International Journal of Applied Mathematics and Computing
Volume 2, Nomor 1, Year 2025
e-ISSN: 3047-146X; Page 59-74
DOI ; https://doi.org/10.62951/ijamc.v2i1.109
Available online at: https://international.arimsi.or.id/index.php/IJAMC

Extended Gamma Distribution to Fitting Breast Tumors

Adel Abbood Najm [1*], Bashar Khalid Ali [2]
[1] Department of Statistics, Administration and Economics College, Sumer University, Iraq
[2] Department of Statistics, Administration and Economics College, Kerbala University, Iraq
adel.abood@uos.edu.iq [1*], bashar.k@uokerbala.edu.iq [2]

Email correspondence: adel.abood@uos.edu.iq

Abstract*. There are many patterns of breast cancer change that make it a global health challenge. The research aims to propose an expanded gamma distribution with parameters and apply it to data for (103) patients with breast cancer. Data normality tests were used, such as the Kolmigrov-Smirnov test, the Anderson-Darling test, and the Chi-square test, to fit the real data. The parameters of the proposed distribution were estimated using the maximum likelihood method. It was found that there is a large difference in the real data with positive Skewness in it. The maximum likelihood estimates reflected the suitability of the data to the proposed distribution, which indicates the accuracy of the obtained results.*

Keywords: *Breast Cancer, Distribution, Normality, Skewness*

1. INTRODUCTION

Due to the high number of breast cancer cases, the world is facing a great disease force that may lead to major consequences because it is the most common type of cancer in the world that affects women to a large extent, which prompts researchers and analysts to study this type of tumor and understand the forms in which it appears. It is necessary to choose the distribution that suits this type of complex data. Many researchers have addressed the issue of the suitability of probability distributions for some diseases, including malignant tumors whose data are characterized by heterogeneity and asymmetry and which do not follow the normal distribution in many cases. This research aims to propose a new probability distribution that we called the extended Gamma distribution, which includes four parameters, as the transformation parameter was added to the Gamma distribution with three parameters to control the changing patterns in cancer data. Cancer is a group of diseases characterized by aggressive cells, which are characterized by unlimited cell growth and division, and the ability of these dividing cells to invade and destroy adjacent tissues or move to distant tissues in a process we call metastasis. These abilities are characteristics of a malignant tumor, unlike a benign tumor, which is characterized by limited growth and inability to invade and does not have the ability to move or metastasize. A benign tumor can also develop into a malignant cancer in some cases. Scientific studies on the factors affecting cancer have revealed many negative effects of these factors that affect patients in general and lead to uncontrollable complications that lead to the death of the patient, When the research is focused on malignant breast tumors, some of the problems that come up are analyzing the incidence rates, the duration of survival or the prevalence of the disease across

Received: Desember 23, 2024; Revised: Januari 07, 2025; Accepted: Januari 29, 2025;
Published: Januari 30, 2025

the broad demographic base. Such data is always heterogeneous and demonstrate skewness or high variance, which continues to pose challenges for their modelling with standard statistical distributions. This data was managed with a four parameter Gamma distribution because it is relatively easy to use for positive skewed data. Nevertheless, the biggest challenge remains in determining the validity of this approach for analyzing malignant breast tumors and the additional challenge of how to estimate its four parameters because the data is sparse and varied. The problem of the research becomes the efficiency of the four parameter gamma distribution in the characterization and analysis of malignant breast tumors data vis-a-vis other distributions, putting stress on the parameter's interpretation as the basis for the analysis of the progression of the disease. By formulating objectives, the level of relevance of the four parameter Gamma distribution on the analysis of malignant breast tumors has been studiously investigated with particular focus emphasis on the degree of accuracy on it's prevalence. The goal of the research is to find the most important factors

2. FAILURE TIME

These distributions and their uses are very important in reliability engineering to facilitate decisions at the system’s design, maintenance and even risk management levels. Failure time, in the context of reliability, or survival time in survival theory, within the boundaries of probability distributions, refers to the time a system or component ceases to perform its intended function. It is important to comprehend the distribution of failure times to be able to predict reliability, schedule maintenance, and design improvements. These are essential in the maintenance and quality engineering discipline because it aids in examining and determining the impact that systems and components can withstand and still function reliably. Such times help to understand when and how components or systems fail and thus enable the take of corrective and preventive measure to enhance performance and efficiency. The use of these failure times and reliability terms can really change a lot in many sectors and fields. For instance, in the production sector, it helps in quality control by finding the flaws in production processes, and as a result, product quality is improved and losses are minimized. In aviation, unplanned downtime can be reduced and production efficiency improved through the use of failure time analysis for maintenance schedule planning. In the same spirit, the aircraft maintenance and repair strategies are based on failure data over a period that has previously been shown to enhance air operations safety and effectiveness. This approach is aimed at the determination of critical components whose reliability analysis over time can pinpoint areas where performance problems or accidents might occur. In the

automotive sector, the application of this strategy seeks to ensure greater design and manufacturing quality so that the reliability is enhanced and the failure rates reduced. It also seeks to ensure internal client satisfaction by introducing the new strategy of post-sale services that includes developing preventive maintenance programs and repairs based on failure time analyses that have been proven to minimize as well as control warranty costs. In the energy sector, this approach has been applied to the maintenance of power plants where reliability analysis is used to formulate the optimum maintenance schedules for components like turbines and generators in order ensure that effective and continuous operation is achieved. And maintenance of electrical networks where the analysis of service interruption and failure data reveals the network weaknesses that have to be improved upon in order to enhance network stability and reduce service interruptions. And the electronics industry in product design: using failure time data to improve the design of electronic devices to ensure higher reliability and reduce failure rates. And environmental testing in evaluating the impact of different environmental conditions on the reliability of electronic devices and improving their resistance to failure. And in healthcare through medical equipment management: ensuring continuous and reliable operation of vital medical equipment through preventive maintenance schedules based on reliability analysis. Improving patient safety by applying reliability concepts to identify and analyze risks associated with medical procedures and equipment to ensure patient safety. In communications in the field of infrastructure maintenance: improving the reliability of communications networks and determining appropriate maintenance and repair schedules to ensure continuous operation and reliable service. And managing data centers by analyzing failure data to improve the design and maintenance of data centers and ensure their high reliability. (Kalbf & L. Prentice, 2002, 1) (Hassanien, 2017, 21)

1. Failure Density Function

If T represents a positive random variable representing the time of failure (Failure time), then it has a function that measures the probability of failure or stopping of the vehicle or unit during the period ($t < T < t + \Delta t$) regardless of the value of the change in time (Δt) and can be expressed mathematically as follows: (Linde, 2016, 56)

$$f_T(t) = \lim_{\Delta t \to 0} \frac{p_r(t < T < t+\Delta t)}{\Delta t} \quad ; \ t \geq 0 \qquad (1)$$

This function has the following properties:

1- $\int_0^{\infty} f_T(t)dt = 1$

2- $0 \leq f_T(t) \leq 1$

3- A single-value functions for each failure time.

2. Failure Density Function

If T represents a positive random variable representing the time of failure, then it has a function that measures the probability of failure or stopping of the vehicle or unit from working during the period ($t < T < t + \Delta t$) regardless of the value of the change in time (Δt) and it can be expressed mathematically as follows: (Linde, 2016, 56)

$$f_T(t) = \lim_{\Delta t \to 0} \frac{p_r(t < T < t+\Delta t)}{\Delta t} \quad ; t \geq 0 \qquad (2)$$

This function has the following properties:

1- $\int_0^\infty f_T(t)dt = 1$

2- $0 \leq f_T(t) \leq 1$

3- Single-valued function for each failure time.

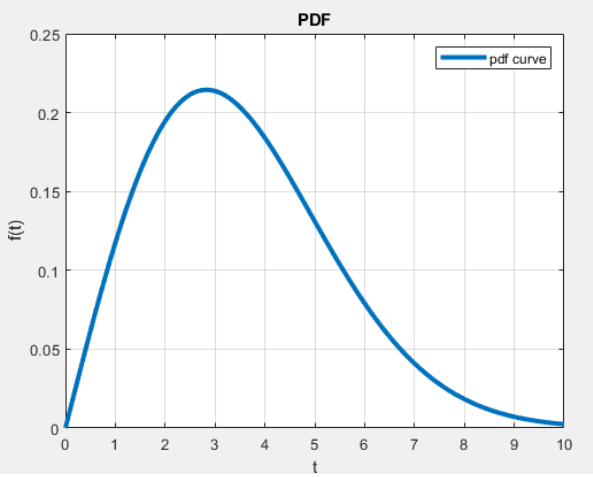

Figure 1 probability density function curve for failure

3. Failure Cumulative Function (FCF)

It is the probability of the vehicle failing or stopping working at the time of failure (t) and is expressed mathematically as follows:

$$F_T(t) = p_r(T \leq t) = \int_0^t f(u)du; \; t \geq 0 \qquad (3)$$

It is called the cumulative probability function (CDF) for failure up to time t. It is a function that does not decrease at any time of failure. (Linde, 2016, 57)

4. Failure Cumulative Function

It is the probability of the vehicle failing or stopping working at the failure time (t) and is expressed mathematically as follows:

$$F_T(t) = p_r(T \leq t) = \int_0^t f(u)du;\ t \geq 0 \quad (4)$$

It is called the cumulative probability function (CDF) of failure up to time t. It is a function that is non-decreasing at any time of failure. (Linde, 2016, 57)

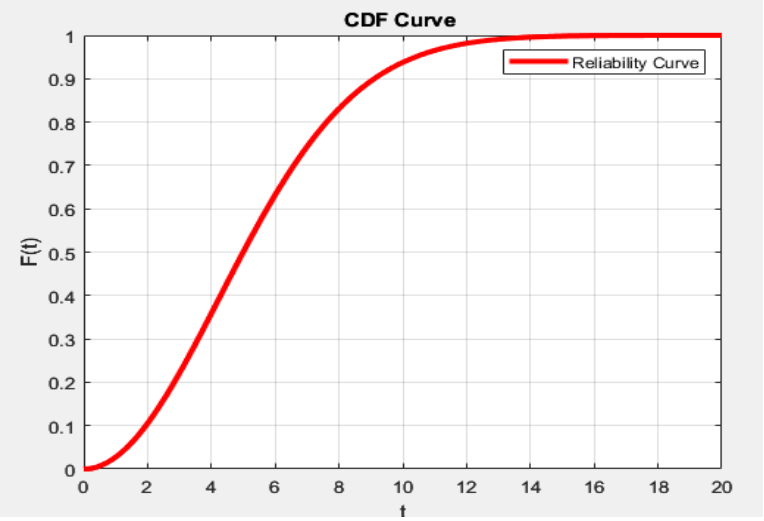

Figure 2 The aggregate density function curve for failure

5. Survival Function

The reliability function, also called the survival function (Survival Function) in the event that the studied phenomenon represents living organisms or the complementary cumulative distribution function (CDF), is one of the basic concepts that measure the probability of the survival of a vehicle, system, device, person or living organism after a certain period of time from its work and evaluating its performance without failure until a specific time. That is, it is the ability of the vehicle or system to perform its function without failure or it is the probability of failure-free performance during the life of the vehicle or system or the time frame of the specified element, under specific environmental operating conditions. (Di Lorenzo, 2008, 2). It is expressed mathematically as follows: (Ali & Neamah, 2022, 278)

$$R_T(t) = P(T > t) = 1 - P(T \leq t) = \int_t^{\infty} f_T(u)du = 1 - \int_0^t f_T(u)du \quad = 1 - F_T(t) = \bar{F}(t) \quad (3)$$

This function has the following properties: (AL-Nasser, 2009, 44)

The probability of the vehicle or system surviving during time (t=0) is equal to 1, i.e. R(0)=1, i.e. $\lim_{t \to 0} R(t) = 1$

The probability of the unit or system surviving during time (t= ∞) is equal to zero, i.e. R(∞)=0, i.e. $\lim_{t\to\infty} R(t) = 0$ meaning that the longer the time, the more susceptible the vehicle or system is to failure, i.e. its reliability decreases. (Neamah & Ali, 2020)

$0 \leq R(t) \leq 1$

$F(t)+R(t)=1$

R(t) is a continuously decreasing function from the left.

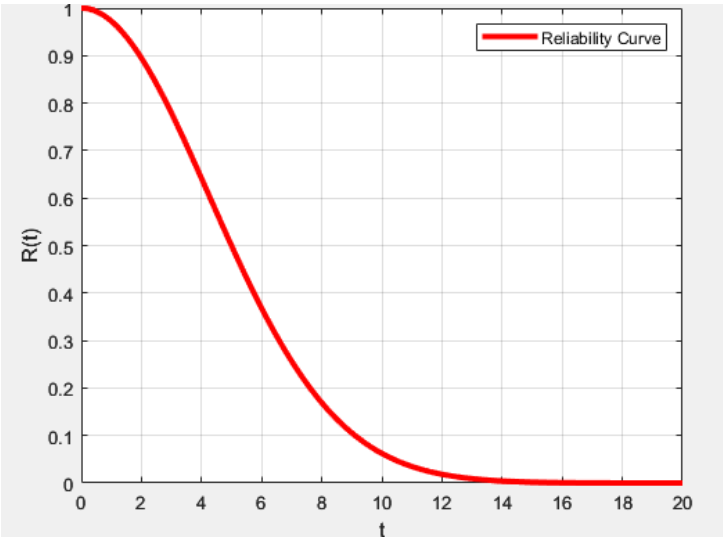

Figure 3 Survival curve

THE PROPOSED FOUR-PARAMETER GAMMA DISTRIBUTION

This distribution was proposed to analyze positive and asymmetric data, to be an effective tool in analyzing medical data that are often characterized by their diversity and Skewness. An extra shift or modification in the data is made with the inclusion of the shift parameter, which can be at the same time used with the μ parameter so that the whole distribution is moved. The flexibility of this distribution is very high as it can be placed and shaped in different ways to fit various medical data such as how long patients survive, how often a given disease occurs, or how responsive biological systems are. It is derived from the extended general gamma distribution by econometric modeling with the use of two additional parameters to control shape and location of the distribution after a fourth parameter, the shift parameter, is added. This distribution finds application in medical statistics and probability in the analysis of data about the survival periods and tumor associated medical data.

The probability density function of the distribution is as follows:

$$f(x;\alpha,\beta,\mu,\theta)=\begin{cases}\frac{\theta}{\Gamma(\alpha)\beta^{\alpha}}(x-\mu)^{\alpha-1}\ e^{-\beta(x-\mu)} & \text{if } x\leq\mu\\ 0 & if\ x>\mu\end{cases}$$

(4)

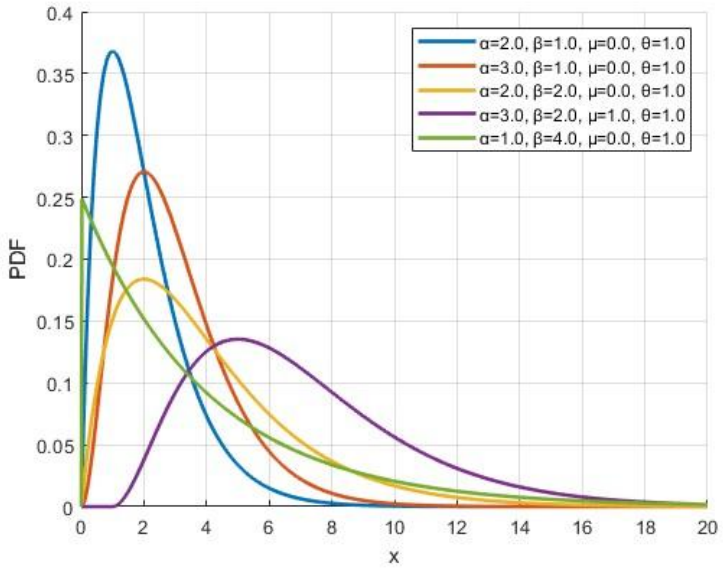

Figure 4 probability density function of the distribution at different values of its parameters

The cumulative probability density function is as follows:

$$F(x;\alpha,\beta,\mu,\theta)=\begin{cases}\frac{\theta.\gamma(.\alpha,\beta x-\mu)}{\Gamma(\alpha)} & \text{if } x>\mu\\ 0 & \text{if } x\leq\mu\end{cases}$$

(5)

α is the shape parameter and determines the shape of the distribution and its symmetry. If α>1, the distribution is skewed to the right, α>1, the distribution is skewed to the left, and if α=1, it turns into an exponential distribution. β is the scale parameter and controls how spread out the distribution is. The larger β, the wider the distribution becomes. μ is the location parameter and determines the starting point of the distribution and values below μ are ignored. θ is the shift parameter and is used to adjust for additional shift or adjustment in the data.

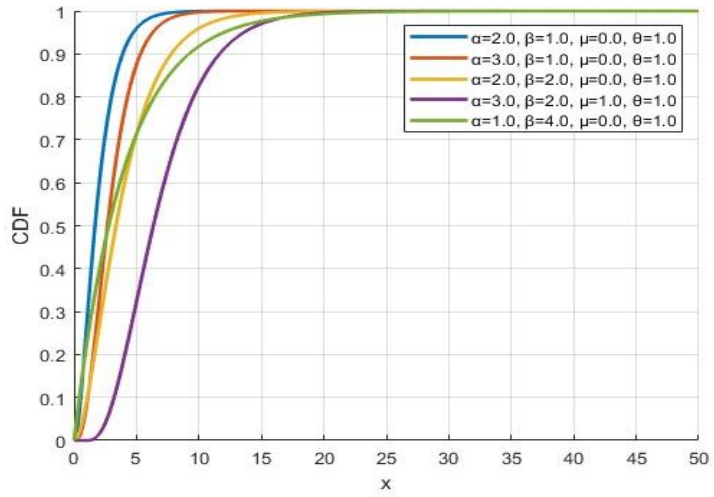

Figure 5 Cumulative probability density function of the distribution at different values of its parameters

The survival function of the distribution is as follows:

$$R(X) = 1 - \ \mathrm{F}(\mathrm{x};\alpha,\beta,\mu,\theta) = 1 - \frac{\theta\gamma(\alpha,\beta\mathrm{x}-\mu)}{\Gamma(\alpha)}$$

(6)

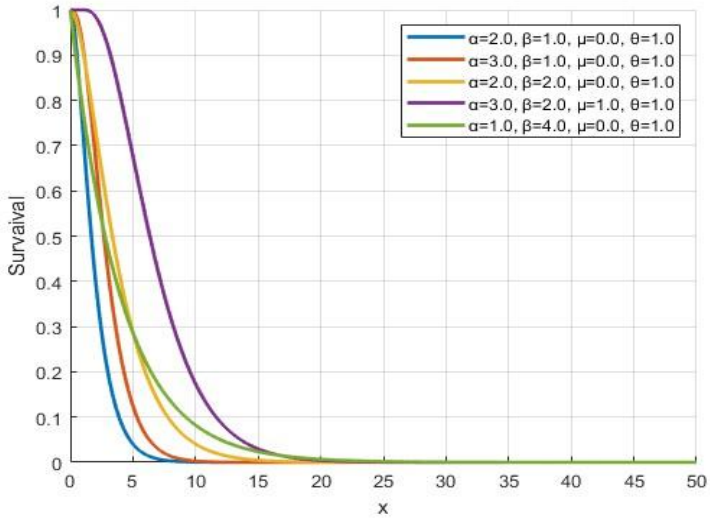

Figure (6) Survival function of the distribution at different values of its parameters

MAXIMUM LIKELIHOOD ESTIMATION OF DISTRIBUTION PARAMETERS

If we have a random sample from a gamma distribution with four $x_1, x_2, .., x_n$ then the likelihood function is:

$$L(x;\alpha,\beta,\mu,\theta) = \prod_{i=1}^{n} f(x;\alpha,\beta,\mu,\theta)$$

$$= \prod_{i=1}^{n} \frac{\theta}{\Gamma(\alpha)\beta^{\alpha}} (\mathrm{x}-\mu)^{\alpha-1} \ e^{-\beta(x-\mu)}$$

(7)

By taking the natural logarithm of the possibility function in Equation (3), we get:

$$l(x;\alpha,\beta,\mu,\theta) = \ln\big(L(x;\alpha,\beta,\mu,\theta)\big) = \ln\left(\prod_{i=1}^{n} f(x;\alpha,\beta,\mu,\theta)\right)$$

$$= nln(\theta) - nln((\mathrm{x}-\mu)^{\alpha-1}) - \ln\big(\Gamma(\alpha)\big) - \alpha\ln(\beta) + (\alpha-1)\sum_{i=1}^{n}\ln(\mathrm{x}-\mu) - \beta\sum_{i=1}^{n}(\mathrm{x}-\mu)$$

(8)

By taking the partial derivative with respect to each parameter, we obtain the maximum likelihood estimators for each parameter as follows:

$$\frac{\partial l}{\partial\theta} = \frac{n}{X_n-\mu} \quad (9)$$

Where X_n the maximum value of the data after arranging the data in ascending order (ordered statistics(

$$\frac{\partial l}{\partial\alpha} = \frac{n\Gamma'(\alpha)}{\Gamma(\alpha)} - nln(\beta) + \beta\sum_{i=1}^{n}(\mathrm{x}-\mu)$$

(10)

$$\frac{\partial l}{\partial\beta} = \frac{-n\alpha}{\beta} + \sum_{i=1}^{n}(\mathrm{x}-\mu)$$

(11)

$$\frac{\partial l}{\partial\mu} = -(\alpha-1)\sum_{i=1}^{n}\frac{1}{(\mathrm{x}-\mu)} + n\beta$$

(12)

The rates (12-9) are solved by numerical methods to find the maximum probability estimates.

5. The applied side:

The research sample consisted of (103) breast cancer patients of both sexes, aged between (29-80) years, which were taken from Al-Hussein (PBUH) Specialized Hospital, Cancer Oncology Department, Karbala Governorate, during the time period of 2023-2024. Table (1) shows the descriptive statistics of the research sample.

Table 1 Descriptive statistics of the research sample

Statistic	Value	Percentile	Value
Sample Size	103	Min	3
Range	897	5%	21.6
Mean	212.92	10%	28.8
Variance	35853.0	25% (Q1)	100
Std. Deviation	189.35	50% (Median)	120
Coef. of Variation	0.88929	75% (Q3)	280
Std. Error	18.657	90%	500
Skewness	1.3607	95%	500
Excess Kurtosis	1.7135	Max	900

The analysis of descriptive statistics in Table (1) shows a large variation among patients, with a range of 897 with values ranging from 3 to 900. The mean was 212.92, which is higher than the median of 120, indicating that there are extreme values that affect the distribution. The high standard deviation (189.35) and high variance (35853.0) reflect a wide spread in the data, while the coefficient of variation (0.88929) shows that the dispersion is relatively high relative to the mean. The percentile values represent the distribution of the data; 25% of the values are less than 100, 50% are less than 120, and 75% are less than 280, showing that the values tend to be concentrated in the lower range despite the presence of some high values. The positive Skewness (1.3607) and high Skewness (1.7135) show that the distribution is asymmetric and tends towards higher values. These statistics indicate a large diversity in the characteristics of the patients and the distribution of the values studied. This matches the characteristics of the proposed distribution. To test the data distribution, the (Kolmogorov-Smirnov test), the (Anderson

Darling test, and the (Squared) test were used. The results are in Table (2)

Table (2) Data Distribution Test

Distribution	Kolmogorov Smirnov	Anderson Darling	Chi-Squared
	Statistic	Statistic	Statistic
Four Parameter Gamma	0.1596	2.2277	49.396
MLE Estimates	$=2.2\hat{\alpha}$	$\hat{\beta}=1.2$	$=1.1\hat{\theta}$ $=0.0004\hat{\mu}$

The results of Table (2) for the data distribution test indicate an assessment of the extent to which the data fit the assumed distribution (Four Parameter Gamma) using the Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared tests. The Kolmogorov-Smirnov test showed a statistical value of 0.1596, indicating a relative difference between the observed and theoretical distribution. The Anderson-Darling test, which is sensitive to tails, gave a statistical value of 2.2277, which reflects a significant deviation between the actual data and the assumed distribution. On the other hand, the Chi-Squared test recorded a high statistical value of 49.396, indicating a significant difference between the observed and expected frequencies. In general, these results indicate that the data do not fully fit the Four Parameter Gamma distribution, which calls for considering other distributions or using non-parametric analysis methods to achieve a better representation of the data. Table (3) shows the details of the appropriate distribution

Table 3 Details of the appropriate distribution

Gamma (4P)					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	103 0.03819 0.13538 1				
Values	0.2	0.1	0.05	0.02	0.01
Critical Value	0.10573	0.12051	0.13381	0.14957	0.16051

Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	4 1.744 0.2538 2				
Value	0.2	0.1	0.05	0.02	0.01
Critical Value	0.9886	0.7794	0.4877	0.668	0.277
Reject?	No	No	No	No	No

The results of Table (3) of the fitting distribution details indicate a high agreement between the data and the Gamma (4P) distribution. In the Kolmogorov-Smirnov test, the statistical value was 0.03819 with a probability value (P-Value) of 0.13538, and the null hypothesis was not rejected at any of the listed significance levels (0.2, 0.1, 0.05, 0.02, and 0.01), which means that the data are consistent with the hypothesized distribution. For the Chi-Squared test, the statistical value was 1.744 with 4 degrees of freedom and a probability value of 0.2538, and the null hypothesis was not rejected at any of the listed significance levels as well. These results indicate that the Gamma (4P) distribution is a suitable distribution to represent the data under investigation.

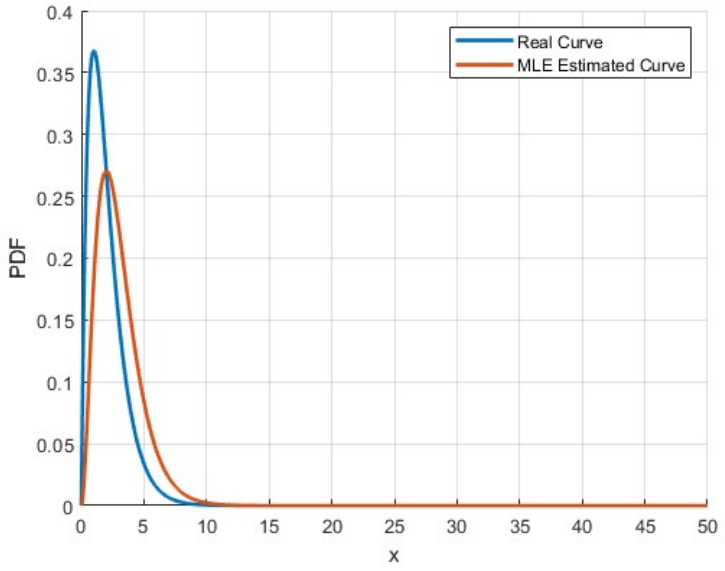

Figure 6 actual curve and the maximum likelihood method estimate of the probability density function of the distribution for real data.

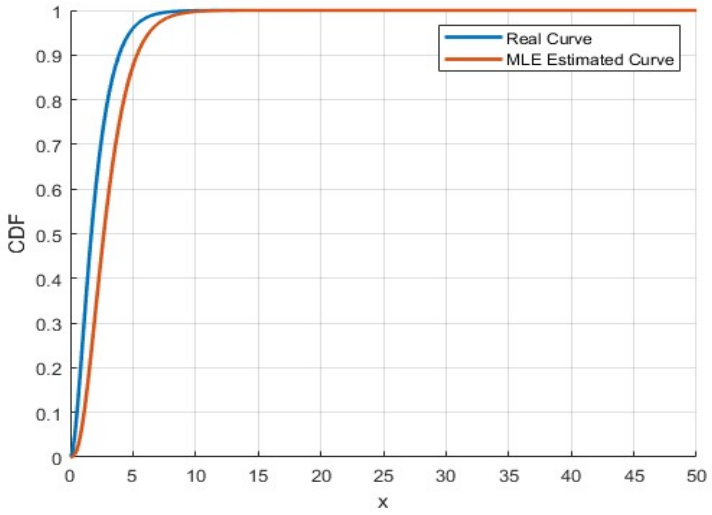

Figure 7 actual and estimated curve using the maximum likelihood method for the cumulative probability density function of the distribution of real data.

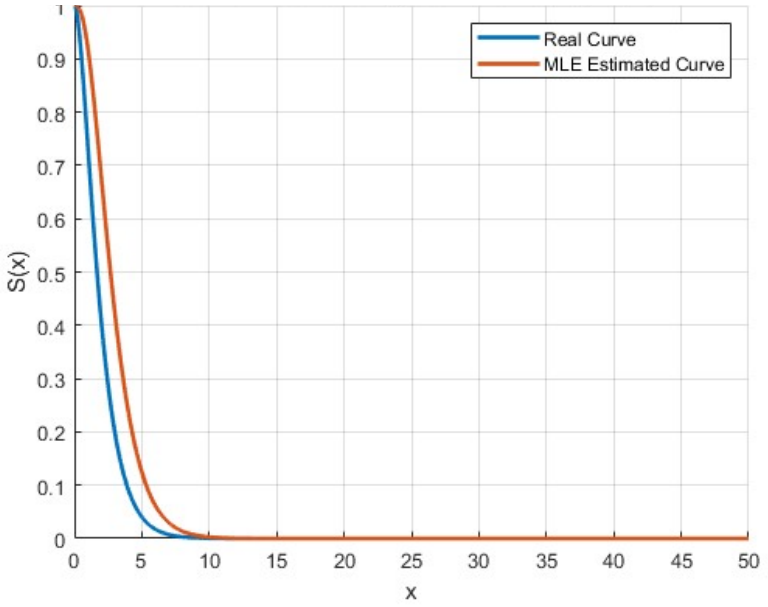

Figure 8 The true and estimated curve using the maximum likelihood method for the survival function of the distribution for real data.

Figure (6) shows the true and estimated curve of the probability density function (PDF): This graph shows a comparison between the true curve of the data (blue line) and the curve estimated using the maximum likelihood method (MLE) (orange line). A large convergence can be observed between the two curves, especially in the peak region, indicating that the maximum likelihood method provides an accurate estimate of the

probability density function. However, there are some slight differences at the ends, where the estimated curve is less accurate. This difference at the ends may be due to the presence of sparse data or the influence of extreme values. Figure (7) shows the true and estimated curve of the cumulative distribution function (CDF): This graph shows a comparison between the true curve of the cumulative distribution function (blue line) and the curve estimated using the maximum likelihood method (orange line). The two curves show excellent agreement across all values, indicating the accuracy of the estimation of the maximum likelihood method in representing the cumulative distribution function. The difference between the two curves is very small, which is an indication that the maximum likelihood method provides a robust estimate of the distribution. Figure (8) Comparison between the actual and estimated curves of the survival function using the maximum likelihood (MLE) method, where the two curves show great convergence at most values, indicating that the maximum likelihood method provides an accurate estimate of the survival function. The two curves match well at low values and across the right tail of high values, with very slight differences that do not affect the accuracy of the estimate. These results reflect the ability of the maximum likelihood method to represent real data robustly and reliably, making it an effective tool for analyzing the survival function of the distribution studied.

3. DISCUSSION OF THE RESEARCH RESULTS:

The research results represent a comprehensive statistical analysis of a sample of patients using the Gamma (4P) distribution and testing its suitability to the real data. The descriptive statistics in Table (1) indicate a large variation between the data values, showing a wide spread and a skew towards higher values due to extreme values. The high standard deviation and variance reflect high levels of dispersion, while the percentage values indicate that most of the data is concentrated in low values, which is consistent with a positively skewed distribution. Data distribution tests in Table (2) showed that the assumed distribution (Four Parameter Gamma) did not fit the data based on the high statistical values of the Kolmogorov-Smirnov, Anderson-Darling, and Chi-Squared tests, indicating a significant deviation between the actual distribution and the proposed distribution. These results prompted a re-evaluation of the distribution using the details of the fitted distribution in Table (3), which showed the results of the Kolmogorov-Smirnov and Chi-Squared tests a strong fit between the real data and the Gamma (4P) distribution based on the low statistical values and high probability values, which means that the Gamma (4P) distribution may fit the data when re-estimating using the maximum likelihood (MLE) method. The

attached graphs (Figures 3, 4, and 5) reinforced these results; Figure (3) showed a clear convergence between the real and estimated curve of the probability density function (PDF) with some slight differences at the ends due to extreme values. Figure (4) showed an excellent fit to the cumulative distribution function (CDF), reflecting the accuracy of the estimation using the maximum likelihood method. Figure (5) confirmed the suitability of the distribution to the survival function, as it showed a strong convergence between the true and estimated curves, which enhances confidence in the maximum likelihood method to represent the data.

4. CONCLUSIONS

a. Descriptive statistics showed a wide variation in data values, reflecting the presence of significant differences between the characteristics of the patient sample. This indicates the diversity of the studied cases, which is common in medical studies.
b. The maximum likelihood (MLE) method is effective in estimating the Gamma (4P) distribution to represent the real data, especially after re-estimation. Although there are some deviations at the ends.
c. The method has proven its ability to accurately represent the probability density functions, cumulative distribution, and survival function. This method can be used in future analysis of medical and statistical data, taking into account improving the distribution model to reduce the influence of extreme values.
d. Some graphs showed slight differences between the real and estimated curve at the ends, indicating the influence of rare or extreme values. These differences can be reduced by using improved models or processing extreme values.
e. The results indicate that statistical analysis using customized distributions such as Gamma (4P) is useful for understanding the distribution of medical characteristics of patients. This approach can be expanded to include larger samples and different population groups.
f. The Kolmogorov-Smirnov and Chi-Squared tests showed an excellent fit to the Gamma (4P) distribution, which enhances the reliability of the results and confirms the importance of these tests in assessing the fit of the distribution.

REFERENCES

Ali, B. K., & Neamah, M. W. (2022). A new robust fuzzy informative standard Bayes estimator for exponential distribution. *Journal of Algebraic Statistics*, *13*(1), 431–442. https://publishoa.com

Al-Nasser, A. M. (2009). *Statistical reliability*. Ithraa Publishing and Distribution, University of Baghdad.

Di Lorenzo, R. A. (2008). *Reliability, maintainability, and availability for engineers*. Defense Acquisition University Mid-West Region.

Hassanien, A. E. (2017). *Handbook of research on machine learning innovations and trends*. IGI Global. https://doi.org/10.4018/978-1-5225-2229-4

Kalbfleisch, J. D., & Prentice, R. L. (2002). *The statistical analysis of failure time data* (2nd ed.). John Wiley & Sons.

Linde, W. (2016). *Probability theory: First course in probability theory and statistics*. De Gruyter. https://doi.org/10.1515/9783110466195

Neamah, M. W., & Ali, B. K. (2020). Fuzzy reliability estimation for Frechet distribution by using simulation. *Periodicals of Engineering and Natural Sciences*, *8*(2), 632–646.