Right-truncated data

For this data, only individuals for whom the event has occurred by a given date are included in the study. Right truncation can occur in infectious disease studies.

Let $T_i$ denote the infection time and $X_i$ the time between infection and onset of the disease (for example, HIV infection versus development of AIDS). We have a right-truncation time of $\tau$, so that only patients who have the disease prior to $\tau$ are in the study.
Right-censoring versus right-truncation

For AIDS, for example, if we followed a cohort of people who had been infected with HIV, and at the end of the study some of them had AIDS and some didn’t, we would consider those who hadn’t developed AIDS by the end of the study to be right-censored rather than right-truncated.

Instead, if the study only includes people who have developed AIDS, and then we determine how long it took to develop AIDS from the time of the HIV infection, then the data is right-truncnated rather than right-censored since we are not allowing for the possibility of people who had HIV but didn’t develop AIDS.
<table>
<thead>
<tr>
<th>infectionTime</th>
<th>inductionTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>5.5</td>
</tr>
<tr>
<td>1.50</td>
<td>2.25</td>
</tr>
<tr>
<td>2.25</td>
<td>3.0</td>
</tr>
<tr>
<td>2.75</td>
<td>1.0</td>
</tr>
<tr>
<td>3.00</td>
<td>1.75</td>
</tr>
<tr>
<td>3.50</td>
<td>0.75</td>
</tr>
<tr>
<td>3.75</td>
<td>0.75, 1, 2.75, 3, 3.5, 4.25</td>
</tr>
<tr>
<td>4.00</td>
<td>1.0</td>
</tr>
<tr>
<td>4.25</td>
<td>1.75</td>
</tr>
<tr>
<td>4.50</td>
<td>3.25</td>
</tr>
<tr>
<td>4.75</td>
<td>1.0, 2.25</td>
</tr>
<tr>
<td>5.00</td>
<td>0.5, 0.75, 1.5, 2.5</td>
</tr>
<tr>
<td>5.25</td>
<td>0.25, 1.0, 1.5</td>
</tr>
<tr>
<td>5.50</td>
<td>0.5, 1.5, 2.5</td>
</tr>
<tr>
<td>5.75</td>
<td>1.75</td>
</tr>
<tr>
<td>6.00</td>
<td>0.5, 1.25</td>
</tr>
<tr>
<td>6.25</td>
<td>0.5, 1.25</td>
</tr>
<tr>
<td>6.50</td>
<td>0.75</td>
</tr>
<tr>
<td>6.75</td>
<td>0.5, 0.75</td>
</tr>
<tr>
<td>7.00</td>
<td>0.75</td>
</tr>
<tr>
<td>7.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>
For this example, the study had a window of 8 years, so data is right-truncated, but the truncation time depends upon when the infection time took place. Let $R_i = \tau - X_i$. Then the trick of thinking of the $R_i$ values as left-truncated can be used, and the method for left-truncation from the previous section can be used. Here $R_i$ is constrained to be greater than $T_i$, so the $R_i$s are left-truncated at $T_i$.

This procedure estimates $\Pr[R > t|R \geq 0]$ which is equivalent to $\Pr[X < \tau - t|X \leq \tau]$. This function is decreasing in $t$ and increasing but increasing on the original timescale, i.e. $\Pr[X < x|X \leq \tau]$ is increasing in $x$. 


Figure 5.2  Estimated conditional distribution of the induction time for AIDS for the 258 adults (---) and 37 children (-----)
In a cohort life table, we follow a group, traditionally from birth, and record when they die within an interval \((a_{j-1}, a_j], \ j = 1, \ldots, k + 1\). Ideally, all individuals are followed until the last one dies; however, censoring may occur due to moving out of the study. The life table can also be applied to animals, and can be applied to other events, such as time to weaning for infants, first evidence of a certain disease, time to first marriage, etc.

We use the following steps to make a Cohort Life Table

Cohort Life Table
Cohort Life Table

- The first column gives the mutually exclusive and adjacent time intervals $I_j = (a_{j-1}, a_j]$, where $a_0 = 0$, and $a_k = \infty$. Intervals do not need to be equal in length.

- The second column gives the number of individuals $Y_j'$ entering interval $I_j$ who have not experienced the event.

- The third column gives the number of individuals lost to follow up (cannot be reached) in interval $I_j$. We assume that the censoring times are independent of the event times.

- The fourth column gives $Y_j$, the number of intervals at risk of experiencing the event in interval $I_j$ assuming that censoring times are uniformly distributed over the interval, so that $Y_j = Y_j' - W_j/2$.

- The fifth column gives the number of individuals $d_j$ who experienced the event in interval $I_j$.

- The sixth column gives the estimated survival function at the start of the interval $\hat{S}(a_{j-1})$.
Cohort Life Table

The survival probabilities follow the Kaplan-Meier formula. The main difference is in how the $Y_i$ values are defined since we average over the interval when dealing with censoring instead of using exact time points.

$$\hat{S}(a_0) = 1; \quad \hat{S}(a_j) = \hat{S}(a_{j-1}) (1 - d_j/Y_j) = \prod_{i=1}^{j} (1 - d_i/Y_i)$$
Cohort Life Table

Other columns:

- The seventh column gives the estimated probability density at the midpoint of the $j$th interval

$$\hat{f}(a_{mj}) = \frac{\hat{S}(a_{j-1}) - \hat{S}(a_j)}{a_j - a_{j-1}}$$

(This looks like a difference quotient from calculus)

- The eighth column gives the estimated hazard rate, based on a linear approximation for the survival function:

$$\hat{h}(a_{mj}) = \frac{\hat{f}(a_{mj})}{\hat{S}(a_{mj})}$$

$$= \frac{\hat{f}(a_{mj})}{\{\hat{S}(a_j) + [\hat{S}(a_{j-1}) + \hat{S}(a_j)]/2\}}$$

$$= \frac{2\hat{f}(a_{mj})}{\hat{S}(a_{j-1}) + \hat{S}(a_j)}$$
Cohort Life Table: other columns

- The 9th column is the standard error estimate at the beginning of the \( j \)th interval, which is the same as for the Kaplan-Meier,
  \[
  \hat{SE}[\hat{S}(a_{j-1})] = \hat{S}_{a_{j-1}} \sqrt{\sum_{i=1}^{j-1} \frac{d_i}{Y_i(Y_i-d_i)}}
  \]

- The 10th column gives the standard error estimate at the midpoint of the \( j \)th interval which is
  \[
  \frac{\hat{S}(a_{j-1})\hat{q}_j}{a_j - a_{j-1}} \sqrt{\sum_{i=1}^{j-1} \frac{\hat{q}_i/(Y_i\hat{p}_i) + \hat{p}_j/(Y_j\hat{q}_j)}{\hat{q}_j = d_j/Y_j \text{ and } \hat{p}_j = 1 - \hat{q}_j}}
  \]

- Finally, the last column gives the estimate standard deviation of the hazard function at the midpoint of the \( j \)th interval as
  \[
  \left\{ \frac{1 - [\hat{h}(a_{mj})(a_j - a_{j-1})/2]^2}{d_j} \right\}^{1/2} \cdot \hat{h}(a_{mj})
  \]
### TABLE 5.6
Life Table for Weaning Example

<table>
<thead>
<tr>
<th>Week weaned (lower, upper)</th>
<th>Number of infants not weaned entering interval</th>
<th>Number lost to follow-up or withdrawn without being weaned</th>
<th>Number exposed to weaning</th>
<th>Number weaned</th>
<th>Est. Cum. proportion not weaned at beginning of interval</th>
<th>Est. p.d.f. at middle of interval</th>
<th>Est. hazard at middle of interval</th>
<th>Est. stand. dev. of survival at beginning of interval</th>
<th>Est. stand. dev. of p.d.f. at middle of interval</th>
<th>Est. stand. dev. of hazard at middle of interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>0– 2</td>
<td>927</td>
<td>2</td>
<td>926</td>
<td>77</td>
<td>1.0000</td>
<td>0.0416</td>
<td>0.0434</td>
<td>0.0091</td>
<td>0.0088</td>
<td>0.0104</td>
</tr>
<tr>
<td>2– 3</td>
<td>848</td>
<td>3</td>
<td>846.5</td>
<td>71</td>
<td>0.9168</td>
<td>0.0769</td>
<td>0.0875</td>
<td>0.0121</td>
<td>0.0055</td>
<td>0.0076</td>
</tr>
<tr>
<td>3– 5</td>
<td>774</td>
<td>6</td>
<td>771</td>
<td>119</td>
<td>0.8399</td>
<td>0.0648</td>
<td>0.0836</td>
<td>0.0149</td>
<td>0.0046</td>
<td>0.0071</td>
</tr>
<tr>
<td>5– 7</td>
<td>649</td>
<td>9</td>
<td>644.5</td>
<td>75</td>
<td>0.7103</td>
<td>0.0413</td>
<td>0.0618</td>
<td>0.0160</td>
<td>0.0027</td>
<td>0.0051</td>
</tr>
<tr>
<td>7–11</td>
<td>565</td>
<td>7</td>
<td>561.5</td>
<td>109</td>
<td>0.6276</td>
<td>0.0305</td>
<td>0.0537</td>
<td>0.0166</td>
<td>0.0021</td>
<td>0.0053</td>
</tr>
<tr>
<td>11–17</td>
<td>449</td>
<td>5</td>
<td>446.5</td>
<td>148</td>
<td>0.5058</td>
<td>0.0279</td>
<td>0.0662</td>
<td>0.0158</td>
<td>0.0014</td>
<td>0.0052</td>
</tr>
<tr>
<td>17–25</td>
<td>296</td>
<td>3</td>
<td>294.5</td>
<td>107</td>
<td>0.3381</td>
<td>0.0154</td>
<td>0.0555</td>
<td>0.0138</td>
<td>0.0008</td>
<td>0.0047</td>
</tr>
<tr>
<td>25–37</td>
<td>186</td>
<td>0</td>
<td>186</td>
<td>74</td>
<td>0.2153</td>
<td>0.0071</td>
<td>0.0414</td>
<td>0.0114</td>
<td>0.0006</td>
<td>0.0066</td>
</tr>
<tr>
<td>37–53</td>
<td>112</td>
<td>0</td>
<td>112</td>
<td>85</td>
<td>0.1296</td>
<td>0.0061</td>
<td>0.0764</td>
<td>0.0059</td>
<td></td>
<td></td>
</tr>
<tr>
<td>53–</td>
<td>27</td>
<td>0</td>
<td>27</td>
<td>27</td>
<td>0.0313</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cohort Life Table: weaning example Hazard function

![Graph showing the estimated hazard rate over time to weaning. The x-axis represents time to weaning in weeks, ranging from 0 to 40. The y-axis represents the estimated hazard rate, ranging from 0.04 to 0.1.]
The hazard function for this example is initially high (indicate high probability that mother will wean the infant soon), then decreases, then goes up again.
Median survival time

If the last observation is right-censored, then the mean survival time isn’t well-defined, so the median is often used. For the median survival time, first determine the interval \( I_j = (a_{j-1}, a_j] \) where \( \hat{S}(a_{j-1}) \geq 0.5 \) and \( \hat{S}(a_j) \leq 0.5 \). Then the median survival time can be estimated by interpolation as

\[
\hat{x}_{0.5} = a_j + \left[ \frac{a_j - a_{j-1}}{\hat{S}(a_{j-1}) - \hat{S}(a_j)} \right] \left( \hat{S}(a_{j-1}) - 0.5 \right) = a_j + \left[ \frac{\hat{S}(a_{j-1}) - 0.5}{\hat{f}(a_{mj})} \right] = a_j + \left[ \frac{\hat{S}(a_{j-1}) - 0.5}{\hat{f}(a_{mj})} \right]
\]

The derivation of this uses college algebra. Draw a line through \((a_{j-1}, \hat{S}a(j - 1))\) and \((a_j, \hat{S}(a_j))\). Determine the point on this line that goes through 0.5 on the \( y \)-axis, and you are done.
You can also determine the median residual life. From a certain time point $a_{j-1}$ the median residual life is the amount of time added to $a_{j-1}$ for which half of the remaining cohort is expected to survive. For example, if 40% of the cohort have died by time $a_{j-1}$, then the median residual life is the additional time expected until 70% have died.
5.8 Individuals seen at a large city sexually transmitted disease (STD) clinic are considered at high risk for acquiring HIV. The following data is recorded on 100 high-risk individuals who are infected with some STD, but did not have HIV, when they were seen at the clinic in 1980. Their records were checked at subsequent visits to determine the time that HIV was first detected.

<table>
<thead>
<tr>
<th>Year Intervals</th>
<th>Number of HIV-Positive</th>
<th>Number Lost to Follow-Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2–4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4–6</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>6–8</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>8–10</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>10–12</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>12–14</td>
<td>3</td>
<td>21</td>
</tr>
</tbody>
</table>

Construct a cohort life table for the incidence of HIV.
To do this problem, note that the number of people entering the study was 100. To construct the table, the first column in the form of time intervals is given.

To construct the second column, the number of individuals entering the interval who have not experienced the event, we consider the number who entered the previous interval without experiencing the event, minus the number who experienced the event in the previous interval, and minus the number who were lost to follow up in the previous interval.
It is easiest to set this up in Excel. Here I just did up to the estimated standard errors for the survival function. For this, it is easiest to construct a column for $d_j / [(Y_j(Y_j - d_j))$, then another column for $\sum_{i=1}^{j} d_j / [(Y_j(Y_j - d_j))$ before getting the standard error.
This chapter considers miscellaneous topics before going into hypothesis testing (Chapter 7). Topics include

- smoothing estimates of the hazard function
- describing excess mortality compared to a reference population
- Bayesian nonparametric approaches for right-censored data.
The Nelson-Aalen estimator (Chapter 4 and week 4 of Notes) is typically used to estimate the cumulative hazard function as

$$\hat{H}(t) = \begin{cases} 0, & \text{if } t \leq t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i}, & \text{if } t > t_1 \end{cases}$$

The survival function is related to the cumulative hazard by

$$\hat{S}(t) = \exp[-\hat{H}(t)].$$

The hazard is the derivative (i.e., slope) of the cumulative hazard, which is crude for a step function. So a smoothed version of the cumulative hazard is often used to make a more reasonable estimate of the hazard function.
Before jumping in to the idea of smoothing the hazard, we’ll talk about smoothing in a simpler setting first.

Smoothing is also used for nonparametric density estimation.

In particular, think of a histogram, and remove the bars underneath. Then the remaining horizontal lines form a rough estimate of the density function.
Histogram as density estimate

> x <- rnorm(100)
> a <- hist(x, prob=T, cex.axis=1.3, cex.lab=1.3, cex.main=1.3,
border=FALSE, col="grey")
names(a)
[1] "breaks" "counts" "density" "mids" "xname" "equidist"
> a$density
[1] 0.02 0.06 0.04 0.22 0.30 0.32 0.40 0.36 0.16 0.08 0.04
> plot(a$breaks, a$density, type="s")
Error in xy.coords(x, y, xlabel, ylabel, log) :
  'x' and 'y' lengths differ
> points(a$mids, a$density, type="s")
> points(a$mids-.25, a$density, type="s")
Histogram as density estimate
Histogram as density estimate
Histogram as density estimate

This isn’t quite right because the height of the histogram should start at the left-endpoint of the bin, not the midpoint, so you need to shift if over or use the `a$breaks` variable

```r
> a <- hist(x, prob=T, cex.axis=1.3, cex.lab=1.3, cex.main=1.3, border=FALSE, col="grey")
> points(a$breaks[1:length(a$density)], a$density, type="s", lwd=3)
```
Histogram as density estimate
Histogram as density estimate

Now we’ll remove the grey, and just think of this as an estimate of the standard normal density.

```r
> a <- hist(x, prob=T, cex.axis=1.3, cex.lab=1.3, cex.main=1.3, border=FALSE)
> x2 <- ((1:100)/100)*6 - 3
> points(x2, dnorm(x2), type="l", lwd=2)
> points(a$breaks[1:length(a$density)], a$density, type="s", lwd=3, lty=3)
> legend(-3,.4, legend=c("true","estimated"), lty=c(1,2), cex=1.5)
> legend(-3,.4, legend=c("true","estimated"), lty=c(1,2), cex=1.5)
```
Histogram as density estimate
Histogram as a density estimate

Of course, with more data, the better the histogram estimates the density.

```r
> x <- rnorm(100000)
> a <- hist(x, prob=T, cex.axis=1.3, cex.lab=1.3, cex.main=1.3, border=FALSE, nclass=50)
> x2 <- ((1:100)/100)*8 - 4
> points(x2, dnorm(x2), type="l", lwd=2)
> points(a$breaks[1:length(a$density)], a$density, type="s", lwd=3, lty=1, col="red")
> legend(-4.2, 0.4, legend=c("true", "estimated"), col=c("black", "red"), cex=1.3, lty=c(1,1))
```
Histogram as density estimate

Histogram of x

true
estimated

Density

x

-4 -2  0  2  4
One way of thinking about what the histogram does is that it stacks rectangular blocks on top of each other for each observation. The location of the block depends on which bin the observation occurs in. If there are 10 observations between -3 and -2, then it stacks 10 blocks in that category. If there are 15 between -2 and -1, then it stacks 15 blocks in that category. Each block is the same height and width (the width of the bin). The heights are chosen so that the total area under the curve is equal to 1.
The blocky of appearance of the histogram is due to this business of stacking blocks. If we want a smoother estimate of the density, we can imagine stacking objects that aren’t so....blocky.

The idea is that instead of adding a rectangle at every data point (within a bin), we instead put the value of a function at the observed data point without worrying about binning, and add up these functions. The functions are scaled in such a way that the total area under the curve will still be 1.0.
Kernel density estimation

If a square Kernel is used, then you add one unit of height, but instead of putting it into a bin, you center it wherever the data happens to be.
Let’s say a normal Kernel is used. Then if you observe data $x_1$, $x_2$, and $x_3$ (only three data points), you draw normal curves centered at $x_1$, $x_2$ and $x_3$. The Kernel density estimate at a point say $x$ distinct from the data points, is the sum of the normal curves evaluated at $x$.

The total density estimate is therefore a sum of normals with different means, but the same variance, and this leads to a smooth, though possibly multimodal, density estimate.
Kernel density estimation: normal Kernel
Kernel density estimation

To see how this works in more detail, we’ll try the example with 3 data points. Say $x_1 = 1.2$, $x_2 = 1.5$, and $x_3 = 3.1$. To use a Kernel function, you need to specify a bandwidth $h$ as well as the Kernel function. The Kernel density estimate at a point $x$ can be written as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - x_i}{h} \right)$$

where $K$ is the Kernel function. For a rectangular (also called uniform) Kernel, we might let $K(u) = 1/2(-1 \leq u \leq 1)$ This adds a rectangular block at each point $x_i$. For example, we have

$$\hat{f}(2) = \frac{1}{3 \cdot 1} \left[ K \left( \frac{2 - 1.2}{1} \right) + K \left( \frac{2 - 1.5}{1} \right) + K \left( \frac{2 - 3.1}{1} \right) \right] = \frac{1}{3} \cdot \left( \frac{1}{2} + \frac{1}{2} + 0 \right)$$
For the same data, if we used the standard normal kernel with bin width 2, we have \( K(u) = \phi(u) \), and we use

\[
\hat{f}(x) = \frac{1}{2n} \sum_{i=1}^{n} K \left( \frac{x - x_i}{2} \right)
\]

We then get

\[
\hat{f}(2) = \frac{1}{6} (\phi(0.4) + \phi(0.25) + \phi(-0.55))
\]

\[
> (1/6)*(dnorm(.4) + dnorm(.25) + dnorm(-.55))
\]

[1] 0.1829804
If instead we had used a binwidth of $h = 1$, then we’d have

$$\hat{f}(2) = \frac{1}{3} (\phi(0.8) + \phi(0.5) + \phi(-1.1)) = 0.2865364$$

This changes the answer quite a bit (and makes it closer to the uniform Kernel answer), but this is a small sample size, making the answer more sensitive to bin width
To get an idea of how to do this in R, we can use, for a bandwidth of 0.5122:

```r
> (1/(3*.5122)) *(dnorm((1-x[1])/0.5122) + dnorm((1-x[2])/0.5122) +
+ dnorm((1-x[3])/0.5122))
[1] 0.4018495
```

to get $\hat{f}(1)$.

```r
> plot(xaxis,(1/(3*.5122)) *(dnorm((xaxis-x[1])/0.5122) +
+ dnorm((xaxis-x[2])/0.5122) + dnorm((xaxis-x[3])/0.5122)),
main="",cex.axis=1.3,cex.lab=1.3,ylab="Density estimate",
type="l",xlab="x")
```

plots the whole curve from -4 to 4.
Kernel density estimation in R: bandwidth = 0.5
Kernel density estimation in R: bandwidth = 1.0

N = 3  Bandwidth = 1
Of course, there is an easier way, which is to use the \texttt{density()} function in R. The default is a Gaussian kernel, but others are possible also. It uses its own algorithm to determine the bin width, but you can override and choose your own. If you rely on the \texttt{density()} function, you are limited to the built-in kernels. If you want to try a different one, you have to write the code yourself.
Kernel density estimation in R: effect of bandwidth for rectangular kernel

```
density.default(x = x, bw = 0.5, kernel = "rec")
```

```
density.default(x = x, bw = 2, kernel = "rec")
```

```
density.default(x = x, bw = 0.1, kernel = "rec")
```

Histogram of x

```
density.default(x = x, bw = 0.5, kernel = "rec")
```

```
density.default(x = x, bw = 2, kernel = "rec")
```

```
density.default(x = x, bw = 0.1, kernel = "rec")
```

```
Histogram of x
```

N = 3  Bandwidth = 0.5

N = 3  Bandwidth = 2

N = 3  Bandwidth = 0.1

N = 3  x
There are lots of popular Kernel density estimates, and statisticians have put a lot of work into establishing their properties, showing when some Kernels work better than others (for example, using mean integrated square error as a criterion), determining how to choose bandwidths, and so on.