Employing an Artificial Intelligence Algorithm to Obtain Survival Functrion Estimator for Breast Cancer Patients

Manal Mahmoud Rashid
Directorate of Education Baghdad Rusafa
Manal4254@gmail.com

Abstract

The interest in studying probability distributions by statistical researchers came as a result of the role played by this statistical method in employing data behavior and knowing its characteristics, as phenomena can be expressed by random variables, and each random variable is expressed by a probabilistic distribution that contains important information for this random variable. Composite, weighted, and mixed distributions appeared due to the statistician's need to use the most appropriate distribution for the data under study. Cancer has spread in Iraq in a large way, especially after 2003, due to environmental pollution as a result of wars and other environmental pollutants, as well as smoking, heredity, and other reasons, and the effects of this disease have been exacerbated Also, due to the presence of large numbers of families living below the poverty line, they cannot secure the therapeutic doses allocated for this disease on a regular basis, in addition to the lack of modern treatment techniques as well as methods of prevention. In this research, the problem of increasing the number of deaths due to breast cancer was addressed. This research included a review of the (Quasi Lindely) distribution for its flexibility and good compatibility with the data of the applied side of this research. (2, θ), and the subject of the research was dealt with by estimating the survival function using the method of Maximum likelihood , and to improve the capabilities of the distribution, a genetic algorithm was employed and applied to breast cancer patients. With the theoretical properties of this function being a diminishing monotonic function.

Keywords: Mixed distribution, Quasi distribution, Lindely distribution, distribution characteristics, Maximum likelihood method, genetic algorithm, breast cancer data, survival function.

توظيف خوارزمية الذكاء الاصطناعي للحصول على مقدرات دالة البقاء لمرضى سرطان الثدي

م.م. منال محمود رشيد
وزارة التربية /المديرية العامة لتربية بغداد الرصافة الثالثة
Manal4254@gmail.com

المستخلص

ان الاهتمام بدراسة التوزيعات الاحتمالية من قبل الباحثين الاحصائيين جاء نتيجة للدور الذي تؤديه هذه الوسيلة الاحصائية في توظيف سلوك البيانات ومعرفة خصائصها ، اذ ان الظواهر يمكن التعبير عنها بمتغيرات عشوائية وكل متغير عشوائي يعبر عنه بتوزيع احتمالي يحتوي على المعلومات المهمة لهذا المتغير العشوائي . وقد ظهرت التوزيعات المركبة والموزونة و المختلطة لحاجة الاحصائي لاستعمال التوزيع الاكثر ملاءمة للبيانات قيد البحث .ان مرض السرطان قد انتشر في العراق بصورة كبيرة وخاصة بعد عام (2003) وذلك بسبب التلوث البيئي نتيجة الحروب والملوثات البيئية الاخرى وكذلك

التدخين والوراثة واسباب اخرى ، وإن اثار هذا المرض قد تفاقمت ايضا بسبب وجود اعداد كبيرة من العوائل تعيش تحت خط الفقر لا تستطيع تأمين الجرعات العلاجية المخصصة لهذا الداء بشكل منتظم ، فضلا عن انعدام تقنيات العلاج الحديثة وكذلك طرائق الوقاية . في هذا البحث تم تناول مشكلة زيادة اعداد الوفيات بسبب الاصابة بمرض سرطان الثدي . تضمَّن هذا البحث استعراض لتوزيع (Quasi Lindely) لمرونته وحسن مطابقته لبيانات الجانب التطبيقي لهذا البحث ، وهو توزيع مختلط من توزيعين احدهما توزيع اسي بمعلمة (θ) والاخر توزيع كاما بالمعلمتين ($2,\theta$) ، وقد تم تناول موضوع البحث تقدير لدالة البقاء بأستعمال طريقة الامكان الاعظم ولتحسين مقدرات التوزيع تم توظيف خوارزمية الجينية و وتطبيقها على مرضى سرطان الثدي و اتضح ان قيم دالة البقاء هي متناقصة بازدياد وقت الاصابة بالنسبة لمجموعة مرضى سرطان الثدي قيد البحث وهذا يتطابق مع الخصائص النظرية لهذه الدالة كونها دالة رتيبة متناقصة.

الكلمات المفتاحية :_التوزيع المختلط ، توزيع Quasi، توزيع Lindely ، خصائص التوزيع ، طريقة الامكان الاعظم ، الخوارزمية الجينية ، بيانات سرطان الثدي ، دالة البقاء .

1. Introduction

The concept of survival function analysis is widely used in biostatistics and most other fields. Survival analysis is defined as the study of the distribution of time from the onset (for example, birth or the beginning of treatment) until the emergence of an event, such as death or recurrence of disease. The interest in studying probability distributions by statistical researchers came as a result of the role What this statistical method plays in describing the behavior of the data and knowing its characteristics, as phenomena can be expressed by random variables, and each random variable is expressed by a probability distribution that contains important information for this random variable. Composite, weighted, and mixed distributions appeared due to the statistician's need to use the most appropriate distribution for the data under study. If the data is homogeneous, it follows a certain probability distribution, but if the data is heterogeneous, it does not follow a single probability distribution, but each part of it follows a specific distribution. The distribution may be the same but with different criteria, or it may be different distributions, and in this case this distribution is called the mixed distribution. Take into account the heterogeneity of the community.

Cancer has spread widely in Iraq, especially after 2003, due to environmental pollution as a result of wars and other environmental pollutants, as well as smoking, heredity and other reasons. The effects of this disease have also been exacerbated by the presence of large numbers of families living below the poverty line who cannot secure therapeutic doses. Allocated to this disease on a regular basis, as well as the lack of modern treatment techniques as well as methods of prevention. In this research, the problem of increasing the number of deaths due to breast cancer was addressed.

This research aims to obtain the most efficient estimation of the survival function for patients with breast cancer by using the traditional estimation method, which is the method of greatest possibility, and to improve its

capabilities, the genetic algorithm was employed in the process of estimating the survival function.

1.1 Literature Review

Ghitany et al. (2008) studied the different characteristics of the Lindley distribution, where the study included finding the moments, the characteristic function, the failure rate function, and the entropy function, and using the methods of moments and the Maximum likelihood in estimating the distribution parameter, and they proved that This distribution is better than the exponential distribution in some applications. While Ghitany et al. (2009) reviewed the Truncated Poisson - Lindley Distribution, and they also calculated the statistical properties of this distribution and estimated the distribution parameter by the methods of Maximum likelihood and moments, and they concluded that The moment estimator is more efficient than the maximum potential estimator. Furthermore, Rama Shanker and A. Mishra (2013) presented the Quasi Lindely distribution and studied some of its properties related to finding moments and the risk function, and they touched on some traditional methods of estimation, namely the method of Maximum likelihood and the method of moments to estimate the two parameters of the distribution, and the researchers compared between This distribution and Lindely distribution through application on data for pigs infected with malignant tuberculosis bacillus and other data for survival times for black birds, and the comparison shows the preference and flexibility of the (Quasi Lindely) distribution. Also, Hiba Z. Muhammed and L.S. Diab (2014) presented the Quasi Lindley Geometric Distribution, which is a composite distribution of the geometric distribution and the Quasi Lindley distribution. The researchers studied a set of distribution properties such as moments, risk function, and survival function, as well as an estimation process Parameters using the methods of Maximum likelihood and least squares. Alwan (2016) conducted a study of the Quasi Lindely distribution. Using the most efficient methods in the process of estimating the risk function for a sample of children with this disease.

2. Materials and Methods

In this topic, the concept of (Q.L) distribution was addressed and its most important characteristics were reviewed, and then the survival function was found and estimated using the greatest possibility method. To improve the capabilities of this function, one of the artificial intelligence algorithms, the genetic algorithm, was used.

2.1 Quasi Lindely Distribution (Q.L)

This distribution was proposed by the two researchers Rama Shanker and A. Mishra (2013) and this distribution has two parameters, and it is a mixed distribution of two random variables, one of which follows an exponential

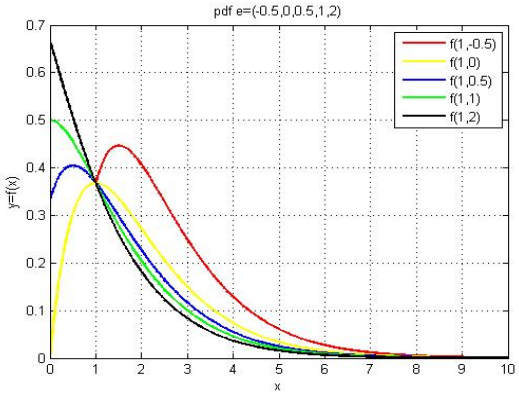
distribution with a parameter (θ) and the other is a Kama distribution with two parameters (2, θ). The probability density function will be For the variable whose distribution (Q.L) is as follows:

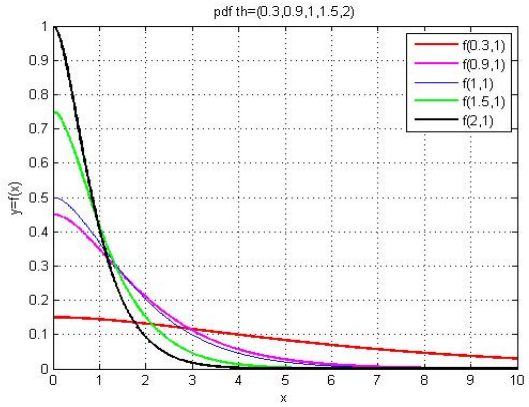
$$f(x,\alpha,\theta) = \frac{\theta(\alpha+x\theta)}{\alpha+1} e^{-\theta x} \quad ; x > 0\ , \theta > 0\ , \alpha > -1 \qquad (1)$$

Since:

θ: scale parameter

α: shape paramete





(1) **(2)**

Figure (1): shows a drawing of a function (pdf) for the (QL) distribution when the location parameter α takes values (-0.5, 0,0.5, 1,2) and the measurement parameter θ takes the value 1

Figure(2): shows a drawing of the (pdf) function of the (QL) distribution when the measurement parameter θ takes the values (0.3,0.9,1,1.5,2) and the parameter α takes the value 1

2.2 Cumulative distribution function for (Q.L) distribution

$$F(x) = \int_0^x f(u)du$$
$$F(x) = \frac{\theta(\alpha+u\theta)}{\alpha+1} e^{-\theta u}$$

$$F(x) = \int_0^x \left(\frac{\alpha\theta e^{-u\theta}}{\alpha+1} + \frac{u\theta^2 e^{-u\theta}}{\alpha+1}\right) du$$

$$F(x) = \frac{\alpha\theta}{\alpha+1}\int_0^x e^{-u\theta}du + \frac{\theta^2}{\alpha+1}\int_0^x ue^{-u\theta}du$$

And using integration by parts we get the cumulative distribution function

$$F(x) = 1 - \frac{(1+\alpha+\theta x)e^{-\theta x}}{\alpha+1} \quad ; x > 0, \theta > 0, \alpha > -1 \qquad (2)$$

This function has a positive skew

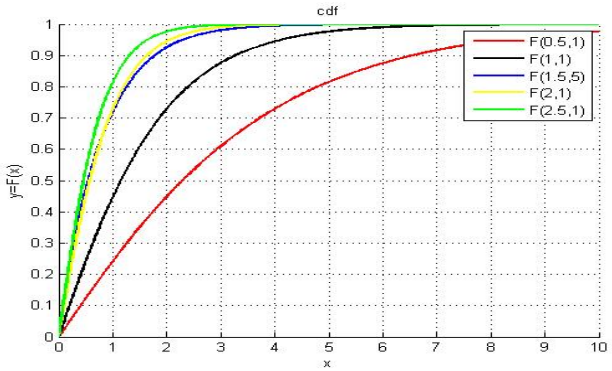


Figure (3): cdf function when the two parameters (θ, α) take different values.

2.3 Survival Function of (Q.L) Distribution:

The survival function is a decreasing function and its formulas are as follows

$$S(x) = 1 - F(x) = \frac{(1+\alpha+\theta x)e^{-\theta x}}{\alpha+1} \quad ; x > 0, \theta > 0, \alpha > -1 \qquad (3)$$

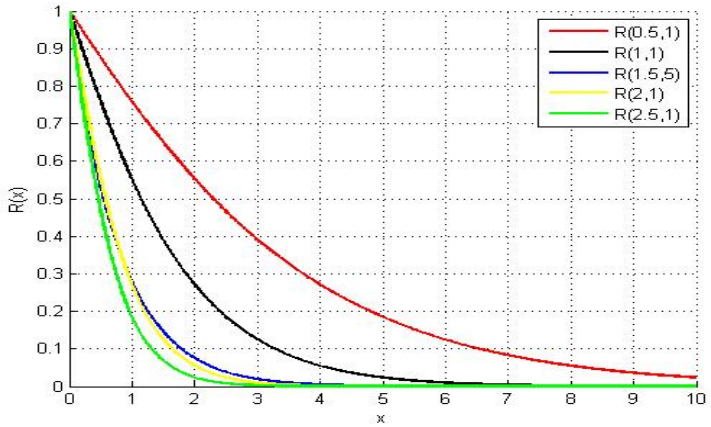


Figure (4): Survival function when the two parameters (θ, α) take different values.

2.4 Distribution Characteristics (Q.L)

2.4.1 Mode

$$\frac{df(x)}{dx} = \frac{\theta^2 e^{-\theta x}}{\alpha+1}(1-\alpha-\theta x)$$

After equating this equation to zero, the value of x will be as follows:

$$x = \frac{1-\alpha}{\theta}$$

If |α|<1, then the value of x is the value that brings f(x) to its maximum.
That is, the value of α is inversely proportional to the mode.
But if α ≥1, then f(x) is decreasing.

2.4.2 The moment about origin

$$E(x^r) = \int_0^\infty x^r f(x)dx$$

$$E(x^r) = \int_0^\infty x^r \frac{\theta(\alpha+x\theta)}{\alpha+1} e^{-\theta x}dx$$

$$E(x^r) = \frac{\theta\alpha}{\alpha+1}\int_0^\infty x^r e^{-\theta x}dx + \frac{\theta^2}{\alpha+1}\int_0^\infty x^{r+1}e^{-\theta x}dx$$

After performing the conversion, we get:

$$E(x^r) = \frac{(\alpha+r+1)\Gamma(r+1)}{(\alpha+1)\theta^r} \quad ; r = 1,2 \ldots. \qquad (4)$$

When we put r = 1, we get the first moment, which is the arithmetic mean:

$$E(x) = \mu'_1 = \frac{\alpha+2}{\theta(\alpha+1)}$$

When we put r = 2 we get the second moment:

$$E(x^2) = \mu'_2 = \frac{2(\alpha+3)}{\theta^2(\alpha+1)}$$

2-4-3-Variation Skewness and Kurtosis Coefficients.

i. The Coefficient of Variation:

$$C.V = \frac{\sigma}{\mu'_1} = \frac{\sqrt{\frac{\alpha^2+4\alpha+2}{\theta^2(\alpha+1)^2}}}{\frac{\alpha+2}{\theta(\alpha+1)}}$$

$$C.V = \frac{\sqrt{\alpha^2+4\alpha+2}}{\alpha+2} \qquad (5)$$

The value of the coefficient of variation increases when the value of α increases,

ii. The Coefficient of Skewness:

$$\sqrt{\beta_1} = \frac{2(\alpha^3+6\alpha^2+6\alpha+2)}{(\alpha^2+4\alpha+2)^{3/2}} \qquad (6)$$

The value of the torsion coefficient is positive when α≥-0.5 and negative otherwise.

iii.Kurtosis The Coefficient :

$$\beta_2 = \frac{\mu_4}{(\mu_2)^2} = \frac{\frac{3(3\alpha^4+24\alpha^3+44\alpha^2+32\alpha+8)}{\theta^4(\alpha+1)^4}}{\left(\frac{\alpha^2+4\alpha+2}{\theta^2(\alpha+1)^2}\right)^2}$$

$$\beta_2 = \frac{3(3\alpha^4+24\alpha^3+44\alpha^2+32\alpha+8)}{(\alpha^2+4\alpha+2)^2} \qquad (7)$$

The value of flattening is β_2>3 when the value of α≥-0.5, that is, the curve of the function becomes tapered, otherwise the curve of the function is flat.

2.5 Methods for estimating the survival function of the (Q.L) distribution

3.5.1 Maximum Likelihood Method

The method of Maximum likelihood depends on the idea of finding the estimator that makes the possibility function at its maximum end, and this method is characterized by the invariance property.

In order to estimate the survival function of the Q.L distribution according to this method, the two parameters of the distribution must first be estimated as follows:

$$L(x_1, x_2, \dots\dots x_n; \alpha, \theta) = \prod_{i=1}^{n} \quad f(x_i; \alpha, \theta)$$

$$L(\theta, \alpha) = \left(\frac{\theta}{\alpha+1}\right)^n \prod_{i=1}^{n} \quad (\alpha + x\theta)e^{-\theta \sum} \quad x_i$$

$$LnL(\theta, \alpha) = nLn\theta - nLn(\alpha + 1) + \sum_{i=1}^{n} \quad Ln(\alpha + x_i\theta) - \theta \sum \quad x_i$$

$$\frac{dLnL(\theta,\alpha)}{d\theta} = \frac{n}{\hat{\theta}} + \sum_{i=1}^{n} \quad \frac{x_i}{\hat{\alpha}+x_i\hat{\theta}} - \sum \quad x_i = 0 \qquad (8)$$

$$\frac{dLnL(\theta,\alpha)}{d\alpha} = \frac{-n}{\hat{\alpha}+1} + \sum_{i=1}^{n} \quad \frac{1}{\alpha + x_i\hat{\theta}} = 0 \qquad (9)$$

From equation (9) we find the value of α:

$$\frac{n}{\hat{\alpha}+1} = \sum_{i=1}^{n} \quad \frac{1}{\alpha+x_i\hat{\theta}}$$

$$n = (\hat{\alpha} + 1) \sum_{i=1}^{n} \quad \frac{1}{\alpha+x_i\hat{\theta}}$$

$$\hat{\alpha} + 1 = \frac{n}{\sum_{i=1}^{n} \quad \frac{1}{\alpha+x_i\hat{\theta}}}$$

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} \quad \frac{1}{\alpha+x_i\hat{\theta}}} - 1 \qquad (10)$$

Substitute Equation No. (10) into Equation No. (9):

$$\frac{n}{\hat{\theta}} + \sum_{i=1}^{n} \quad \frac{x_i}{\left[\frac{n}{\sum_{i=1}^{n} \quad \frac{1}{\alpha+x_i\hat{\theta}}} - 1\right] + x_i\hat{\theta}} - \sum \quad x_i = 0$$

After simplifying this equation, it results:

$$\frac{n}{\hat{\theta}} + \sum_{i=1}^{n} \quad \frac{x_i}{\left[\frac{n}{\sum_{i=1}^{n} \quad (\alpha+x_i\hat{\theta})^{-1}} - 1\right] + x_i\hat{\theta}} - \sum \quad x_i = 0 \qquad (11)$$

By solving the above equation by means of one of the numerical methods - such as Newton Raphson's method, we find the two greatest potential estimators (α,

θ), and by substituting these two estimates in the survival function, we get the Maximum Likelihood estimator for this function.

$$\hat{S}(x) = \frac{(1+\hat{\alpha}+\hat{\theta}x)e^{-\hat{\theta}x}}{\hat{\alpha}+1}$$

2.5.2 Genetic Algorithm Method

It is one of the methods of artificial intelligence, which can be defined to obtain the best solution to the issue under study in an effective and fast manner. Its idea came from Professor of Computer Science John Holland, which aims to modernize the concept of the natural evolution process and design industrial systems with similar characteristics to natural systems and his continuous ambition to improve The performance of the computational systems made the genetic algorithms more effective in solving optimization issues.

The following is a summary of the idea of a genetic algorithm:

i. Building the basic solution by means of the defined methods and the number of parameters. The method used is the Maximum Likelihood method (MLE). The number of parameters for the Q.L distribution is two parameters.
ii. Extracting the comparison scale (MSE) between the standard method (MLE) and the improved genetic method (MLE.AG), and then the best method is determined according to the comparison scale above.
iii. We make combinations between the above parameters according to the comparison scale.
iv. We do mixing or mating between the parameters of the used methods and calculating the comparison measures for these new models and comparing them with the best (MSE) that is determined.
v. A number of models are generated for each distribution that has been used in a specific direction towards the positive and negative for (m) times, and thus a large number of generated distributions will be generated. (MSE) is calculated for each distribution and compared with the parameters (optimal value) and upon completion of this work we will have a smaller (MSE) and the parameters by which this measure was calculated (the best parameters).

The following diagram represents the application of the stages of the genetic algorithm to the distribution of (Q.L) according to the method that we presented in the research

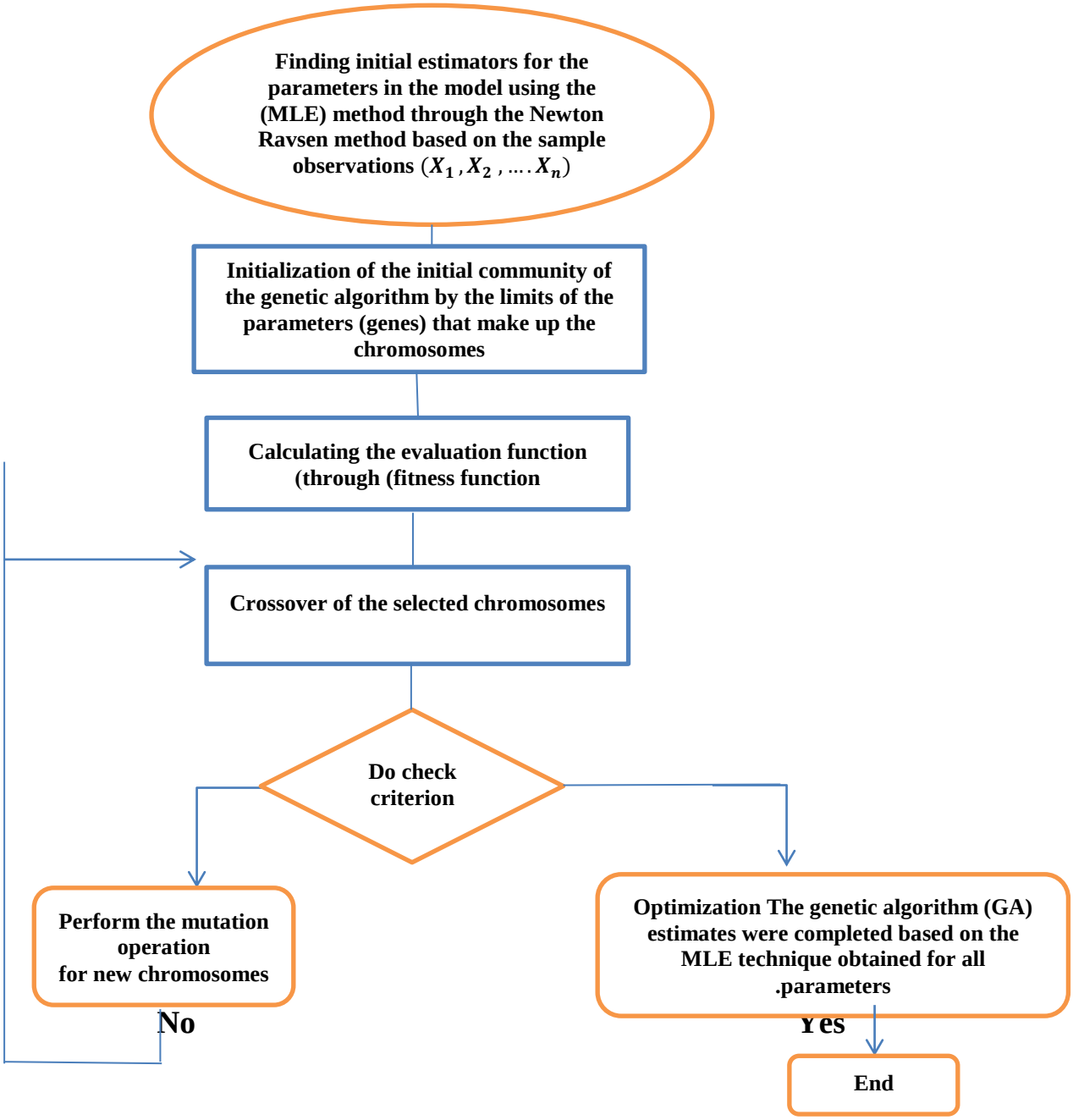


Diagram (1) :shows the flowchart of the genetic algorithm based on the MLE estimation technique.

3. Discussion of Results

In this research, real data will be used to estimate the survival function for the (Q.L) distribution. The Medical City Hospital in Baghdad affiliated to the Ministry of Health was selected to collect data related to the subject. Patients with breast cancer were selected. The process of collecting data for this disease was carried out with a sample size of 42 for the year 2022

3.1 Goodness of Fit Test

To find out that the data has a (Q.L) distribution, the (Kolmogorov-Smirnov) test and the Chi-squared test were used to test the good fit of the data. It was found that the data is distributed according to the (Q.L) distribution. The hypothesis that the data is distributed (FWMED) was accepted. :

H_0: the data are Q.L.

H_1: the data are not Q.L

Table (1): The results of the two tests of good conformity**.**

The test	I appreciate the two teachers		Calculated test statistic value	tabular value	The decision
	$\hat{\theta}$	$\hat{\alpha}$			
Kolmogorov-Smirnov	0.1195	0.4564	0.1321	0.2529	Accept H_0
Chi-Squared	0.2135	0.2137	0.7167	7.64	Accept H_0

3.2 Real data program results

In this paragraph, the results of the real data program and the analysis of the survival function estimation results are presented as in the following table.

Table (2): shows the values of the survival function estimators

(t_i)	$\hat{S}(t)_{MLE.G.E}$	(t_i)	$\hat{S}(t)_{MLE.G.E}$
9.2	0.217695	10.9	3.09E-02
4.5	0.107034	9.7	1.90E-02
7.4	8.89E-02	1.8	9.61E-03
2.9	8.89E-02	1.8	9.08E-03
10.5	8.89E-02	3.1	8.58E-03
10.5	8.89E-02	8.6	8.58E-03
9.6	8.89E-02	1.8	7.67E-03
3.6	8.89E-02	3.1	5.87E-03
7.9	8.89E-02	8.6	5.87E-03
6	6.82E-02	1.8	5.87E-03
3.4	4.17E-02	1.7	5.57E-03
3.7	3.86E-02	9	5.03E-03
0.7	3.58E-02	12.1	5.03E-03
4.1	3.58E-02	11	5.03E-03
10.1	3.58E-02	11	4.78E-03
5.9	3.58E-02	9.2	3.41E-03
5.8	3.09E-02	9.2	2.26E-03

5	3.09E-02	11.7	2.01E-03
8	3.09E-02	11.9	1.93E-03
11	3.09E-02	5.2	1.85E-03
6.8	3.09E-02	6.4	1.51E-03

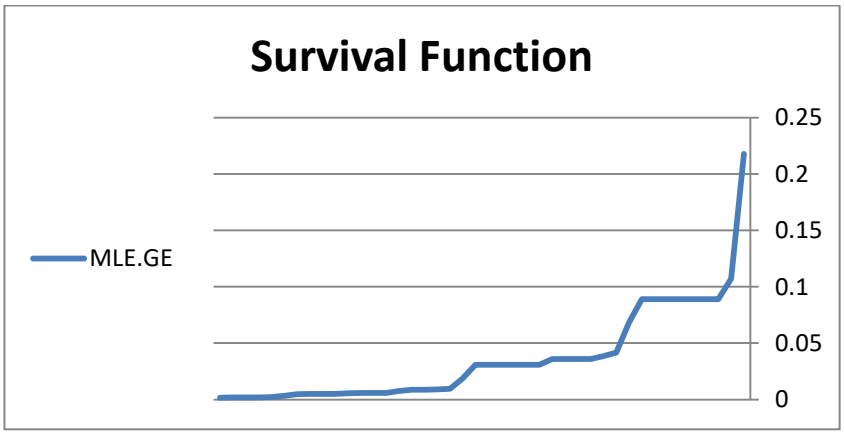


Figure (5): represents the behavior of survival function estimators using the genetic algorithm method

. Conclusion

From the conclusions of the applied side, it turned out that the values of the survival function are decreasing with increasing time of infection for the group of breast cancer patients under study, and this matches the theoretical characteristics of this function being a monotone decreasing function. Finally, we hope that the proposed model will attract more comprehensive applications in the fields of And different estimation methods, and the researcher recommends the use of mixed, weighted, and compound distributions in practical application in the health field and the industrial field, especially with regard to estimating the survival function and the risk function for living organisms, as well as the reliability function for serial and parallel systems, because these distributions are more flexible and accurate in describing and representing data than common distributions.

References

1- Saleh, Ahmed Alwan (2016) "Methods of estimating the risk function for the Quasi-lindely distribution, a comparative research with a practical application." A master's thesis in statistics submitted to the College of Administration and Economics at the University of Baghdad.

2- Rashid, Manal Mahmoud (2021) "Employment of the artificial intelligence algorithm in generalized beta commer estimations and its comparison with classical methods with a practical application." A master's thesis in

statistics submitted to the College of Administration and Economics at the University of Baghdad.

3- Al-Mashhadani, Mahmoud Hassan, Amir Hanna Hormuz, “Statistics,” (1989), The National Library, Baghdad.

4- Al-Sabawi, Ahmed Mahmoud, Khalil, Zaidoun Muhannad (2014), “A proposal for a hybrid algorithm by linking the genetic algorithm and the simulation algorithm annealing to solve quadratic allocation problems,” the Iraqi Journal of Statistical Sciences, pp. 117-136.

5- Diab L.S. and , Muhammed H. Z. , (2014) " Quasi Lindley Geometric Distribution " International Journal of Computer Applications (0975 – 8887) Volume 95– No. 13, June.

6- Shanker R., and Mishra A. (2013) " A quasi Lindley distribution" , African Journal of Mathematics and Computer Science Research, Vol. 6(4), pp. 64-71, April 2013.

7- Hooge , R .B. and craig , A . T. (1966) , "Introduction to mathematical statistics", 3rded, the Macmillan company , New York.

8- Mood, A.M., Graybill, F.A. & Bocs, D.C. (1985), "Introduction to The Theory of Statistics", McGraw-Hill, Inc

9- Demir , E. , Akkus ,O., (2015), " An Introductory Study on How the Genetic Algorithm Works in the Parameter Estimation of Binary Logit Model", IJS:BAR, pp.162-180.

10-Akkus , Demir , E. , (2016), " Comparison oF Som Classical And Meta-Heuristic Optimazation Techniques in The Estimation Of The Logit Model Parameters", IJAR, pp.1026-1042

11- Pasia, J., Hermosilla, A., & et al., (2005)," A useful tool for statistical estimation genetic algorilhm", JSCS,pp. 237 – 251.

12- Hadji, S. & et al. ,(2015)," Theoretical and experimental analysis of genetic algorithms based MPPT for PV systems", ELSEVIER, pp. 772-787.

13- Goldberg ,D. E.,(1989) ,"Genetic Algorithms in Search ,Optimization and Machine Learning", AWPC, Reading, MA.

14- Mitchell ,M., (1999) ," An Introduction to Gentic Algorithms ",London , England fifth printing: Abradford Book the MIT Press Cambridge Massachusetts.

15- Raghupathikumar, D., & Raja, K., (2012)," A Ggenetic Algorithm based Scheduling of an Input Queued Switch",IJCA,pp.37-42.

16-Ghitany M.E., Atieh B., Nadarajah S.,(2008)," Lindley distribution and its application",Mathematics and Computers in Simulation 78- 493–506

17- Ghitany , M.E. , Al-Mutairi , D.K. &Nadaraja , S. (2009) ."Zero-truncated Poisson-Lindley distribution and its application" , Mathematics and Computers in Simulation , Vol. 79 , PP.279-287.