BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1795–1802
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1795-1802

SURVIVAL FUNCTION AND HAZARD FUNCTION ANALYSIS OF EXPONENTIAL DISTRIBUTION IN TYPE I CENSORED SURVIVAL DATA: A CASE STUDY OF BREAST CANCER PATIENTS

Ardi Kurniawan[1*], **Anggara Teguh Previan**[2], **Zidni 'Ilmatun Nurrohmah**[3]

[1,2,3]*Departement of Mathematics, Faculty of Sains and Technology, Airlangga University*
Kampus C Universitas Airlangga, Surabaya, 60115, Indonesia

Corresponding author's e-mail: ***ardi-k@fst.unair.ac.id***

ABSTRACT

Article History:

Received: 31st May 2023
Revised: 20th August 2023
Accepted: 29th August 2023

Keywords:

Analysis of survival;
Breast cancer;
Exponential distribution;
Maximum likelihood estimation

Breast cancer is the most common cancer in women and the leading cause of cancer-related death in Indonesia. Analysis of survival data is important for improving the treatment and care of breast cancer patients. This study aims to estimate the parameters, find the survival function, and hazard function of breast cancer patients using a parametric method with an exponential distribution. Previous studies have shown that the Maximum Likelihood Estimation (MLE) method is suitable for estimating the survival function from exponential survival data by censoring. In this study, the exponential distribution was found to be the best for data on breast cancer patients from Surabaya Ontology Hospital. The estimated parameters are θ = 33.9157, and the survival function is calculated using $S(x) = e^{\left(-\frac{x}{33.9157}\right)}$. The estimated hazard function for patient death or failure is 0.0295. The results of this study can contribute to the development of better treatment and care strategies for breast cancer patients. However, further research is needed because this study only used monthly time units.

How to cite this article:

A. Kurniawan, A. T. Previan and Z. I. Nurrohmah., "SURVIVAL FUNCTION AND HAZARD FUNCTION ANALYSIS OF EXPONENTIAL DISTRIBUTION IN TYPE I CENSORED SURVIVAL DATA: A CASE STUDY OF BREAST CANCER PATIENTS," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1795-1802, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Based on reports from the World Health Organization (WHO) and the Ministry of Health of the Republic of Indonesia in 2020, breast cancer is the most common type of cancer suffered by women. Breast cancer is also one of the biggest contributing factors to the cause of death in cancer patients [1]. Efforts to cure cancer patients need to be developed to reduce mortality, because this is in line with one of the points in the third Sustainable Development Goals (SDGs) regarding Good Health and Well-Being which aims to reduce mortality due to preventable and treatable diseases [2]. As well as increasing people's access to quality health care. Therefore, an appropriate method is needed in analyzing breast cancer patient survival data in order to improve the management and care of these patients.

Analysis of life test data or survival analysis is one of the statistical analyzes that discusses the survival of an object or individual. Analysis of survival test data aims to estimate the probability of survival, recurrence, death and other events up to a certain period. The life time period is recorded as data about the duration of an event from the beginning to the end. The definition of survival time here is the time the patient is first diagnosed with a disease until the patient dies. In the survival analysis, there are frequently encountered censored data, where information about the patient's time of death is not yet available at the end of the observation period [3]. To overcome this, parametric methods, such as the exponential distribution, can be used to model the distribution of patient survival time [4].

Studies related to the estimation of survival model parameters or survival analysis from exponential distributions have been carried out by many previous researchers. Among them is Chen et.al [5] which discusses the use of Maximum Likelihood Estimation (MLE) to estimate the survival function in exponential type I censored life test data. Another study was a study conducted by Yu et.al [6] which used a progressive type I cencoring technique to estimate parameters from the exponential distribution and calculate the survival function from survival test data. Then, Burdiantoro [7] conducted a study of type I censored survival data with exponential distribution and six sigma. From these studies it can be concluded that the use of MLE in estimating the survival function of exponential type I censored survival test data has been extensively studied and can provide accurate results.

This study aims to analyze the survival function and hazard function in breast cancer patient survival data using exponential distribution parametric method. This modeling is expected to provide a better understanding of the selection of the most suitable survival time distribution model for the data. The results of this study can contribute to the development of better management and care strategies for breast cancer patients.

2. RESEARCH METHODS

2.1 Study of Literature

Literature study in this study was carried out by reading and studying the concepts related to survival analysis, such as the probability density function, survival function, hazard function, censored data, and the MLE method. The following is an explanation of the definitions or basic theories of these concepts.

2.1.1 Distribusi Eksponensial

The exponential distribution is a distribution commonly used in survival test cases. As for the probability density function of the probability distribution with parameters Ө is as follows

$$f(x) = \theta\ exp(-x\theta) \text{ with } x > 0, \theta > 0 \tag{1}$$

The mean value and variance of the exponential distribution are,

$$E(x) = \theta^{-1} \text{ and } Var(x) = \theta^{-2} \tag{2}$$

2.1.2 Survival Function

The survival function is the probability of an individual surviving beyond x time. The survival function is denoted by $S(x) = Pr(X > x)$. Mathematically stated as follows [8],

$$S(x) = Pr(X > x) = 1 - F(x) \tag{3}$$

2.1.3 Hazard Function

The hazard function is the probability that an individual will die or fail if it is known that the individual will survive until time *x*. The hazard function is denoted by *h(x)* and is formulated as follows [9],

$$h(x) = \frac{f(x)}{S(x)} \tag{4}$$

Meanwhile, the cumulative hazard function is a description of the cumulative disturbance or risk of a product during the time interval [0, x]. The cumulative hazard function is denoted by *H(x)* and its formula is,

$$H(x) = -ln\, S(x) \tag{5}$$

2.1.4 Censored Sample

In the analysis of live test data, the presence of censored sample data is the difference between this analysis and other statistical analyses. Censored samples are obtained if not all observation units observed during a certain research period fail so that the actual survival time of some observations is unknown [10]. Censorship on research samples can be done to achieve experimental efficiency because measuring the time of individual failure or death takes a long time and costs a lot. Based on the limited time of the research, there are three types of censorship in the live test data samples, namely:

Type I Censored Sample

The sample is said to be type I censored if the research time has been determined and the research object *(n)* enters the study at the same time then the research is stopped at a certain time limit [11]. The weakness of type I censorship is that live test data from the object under study will not be obtained if all the research objects are still alive by the time limit [4].

Type II Censored Sample

The sample is said to be type II censored if individuals enter the study at the same time and the study is stopped if the number of deaths *(r)* that occurs matches what has been determined from *(n)* the objects of observation [11]. The weakness of type II censorship is that the time needed to obtain r deaths is likely to be very long, but survival data from the object under study will definitely be obtained.

Type III Censored Sample

The sample is said to be type III censored if each individual enters the study at different times during the study period [12]. Some individuals who died or failed before the observation ended had survival data, some were still alive until the end of the study, and some were still alive but left the study.

2.1.5 Likelihood Function

The joint density function of n random variables $X_1, X_2, \ldots, X_n$ which yields $X_1, X_2, \ldots, X_n$ say $f(x_1, x_2, \ldots, x_n; \theta)$ is the Likelihood function. Often the Likelihood function for parameter θ is denoted $L(\theta)$. If $X_1, X_2, \ldots, X_n$ is a random variable that has pdf$f(x;\theta)$ which is **IID** then,

$$L(\theta) = f(x_1;\theta)\, f(x_2\,;\theta) \ldots f(x_n;\theta) \tag{6}$$

In survival data that contains censored data, the form of the likelihood function is slightly different from the general one. This is because survival data is influenced by life time and sensor time which are independent of each other, denoted as follows,

$$L(\theta) = \prod_{i=1}^{n} [f(x_i)]^{\delta_i} [S(x_i)]^{1-\delta_i} \tag{7}$$

With δ_i=1 indicating uncensored data, while δ_i=0 indicating censored data. $S(x_i)$ is the survival function for x_i [13].

2.1.6 Maximum Likelihood Estimation (MLE) Method

Maximum Likelihood Estimation (MLE) is a method of parameter estimation. Suppose $X_1, X_2, \dots, X_n$ are random variables of a distribution with density function $f(x;\theta), \theta\epsilon\Omega$ are parameter spaces. The shared PDF between $X_1, X_2, \dots, X_n$ is $f(x_1;\theta)\, f(x_2\,;\theta) \dots f(x_n;\theta)$. If the joint PDF is expressed as a function of θ it is called the likelihood function which is written as follows,

$$L(\theta; x_1, x_2, \dots, x_n) = f(x_1;\theta)\, f(x_2\,;\theta) \dots f(x_n;\theta) \tag{8}$$

The statistic $\hat{\theta} = t(x_1, x_2, \dots x_n)$ that maximizes the likelihood function $L(\theta; x_1, x_2, \dots, x_n), \theta\ \epsilon\Omega$ is the statistic $\hat{\theta} = t(x_1, x_2, \dots x_n)$ which is called the Maximum Likelihood Estimation (MLE) of θ. However, difficulties are usually found when the first decrease in the likelihood function for the parameter is equal to zero. Therefore, it can be done by determining the maximum value of the natural logarithm of the likelihood function or log likelihood with the aim of facilitating parameter estimation. Notated as follows [13].

$$\ln L(\theta; x_i) = \prod_{i=1}^{n} ln\, f(x_i, \theta) \tag{9}$$

2.2 Data Types and Sources

In this study, the data to be used is life time data of breast cancer patients which are calculated from the time they are diagnosed with breast cancer to the fatality phase (death) in months. This data is secondary data of 15 data sourced from the Surabaya Oncology Hospital [14].

2.3 Data processing

Data processing is carried out based on quantitative methods both descriptively and inference with the formulas obtained in point **2.1 Study of Literature.** The stages of processing are as follows:

1. Determine type I censored data.
2. Determine the distribution of data using the Anderson-Darling test.
3. Estimating parameters.
4. Determine the survival function of exponentially distributed life test data.
5. Determine the hazard function of exponentially distributed live test data.

3. RESULTS AND DISCUSSION

3.1 Anderson Darling Test

The Anderson Darling test is used to determine whether a data has a certain distribution. In this study, data analysis was carried out on the exponential distribution [15].

Hypothesis:

H_0:Samples taken from exponential distribution

H_a: Samples not taken from exponential distribution

Testing criteria: Reject H_0 when the $A^2_{test} > A^2_{table}$

Table 1. Anderson Darling Test Calculation

Distribution	A^2_{Test}
Exponential	1.288

Based on **Table 1** it can be seen the results of the Anderson Darling test calculations. In these calculations it is known that the data fulfills the Exponential distribution. This is because the value of A^2_{test} is less than $A^2_{table} = 2.502$ so that H_0 which states that the data follows an Exponential distribution is accepted.

3.2 Determine Parameter Estimation

After the data is known to follow an exponential distribution, the next step is to bootstrap 15 times to get more accurate results and estimate parameters based on the exponential distribution [15]. The data to be tested will be censored when the lifetime is more than 50 months.

Table 2. Bootstrap Results Research Data

No	Time (x) in month	Censored (δ)
1	18	1
2	56	0
3	56	0
4	56	0
5	66	0
6	66	0
7	8	1
8	6	1
9	35	1
10	5	1
11	18	1
12	18	1
⋮	⋮	⋮
749	56	0
750	18	1

Based on Table 2, the censored data is data that has a value of $\delta_i = 0$. By using the maximum likelihood estimation method, the estimated parameters are obtained as follows. The first step is to determine the likelihood function as in Equation (7).

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}\left([f(x_i)]^{\delta_i}[S(L_i)]^{1-\delta_i}\right) \\ &= \prod_{i=1}^{n}\left(\left(\frac{1}{\theta^{\delta_i}}e^{-\frac{x_i\delta_i}{\theta}}\right)e^{-\frac{L_i(1-\delta_i)}{\theta}}\right) \\ &= \prod_{i=1}^{n}\left(\left(\frac{1}{\theta^{\delta_i}}e^{-\frac{x_i\delta_i}{\theta}}\right)e^{-\frac{x_i(1-\delta_i)}{\theta}}\right) \\ &= \frac{1}{\theta^{\sum_{i=1}^{n}\delta_i}}e^{-\sum_{i=1}^{n}\frac{x_i}{\theta}} \\ &= \frac{1}{\theta^{(1+0+0+\cdots+1)}}e^{-\frac{(18+56+56+\cdots+18)}{\theta}} \\ L(\theta) &= \frac{1}{\theta^{605}}e^{-\frac{20519}{\theta}} \end{aligned} \tag{10}$$

To facilitate the derivation of the likelihood function, the logarithm is given for the two sides of the likelihood function.

$$\begin{aligned} \ln\big(L(\theta)\big) &= \ln\left(\frac{1}{\theta^{605}}e^{-\frac{20519}{\theta}}\right) \\ &= \ln\left(\frac{1}{\theta^{605}}\right)\ln\left(e^{-\frac{20519}{\theta}}\right) \\ &= \ln\left(\frac{1}{\theta^{605}}\right)\left(-\frac{20519}{\theta}\right) \\ \ln\big(L(\theta)\big) &= -605\ln(\theta) - \frac{20519}{\theta} \end{aligned} \tag{11}$$

After that, the reduction is carried out in Equation (11) and the maximum value of the derivative is sought, by solving the equation, the estimated parameters are obtained as follows.

$$\frac{\partial\big(\ln\big(L(\theta)\big)\big)}{\partial\theta} = 0$$

$$\frac{\partial\left(-605\ln(\theta)-\frac{20519}{\theta}\right)}{\partial\theta}=0$$
$$-\frac{605}{\theta}+\frac{20519}{\theta^2}=0$$
$$\frac{-605\theta+20519}{\theta^2}=0$$
$$-605\theta+20519=0$$
$$-605\theta=-20519$$
$$\theta=\frac{20519}{605}$$
$$\theta=33.9157$$

So that the estimated parameter is $\theta = 33.9157$. This means that the longer a patient survives, the less likely it is to fail or die. This parameter estimation can also be used to predict the patient's chances of survival in the future, as well as assist doctors in determining the right type and duration of treatment for the patient.

3.3 Determine the Survival Function and Hazard Function

After the parameter estimation values are obtained, the survival function can be determined through **Equation (7)**.

$$\begin{aligned} S(x) &= 1-F(x)\\ &= 1-\left(1-e^{\frac{-x}{\theta}}\right)\\ &= e^{\frac{-x}{\theta}} \end{aligned}$$
$$S(x)=e^{\frac{-x}{33.9157}} \tag{12}$$

Then, it can be seen that the survival function when time *t* can be seen in **Table 3**.

Table 3. Survival Probability

Time (x) in month	Survival Probability
0	1
1	0.970946
2	0.942735
3	0.915345
⋮	⋮
49	0.235803
50	0.228952

In addition, the survival chances of patients with breast cancer based on the time x are able to survive can be seen in **Figure 1**.

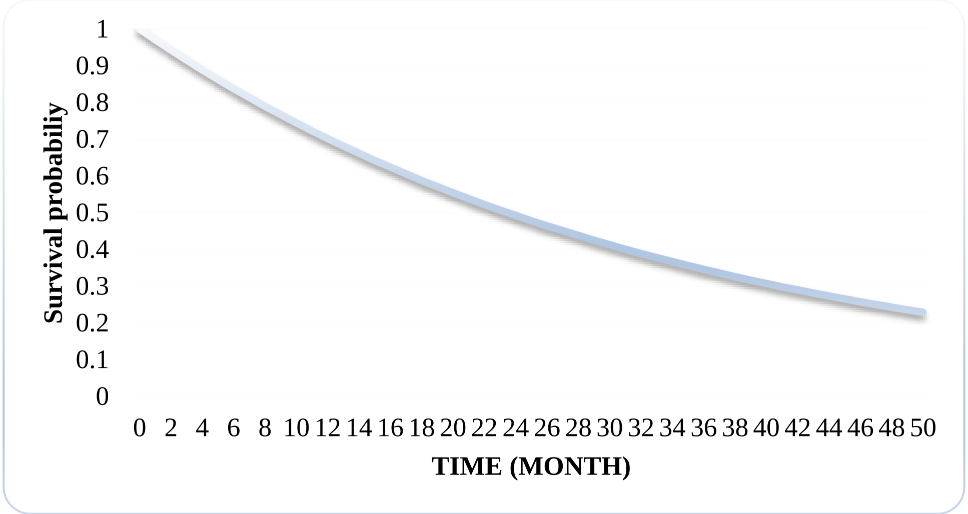

Figure 1. Survival Probability Graph

Based on Table 3 and Figure 1, we can see the survival chances of breast cancer patients at any given time in months, starting from 0 to 50 months. At the beginning of the observation (time 0 months), all patients were still alive, so the chance of survival was 1. However, as time went on, the chance of survival for patients decreased, which can be seen from the value of the chance of survival which was getting lower at later times. At 50 months, the patient's survival chance is only 0.228952, meaning that most patients with breast cancer die within 50 months after diagnosis. This shows that breast cancer is a serious disease and needs to be treated seriously and effectively to increase the patient's chances of survival. It is linear with the statement from World Health Organization that majority of breast cancer sufferers only survive for less than 50 months. After determining the survival function, the hazard function is calculated using Equation (4).

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= \frac{\frac{1}{\theta}e^{-\frac{x}{\theta}}}{e^{\frac{-x}{\theta}}} \\ &= \frac{1}{\theta} \\ h(x) &= \frac{1}{33.9157} \approx 0.0295 \end{aligned} \tag{13}$$

The hazard function, which describes the failure rate at a certain time, can be obtained from the survival function using the formula $h(x) = \frac{f(x)}{S(x)} = \frac{1}{\theta}$. However, because the data is censored, the hazard function can only be estimated up to the last observed time, which in this case was 49 months. Therefore, the estimated hazard function can be represented as $h(x) = 0.0295$, for $0 \leq x \leq 49$.

4. CONCLUSIONS

Based on data on breast cancer patients at Surabaya Ontology Hospital and the analysis provided use Anderson Darling test, the exponential distribution was found to be the best fit for the data. The next step is to estimate the parameters of the exponential distribution using the maximum likelihood method. Data is censored for lifetimes greater than 50 months. The estimated parameter was found to be θ=33.9157, it means that the average of the patient lifetime is about 33.9157≈ 40 months. Using these estimated parameters, the survival function is then determined using the formula $S(x) = e^{-\frac{x}{\theta}}$. The calculated values of the survival function at various time points are presented in Table 3. The analysis results also show that the estimated hazard function for patient failure or death is 0.0295, this hazard function value at a specific time point indicates the instantaneous risk of patient failure or death for individuals who have survived up to that time. Even so, it should be noted that this research was conducted using time data in months. Therefore, further research is needed using more accurate time data in order to provide more representative and reliable results. Even so, the results of this study can make an important contribution in increasing understanding of the risk of death or failure in patients with breast cancer, so that it can help doctors provide better and more effective care. In this case, information about the estimation of the hazard function can assist doctors and health professionals in planning appropriate and optimal treatment strategies for patients with breast cancer. Thus, the results of this study have important implications in improving the quality of life of patients with breast cancer.

REFERENCES

[1] Kementerian Kesehatan Republik Indonesia, "Kanker Payudara Paling Banyak di Indonesia, Kemenkes Targetkan Pemerataan Layanan Kesehatan," Kementerian Kesehatan Republik Indonesia, May 2021. [Online]. Available: https://www.kemkes.go.id/article/view/22020400002/kanker-payudara-paling-banyak-di-indonesia-kemenkes-targetkan-pemerataan-layanan-kesehatan.htm. [Accessed 15 April 2023].

[2] iTech Mission, "SDGs Global Dashboard," iTech Mission, 2021. [Online]. Available: https://www.sdgsdashboard.org/. [Accessed 19 June 2023].

[3] T. G. Ieren and P. E. Oguntunde, "A Comparison between Maximum Likelihood and Bayesian Estimation Methods for a Shape Parameter of the Weibull-Exponential Distribution," *AJPAS*, vol. 1, no. 1, pp. 1-12, May 2018.

[4] A. Wiranto, A. Kurniawan, D. A. Fitria and N. Chamidah, "Estimation of type I censored exponential distribution parameters using objective bayesian and bootstrap methods (case study of chronic kidney failure patients)," *Journal of Physics: Conference Series,* vol. 1397, no. 1, p. 012060, December 2019.

[5] W. Chen, Y. Zhang and J. He, "Estimating the survival function for exponential distribution with type-I censoring," *Journal of Statistical Computation and Simulation,* vol. 85, no. 3, pp. 563-574, 2015.

[6] Z. Yu, W. Tang and J. Huang, "Bayesian Survival Analysis For Type-I Censored Data with Exponential Distribution," *Journal of Statistical Computation and Simulation,* vol. 87, no. 9, pp. 1765-1779, 2017.

[7] A. Budiantoro, "Kajian Data Ketahanan Hidup Tersensor Tipe-I Dengan Distribusi Eksponensia," Neliti, 2016. [Online]. Available: https://media.neliti.com/media/publications/136888-ID-kajian-data-ketahanan-hidup-tersensor-ti.pdf [Accessed: April 15, 2023].. [Accessed 15 April 2023].

[8] D. Collett, Modelling survival data in medical research, Boca Raton: CRC press, 2023.

[9] A. J. Turkson, "Perspectives on Hazard Rate Functions: Concepts; Properties; Theories; Methods; Computations; and Application to Real-Life Data," *Open Access Library Journal,* vol. 9, no. 1, pp. 1-23, 2022.

[10] A. Kurniawan, "Analisis Data Uji Hidup," Surabaya, Universitas Airlangga, 2023.

[11] A. J. Turkson, F. Ayiah-Mensah and V. Nimoh, "Handling censoring and censored data in survival analysis: a standalone systematic literature review," *International journal of mathematics and mathematical sciences,* no. 2021, pp. 1-16, 2021.

[12] H. Panahi, "Estimation of the Burr type III distribution with application in unified hybrid censored sample of fracture toughness," *Journal of applied Statistics,* vol. 44, no. 14, pp. 2575-2592, 2017.

[13] S. Su, "Flexible modelling of survival curves for censored data," *Journal of Statistical Distributions and Applications,* vol. 3, pp. 1-20, 2016.

[14] N. H, Estimator Bayes Distribusi Eksponensial Terpotong Kiri pada Data Uji Hidup Tersensor Tipe I Berdasarkan Prior Non-Informatif, Skripsi, Universitas Airlangga, 2009.

[15] L. Jäntschi and S. D. Bolboacă, "Computation of probability associated with Anderson–Darling statistic," *Mathematics,* vol. 6, no. 6, p. 88, 2018.