Selection Biases: Truncation and Censoring

Jogesh Babu and Woncheol Jang

Penn State University and Duke University

March 8, 2006
Outline

1. Bias
2. Censoring
3. Statistical inferences for censoring
4. Truncation
5. Statistical inferences for truncation
6. Statistical inferences for Doubly truncated data
7. Summary
Bias

Statistical (Mathematical) bias $E(\hat{\theta}) - \theta$

$E(\bar{X}) = \mu \rightarrow$ No bias (Unbiased estimate)

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

$E(S^2) - \sigma^2 = 0 \rightarrow$ No bias.

But $E(S') - \sigma \neq 0$. Biased estimate.

Transformation Bias! Similar to the one expressed in one of the talks earlier.
Censoring vs Truncation

- **Censoring**: Sources/events can be detected, but the values (measurements) are not known completely. We only know that the value is less than some number.

- **Truncation**: Objects can be detected only if their values is greater than some number; and the value is completely known in the case of detection. For example, objects of certain type in a specific region of the sky will not be detected by the instrument if the apparent luminosity of objects is less than a certain lower limit. This often happens due to instrumental limitations or due to our position in the universe.

- The main difference between censoring and truncation is that censored object is detectable while the object is not even detectable in the case of truncation.
Example: Left/Right Censoring

- **Right Censoring**: the exact value \( X \) is not measurable, but only \( T = \min(X, C) \) and \( \delta = I(X \leq C) \) are observed.

- **Left Censoring**: Only \( T = \max(X, C) \) and \( \delta = I(X \geq C) \) are observed.
Example: Interval/Double Censoring

- **Double Censoring**: This occurs when we do not observe the exact time of failure, but rather two time points between which the event occurred:

  \[
  (T, \delta) = \begin{cases} 
  (X, 1) & : \quad L < X < R \\
  (R, 0) & : \quad X > R \\
  (L, -1) & : \quad X < L
  \end{cases}
  \]

  where \( L \) and \( R \) are left and right censoring variables.

- HIV vaccine trial with 4 monthly blood testing
Survival Function

- Cumulative failure function: \( F(t) = P(T \leq t) \)
- Density function: \( f(t) = \frac{dF(t)}{dt} \)
- Survivor function:

\[
S(t) = P(T > t) = 1 - F(t)
\]
Hazard function

- Hazard rate:

\[ h(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta | T \geq t)}{\Delta} \]

The hazard rate is the instantaneous probability of having an event at time \( t \) given that one has survived up to time \( t \).

- Cumulative hazard function: \( H(t) = \int_0^t h(u)du \)

- \( h(t) = \lambda \) if and only if \( F \) is exponential distribution.
Relationship between those functions

- Hazard from density and survival: $h(t) = f(t)/S(t)$
- Survival from density: $S(t) = \int_{t}^{\infty} f(u)du$
- Density from survival: $f(t) = -dS(t)/dt$
- Density from hazard: $f(t) = h(t) \exp\left(-\int_{0}^{t} h(u)du\right)$
- Survival from hazard: $S(t) = \exp\left(-\int_{0}^{t} h(u)du\right)$
- Hazard from survival: $h(t) = -\frac{d}{dt} \log S(t)$
Kaplan-Meier Estimator

- Nonparametric estimate of survivor function
  \[ S(t) = P(T > t) \]
- Intuitive graphical presentation
- Commonly used to compare two study populations
Survival Data (right-censored)
Corresponding Kaplan-Meier Curve

100%

Probability of surviving to just before 4 months is 100% = 5/5

Subject E dies at 4 months

Fraction surviving this death = 4/5

→ Time in months →
Survival Data

- Subject A: drops out after 6 months
- Subject B
- Subject C: dies at 7 months
- Subject D
- Subject E: dies at 4 months

Beginning of study → Time in months → End of study
Corresponding Kaplan-Meier Curve

subject C dies at 7 months

Fraction surviving this death = 2/3

→ Time in months →
Survival Data

Subject A
2. subject A drops out after 6 months

Subject B

Subject C
3. subject C dies at 7 months

Subject D

Subject E
1. subject E dies at 4 months

Beginning of study
→ Time in months →
End of study

4. Subjects B and D survive for the whole year-long study period
Corresponding Kaplan-Meier Curve

Product limit estimate of survival =
\[ P(\text{surviving/at-risk through failure 1}) \times 
P(\text{surviving/at-risk through failure 2}) = 
\frac{4}{5} \times \frac{2}{3} = .5333 \]
Kaplan-Meier Estimator (continued)

- Let
  - $t_i$: $i$th ordered follow-up time
  - $d_i$: number of ‘censored’ events at $i$th ordered time
  - $R_i$: number of subjects at-risk at $i$th ordered time

- The Kaplan Meier estimator is
  \[
  KM(t) = \prod_{t_i \leq t} \left( 1 - \frac{d_i}{R_i} \right)
  \]

- Standard error of $KM(t)$ (Greenwood’s formula)
  \[
  \hat{\text{Var}}\left( \hat{S}(t) \right) = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{R_i(R_i - d_i)}
  \]
Truncation

- **Left Truncation**: An event/source is detected if its measurement is greater than a truncation variable.
- **Right Truncation**: An event/source is detected if its measurement is less than a truncation variable.
- **Double Truncation**: This occurs when the time to event of interest in the study sample is in an interval.

The pair $(X, Y)$ is observed only if $X \geq Y$, $X$ is the measurement of interest and $Y$ is the truncation variable.

$$M = m + 5 \log P - 5$$

$P$ parallax

Object is detected only if $P \geq \ell$. 
Lynden-Bell- Woodroofe Estimator

- Model: observe $y$ only if $y > u(x)$.
- Data: $(x_1, y_1), \ldots, (x_n, y_n)$.
- Risk set numbers:

$$N_j = \#\{ : u_i \leq y(j) \text{ and } y_i \leq y_j \}$$

where $u_i = u(x_i)$ and $y(i)$ is $i$th ordered value of $y = (y_1, \ldots, y_n)$

- In the KM estimator, $N_j$ is the number of points at risk just before the $j$th event.

- The only differences in comparable points between truncated cases and censoring cases is that points with $y_i > y_j$ but $t(x_i) > y_j$ are not at risked in the truncated case because it cannot be observed.
Bias  Censoring  Statistical inferences for censoring  Truncation  Statistical inferences for truncation  Statistical inferences for Doubly truncated data

Summary
Lynden-Bell- Woodroofe Estimator (continued)

- Hazard rate estimate:

\[ \hat{h}_j = -\log \left( 1 - \frac{1}{N_j} \right) \]

- Lyden-Bell-Woodroofe survival estimator:

\[ LBW(t) = 1 - \exp \left( -\sum_{i \leq j} \hat{h}_i \right) \]
Efron’s Nonparametric MLE

- Model: observe $y$ only if $y \in [u(x), v(x)]$.
- Data: $(x_1, y_1), \ldots, (x_n, y_n)$.
- Let

\[
\hat{Q}_i = \frac{\hat{G}_{v_i+}}{\hat{F}_i}
\]
\[
\hat{G}_{v_i+} = \sum \left\{ \hat{f}_k : y_k > v_i \right\}
\]
\[
J_{ij} = I(y_j \in [u(x_i), v(x_i)])
\]
\[
\hat{f}_i = \left( \sum_{j=1}^{n} J_{ji} F_j^{-1} \right)^{-1}
\]
Efron’s Nonparametric MLE (continued)

- hazard rate estimate:

\[
\hat{h}_j = \left( N_j + \sum_{i=1}^{n} J_{ij} \hat{Q}_i \right)^{-1}
\]

where \( N_j \) are the risk set numbers.

- survival estimator:

\[
\hat{S}(t) = 1 - \exp \left( - \sum_{i \leq j} \hat{h}_i \right)
\]