TMA4275 Life time analysis
Obligatory project 1, Spring 2021

Out: Wednesday February 10
In: Wednesday March 3 at (latest) 18.00

Important information: The project are to be done using R. An introduction to R can be found in the course web page (see Statistical software). The project report should consist of one (and only one) pdf-file, and should be uploaded via Blackboard. The project report should include the R code you have used to solve the project and the plots you have generated. Associated to the various plots there should be captions explaining the content of the plots, and in addition all the plots should be explained and discussed in the main text of the project.

Problem 1:

Breast cancer is one of the most common forms of cancer occurring in women living in the Western world. However, the biological behaviour of the tumour is unpredictable, and there is at present no reliable method for determining whether or not a tumour is likely to have metastasised, or spread, to other organs in the body. In this exercise we consider results from an old investigation to evaluate a histochemical marker that discriminates between primary breast cancer that has metastasised and that which has not. The marker under study is denoted HPA. In order to investigate whether the marker can be used to predict the survival experience of women with breast cancer, a retrospective study was carried out, based on the records of women who had received surgical treatment for breast cancer. Sections of the tumours of these women were treated with HPA and each tumour was subsequently classified as being positively or negatively stained; positive staining corresponding to a tumour with the potential for metastasis. The study was concluded in July 1987, when the survival times of those women who had died of breast cancer were calculated. For those women whose survival status in July 1987 was unknown, the time from surgery to the date on which they were last known to be alive is regarded as a censored survival time. The survival times of women who had died from causes other than breast cancer are also regarded as right-censored. A subset of the data are given below. The survival times of each woman is given in months and classified according to whether their tumour was negatively or positively stained. Censored survival times are labelled with an asterisk (*). There are totally 13 negative stained and 20 positive stained cases in the data. In the analysis one was particularly interested in whether or not there was a difference in the survival experience of the two groups. An evidence that those women with negative HPA staining tended to live longer after surgery than those with a positive staining, would be an indication that the prognosis for a breast cancer patient was dependent on the outcome of the staining procedure. The investigation is documented in the article: Leathem, A.J. and Brooks, S.A. (1987) Predictive value of lectin binding on breast cancer recurrence and survival. The Lancet, I, 1054-1056.

Negative	Positive
23	5
47	8
69	10
70*	13
71*	18
100*	24
101*	26
148	31
181	35
198*	50
208*	59
212*	61
224*	76*
	109*
	116*
	118
	143
	154*
	162*
	225*

In the following we denote patients with negatively stained tumours as group 1 and patients with positively stained tumours as group 2.

a) Make plots of the number of patients at risk for group 1 and for group 2, i.e. $Y_1(t)$ and $Y_2(t)$.

b) Make an R function that takes as input the observed data (and if necessary other relevant information) for one of the two groups and returns a suitable representation of the associated Nelson-Aalen estimator for the integrated hazard rate together with the estimated variance of this estimator.

Using your R function, make a plot of the estimated integrated hazard rate with associated confidence interval for each of the two groups. Include the results for both groups in the same plot so that it is easy to compare the results.

c) Make an R function that takes as input the observed data (and if necessary other relevant information) for one of the two groups and returns a suitable representation of the Kaplan-Meier estimator together with the estimated variance of this estimator.

Using your R function, make a plot of the estimated survival function with associated confidence intervals for each of the two groups. Include again the results for both groups in the same plot, to make it easier to compare the results.

Check your R function and your results by using also the R function "survfit" to produce the Kaplan-Meier estimators for each group, see the description in Chapter 3.1 in More (2016).

Based on the results so far, discuss briefly whether you think it is reason to conclude that the group of patients with negative HPA tends to live longer after surgery than those with a positive

HPA.

d) Estimate the median survival time for each of the two groups and find the associated confidence interval. You can either make your own R code to find this or you can use the R functions discussed in Chapter 3.2 in More (2016).

e) Use the log-rank test to test $H_0 : \alpha_1(t) = \alpha_2(t)$ for $t \in [0, 200]$. Write up the test statistic you are using and write your own R code to evaluate this test statistic. Find also the p-value of the test and discuss briefly your results.

Problem 2:

In this problem you will use stochastic simulation to evaluate the quality of the normal approximation of the log-rank test statistic when testing $\alpha_1(t) = \alpha_2(t), t \in [0, t_0]$. In the last item of this problem you will also estimate the power of the test when H_0 is not true.

Assume we have two groups of individuals with n individuals in each group. Except in the last item of this problem we will assume the hazard rates for the individuals in the two groups to be identical, so that H_0 is true. Assume the life time, T, for each individual is independently Weibull distributed

$$f_T(t; \alpha, \beta) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}, t \geq 0,$$

with $\alpha = 0.01$ and $\beta = 1.1$. In the simulation study we also include right-censoring. For each individual assume we have a censoring time, and that the censoring times, C, are independently exponentially distributed,

$$f_C(c; \lambda) = \lambda e^{-\lambda c}, c \geq 0,$$

with $\lambda = 0.005$. Moreover, we assume our study is terminated at time $t = 225$. Thus, if we let T_i and C_i denote the life time and censoring time, respectively, for individual number i, we observe for this individual the right censored survival time

$$\widetilde{T}_i = \min\{T_i, C_i, 225\}$$

and the censoring indicator

$$D_i = \begin{cases} 1 & \text{if } T_i \leq \min\{C_i, 225\}, \\ 0 & \text{otherwise.} \end{cases}$$

a) Make an R function that simulates $\widetilde{T}_i$ and D_i for n individuals. To simulate Weibull and exponentially distributed random variates you can either use build-in functions in R (but if so you must carefully check the parameterisations used by R for these distributions) or you can use the probability integral transform method (which you know from TMA4300 if you take that course). Your R function should return a representation of $(\widetilde{T}_i, D_i), i = 1, \ldots, n$, which is suitable as input to the R function you made in Problem 1b).

b) For $n = 15$, make (at least) three simulated data sets for the situation described above, i.e. since each of the three data sets should contain data for two groups, you need to call your R

function from Problem 2a) two times for each data set.

For each of your simulated data set, make a plot of the Nelson-Aalen estimators and associated confidence intervals corresponding to what you did in Problem 1b). For each simulated data set evaluate also the test statistic for the log-rank test to test $H_0 : \alpha_1(t) = \alpha_2(t), t \in [0, 200]$ and the associated p-values. Compare the plots and the test statistics you got here with the ones you found in Problems 1b) and 1e).

c) For $n = 15$ make $M = 1000$ simulated data sets for the situation described above and evaluate the test statistic and the p-value for the same log-rank test as in Problem 2b). Use a Q-Q plot for the test statistics to evaluate the normal approximation in the test. Make also a histogram of the resulting p-values. Based on these plots, discuss how good the normal approximation of the test statistic is for this amount of data.

d) If you in Problem 2c) found the test statistic to be well approximated with the standard normal distribution, repeat the simulation exercise with smaller values for n to check how small n can be for the normal approximation to be good.

If you in Problem 2c) found that the test statistic was not well approximated with the standard normal distribution, repeat the simulation exercise with larger values for n to check how large n must be for the normal approximation to be good.

e) In this last item of the problem, we will consider a situation where H_0 is not true, i.e. a situation where $\alpha_1(t)$ is not identical to $\alpha_2(t)$, and use simulation to estimate the power of the log-rank test for such a situation.

For group 1 let the distribution of the life time and censoring times be as before, but for group 2 assume that the life times are exponentially distributed with $\lambda = 0.01$ and the censoring times are distributed as before. Make then a plot with the true hazard rates $\alpha_1(t)$ and $\alpha_2(t)$.

Using simulation, estimate the power of the log-rank test for testing $H_0 : \alpha_1(t) = \alpha_2(t)$ for $t \in [0, 200]$, for this situation for different values of n.

Finally, define at least one other situation where H_0 is not true and estimate the power of the log-rank test for different values of n also in that situation.