Available online at http://scik.org

J. Math. Comput. Sci. 2022, 12:106

https://doi.org/10.28919/jmcs/7085

ISSN: 1927-5307

# MODELING THE AGE DISTRIBUTION OF BREAST CANCER PATIENTS OF NORTH-EAST INDIA

SWAPAN BHATTACHARJEE*, SUROBHI DEKA

Department of Statistics, University of Cotton, Panbazar, Guwahati-781001, India



**Abstract:** In this study four types of probability models are used, in order to find the best fitted model, namely Exponential, Gamma, Lognormal and Weibull. Goodness of fit measures are compared for each distribution using R programming. It is found that Gamma distribution provides the best fitted model.

**Keywords:** breast cancer; probability models; AIC; K-S statistic; AD statistic.

**2010 AMS Subject Classification:** 46N30, 62H10.

## 1. INTRODUCTION

Breast cancer is the most common invasive cancer in females. Breast cancer is the second leading cause of death among women worldwide [5]. Subramanian et al. [8] showed that breast cancer accounts for 34% of all cancer cases among women in India. Breast cancer contributes to 10.4 % of the global burden [6]. Statistical distribution can provide knowledge on the probability behavior of the magnitude of age distribution data of breast cancer patients. Bhattacharjee and Deka [4] showed that the log-logistics model was the best fit among different parametric model

*Corresponding author

E-mail address: swapanbhatta73@gmail.com

Received December 14, 2021

of the survival analysis of breast cancer patients. Rajbongsh et al. [7] found that the gamma model was the best fit to the age distribution data of the breast cancer patients. Srividhya and Radhika [10] showed that the Weibull distribution, Log - Normal distribution, Log - Logistic distribution and Generalized Gamma distribution had given the approximate results of the cancer data when compared to Exponential and Gompertz distribution.

## 2. MATERIALS AND METHODS

This is a prospective cohort study on 313 breast cancer patients who admitted at State Cancer Institute Hospital, Guwahati Assam India, during 2016 to 2018 and followed until December 2019. Death reported was 37. Some patients are still alive and are lost due to follow up. Data were collected on demographic characteristics and clinical characteristics of the patients. Patients' data including age, place of residence, religion, marital status, stage, grade, different communities, districts, treatment taken were collected. This study is used to compare the performance of parametric model by using exponential, Weibull, gamma and lognormal distributions. In order to determine the best parametric model of the breast cancer patients , Akaike Informantion Criterion (AIC) , Kolmogorv- Smirnov statistic and Anderson Darling statistic are calculated and compared. Some statistical models have been selected to describe the best fit of age distributions of breast cancer patients

### 2.1 Exponential Distribution

Exponential distribution is the simple statistical distribution that has a one parameter. The probability density function of exponential distribution is

$$f(x) = \theta e^{-\theta x}$$

where $\theta$ is a scale parameter. Its cumulative density function (CDF) is given by

$$F(x) = 1 - \mathrm{e}^{-\theta \mathrm{x}}$$

### 2.2 Gamma Distribution

Gamma distribution is a continuous probability distribution that is widely used to model continuous data. The probability density function (pdf) is given by

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} e^{-\beta x} x^{\alpha-1}$$

where $\alpha$ is the shape parameter and $\beta$ is the scale parameter. The cumulative density function (cdf) is given by

$$F(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)}$$

**2.3 Lognormal Distributions**

Lognormal distribution is a statistical distribution of logarithmic values derived from related normal distributions [1]. The pdf of lognormal distribution is given by

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[\frac{-(\ln(x) - \mu)^2}{2\sigma^2}\right]$$

where $\mu$ is the location parameter and $\sigma$ is the shape parameter. The cdf of lognormal distribution is

$$F(x) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left[\frac{\ln(x) - \mu}{\sigma\sqrt{2}}\right]$$

Where erf(.) is a complementary error function.

**2.4 Weibull Distribution**

The Weibull distribution is one of the most popular distributions in the analysis of time to event data. The pdf of Weibull distribution is given by

$$f(x) = \frac{\alpha}{\beta}\left(\frac{\alpha}{\beta}\right)^{\alpha-1} \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

where $\alpha$ is a shape parameter and $\beta$ is a scale parameter. The cdf for Weibull distribution is given by

$$F(x) = 1 - \exp\left[-\left(\frac{x}{\beta}\right)^{\alpha}\right]$$

**2.5 Goodness- of- fit measurement**

The performance of each distributions will be evaluated by using three different goodness of fit measures known as Akaike ‘s Information Criterion (AIC), Kolmogorov – Smirnov Statistic (K-S statistic) and Anderson- Darling statistic. The AIC formula is given as follows

$$\text{AIC} = -2\log(\text{L}) + 2\text{k}$$

Where $k$ is the number of the parameter and $L$ is the likelihood function on each fitted model respectively [2]. The formula for K-S statistic and Anderson- Darling statistic (AD) [3] is given as follows

$$D_n = \sup|F_N(x_i) - F(x_i)|$$

The Anderson and Darling (1954) [9] test statistic was

$$AD = -\sum_{i=1}^{n} \frac{(2i-1)}{n}\left[\log F(Y_i) + \log\left(1 - F(Y_{n+1-i})\right)\right] - n,$$

where F is the CDF of the specified distribution and $Y_i$ is the ordered data and n is the number of observation.

## 3. RESULTS AND DISCUSSION

A total of 313 clinically diagnosed cases of the female breast cancer patients are found between the study periods 2016-2018. Table 1 reveals the descriptive statistics of breast cancer patients. The mean and median age of patients is 47.555 and 47. The minimum age of the breast cancer patients is 18 years and maximum age is 80 years. The skewness and kurtosis are unequal to zero, suggesting that the age distribution data do not follow a normal distribution. Table 2 provides the results of the estimated parameter of each model using Maximum Likelihood Estimation approach. On the basis of the estimated parameters, figure 1 illustrates the graphical representation of the PDF plot, CDF plot, P-P plot and Q-Q plot on each fitted statistical model to the age distribution of data.

Table 1. Descriptive statistics for the age distribution of breast cancer patients

| Variable | Mean | Median | Standard Deviation | Minimum Value | Maximum Value | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|
| Duration | 47.555 | 47 | 10.937 | 18 | 80 | 0.165 | 3.247 |

Table 2. Results of the estimated parameters for each model

| Estimated Parameters | | | | | | | |
|---|---|---|---|---|---|---|---|
| Variable | Exponential | Gamma | | Lognormal | | Weibull | |
| | θ | α | β | μ | σ | α | β |
| Duration | 0.021 | 17.92 | 0.37 | 3.83 | 0.24 | 4.68 | 51.86 |

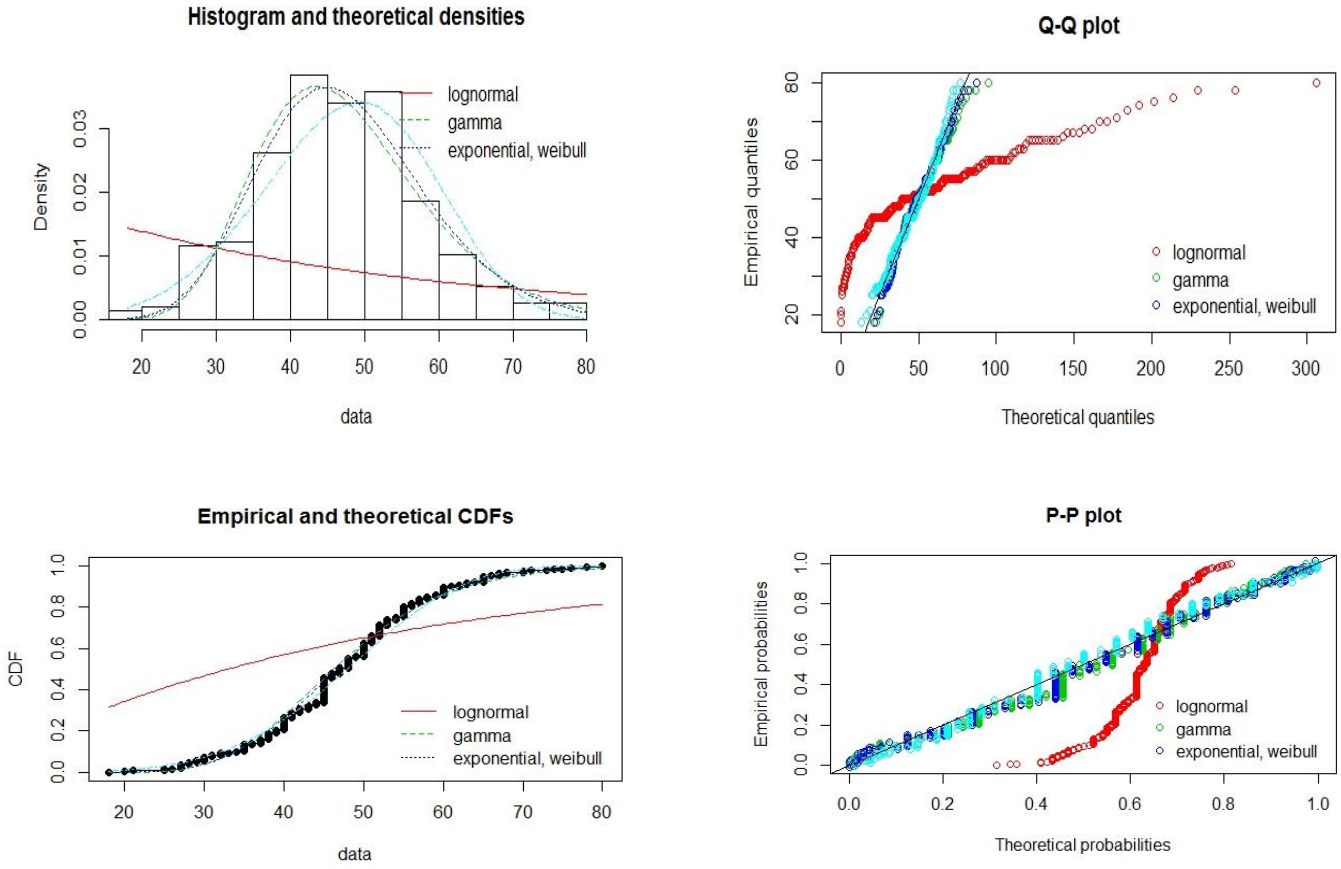


Figure 1. PDF, CDF, P-P plot and Q-Q plots on each fitted statistical model for the age distribution of breast cancer patients.

However on the basis of graphical representation it is difficult to decide which model is the best approximated. Therefore there are other measures such as Akaike's Information Criterion (AIC), Kolmogorov- Smirnov (K-S) statistic and Anderson Darling statistic. A low value of AIC indicates less information on a particular fitted model. A low value of K-S statistic indicates a high level of similarity between empirical CDF and the CDF of a fitted model and also a low value of AD statistic indicates the best fit. Table 3 presents the results of the goodness of fit for all fitted distribution models. As show in the table, it is found that from all the goodness of fit measures, gamma distribution is the best fitted model of age distribution of breast cancer patients.

Table 3. Results of the goodness of fit on each fitted model

| Fitted Distribution | AIC | K-S statistic | AD statistic |
|---|---|---|---|
| Exponential | 2407.469 | 0.1170051 | 2.5168499 |
| **Gamma** | **2394.525** | **0.1005963** | **1.4957212** |
| Lognormal | 3045.553 | 0.4187258 | 86.5687702 |
| Weibull | 2399.585 | 0.07569967 | 1.91923301 |

## 4. CONCLUSION

Based on the empirical data on age, this study investigates a probability model that can be used to represent the data on age distribution. The method of maximum likelihood is used as parameter estimation, where AIC, KS statistic and AD statistic are employed to evaluate the goodness of fit of each fitted model. In table 3, the results reveal that gamma distribution provides a best fitted model to the age distribution of breast cancer patients.

## ACKNOWLEDGEMENT

We are grateful to Dr. Kanakeswar Bhuyan, Medical Superintendent, State Cancer Institute, Guwahati, for providing us the Breast Cancer data.

## CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

## REFERENCES

[1] N. Masseran, M.A.M. Safari, S.I. Hussain, Modeling the distribution of duration time for unhealthy air pollution events, J. Phys: Conf. Ser. 1988 (2021), 012088.

[2] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Contr. 19 (1974), 716–723.

[3] R. B. D'Agostino and M. A. Stephens, Eds., Goodness-of-fit Techniques, ser. Statistics: Textbooks and Monographs. Marcel Dekker, Inc. New York, 1986, vol. 68.

[4] Swapan Bhattacharjee and Surobhi Deka, Application of parametric models to a survival analysis of breast cancer patients of North-east India, JP Journal of Biostatistics 18(2) (2021), 295-303.

[5] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control.

[6] http://www.dep.iarc.fr/globocan/globocan.htm.

[7] N. Rajbongshi, D.C. Nath, L.B Mahanta, Exploring age distribution pattern of female breast cancer patients in Assam, India using gamma probability model, J. Appl. Sci. 16 (2016), 496-503.

[8] L. Subramanian, V.U. Salini, H. Anandan and U. Insuvai, Breast cancer awareness in South India, Int. J. Sci. Stud. 6 (2018), 39-42.

[9] T. W. Anderson, D. A. Darling, A test of goodness of fit, J. Amer. Stat. Assoc. 49 (1954), 765-769.

[10] K. Srividhya, A. Radhika, Comparison of different parametric modeling for time-to-event data among cancer patients, Int. J. Sci. Res. Math. Stat. Sci. 6 (2019), 187-192.