:

$$\frac{(\sum_j \alpha_j w_j)^2}{\sum_j \alpha_j^2 v_j}$$

where $\{\alpha_j\}$ are weights chosen to emphasize or deemphasize various time points.

- **Gehan-Breslow test**: weighting by the number at risk at time t_j:

  $$\alpha_j = n_j$$

  which is also known as *Gehan test*.
- **Peto-Prentice test**: weighting by the pooled survival estimate:

  $$\alpha_j = \hat{S}(t_j)$$

  which is also known as *Peto-Peto test*.
- Both tests place more emphasis on earlier failures compared to the log-rank test.

Logrank test: Summary

- The above test is known as the **log-rank test**
- The idea behind the test is essentially the same as that of the Cochran-Mantel-Haenszel test in categorical data analysis, with time as the stratification variable
- The *log-rank test* is the most widely used test for comparing 2+ survival time distributions, in part due to the simple "observed - expected" form.
- The log-rank test is particularly powerful when the ratio between two hazard functions being compared is constant over time.

Logrank test of the survival time

```
fit.diff <- survdiff(Surv(Time, Status==0) ~ TRT, data=carcinoma, rho=0)
# print the result
fit.diff


## Call:
## survdiff(formula = Surv(Time, Status == 0) ~ TRT, data = carcinoma,
##     rho = 0)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## TRT=S+CT      11        6     3.64   1.52842   2.12654
## TRT=S+CT+IT 10        3     5.19   0.92444   1.51837
## TRT=S+IT      10        5     5.17   0.00549   0.00887
##
##  Chisq= 2.5  on 2 degrees of freedom, p= 0.3
```