Lifetime Data Analysis with R

Session 1: Nonparametric and Parametric Inference

Exercise 1 (Nonparametric Survival Analysis with the package survival)

Load the Hahn.csv dataset. Elements of presentation of this dataset are given in the file Hahn.txt.

The aim is to draw nonparametric inference using some R packages.

1. Load the package survival. Using the function Surv create a survival object with the above dataset.

2. Using the function survfit, calculate and plot the Kaplan-Meier estimates of the Survival (Reliability) function and the 95% pointwise confidence intervals. Play with the options conf.type = "none" and conf.int=0 in the commands survfit and plot respectively. Compare the confidence intervals you obtain with confidence equal to 95% and 99%.

   On peut utiliser la fonction plot de base ou encore la fonction ggsurvplot disponible avec le package survminer. On peut avoir en plus l'information sur l'évolution du nombre à risque au cours du temps qui permet d'ajuster la qualité de l'estimation non paramétrique.

3. Using the option conf.type, compare the pointwise confidence intervals linear and log-log. The first one has been seen in class, the second is given by

$$[\hat{R}^{1/u(t)}(t), \hat{R}^{u(t)}(t)]$$

   with

$$u(t) = \frac{z_{1-\alpha/2}\sqrt{\hat{\sigma}^2_{GW}}}{\hat{R}(t)\log(\hat{R}(t))}$$

   where $\hat{\sigma}^2_{GW}$ is the Greenwood estimator of the asymptotic variance of the KM estimator.

4. Compare the Kaplan-Meier and the Fleming-Harrington estimates of the Reliability function. This latter is given by

$$\hat{\bar{F}}(t) = e^{-\hat{\Lambda}(t)}$$

   where $\hat{\Lambda}(t)$ is the Nelson-Aalen estimator of the cumulative hazard rate function. Use the option `type='fh'` to obtain the Fleming-Harrington estimator of the Reliability function.

5. Confidence bounds are available with the package km.ci. Then using the function km.ci in order to get the values of the confidence band, plot in a same Figure a confidence band and a pointwise confidence interval for the Reliability function at a level 95%. Comment your results.

6. Compare the estimations obtained when one considers the full censored dataset and when only the failure times are considered.

7. With the function print applied to your kmfit object, find the estimate of the median and the expectation of the lifetime. Understand the comment *"restricted mean with upper limit = 135000"*.

8. Plot the Nelson-Aalen estimates of the cumulative hazard rate function. Compare with the estimate obtained through the Kaplan-Meier estimator of the Reliability.

9. The library muhaz allows to obtain a smooth estimation of the hazard rate. Plot the empirical estimates of the instantaneous hazard rate (used in Nelson-Aalen estimator of the cumulative hazard rate) and the smooth estimates using the function muhaz. What property do you observe on the smooth estimate of the hazard rate function? Can you see this result on the plot of the Nelson-Aalen estimates of the cumulative hazard rate function?

Exercise 2 (Simulations)

Let us now work on simulated data. In this case we know exactly what is the *true* distribution of the simulated dataset. The aim of this exercise is to compare the estimates of the Reliability with the theoretical one when the sample size or the percentage of censoring change.

a Simulate (nested) samples with sizes 50, 100, 200 and 500 under the following hypotheses.

- $X \sim \mathcal{W}(4, 2)$ and $C \sim \mathcal{E}(0.4)$. In this case we have approximatively 70% of censoring.
- $X \sim \mathcal{W}(4, 2)$ and $C \sim \mathcal{E}(0.22)$. In this case we have approximatively 50% of censoring.
- $X \sim \mathcal{W}(4, 2)$ and $C \sim \mathcal{E}(0.105)$. In this case we have approximatively 30% of censoring.
- $X \sim \mathcal{W}(4, 2)$ and $C \sim \mathcal{E}(0.03)$. In this case we have approximatively 10% of censoring.

b For each censoring percentage, plot a single figure the Kaplan-Meier estimates of the Reliability function for the different sample sizes. Add the true Reliability function.

c For each sample size, plot a single figure the Kaplan-Meier estimates of the Reliability function for the different percentage of censoring. Add also the true Reliability function.

Exercise 3 (Hazard Plotting)

Load the datasets bmt and tongue. In the first one, consider t2 as the lifetime and d3 as censoring indicator.

1. With the datasets bmt and tongue consider the following points.

   (a) Plot the Nelson-Aalen estimate of the cumulative hazard rate function. What can you deduce from this plot? Confirm your intuition by the plot of smoothed instantaneous hazard rate function.

   (b) Use the Hazard Plotting method to check adequacy to the following models: exponential, Weibull, Log-normal. One can use the R function `lm`. A regression line can be added using the function `abline` on the result of `lm`.

   (c) First show that a r.v. with log-logistic distribution with parameter μ and σ has a reliability function which can be written like

   $$R(x) = \frac{1}{1 + \alpha x^{\beta}},$$

   where α and β are real parameters. Use the Hazard Plotting method to check adequacy to this distribution on the dataset under consideration.

2. Which model would you choose for each dataset?

Exercise 4 (Test of homogeneity)

1. Load the library MASS and the dataset Melanoma.

2. Consider status 2 and 3 as a form of censoring. Using survfit (option formula), plot separate Kaplan-Meier curves for males and females.

3. Using the function survdiff, test if there is a significant difference between the distributions of the survival times for male and female. Use the log-rank test as well as the Gehan-Wilcoxon test. It is also of interest in this case to use the ggsurvplot function with the option pval=TRUE.

4. Load the library asaur and the dataset prostateSurvival. Test if the survival distributions depend on the covariable stage and then on the covariable ageGroup.