

# Computer Assignment 1

*Jan Alexandersson, Anton Strähle & Max Sjödin*

*September 19, 2020*

- Argue for the performance of the estimators relative to the true theoretical model used, is any of the two preferable?
- What is the effect of censoring?
- What is the effect of  $n$ ?

In this computer assignment we aim to estimate the cumulative hazard rate with the Nelson-Aalen estimator and the Kaplan-Meier estimator. However, the Kaplan-Meier estimator estimate the survival function and not the cumulative hazard rate, but this can be obtained by taking the negative logarithm of the Kaplan-Meier estimator. We chose to present both estimators as cumulative hazard rates so that they can easily be compared.

We begin by generating  $n$  Weibull distributed random numbers from Weibull distribution given by the density function

$$f(t; a, b) = \frac{a}{b} \left(\frac{t}{b}\right)^{a-1} \exp\left\{-\left(\frac{t}{b}\right)^a\right\}, t, a, b \geq 0,$$

where  $a = 4.5$  and  $b = 22.5$ . This is done for  $n = 10, 100, 200, 500$  and  $1000$ .

The Nelson-Aalen estimator is given by

$$\hat{A}(t) = \sum_{T_j \leq t} \frac{d_i}{Y(T_j)},$$

where  $Y(T_j)$  is the number of individuals at risk at time  $T_j$  and  $d_i$  is the number of events if there is a tie and 1 if not.

The Kaplan-Meier estimator is given by

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_i}{Y(T_j)}\right).$$

As written earlier, we get the Kaplan-Meier estimate for the cumulative hazard rate by  $-\log(\hat{S}(t))$ . In figures 1-5a) we can see the Nelson-Aalen estimate and the Kaplan-Meier estimate compared with the true cumulative hazard rate, in the case of no ties and no censoring. We can also see 95% confidence intervals for the estimates, which are log-transformed Wald intervals. We can see that when increasing  $n$  our estimates are closer to the theoretical cumulative hazard rate and our confidence intervals less wide. We also see that there is not much difference between using the Nelson-Aalen estimate compared to Kaplan-Meier estimate.

We continued by also computing the Nelson-Aalen and Kaplan-Meier estimates when we only observe total number of events at the end of each interval of length 0.1, therefore creating “ties”. We regard these ties as “true ties”. We see in figures 1-5b) that the Kaplan-Meier estimate is slightly more accurate compared to the theoretical and is therefore preferred, however the difference to the Nelson-Aalen estimate is not large.

Overall the effect of  $n$  is that the Nelson-Aalen and Kaplan-Meier estimates are underestimating the cumulative hazard rate for smaller  $n$  and also deviates more from the theoretical cumulative hazard rate.

We then add censoring by generating  $n$  exponentially distributed random numbers with mean 80. We denote these by  $C_i$ . If  $T_i > C_i$  the observation is censored. We can see in figures 1-5c) that when the estimates deviate from the theoretical distribution for when  $t$  is large ( $t > 25$ ).

ADD MORE HERE ABOUT CENSORING

## Appendix

**Figure 1 : For 10 observations**

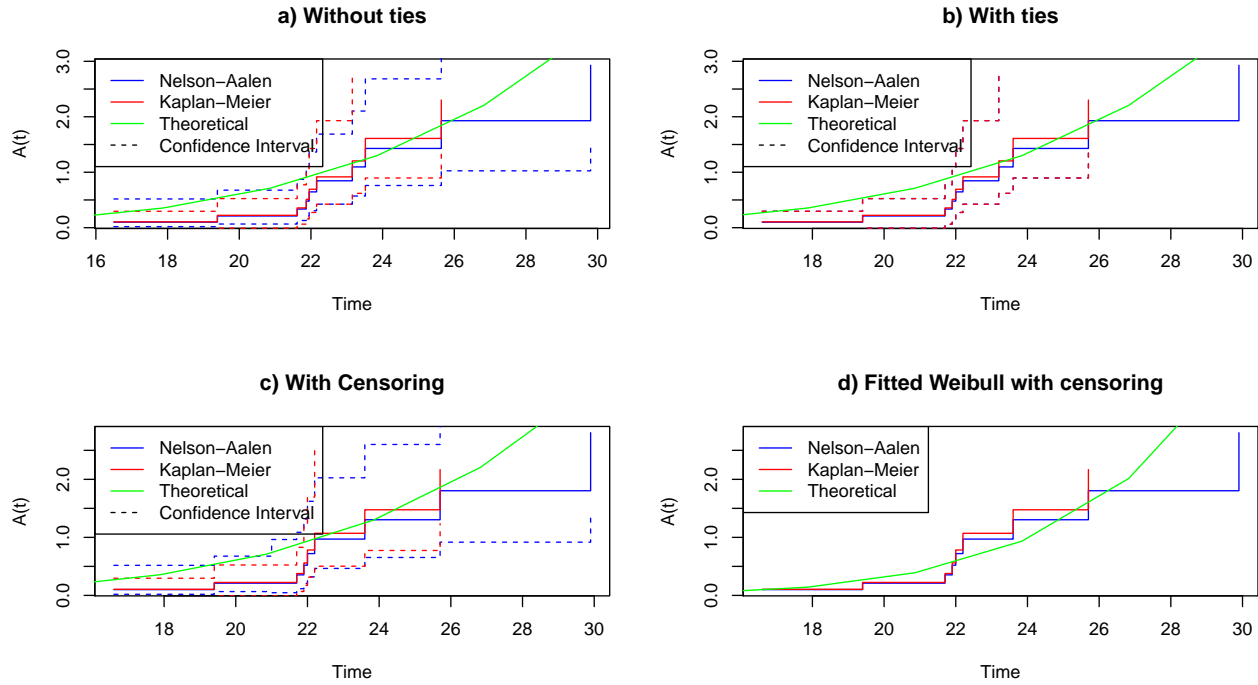


Figure 2 : For 100 observations

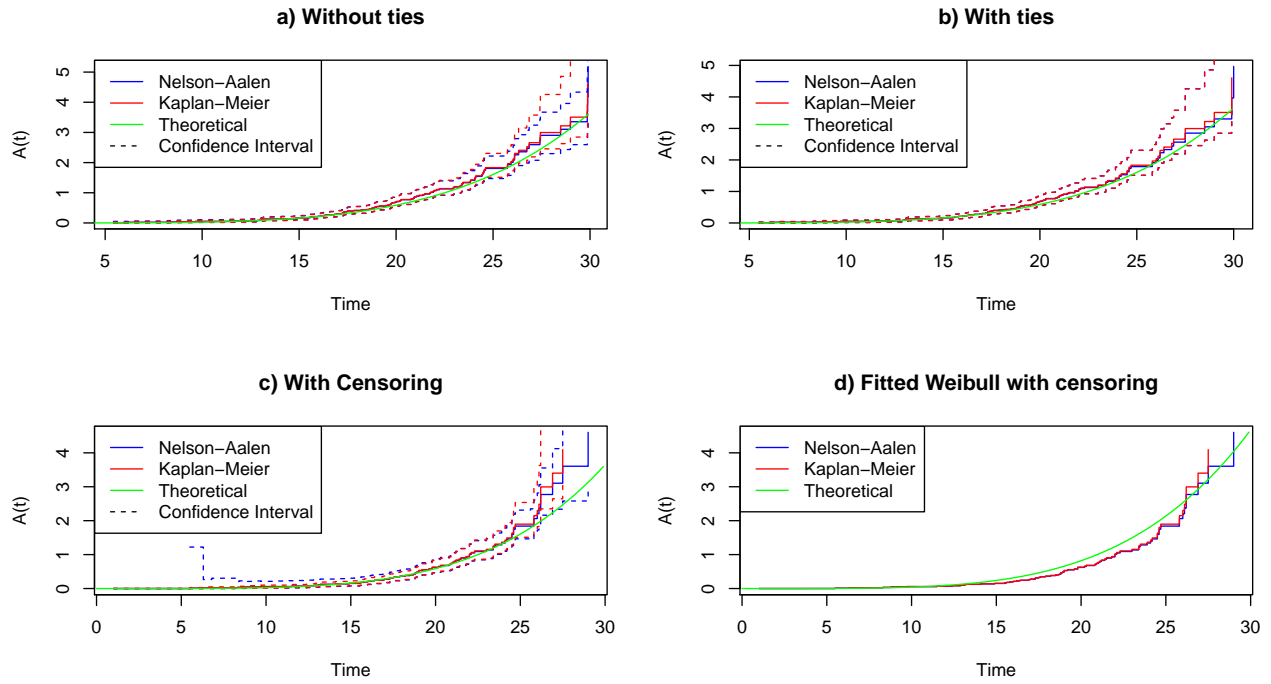


Figure 3 : For 200 observations

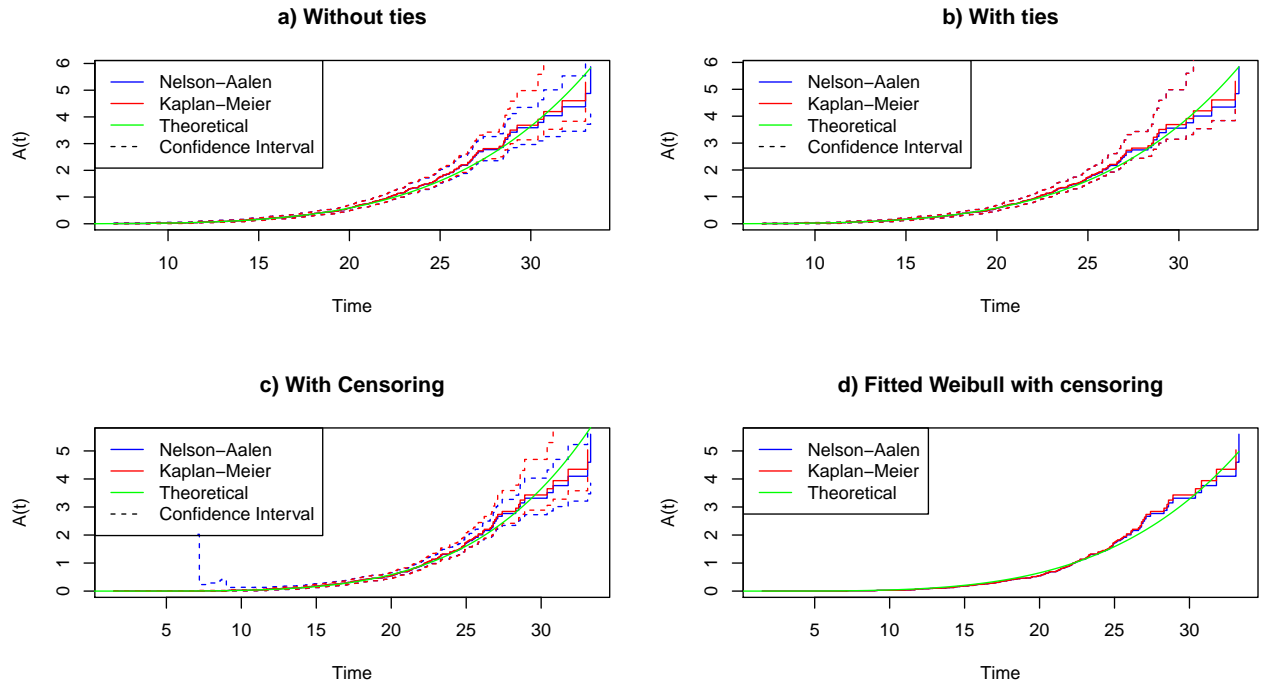


Figure 4 : For 500 observations

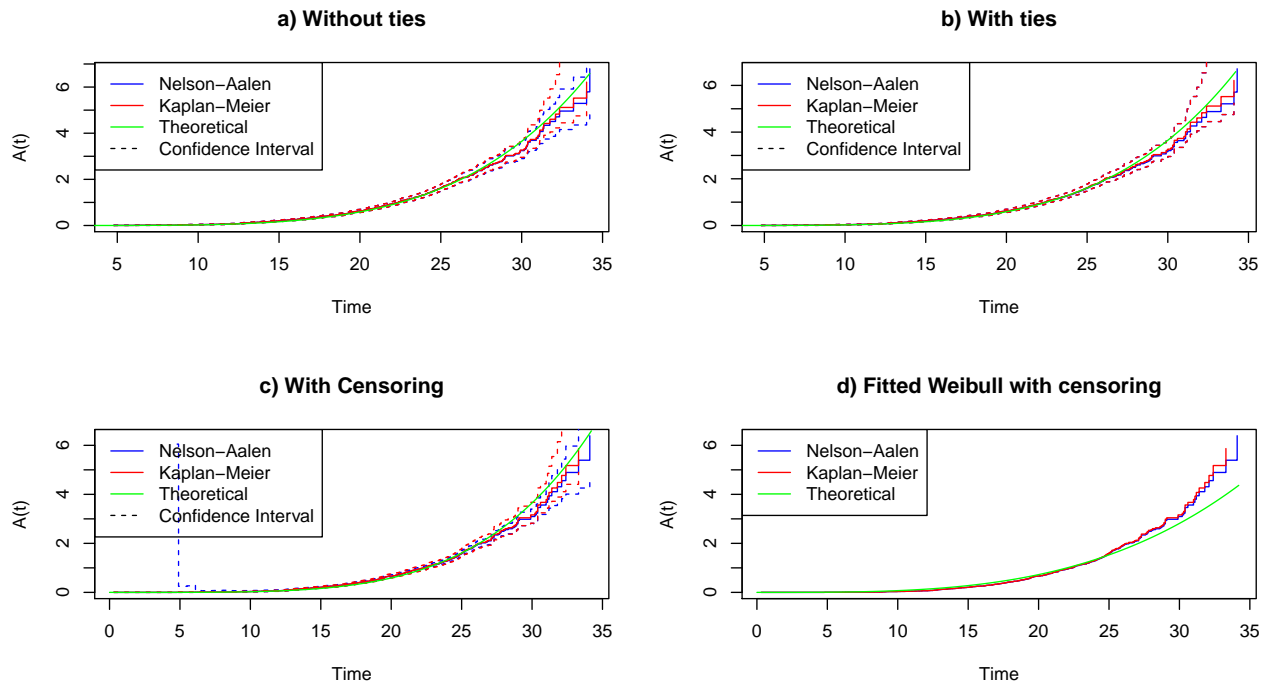


Figure 5 : For 1000 observations

