Selection Biases: Truncation and Censoring

Jogesh Babu

Penn State University

July 23, 2010

Outline

1. Censoring vs Truncation
2. Censoring
3. Statistical inferences for censoring
4. Truncation
5. Statistical inferences for truncation
6. Doubly truncated data
7. Recent methods

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: An object can be detected only if its value 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

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.

Survival Function

- Cumulative failure function:

$$F(t) = P(T \leq t)$$

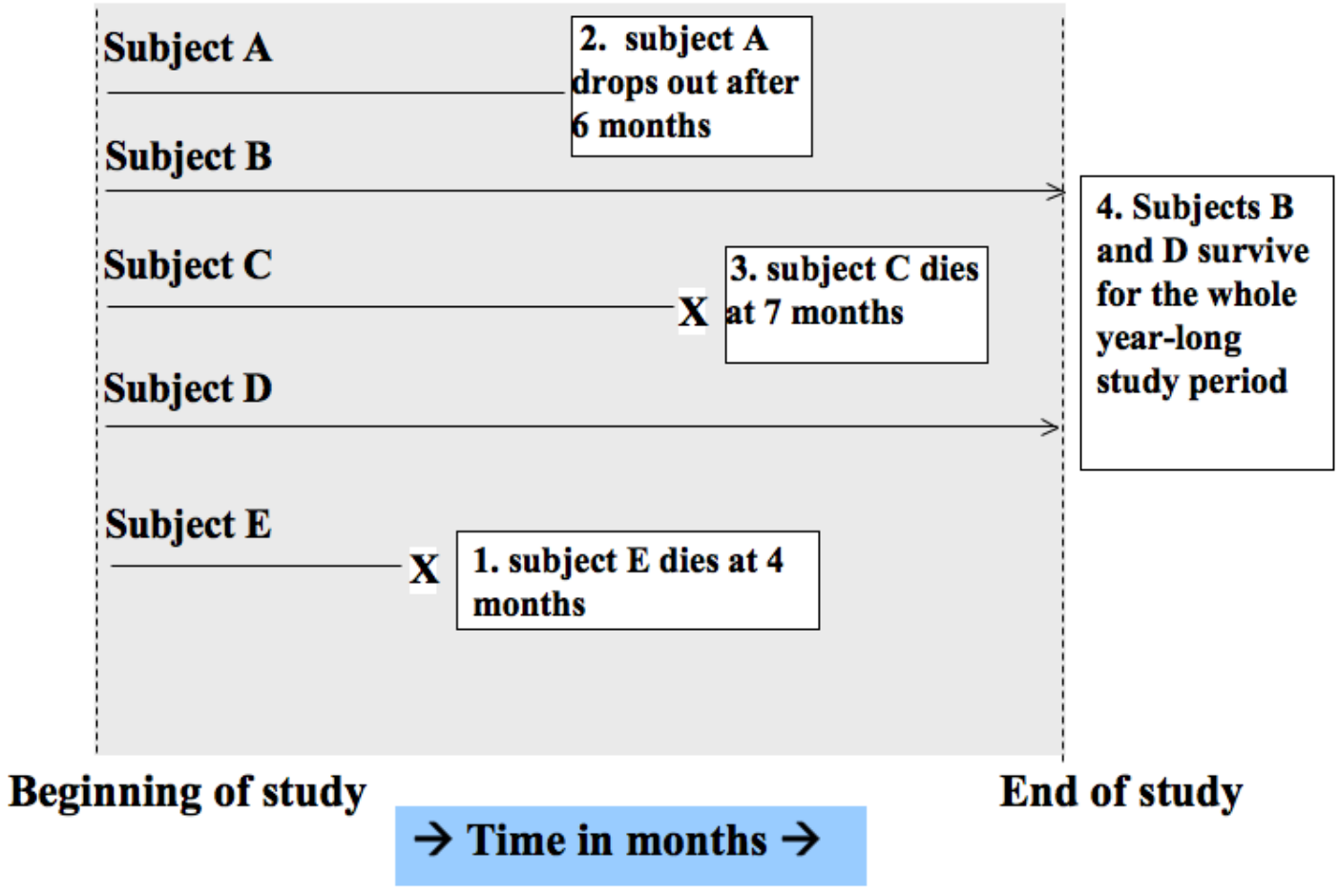
- Survivor function:

$$S(t) = P(T > t) = 1 - F(t)$$

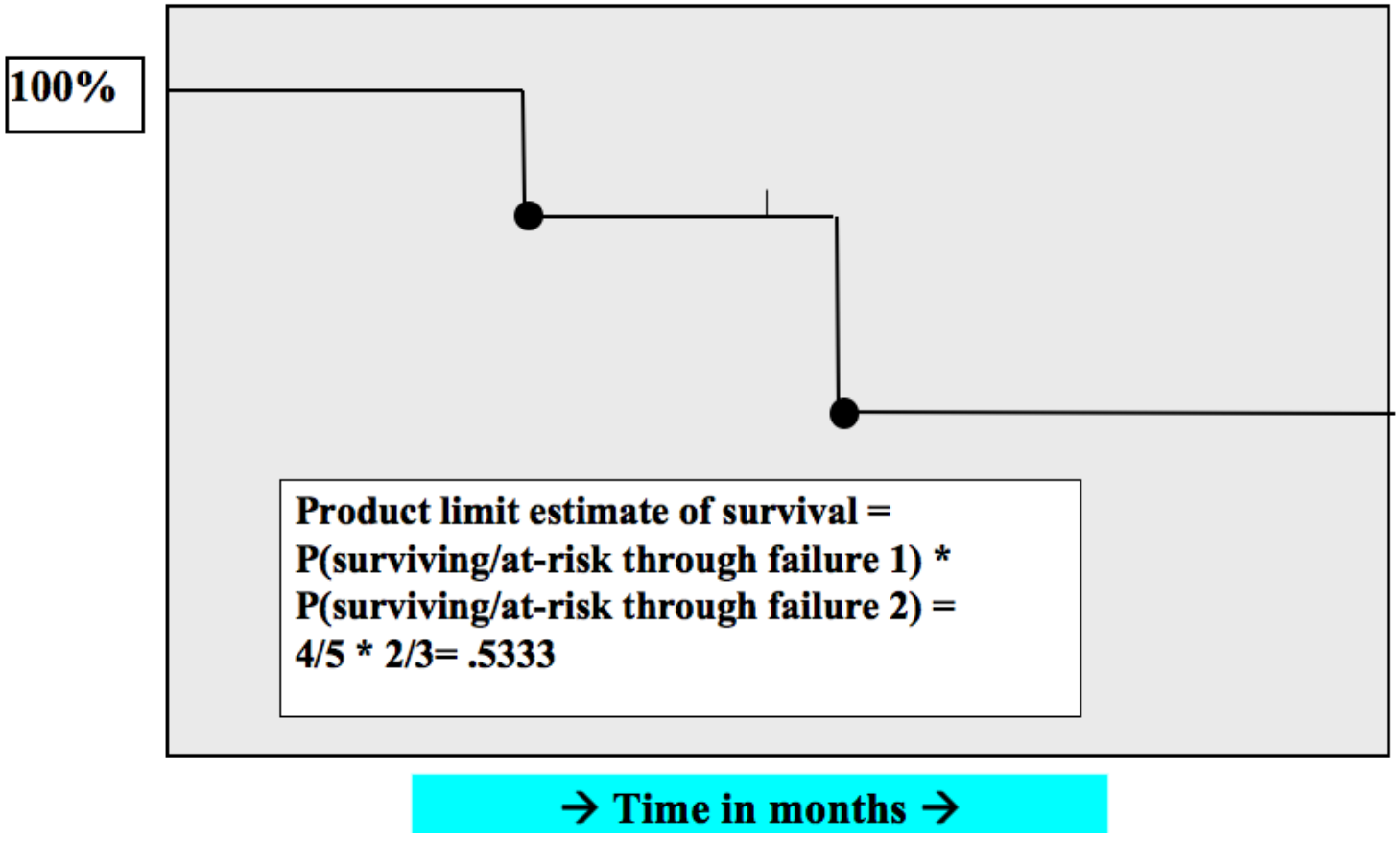
Kaplan-Meier Estimator

- Nonparametric estimate of survivor function $S(t) = P(T > t)$
- Intuitive graphical presentation
- Commonly used to compare two populations

Survival Data



Corresponding Kaplan-Meier Curve



Kaplan-Meier Estimator (continued)

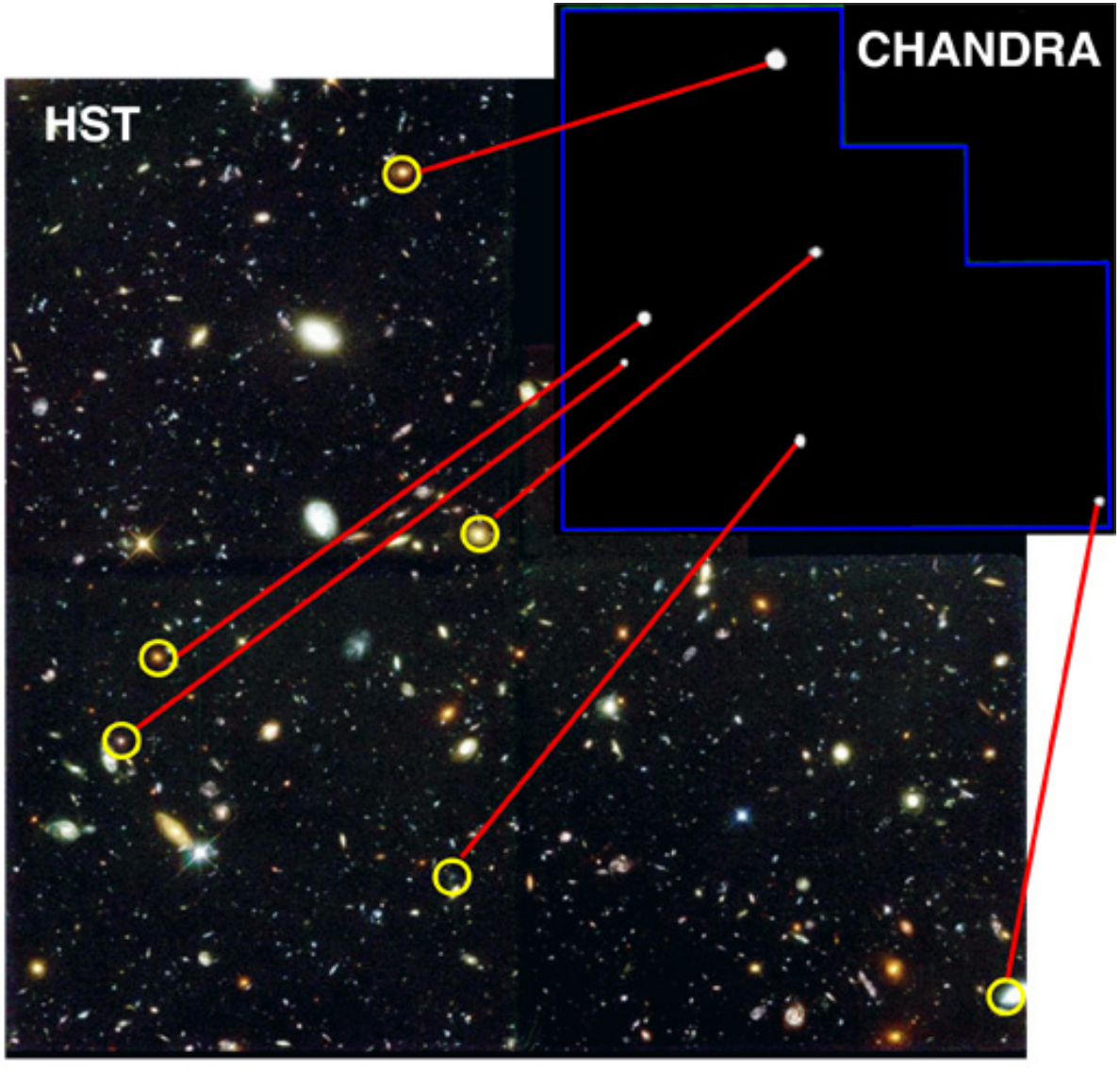
Let

- t_i: ith ordered observation
- d_i: number of 'censored' events at ith ordered observation
- R_i: number of subjects 'at-risk' at ith ordered observation

The Kaplan Meier estimator of the survival function is

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{R_i}\right)$$

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 $P \geq \ell$.

Forty years ago, distinguished astrophysicist Donald Lynden-Bell derived a fundamental statistical result in the appendix to an astronomical study: the unique nonparametric maximum likelihood estimator for a univariate dataset subject to random truncation. The method is needed to establish luminosity functions from flux-limited surveys, and is far better than the more popular heuristic $1/V_{max}$ method by Schmidt (1968). Two young astrostatisticians are now developing Lynden-Bell's method further. Schafer (2007) gives a nonparametric estimation for estimating the bivariate distribution when one variable is truncated. Kelly et al. use a Bayesian approach for a normal mixture model (combination of Gaussians) to the luminosity function.
Lynden-Bell (1971, MNRAS.155, 95)

Lynden-Bell Estimator

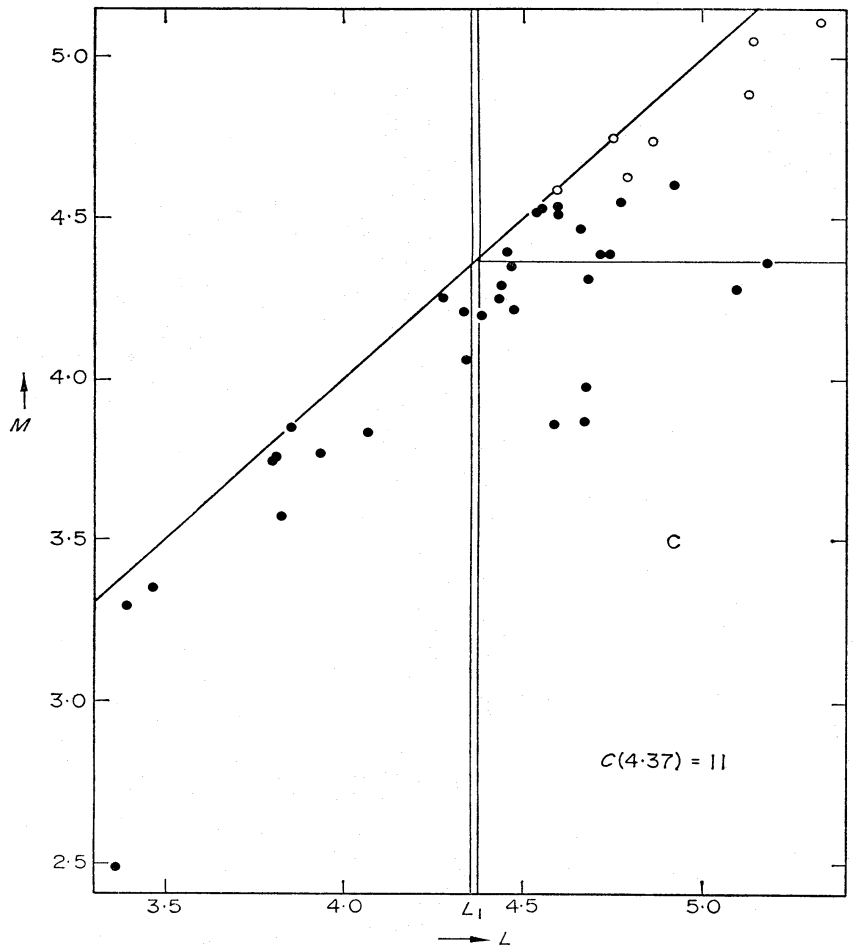


FIG. 1. *Plot of the 40 3CR quasars in the L, M plane. L = P = log* (F_R/F_0) *M = limiting value of L beyond which an object of the same optical flux is rejected from 3CR as being too radio weak. The number of points in the box is denoted by C, the number* (0) *in the infinitesimal column is dX.*

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:

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

  where $u_i = u(x_i)$ and $y_{(i)}$ is ith 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 jth event.
- The only differences in comparable points between truncated cases and censoring cases is that points with $y_i > y_k$ but $t(x_i) > y_k$ are not considered at risk in the truncated case. This is because these points cannot be observed.

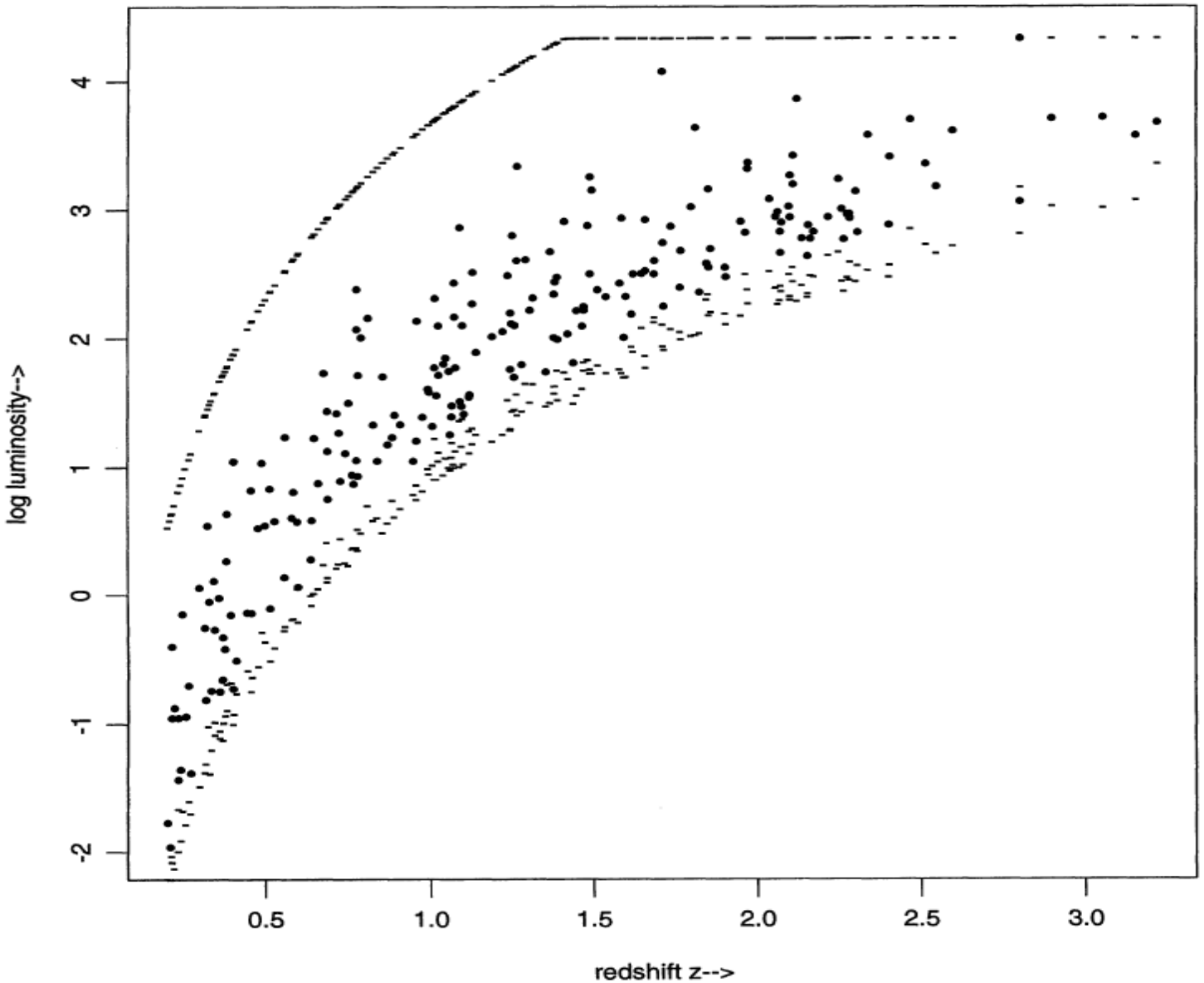
Lynden-Bell- Woodroofe Estimator (continued)

Lyden-Bell-Woodroofe survival function estimator:

$$LBW(t) = \prod_{y_{(j)} \leq t} \left(1 - \frac{1}{N_j}\right)$$

Lynden-Bell, D. 1971, MNRAS, 155, 95
Woodroofe, M. 1985, Ann. Stat., 13, 163

Doubly truncated data: Efron's Nonparametric MLE



15,343 quasars

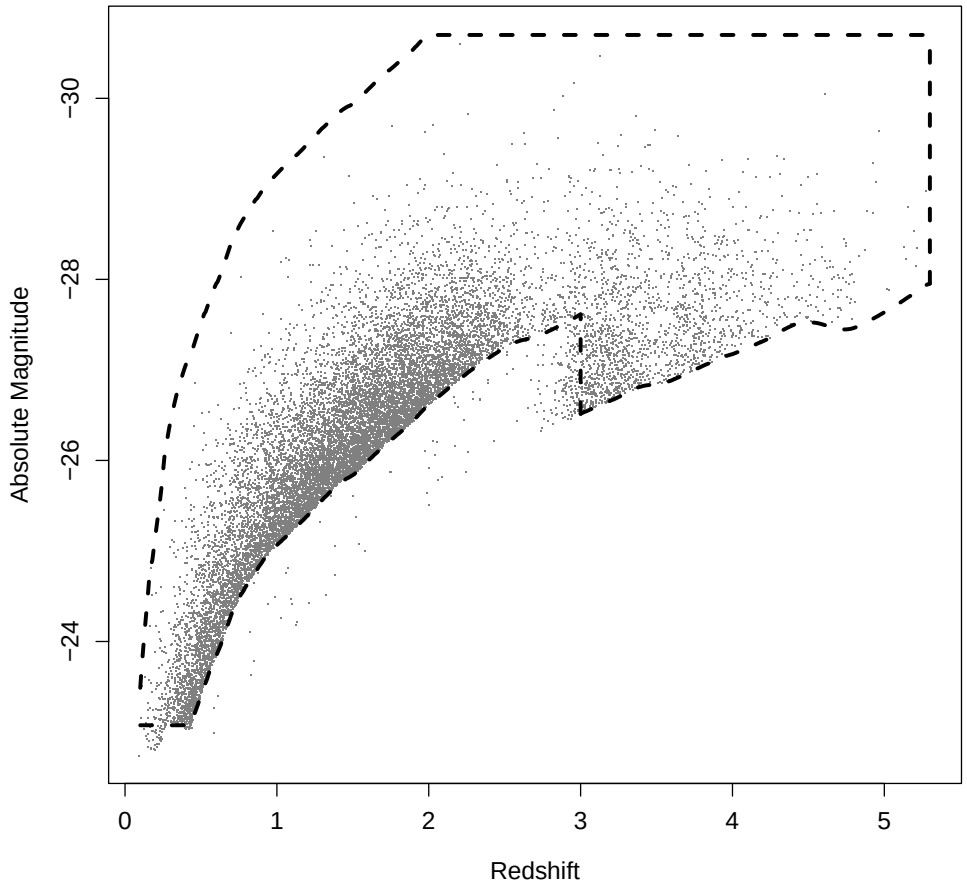


Fig. 1.— Quasar data from the Sloan Digital Sky Survey, the sample from Richards et al. (2006). Quasars within the dashed region are used in this analysis. The removed quasars are those with $M \leq -23.075$, which fall into the irregularly-shaped corner at the lower left of the plot, and those with $z \leq 3$ and apparent magnitude greater than 19.1, which fall into a very sparsely sampled region.

Data

The data shown in the figure, consists of 15,343 quasars. From these, any quasar is removed if it has $z \geq 5.3, z \leq 0.1, M \geq -23.075$, or $M \leq -30.7$. In addition, for quasars of redshift less than 3.0, only those with apparent magnitude between 15.0 and 19.1, inclusive, are kept; for quasars of redshift greater than or equal to 3.0, only those with apparent magnitude between 15.0 and 20.2 are retained. These boundaries combine to create the irregular shape shown by the dashed line. This truncation removes two groups of quasars from the Richards et al. (2006) sample. There are 15,057 quasars remaining after this truncation.

- Schafer, C. M. (2007, ApJ, 661, 703) uses semi-parametric methods $\log \phi(z, M) = f(z) + g(M) + h(z, M.\theta)$, (z_i, M_i) observed.
- Kelly et al. (2008, ApJ, 682, 874) use Bayesian approach for normal mixture model.
- The results obtained are better than the heuristic $1/V_{\max}$ of Schmidt (1968, ApJ, 151, 393)