TMA4275 Lifetime analysis
Spring 2020
Obligatory Exercise 1

Out: Wednesday February 12
In: Wednesday February 26 at (latest) 15.00.

Important information: *The instructions in this exercise tell you to use MINITAB. You are however free to use R (or another program that you prefer). An introduction to survival analysis in R can be downloaded from the course webpage (see Statistical software). Two txt-files with R-commands for survival analysis may also be downloaded from Lecture plan and progress on the course website. Please note that if you use another program than MINITAB, then the source code should be included in the report. The reports should be uploaded via Blackboard. If two students have been working together, they must still upload individual reports. Write at the start of the report the student number for your collaborator.*

Exercise

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 (*neg.*) or positively (*pos.*) stained. Censored survival times are labeled 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.

Neg.	Pos.
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 order to analyze the data you can download a MINITAB worksheet from the course webpage "Data sets". Look for `Oblig1-dat.mwx`. The first two columns show the data for the negatively stained case, respectively giving the observed times and the corresponding censoring status. Columns 3-4 are in the same way the results for the positively stained data. In columns 5-7, the data are stacked with column 5 giving all the observed times; column 6 giving the corresponding censoring statuses, while column 7 gives the group number (1 for "neg.", 2 for "pos.") These last columns are convenient for use of the "By" option in MINITAB, if simultaneous plots or analyses of the two cases are asked for, see subpoint (d) below.

If you use R, you may download the txt-file `Oblig1-dat.txt` which contains the data as explained for columns 5-7 in the MINITAB worksheet.

Let T_1 be the time to death of a woman with a negatively stained tumour, with survival function $R_1(t) = P(T_1 > t)$, and let T_2 and $R_2(t)$ be the corresponding time and survival function in case of a positively stained tumour.

When you below are asked to do a computation or a plot "by hand" it is meant that you should provide a detailed solution clearly showing the relevant aspects of the task (and not merely copying directly from a MINITAB output).

a) In this point we consider the Kaplan-Meier estimator $\hat{R}_1(t)$ of the survival function $R_1(t)$ for T_1.

Use the data to compute $\hat{R}_1(t)$ "by hand" using the standard formula for the Kaplan-Meier estimator. In addition, compute the estimated standard error of the estimates, again "by hand", using Greenwood's formula.

Then use the "Redistribution of Mass" algorithm presented in the note "Extra on Kaplan-Meier" (see "Lecture plan and progress") to compute the KM-estimator. Check that you get the same result as above.

(Hint: You may simplify the last calculation by noting that there are four censored observations in a row from 70 to 101.)

Then use MINITAB to do the estimation and to draw the graph of $\hat{R}_1(t)$. Also let the MINITAB plot include the corresponding 95% confidence intervals.

b) Why cannot the median the interquartile range (IQR) of T_1 be directly estimated from the Kaplan-Meier plot in subpoint (a)?

Compute "by hand" the estimate for the expected lifetime $E(T_1)$ obtained from the plot. Check that you get the result that is displayed by MINITAB. Looking at the plot, do you think the obtained estimate for $E(T_1)$ is reasonable? Comment.

(*MINITAB-hint: The Kaplan-Meier plot would look more informative if you force the y-axis to be labeled from 0 to 100. You can achieve this by right-clicking in the graph and choosing "Edit Graph" in the menu. Then right click on the vertical axis and choose "Edit Y Scale".*)

c) As mentioned in the beginning of the exercise, one will be particularly interested in finding a possible difference in the survival of the two groups of women. For this we want to compare the survival functions $R_1(t)$ and $R_2(t)$.

Use the "By" option in MINITAB to display the Kaplan-Meier plots for both $R_1(t)$ and $R_2(t)$ in the same figure. Can you give a preliminary conclusion based on this figure?

For a formal check of a possible difference one wants to test the hypotheses $H_0: R_1(t) = R_2(t)$ *for all* t versus $H_1: R_1(t) \neq R_2(t)$ *for at least one* t. Perform the test "by hand" using the logrank test from the lectures. What is the conclusion?

d) In this subpoint we consider the Nelson-Aalen estimators of the cumulative hazard functions $Z_1(t)$ for T_1 and $Z_2(t)$ for T_2.

Compute $\hat{Z}_1(t)$ "by hand" and make a plot. Does the plot give any indications regarding the hazard rate of T_1?

Then use the Minitab macro for Nelson-plot (see "Statistical software" on the course webpage) to compute and graph the Nelson plot also for the positively stained case, resulting in $\hat{Z}_2(t)$.

Do you see a pattern in this plot?

(*MINITAB-hint: In Minitab 19 you click on "View" in the top menu, and then "Command Line/History". Then the command line will appear to the right. In order to have the macro work properly, you first need to open a new MINITAB session where only the two columns for the "pos" data are given. You may download the needed worksheet from the "Data sets" webpage as* `Oblig1-T2.mwx`. *If you get an error message when running the macro, you might try to change the quotes ' to "*)

e) Compute and plot "by hand" the TTT-plot for the data for the negatively stained cases. Also perform "by hand" Barlow-Proschan's test for H_0: T_1 is exponentially distributed versus H_1: T_1 has a monotone hazard rate.

What is the conclusion based on this test?

Then use the MINITAB macro for TTT-plot and Barlow-Proschan's test (which you again find under "Statistical software") to make a TTT plot and to perform Barlow-Proschan's test also for the positively stained case.

What is your conclusion from this?

f) From the Nelson-Aalen plot, the TTT-plot, and the Barlow-Proschan test, it may be argued that the exponential distribution is not the best model for the T_2-data.

Consider instead the Weibull distribution, which is obtained from the exponential distribution by adding an extra ("shape") parameter α. The density is then

$$f_1(t;\theta,\alpha) = \frac{\alpha t^{\alpha-1}}{\theta^\alpha} e^{-(t/\theta)^\alpha}.$$

Find numerical values of the maximum likelihood estimates of the parameters by using MINITAB.

Does the probability plot in MINITAB indicate that the Weibull distribution is a suitable distribution for these data?

Explain briefly how the Weibull probability plot in MINITAB is constructed.

g) Consider again the T_2-data. As an alternative to the Barlow-Proschan test for the null hypothesis of exponential distribution, one may use a test based one (log) likelihoods. Asssuming a Weibull model for T_2 as in the previous subpoint, one may then test the hypotheses

$$H_0 : \alpha = 1 \text{ versus } H_1 : \alpha \neq 1$$

Explain how you can perform this test by using likelihoods displayed by MINITAB. Use significance level 5%. Also find the p-value (this can be done by MINITAB using *Calc* > *Probability Distributions* > *Chi Square* from the top menus).

Which results from likelihood testing are you using?

h) In MINITAB, use *Stat > Reliability/Survival > Distribution Analysis (Right Censoring) > Distribution ID Plot* to fit the four parametric models Exponential, Weibull, Lognormal, Loglogistic to the T_2-data. Which distribution do you think gives the best fit?

Do the same for the "neg." data. What is now the conclusion?

Look at the MINITAB output for each of these cases, where estimates for the expected survival time ("MTTF") is given for each model. Why do you think that these values differ so much from model to model?

i) The four models considered in the previous point are so-called *log-location-scale families* of distributions.

How can these families be described? What is meant by the location and scale parameter, respectively, of these models?

What is the definition of the *log-logistic distribution*? Compute the hazard rate function $z(t)$ for the log-logistic distribution.