B.5 Censoring and truncation, Kaplan-Meier estimator

1. (a) Let \( T_j \) be the total time spent in state \( j \) before death, and \( W \) the total value of a life. Then letting \( X(t) \) be the state of the individual at age \( t \) (and \( w_D = 0 \) for the absorbing state),

\[
W = \int_0^{\infty} w_{X(t)} dt = \int_0^{\infty} \sum_j w_j 1_{\{X(t)=j\}} dt = \sum_j w_j T_j.
\]

Thus \( E[W] = \sum w_j E[T_j] \).

We have the general result that \( E_i[T_j] = (-Q^{-1})*_{i,j} \) (see for example Theorem 3 in Section 10.5 of the lecture notes online). If we start in state \( i \) with probability \( p_i \), this expected total time becomes \( \sum_i p_i(-Q^{-1})*_{i,j} \), and

\[
E[W] = \sum_{i,j} p_i(-Q^{-1})*_{i,j} w_j = -p^T Q^{-1} w.
\]

(b) We have the Healthy-Sick-Dead model with \( \sigma = 0.1, \rho = 0, \delta = 0.01, \) and \( \gamma = 0.2 \). What is the life expectancy of someone initially healthy?

We have

\[
Q = \begin{pmatrix} -0.11 & 0.1 & 0.01 \\ 0 & -0.2 & 0.2 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q_* = \begin{pmatrix} -0.11 & 0.1 \\ 0.2 & -0.4 & 0.2 \\ 0 & 0 & 0 \end{pmatrix}, \quad -Q_*^{-1} = \begin{pmatrix} 9.1 & 4.55 \\ 0 & 5 \end{pmatrix}
\]

So, with \( w = (1, 1)^T \) and \( p = (1, 0)^T \) for the healthy start, the healthy person has, on average 13.65 remaining years.

In this case, the calculation would have been easy to do directly. The remaining lifetime of an initially healthy person consists of a stay in state \( H \) which is exponential with mean \( (\sigma + \delta)^{-1} \), followed by, with probability \( \sigma/(\sigma + \delta) \), a stay in state \( S \) which is exponential with mean \( \gamma^{-1} \).

(c) The healthy person has, on average, 9.1 remaining healthy years and 4.55 sick years, which count for 2.3 QALYs. Thus, 11.4 QALYs altogether. The sick person has 2.5 QALYs on average.

(d) We now have

\[
Q = \begin{pmatrix} -0.11 & 0.1 & 0.01 \\ 0.2 & -0.4 & 0.2 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q_* = \begin{pmatrix} -0.11 & 0.1 \\ 0.2 & -0.4 \\ 0 & 0 \end{pmatrix}, \quad -Q_*^{-1} = \begin{pmatrix} 16.7 & 4.2 \\ 8.3 & 4.6 \end{pmatrix}
\]

Thus, we would be increasing the QALY expectancy for healthy people to 18.8, and for sick people to 10.6.

Obviously, we could achieve unlimited life-expectancy increase by reducing \( \gamma \) down to 0. We can compute this exactly by

\[
Q_* = \begin{pmatrix} -0.11 & 0.1 \\ 0 & -\gamma \end{pmatrix}, \quad -Q_*^{-1} = \begin{pmatrix} \frac{1}{19.4} & \frac{0.91}{19.4} \\ 0 & \frac{1}{19.4} \end{pmatrix}
\]

Thus, to get sick people up to 10.6 expected QALYs we need \( \gamma = 1/21.2 = 0.047 \); and to get healthy people up to 18.8 QALYs we need \( 0.91/\gamma = 19.4 \), which is also \( \gamma = 0.047 \).
(e) We have a general method for questions like this as covered in lectures. We write the state space with absorbing states $D_H$ and $D_S$, representing what state the individual died from. The generator then becomes

$$Q = \begin{pmatrix}
-0.11 & 0.1 & 0.01 & 0 \\
0.2 & -0.4 & 0 & 0.2 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{pmatrix}. $$

We want to find the probability of being absorbed in state $D_S$. Let $v$ be the vector defined by letting $v_i$ be the probability of absorption in $D_S$ when starting from state $i$. Then the recursions for absorption probabilities (see also Theorem 4 in Section 10.5 of the notes) correspond to the relation $Qv = 0$, i.e. $v$ is a right eigenvector of $Q$ with eigenvalue 0.

We can find a basis for the eigenvectors with eigenvalue 0:

$$\begin{pmatrix}
0.17 \\
0.083 \\
1 \\
0
\end{pmatrix} \quad \begin{pmatrix}
0.83 \\
0.917 \\
0 \\
1
\end{pmatrix}. $$

But by definition of $v$ we also know that $v$ takes value 0 for state $D_H$ and 1 for state $D_S$. So $v$ must be the second vector given above.

So an individual now healthy will die sick with probability 0.83. Asymptotically, this will be the fraction of all individuals initially healthy. For those initially sick, it’s 0.917. If the initial population contains a proportion $\beta$ of sick people, then it’s the weighted average $0.917\beta + 0.83(1 - \beta)$.

(f) We want to estimate the average number of years spent in the healthy and sick states. We begin at the end: The remaining life of a healthy or sick “old” person. We get the matrix

$$-(Q^{\text{old}})^{-1} = \begin{pmatrix}
3.85 & 4.81 \\
1.92 & 4.90
\end{pmatrix},$$

so that the worth of the remaining life is 6.25 or 4.37 when starting from H or S respectively.

For the first 30 years, we can calculate $P_s(30)$ either using forward equations $P'(t) = P(t)Q$ or using $P(t) = e^{tQ}$. For the latter, we diagonalise the $Q_s$ for young individuals as

$$Q_s = \begin{pmatrix}
-0.11 & 0.1 \\
0.2 & -0.4
\end{pmatrix} = \begin{pmatrix}
-0.28 & -0.87 \\
0.96 & -0.50
\end{pmatrix} \begin{pmatrix}
-0.46 & 0 \\
0 & -0.052
\end{pmatrix} \begin{pmatrix}
-0.51 & 0.89 \\
0.99 & 0.28
\end{pmatrix}.$$  

Actually the resulting distribution after 30 years will depend only on the eigenvector with eigenvalue $-0.052$, but let us give the exact computation here:

$$P_s(30) = e^{30Q_s} = \begin{pmatrix}
-0.28 & -0.87 \\
0.96 & -0.50
\end{pmatrix} \begin{pmatrix}
e^{-30 \times 0.46} & 0 \\
e^{-30 \times 0.052}
\end{pmatrix} \begin{pmatrix}
-0.51 & 0.89 \\
0.99 & 0.28
\end{pmatrix} = \begin{pmatrix}
0.18 & 0.05 \\
0.10 & 0.03
\end{pmatrix}.$$  

This gives the probability that, after 30 years, a person is in states H and S (columns) when starting from H or S (rows).
As above (via Theorem 3) we have that
\[ E(30) \left( \begin{array}{c} 1 \\ 0.5 \end{array} \right) = Q^{-1} \left( P_s(30) - I \right) \left( \begin{array}{c} 1 \\ 0.5 \end{array} \right) = \left( \begin{array}{c} 14.89 \\ 8.37 \end{array} \right) \]

Thus, the expected number of QALYs before age 30 for a healthy person is 14.89, and after age 30 is
\[ 0.18 \cdot 6.25 + 0.05 \times 4.37 = 1.34, \]
so 16.23 in all. For a sick person, the corresponding numbers are 8.37 before age 30, and
\[ 0.10 \cdot 6.25 + 0.03 \times 4.37 = 0.76, \]
so 9.13 in all.

2. (a) Observations could be either left-censored (children who already possess the skill at 12 months) or right-censored (children who do not gain the skill until after 18 months). Right censoring could also result from a child leaving the trial before the age of 18 months, for whatever reason. In addition, one could say there is interval censoring (we see only in which month the skill was acquired), but this is essentially just standard discretisation.

(b) There is right censoring because the study is of fixed length (and also individuals might leave the study because of death or some other reason). There is left truncation, because if an individual experiences a second episode too soon (before the start of the trial) they are not enrolled in the trial.

(c) For an observed individual, an observation is interval censored; if the time in the region is reasonably long, then this is effectively right censoring.

There is also left truncation; if the incubation time is short enough, the symptoms appear abroad rather than in the UK.

[Analysis could be difficult in this case since for the individuals that we observe, we cannot directly measure their left truncation time – i.e. their “time” of entry into the number at risk, where “time” is the time since infection.]

There is also theoretically right truncation; when doing the study, there could be potential patients we fail to observe because they have not yet developed symptoms; but since incubation times are typically short, this is not likely to be a major concern.

3. \( L \propto f(x_1)f(x_2)\ldots f(x_k)S(x_k)^{n-k}. \)

For the Kaplan Meier curve: \( \hat{F}(t) = \frac{1}{n} \# \{ i \leq k : x_i \leq t \} \) for \( t \leq x_k. \)

For \( t > x_k \) we have no estimate.

4. Suppose \( t_1 < t_2 < \cdots < t_k \) are times at which events occur. The Kaplan-Meier estimator for the survival function is
\[ \hat{S}(t) = \prod_{t_i \leq t} \left( 1 - \frac{d_i}{n_i} \right), \]
where \( n_i \) is the number at risk at time \( t_i \) and \( d_i \) is the number of events at \( t_i \). If there is
no censoring, then \( n_{j+1} = n_j - d_j = n_0 - \sum_{i=1}^{j} d_i \). Thus

\[
\hat{S}(t) = \prod_{t_i \leq t} \left( \frac{n_i - d_i}{n_i} \right) = \prod_{t_i \leq t} \left( \frac{n_{i+1}}{n_i} \right) = \frac{n_{j+1}}{n_0} \text{ where } j = \max\{i : t_i \leq t\}
\]

\[= 1 - n_0^{-1} \sum_{i : t_i \leq t} d_i = 1 - \hat{F}(t).\]

5. (a) As we did for Greenwood’s formula in the lecture, we approximate by assuming a discrete lifetime distribution, with only a finite number of values less than \( t \) which the lifetime can take, say \( t_1, t_2, \ldots, t_m \).

The Nelson-Aalen estimator is

\[
\tilde{S}(t) = \exp \left( - \sum_{t_i \leq t} \frac{d_i}{n_i} \right)
\]

where \( d_i \) is the number of deaths at \( t_i \) and \( n_i \) is the number at risk at \( t_i \).

Let \( h_i \) be the discrete hazard of the distribution at time \( t_i \); i.e. given \( n_i, d_i \sim \text{Binomial}(n_i, h_i) \). So \( h_i = 1 - S(t_i+)/S(t_i-) \).

We have the maximum likelihood estimator \( \hat{h}_i = d_i/n_i \), and as in the lecture we have by considering the Fisher information that

(i) \( \hat{h}_i \) and \( \hat{h}_j \) are asymptotically independent for \( i \neq j \);

(ii) \[
\text{Var}(\hat{h}_i) \approx \frac{1}{n_i^2} n_i h_i (1 - h_i) \approx \frac{1}{n_i} \hat{h}_i (1 - \hat{h}_i) = \frac{1}{n_i} \frac{d_i}{n_i} \left( 1 - \frac{d_i}{n_i} \right) = \frac{d_i (n_i - d_i)}{n_i^3}.
\]

So

\[
\text{Var} \log \tilde{S}(t) = \text{Var} \left( - \sum_{t_i \leq t} \hat{h}_i \right) \\
\approx \sum_{t_i \leq t} \frac{d_i (n_i - d_i)}{n_i^3}.
\]

As in the case of Greenwood’s formula, we might well be happy to work with \( \text{Var} \log \tilde{S}(t) \) rather than with \( \text{Var} \tilde{S}(t) \) itself; but for the latter we can use the delta method with \( \log \tilde{S}(t) \approx \log S(t) \) to give

\[
\text{Var} \tilde{S}(t) = \text{Var} e^{\log \tilde{S}(t)} \\
\approx \left( \frac{d}{dx} e^x \bigg|_{x = \log S(t)} \right)^2 \text{Var} \log \tilde{S}(t) \\
= S(t)^2 \text{Var} \log \tilde{S}(t) \\
\approx S(t)^2 \sum_{t_i \leq t} \frac{d_i (n_i - d_i)}{n_i^3}.
\]
(b) We start from

\[ \text{Var} \log \hat{S}(t) \approx \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \]

Using the delta method,

\[ \text{Var} \log(- \log \hat{S}(t)) \approx \left( \frac{d}{dx} \log x \bigg|_{x = -\log \hat{S}(t)} \right)^2 \text{Var}(- \log \hat{S}(t)) \]

\[ \approx \left( \log \hat{S}(t) \right)^{-2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \]

Calling this \( \sigma^2 \), we have an approximate \((1 - \alpha)\) confidence interval for \(- \log S(t)\) given by

\[ \log(- \log \hat{S}(t)) \pm \sigma z \]

where \( z = z_{1-\alpha/2} \) is the appropriate standard normal quantile, giving an approximate C.I. of

\[ e^{\pm \sigma z} \log \hat{S}(t) \]

for \( \log S(t) \), and finally

\[ (\hat{S}(t)^{\exp(\sigma z)}, \hat{S}(t)^{\exp(-\sigma z)}) \]

for the survival probability \( S(t) \) itself.

6. \( \hat{S}(4) = \frac{14}{15} \frac{13}{14} \frac{12}{13} \frac{10}{11} \frac{7}{8} = 0.6364 \). We have

\[ \text{Var}(\log \hat{S}(4)) \approx \sum_{t_i < 4} \frac{d_i}{n_i(d_i - n_i)} \]

\[ = \frac{1}{15 \times 14} + \frac{1}{14 \times 13} + \frac{1}{13 \times 12} + \frac{1}{11 \times 10} + \frac{1}{8 \times 7} \]

\[ = 0.04361, \]

with \( \log \hat{S}(4) = -0.4519 \) and \( \log(- \log \hat{S}(4)) = -0.7942 \). Then Greenwood’s formula is

\[ \text{Var}(\hat{S}(4)) \approx \hat{S}(4)^2 \sum_{t_i < 4} \frac{d_i}{n_i(d_i - n_i)} \]

\[ = 0.6364^2 \times 0.04361 \]

\[ = 0.01766. \]

If we wanted to follow the approach in part (b) of the previous question, we could also calculate

\[ \text{Var} \log(- \log \hat{S}(4)) \approx \frac{1}{(\log \hat{S}(4))^2} \sum_{t_i < 4} \frac{d_i}{n_i(d_i - n_i)} \]

\[ = 0.2135. \]

These lead variously to approximate confidence intervals of

\[ 0.6364 \pm 1.96\sqrt{0.01766} = (0.3759, 0.8969) \]

\[ \exp \left( -0.4519 \pm 1.96\sqrt{0.04361} \right) = (0.4227, 0.9583) \]

\[ \exp \left( - \exp \left( -0.7492 \pm 1.96\sqrt{0.2135} \right) \right) = (0.3270, 0.8330). \]
Of course the sample size is small, so we should not put too much faith in any of these approximate intervals.

7. (a) Right censoring and left truncation.

(b) We use the actuarial estimator. If \( n_i \) is the number at risk at exact age \( i \), then we approximate the number at risk in the interval \((i, i+1)\) by \( \tilde{n}_i = (n_i + n_{i+1})/2 \).

Put another way, \( \tilde{n}_i = n_i + (a_i - d_i - c_i)/2 \), where \( a_i \) is the number arriving at curtate age \( i \), \( d_i \) is the number dying at that age, and \( c_i \) is the number censored (i.e. leaving the population for other reasons, or surviving at the end of the period studied) at that age.

This gives the following estimates for the number at risk:

<table>
<thead>
<tr>
<th>Age</th>
<th>Arrivals</th>
<th>Deaths</th>
<th>Other Exits</th>
<th>Number at risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>69</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>70</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>7.5</td>
</tr>
<tr>
<td>71</td>
<td>24</td>
<td>5</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>72</td>
<td>24</td>
<td>6</td>
<td>3</td>
<td>34.5</td>
</tr>
<tr>
<td>73</td>
<td>11</td>
<td>9</td>
<td>9</td>
<td>38.5</td>
</tr>
<tr>
<td>74</td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>35</td>
</tr>
<tr>
<td>75</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>32.5</td>
</tr>
<tr>
<td>76</td>
<td>4</td>
<td>6</td>
<td>2</td>
<td>28</td>
</tr>
</tbody>
</table>

(c) Our estimate of the probability of surviving from age 71 to age 76 is then

\[
\hat{S} = \prod_{i=71}^{75} \left( 1 - \frac{d_i}{\tilde{n}_i} \right) = \left( 1 - \frac{5}{19} \right) \cdots \left( 1 - \frac{5}{32.5} \right) = 0.3044,
\]

giving \( \log \hat{S} = -1.1893 \). We can estimate variances by

\[
\text{Var} \hat{S} \approx \hat{S}^2 \sum_{i=71}^{75} \frac{d_i}{\tilde{n}_i(\tilde{n}_i - d_i)} = 0.004345
\]

\[
\text{Var} \log \hat{S} \approx \sum_{i=71}^{75} \frac{d_i}{\tilde{n}_i(\tilde{n}_i - d_i)} = 0.04688.
\]

Using the first one gives an approximate 95% confidence interval for \( S \) of \( \hat{S} \pm 1.96 \sqrt{0.004345} = (0.1752, 0.4336) \).

Alternatively, the second one gives a C.I. for \( \log S \) of \((-1.6137, -0.7649)\), leading to a C.I. for \( S \) of \((0.1992, 0.4654)\).

(d) We assume that both the censoring and the truncation are non-informative. For example, a typical 74-year-old just arriving in the population has the same mortality as a typical 74-year-old who has been in the population since age 70. Since we have little information about the context, it’s hard to say, but in many cases these assumptions might be unreliable.