APTS - Survival Analysis
Lab Session 1

Ingrid Van Keilegom

August 23, 2017

Before starting, you will have to download the package 'survival' in R :

```
install.packages("survival")
library("survival")
```

1. A study was conducted on the effects of ploidy on the prognosis of patients with cancers of the mouth. Patients were selected who had a paraffin-embedded sample of the cancerous tissue taken at the time of surgery. Follow-up survival data were obtained for each patient. The tissue samples were examined using a flow cytometer to determine if the tumor had an aneuploid (abnormal) or diploid (normal) DNA profile. The data can be found as follows in R :

   ```
   install.packages("KMsurv")
   library("KMsurv")
   data(tongue)
   ```

   Explanation about the data set can be found at

   https://cran.r-project.org/web/packages/KMsurv/KMsurv.pdf

   (a) Plot the survival functions and their standard errors for both the diploid and aneuploid groups (in separate graphs).

   (b) Estimate the median time to death, and find a 95% confidence interval for the median survival time for both the diploid and aneuploid groups. Interpret the results.

   Hint : Use 'survfit' in R (and type 'help(survfit.formula)' for more details).

2. Consider data on the times until staphylococcus infection of burn patients, also available in the 'KMsurv' package as follows :

   ```
   data(burn)
   ```

   (a) Using the log-rank test, test the hypothesis of no difference in the time to staphylococcus infection between patients whose burns were cared for with a routine bathing care method versus those whose body cleansing was initially performed using 4% chlorhexidine gluconate. Use a two-sided test and a 0.05 significance level.

   (b) Solve the same exercise by stratifying on the percentage of surface area burned (0-29%, 30-50%, 51-100%).

   Hint : Use 'survdiff' in R.

If time allows, the following exercises are also very useful :

3. Consider the following dataset:

subject	1	2	3	4	5	6	7	8	9	10
T_i	3^+	5	7	2	18^+	16	2	9	16	5^+,

where $^+$ means that the observation is censored.

(a) Specify:
   (i) the number of distinct event times ?
   (ii) the ordered event times ?
   (iii) the size of the risk set at time $y_{(1)}$?

(b) Calculate by hand (without using R),
   (i) the Kaplan-Meier estimator,
   (ii) the Nelson-Aalen estimator.

(c) Represent graphically the Kaplan-Meier estimator.

(d) For the Kaplan-Meier estimator, compare the results to those obtained via the function `survfit` in R.

4. Consider the following dataset :

$$
\begin{array}{llllll}
\text{Group 1:} & 4.1 & 7.8^+ & 10 & 10 & 12.3^+ & 17.2 \\
\text{Group 2:} & 9.7 & 10 & 11.1^+ & 13.1^+ & 19.7 & 24.1^+
\end{array}
$$

where $^+$ means that the observation is censored.

(a) Calculate by hand the log-rank test in order to compare the survival curves of the two groups.

(b) Is the result the same when using the function `survdiff` in R ?