Comparison of distributions Log Normal, Weibull and Exponential using HIV data.

Vinicius Raniero Angelo
Universidade Federal de Mato Grosso, Cuiabá, Brazil – viniciusuf@gmail.com

Abstract

Raises the question which of these three distributions (Exponential, Lognormal, and Weibull) best fits the survival data of people with HIV virus. This paper aims to address the survival analysis comparing those distributions using the criteria AIC, BIC, CAIC and Kaplan-Meier estimator, to decide the best distribution that fits the survival time of people with HIV. The results showed that the Exponential distribution fitted better than others distributions.

Keywords: Survival analysis; HIV; Kaplan-Meier; Exponential.

1. Introduction

The Acquired Immune Deficiency Syndrome, is caused by HIV. Since this virus attacks our body's defense cells, the body is more vulnerable to various diseases, from the common cold to more serious infections such as tuberculosis or cancer. Even the treatment of these diseases is impaired. A few years ago, being diagnosed with AIDS was a death sentence. But today, it is possible to be HIV positive and living with quality of life. Because of this it's interesting to estimate the survival time in HIV carriers.
In survival analysis, by wanting to set a distribution curve with the Kaplan-Meier curve, it's necessary to know what distribution best fits the model. It is known that when the total test time curve (TTT) is decreasing, one can use the parametric distributions. We know that the exponential and Weibull distributions have similarities DAS (2008) and that both can be used to model survival time of people with HIV, but it is unclear what the behavior of the lognormal distribution for this type of variable. Therefore this study aims to evaluate which of the three distributions, Log Normal, Exponential and Weibull, best fit to the data set of HIV patients, comparing its effectiveness through the statistics AIC, BIC and CAIC.

2. Methodology

Data were obtained from the Institute of Clinical Research Evandro Chagas (IPEC / Fiocruz) and comes to 193 individuals with HIV treated between 1986 and 2000. The data set is the AIDS-Classical obtained on the site "Sobrevida" of Fiocruz (2015). Were considered for this study only three variables : Individual identification, treatment time and censors.
The construction of the total time of the test chart (TTT curve) proposed by Aarset (1987) was used to observe the behavior of risk function. The TTT curve was obtained by the following formula:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{r} T_{i:n} + (n-r)T_{r:n}}{\sum_{i=1}^{n} T_{i:n}}, \text{ by } \frac{r}{n},$$

Where n is the size of the sample, r = 1, ..., n Ti: n, i = 1, ..., n are sample order statistics. To estimate the parameters we used the maximum likelihood method for the three distributions.
Distributions were used Log Normal, Exponential and Weibull, they have the following densities respectively: Casella (2002)

$$f(x;\mu,\sigma) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left[\dfrac{-(\log(x)-\mu)^2}{2\sigma^2}\right], & \text{se } x > 0 \\ 0 \text{ caso contrário} \end{cases}$$

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ se } x \geq 0 \\ 0 \text{ se } x < 0 \end{cases}$$

$$f(x) = \begin{cases} \frac{\alpha}{\beta^\alpha} x^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^\alpha\right], & \text{se } x \geq 0 \\ 0 \text{ se } x < 0 \end{cases}$$

The estimates of the parameters of the distributions cited above were made by the maximum likelihood method comprising the following generic form:

$$L(\theta; x_1, \ldots, x_n) = f(x_1;\theta) \times \ldots \times f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

Being θ the unknown parameter and xn the observed values and the estimators that maximize $\boldsymbol{L(\theta)}$ are found by solving the following system of equations: Casella (2002)

$$U(\theta) = \frac{\partial \log(L(\theta))}{\partial\theta} = 0$$

The criteria for comparing the complexity of the model and test which distribution that best fits the Kaplan-Meier curve were the AIC, BIC and CAIC, given by their formulas: COLOSIMO (2006)

$$AIC_p = -2log(L_p) + 2[(p+1)+1],$$

$$BIC_p = -2log(L_p) + [(p+1)+1]log(n).$$

$$\boldsymbol{CAIC_p = AIC + \frac{log(p+2)(p+3)}{n-p-3}}$$

The $\boldsymbol{L_p}$ is the function of maximum likelihood of the model and p is the number of explanatory variables considered in the model and n is the sample size.
Statistical analysis and graphs were performed in R. Software.(R Development Core Team (2012))

3. Results and Discussion

For the variable survival time of patients with HIV, was designed to TTT curve (Figure 2). Analyzing it we can see that it indicates a risk function in decreasing order, so we can use the Log Normal distributions, Weibull and Exponential.
The Figure 1 shows a comparison of the estimates of Kaplan-Meier survival function and the second to the models Log Normal, Weibull and Exponential, for the variable survival time of patients with HIV, through it we can see that the model that best fits Kaplan Meier survival function is exponential.
Table 1 shows the AIC, CAIC and BIC statistics between the adjusted models, we can see that the model with the smallest statistics is the exponential model.

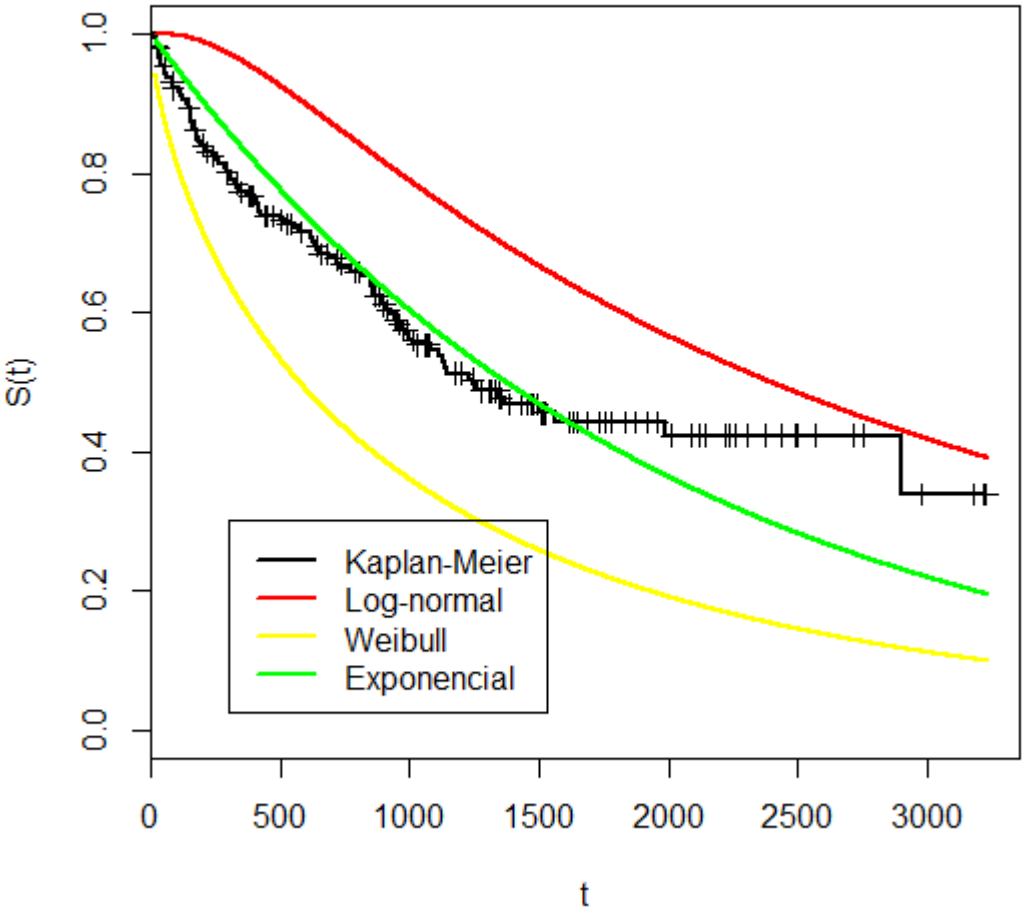
S(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
t
Kaplan-Meier
Log-normal
Weibull
Exponencial

Figure 1: survival function estimates according to Kaplan-Meier and according to the Log-Normal models, Exponential and Weibull, to the variable survival time of patients with HIV.

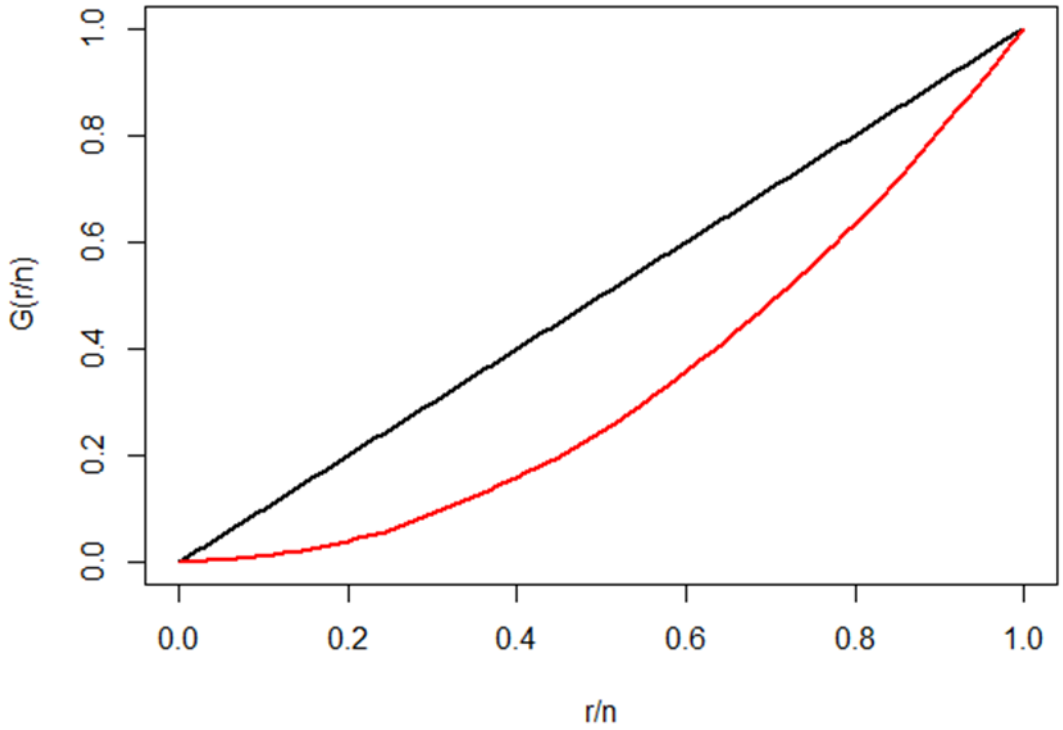

Figure 2: TTT curve for the variable survival time of patients with HIV.

Table 1: Adjustment of models compared to the variable survival time of patients with HIV.

Model	AIC	CAIC	BIC
Exponential	1551,269	1551,396	1554532
LogNormal	2474,556	2474,769	2481,081
Weibull	1595,522	1595,735	1602,437

5. Conclusions

The exponential distribution was better than the Log Normal and Weibull distributions in the present study. We can evaluate the exponential distribution was better than the other distributions, because analyzing Figure 1, it is the one that best fits the Kaplan-Meier also the AIC statistics, BIC and CAIC of Exponential were lower than the statistics other tested distributions.

References

DAS, K. A comparative study of exponential distribution vs Weibull distribution in machine reliability analysis in a CMS design. **Computers & Industrial Engineering**, v. 54, n. 1, p. 12-33, 2008.

Aarset, M. V., 1987. How to identify bathtub hazard rate. Transactions on Reliability 36, 106–108.

Casella, George, and Roger L. Berger. *Statistical inference*. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

COLOSIMO, Enrico Antônio; GIOLO, Suely Ruiz. Análise de sobrevivência aplicada. In: **ABE-Projeto Fisher**. Edgard Blücher, 2006.

R Development Core Team, 2012. R: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org