

# Applied Survival Analysis - January 2016

## 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<sup>+</sup>, 3, 5, 6, 7<sup>+</sup>, 8, 16<sup>+</sup>, 17, 34<sup>+</sup>.

- (a) Using the above data, calculate the KM estimate of the survival function,  $S(t^+)$ , by hand. Summarize your calculations in a table with columns for  $d_j$ ,  $c_j$  and  $r_j$ . Import the data set *nhl1* in R and verify your calculations using the function `survfit`. **Optional:** Get the number at risk from the object created by the `survfit` function and write your own R code to obtain the KM estimates.
- (b) Calculate the estimated standard error of the KM Survival estimate for times  $t = 1$  and  $t = 3$  using the Greenwood's formula. Show how these standard errors are used in calculating the confidence intervals by R. Use the `plot` to produce a graph of the estimated survival function along with pointwise 95% CIs using the *log-log* approach.
- (c) Identify the estimated median, 25%-ile and 75%-ile survival time.
- (d) Find the *Nelson-Aalen* estimator of the cumulative hazard function using the `survfit` function. Produce a plot using the `plot` function. One could use an exponential model for analysing these data. Do you think it would be appropriate? Justify your opinion.
- (e) An alternative estimator of survival when the data are grouped is the Lifetable estimator. Import the *nurshome* dataset in R (*nurshome.csv*). Please have a look at the lecture 1 for the description of the dataset. Group the length of stay into 100 day intervals using the `floor` function (see `?floor`). Then construct a lifetable using the `lifetab` function presented in package *KMsurv*. You can download the package from CRAN by typing from the R prompt `install.packages("KMsurv")`.
- (f) **Optional:** Construct an R function to obtain KM estimates. Your function should accept the survival times along with the failure indicator as input, and it should return

a data frame with the distinct survival times and the survival probabilities. Simulate 30 lognormal survival times with  $\text{mean}=-1$  and  $\text{sd}=0.5$  (on the log scale). To allow censoring, simulate random censoring times from the exponential distribution with a constant hazard rate 2, and find the failure indicator. Compare your results with those obtained from `survfit`. Since both functions are computing exactly the same thing, the results should be identical, are they? Use `set.seed` for reproducibility.