# TPK4120 - Lecture summary

Jørn Vatn
eMail: jorn.vatn@ntnu.no

Updated 2021-11-17

## Chapter 14 - Reliability Data Analysis

Methods and models introduced in previous chapters require numbers for the reliability parameters such as the failure rates, repair rates, ageing parameters and so on. A parameter in this context is a quantity in a reliability model for which we assign numerical values. To obtain such numeric values several principle for parameter *estimation* exist.

In this presentation only the principle of maximum likelihood estimation will be addressed.

### The MLE principle

The basic idea behind the Maximum Likelihood Estimation (MLE) principle is to choose the numerical values of the parameters that are the most likely ones in light of the data. The procedure goes as follows:

- Assume that we know the probability density function of the observations for which we have data. Let this distribution be denoted $f(t;\theta)$.
- The involved parameters, are unknown, and are generally denoted $\theta$.
- We have $n$ independent observations (data points) that we denote $T_1$, $T_2, \ldots, T_n$. When we refer to the actual numerical values observed, we use the (lowercase) notation $t_1$, $t_2, \ldots, t_n$.

The MLE principle now tells us to estimate $\theta$ by the value which is most likely given the observed data. To define "likelihood" we use the probability density function. The simultaneous probability density for $T_1$, $T_2, \ldots, T_n$ is given by:

$$f(t_1;\theta)f(t_2;\theta)\ldots f(t_n;\theta) = \prod_{i=1}^{n} f(t_i;\theta) \tag{1}$$

This density express how likely a given combination of the $t$-values are, given the value of $\theta$. However, in our situation the t-values are given, whereas θ is unknown. We therefore interchange the arguments, and consider the expression as a function of θ:

$$L(\theta;t_1,t_1\ldots,t_n)=\prod_{i=1}^{n} f(t_i;\theta) \tag{2}$$

where $L(\theta;t_1,t_1\ldots t_n)$ in equation (2) denotes the likelihood function. The MLE principle will now be formulated as to choose the θ-value that maximizes the likelihood function. To denote the MLE *estimator* we write a "hat" over θ, $\hat{\theta}$. Generally, θ will be a function of the observations:

$$\hat{\theta}=\hat{\theta}(T_1,T_2,\ldots,T_n) \tag{3}$$

When we insert numerical values for the $t$-values we denote the result as the parameter *estimate*.

## Estimation in the exponential distribution

We consider the situation where we have observed $n$ failure times, and we will estimate the failure rate, $\lambda$, under the assumption of exponentially distributed failure times.The observed failure times are denoted $t_1$, $t_2$, ..., $t_n$. Equation (2) gives:

$$L(\lambda;t_1,t_2,\ldots,t_n)=\prod_{i=1}^{n}\lambda e^{-\lambda t_i}$$

Note that the parameter is denoted $\lambda$, whereas we generally use $\theta$**.** Further we denote the observations with $t$ because we here have failure *times*. The probability density function in the exponential distribution is given by $f(t)=\lambda \mathrm{e}^{-\lambda t}$. A common "trick" when maximising the likelihood function is to take the logarithm. Because the logarithm (ln) function is monotonically increasing, $\ln L$ will also be maximised for the same value as for which $L$ is maximised. We could then find:

$$l(\lambda;t_1,t_2,\ldots,t_n)=\ln L(\lambda;t_1,t_2,\ldots,t_n)=n\ln\lambda-\sum_{i=1}^{n}\lambda t_i$$

By taking the derivative wrt $\lambda$ and equate to zero, we easily obtain:

$$\hat{\lambda}=n/\sum_{i=1}^{n}t_i$$

## How to obtain the data?

In some situations we are able to conduct experiments to get access to reliability data. We can imagine that we put $n$ identical lightbulbs in $n$ sockets and observe the failure times. A challenge might be that we do not have time to wait for all light bulbs to fail. This means that we will have some "real" life times and some "censored" life times. The censored life times are then the period they have survived. The fact that some light bulbs might have survived the time period of our experiment is also an information we will utilize. In the text book different types of censoring is discussed.

In most cases we do not have access to data in such a controlled manner. But very often we will have access to data from computerized maintenance management systems (CMMS) in terms of failure reports and reports from preventive maintenance.

From the CMMS it is to some extent possible to extract life time data. Several challenges are encountered in such an attempt to get life time data to use in our parameter estimation:

- Data is not reported on the appropriate level, for example we are seeking the failure rate of a pump bearing, but failures are only reported on the pump level
- There are several failure modes reported for an item, and we do not have any information regarding if the item is "as good as new" with respect to all failure modes after a corrective repair action
- Preventive maintenance is carried out, and hence we have very few "real" life times
- We have data for several items, but they are not operated under "identical" conditions, hence merging the data to get a sufficient number of data points is not easy

### Failures vs censoring life times

In experiments as well as in real life there are situations where we are not able to observe the time of failure of an item. The reasons for this could be that the experiment is terminated before all items have failed, or for a real life item, the item is replaced preventively before a failure occurs. In this situations we will usually know that the item has survived a certain time period. The point of time representing this survival period is denoted a *censoring* life time. It is obvious that a censoring life time has less informative value than a real life time in order to assess the underlying reliability parameters. However, the censoring life time represent some information we will not discard in the parameter estimation. We often put a star (*) on the censoring life times to distinguish them from the real life times. In the following we also use an indicator variable to distinguish censoring and real life times, where the value 1 means a real life time and the value 0 means a censoring life time.

### Estimation when life times are Weibull distributed

Now assume that we have been able to extract life time data from either controlled experiments or from our CMMS.

Let $t_1$, $t_2$, $\dots t_n$ denote the observed life times including censored life times. Further let $I_1$, $I_2$, $\dots I_n$ be indicator variables equal to one if the corresponding life time is a real life time, and equal to zero if it is a censored life time.

The censored life times are assumed to be "right censored" life times in the meaning that we know the "birth" of the item, but not the "death". The only thing we know is thus the fact that the item has survived the censored life time. To get "something" to put into the likelihood function, we then use the survivor function, $R(t)$. $R(t)$ is the likelihood that an item survives $t$, and this is what we need, i.e., what is "the likelihood of observing what we observed"?

Recall that the pdf of the Weibull distribution is given by $f(t;\alpha,\lambda) = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha}$ and the survivor function is given by $R(t;\alpha,\lambda) = e^{-(\lambda t)^\alpha}$. Thus the likelihood function is given by:

$$L(\alpha,\lambda;t_1,t_2,\dots,I_1,I_2,\dots) = \prod_i \left(I_i\alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha} + (1-I_i)e^{-(\lambda t)^\alpha}\right) \tag{4}$$

taking logarithm we obtain:

$$\begin{aligned} l(\alpha,\lambda;t_1,t_2,\dots,I_1,I_2,\dots) &= \ln L(\alpha,\lambda;t_1,t_2,\dots,I_1,I_2,\dots) \\ &= \sum_{i=1}^{n} I_i\left[\ln\alpha + \alpha\ln\lambda + (\alpha-1)\ln t_i\right] - \sum_{i=1}^{n}(\lambda t_i)^\alpha \end{aligned} \tag{5}$$

Numerical methods are required for maximizing equation (5)

**Example**

Assume we have observed the following life times: 8,9,7,6,12,18,14,18*,6,9,11, 24,30* and 28*. Here a star (*) indicates that the life time is a censored life time. The MLE estimates are given by:

$$\hat{\alpha} \approx 1.61$$
$$\hat{\lambda} \approx 0.0555$$

obtained by the "Solver" in Excel.

## Graphical techniques

Several graphical techniques may be used to analyse life time data. In the textbook the total time on test (TTT) plot, the Nelson plot and the Kaplan-Meier plot are discussed. In the following we will demonstrate the use of the Kaplan-Meier estimator. This estimator is estimating the survivor function and can handle both real life times and censoring life times.

It may be shown that we always may sort our life time data since the ordering of collecting data will in any case be arbitrary given that data are independent and identically distributed. Let the sorted data be denoted $t_{(1)}$, $t_{(2)}, \dots, t_{(n)}$ where also censored life times are included. Further let $n_i$ be the number of items "under risk" at time $t_{(i)}$, i.e., the number of items not failed just prior to $t_{(i)}$ . Now at time $t_{(i)}$ there might be no failure if this was a censoring time, it might be one failure, or it might even be more than one failure if two failures occurred at the same time. Theoretically it is not possible to have two failures exactly at the same time, but due to limitation in "number of digits" to represent the failure times, we may have more than one failure at the same time. Let $s_i$ be the number of life times observed at time $t_{(i)}$.

To obtain the Kaplan-Meier estimator we use more or less the same type of arguments as given in the text book.

First consider a small time interval around time $t_{(i)}$. In the beginning of this interval it will be $n_i$ items at risk. Let $p_i$ be the probability that one arbitrary of these items will survive this small interval. A natural estimator for $p_i$ is given by

$$\hat{p}_i = \frac{n_{(i)} - s_{(i)}}{n_{(i)}} \tag{6}$$

since $n_{(i)} - s_{(i)}$ of the items we had survived this interval. Now, assume that we have an estimate, $\hat{R}_i^-$ for the probability that an item has survived *up to* the interval we are considering, then it follows that an estimate for the probability that an item will survive from $t = 0$ *to the end* of the interval is given by

$$\hat{R}_i^+ = \hat{R}_i^- \hat{p}_i \tag{7}$$

Following such arguments we obtain the Kaplan-Meier estimator for the survivor function at time $t$:

$$\hat{R}(t) = \prod_{t_{(i)} < t} \frac{n_{(i)} - s_{(i)}}{n_{(i)}} \tag{8}$$

**Example**

Assume we have observed the following life times: 8,9,7,6,12,18,14,18*,6,9,11, 24,30* and 28*. Here a star (*) indicates that the life time is a censored life time. The tableau for the Kaplan-Meier plot now reads:

The following link shows the Excel file: `http://folk.ntnu.no/jvatn/eLearning/TPK4120/Excel/MLE_Kaplan_Meier.xlsx`.

Table 1: Kaplan Meier plot

| $t_i$ | $I_i$ | $n_i$ | $s_i$ | $(n_i - s_i)/n_i$ | $R(t_i)$ |
|---|---|---|---|---|---|
| 6 | 1 | 14 | 2 | 12/14=0.86 | 0.86 |
| 7 | 1 | 12 | 1 | 11/12=0.92 | 0.79 |
| 8 | 1 | 11 | 1 | 10/11=0.91 | 0.71 |
| 9 | 1 | 10 | 2 | 8/10=0.8 | 0.57 |
| 11 | 1 | 8 | 1 | 7/8=0.88 | 0.5 |
| 12 | 1 | 7 | 1 | 6/7=0.86 | 0.43 |
| 14 | 1 | 6 | 1 | 5/6=0.83 | 0.36 |
| 18 | 1 | 5 | 1 | 4/5=0.8 | 0.29 |
| 18 | 0 | 5 | 0 | 5/5=1 | 0.29 |
| 24 | 1 | 3 | 1 | 2/3=0.67 | 0.19 |
| 28 | 0 | 3 | 0 | 3/3=1 | 0.19 |
| 30 | 0 | 3 | 0 | 3/3=1 | 0.19 |