Math 659: Survival Analysis

Chapter 3 – Censoring and Truncation

Wenge Guo

September 14, 2011

Introduction on Censoring

- One special feature of time-to-event data is censoring
- Censoring means some lifetimes are known to have occurred only with certain intervals, while the rest of the lifetime are known exactly
- There are various forms of censoring, such as, right censoring, left censoring, interval censoring, or a combination of these
- Why censoring, for right censoring, is that one could not wait until the interested event occurs for all subjects under study, due to limited time, or limited budget, ...

Introduction on Truncation

- Another feature is the present of truncation in some study
- Truncation occurs when only those individuals whose event time lies with a certain observational window (Y_L, Y_R) are observed
- Truncation could be left or right
- Must use the conditional distribution because we observe $X = x$ is conditional on that $Y_L \leq X \leq Y_R$
- Different form of censoring and truncation have different likelihood constructions

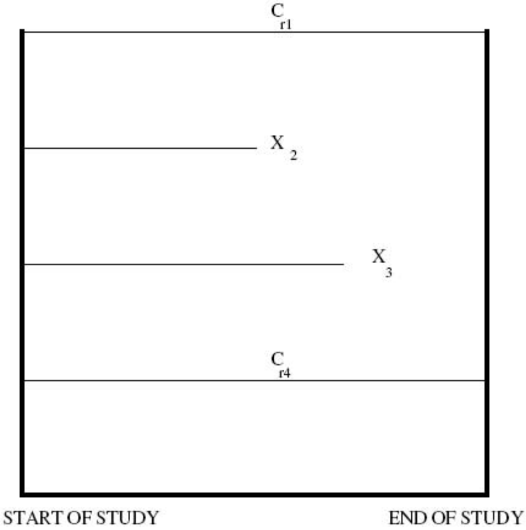
Right Censoring

- Denote the lifetime by X and the censoring time by C_r, we only observe $T = \min\{X, C_r\}$
- The time-to-event data are often denoted by (T, δ) where δ is called the censoring indicator
- In particular, $\delta = 1$ if $T = X$ and $\delta = 0$ if $T = C_r$

Type I Censoring

- For Type I censoring (time censoring), the event is observed only if it occurs some prespecified time
- That is the censoring time C_r is non-random
- The censoring time could vary from individual to individual
- Example:
    - A large scale animal experiment conducted at the National Center for Toxicological Research in which mice were fed a particular dose of carcinogen
    - To study the effect of carcinogen on the survival
    - Carcinogen refers to any substance, radionuclide or radiation that is an agent directly involved in the promotion of cancer or in the increase of its propagation
    - Mice were followed from the beginning of the experiment until death or until a prespecified censoring time was reached

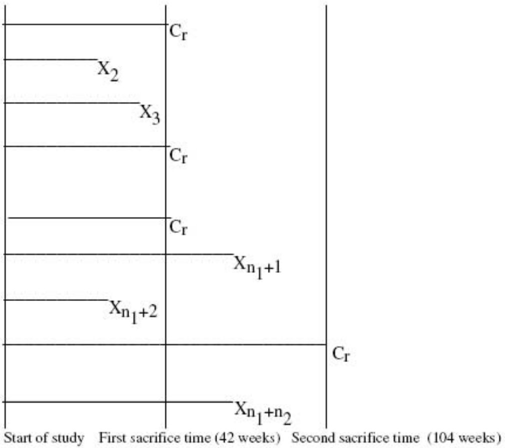
Example of Type I Censoring



Progressive Type I Censoring

- When animals have different prespecified follow-up times, this form of Type I censoring is called progressive Type I censoring
- Example:
  - Consider a mouse study where, for each sex, 200 mice were randomly divided into four dose-level groups
  - Each mouse was followed until death or until a prespecified time (42 weeks or 104 weeks) was reach

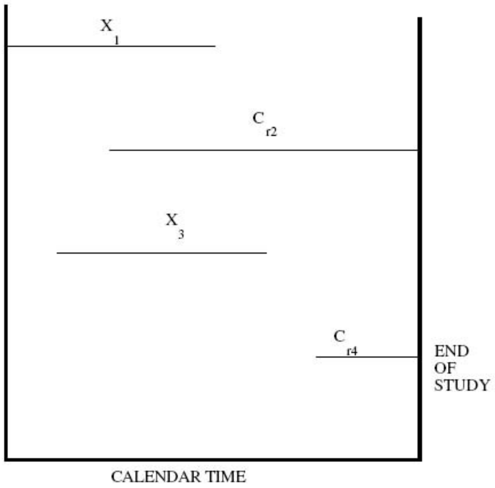
Example of Progressive Type I Censoring



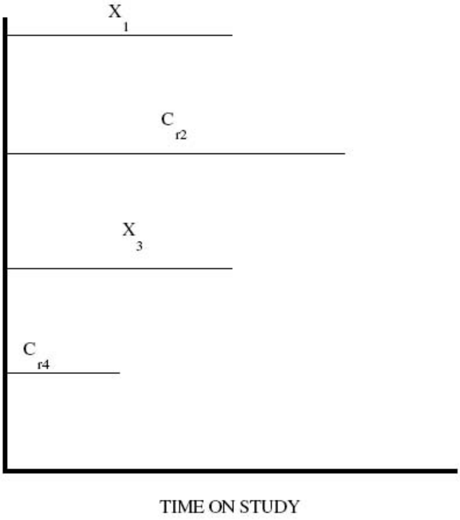
Generalized Type I Censoring

- Patients enter into the study at different time and the terminal point of the study is predetermined by the investigator or the study protocol
- Each individual has their own censoring time which is determined by the study terminal point minus entry point
- This type of censoring is called generalized Type I censoring

Example of Generalized Type I Censoring



Time on Study



Type II Censoring

- It is possible that a study using Type I censoring yields no events (failures) or very few events that contains very little information
- An alternative strategy is to stop the study until r number of events were observed
- This type of censoring is called Type II censoring (failure censoring). Type II censoring can save some money and get enough information needed
- Type II censoring is rare in medical studies
- More popular in engineering studies

Random Censoring

- Random censoring occurs when C_r is also a random variable
- $T = \min\{X, C_r\}$, $\delta = 1$ if $X \le C_r$, $\delta = 0$ otherwise
- Example
  - X is the death time due to a particular cancer (interested event)
  - C_r is the death time due to other causes (heart disease, accidents, etc.)
  - Only observe X if $X \le C_r$
  - X is randomly censored by C_r

Left Censoring

- A lifetime X is left censored if it is less than a censoring time C_l
- The event of interest has already occurred for the individual before that person is observed in the study at time C_l
- We know the event occurred at some time point before C_l, but don't know it exactly
- $T = \max\{X, C_l\}$
- Example:
  - A study was done to determine the distribution of the time until first marijuana use among high school boys in California
  - The question was asked "When did you first use marijuana?"
  - One response was "I have used it but can not recall when" (left)
  - "I never used it" (right)
  - "When I was at 15" (exact event time)

Interval Censoring

- Interval censoring: lifetime only known to occur within an interval
- In a clinical trial or longitudinal study have periodic follow-up and the patient's event time is only known to fall in an interval $(L_i, R_i]$
- In industrial experiments, periodic inspection for the proper functioning of equipment items
- Data precision: a high-voltage power transformer failed in 1988 means it failed some where between 01/01/1988 and 12/31/1988

Truncation

- Only those individual whose event time lies within a certain observational window (Y_L, Y_R) are observed
- An individual whose event time is not in this interval is not observed and no information on this subject is available
- For censoring, we have at leat have partial information on each subject
- The inference for truncated data is restricted to conditional estimation

Left Truncation

- When $Y_R = \infty$, that is the observational window is (Y_L, ∞), we have left truncation
- We only observe X if and only if $X > Y_L$
- Example 1:
  - In the problem of estimating the distribution of the diameter of microscopic particles
  - Only particles big enough to be seen based on the resolution of the microscope are observed
  - Small particles do not come to the attention of the investigator
- Example 2:
  - A survival study of residents of the Channing House retirement center located in California
  - Ages at death are recorded
  - An individual must survive to a sufficient age to enter the retirement center, so there is left truncation

Right Truncation

- Right truncation occurs when Y_L is equal to 0
- We observe the survival time only when $X \leq Y_R$
- Example 1:
  - In estimating the distribution of stars from the earth
  - Stars too far away are not visible and are right truncated
- Example 2:
  - Example 1.19 [KM], Lagakos et al. (1988) report data on the infection and induction times for 258 adults and 37 children who were infected with the AIDS virus and developed AIDS by June 30, 1986
  - Those infected but not developed AIDS by June 30, 1986 were not report, thus were right truncated

Likelihood Construction

- The likelihood function can be viewed as the probability of observing the data (or a function that is proportional to the probability of the data)
- The likelihood is written as a function of the unknown model parameters
- Denote the unknown parameters by θ and suppose there is n observations, assuming all observations are independent
- The likelihood is

$$L(\theta|DATA) = \prod_{i=1}^{n} L_i(\theta)$$

- $L_i(\theta)$ is the likelihood contribution from observation i

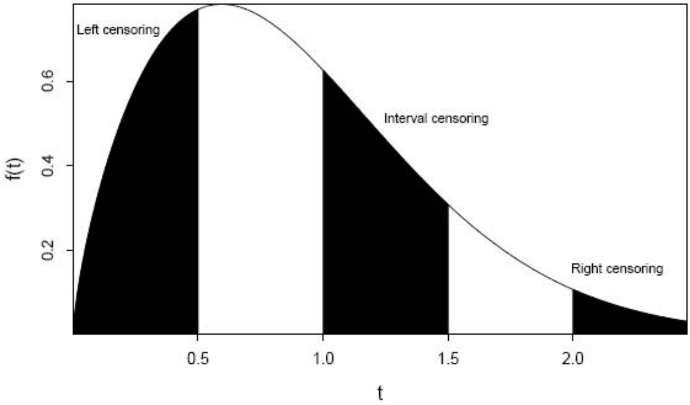
Likelihood Contribution for Exact Lifetimes

- The likelihood contributions for various types of censoring schemes are different
- Exact lifetimes
  - What is the probability of observing an exact lifetime?
  - Is it $\Pr(X = x) = 0$ for continuous distribution?
  - The recorded data are always discrete due to rounding
  - For example, $x = 3.1$ could mean x is any number between 3.1 and 3.1999999999....
  - The correct probability of observing an exact lifetime is $\Pr(x \leq X \leq x + \Delta x) = \int_x^{x+\Delta x} f(t)dt \approx f(x)\Delta x$
  - In practice, $f(x)$ is often used as the likelihood contribution for exact lifetimes

Likelihood Contributions

- For exact lifetimes: $f(x)$
- For right-censored observations: $S(C_r)$
- For left-censored observations: $1 - S(C_l)$
- For interval-censored observations: $[S(L) - S(R)]$
- For left-truncated observations: $\frac{f(x)}{S(Y_L)}$
- For right-truncated observations: $\frac{f(x)}{[1-S(Y_R)]}$
- For interval-truncated observations: $\frac{f(x)}{[S(Y_L)-S(Y_R)]}$

Illustration of Likelihood Contributions



Likelihood Function for Type I Censoring

- The data are represented by (T, δ)
- $T = \min\{X, C_r\}$ and δ is called the censoring indicator
- In particular, $\delta = 1$ if $T = X$ and $\delta = 0$ if $T = C_r$
- $\Pr(T, \delta = 0)$
- $\Pr(T, \delta = 1)$
- $\Pr(t, \delta) = [f(t)]^{\delta}[S(t)]^{1-\delta}$
- If we have n observations (T_i, δ_i), the likelihood function is $L = \prod_{i=1}^{n} \Pr(t_i, \delta_i) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i}$

Likelihood Function for Random Censoring

- The data are represented by (T, δ)
- $T = \min\{X, C_r\}$ and C_r is with pdf $g(c_r)$ and cdf $G(c_r)$
- X and C_r are independent
- $\Pr(T, \delta = 0) =$

Assumption of Censoring Distribution

- $\Pr(T, \delta = 1) =$

- The likelihood is $L =$

- The assumption is censoring should be noninformative, that is the censoring times of units provide no information about the failure-time distribution

Numerical Example

- Lifetime data from Example 3.8 of Meeker and Escobar (1998), *Statistical Methods for Reliability Data*, Wiley
- The data give the distance (in km) to failure for 38 vehicle shock absorbers, with 11 failures
- The data were recorded as: 6700, 6950+, 7820+, 8790+, 9120, 9660+, 9820+, 11310+, 11690+, 11850+, 11880+, 12140+, 12200, 12870+, 13150, 13330+, 13470+, 14040+, 14300, 17520, 17540+, 17890+, 18450+, 18960+, 18980+, 19410+, 20100, 20100+, 20150+, 20320+, 20900, 22700, 23490+, 26510, 27410+, 27490, 27890+, 28100+

Example: Shock Absorber Data

- The Weibull cdf and pdf can expressed as $F(t;\mu,\sigma) = \Phi_{\text{sev}}\left[\frac{\log(t)-\mu}{\sigma}\right]$ and $f(t;\mu,\sigma) = \frac{1}{t\sigma}\phi_{\text{sev}}\left[\frac{\log(t)-\mu}{\sigma}\right]$ with $\Phi_{\text{sev}}(z) = 1-\exp(-\exp(z)), \phi_{\text{sev}}(z) = \exp(z-\exp(z))$
- $\mu = -\frac{\ln(\lambda)}{\alpha}$, $\sigma = \frac{1}{\alpha}$, λ, α is the parameter used in $S(t) = \exp(-\lambda t^{\alpha})$
- The data are denoted by (t_i, δ_i)
- The likelihood is $L(\mu,\sigma|DATA) = \prod_{i=1}^{n} F(t_i;\mu,\sigma)^{\delta_i}[1-F(t_i;\mu,\sigma)]^{1-\delta_i}$
- The MLE can be obtained by finding values of μ, σ that maximize $L(\mu,\sigma|DATA)$

Use of `survreg` in `survival` Package

- `survreg(formula = formula(data), data = yourdata, dist = "weibull")`
-
```
> fit=survreg(Surv(distance, failed)~1,
data=shock.absorder.dat, dist='weibull')
> summary(fit)
Call:
survreg(formula = Surv(distance, failed) ~ 1,
data = shock.absorder.dat,
    dist = "weibull")
             Value Std. Error     z        p
(Intercept) 10.23      0.110 93.09 0.00e+00
Log(scale)  -1.15      0.231 -4.98 6.48e-07
Scale= 0.316
Weibull distribution
```
- Here $\widehat{\mu} = 10.23, \widehat{\sigma} = 0.316$