Obligatorisk øving 1 - TMA4275

Emil Myhre - 477789

April 2020

a)

We will calculate the Kaplan-Meier estimator $\hat{R}_1(t)$ for the survival function $R(t)_1$ for T_1, with T_1 being the lifetime of a woman with negatively stained tumour. Let d_i be the number of the deaths, n_i be the number of individuals at risk at point i and n_{ci} the number of censorings in a given time interval. The Kaplan-Meier estimator is defined in the following way

$$\hat{R}(t) = \prod_{i:\mathrm{T}_{(i)} \leq t}^{n} \frac{n_i - d_i}{n_i}, \tag{1}$$

The lifetimes T_1 in our data set are $23, 47, 69, 148, 181$ respectively. Using this formula on our data, we obtain the following estimate for the survival function

$$\hat{R}(t) = \begin{cases} 1 & \text{for } 0 \leq t < 23 \\ 0.92 & \text{for } 23 \leq t < 47 \\ 0.84 & \text{for } 47 \leq t < 69 \\ 0.77 & \text{for } 69 \leq t < 148 \\ 0.64 & \text{for } 148 \leq t < 181 \\ 0.51 & \text{for } 181 \leq t \leq 224. \end{cases} \tag{2}$$

To calculate the estimated standard error we use Greenwoods formula given by

$$\widehat{\mathrm{Var}(\hat{\mathrm{R}}(t))} = \hat{\mathrm{R}}(t)^2 \sum_{T_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{3}$$

yielding the following standard errors for the estimate

$$\hat{\sigma} = \begin{cases} 0, & \text{for } 0 \leq t < 23 \\ 0.071, & \text{for } 23 \leq t < 47 \\ 0.096, & \text{for } 47 \leq t < 69 \\ 0.112, & \text{for } 69 \leq t < 148 \\ 0.142, & \text{for } 148 \leq t < 181 \\ 0.153, & \text{for } 181 \leq t < 224 \end{cases}. \tag{4}$$

This can also be done by the redistribution of mass algorithm, which consists of four steps. First the data is arranged in increasing order. Secondly every observation is given mass $\frac{1}{n}$. Thirdly we move from left to right through the observations, and when we reach a censored observation, its mass is distributed equally over all the remaining observations to the right. Finally we repeat this step until all censored observations have no mass. By using this method, and constructing a table as in the provided paper "Extra on Kaplan-Meier", we get the following estimator

$$\hat{R}(t) = \begin{cases} 1 & \text{for } 0 \leq t < 23 \\ 1 - 1/13 = 0.92 & \text{for } 23 \leq t < 47 \\ 0.92 - 1/13 = 0.84 & \text{for } 47 \leq t < 69 \\ 0.84 - 1/13 = 0.77 & \text{for } 69 \leq t < 148 \\ 0.77 - 0.128 = 0.64 & \text{for } 148 \leq t < 181 \\ 0.64 - 0.128 = 0.51 & \text{for } 181 \leq t \leq 224. \end{cases} \tag{5}$$

Which corresponds nicely with the one we found with the definiton of Kaplan-Meier estimator.

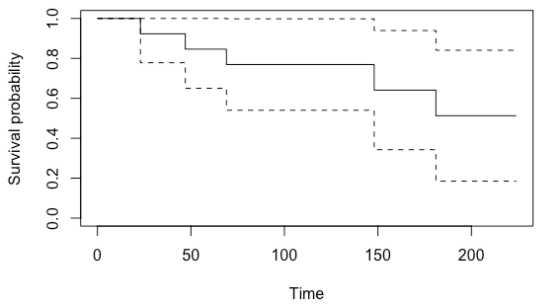

Figure 1: $\hat{R}_1(t)$ plotted with 95% confidence interval.

b)

Only from the Kaplan-Meier plot we dont really have enough information to estimate the IQR accurately. Only 5 of the total of 13 lifetimes are marked. Using the definition of the expected lifetime, we can calculate it in the following way

$$\text{MTTF} = E[T_1] = \int_0^\infty R_1(t)\mathrm{d}t \approx \int_0^{224} R_1(t)\mathrm{d}t. \tag{6}$$

where we have used the approximation because we dont have data for $t \to \infty$, so we substitute ∞ with the time of our last observation. I also assume that due to this, our estimation might be too weak, as we are essentially setting 224 as an upper limit for lifetimes, although in reality it would certainly be a non-zero

probability of observing a lifetime greater than 224. Using the numerical values from the plot and summing up, we get

$$E[T_1] = 167.7.$$

c)

We are interested in testing the hypothesis, whether the survival times R_1 and R_2 are equal or not. That is

$$H_0 : R_1(t) = R_2(t) \ \ \forall t \quad \text{vs} \quad H_1 : R_1(t) \neq R_2(t) \ \ \forall t$$

in order to do this we are using the logrank test described in the notes "The logrank test for comparison of survival funcitons". The idea is to compare expected and observed values of failures with a test statistic. Let $\mathrm{O}_1 = \sum_{j=1}^{k} \mathrm{O}_{1,j}$ and $\mathrm{O}_2 = \sum_{j=1}^{k} \mathrm{O}_{2,j}$ be the observed failures of the two groups and $\mathrm{E}_1 = \sum_{j=1}^{k} \mathrm{E}_{1,j}$ and $\mathrm{E}_2 = \sum_{j=1}^{k} \mathrm{E}_{2,j}$ the expected failures. The test statistic is given by

$$\mathrm{F} = \frac{(\mathrm{O}_1 - \mathrm{E}_1)^2}{\mathrm{E}_1} + \frac{(\mathrm{O}_2 - \mathrm{E}_2)^2}{\mathrm{E}_2} \tag{7}$$

and is approximately chi-square distributed with 1 degree of freedom. Numerical values from our data yields a test statistic of $F = 4.7$. The upper 0.05 quantile of this distribution is $3,84$, and since we observed a greater test statistic, we reject the null hypothesis. From the graph (4) we see that this seems reasonable, as it looks like R_2 is decaying significantly faster than R_1.

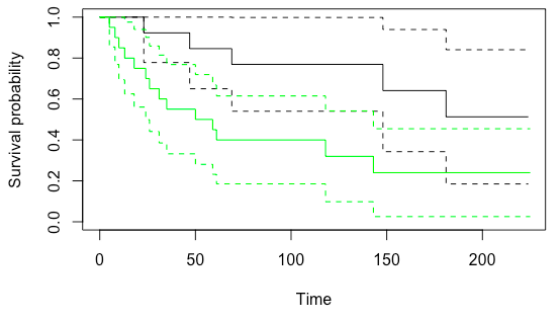

Figure 2: $\hat{R}_1(t)$ (black) and $\hat{R}_2(t)$ (green) in the same plot with confidence intervals.

d)

Now we consider the cumulative hazard functions $Z_1(t)$, $Z_2(t)$ for T_1 and T_2 respectively. We will compute the Nelson-Aalen estimator for Z_1. The estimator

is defined by

$$\hat{Z}_{NA}(t) = \sum_{T_{(i)} \leq t} \frac{d_i}{n_i}. \tag{8}$$

with d_i and n_i still being defined as previously. Numerical data from our data set yields the following estimator for Z_1

$$\hat{Z}_1(t) = \begin{cases} 0, & \text{for } 0 \leq t < 23 \\ 1/13, & \text{for } 23 \leq t < 47 \\ 1/13 + 1/12 = 0.160, & \text{for } 47 \leq t < 69 \\ 0.160 + 1/11 = 0.251, & \text{for } 69 \leq t < 148 \\ 0.251 + 1/6 = 0.418, & \text{for } 148 \leq t < 181 \\ 0.418 + 1/5 = 0.618, & \text{for } 181 \leq t < 224 \end{cases}. \tag{9}$$

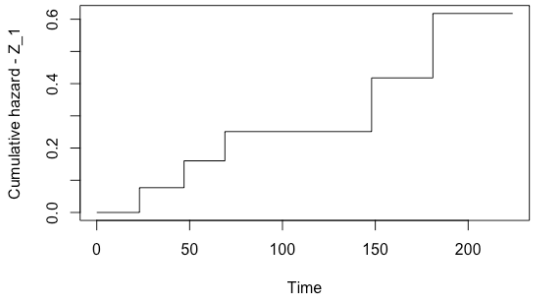

Figure 3: $\hat{Z}_1(t)$ for negatively stained patients.

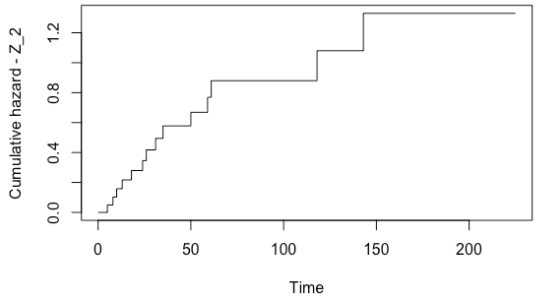

Figure 4: $\hat{Z}_2(t)$ for positively stained patients.

e)

Next we will compute the TTT-plot for the negatively stained cases. Lets denote the Total Time on Test (TTT) at times T_i as Y_i. These can be computed

$T_{(i)}$	Y_i
23	299
47	587
69	829
70*	839
71*	848
100*	1080
101*	1087
148	1369
181	1534
198*	1602
208*	1632
212*	1640
224*	1652

Table 1: TTT values, *positively stained cases

iteratively in the following way

$$\begin{aligned}
Y_1 =& nT_{(1)} \\
Y_2 =& nT_{(1)} + (n-1)(T_{(2)} - T_{(1)}) \\
Y_3 =& nT_{(1)} + (n-1)(T_{(2)} - T_{(1)}) + (n-2)(T_{(3)} - T_{(2)}) \\
& \vdots \\
Y_n =& T_{(1)} + T_{(2)} + \ldots + T_{(n)}
\end{aligned}$$

the results by hand are presented in table (1).

The plot of these points are made by plotting the points

$$\left(\frac{i}{k}, \frac{Y_i}{Y_k}\right) \text{ for all } i : 1, \ldots, k.$$

Now we want to do the Barlow Proschans test on the following hypothesis

$$H_0 : \ T_1 \sim \exp(\lambda) \text{ for some } \lambda$$

versus

$$H_1 : \begin{cases} T_1 \text{ is IFR: Reject } H_0 \text{ if } Z \geq z_\alpha \\ \text{or } T_1 \text{ is DFR: Reject } H_0 \text{ if } Z \leq -z_\alpha \\ \text{or } T_1 \text{ has a monotone hazard rate: Reject } H_0 \text{ if } -z_{\alpha/2} \geq Z \text{ or } Z \geq z_{\alpha/2} \end{cases}$$

We use the following test statistic

$$W = \frac{1}{Y_k}\left(Y_1 + Y_2 + \ldots + Y_{k-1}\right) \tag{10}$$

and use the values from our TTT table, to obtain $W = 2.01$. We compute

$$Z = \frac{W - \frac{k-1}{2}}{\sqrt{\frac{k-1}{12}}} \tag{11}$$

which is $\sim N(0,1)$ under H_0. Inserting our value for W and $k = 5$ we get $Z = 0.017$. With significance level $\alpha = 0.05$, we have $\pm z_\alpha = \pm 1.64$ and $\pm z_{\alpha/2} = \pm 1.96$. Hence we do not reject H_0 in this case and based on this we can assume exponentially distributed lifetimes for negatively stained case.

f)

In this section we want to fit the Weibull distribution as a possible distribution for our data. Its density is

$$f(t;\theta,\alpha) = \frac{\alpha t^{\alpha-1}}{\theta^\alpha} e^{-(t/\theta)^\alpha}, \tag{12}$$

The survival function for a Weibull distributed parameter is

$$R(t) = \exp\left(-(t/\theta)^\alpha\right) \tag{13}$$

By rewriting this with logarithms we can obtain a following linear relation

$$\ln(-\ln R(t)) = \alpha \ln t - \alpha \ln \theta \tag{14}$$

If this was to be done in MINITAB, then by plotting our numerical data and comparing with this line, we can gain information about whether it is reasonable for the data to follow a Weibull distribution.

g)

We want to test the hypothesis

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

In order to do so, we can alternatively use the test statistic

$$W(\alpha) = 2(l(\hat{\theta}, \hat{\alpha}) - l(\hat{\theta}(\alpha), \alpha)) \tag{15}$$

which is approximately chi-square distributed for some estimator $\hat{\alpha}$, and $\tilde{l}(\alpha)$ is the profile log likelihood function of α. We can use MINITAB to calculate $\tilde{l}(\hat{\alpha})$ and $\tilde{l}(1)$ and calculate the test statistic W, then we can use the chi-square distribution of W to determine a p-value in order to reject the null-hypothesis or not.

i)

Log-location-scale families all have something in common. A lifetime T has a log-location-scale family of distributions if $\ln T$ has a location-scale family of distributions. That is, T has a log-location-scale family of distributions if

$$\ln T = \mu + \sigma U.$$

has a location-scale family of distributions. U is a random variable with zero mean and unit variance. Here μ is called the location parameter, and σ the scale parameter.
The four families we have been working with is displayed under, with scale and location parameters

$$\mathrm{U} \sim \mathrm{N}(0,1) \Rightarrow T \sim \text{lognormal}(\mu, \sigma) \tag{16}$$
$$\mathrm{U} \sim \text{logistic}(0,1) \Rightarrow T \sim \text{log-logistic}(\mu, \sigma) \tag{17}$$
$$\mathrm{U} \sim \text{Gumbel}(0,1) \rightarrow T \sim \text{Weibull}(\theta, \alpha), \text{ with } \mu = \ln\theta, \sigma = 1/\alpha \tag{18}$$
$$\mathrm{U} \sim \mathrm{N}(0,1) \Rightarrow T \sim \text{exponential}(\lambda), \text{ with } \mu = -\ln\lambda, \sigma = 1. \tag{19}$$

Recall than the hazard rate is defined by

$$z(t) = \frac{f(t)}{R(t)}$$

T is defined to be log-logistic with parameters μ and σ if

$$\mathrm{R}(t) = \frac{1}{1 + (t/e^{\mu})^{1/\sigma)}}.$$

Then, since $\ln(T)$ has a standardized logistic distribution, it can be shown that

$$\mathrm{R}(t) = 1 - \mathrm{H}\Big(\frac{\ln t - \mu}{\sigma}\Big),$$

where H is the cdf of the logistic distribution. Let h be the pdf of the logistic distribution, then it follows that

$$\mathrm{f}(t) = h\Big(\frac{\ln t - \mu}{\sigma}\Big) \cdot \frac{1}{\sigma t}$$

Substituting $u = \Big(\frac{\ln t - \mu}{\sigma}\Big)$, using the expression for H and h, and using the definition of hazard rate we obtain

$$\mathrm{z}(t) = 1 + e^{u}.$$