Article citation info:
FERREIRA LA, SILVA JL. Parameter estimation for Weibull distribution with right censored data using EM algorithm. Eksploatacja i Niezawodnosc – Maintenance and Reliability 2017; 19 (2): 310–315, http://dx.doi.org/10.17531/ein.2017.2.20.

Luís Andrade FERREIRA
José Luís SILVA

PARAMETER ESTIMATION FOR WEIBULL DISTRIBUTION WITH RIGHT CENSORED DATA USING EM ALGORITHM

ZASTOSOWANIE ALGORYTMU MAKSYMALIZACJI WARTOŚCI OCZEKIWANEJ DO ESTYMACJI PARAMETRÓW ROZKŁADU WEIBULLA W PRZYPADKU DANYCH OBCIĘTYCH PRAWOSTRONNIE

The maximum-likelihood estimation (MLE) is a method of estimating the parameters of a statistical model for given data. This method allows us to estimate the unknown parameters of a statistical model. These parameters are obtained by maximizing the likelihood function of the model in question. In many practical situations the likelihood function is associated with complex models and the likelihood equation has no explicit analytical solution, it is only possible to have its resolution through numerical methods. The estimation of the parameters of the Weibull distribution by maximum-likelihood method based on information from a historical record with right censored data shows this difficulty. The solution presented in this article entails using the Expectation-Maximization (EM) algorithm. Actual data from the historical record of 5 centrifugal pumps failures of a petrochemical company were analyzed for application of the methodology.

***Keywords**: EM algorithm, parameter estimation, maximum likelihood estimates, reliability.*

Metoda największej wiarygodności (MLE) służy do estymacji parametrów modelu statystycznego dla zadanych danych. Metoda ta pozwala na estymację nieznanych parametrów modelu statystycznego. Parametry te otrzymuje się poprzez maksymalizację funkcji wiarygodności rozważanego modelu. Często w praktyce metoda ta może jednak nastręczać trudności związane z wielomodalnością funkcji wiarygodności oraz niemożnością uzyskania jawnych analitycznych rozwiązań równań wiarygodności. Równania takie można jedynie rozwiązywać za pomocą metod numerycznych. Trudności te dobrze ilustruje estymacja parametrów rozkładu Weibulla z wykorzystaniem metody największej wiarygodności wykonywana w oparciu o prawostronnie cenzurowane dane z eksploatacji. Rozwiązanie przedstawione w niniejszej pracy opiera się na zastosowaniu algorytmu maksymalizacji wartości oczekiwanej (EM). Możliwości aplikacyjne proponowanej metodyki badano na przykładzie danych eksploatacyjnych uzyskanych z przedsiębiorstwa petrochemicznego, dotyczących awarii pięciu pomp odśrodkowych.

***Słowa kluczowe**: algorytm EM, estymacja parametrów, estymator największej wiarygodności, niezawodność.*

1. Introduction

The estimation process is supported by a number of statistical techniques, methods and procedures for analyzing the data on the variable of interest that may be the time that elapses from the well-defined initial instant, for example, installation of equipment, until the occurrence of a specific event, such as the failure of the equipment or component under consideration.

The process aims to estimate the distribution parameters for modeling the system under review. The parameters are the distribution characteristics that show the behavior of a given population and therefore fixed for a specific system. Thus, the estimation of system parameters is obtained from the data collected from the population.

System data analysis can be obtained from various possible sources, namely, laboratory tests or a recording of occurrences along its use (historical record).

The parametric analysis assumes that the data fits a specific distribution, such as the Weibull distribution.

The Weibull distribution has a wide application in various fields. These applications include its use to model the distribution phenomena of fatigue and the life of many devices, such as bearings, shaft and motor [1, 14, 20].

The popularity of the distribution of Weibull due to its great flexibility, i.e., as it can describe functions with failure intensity constant, increasing and decreasing, for different values of the shape parameter.

Since the distribution of Weibull became widely recognized, various methods have been proposed to estimate its parameters [2,17,19].

However, the maximum-likelihood estimation method is currently one of the most used methods of estimation, for its versatility and provides reliable results [10].

The maximum-likelihood estimation method allows us to estimate the unknown parameters of a statistical model. These parameters are obtained by maximizing the likelihood function of the model in question [4].

2. Types of data

With the information from a historical record of data is possible to obtain indicators to estimate and understand the behavior of the equipment with respect to failures. Therefore, using appropriate methodologies, it will be possible to set the proper maintenance policies to any equipment and their components.

The data is considered complete when it is known the exact time of each system failure. In many cases the data contain uncertainties, i.e., it is not known the exact moment when the failure occurred. The data containing such uncertainty as to when the event occurred are regarded as incomplete or partial. Incomplete data can be classified into censored or truncated [8, 13].

Incomplete data give only part of the information about the failure time of the units under review. However, this information should not be ignored or treated as failure. In the absence of such data, it would not be possible to make good estimation parameters and thus make a proper analysis.

One of the most common types of censored data, which may arise in real cases, is Type-1 right censored data [9]. For Type-1 right censored data, all units of a system are observed up to the date of completion of the study. For this censorship scheme the time each unit is under observation is fixed, while the number of units that fail (uncensored observations) is random.

If *T* is a random variable representing the failure time and *Cd* another random variable independent of *T* which corresponds to the end of the registration information (observation time). It is said that the time to failure is right censored when one does not know its exact value, only that its value is greater than *Cd*, with regard to item i (i = 1, 2, ..., n). Therefore:

$$t_i = \min\left(T_i, Cd_i\right) and\ \delta_i = \begin{cases} 1\ if\ T_i \le Cd_i \\ 0\ if\ T_i > Cd_i \end{cases} \tag{1}$$

The δ_i variable (censorship indicator) indicates whether T_i is censored or not. The obtained data can be represented by the pair (t_i, δ_i) i.e. t_i the failure time or censored time and δ_i the variable that indicates whether it concerns a failure or censorship, that is, (2):

$$\delta_i = \begin{cases} 1,\ for\ uncensored\ data \\ 0,\ for\ censored\ data \end{cases} \tag{2}$$

In the right censored data the failure time of the units with censored data it is just known to be greater than the operating time of the conclusion of the registration information. These right censored data are further classified into Type-1 if the recording of information is interrupted at a predetermined time and Type-2 censure if registration is completed when a predetermined number of failures occur [12].

3. The maximum-likelihood estimation method

As we have seen, in the particular case of the right censored data classified as Type-1 censorship, the recording of information is interrupted at a predetermined time $Cd > 0$, such that t_i is observed to occur before *Cd*, otherwise, one only knows that the failure time is greater than the observation time.

Let $t_i = t_1, t_2, ..., t_n$, *n* independent observations, where *r* records are failure times and (*n* - *r*) are censured information.

Let the probability density function, $f(x, \theta)$ and the cumulative probability function $F(x, \theta)$, with the distribution parameters denoted by θ, the likelihood function for right censored data Type-1 is given by [18]:

$$L\left(\theta_1, \theta_2, \ldots, \theta_k\right) = \prod_{\delta_i = 1} f\left(x_i | \theta_1, \theta_2, \ldots, \theta_k\right) \prod_{\delta_i = 0} \left[1 - F\left(x_i | \theta_1, \theta_2, \ldots, \theta_k\right)\right] \tag{3}$$

Where ($x_1, x_2, ..., x_n$) is a sample of *n* independent observations of the random variable *X* and $\theta = (\theta_1, \theta_2, ..., \theta_k)$ is the vector of unknown distribution parameters.

For the Weibull distribution with scale parameter *η* and shape parameter *β*, the likelihood function for the right censored data Type-1 is given by:

$$L(\eta, \beta) = \prod_{i=1}^{n} \left\{ \frac{\beta}{\eta} \left(\frac{t_i}{\eta}\right)^{\beta - 1} exp\left(-\left(\frac{t_i}{\eta}\right)^{\beta} \right) \right\}^{\delta_i} \left\{ exp\left(-\frac{t_i}{\eta} \right)^{\beta} \right\}^{1 - \delta_i} \tag{4}$$

In many situations it is easier to achieve the maximization of the log of the likelihood function and, since the logarithm function is a monotonically increasing function, is equivalent to maximizing the likelihood function or the log-likelihood function.

By applying logarithm to eq. 4, it becomes:

$$\ln L(\eta, \beta) = l(\eta, \beta) = \sum_{i=1}^{n} \left(n\delta_i \ln\beta - n\beta\delta_i ln\eta\right) + (\beta - 1) \sum_{i=1}^{n} \left(\delta_i \ln t_i\right) - \sum_{i=1}^{n} \left(\frac{t_i}{\eta}\right)^{\beta} \tag{5}$$

As the maximum-likelihood equations in many cases have no analytical solution to determine their solutions, it is needed to use numerical optimization methods. Among the possible numerical solutions, in this paper the Expectation-Maximization (EM) algorithm was selected [3, 5, 15].

3.1. The Expectation-Maximization algorithm

The EM algorithm is an iterative process that can be used to calculate the maximum likelihood estimators in cases with incomplete data.

Let *X* be the set of complete data with the probability density function f_c (x, θ) and *Y* the observed data. The corresponding log-likelihood function to the full sample is represented by:

$$\ln L_c(x, \theta) = l_c(x, \theta) \tag{6}$$

Each iteration of the EM algorithm involves two steps: step E (expectation) and step M (maximization), defined by [15],

Step E: To calculate $Q\left(\theta, \theta^{(k)}\right)$, where:

$$Q\left(\theta, \theta^{(k)}\right) = E_{\theta^{(k)}}\left[l_c(x, \theta) \middle| y, \delta, \theta^{(k-1)} \right] \tag{7}$$

Step M: To find $\theta^{(k+1)}$ that maximizes $Q\left(\theta, \theta^{(k)}\right)$, that is:

$$\theta^{(k+1)} = \arg\max Q\left(\theta, \theta^{(k)}\right)$$
$$Q\left(\theta^{(k+1)}, \theta^{(k)}\right) \ge Q\left(\theta, \theta^{(k)}\right) \tag{8}$$

The procedure is performed until the difference between the iteration *k* and iteration *k*+1:

$$\epsilon = L\left(\theta^{(k+1)}\right) - L\left(\theta^{(k)}\right) \tag{9}$$

decrease to an acceptable value, with $\epsilon > 0$

In Step E the algorithm calculates the conditional expected value of the logarithm of the likelihood function for complete data given the observed sample and step M calculates its maximum.

This algorithm requires an initial solution for the values of the distribution parameters, designated by $\theta^{(0)}$. The selection of this initial solution requires particular attention to the extent that the algorithm convergence speed may become extremely slow due to a poor choice. Another aspect to take into account is the maximum likelihood equation can have multiple solutions corresponding to local maxima, so the choice of the starting solution becomes important.

A comparative study of various strategies in the choice of initial values can be found in Karlis and Xekalaki [11]. The results obtained by these authors demonstrate the importance of the choice of initial values, to have a reasonable convergence speed.

The figure 1 illustrates the EM algorithm procedure.

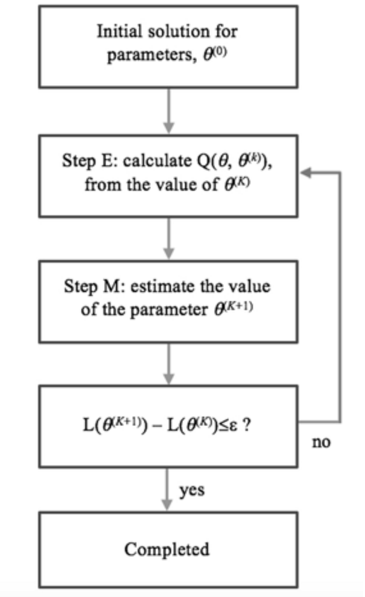
Initial solution for parameters, $\theta^{(0)}$
Step E: calculate Q(θ, $\theta^{(k)}$), from the value of $\theta^{(K)}$
Step M: estimate the value of the parameter $\theta^{(K+1)}$
L($\theta^{(K+1)}$) – L($\theta^{(K)}$)≤ε ?
no
yes
Completed

Fig. 1. EM algorithm procedure

The EM algorithm has several advantages which stand out relative to other iterative algorithms:

- The EM algorithm converges in a large variety of conditions, that is, from an arbitrary data, $\theta^{(0)}$ the algorithm usually finds a local maximum, except for a poor choice of initial solution $\theta^{(0)}$ or in the wrong formulation of the likelihood function.
- The required analytical work is simpler than with other methods, since only it is necessary to maximize the conditional expected value of the log-likelihood for complete data.
- The EM algorithm is relatively easy to program and implement.
- During the iterations it is possible to control the convergence and programming errors

However it also has some disadvantages:

- The EM algorithm can converge slowly, even in some apparently simple problems and on issues where there is a lot of incomplete information.
- The EM algorithm does not have an integrated process for producing an estimate of the covariance matrix of the estimated parameters. But this disadvantage can be overcome by using appropriate methodology.
- The EM algorithm does not guarantee convergence to the global maximum when there are multiple maximum sites. The obtained estimate depends on the initial solution

3.2. EM algorithm with right censored data

The observed data consists of (Y_i, δ_i), with Y_i = min (T_i, Cd_i), where Cd_i corresponds to the observation times, $\delta_i = 1$ ($T_i \le Cd_i$), corresponds to the non-censored data and $\delta_i = 0$ ($T_i > Cd_i$), to the censored data.

In the step E of the algorithm it is required the calculation of the conditional expected value of the log-likelihood for complete data, given the observed sample. In this case the log-likelihood function to complete data is given by equation (10), for n data collected:

$$\ln L(\eta,\beta) = l(\eta,\beta) = n\,ln\,\beta - n\beta\ln\eta + (\beta-1)\sum_{i=1}^{n}(\ln t_i) - \sum_{i=1}^{n}\left(\frac{t_i}{\eta}\right)^{\beta} \quad (10)$$

The $\theta = (\eta, \beta)$ are the distribution parameters, $\delta = (\delta_1, \delta_2, ..., \delta_n)$ is the vector indicator of censorship and $y = (y_1, y_2, ..., y_n)$ is the vector of observed data.

We have from equation (7):

$$Q\left(\theta,\theta^{(k)}\right) = n\ln\beta - n\beta\ln\eta + (\beta-1)\sum_{i=1}^{n}A_i - \sum_{i=1}^{n}B_i \quad (11)$$

Where:

$$A_i = E_{\theta^{(k)}}\left[\ln t_i | y_i\delta\right] = \delta_i \ln y_i + (1-\delta_i)E_{\theta^{(k)}}\left[\ln t_i | t_i > y_i\right] \quad (12)$$

$$B_i = E_{\theta^{(k)}}\left[t_i^{\beta} | y_i\delta\right] = \delta_i y_i^{\beta} + (1-\delta_i)E_{\theta^{(k)}}\left[t_i^{\beta} | t_i > y_i\right] \quad (13)$$

For the conditional expected value is first necessary to determine the expression of the corresponding conditional probability density function, $f_{(y|y')}$ $(y|y')$. For this case is given by [16]:

$$f(t_i|y_i) = \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1}\exp\left[\left(\frac{y_i}{\eta}\right)^{\beta} - \left(\frac{t_i}{\eta}\right)^{\beta}\right], t_i > y_i \quad (14)$$

So,

$$E_{\theta^{(k)}}\left[\ln t_i | t > y_i\right] = \ln y_i + \frac{1}{\beta^{(k)}}\exp\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}} \text{“}\left[0,\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\right] \quad (15)$$

$$E_{\theta^{(k)}}\left[t_i^{\beta} | t > y_i\right] = y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}} \quad (16)$$

Where:

$\Gamma(p,x) = \int_{x}^{\infty} u^{p-1}e^{-u}du$, *is the upper incomplete gamma function.*

Introducing equations (15) and (16) in equations (12) and (13):

$$A_i = \delta_i \ln y_i + (1-\delta_i)\left\{\ln y_i + \frac{1}{\beta^{(k)}}\exp\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}} \text{“}\left[0,\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\right]\right\}$$

(17)

$$B_i = \delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}} \right] \tag{18}$$

Thus, the expression referring to step E of the algorithm EM, $Q(\theta,\ \theta^{(k)})$ with censored data to the right is given by the following equation:

$$Q\left(\theta,\theta^{(k)}\right) = n\,ln\,\beta - n\beta\ln\eta + (\beta-1)\sum_{i=1}^{n}\left[\delta_i \ln y_i + \left(1-\delta_i\right)\left\{\ln y_i + \frac{1}{\beta^{(k)}}\cdots\right.\right.$$
$$\left.\left.\exp\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}} \Gamma\left[0,\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\right]\right\}\right] - \frac{1}{\eta^{\beta}}\sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\} \tag{19}$$

As mentioned above, the step M is intended to find the solution $\theta^{(k+1)}$ which maximize $Q(\theta,\ \theta^{(k)})$.

In order to get the points that maximize the function it is necessary to solve the partial derivatives of the above equation and equal them to zero, as shown in the following equations:

$$\frac{\partial Q\left(\theta,\theta^{(k)}\right)}{\partial\eta} = -\frac{n\beta}{\eta} + \frac{\beta}{\eta^{\beta+1}}\sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\} \tag{20}$$

$$\frac{\partial Q\left(\theta,\theta^{(k)}\right)}{\partial\beta} = \frac{n}{\beta} - n\,ln\,\eta + \sum_{i=1}^{n}\left[\delta_i \ln y_i + \left(1-\delta_i\right)\left\{\ln y_i + \frac{1}{\beta^{(k)}}\exp\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\cdots\right.\right.$$
$$\left.\left.\Gamma\left[0,\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\right]\right\}\right] + \sum_{i=1}^{n}\left[\frac{\ln\eta}{\eta^{\beta}}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\} - \frac{1}{\eta^{\beta}}\delta_i y_i^{\beta}\ln y_i\right] \tag{21}$$

Thus, the solution for the estimation of the distribution parameters η is given by the following equation:

$$\eta = \sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left(y_i^{\beta^{(k)}} + \eta^{(k)}\exp\left(\frac{y_i}{\eta^{(k)}}\right)^{\beta^{(k)}}\right)\right\}^{\frac{1}{\beta}} .\eta^{\frac{1}{\beta}} \tag{22}$$

With equation (22) for η parameter, it is possible to calculate the β parameter of the distribution.

The second derivative must be negative to ensure that the results obtained correspond to a maximum point. The second derivative equations are as follows:

$$\frac{\partial^2 Q\left(\theta,\theta^{(k)}\right)}{\partial\eta^2} = \frac{n\beta}{\eta^2} - \frac{\beta\left(\beta+1\right)}{\eta^{\beta+2}}\sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\} \tag{23}$$

$$\frac{\partial^2 Q\left(\theta,\theta^{(k)}\right)}{\partial\beta^2} = -\frac{n}{\beta^2} - \frac{\left(\ln\eta\right)^2}{\eta^{\beta}}\sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\}\cdots$$
$$+\frac{2\ln\eta}{\eta^{\beta}}\sum_{i=1}^{n}\left(\delta_i y_i^{\beta}\ln y_i\right) - \frac{1}{\eta^{\beta}}\sum_{i=1}^{n}\left[\delta_i y_i^{\beta}\left(\ln y_i\right)^2\right] \tag{24}$$

$$\frac{\partial^2 Q\left(\theta,\theta^{(k)}\right)}{\partial\eta\partial\beta} = \frac{\partial^2 Q\left(\theta,\theta^{(k)}\right)}{\partial\beta\partial\eta} = -\frac{n}{\eta} + \frac{1-\beta\ln\eta}{\eta^{\beta+1}}\sum_{i=1}^{n}\left\{\delta_i y_i^{\beta} + \left(1-\delta_i\right)\cdots\right.$$
$$\left.\left[y_i^{\beta^{(k)}} + \left(\eta^{(k)}\right)^{\beta^{(k)}}\right]\right\} + \frac{\beta}{\eta^{\beta+1}}\sum_{i=1}^{n}\left(\delta_i y_i^{\beta}\ln y_i\right) \tag{25}$$

With this analysis it is possible to calculate the Weibull distribution parameters β and η, through the knowledge of the maximum values of function Q.

4. Case study

The case study concerns the history of five centrifugal pumps failures of a petrochemical company, used to pump oil with similar density for the period 2006 to 2013.

From the collected data it can be seen that one of the pumps components is responsible for an important number of failures. It is the mechanical seal which represents approximately 45% of the total number of failures. Thus, based on the assessment of these results, it justifies the need of further study on this component.

For the purposes of this study, we select only the data related to failures due to excessive leakage of fluid to the outside of the pumps. The reason for this consideration is due to the significant number of failures and to restrict the study for just one failure mode.

The possible failures between inspections are treated as complete data, as the centrifugal pumps in question are visually inspected by the user of the equipment at least every 8 hours. So, it was not considered the possible mistake between two inspections compared to the total time of the observation and it was assumed that the exact moment of failure was well known.

The last record in each pump does not correspond to a fault but at the end of the test, since the pumps continued in operation. Thus the last recording time of each pump was considered as right censored data.

For the estimated values of the Weibull distribution parameters by the maximum-likelihood method it was used the EM algorithm for right censored data as described in section 3.2.

For the algorithm implementation process it was used the statistical program R.

The autors used a manually written code in R software and also used some packages like for exemple, maxNR, boot and mass.

The initial solution $\theta^{(0)}$ was attributed by the result obtained by the least squares method. The iterative process stopped when the difference between the iteration k and iteration k+1, given by equation (9), was less than 0,1.

The following Table 1 shows the expected value for $\hat{\beta}$ and $\hat{\eta}$ for each of the mechanical seals applying the maximum-likelihood method with the EM algorithm, as described in this paper. The respective confidence intervals are also presented.

Table 1. Expected value of $\hat{\beta}$ and $\hat{\eta}$ (days) for each of the mechanical seals obtained with the maximum-likelihood method (EM) and respective confidence interval by the bootstrap-T method

Mechanical seals	$\hat{\beta}$			$\hat{\eta}$ (days)		
1	6,12	9,21	14,09	365,51	392,96	425,53
2	2,01	5,04	10,12	301,49	345,31	392,67
3	10,85	13,92	17,19	385,03	403,68	423,21
4	4,93	8,17	12,98	352,85	381,44	416,05
5	3,96	7,56	12,51	335,62	368,37	406,99

Because the sample of data is small, it was used the bootstrap-t method in determining the confidence intervals with a confidence level of 95% [6, 7].

The bootstrap-t method allows the calculation of the confidence interval of the parameters of interest in particular in the case where the sample is small (n <30).

New samples are taken randomly from resampling the original sample. For the randomization of the process be minimized it is necessary to perform a large number of resampling, B. With the generated bootstrap samples it is possible to calculate the standard deviation of B repetitions that will be used in the confidence intervals:

$$\hat{\sigma}_{boot}\left(\hat{\theta}^*\right) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}\left[\hat{\theta}_i^* - \hat{\theta}^*(.)\right]^2} \qquad (26)$$

Where:

$$\hat{\theta}^*(.) = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^* \qquad (27)$$

From the table of the t-Student distribution gets the value t_c that:

$$P\left[-t_c \le \frac{\hat{\theta}-\theta}{\hat{\sigma}_{boot}\left(\hat{\theta}^*\right)} \le +t_c\right] = (1-\alpha) \qquad (28)$$

Thus, the bootstrap-t confidence interval with confidence level 100 (1-α)% is given by:

$$\left[\hat{\theta} - t_c.\hat{\sigma}_{boot}\left(\hat{\theta}^*\right), \hat{\theta} + t_c.\hat{\sigma}_{boot}\left(\hat{\theta}^*\right)\right] \qquad (29)$$

The number of resampling for the bootstrap method is equal to 1000. From the results obtained and presented in table 1, it can be noted that, all mechanical seals feature the shape parameter β>1. The values of the scale parameter η vary between 345.31 and 403.68 days of operation.

With the information obtained by the bootstrap method, it is possible to represent the confidence interval around the probability density function of the estimated Weibull distribution, as shown in Figure 2 to seal 1.

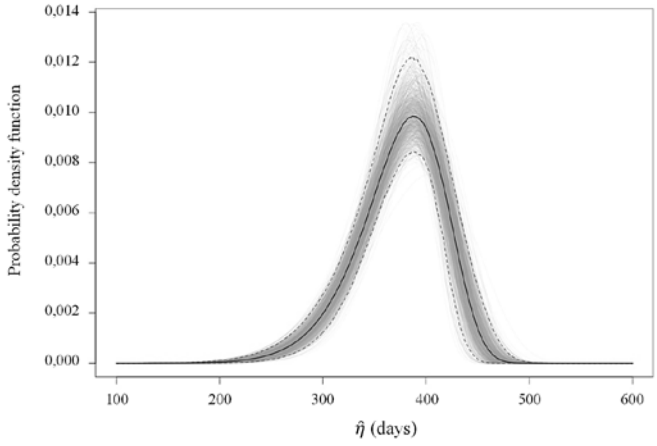

Fig. 2. Confidence interval around the probability density function of the estimated Weibull distribution by the bootstrap method, seal 1

Table 2. Expected value of $\hat{\beta}$ and $\hat{\eta}$ (days) for each of the mechanical seals obtained with the least squares method

Mechanical seals	$\hat{\beta}$	$\hat{\eta}$ (days)
1	8,76	388,72
2	4,26	338,17
3	12,54	415,15
4	7,68	386,71
5	6,55	363,42

The estimation of the unknown parameters of the Weibull distribution was also obtained, in the same system, with the least squares method, to compare with the results obtained by the EM algorithm.

In Rinne is indicated how to obtain the estimates of the unknown parameters using least squares method.

The following Table 2 shows the expected value for $\hat{\beta}$ and $\hat{\eta}$ for each of the mechanical seals applying the least squares method.

The shape parameter, β, is higher by maximum likelihood method (EM algorithm) than the least squares method and in both is higher than 1.

For the scale parameter, η, the values obtained by the two methods are similar.

5. Conclusion

The prediction of failures of equipment for a given time horizon can only be considered in probabilistic terms, because there is always uncertainty about the time they will happen. Thus, this work has focused on the presentation of statistical techniques inherent to an estimation process of the parameters of a theoretical probabilistic model that best fits the observed data.

Also, it showed the importance of the record of occurrences of failures during the use of equipment (historical record). This information is essential for monitoring a system, and also for the correct performance of the maintenance activities.

However, some data may contain uncertainty regarding the time of occurrence of the failures and thus are regarded as incomplete or partial.

Due to the complexity of the analyzed data, in particular in the presence of right censored data, it was found that the equation from the maximum-likelihood method showed no analytical solution. So the solution presented in this work was using the EM algorithm.

In this work it was possible to validate the applicability of the EM algorithm to determine the solutions of the equation that is derived from the maximum-likelihood method in the presence of incomplete data.

Acknowledgement

The research work partially financed by FCT (portuguese Foundation for Science and Technology.

References

1. Abernethy R B. The New Weibull handbook. Florida: Robert B. Abernethy, 2006.
2. Akram M, Hayat A. Comparison of estimators of the Weibull distribution. Journal of Statistical Theory and Practice 2014; 8(2): 238-259, https://doi.org/10.1080/15598608.2014.847771.
3. Balakrishnan N, Mitra D. Left truncated and right censored Weibull data and likelihood inference with an illustration. Computational Statistics and Data Analysis 2012; 56: 4011-4025, https://doi.org/10.1016/j.csda.2012.05.004.
4. Balakrishnan N, Kundu D, Ng H K T. Point and interval estimation for a simple step-stress model with type-II censoring. Journal of Quality Technology 2007; 39: 35-47.

5. Chambers R L, Steel D G, Wang S, Welsh A H. Maximum likelihood estimation for sample surveys. Boca Raton, Florida: Chapman and Hall/CRC Press, 2012.
6. Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 1977; 39(1): 1-38.
7. Efron B, Tibshirani R J. An introduction to the bootstrap. London: Chapman & Hall, 1993, https://doi.org/10.1007/978-1-4899-4541-9.
8. Fang L Y, Arasan J, Midi H, Bakar M R A. Jackknife and bootstrap inferential procedures for censored survival data. AIP Conference Proceedings 2015; 1682: 1-6, https://doi.org/10.1063/1.4934631.
9. Gijbels I. Censored data. Wires Computational Statistics 2010; 2: 178 - 188, https://doi.org/10.1002/wics.80.
10. Guure C B, Ibrahim N A. Methods for estimating the 2-parameter Weibull distribution with type-I censored data. Journal of Applied Sciences, Engineering and Technology 2013; 5(3): 689-694.
12. Karlis D, Xekalaki E. Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics and Data Analysis 2003; 41: 577-590, https://doi.org/10.1016/S0167-9473(02)00177-9.
13. Kinaci I, Akdogan Y, Kus C, Ng H K T. Statistical inference for Weibull distribution based on a modified progressive type-II censoring scheme. Sri Lankan Journal of Applied Statistics 2014; 1: 95-116, https://doi.org/10.4038/sljastats.v5i4.7786.
14. Lawless J F. Statistical models and methods for lifetime data. New Jersey: John Wiley & Sons, 2003.
15. McCool J I. Using the Weibull distribution, reliability, modeling and inference. New York: John Wiley & Sons, 2012, https://doi.org/10.1002/9781118351994.
16. McLachlan G J, Krishnan T. The EM algorithm and extensions. New Jersey: John Wiley & Sons, 2008, https://doi.org/10.1002/9780470191613.
17. Ng H K T, Chan P S, Balakrishnan N. Estimation of parameters from progressively censored data using EM Algorithm. Computational Statistics and Data Analysis 2002; 39: 371-386, https://doi.org/10.1016/S0167-9473(01)00091-3.
18. Procaccia H, Ferton É, Procaccia M. Fiabilité et maintenance des matériels industriels réparables et non réparables. Paris: Ed. Tec & Doc, 2011.
19. Rinne H. The Weibull distribution - A handbook. Florida: CRC Press, 2009.
20. Teimouri M, Hoseini S M, Nadarajah S. Comparison of estimation methods for the Weibull distribution. Statistics: Journal of Theoretical and Applied Statistics 2013; 47(1): 93-109, https://doi.org/10.1080/02331888.2011.559657.
21. Tobias P A, Trindade D C. Applied reliability. Florida: Chapman & Hall/CRC Press, 2011.

Luís Andrade FERREIRA
FEUP - Faculdade de Engenharia da Universidade do Porto
Department of Mechanical Engineering
Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

José Luís SILVA
ESTV – Escola Superior Tecnologia de Viseu
Department of Mechanical Engineering and Industrial Management
Campus Politécnico, 3504-510, Viseu, Portugal
E-mail: lferreir@fe.up.pt, jsilva@ipv.pt