Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

Tikrit Journal of Administrative and Economics Sciences
مجلة تكريت للعلوم الإدارية والاقتصادية
EISSN: 3006-9149
PISSN: 1813-1719

Estimating Gaussian Distribution Parameters Using Rank Regression and comparing them with the maximum likelihood estimators

Esraa Awni Haydier*
College of Administration and Economics/Salahaddin University, Erbil

Abstract: The representation of data through a Gaussian model allows it to be more flexible, effectively handle distortions, and avoid over-adaptation to the data. This approach helps improve the model's accuracy and understand how outliers impact the accuracy of the distribution parameters estimator **.**

In this research, a comparison was made between some methods for estimating Gaussian distribution parameters in analyzing the Survival Function using Rank Regression on the dependent variable and then on the independent variable, in addition to the maximum likelihood estimator method in the presence (and absence) of outliers in the distribution data. The mean square error criterion for the estimated parameters and the chi-square goodness-of-fit test were relied upon to compare the three methods through a simulation study for several parameter values and different sample sizes repeated (1000) times, in addition to real data representing the survival time of early breast cancer patients through a program in the MATLAB program designed for this purpose, in addition to the Easy-Fit program. The results of the study showed that the method of maximum likelihood estimators was superior in the absence of outliers in the Gaussian distribution data, while the method of estimating the rank regression on the independent variable was superior in the presence of outliers.

Keywords:
Gaussian distribution, Maximum Likelihood Estimation, Survival Function, Rank Regression, and outliers

A R T I C L E I N F O

Article history:
Received 29 Ouc. 2023
Accepted 21 Jan. 2025
Available online 31 Mar. 2025

*Corresponding author:
Esraa Awni Haydier
College of Administration and Economics/Salahaddin University, Erbil

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

تقدير معلمات توزيع كاوسيان باستخدام انحدار الرتبة ومقارنتها مع مقدرات الامكان الأعظم

إسراء عوني حيدر
كلية الإدارة والاقتصاد/ جامعة صلاح الدين، أربيل

المستخلص

يسمح تمثيل البيانات من خلال نموذج كاوسيان بأن يكون أكثر مرونة، ويتعامل بشكل فعال مع التشوهات، ويتجنب الإفراط في التكيف مع البيانات. يساعد هذا النهج على تحسين دقة النموذج وفهم كيفية تأثير القيم المتطرفة على دقة مقدّر معلمات التوزيع. تم في هذا البحث المقارنة بين بعض طرائق تقدير معلمات توزيع كاوسيان في تحليل دالة البقاء باستخدام انحدار الرتبة على المتغير التابع ومن ثم على المتغير المستقل فضلاً عن طريقة مقدرات الامكان الأعظم وذلك بوجود (وعدم وجود) القيم الشاذة في بيانات التوزيع. تم الاعتماد على معيار متوسط الخطأ التربيعي للمعلمات المقدرة واختبار جودة المطابقة كاي تربيع للمقارنة بين الطرائق الثلاث من خلال دراسة المحاكاة لعدة قيم معلمات واحجام عينات مختلفة مكرره (1000) مرة فضلاً عن بيانات حقيقية تمثل وقت البقاء لمرضى سرطان الثدي المبكر من خلال برنامج بلغة ماتلاب صمم لهذا الغرض فضلاً عن برنامج (Easy-Fit). توصلت نتائج الدراسة الى تفوق طريقة مقدرات الإمكان الأعظم في حالة عدم وجود قيم شاذة في بيانات توزيع كاوسيان في حين تفوقت طريقة تقدير انحدار الرتبة على المتغير المستقل في حالة وجود القيم الشاذة.

الكلمات المفتاحية: توزيع كاوسيان، تقدير الامكان الأعظم، دالة البقاء، انحدار الرتبة والقيم الشاذة.

1. Introduction

The Gaussian distribution is one of the most important statistical distributions in mathematics and statistics. The Gaussian distribution is characterized by a bell-shaped curve, where the majority of values cluster around a central mean, gradually decreasing towards lower and higher values. The survival function, often denoted as S(t) or survival probability, is a concept commonly used in statistics, probability theory, and survival analysis. It is a fundamental concept in understanding the probability of an event or entity surviving beyond a certain time or age (Ahn & Reinsel, 1988, p.852). Survival analysis is a statistical approach that is particularly useful in fields like epidemiology, medicine, biology, and engineering, where the focus is on the time until an event of interest occurs, such as the failure of a system, the onset of a disease, or the death of individuals in a population. The S(t) is defined as the probability that a random variable T (representing the time until the event of interest occurs) is greater than or equal to a specific time t.

The estimation process is considered one of the pillars of the statistical inference process, in addition to testing hypotheses. Through estimation, information and conclusions are collected about a parameter or parameters of the population based on the results extracted from the sample drawn from that population (Raza et al. 2018, p. 134). To obtain estimators with good characteristics, especially if there is more than one way to estimate a parameter, this leads to studying the comparison between these estimators to choose the best one, based on statistical criteria, the most important of which is the mean square error.

Outliers can have a significant impact on estimating the parameters of a Gaussian distribution (also known as a normal distribution). The presence of outliers can skew parameter estimates and lead to inaccurate results. Here's how outliers affect parameter estimation in a Gaussian distribution (Mustafa & Ali, 2013: 194), Maximum Likelihood Estimators (MLEs) are commonly used to estimate the parameters of a Gaussian distribution. The MLEs for the mean and variance are sensitive to outliers. When outliers are present, these estimators may be biased, and the parameter estimates may not accurately represent the underlying Gaussian distribution. Rank regression is a statistical technique that is used when you want to estimate a relationship between variables, but the assumptions of traditional linear regression may not be met. Instead of modelling the relationship between the variables in terms of their means (as in ordinary least squares regression), rank regression focuses on estimating the relationship based on the ranks of the observations. In the context of estimating a Gaussian (normal) distribution, rank regression can be useful when you have data that may not meet the assumptions of normality or when you suspect outliers in your data. Rank-based methods can be more robust in such cases. This introduces two methods for the parameter estimation of lifetime distributions. Rank Regression (RR) fits a straight line through transformed plotting positions and Maximum likelihood (ML) strives to maximize a function of the parameters given the sample data (Murali, 2016, p. 167). If the parameters are obtained, a cumulative distribution function (CDF) can be computed and added to a probability plot.

2. **Theoretical Aspect:** The theoretical aspect included the basic concepts of the Gaussian distribution and some methods for estimating its parameters and criteria for the efficiency of the estimated models.

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

2-1. Survival Function: The survival function is a function that gives the probability that a patient, device, or other object of interest will survive past a certain time. The survival function is also known as the survivor function or reliability function (Ali & Jwana, 2022, p.18).

Let the lifetime T be a continuous random variable with cumulative hazard function F(t) and hazard function f(t) on the interval [0,∞). Its survival function or reliability function is:

$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)du = 1 - F(t) \qquad (1)$$

2-2. Gaussian Distribution: Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. In graphical form, the normal distribution appears as a "bell curve", (Ali et al. 2022, p. 441). Normal distributions are important in statistics and are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. In the field of statistics, the normal distribution, also known as the Gaussian distribution, is a fundamental concept. It serves as a continuous probability distribution model for real-valued random variables (Hussein et al. 2023, p. 41). The probability density function that defines this distribution is expressed as follows:

$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} \qquad (2)$$

Where:

(μ): The average, or mean, of the normal distribution representing the time-to-failure, is commonly denoted as $(\overline{T})$.

(σ): Is the symbol representing the standard deviation for the times-to-failure.

It is a 2-parameter distribution with parameters (μ or $\overline{T}$) and σ (i.e., the mean and the standard deviation, respectively).

2-3. The Estimation: Estimation is the process by which the numerical value of unknown population values is inferred from incomplete data, such as a sample (Shahla et al. 2023, p. 140). Parameter estimation means using sample data (like times-to-failure or success data) to make educated guesses about distribution parameters. There are various methods available for parameter estimation (Esraa et al. 2023). There are several ways to estimate normal distribution parameters:

-Maximum Likelihood Estimation – MLE
- Method of Moments Estimation
- Kernel Density Estimation
- Graphical Methods
- Rank Regression on Y
- Rank Regression on X

2-3-1. Maximum Likelihood Estimation: For many distributions, maximum likelihood estimation (MLE) is a common and powerful method to estimate parameters. MLE aims to find parameter values that maximize the likelihood of observing the given data (Omar et al. 2020, p. 58). The equations for the partial derivatives of the log-likelihood function are derived and given next:

$$\frac{\partial \Lambda}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(t_i - \mu) = 0$$

And:

$$\frac{\partial \Lambda}{\partial \sigma} = \sum_{i=1}^{N}\left(\frac{t_i - \mu}{\sigma^3} - \frac{1}{\sigma}\right) = 0$$

2-3-2. Rank Regression on Y: Performing rank regression on Y requires that a straight line be fitted to a set of data points such that the sum of the squares of the vertical deviations from the points to the line is minimized (Kareem et al. 2020, p. 251). The following equations for regression on Y were derived**:**

$$\mathbf{\hat{a} = \bar{b} - \hat{b}\bar{x}}$$

$$\mathbf{= \frac{\sum_{i=1}^{N} y_i}{N} - \hat{b}\frac{\sum_{i=1}^{N} x_i}{N}}$$

And:

$$\mathbf{\hat{b} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} x_i^2 - \frac{(\sum_{i=1}^{N} x_i)^2}{N}}}$$

In the case of the normal distribution, the equation for y_i and x_i are (Chen et al., 2016: 3337):

$$\mathbf{y_i = \phi^{-1}[F(t_i)]} \qquad \mathbf{(3)}$$

And: $\mathbf{x_i = t_i}$, where the values $[\mathbf{F}(\mathbf{T_i})]$ are estimated from the median ranks, solve the above linear equation for the unknown value of y which corresponds to (George & Roger, 2020, p. 315):

$$\mathbf{x} = -\frac{\hat{\mathbf{a}}}{\hat{\mathbf{b}}} + \frac{\mathbf{1}}{\hat{\mathbf{b}}}\mathbf{y} \qquad (4)$$

Solving for the parameter, we get:

$$\hat{\boldsymbol{\mu}} = -\hat{\mathbf{a}}\,\hat{\boldsymbol{\sigma}} \qquad (5)$$

And:

$$\hat{\boldsymbol{\sigma}} = 1/\hat{\mathbf{b}} \qquad (6)$$

2-3-3. Rank Regression on X: Performing rank regression on X requires that a straight line be fitted to a set of data points such that the sum of the squares of the vertical deviations from the points to the line is minimized (Jasim et al. 2023. P. 99). The best–fitting straight line for the data, for regression on X, is the straight line:

$$\boldsymbol{x} = \hat{\boldsymbol{a}} - \hat{\boldsymbol{b}}\boldsymbol{y} \qquad (7)$$

The corresponding equations (David & Smith, 1972, p. 115), for $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are:

$$\hat{\boldsymbol{a}} = \bar{\boldsymbol{x}} - \hat{\boldsymbol{b}}\bar{\boldsymbol{y}} = \frac{\sum_{i=1}^{N} x_i}{N} - \hat{\boldsymbol{b}}\frac{\sum_{i=1}^{N} y_i}{N}$$

And:

$$\hat{\boldsymbol{b}} = \frac{\sum_{i=1}^{N} x_i y_i - \frac{\sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N}}{\sum_{i=1}^{N} y_i^2 - \frac{(\sum_{i=1}^{N} y_i)^2}{N}}$$

Where:

$$\mathbf{y_i} = \boldsymbol{\phi}^{-1}[\mathbf{F}(\mathbf{t_i})]$$

And: $\mathbf{x_i = t_i}$

Where the values $[\mathbf{F}(\mathbf{T_i})]$ are estimated from the median ranks, solve the above linear equation for the unknown value of y which corresponds to (Douglas, 2012: 86):

$$\mathbf{y} = -\frac{\hat{\mathbf{a}}}{\hat{\mathbf{b}}} + \frac{\mathbf{1}}{\hat{\mathbf{b}}}\mathbf{x} \qquad (8)$$

Solving for the parameter, we get:

$$\mathbf{a} = -\frac{\hat{\mathbf{a}}}{\hat{\mathbf{b}}} = -\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}} \Rightarrow \hat{\boldsymbol{\mu}} = \hat{\mathbf{a}} \qquad (9)$$

And:

$$\mathbf{b} = \frac{\mathbf{1}}{\hat{\mathbf{b}}} = \frac{\mathbf{1}}{\boldsymbol{\sigma}} \Rightarrow \hat{\boldsymbol{\sigma}} = \hat{\mathbf{b}} \qquad (\mathbf{10})$$

2-4. Outliers in the data: In statistics, an outlier is a data point that differs significantly from other observations. An outlier may be due to a variability in the measurement, an indication of novel data, or it may be the result of experimental error, the latter are sometimes excluded from the data set (Represent Scores that are unusually large or small relative to other scores). Outliers can seriously affect the integrity of data and result in biased or distorted sample statistics and faulty conclusions (Liu et al. 2012, 176). Several criteria have been suggested for identifying obvious and not-so-obvious outliers. According to one criterion, an outlier is any score that falls outside of the interval given by:

$$Md_n \pm 2(Q_3 - Q_1) \qquad (11)$$

Another criterion identifies an outlier as any score that falls outside of the interval (Chen et al. 2014, p. 313):

$$\overline{X} \pm 2.5\,S \qquad (12)$$

2-5. Mean Squared Error and Goodness of Fit: Tests of the three null hypotheses just described all use the chi-square sampling distribution. The chi-square distribution, like the t and F distributions, is a family of distributions whose shape depends on its degrees of freedom, n. The chi-square distribution like the F distribution is positively skewed, but as n increases (Ali et al. 2022, P. 394), the distribution approaches a normal distribution with mean and variance, respectively:

$$E(\chi^2_v) = v \quad \text{and} \quad Var\,(\chi^2_v) = 2v$$

Because χ^2_v is a squared quantity, it can range over only non-negative numbers, zero to positive infinity, whereas t and z can range over all real numbers. The goodness-of-fit test was developed to test the hypothesis that a population distribution estimated by a random sample is identical to a hypothesized or expected distribution. Let O_1, O2,..., O_k represent observed sample frequencies and E1, E2,..., E_k represent expected frequencies. The null hypothesis is rejected if Pearson’s statistic, exceeds or equals the critical value of chi square, $\chi_{\alpha,v}$, at a level of significance for $\upsilon = k - 1$ degrees of freedom (Anderson, 2011, P. 53).

$$\chi^2 = \sum_{j=1}^{k} \frac{\left(O_j - E_j\right)^2}{E_j} \qquad (13)$$

The mean squared error of an estimator of the parameter $\widehat{\Theta}$ is defined as:

$$MSE(\widehat{\Theta}) = E(\Theta - \theta)^2 \qquad (14)$$

3. Application Aspect: To compare the methods of estimating rank regression on the dependent variable (RRY) and then on the independent variable (RRX) with the maximum likelihood estimation method (MLE) when there are outliers in the Gaussian distribution data and in their absence, simulation was used in addition to the real data.

3-1. Simulation Study: Data having a Gaussian distribution with the Location (μ) and scale (σ) parameters (for several different values) and for different sample sizes (50, 100, 200, 500, and 1000) were generated using a program designed for this purpose in the MATLAB program.

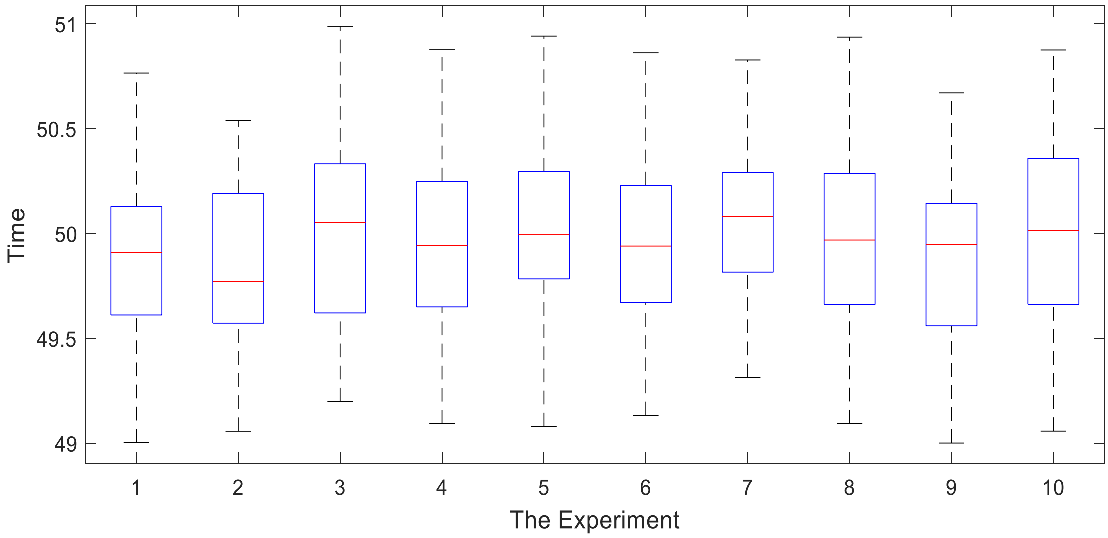

Figure (1): Box plot for the first ten simulation experiments

The simulation experiments for the first ten without outliers (n = 100) are shown in Figure 1, using the Box plot. The simulation experiments were repeated (1000) times, the parameters of the Gaussian distribution were estimated, and the MSE average (for MSE(Mu) and MSE(Sigma)). The results were summarized in Tables 1-3:

Table (1): Average of MSE when Mu = 50 & Sigma = 0.5

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	0.0543	0.0543	0.0543
	MSE(Sigma)	0.0407	0.0438	0.0412

Doi: www.doi.org/10.25130/tjaes.21.69.1.18

Sample Size	Criterion	MLE	RRY	RRX
100	MSE(Mu)	0.0381	0.0381	0.0381
	MSE(Sigma)	0.0292	0.0304	0.0294
200	MSE(Mu)	0.0275	0.0275	0.0275
	MSE(Sigma)	0.0209	0.0215	0.0211
500	MSE(Mu)	0.0178	0.0178	0.0178
	MSE(Sigma)	0.0136	0.0138	0.0136
1000	MSE(Mu)	0.0126	0.0126	0.0126
	MSE(Sigma)	0.0091	0.0091	0.0091

Table (2): Average of MSE when Mu = 10 & Sigma = 1

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	0.1085	0.1085	0.1085
	MSE(Sigma)	0.0814	0.0876	0.0823
100	MSE(Mu)	0.0762	0.0762	0.0762
	MSE(Sigma)	0.0583	0.0609	0.0588
200	MSE(Mu)	0.0550	0.0550	0.0550
	MSE(Sigma)	0.0419	0.0431	0.0423
500	MSE(Mu)	0.0355	0.0355	0.0355
	MSE(Sigma)	0.0271	0.0275	0.0273
1000	MSE(Mu)	0.0253	0.0253	0.0253
	MSE(Sigma)	0.0183	0.0183	0.0182

Table (3): Average of MSE when Mu = 100 & Sigma = 10

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	1.0850	1.0850	1.0850
	MSE(Sigma)	0.8135	0.8760	0.8234
100	MSE(Mu)	0.7621	0.7621	0.7621
	MSE(Sigma)	0.5832	0.6087	0.5879
200	MSE(Mu)	0.5501	0.5501	0.5501
	MSE(Sigma)	0.4188	0.4306	0.4226
500	MSE(Mu)	0.3553	0.3553	0.3553
	MSE(Sigma)	0.2710	0.2752	0.2726
1000	MSE(Mu)	0.2526	0.2526	0.2526
	MSE(Sigma)	0.1826	0.1829	0.1823

The three Tables (1-3) show that the three methods that estimated the Location parameter have the same estimated values and certainly had the

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

same criterion average (MSE), while the MLE outperformed the regression rank methods in all simulation cases except the case when the sample size was large (1000), the preference was for the rank regression method on the independent variable (RRX) and with note that the results of the three methods converge greatly when the sample size is large (1000). The accuracy of estimating Gaussian distribution parameters increases when the sample size increases. The accuracy of estimating the Gaussian distribution parameters decreases as its assumed value increases for all simulation cases. The estimators of the (RRY) method were inefficient in estimating the scale parameter compared to the (RRX) method for all simulation cases except for the case of large sample size (1000).

Also, outliers have been added to the generated data, using the (randperm (4,1)) function plus Mu to the data generated. The simulation experiments for the first ten with outliers (n = 100) are shown in Figure 2, using the Box plot.

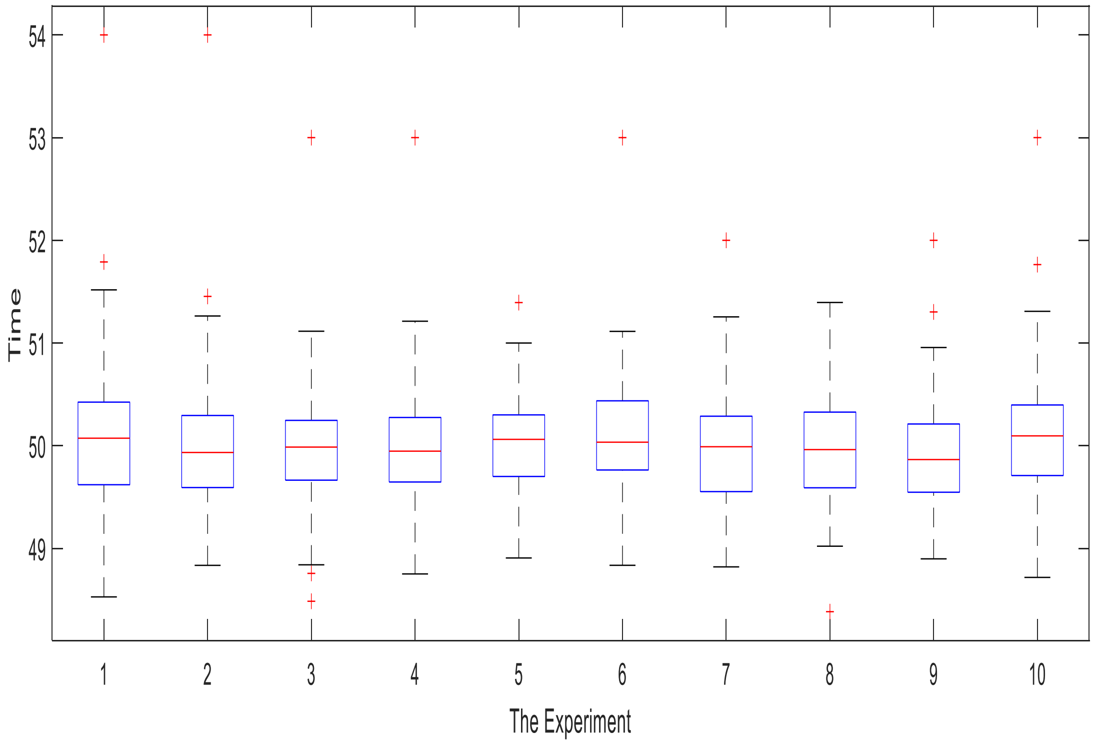

Figure (2): Box plot for the first ten simulation experiments with outliers

The simulation experiments were repeated (1000) times, the parameters of the Gaussian distribution were estimated, and the MSE average (for MSE(Mu) and MSE(Sigma)). The results were summarized in Tables 4-6:

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

Table (4): Average of MSE when Mu = 50 & Sigma = 0.5, (The presence of outliers)

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	0.0697	0.0697	0.0697
	MSE(Sigma)	0.1241	0.1966	0.0968
100	MSE(Mu)	0.0441	0.0441	0.0441
	MSE(Sigma)	0.0717	0.1077	0.0560
200	MSE(Mu)	0.0299	0.0299	0.0229
	MSE(Sigma)	0.0383	0.0542	0.0316
500	MSE(Mu)	0.0182	0.0182	0.0182
	MSE(Sigma)	0.0188	0.0245	0.0164
1000	MSE(Mu)	0.0125	0.0125	0.0125
	MSE(Sigma)	0.0114	0.0138	0.0106

Table (5): Average of MSE when Mu = 10 & Sigma = 1, (The presence of outliers)

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	0.1358	0.1358	0.1358
	MSE(Sigma)	0.2094	0.3251	0.1708
100	MSE(Mu)	0.0865	0.0865	0.0865
	MSE(Sigma)	0.1216	0.1785	0.0996
200	MSE(Mu)	0.0592	0.0592	0.0592
	MSE(Sigma)	0.0702	0.0978	0.0598
500	MSE(Mu)	0.0364	0.0364	0.0364
	MSE(Sigma)	0.0357	0.0452	0.0321
1000	MSE(Mu)	0.0253	0.0253	0.0253
	MSE(Sigma)	0.0218	0.0257	0.0206

Table (6): Average of MSE when Mu = 100 & Sigma = 10, (The presence of outliers)

Sample Size	Criterion	MLE	RRY	RRX
50	MSE(Mu)	1.6373	1.6373	1.6373
	MSE(Sigma)	3.9644	6.6530	2.5848
100	MSE(Mu)	0.9730	0.9730	0.9730
	MSE(Sigma)	2.2913	3.7191	1.4435
200	MSE(Mu)	0.6327	0.6327	0.6327
	MSE(Sigma)	1.2992	2.0317	0.8248

Doi: www.doi.org/10.25130/tjaes.21.69.1.18

Sample Size	Criterion	MLE	RRY	RRX
500	MSE(Mu)	0.3764	0.3764	0.3764
	MSE(Sigma)	0.6065	0.8984	0.4061
1000	MSE(Mu)	0.2567	0.2567	0.2567
	MSE(Sigma)	0.3324	0.4680	0.2430

The three Tables (4-6) show that the three methods that estimated the Location parameter have the same estimated values and certainly had the same criterion average (MSE), while the RRX outperformed the (RRY) and (MLE) methods in all simulation cases. The accuracy of estimating Gaussian distribution parameters increases when the sample size increases. The accuracy of estimating the Gaussian distribution parameters decreases as its assumed value increases for all simulation cases. The estimators of the (RRY) method were inefficient in estimating the scale parameter compared to the (MLE) method for all simulation cases when there are outliers.

3-2. The Real Data: In a breast cancer study, observed times in months for time to breast retraction of early breast cancer patients (a subset of the total data set). The real data is taken from (Iqbal et al. 2022, P. 148) shown in Table 7 to (22) patients.

Table (7): Breast cancer time data

5	7	4	4	8	12	7	5	7	6	7
7	7	14	12	8	7	9	8	16	7	7

The box plot of real data for the breast cancer study shows the presence of (5) outliers as shown in Figure 3.

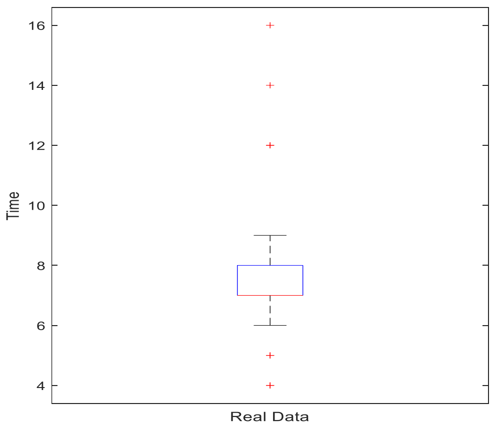

Figure (3): Box plot for the real data

After estimating the location and scale parameters of the Gaussian distribution for the three methods, they are used to estimate survival times (expected values) and then the goodness-of-fit test (Chi-Square) is used to test the efficiency of the estimated models. This test assesses the model's overall fit by comparing the observed data with the model's predicted values. The associated p-value obtained from this test measures the statistical significance of the differences between the observed and predicted values. Thus, by testing statistics, we can comprehensively evaluate the validity of the proposed model. The model that exhibits the best fit, as indicated by the minimum values of the Chi-Square and non-significant p-values from the test statistic, can be considered the most suitable for the given data set. Table 8 summarizes the estimation and testing results for the three methods, and shows that method (RRX) was better than methods (MLE, and RRY) because the value of the test statistic was equal to (7.8128), which is less than the critical value (9.2103) under significance level (0.01), and the degrees of freedom (2), (also, it less than test statistics (7.8893) and (8.7601) for MLE, and RRY, respectively), and this is confirmed by the p-value (0.020), which was not significant, indicating the efficiency of the RRX model. Figure A-C in the appendix shows the probability and cumulative density function with the Survival function of the Gaussian distribution for breast cancer using the three methods.

Table (8): Results of analysis for the real data

Method	Mean parameter	Sigma parameter	Chi-Square Statistic	p-value	Critical Value
MLE	7.9091	2.9834	7.8893	0.019	9.2103
RRY	7.9091	3.5044	8.7601	0.013	9.2103
RRX	7.9091	2.9287	7.8128	0.020	9.2103

4. Conclusion & Recommendations: Through the simulation study and real data, the following main conclusions and recommendations were summarized:

4-1. Conclusions:

1. The three methods (MLE, RRY, and RRX) that estimate the Location parameter have the same estimated values in the presence and absence of outliers for all simulation cases and real data.
2. The MLE outperformed the regression rank methods in the absence of outliers for all simulation cases except the case when the sample size was large (1000), the preference was for the RRX method.
3. The RRX outperformed the MLE and RRY methods in the presence of outliers for all simulation cases and real data.

4. The accuracy of estimating Gaussian distribution parameters increases when the sample size increases for the three methods.
5. The accuracy of estimating the Gaussian distribution parameters decreases as its assumed value increases for the three methods.
6. The estimators of the (RRY) method were inefficient in estimating the scale parameter compared to the (RRX) method for all simulation cases except for the case of large sample size (1000).

4-2. Recommendations

1. *Using the RRX method to estimate two-parameter Gaussian distribution* when there are outliers.
2. Using the MLE method to estimate two-parameter Gaussian distribution when there are no outliers.
3. Conducting a prospective study on the use of the robust rank regression method to estimate two-parameter Gaussian distribution.
4. Conducting a prospective study on the use of the rank regression method to estimate two-parameter Exponential and Weibull distribution.

References

1. Ahn, S. K. and Reinsel, G. C. “Nested reduced-rank autoregressive models for multiple time series.” Journal of the American Statistical Association, 83 (1988), 849–856.
2. Ali, Taha Hussein and Jwana Rostam Qadir. “Using Wavelet Shrinkage in the Cox Proportional Hazards Regression model (simulation study)”, Iraqi Journal of Statistical Sciences, 19, 1, 2022, 17-29.
3. Ali, Taha Hussein, Saman Hussein Mahmood, and Awat Sirdar Wahdi. "Using Proposed Hybrid method for neural networks and wavelet to estimate time series model." Tikrit Journal of Administration and Economics Sciences 18.57 part 3 (2022).
4. Ali, Taha Hussein. “Modification of the adaptive Nadaraya-Watson kernel method for nonparametric regression (simulation study).” Communications in Statistics-Simulation and Computation 51.2 (2022): 391-403.
5. Anderson, Theodore W. "Anderson-Darling Tests of Goodness-of-Fit." International Encyclopedia of Statistical Science, vol. 1, 2011, pp. 52-54.
6. Chen T, Kowalski J, Chen R, Wu P, Zhang H, Feng C, Tu XM. "Rank-preserving regression: a more robust rank regression model against outliers. " Stat Med. (2016) Aug 30;35(19):3333-3346. doi: 10.1002/sim.6930. Epub 2016 Mar 2. PMID: 26934999.
7. Chen T., Tang W., Lu Y., and Tu X. M. Rank regression: an alternative regression approach for data with outliers. Shanghai Archives of Psychiatry 26.5 (2014): 310-315.
8. David, J. Smith "Reliability Engineering" Pitman publishing, (1972).
9. Douglas C. Montgomery "Introduction to Statistical Quality Control" Seventh Edition, John Wiley & Sons, Inc, (2012).

10. George, C., Roger L. Berger "Statistical Inference" second edition, United State of America (2020).
11. Hussein, D. N., S. H. D. AL-Zakar, and A. M. Yonis. “Estimating the Intensity Equations for Rain Intensity Frequency Curves (Mosul /Iraq): Intensity Equations for Rain”. Tikrit Journal of Engineering Sciences, vol. 30, no. 3, Sept. 2023, pp. 38-48, doi:10.25130/tjes.30.3.5.
12. Iqbal, M.S.; Ahmad, W.; Alizadehsani, R.; Hussain, S.; Rehman, R. “Breast Cancer Dataset, Classification and Detection Using Deep Learning." Healthcare (2022): 10, 2395. https://doi.org/10.3390/healthcare10122395
13. Jasim, N. A., Ibrahim, A. A., & Hatem, W. A. (2023). Forecasting the Performance Measurement for Iraqi Oil Projects using Multiple Linear Regression. Tikrit Journal of Engineering Sciences, 30(2), 94–102. https://doi.org/10.25130/tjes.30.2.10
14. Kareem, Nazeera Sedeek and Mohammad, Awaz Shahab, and Ali, Taha Hussein, “Construction robust simple linear regression profile Monitoring” Journal of Kirkuk University for Administrative and Economic Sciences, 9.1. (2020): 242-257.
15. Liu, G., Lin, Z., Yan, S., Sun, J., Yu, Y. and Ma, Y. "Robust recovery of subspace structures by low-rank representation." IEEE Transactions on Pattern Analysis and Machine Intelligence (2012): 35 171–184.
16. Murali, Krishna Aragonda, "Life Data Analysis Reference", ReliaSoft Corporation Worldwide Headquarters 1450 South Eastside Loop Tucson, Arizona (2016): 85710-6703, USA.
17. Mustafa, Qais, and Ali, Taha Hussein. "Comparing the Box Jenkins models before and after the wavelet filtering in terms of reducing the orders with application." Journal of Concrete and Applicable Mathematics 11 (2013): 190-198.
18. Omar, Cheman, Taha Hussien Ali, and Kameran Hassn, Using Bayes weights to remedy the heterogeneity problem of random error variance in linear models, Iraqi Journal of Statistical Sciences, 17, 2, 2020, 58-67.
19. Raza, Mahdi Saber, Taha Hussein Ali, and Tara Ahmed Hassan. “Using Mixed Distribution for Gamma and Exponential to Estimate of Survival Function (Brain Stroke).” Polytechnic Journal 8.1 (2018).
20. Shahla Hani Ali, Heyam A.A.Hayawi, Nazeera Sedeek K., and Taha Hussein Ali, (2023) “Predicting the Consumer price index and inflation average for the Kurdistan Region of Iraq using a dynamic model of neural networks with time series”, The 7th International Conference of Union if Arab Statistician-Cairo, Egypt 8-9/3/2023:137-147.
21. Esraa Awni Haydier; Nasradeen Haj Salih Albarwari; Taha Hussein Ali. "The Comparison Between VAR and ARIMAX Time Series Models in Forecasting". Iraqi Journal of Statistical Sciences, 20, 2, 2023, 249-262. doi: 10.33899/iqjoss.2023.181260

Tikrit Journal of Administrative and Economic Sciences, Vol. 21, No. 69, Part (1): 310-327
Doi: www.doi.org/10.25130/tjaes.21.69.1.18

Appendix

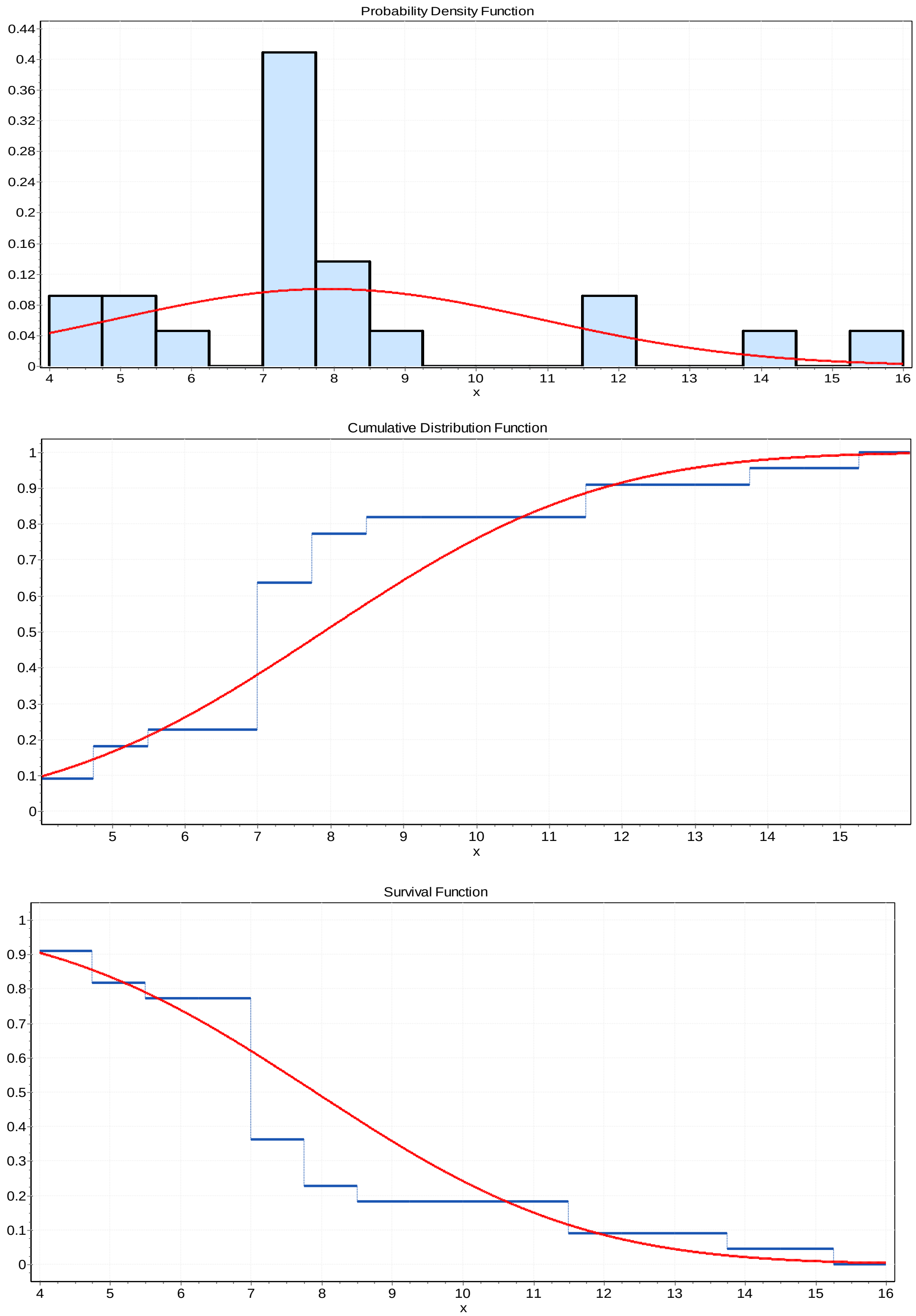

Figure (A): Pdf, Cdf, and Survival function for MLE

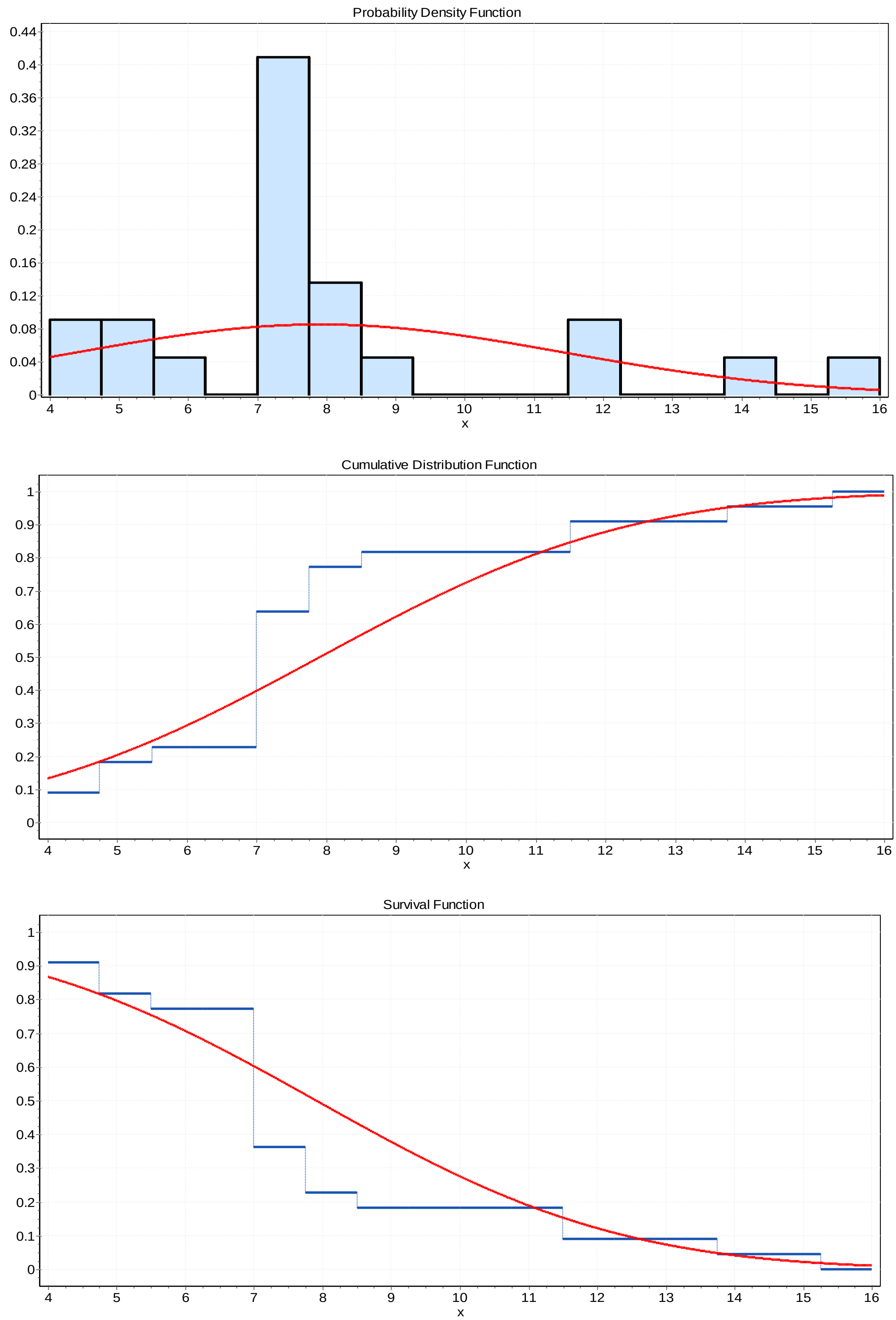

Figure. (B): Pdf, Cdf, and Survival function for RRY

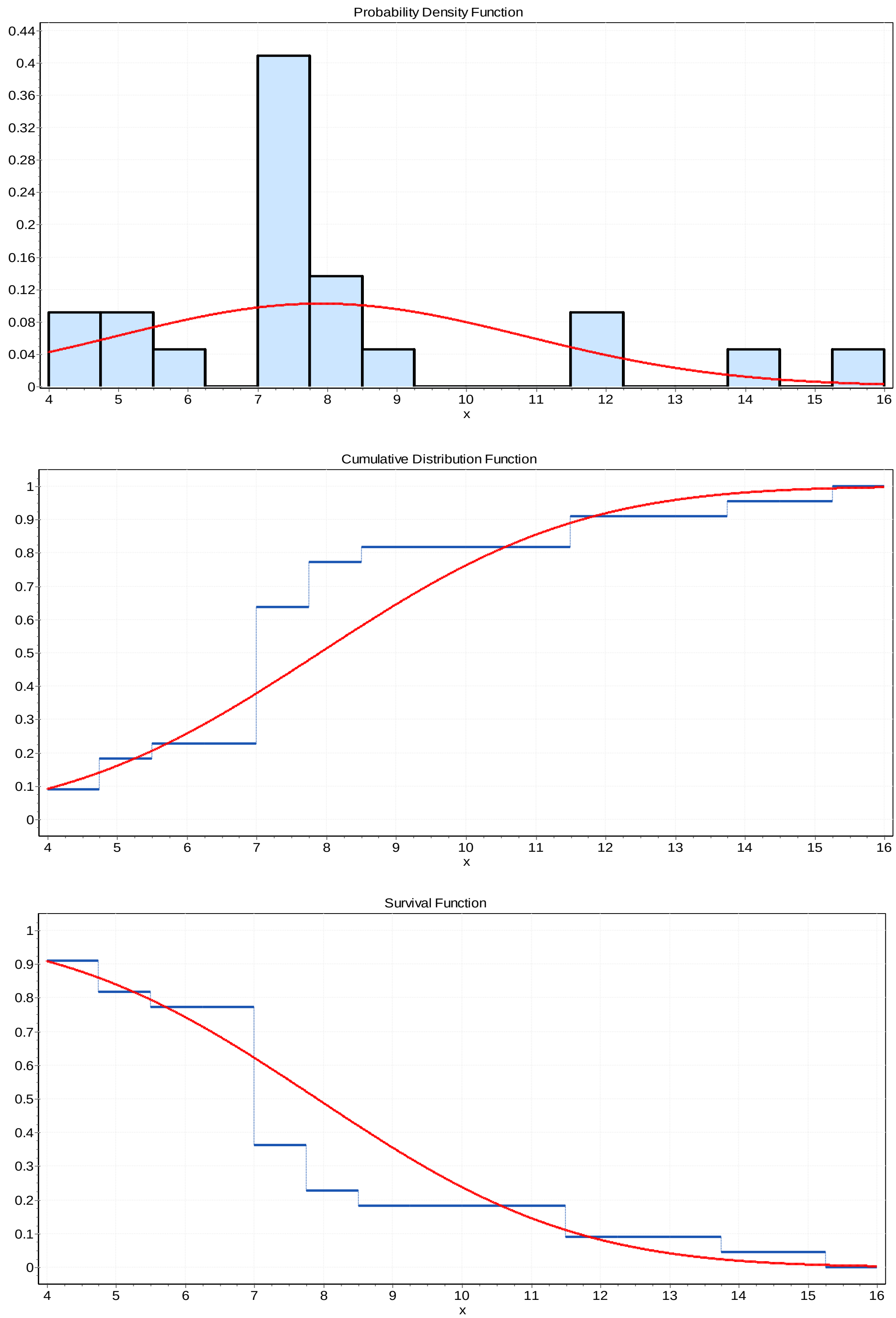

Figure. (C): Pdf, Cdf, and Survival function for RRX