Applied Survival Analysis
Lab 2: Kaplan-Meier survival estimate

For this lab, we will use the same data from the Ohio State University Bone Marrow Transplant Unit on the time to death or relapse (in months) after bone marrow transplant for 12 patients with non-Hodgkin's lymphoma. However, now the event times that were actually censored are noted by "$^+$".

$$1, 2, 2, 2^+ 3, 5, 6, 7^+, 8, 16^+, 17, 34^+$$

Kaplan—Meier Survival Estimate:

(a) Using the above data, calculate the Kaplan-Meier estimate of the survival function, $\hat{S}(t^+)$ by hand. Summarize your calculations in a table with columns for d_j ,c_j , and r_j.

t^+	d_j	c_j	r_j	$(1-d_j / r_j)$	$\hat{S}(t^+)$

Next we are going to use STATA to estimate the Kaplan-Meier survival estimate as in part (a). So, first we have to input the data. We need one variable `nhltime` for the event time and another `fail` for the event indicator as the last time. But in this case we do have censorings (not everybody relapsed or died) so for the censored values the event indicator will be assigned to zero:

```
input nhltime fail

        nhltime       fail
  1. 1 1
  2. 2 1
  3. 2 1
  4. 2 0
  5. 3 1
  6. 5 1
  7. 6 1
  8. 7 0
  9. 8 1
 10. 16 0
 11. 17 1
 12. 34 0
 13. end
```

Now we have the data and again in order to tell STATA that these are survival-time variables we have to `stset` our data with the following command:

`stset nhltime, failure(fail)` (Option `failure()` specifies failure indicator)

```
     failure event:  fail ~= 0 & fail ~= .
obs. time interval:  (0, nhltime]
 exit on or before:  failure

--------------------------------------------------------------------------
       12  total obs.
        0  exclusions
--------------------------------------------------------------------------
       12  obs. remaining, representing
        8  failures in single record/single failure data
      103  total analysis time at risk, at risk from t =         0
                            earliest observed entry t =         0
                                 last observed exit t =        34
                            earliest observed entry t =         0
                                 last observed exit t =        34
```

Note that this time we have 8 failures and not 12 failures as we had the last time since 4 observations are censored in this dataset.
To get the survivor function you type again the following command:

`sts list`

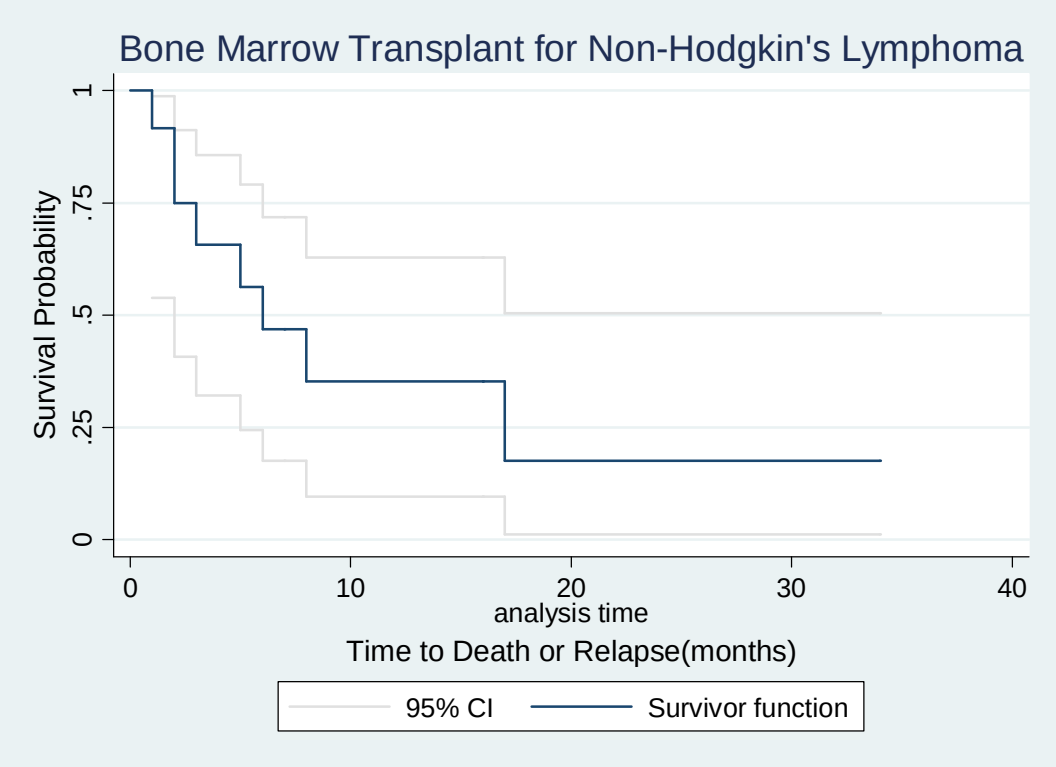
```
         failure _d:  fail
  analysis time _t:  nhltime

          Beg.          Net            Survivor      Std.
  Time    Total   Fail   Lost           Function     Error     [95% Conf. Int.]
-------------------------------------------------------------------------------
     1       12      1      0             0.9167    0.0798     0.5390    0.9878
     2       11      2      1             0.7500    0.1250     0.4084    0.9117
     3        8      1      0             0.6562    0.1402     0.3204    0.8557
     5        7      1      0             0.5625    0.1482     0.2437    0.7910
     6        6      1      0             0.4688    0.1503     0.1762    0.7185
     7        5      0      1             0.4688    0.1503     0.1762    0.7185
     8        4      1      0             0.3516    0.1517     0.0956    0.6278
    16        3      0      1             0.3516    0.1517     0.0956    0.6278
    17        2      1      0             0.1758    0.1456     0.0120    0.5049
    34        1      0      1             0.1758    0.1456     0.0120    0.5049
```

(b) Calculate the estimated standard error of the KM Survival estimate for times $t = 1$ and $t = 3$ using Greenwood's formula. Show how these standard errors are used in calculating the confidence intervals by STATA for $t = 1$.

Next we want to produce the graph of the estimated survival function $\hat{S}(t)$ and pointwise 95% confidence intervals using the "log-log" approach, so we add the option **gwood** after the **sts graph** command:

```
sts graph, gwood title(Bone Marrow Transplant for Non-Hodgkin's
Lymphoma) l1(Survival Probability) b2(Time to Death or Relapse
(months))
```



(c) Identify the estimated median , 25%-ile and 75%-ile survival time. What is the actual estimated survival probability corresponding to the estimated median survival time? (i.e., if the median was 8, then what is $\hat{S}(8)$?).

***Cumulative Hazard Estimate*:**

Now lets move on to the *Cumulative Hazard Estimator*. To get the *Nelson-Aalen* estimator in STATA you still use the **`sts list`** command but you add the option **`na`** :

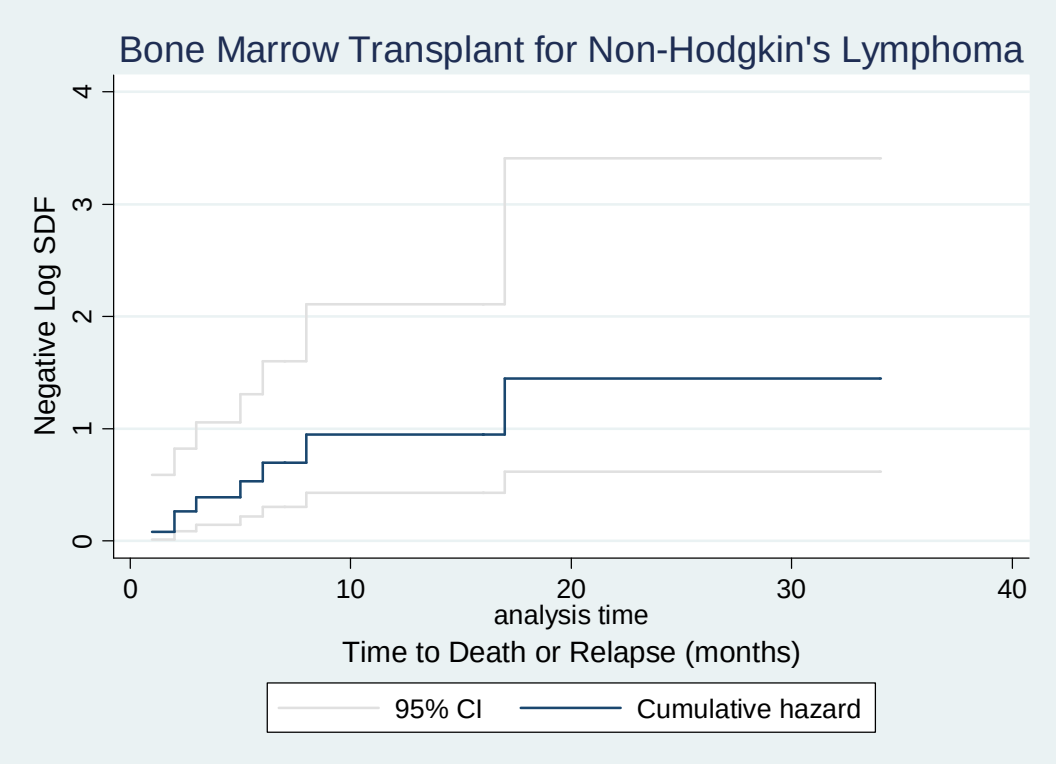
```
sts list, na
```

```
        failure _d:  fail
  analysis time _t:  nhltime
```

Time	Beg. Total	Fail	Net Lost	Nelson-Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
1	12	1	0	0.0833	0.0833	0.0117	0.5916
2	11	2	1	0.2652	0.1532	0.0854	0.8229
3	8	1	0	0.3902	0.1977	0.1445	1.0535
5	7	1	0	0.5330	0.2439	0.2174	1.3071
6	6	1	0	0.6997	0.2954	0.3058	1.6007
7	5	0	1	0.6997	0.2954	0.3058	1.6007
8	4	1	0	0.9497	0.3870	0.4273	2.1109
16	3	0	1	0.9497	0.3870	0.4273	2.1109
17	2	1	0	1.4497	0.6323	0.6166	3.4082
34	1	0	1	1.4497	0.6323	0.6166	3.4082

If we would like to produce a graph of the *Nelson-Aalen Cumulative Hazard* with 95% confidence intervals we add two options after the **`sts graph`** command one for the *Nelson-Aalen Cumulative Hazard* **`na`** and one for the confidence intervals **`cna`** :

```
sts graph, na cna title(Bone Marrow Transplant for Non-Hodgkin's
Lymphoma) l1(Negative Log SDF) b2(Time to Death or Relapse (months))
```



(d) This graph can also be used to check if an exponential model is appropriate for the data. Do you think is an exponential distribution is appropriate or not?