

# Statistics for Biology and Health

## Chapter 2 Censoring and Truncation

Qi Guo

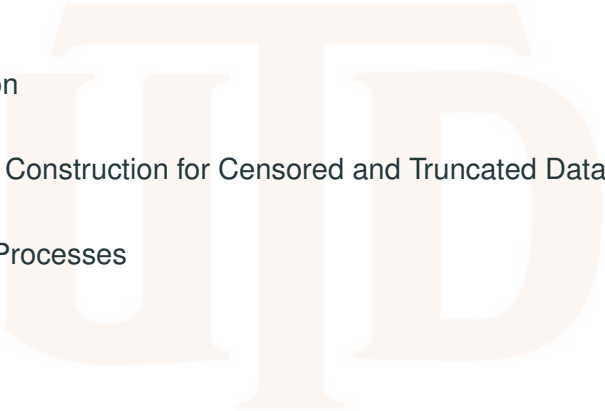
---

July 19,2019



THE UNIVERSITY OF TEXAS AT DALLAS

School of Natural Sciences and Mathematics

- 
1. Introduction
  2. Likelihood Construction for Censored and Truncated Data
  3. Counting Processes

# Introduction

---

- Time-to-event data present themselves in different ways which create special problems in analyzing such data. There are two main features, one is known as censoring, occurs when some lifetimes are known to have occurred only within certain intervals, and the other one is truncation.
- About the censoring and truncation, more detail information are in Chapter 1 and notes before.

# Notations

- It's convention that random variables are denoted by upper case letters and fixed quantities or realizations of random variables are denoted by lower case letters.
- About censoring,  $C_r$  is for a fixed right censoring time;  $C_l$  is for a fixed left censoring time.
- About truncation, survival data occurs when only those individuals whose event time lies within a certain observational window ( $Y_L, Y_R$ ) are observed. When  $Y_R$  is infinite then it's left truncation, and right truncation occurs when  $Y_L$  is equal to zero.

# **Likelihood Construction for Censored and Truncated Data**

---

# Likelihood for Censored and Truncated Data

- A critical assumption is that the lifetimes and censoring times are independent.
- The likelihoods for various types of censoring schemes may all be written by incorporating the following components:
  - exact lifetimes -  $f(x)$
  - right-censored observations -  $S(C_r)$
  - left-censored observations -  $1 - S(C_l)$
  - interval-censored observations -  $[S(L) - S(R)]$
  - left-truncated observations -  $f(x)/S(Y_L)$
  - right-truncated observations -  $f(x)/[1 - S(Y_R)]$
  - interval-truncated observations -  $f(x)/[S(Y_L) - S(Y_R)]$

# Likelihood for Censored and Truncated Data

- The likelihood function may be constructed by putting together the component parts as:

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_r) \prod_{i \in L} (1 - S(C_l)) \prod_{i \in I} [S(L_i) - S(R_i)]$$

where  $D$  is the set of death times,  $R$  is the set of right-censored observations,  $L$  is the set of left-censored observations, and  $I$  is the set of interval- censored observations.



# Counting Processes

---

# Counting Processes

- An alternative approach to developing inference procedures for censored and truncated data is by using counting process methodology.
- Define a counting processes  $N(t)$ ,  $t \geq 0$ , as a stochastic process with the properties that  $N(0)$  is zero;  $N(t) < \infty$ , with probability 1.
- $N(t)$  are right-continuous and piecewise constant with jumps of size +1.
- Given a right-censored sample,  $N_i(t) = I[T_i \leq t, S_i = 1]$ , which are zero until individual  $i$  dies and then jump to 1, are counting process.
- The process  $N(t) = \sum_{i=1}^n N_i(t) = \sum_{t_i \leq t} \delta_i$  counts the number of deaths in the sample at or prior to time  $t$  is another form of counting process.

# History

- The accumulated knowledge about what has happened to patients up to time  $t$  is called the history or filtration of the counting process at time  $t$  and is denoted by  $F_t$ .
- And we know that  $F_s \subset F_t$  when  $s \leq t$ .
- For right-censored data, the history at time  $t$ ,  $F_t$ , consists of knowledge of the pairs  $(T_i, \delta_i)$  provided  $T_i \leq t$  and  $T_i > t$  for those individuals still under study at time  $t$ , and denote the history at an instant just prior to time  $t$  by  $F_{t-}$ .
- If we define the process  $Y(t)$  as number of individuals with a study time  $T_i \geq t$ , then,

$$E[dN(t)|F_{t-}] = Y(t)h(t)dt$$

## The intensity process

- The process  $\lambda(t) = Y(t)h(t)$  is called the intensity process of the counting process.  $\lambda(t)$  is itself a stochastic process that depends on the information contained in the history process.
- The stochastic process  $Y(t)$  is the process which provides us with the number of individuals at risk at a given time and, along with  $N(t)$ .
- And define the process  $\Lambda(t) = \int_0^t \lambda(s)ds$ ,  $t \geq 0$  as the cumulative intensity process, has the property that  $E[N(t)|F_{t-}] = E[\Lambda(t)|F_{t-}] = \Lambda(t)$ .
- The stochastic process  $M(t) = N(t) - \Lambda(t)$  is called the counting process martingale, it has the property that increments of this process have an expected value, given the strict past,  $F_{t-}$ , that are zero.

# Martingale

- A stochastic process with the property that its expected value at time  $t$ , given its history at time  $s < t$ , is equal to its value at time  $s$  is called a martingale, that is,  $M(t)$  is a martingale if

$$E[M(t)|F_s] = M(s), \text{ for all } s < t$$

- The counting process martingale  $M(t) = N(t) - \Lambda(t)$  is made up of two parts.  $N(t)$  is a nondecreasing step function, and  $\Lambda(t)$  is a smooth process which is predictable in that its value at time  $t$  is fixed just prior to time  $t$ . This random function is called a compensator of the counting process.
- The martingale can be considered as mean zero noise which arises when we subtract the smoothly varying compensator from the counting process.

## The predictable variation process

- The predictable variation process of  $M(t)$ , denoted by  $\langle M \rangle(t)$ , is defined as the compensator of the process  $M^2(t)$ .
- Although  $M(t)$  reflects the noise left after subtracting the compensator,  $M^2(t)$  tends to increase with time, so here  $M(t)$  is the systematic part of this increase and is the predictable process needed to be subtracted from  $M^2(t)$  to produce a martingale.
- So we introduce the predictable variation process:  
$$\text{var}(dM(t)|F_{t-}) = d\langle M \rangle(t).$$
- To find the variance recall that  $dN(t)$  is a zero-one random variable with a probability, given the history, of time  $\lambda(t)$  of having a jump of size one at time  $t$ . The variance of such a random variable is  $\lambda(t)[1 - \lambda(t)]$ .
- If there are no ties in the censored data case,  $\lambda(t)^2$  is close to zero so that  $\text{Var}(dM(t)|F_{t-}) \cong \lambda(t) = Y(t)h(t)$ .