3.2-3.4 Formal introduction to censoring and truncation

1. Right censoring: An observation is right censored if it is not known exactly, but greater than some value.

- $X =$ time to event of interest, $C =$ censoring time
- We observe $(T, \delta)$, where $T = \min(X, C), \delta = I(X < C)$.
- Data with $n$ realizations of $(T, \delta)$, denoted as $\{(T_i, \delta_i), i = 1, \ldots, n\}$
- Many types of right censoring, some provide information on $C$, even when $X < C$.

(a) Type I right censoring: The event is observed only if it occurs prior to some pre-specified time.

- Fixed type I censoring
  - $C_i = \tau$ for all $i$
  - Often occur in controlled experiment settings
  - E.g. Mice study — time to death after a fixed time period following intense exposure to carcinogen
- Progressive type I censoring: $C$ is fixed in advance, but may take different values for each $i$. There is only finite number of values of $C$.
  - $X =$ time at which there is a certain % weight loss
  - Weights are measured every 12 hours.
  - Tumor size is also of interest.
  - Tumor sizes only observable at sacrifice times
  - Sacrifice scheme: Mice in 1st and 2nd cages are sacrificed at the end of week 1 and in 3rd and 4th cages are sacrificed by the end of week 2.

- Generalized type I censoring
  - $C$ may be different for all individuals, but are known in advance.
  - It often occurs in human study with staggered entry into study with complete follow-up until the end of the study.
  - E.g., Phase I study of chemotherapy, $X =$ time to toxicity
- $C$ can be computed at time patient enrolled on study, assuming length of study fixed and no dropout.
- $C$ can take infinite number of values.
- With type I censoring, $C$ is always observed, even though $X$ may not be; this is not always realistic.
- Methods were originally developed for type I censored data.
- May not be applicable when $C$ is not observed on all subjects.

(b) Type II right censoring

- Simple type II censoring
  - Sample size $n$
  - Study continues until the failure of the first $r$ individuals.
  - Theory of order statistics is directly applicable to determine the likelihood.
  - Note that $r$ the number of failures and $n - r$ the number of censored observations are fixed, and the censoring time $T(r)$ is random.

- Progressive type II censoring
  - First $r_1$ failures observed, $n_1 - r_1$ of the remaining $n - r_1$ unfailed items are removed/sacrificed.
  - When the next $r_2$ failures observed, $n_2 - r_2$ of the unfailed items are removed/sacrificed.

(c) Competing risks censoring: the censoring time may be correlated with the event time.

- E.g. dementia/death
- E.g. informative dropout
- Survival function for the event of interest is not identifiable.
• Special quantities (cause-specific hazard / cumulative incidence function) may be used instead.

(d) Random censoring

• usual setup in human study
• \( C \sim G(t) \), not fixed in advance, may not be known
• Censoring at end of study, known as “administrative loss to follow-up” fixed in advance, similar to type I censoring.
• Censoring prior to end is “drop out”, due to other factors, random
• There is some ambiguity in the definitions of random censoring.
  – P69 L15: A special case of competing risks censoring is random censoring.
  – P76 in Example 3.10: A simple random censoring occurred if each subject has a lifetime \( T \) and a censoring time \( C_r \), \( T \) and \( C_r \) being independent random variables.
  – In this class, if \( T \) and \( C \) are independent, while \( C \) is random, we call it random censoring. If \( T \) and \( C \) are dependent, we call it competing risks censoring.

2. Left censoring: An observation is left censored if it is not known exactly, but less than some value.

E.g., some patients with cystic fibrosis are found to have *pseudomonas aeruginosa* (PA) infection at their first recorded clinical visits.

3. Double censoring

• \( C_l = \) left censoring time
• \( C_r = \) right censoring time
• \( T_* = \max\{\min(X, C_r), C_l\} \)
\[ \delta_s = \begin{cases} 1 & if \ C_l \leq X \leq C_r \\ -1 & if \ X < C_l \\ 0 & if \ X > C_r \end{cases} \]

4. Interval censoring: An observation is interval censored if it is only known to happen within a certain interval.

5. Left truncation: An observation is left truncated if it is observed exactly, but conditionally on > some value.

6. Right truncation: An observation is right truncated if it is observed exactly, but conditionally on < some value.

**Exercise**

To estimate the distribution of the ages at which post menopausal women develop breast cancer, a sample of eight healthy 50-year-old women were given yearly mammograms for a period of 10 years. At each exam, the presence or absence of a tumor was recorded. In the study, no tumors were detected by the women by self-examination between the scheduled yearly exams, so all that is known about the onset time of breast cancer is that it occurs between examinations. For four of the eight women, breast cancer was not detected during the 10 year study period. The ages of onset of breast cancer for the eight subjects were in the following intervals: \((55, 56], (58, 59], (52, 53], (59, 60], > 60, > 60, > 60, > 60, > 60\).

**Question:** What type of censoring or truncation is represented in this sample?
### 3.5 Likelihood construction for censored and truncated data

Let $X$ be a continuous lifetime with pdf $f(x)$, cdf $F(x)$ and survival function $S(x) = 1 - F(x)$. Let

- $C_r = \text{right censoring point}$
- $C_l = \text{left censoring point}$
- $(L, R] = \text{censoring interval}$
- $Y = \text{truncation point}$

We consider a random sample $X_1, X_2, \ldots, X_n$, and want to determine a likelihood $L$. For data subject to censoring, the general form is

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in RC} S(C_{r,i}) \prod_{i \in LC} F(C_{l,i}) \prod_{i \in IC} \{S(L_i) - S(R_i)\}, \quad (3.5.10)$$

where

- $D = \text{death times (i.e. observed values of } X_i)$
- $RC = \text{right censored observations}$
- $LC = \text{left censored observations}$
- $IC = \text{interval censored observations}$

**How about truncation?**

**Type I RC:** We use $(T_i, \delta_i), i = 1, \ldots, n$, where $T_i = \min(X_i, C_i), \delta_i = I(T_i = X_i)$. The likelihood is expressed below:

$$L = \prod_{i=1}^n Pr(t_i, \delta_i) = \prod_{i=1}^n (f(t_i))^{\delta_i} (S(t_i))^{1-\delta_i} = \prod_{i=1}^n (h(t_i))^{\delta_i} e^{-H(t_i)}.$$
Example: $X \sim \text{exponential}(\lambda)$, and $X$ is subject to right censoring. Let $T = \min(X, C)$ and $\delta = I(X < C)$. What is the likelihood function?

<table>
<thead>
<tr>
<th>$T_i$</th>
<th>0.5</th>
<th>1</th>
<th>0.75</th>
<th>0.25</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta_i$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Random censoring: $X_i$ and $C_i$ are both random, but are assumed to be independent. $X_i \sim \text{pdf } f$ and survival $S$; $C_i \sim \text{pdf } g$ and survival $G$; and the joint pdf of $(X_i, C_i)$ is denoted as $f(u, v)$. We still observe $T_i = \min(X_i, C_i)$, and $\delta_i = I(T_i = X_i)$. The likelihood function can be expressed as following:

$$L = \prod_{i=1}^{n} Pr(t_i, \delta_i)$$

$$= \prod_{i=1}^{n} (f(t_i) G(t_i))^{\delta_i} (g(t_i) S(t_i))^{1-\delta_i}$$

$$= \prod_{i=1}^{n} (f(t_i))^{\delta_i} (S(t_i))^{1-\delta_i} \prod_{i=1}^{n} (g(t_i))^{1-\delta_i} (G(t_i))^{\delta_i}$$

$$\overset{d}{=} L_X L_C.$$

- Suppose that $f(t) = f(t; \theta)$, where $\theta$ is unknown.

- How to estimate $\theta$?

- Approach 1: ignore $L_C$, maximize $L_X$ to estimate $\theta$

- Approach 2: Simultaneously maximize $L_X$ and $L_C$ to estimate $\theta$

- Valid inferences for $\theta$ by ignoring $L_C$, regardless of whether $g$ depends on $\theta$

- Don’t have to specify $g$. 
Type II censoring: Suppose we have $n$ samples $X_1, \ldots, X_n$, with pdf $f(x)$. For the simple type II censoring, recall that we have $r$ ordered values $X_{(1)} < X_{(2)} < \cdots X_{(r)}$ and $n-r$ values censored at $X_{(r)}$. From a standard result on order distribution, the joint distribution of the first $r$ ordered statistics and the last $n-r$ censored values is

$$L = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(x_{(i)}) \{S(x_{(r)})\}^{n-r}.$$

Progressive Type II censoring: Two repetitions only, i.e., first $r_1$ ordered failures $X_{(1)}, \ldots, X_{(r_1)}$, then $n_1 - r_1$ additional items removed. The next $r_2$ ordered statistics observed, denoted by $X^*_{(1)}, \ldots, X^*_{(r_2)}$, then the remaining $n - n_1 - r_2$ removed. We have

$$L \propto \prod_{i=1}^{r_1} f(x_{(i)}) \{S(x_{(r_1)})\}^{n_1-r_1} \prod_{i=1}^{r_2} f(x^*_{(i)}) \{S(x^*_{(r_2)})\}^{n-n_1-r_2}.$$