Q2

Simon Jorstedt

2023-12-15

Question 2 - EM Algorithm

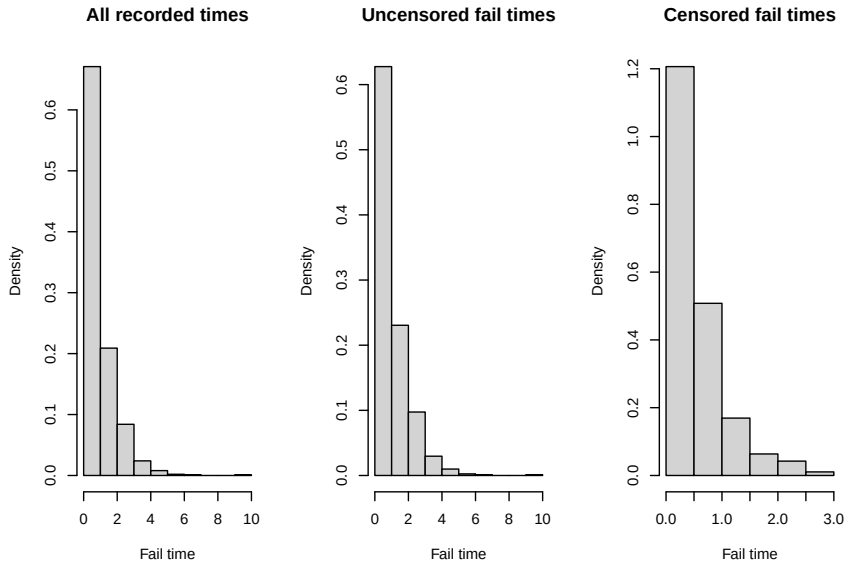
A certain type of product fails after random times. We are given a dataset of observations of these fail times. Most observations represent the exact fail time, but for some observations, the time that a product was discovered to have failed is recorded instead. In other words, these observations are censored. Assuming that fail times and time until discovery are independent, this would imply that the uncensored observations can be modeled by some probability distribution $X \sim D$, and the censored observations can be modeled by the distribution of a random variable $X + Y$ where Y is a random variable from the distribution of the time between a product failing and it being discovered. In the instructions, it is stated that the censored measurements are left censored, which makes sense in this context.

In the instructions it is also stated, that if the uncensored observations are assumed to be exponentially distributed, then the censored observations follow a truncated exponential distribution. This makes less sense. In addition, when the distribution of the observed fail times is investigated visually in Figure 2.1 below, we find that the *allegedly* left censored and/or truncated censored observations definetly are not left truncated, and likely not left censored either. Huh. When asked to clarify this conundrum, the teacher responsible for the course and the lab instructions, Krzysztof Bartoszek, stated that perhaps the censored data is in fact right-truncated. This interestingly makes even less sense. In addition, we the students were encouraged to (paraphrasing) "just do the work and follow the instructions regardless of what the data looks like".

... Anyway, let us proceed with this *interesting* lab assignment.

Question 2.1

Figure 2.1: Density of fail times



All three histograms in Figure 2.1 appear approximately exponential. The density for all recorded times is not significantly different from the uncensored fail times, but it should be noted that the censored fail times only make up about 23 % of the observations.

Question 2.2

Let us now assume that the fail times are exponentially distributed with density $f(x) = \lambda e^{-\lambda x}$. **According to the instructions**, this means that the censored observations must follow a truncated exponential distribution (again, this does not make much sense, especially with respect to Figure 2.1). Under these (blatantly incorrect) assumptions, we can formalise the likelihood of our data below. Let us assume that the truncation has restricted the exponential distribution into the interval $[a, b]$, where $0 \leq a < b$. We introduce an indicator variable δ_i which takes the value 1 if observation x_i is censored, and 0 if it is not. The likelihood of a censored observation x_i is

$$\frac{g(x_i)}{F(b) - F(a)}$$

where $g(x) = f(x)$ for $a \leq x < b$, and $g(x) = 0$ otherwise. Using this, we can formulate the likelihood function for the data in the following way,

$$L(\lambda) = \prod_{i=1}^{n} f(x_i)^{(1-\delta_i)} \cdot \left(\frac{f(x_i)}{F(b) - F(a)} \right)^{\delta_i} = \Big(F(b) - F(a) \Big)^{-n\bar{\delta}} \cdot \prod_{i=1}^{n} \Big(\lambda e^{-\lambda x_i} \Big)^{(1-\delta_i)} \cdot \Big(\lambda e^{-\lambda x_i} \Big)^{\delta_i} \propto \lambda^n e^{-\lambda n \overline{x}}$$

where $n\bar{\delta}$ is the number of censored observations in the dataset. Interestingly, the maximum likelihood estimate can be analytically determined from this Likelihood function. It comes out to

$$\hat{\lambda}_{MLE} = \frac{1}{\overline{x}} \approx 1.1$$

Question 2.3

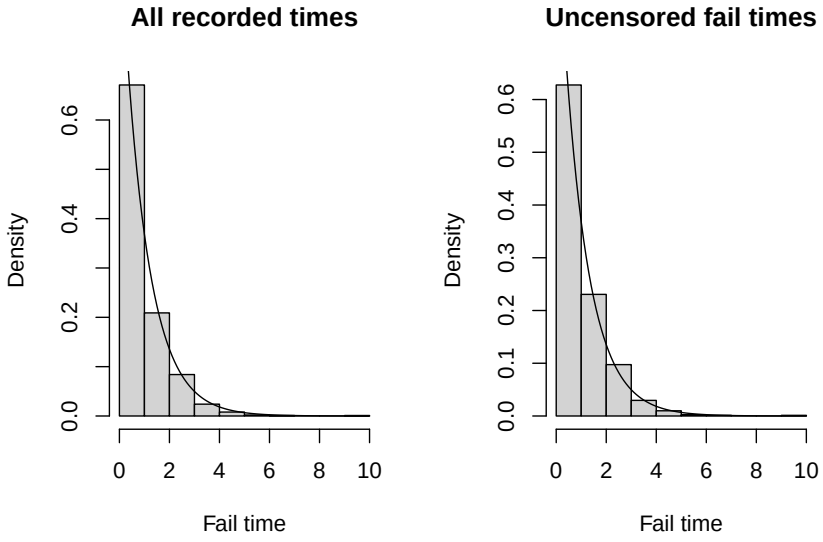
Let

Question 2.4

```r
Likelihood <- function(lambda, data){
  first_term <- (lambda*exp(-lambda*data$time))^(data$cens == 1)

  second_term <- (1-exp(-lambda*data$time))^(1 - (data$cens == 1))

  prod(first_term, second_term)
}

Q <- function(theta, theta_k){

}


EM_algorithm <- function(lambda_start = 100, epsilon = 0.001){
  # Setup
  k <- 0
  k_max <- 100
  lambda =

  # Keep iterating until convergence is reached
  while(k <= k_max & F){
    # E Step

    # M Step

  }
}
```

Question 2.5

Figure 2.2: Density of fail times with fitted exponential density



TESTING