Obligatory excercise 1, TMA4275 Lifetime analysis

February 17, 2017

You may use either Minitab or R to solve the exercise (instructions for doing the analysis in R is given below). For MINITAB, use the help system in MINITAB and the slides on the course web page. Write your **candidate** (or student number if you have not yet been assigned a candidate number), not names, on the report. We recommend using latex + knitr (menu option 'File/New file/R Sweave' in Rstudio) for writing the report (see for example, `https://support.rstudio.com/hc/en-us/articles/200552056-Using-Sweave-and-knitr`). You may work on your own or write the report jointly with another student.

We will study the survival times after surgery of a number of gastic cancer patients. Load the data into R with the command

```
gastricXelox <- read.csv("https://www.math.ntnu.no/~jarlet/levetid/gastricXelox.csv")
```

Also load (and if necessary first install) the `survival` pacakge.

1. Using the function `survfit`, compute the Kaplan-Meier estimate of the survival function $R(t)$ and assign the returned object to a variable in R. Plot the Kaplan-Meier estimate.

2. Inspect the object returned by survfit (a list) and use the various components in the list (`n.event` and `n.risk`) to compute the Kaplan-Meier and the Nelson-Aalen estimates of the cumulative hazard function. Also compute a TTT-plot for the data (see slides and example code in `demo.R` on the course web page for details). Discuss how the TTT-plot and N-A and K-M estimaes of the cumulative hazard agree or deviate from what you would if expect if the survival times follow an exponential model.

3. Briefly explain the theory behind the Barlow-Prochan test of exponentiality and carry out the test on the observed data.

4. Use the command `print.survfit` to obtain an estimate of the 'restricted' mean survival time (see the help page of `print.survfit` for details) based on above non-parametric estimate of $R(t)$ as well as the median survival time. Compare these to point estimates and approximate confidence intervals for the same mean and median survival time based on the parametric exponential model computed using your own R-code. You'll need to compute the total time on test at the time of the last right censoring event. Compute the confidence intervals based on approximate normality of $\ln \hat{\theta}$ where θ is the MLE of the scale parameter of the exponential distribution.

   Briefly discuss the differences.

5. Based on the observed total time on test and observed number of failures, make a plot (see the `curve` function) of the log-likelihood function under the exponential survival model. Based on the asymptotic, 1 degree of freedom, chi-square distribution of $2(l(\hat{\theta})-l(\theta))$, explain how this approximate pivotal quantity can be used to construct another approximate confidence interval for θ. Use the plot to find the resulting 95% confidence limits for θ for the observed gastric cancer data or compute the limits numerically using the `uniroot` function.

6. Fit different parameteric models available via the `survreg` function (except the gaussian and logistic) and overlay corresponding parametric estimates of the survival function $R(t)$ to plots showing the non-parametric Kaplan-Meier estimate. Based on plots, briefly discuss which model(s) provide a reasonable fit to the data.

7. Compute a new estimate of ET based on your selected model in point 6 and briefly comment on any difference from the estimates obtained using the exponential and non-parametric models. Note that for distributions such as the log-normal and the log-logistic, ET (if finite) can be easily be found via the moment generating function of $\ln T$ (if this is known), since $ET = Ee^{\ln T} = M_{\ln T}(1)$.