Chapter 2: Survival Function, Censoring and Truncation

Chien-Fu Jeff Lin, MD. PhD.

cflin@mail.ntpu.edu.tw

http://web.ntpu.edu.tw/~cflin

Department of Statistics
National Taipei University, Taipei, Taiwan

Spring 2007

Outline I

1. Basic Survival Function and Hazard Function
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation

Outline II

Functions Characterize Lifetime Random Variable X I

1. Each member of the population has a survival time, X.
2. Let X denote the lifetime for a random member of the population.
3. Several functions are commonly used to characterize the distribution of values of X in the population:

Functions Characterize Lifetime Random Variable X II

1. Survival function, $S(x)$.
2. (Cumulative) Distribution function, $F(x)$.
3. Density function, $f(x)$.
4. Hazard function, $h(x)$.
5. Cumulative hazard function, $H(x)$.
6. Mean residual life, $\mathbf{mrl}(x)$.

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Random Variable, X, of Survival Time I

1. Let X be the time from a well-defined time point zero to a well-defined time point when some specified event occurs.
2. A single nonnegative random variable, X.
3. Let $X \geq 0$ and $f(X)$ be the probability density (mass) function.

Definition 1.1 (**Probability Density Function)**

Probability density function (p.d.f) of X is

$$f(X=x)=\lim_{\Delta x\to 0}\frac{Pr(x\le X<x+\Delta x)}{\Delta x}=\frac{dF(x)}{dx} \tag{1.1}$$

with $\int_0^\infty f(x)dx=1$ and $x\in[0,\infty)$.

The range of X is $[0,\infty)$, and this should be understood as the domain of definition for function of x.

Definition 1.2 (**Survival Function**)

Survival function is the probability of an individual surviving beyond time x (experiencing the event after time x).

$$\begin{aligned} S(x) &= Pr(X > x) && (1.2)\\ &= \int_x^\infty f(t)\,dt; \quad \text{for a continuous random variable.} && (1.3) \end{aligned}$$

In the context of equipment item failures, $S(x)$ is referred to as the **reliability function.**

Survival Function I

1. $S(x_1) - S(x_2)$ is the fraction of the population that dies between ages x_1 and x_2 for $x_1 < x_2$.
2. Survival functions are monotone, decreasing (nonincreasing) functions equal to one at zero and zero at the time approaches infinity.
3. If $x_1 < x_2$, then $S(x_1) > S(x_2)$.
4. $S(0) = 1$.

Survival Function II

5. If the every member of the population eventually has an event, then $S(\infty) = 0$.
6. If some member of the population never have the event, then it is possible that the survival curve does not approach 0 as time increase.

Survival Function III

7. The notation dealing with this is not standardized, but one practical implication is that a survival curve estimate need not reach 0 by the end of follow-up.
8. When X is a continuous random variable, the survival function is the complement of the cumulative distribution function.

In practice, survival curves are often shown at discrete times (ages), as in the following Table 1.

Table 1: Male survival 1990

Age range				Percent Alive at Start
0	to	$<$	5	100.00
5	to	$<$	10	98.90
10	to	$<$	20	98.78
20	to	$<$	30	98.06
30	to	$<$	40	96.56
40	to	$<$	50	94.56
50	to	$<$	60	91.15
60	to	$<$	70	83.14
70	to	$<$	80	66.17
80 plus				38.36

Definition 1.3 (**Failure Function)**

Failure function is the cumulative distribution.

$$F(x) = Pr(X \leq x) = 1 - S(x) \tag{1.4}$$

Thus,

$$f(x) = \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x)}{\Delta x} = \frac{dF(x)}{dx} \tag{1.5}$$

$$= -\frac{dS(x)}{dx} = \lim_{\Delta x \to 0} \frac{S(x) - S(x + \Delta x)}{\Delta x} \tag{1.6}$$

Failure Function I

1. $f(x)dx \approx$ fraction who die between age x and $x + \Delta x$ when Δx is a short interval of time.
2. The density is positive. $F(x) = int_0^x f(t)dt$, and $S(x) = \int_x^\infty f(t)dt$.
3. When we use the density function, we generally write it as a function of parameters and require $\int_0^\infty f(t)dt = 1$.

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Hazard Function I

A fundamental in survival analysis is the **hazard function.**

Hazard Function is Known As: I

1. The **hazard rate** in survival analysis
2. The **conditional failure rate** in reliability
3. The **force mortality** in demography
4. The **intensity function** in stochastic process
5. The **age-specific failure rate** in epidemiology
6. The **inverse of Mill's ratio** in economics

Definition 1.4 (**Hazard Function (Hazard Rate)**)

Hazard function (hazard rate) is conditional probability that specifies the instantaneous rate of failure at $X = x$ conditional upon survival to time x, and is defined as

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{Pr(x \leq X < x + \Delta x \mid X \geq x)}{\Delta x} && (1.7) \\ &= \frac{f(x)}{S(x)} = -\frac{d}{dx} \log[S(x)] && (1.8) \\ f(x) &= h(x)S(x) && (1.9) \end{aligned}$$

Hazard Function (Hazard Rate) I

1. Note that **death rates** are generally reported among those still surviving, and are the same as the hazard function.
2. The concept of the hazard function has been discovered in many field has many names.
3. This function is known as conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic process, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate.

Definition 1.5 (**Cumulative Hazard Function)**

Cumulative hazard function is defined

$$
\begin{aligned}
H(x) &= \int_0^x h(u)\, du && (1.10) \\
&= -\log[S(x)] && (1.11)
\end{aligned}
$$

Survival Function and Hazard Function I

For continuous survival time,

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)\ du\right] \tag{1.12}$$

Hazard Function I

1. Hazard function is particularly useful in determining the appropriate failure distribution utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time.
2. There are many general shapes for the hazard rate.

Hazard Function II

3. The only restriction on $h(x)$ is that it be nonnegative, i.e.,

$$h(x) \geq 0. \tag{1.13}$$

4. One may believe that the hazard rate for the occurrence of a particular event is increasing, decreasing, constant, bathtub-shaped, hump-shaped or possessing some other characteristic which describes the failure mechanism, ash shown in Figure 1.

Hazard Function III

5. $H(x)$ is the expect number of events when following a single person to time x, with replacement at death.
6. It is easy to estimate $S(x)$.
7. This makes it easy to examine the shape of $H(x)$ graphically, which tells us about the hazard function as the slope of $H(x)$.

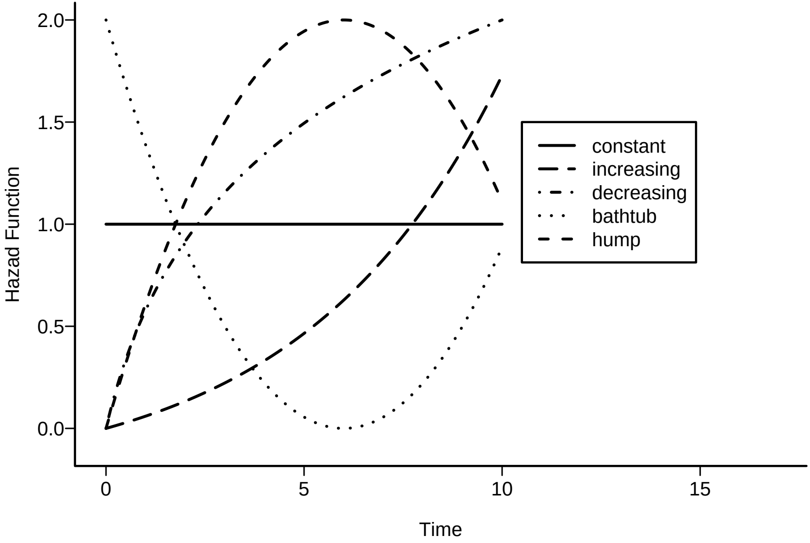

Figure 1: Various Hazard Function

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Discrete Random Variable I

1. When X is a discrete random variable, different techniques are required.
2. Now, let X is a discrete random variable.

Definition 1.6 (**Discrete Probability Mass Function)**

Suppose that X take on values $x_j, j = 1, 2, \ldots, n$ with **probability mass function (p.m.f)**,

$$p(x_j) = Pr(X = x_j), \tag{1.14}$$

where $x_1 < x_2 < \ldots < x_n$.

Definition 1.7 (**Survival Function)**

Survival Function for a discrete random variable X is

$$S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j) \tag{1.15}$$

where $S(0) = 1$ and $p(x_j) = S(x_{j-1}) - S(x_j)$.

Definition 1.8 (**Discrete Hazard Function)**

Discrete hazard function for a discrete random variable is defined as

$$h(x) = Pr(X = x_j \mid X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})} \quad j = 1, 2, \ldots \qquad (1.16)$$

Discrete Survival and Hazard Function I

$$h(x_j) = \frac{S(x_{j-1}) - S(x_j)}{S(x_{j-1})} = 1 - \frac{S(x_j)}{S(x_{j-1})} \tag{1.17}$$

$$S(x_{j-1}) \times h(x_j) = S(x_{j-1}) - S(x_j) \tag{1.18}$$

$$S(x_j) = S(x_{j-1})[1 - h(x_j)] \tag{1.19}$$

Discrete Survival and Hazard Function I

Thus, for discrete survival time, the survival function is the product of conditional survival probability as

$$
\begin{aligned}
S(x) &= \sum_{x_j > x} p(x_j) && (1.20)\\
&= \prod_{x_j \le x} \left[1 - h(x_j)\right] && (1.21)\\
&= \prod_{x_j \le x} \frac{S(x_j)}{S(x_{j-1})} && (1.22)
\end{aligned}
$$

Definition 1.9 (**Discrete Cumulative Hazard Function)**

The cumulative hazard function for discrete random variable is

$$\begin{aligned} H(x) &= -\log S(x) && (1.23)\\ &= \sum_{x_j \le x} \log[1 - h(x_j)] && (1.24)\\ &\approx \sum_{x_j \le x} h(x_j); \quad \text{if } h(x_j) \text{ is small for } j = 1, 2, \ldots && (1.25) \end{aligned}$$

Discrete Cumulative Hazard Function I

1. The equation (1.24) is based on the relationship for continuous lifetimes $S(x) = \exp[-H(x)]$ will be preserved for discrete lifetimes.
2. The equation (1.25) is directly estimable from a sample of censored or truncated lifetimes and the estimator has a very desirable statistical properties, however, the relationship $S(x) = \exp[-H(x)]$ for the equation (1.25) no longer holds true.

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Expected Remaining Life Time I

For individuals of age x, we are interested in their expected remaining life time.

Definition 1.10 (**Mean Residual Life Function (Time)**)

Mean residual life function (Time) is defined as expected residual life at time *x*

$$\mathbf{mrl}(x) = \mathcal{E}(X - x|\ |X > x) \tag{1.26}$$

Mean Residual Life Function (Time) I

For a continuous random variable, the mean residual life function, mean and variance of X can be calculated as

$$\mathbf{mrl}(x) = \frac{\int_x^\infty (t-x)\ f(t)\ dt}{S(x)} = \frac{\int_0^\infty t\ f(x+t)dt}{S(x)} = \frac{\int_x^\infty S(t)\ dt}{S(x)} \tag{1.27}$$

$$\mu = \mathcal{E}(X) = \mathbf{mrl}(0) = \int_0^\infty t\ f(t)\ dt = \int_0^\infty\ S(t)\ dt \tag{1.28}$$

$$\mathbf{Var}(X) = 2\int_0^\infty t\ S(t)\ dt - \left[\int_0^\infty S(t)\ dt\right]^2 \tag{1.29}$$

by using $f(t)dt = -dS(t)dt$ and integrating by parts technique.

Definition 1.11 (**Percentiles of Survival Time**)

Percentiles The p^{th} quantile (the $100p\%$ percentile) of distribution of X is the value x_p such that

$$F(x_p) \geq p \text{ and } S(x_p) \geq 1 - p \tag{1.30}$$

If X is a continuous random variable, the p^{th} quantile is found by solving the equation $S(x_p) = 1 - p$.

Definition 1.12 (**Median Life)**

Median life is the 50^{th} percentile $x_{0.5}$ of the distribution of X such that

$$S(x_{0.5}) = 0.5 \tag{1.31}$$

Definition 1.13 (**Median Remaining Lifetime)**

1. **Median remaining lifetime** is the length of the time interval after time x when half of the population that was alive at time x will have died.
2. This quantity is easily computed from the survival curve and equal $\mathbf{med}(x) - x$ where $\mathbf{med}(x)$ satisfies

$$S(\mathbf{med}(x)) = S(x)/2. \tag{1.32}$$

Measuring Failure Time I

Three basic requirements for measuring failure time

1. Time origin
2. Scale of measuring time
3. meaning of point event

Time Origin I

1. The time origin should be precisely defined.
2. The time origin need not be and usually is not at the same calendar time.
3. Most randomized clinical trials have staggered entry, so time origin is usually his own date of entry.

Scale of Measuring Time I

The scale of measuring time is often clock time (real time), although other possibility certainly arise, such as operating time of a system, mileage of a car.

Meaning of Point Event I

The meaning of point event of failure must be **defined precisely** such as death.

Incomplete Observation of the Failure Time I

1. The tools of survival analysis are designed to yield inferences about the distribution of the times to event, X, (lifetime) in a population.
2. A special source of difficulty in survival analysis is that some individuals may not be observed for the full time to failure.
3. Some lifetimes are known to have occurred only within certain interval.
4. Such incomplete observation of the failure time is called censoring.

Incomplete Observation of the Failure Time II

5. In practice, we often do not observe X for a random sample, but only known that X
    1. is a point event and that the period of observation for censored individuals must be recorded (right censoring).
    2. lies in an observed interval (L, R) (interval censoring)
    3. might only observe a subject conditional on certain conditions (truncation)

Censoring I

1. X is unobservable.
2. Formally, an observation is said to be **right censored** at time R if the exact value of the observation is not known but only that it is greater than or equal R, i.e., $X > R$.
3. Similarly, an observation is said to be **left censored** at time L if it is known only that the observation is less than or equal to L, i.e., $X < L$.
4. Right censoring is very common.
5. We use different notation for the observed data to clarify that it is different from the measure, X, that we are interested in.

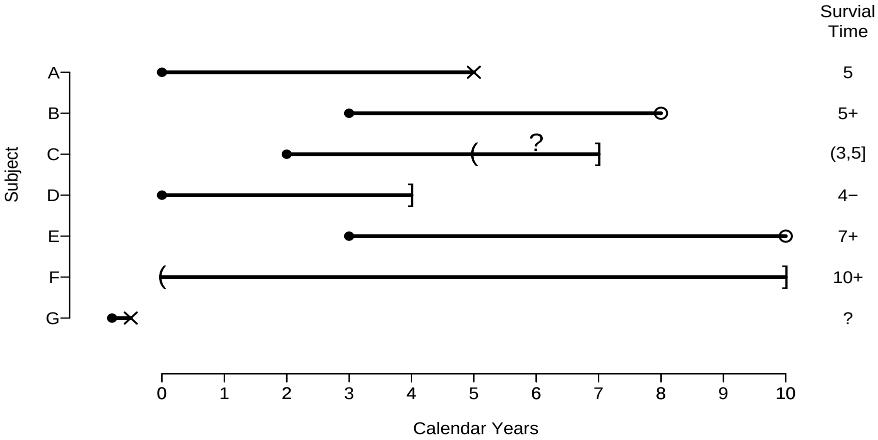

Figure 2: Types of Censoring Time

Censoring I

1. Subject A died at calendar time 5 and the exactly observed patient time is 5.
2. Subject B was right-censored at calendar time 8; the observed patient time is 5 and is the censoring time (as $5+$); however, the exact death time is unknown,
3. Subject C died between calender time 5 and 7, and the exact death time is unknown,
4. The patient time for subject C is between 3 and 5. Let $(3, 5]$ denotes the interval-censored time, such that the death time falls within the interval, $(3, 5]$.

Censoring II

5. Subject *D* died before calender time 4, Subject *D* was left-censored, the exact death time was unknown and less than calendar time 4 (denoted as $4-$).
6. Subject *E* was alive at the end of the study, calendar time 10,
7. subject *E* was right-censored, the death time is greater the patient time 7 (as $7+$).
8. Subject *F* was similar to subject *E*, right-censored survival time as $10+$.
9. Subject *G* died before the beginning of the study, how should be dealing this subject?

Review of Survival Function I

1. X denotes the lifetime for a random variable member of the population.
2. Then survival function

$$
\begin{aligned}
S(x) &= Pr(X > x) && (2.1)\\
&= \int_0^x f(t)dt && (2.2)\\
&= \exp(-H(x)) = \exp[\int_0^x h(t)dt] = 1 - F(x) && (2.3)
\end{aligned}
$$

is the fraction of the population that lives beyond age x.

Review of Survival Function II

3. Distribution function

$$F(x) = 1 - S(x) = Pr(X \leq x) \tag{2.4}$$

is the fraction who die by time x.

Review of Survival Function III

4. Probability density function $f(x)dx \approx$ fraction who die between age x and $x + dx$ when dx is a short interval.

$$f(x) = \frac{dF(x)}{dx} = -\frac{dS(x)}{dx} \tag{2.5}$$

$$= h(x)S(x) = \frac{S(x) - S(x + \Delta(x))}{\Delta x} \tag{2.6}$$

Review of Survival Function IV

5. The hazard function is

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d\log(S(x))}{dx}. \tag{2.7}$$

The cumulative hazard function is

$$H(x) = \int_0^x h(t)dt = -\log(S(x)). \tag{2.8}$$

Observed Data: Right Censoring I

1. The event occurs after the follow-up time.
2. Follow-up starts at time 0 and continue until the event, x, or censoring time, C_r, whichever comes first.
3. The observed data are
    1. the follow-up time $T = \min(X, C)$ and
    2. an indicator for status at end of follow-up as $\delta = I(T = X)$ which equal 1 if a death occurs and equal 0 if the observation is censored.

Observed Data: Right Censoring I

1. Note that we observe either X or C_r, but not both.
2. In some cases, the censoring time, C_r, is pre-specified and is known from the experimental design for all subjects (Type I or Type II censoring).
3. In other cases, the censoring time C_r is random (varies unpredictably among observations) and is statistically independent from X.

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Definition 3.1 (**Type I Censoring**)

Type I Censoring is that event is observed only if it occurs prior to some pre-specified time. This censoring time may vary from individual to individual.

Type I Censoring I

1. Let $X_1, X_2, X_3, \ldots, X_n$ be independently identical distributed (i.i.d.) each with density function F,
2. Let C_r is a fixed censoring time (Type I censoring), then we can only observed $T_1, T_2, T_3, \ldots, T_n$.

Type I Censoring II

$$T_i = \begin{cases} X_i & \text{if } X_i \leq C_r \ . \\ C_r & \text{if } X_i > C_r. \end{cases} \tag{3.1}$$

$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq C_r. \\ 0 & \text{if } X_i > C_r. \end{cases} \tag{3.2}$$

3. So the data can be represented by pairs of random variable (T, δ), such that $T = \min(X, C_r)$ and δ is a censored indicator.

Example 3.2

Type I Censoring (a)

1. An animal experiment with all animals are followed until time C_r, which is the same for all animals.
2. This is type I censoring since the censoring times are pre-specified and known for all subjects.
3. Note that such data could be analyzed directly to estimate the binomial outcome $X \leq C_r$ versus $X > C_r$.
4. Survival analysis yields the information about the probabilities of survival at all times in the interval $(0, C_r)$.

Example 3.3

Type I Censoring (b)

1. Another example for generalized type I censoring is that individuals enter the study at different times (random entry time) and the follow-up stops at a pre-determined specified calendar time.
2. So the censoring times for each individual is the time from entry until end of study.
3. This is **generalized Type 1 censoring**.

Outline

1. Basic Survival Function and Hazard Function
   - Survival Function
   - Hazard Function
   - Discrete Survival and Hazard Function
   - Mean Residual Life Function and Median Life
2. Introduction to Censoring and Truncation
3. Observed Data: Right Censoring
   - **Type I Censoring**
   - Type II Censoring
4. Observed Data: Random Censoring
5. Observed Data: Left, Double and Interval Censoring
6. Observed Data: Truncation
7. Likelihood Construction for Censored and Truncated Data

Definition 3.4 (**Type II Censoring**)

Type II Censoring is in which the study continues until the failure of the first r individuals, where r is some predetermined integer $(r < n)$.

Type II Censoring I

1. Let r, $r < n$, be fixed.
2. Let $X_{(1)} < X_{(2)} < X_{(3)} < \ldots < X_{(n)}$ be the order statistics of $X_1, X_2, X_3, \ldots, X_n$.
3. Observation ceases after the r^{th} failure so we can observed $T_{(1)}, T_{(2)}, T_{(3)}, \ldots, T_{(r)}$.

Type II Censoring II

4. The full ordered observed sample is

$$
\begin{aligned}
T_{(1)} &= X_{(1)} \\
T_{(2)} &= X_{(2)} \\
\vdots &= \vdots \\
T_{(r)} &= X_{(r)} \\
T_{(r+1)} &= X_{(r)} = C_{(r+1)} \\
\vdots &= \vdots \\
T_{(n)} &= X_{(r)} = C_{(n)}
\end{aligned}
$$

Example 3.5

Type II Censoring

1. For example, animal experiment with all 100 animals are followed until 50 animals die, which is the same for all animals.
2. This is type II censoring.
3. It should be stressed that with type II censoring the number of observations r (in the above example, $r = 50$) is decided before the data are collected.
4. Formally the data consist of the r smallest lifetime $T_{(1)} < T_{(2)} < \ldots < T_{(r)}$ out of random sample n lifetime $T_1, T_2, \ldots, T_n$ from lifetime distribution.
5. A generalized progressive type II censoring has different sacrifice times.

Random Censoring I

1. Censoring time are often effectively random.
2. Sometimes, individuals will experience some other competing event of interest which causes them to be removed from the study.
3. Some events which cause the individual to be randomly censored, with respect to event of interest, are accident death, migration of human population.

Random Censoring II

4. Let $C_1, C_2, \ldots, C_n$ be i.i.d. each with distribution function G. C_i is the censoring time associated with T_i.
5. We can only observe $(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)$, where

$$T_i = \min(X_i, C_i) = X_i \wedge C_i \tag{4.1}$$

$$\delta_i = \mathrm{I}(X_i \leq C_i) \begin{cases} 1 & \text{if } X_i \leq C_i \text{, that is, } T_i \text{ is not censored,} \\ 0 & \text{if } X_i > C_i \text{, that is, } T_i \text{ is censored.} \end{cases} \tag{4.2}$$

Random Censoring III

6. Random censoring often arises in medical application.
7. In a clinical trial, patients may enter the study at different times; then each is treated with one of several possible therapies.
8. We want to observe their lifetimes, but censoring occurs in one of the following forms: loss to follow-up, drop out and termination of the study.
9. We often assume that X_i and C_i are independent.
10. If the reason for dropping out is related to the course of the therapy, there may well be dependent between X_i and C_i.

Example 4.1

Random Censoring: Competing Risks

1. Consider conceptual times until death from various causes.
2. $X =$ time to death from cancer.
3. $C_R =$ time to death from other causes.
4. We observe the time until death $T = min(X; C_R)$.

Definition 5.1 (**Left Censoring)**

A life time X associated with a specific individual in a study if it is less than a censoring time C_l.

Left Censoring I

1. That is, the event of interest has already occurred before that person is observed in the study at time C_l.
2. We know that they have experienced the event sometime before time C_l, but their exact event time is unknown.

Left Censoring II

3. We only observe $(T_1, \varepsilon_i), (T_2, \varepsilon_i), \ldots, (T_n, \varepsilon_n)$, where

$$T_i = \max(X_i . C_i) = X_i \vee C_i, \tag{5.1}$$

$$\varepsilon_i = \mathrm{I}(C_i \leq T_i) \tag{5.2}$$

4. The event time occurred before the observation time.
5. Follow-up starts at time C_l.
6. If $X \leq C_l$ then follow-up stops immediately and we observed the time C_l and the value of $I(X \leq C_l)$.
7. If $X > C_l$ then follow-up continues until the event at time X.

Definition 5.2 (**Doubly Censoring)**

If left censoring occurs in a study, right censoring may also occur, that is combination of right and left censoring. (Turnbull, 1974).

Doubly Censoring I

1. The data can be represented by a pair of variables (T, δ), where

$$T = \max[\min(X, C_r),\ C_l] \tag{5.3}$$

$$\delta = \begin{cases} 0, & \text{if } T \text{ is a death time,} \\ 1, & \text{if } T \text{ is a right-censored time,} \\ -1, & \text{if } T \text{ is a left-censored time.} \end{cases} \tag{5.4}$$

2. Followup starts at time C_l and stops at time $T = \min(X, C_r)$.

Example 5.3

Left, Double, and Interval Censoring

1. For example, in a study to determine the distribution of the time until first smoking use among high school boys in Taiwan, the question was asked “When did you first smoke?”
2. One of the responses was “I have smoked but can not recall just when the first time was”.
3. The event had occurred prior to the boy’s age at interview but the exact age is not known.
4. This is **left censoring**.

5. Another possible response was “I never smoke”.
6. This is **right censoring** at the age of interview.
7. Both right censored observations and left censored observation are present as **double censoring**.

Interval Censoring I

Definition 5.4 (**Interval Censoring)**

Event time of an individual, X, is only known to fall into an interval $(L, R]$.

1. Observation is not continual, but occurs at discrete times and we only know the time between which the event occurred.
2. The data consist of the interval that includes the event: $(L < X \leq R]$.
3. For example, in longitudinal screening studies, patients or machines have periodic examinations and the event time is only known to fall into an interval.

Survival Analysis and Censoring I

1. The standard tools of survival analysis are only valid when the censoring times are “unrelated” to the survival times.
2. There are several ways to specify sufficient definitions for “unrelated”.
3. One sufficient conditions is that the death rate among those who were censored equal the death rate among those still in the study.
4. This allow us to estimate death rates based on those are being followed at each moment in time.

Definition 6.1 (**Truncation)**

1. **Truncation** is a condition which screens certain subjects so that the investigator will not be aware of their existence.
2. Only part of the population is observed.
3. The observable subset is defined by the value of X.
4. The fraction unobserved is unknown.

Truncation I

1. If Y_l is the time of the event truncates individuals, then for **Left Truncation** samples, only individuals with $X \geq Y_l$ are observed.
2. We can only estimate $S_X(x \mid X \geq Y_l)$.

Truncation II

3. The most common type of left truncation occurs when subjects enter a study at random ages and are followed from this **delayed entry time** until the event occurs or until the subject is right-censored.
4. Subjects who die before this delayed entry time are not known.
5. We observed the bivariate vector (X, Y_l) when $X > Y_l$.
6. We do not even know how many truncated observations with $X \leq Y_l$ there are.

Truncation III

7. For example, let X be the lifetime from first symptom conditional on survival until referral to a treatment center.
8. We don't know how many died before they were referred to the study center.
9. For another example, disease registry which includes the time of disease among those who have gotten the disease.
10. How many did not get the disease?

Right Truncation I

1. **Right truncation** occurs when only individuals who have experienced the event are included in the sample, and any individual who yet to experience the event is not observed.
2. This definition of right truncation is poorly defined in the text and is not standardized.
3. That is, if Y_r is the time of the event truncates individuals, only individuals with $X \leq Y_r$ are observed.
4. We can only estimate $S_X(x \mid X \leq Y_r)$.
5. The mechanism that leads to this is not specified. It should be.

Right Truncation I

1. An example of a right-truncated data is particular relevant to studies of AIDS.
2. For example, in a AIDS mortality study, the number of infected individuals is unknown and information is available only for those who became infected and developed AIDS within a certain time period (i.e. 1985-1990).
3. Individuals who have yet to develop AIDS are not known to the investigator and are not included in the sample.

Example 6.2

Left and Right Truncation

1. Consider the follow-up schematic Figure 3 below.
2. Let X be the age of mumps (a disease that occurs only once, at most).
3. School records of all students with the mumps were reviewed and the age was recorded.
4. Data included 6 years of elementary school (during which records are reviewed), and 2 years after that.

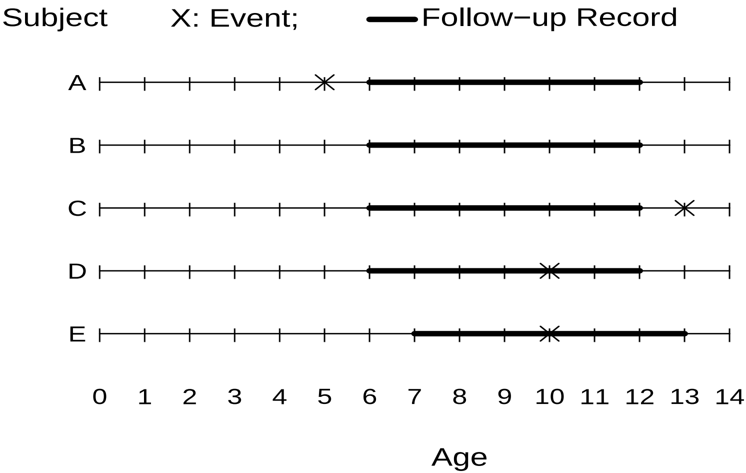

Figure 3: Left and Right Truncation

Example: Left and Right Truncation I

1. Each line segment represents 1 year.
2. Dark horizontal lines represent time during which records are abstracted. × represent ages of mumps.
3. There are 6 years prior to school, except for the last observation who started school at age 7.

Example: Left and Right Truncation II

4. The first three subjects (A, B, C) are not observed (we do not even know how many such observations there are).
5. The last two subjects (D, E) yield left and right truncated observations: $(X = 10 \mid 6 < X \leq 12)$ and $(X = 10 \mid 7 < X \leq 13)$

Likelihood Construction I

1. Many of the tools of survival analysis use maximum likelihood methods to estimate the distribution of X.
2. We write down the likelihood of the data, which yields a function of the parameters to be estimated.
3. Maximizing that function yields estimates of parameters.

Likelihood Construction II

4. To write down likelihood function, the first is to decide what is random variable X.
5. Then choose a family of distributions for the random quantities (parametric or non-parametric).
6. Decide what is independent of what (different subjects, often).
7. Decide what is fixed by design.

Likelihood Construction III

8. Write down the likelihood that the random quantities take the values observed for each independent unit, and write he product over independent units.
9. Find the parameter values that maximize the resulting likelihood.
10. Estimate the parameters of interest and decide what to do about the other parameters.
11. The likelihood for various type of censoring schemes may all be written by incorporating the following components:

Likelihood Construction IV

Observations	Probability	Likelihood
Exact lifetimes	$P(X = x)$	$\Rightarrow f(x)$
Right-censored	$P(X > C_r)$	$\Rightarrow S(C_r)$
Left-censored	$P(X \leq C_l)$	$\Rightarrow 1 - S(C_l)$
Left-truncated	$P((X = x \mid X > Y_l)$	$\Rightarrow f(x) \ / \ S(Y_l)$
Right-truncated	$P(X = x \mid X \leq Y_r)$	$\Rightarrow f(x) \ /[1 - S(Y_r)]$
Interval-censored	$P(L < X \leq R)$	$\Rightarrow [S(L) - S(R)]$

Likelihood Construction I

1. In general case, $\delta = 1$ if $X = x$ exactly or $\delta = 0$ if censored.
2. There are several questions need to be addressed first:

Likelihood Construction II

- Are subjects with no event in the sample?
- Does potential follow-up start at time 0 or at some time after 0?
- Are subjects with an event before entry into follow-up in the sample?
- Does follow-up stop before the event for some subjects?

Likelihood Construction III

- Does we know about subjects whose event is before the start or after the end of follow-up?
- Is the event time know exactly, or only to within an interval?
- Is the censoring process independent of the event time?

Likelihood: Censored Data I

1. The likelihood function may be constructed by putting together the components:

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(C_{ir}) \prod_{i \in L} [1 - S(C_{il})] \prod_{i \in I} [S(L_i) - S(R_i)] \quad (7.1)$$

2. Where *D* is the set of death time, *R* the set of right-censored observations, *L* the set of left-censored observations, and *I* the set of interval-censored observations.

Likelihood: Left-Truncated Data I

For **left-truncated data**, with truncation time Y_l independent form the death time, we

$$\begin{aligned} \text{replace} \quad f(x_i) \quad &\text{by} \quad f(x_i)/S(Y_{il}); && (7.2) \\ \text{and replace} \quad S(C_{ir}) \quad &\text{by} \quad S(C_{ir})/S(Y_{il}). && (7.3) \end{aligned}$$

Likelihood: Right-Truncated Data I

For **right truncated data**, only deaths are observed, so that the likelihood is of the form

$$L \propto \prod_i f(x_i)/[1 - S(Y_{ir})] \tag{7.4}$$

Assumptions for Likelihood I

Two critical assumptions are that

1. Indenendent between Lifetime and Censoring Times Distributions
2. Non-Informative Censoring

Definition 7.1 (**Independent Censoring)**

The likelihood approach above assume that the distribution of event times for those in the study (uncensored and not truncated) is the same as for those not in the study.

1. For example, for right censored data, this assumption is the same as $h(x|C \geq x) = h(x|C < x)$.
2. That is, the death rate is the same for subjects who were previously censored as it is for subjects who are still being followed.
3. Always think about this assumption to determine logically if it is reasonable. Better yet, use empirical evidence to determine if it is plausible.

Definition 7.2 (**Non-Informative Censoring)**

1. When the censoring mechanism is random, it should be modeled in the likelihood.
2. The distribution of censoring times could conceivable shared parameters with the survival distribution.
3. We generally assume that there are no such shared parameters (non-informative censoring).
4. This concept is made more explicit in the random censoring model.

Suppose that $X \sim F(\underline{\boldsymbol{\theta}} \mid ; X)$ and $C \sim G(\underline{\boldsymbol{\theta}}^{\star} \mid C)$, under the non-informative censoring assumption, we have

$$\underline{\boldsymbol{\theta}} \cap \underline{\boldsymbol{\theta}}^{\star} = \emptyset. \tag{7.5}$$

Definition 7.3 (**Random Censoring Likelihood**)

Random censoring in likelihood builds in independence and non-informativeness.

Random Censoring Likelihood I

1. A important special situation is random censoring with right censoring.
2. Let lifetime random variable X with density $f_{\underline{\theta}}$, distribution $F_{\underline{\theta}}$, survival function $S_{\underline{\theta}}$, and censoring time random variable C with density $g_{\underline{\alpha}}$, survival function $G_{\underline{\alpha}}$.

Random Censoring Likelihood II

3. Let observed data represent pairs of random variable (T, δ) such that $T = \min(X, C), \delta = \mathrm{I}(T = X)$ is 1 if follow-up is ended by the event and 0 if censored.
4. Assume that X and C are independent.
5. Then $Pr(T > t) = Pr(X > t \text{ and } C > t) = S_{\underline{\boldsymbol{\theta}}}(t) G_{\underline{\boldsymbol{\alpha}}}(t)$.

Random Censoring Likelihood III

6. For example,

$$Pr(T = t, \delta = 1) = Pr(X = t \text{ and } C > t) = f_{\underline{\theta}}(t) G_{\underline{\alpha}}(t) \quad (7.6)$$

$$Pr(T = t, \delta = 0) = Pr(X > t \text{ and } C = t) = S_{\underline{\theta}}(t) g_{\underline{\alpha}}(t) \quad (7.7)$$

Random Censoring Likelihood IV

7. Assume different observation are independent.
8. Let random sample of pairs $(T_i, \delta_i), i = 1, 2, \ldots, n$, are i.i.d. with distribution as described above, then likelihood function is

$$\mathbf{L} = \prod_i^n Pr[t_i, \delta_i] = \prod_i^n [f_{\underline{\boldsymbol{\theta}}}(t_i) G_{\underline{\boldsymbol{\alpha}}}(t_i)]^{\delta_i} \, [S_{\underline{\boldsymbol{\theta}}}(t_i) g_{\underline{\boldsymbol{\alpha}}}(t_i)]^{1-\delta_i} \tag{7.8}$$

$$\mathbf{L} = \prod_i^n Pr[t_i, \delta_i] = \prod_i^n [f_{\underline{\boldsymbol{\theta}}}(t_i)]^{\delta_i} \, [S_{\underline{\boldsymbol{\theta}}}(t_i)]^{1-\delta_i} \prod_i^n [G_{\underline{\boldsymbol{\alpha}}}(t_i)]^{\delta_i} \, [g_{\underline{\boldsymbol{\alpha}}}(t_i)]^{1-\delta_i} \tag{7.9}$$

Random Censoring Likelihood V

9. If $\underline{\boldsymbol{\theta}}$ and $\underline{\boldsymbol{\alpha}}$ share no common components, then the second factor in the equation (7.9) is a constant respect to $\underline{\boldsymbol{\theta}}$ and can be ignored when maximization the likelihood with respect to $\underline{\boldsymbol{\theta}}$.
10. That is

$$\mathbf{L} = \prod_{i}^{n} Pr[t_i, \delta_i] \propto \prod_{i}^{n} [f_{\underline{\boldsymbol{\theta}}}(t_i)]^{\delta_i} \; [S_{\underline{\boldsymbol{\theta}}}(t_i)]^{1-\delta_i} \tag{7.10}$$

$$\mathbf{L} = \prod_{i}^{n} Pr[t_i, \delta_i] \propto \prod_{i}^{n} [\; h_{\underline{\boldsymbol{\theta}}}(t_i) \;]^{\delta_i} \; \exp[\; -H_{\underline{\boldsymbol{\theta}}}(t_i) \;] \tag{7.11}$$